

Point Space and Interface: A Holistic Approach to Search Result Visualisation

by

Kenneth R.F. Treharne, *B.IT. (Hons)*
School of Computer Science, Engineering and Mathematics,
Faculty of Science and Engineering

October, 2012

A thesis presented to the
Flinders University of South Australia
in fulfilment of the requirements for the degree of
Doctor of Philosophy

CONTENTS

<i>Abstract</i>	ix
<i>Certification</i>	xi
<i>Acknowledgements</i>	xiii
<i>1. Introduction</i>	1
1.1 Statement of Problem	1
1.2 Contributions of Thesis	2
1.3 Outline of Chapters	3
<i>2. Review of Area</i>	5
2.1 Introduction	5
2.1.1 Information Tools	6
2.1.2 Three Problems	14
2.1.3 Three Explanations	16
2.1.4 A Proposed Solution	23
2.2 Document Attribute Visualisation	32
2.3 Static and Non-Static Attribute Visualisation	37
2.3.1 The Perception of Motion	38
2.3.2 The Dimensions of Motion	39
2.3.3 Studies of Motion	42
2.3.4 A Gap in Understanding - Encoding Data with Motion Frequency	44
2.4 Natural Encoding Paradigms	46
2.4.1 The Principle of Best Foot Forward	46
2.4.2 A Gap in Understanding Semantically Motivated Encoding Paradigms	51
2.5 Visualisation of Inter-Document Relationships	51
2.5.1 The Semantic Relationship	52

2.5.2	Spatial Arrangement	53
2.5.3	Alternative Result Presentation Paradigms	54
2.5.4	The Ranked-list Result Presentation Paradigm	63
2.5.5	Perspectives on Alternative Result Presentation Paradigms	68
2.5.6	Meeting the Advantages of the Current Ranked-list Paradigm	73
2.5.7	The Absence of Guidelines for Spatially-Organised Results	77
2.6	Summary	78
3.	<i>On the Role of Motion in Attribute Visualisation</i>	81
3.1	Introduction	81
3.2	Motivation for Research	82
3.3	The Breadth of Reliable Encoding	83
3.3.1	Use of Perception Research for Encoding	85
3.4	Motion in Encoding	88
3.5	Web-Based Experimentation	92
3.5.1	Experimenter Motivations	92
3.5.2	Online Laboratories	94
3.5.3	Apparatus and Delivery Mechanisms	96
3.5.4	Data Integrity	97
3.5.5	Participation Rewards	97
3.5.6	Ethical Considerations	99
3.5.7	Data Security in Storage and Transmission	100
3.5.8	Weighing Up Web-based Experiment Methodologies	101
3.6	A Web-Based Evaluation of Motion Frequency Encoding	101
3.6.1	Graphical Features for Encoding Paradigm	102
3.6.2	Data Features for Encoding Paradigm	110
3.6.3	Method	111
3.6.4	Results	122
3.6.5	Discussion	152
3.7	Summary	161

4. <i>On the Role of Naturalness in Attribute Visualisation</i>	163
4.1 Introduction	163
4.2 Encoding Paradigms in Attribute Visualisation	165
4.2.1 Diagrams and Pictorial Representation	165
4.2.2 Semantics of Charts and Graphics	167
4.2.3 Encoding by Social Norms and Conventions	168
4.2.4 Encoding by Metaphor	169
4.2.5 Encoding by Data Type	170
4.2.6 Encoding by Perceptual-Cognitive Psychology	171
4.2.7 Encoding by Natural Encoding	175
4.3 A Web-Based Evaluation of Natural Encoding Paradigms	181
4.3.1 Method	181
4.3.2 Results	198
4.3.3 Discussion	216
4.4 Summary	224
5. <i>On the Role of Space and Interface</i>	225
5.1 Introduction	225
5.2 The Concept of Information Space	227
5.3 Formalising Information Space	228
5.4 Models of Navigation in Information Space	229
5.5 Spatial Metaphor	230
5.6 Spatial Configuration	232
5.7 Evaluation Results	233
5.8 Open Research Questions	236
5.9 Deciding on a Spatialisation Construction Approach	238
5.9.1 Document Corpus and Topics	239
5.9.2 Pre-Processing	241
5.9.3 Algorithms	242
5.9.4 Evaluation	246
5.9.5 Selected Approach for Information Space Construction	255
5.10 Summary	257

6.	<i>A Laboratory-Based Evaluation of Information Space Usability</i>	259
6.1	Introduction	259
6.2	Apparatus	260
6.2.1	Document Full-text View	260
6.2.2	Ranked-list Interface	270
6.2.3	Theme Map Interface	272
6.2.4	Theme Cloud Layout Control	290
6.2.5	Search Facilities of the Apparatus	295
6.2.6	Summary of Apparatus Design	298
6.3	Exploratory Hypotheses	299
6.3.1	Hypothesis One - Pop-up Transparency	299
6.3.2	Hypothesis Two - Document Full-text Integration	300
6.3.3	Hypothesis Three - Theme Map Layout Control	301
6.3.4	Exploratory Hypotheses - Summary	301
6.4	Method	303
6.4.1	Participants	303
6.4.2	Materials	303
6.4.3	Procedure	312
6.4.4	Design	314
6.5	Results	316
6.5.1	Analysis One: Analysis of Document Full-text View in Ranked-list Interface	316
6.5.2	Analysis Two: Analysis of Document Full-text View, Pop-up Transparency and Projection Dimension Control in Theme Map	322
6.5.3	Subjective Response	331
6.6	Discussion	333
6.6.1	Relevance Assessment Methodology	333
6.6.2	Analysis One	334
6.6.3	Analysis Two	339
6.7	Summary	344

7. <i>Discussion and Conclusions</i>	347
7.1 Introduction	347
7.2 On the Role of Motion and Natural Encoding in Attribute Visualisation	347
7.3 On the Role of Space and Interface	350
7.4 A Holistic Perspective on Search Result Visualisation	356
7.4.1 The Point: Representing an Individual Search Result	356
7.4.2 The Space: Representing Semantic Relationships	358
7.4.3 The Interface: Facilitating Interaction with Information Space .	359
7.4.4 The User	360
7.4.5 The Evaluation	361
7.5 Future Experimental Work	368
7.6 Contribution of Thesis	372
7.7 Conclusion	374
<i>Appendix</i>	375
A. <i>Motion Encoding Experiment: Instructions</i>	377
B. <i>Natural Encoding Experiment: Instructions I</i>	379
C. <i>Natural Encoding Experiment: Instructions II</i>	381
D. <i>Space and Interface Experiment: Training</i>	384
E. <i>Space and Interface Experiment: Handout for Integrated Full-text Interface</i> .	389
F. <i>Space and Interface Experiment: Handout for Modal Full-text Interface</i> . . .	391
G. <i>Space and Interface Experiment: Spatialisation Qualitative Evaluation</i>	393
H. <i>Statistical Methods</i>	406
H.1 Object Measures	406
H.1.1 Time	406
H.1.2 Accuracy	406
H.2 Statistical Techniques	410
H.2.1 Significance Testing	410
H.2.2 Multiple Comparisons	410
H.2.3 Use of Error Bars	411

<i>I. Select Publications</i>	414
<i>Bibliography</i>	416

LIST OF FIGURES

2.1	A search result visualisation taxonomy derived from Bonnel, Cotarmanac’h, and Morin (2005) and Drori (2000)	24
2.2	A survey of search result interfaces incorporating a visualisation-based component	59
3.1	The size palette	106
3.2	The orientation palette	106
3.3	The hue palette	106
3.4	The saturation palette	106
3.5	The grow palette	108
3.6	The pulse palette	108
3.7	The rotate palette	109
3.8	The shuffle palette	109
3.9	A screen shot of the motion experiment	113
3.10	The two stages of the experiment	116
3.11	Personalised performance report	117
3.12	A graph of preparation time for dimensionality	124
3.13	A graph of answer time for dimensionality	124
3.14	A graph of error for dimensionality	125
3.15	A graph of preparation time for number of motion features and dimensionality	128
3.16	A graph of answer time for number of motion features and dimensionality	128
3.17	A graph of error for number of motion features and dimensionality	129
3.18	A graph of preparation time for one dimensional trials	132
3.19	A graph of answer time for one dimensional trials	132
3.20	A graph of error for one dimensional trials	133
3.21	A graph of preparation time for two dimensional trials	136
3.22	A graph of answer time for two dimensional trials	136

3.23	A graph of error for two dimensional trials	137
3.24	A graph of preparation time for three dimensional trials	141
3.25	A graph of answer time for three dimensional trials	141
3.26	A graph of error for three dimensional trials	142
3.27	A graph of preparation time for four dimensional trials	145
3.28	A graph of answer time for four dimensional trials	146
3.29	A graph of error for four dimensional trials	146
3.30	A graph of participant drop out for experiment stage	149
4.1	A screen shot of the apparatus for naturalness experiment	189
4.2	Hue palette for naturalness experiment	190
4.3	A graph of time for task set	202
4.4	A graph of pop-ups triggered for task set	202
4.5	A graph of time for task set presentation	203
4.6	A graph of pop-ups triggered for task set presentation	203
4.7	A graph of time for task set and naturalness of encoding	204
4.8	A graph of pop-ups triggered for task set and naturalness of encoding .	204
4.9	A graph of time for task set presentation and naturalness of encoding .	205
4.10	A graph of pop-ups triggered for task set presentation and naturalness of encoding	205
4.11	A graph of time for naturalness of encoding	206
4.12	A graph of pop-ups triggered for naturalness of encoding	206
4.13	A graph of time for trial type and naturalness of encoding	207
4.14	A graph of pop-ups triggered for trial type and naturalness of encoding	207
4.15	A graph of participant drop out for experiment stage	216
4.16	A proposal for future encoding legend	219
4.17	A colour mixing guide.	219
5.1	A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Hong Kong Hand Over Task Set	253
5.2	A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Newspaper Circulation Task Set	253
5.3	A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Recycled Materials Task Set	254
5.4	A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Space Shuttle Task Set	254

6.1	A screen shot of the ranked-list interface with integrated full-text	263
6.2	A screen shot of the ranked-list interface with modal full-text	264
6.3	A screen shot of the modal full-text view	265
6.4	A screen shot of the theme map interface with integrated full-text, transparent pop-up windows and theme cloud control	266
6.5	A screen shot of the theme map interface with integrated full-text, transparent pop-up windows and theme list control	267
6.6	A screen shot of the theme map interface with modal full-text, transparent pop-up windows and theme cloud control	268
6.7	A screen shot of the theme map interface with modal full-text, non-transparent pop-up windows and theme list control	269
6.8	A screen shot of the ranked-list	271
6.9	A theme map of the <i>Recyclable Materials</i> task set	273
6.10	A theme map of the training task set	274
6.11	Colour set for document icons; colour denotes differences in cluster membership	276
6.12	Five techniques to display message text in a desktop computing application	280
6.13	Non-transparent pop-up windows	287
6.14	Transparent pop-up windows	287
6.15	The theme cloud control consisting of theme tags and descriptor tags . .	293
6.16	The theme list control consisting of horizontal theme tags and descriptor tags and vertical theme tags and descriptor tags.	296
6.17	A screen shot of the relevance judgement task interface	306
6.18	Progression of participant through training stages	313
6.19	Progression of participant through experiment stages	314
6.20	A graph of time (in seconds) for full-text integration	317
6.21	A graph of Bookmaker for full-text integration	318
6.22	A graph of the number of documents opened for full-text integration . .	319
6.23	A graph of the number of ranked-list resort actions for full-text integration	321
6.24	A graph of the average length of the re-sort vector for full-text integration	321
6.25	A graph of time (in seconds) for projection control, pop-up transparency and full-text integration	326
6.26	A graph of Bookmaker score for projection control, pop-up transparency and full-text integration	326
6.27	A graph of the number of documents opened for projection control, pop-up transparency and full-text integration	327

6.28	A graph of the number of projection dimension configurations for projection control, pop-up transparency and full-text integration	327
6.29	A graph of the proportion of trial time spent with the multi pop-up facility active for projection control, pop-up transparency and full-text integration	328
7.1	The Point, Space and Interface of Search Result Visualisation	357

LIST OF TABLES

2.1	The information seeking model adopted in this research	12
2.2	Motion dimensions and expressive capacity	40
2.3	Interpolation functions used in the production of motion	41
2.4	Intuitive and conventional encoding seen across literature	50
2.5	Feature definition and description for survey	57
2.6	A survey of search user interface systems	58
2.7	Desirable characteristics of a ranked-list interface	64
2.8	To what degree do alternative result presentation techniques meet the desirable characteristics of a ranked-list interface	76
3.1	Description of the motion under investigation	107
3.2	Parametrisation of the motion under investigation	107
3.3	Trial conditions from 4 static and 4 motion feature combinations; motion and static features unmixed	119
3.4	Trial conditions from 4 static and 4 motion feature combinations; motion and static features mixed	120
3.5	Trial conditions from 4 static and 4 motion feature combinations; motion and static features mixed	121
3.6	Trial outcome by success category	129
3.7	Success outcome for 1 dimensional trials.	131
3.8	Dependent variables by feature type for 1 dimensional trials.	131
3.9	Success outcome for 2 dimensional trials.	135
3.10	Dependent variables by feature type for 2 dimensional trials.	138
3.11	Success outcome for 3 dimensional trials.	140
3.12	Dependent variables by feature type for 3 dimensional trials.	143
3.13	Success outcome for 4 dimensional trials.	145
3.14	Dependent variables by feature type for 4 dimensional trials.	147
3.15	Subjective questionnaire data	148
3.16	Demographics and drop-out incidence	151

3.17	Preparation time and answer time for unmixed static and dynamic feature trials	152
3.18	Feature ranking based on average answer time	154
3.19	Feature ranking based on average error rate	154
4.1	A natural encoding scheme for use in metadata visualisation	176
4.2	Encoding paradigms based on approaches discussed in chapter	179
4.3	Experiment tasks for the Dog Train Security (DTS) task set	192
4.4	Experiment tasks for the Australian Music Festival (AMF) task set . . .	192
4.5	Experiment condition configuration for cluster icon shape and size . . .	193
4.6	Encoding of icon shape for word count and cardinality	196
4.7	Encoding of icon size for word count and cardinality	196
4.8	Complete cluster colour-coding scheme and interpretation	197
4.9	Objective performance measures for condition and question type	199
4.10	Accuracy over condition and question type	208
4.11	Accuracy for condition, task set and task type	209
4.12	Correct encoding recall for condition	210
4.13	Diversity and frequency of incorrect responses to recall question for condition	210
4.14	Subjective response data sorted by question type	213
4.15	Subjective response data sorted by question type	215
5.1	Topics and queries that form the basis for task sets in experiment	240
5.2	A summary of the qualitative evaluation	248
5.3	Mean Spearman Rank Coefficient for document distance ranks	255
6.1	Interactive capabilities of interface	297
6.2	Search facilities and strategies afforded by the experiment apparatus . .	298
6.3	Dependent variable outcome predictions for manipulated factor and levels in analysis one	302
6.4	Dependent variable outcome predictions for manipulated factors and levels in analysis two	302
6.5	Topics and queries that form the basis for task sets in experiment	304
6.6	Relevance score ranges for relevance calculations	308
6.7	Task set relevance ratings source from experiment population and one expert	308

6.8	Task set relevance ratings for corrected model	308
6.9	Fictitious document ratings contributing to the production of Gold Standard	308
6.10	Agreement between expert and crowd	309
6.11	Correlation calculations for expert and crowd	310
6.12	Schedule of experimental factor randomisation; full-text integration and pop-up transparency are combined in addition to randomisation of the interface presentation order	315
6.13	Descriptive statistics for document full-text view factor in analysis one.	317
6.14	Proportion of participants who did and did not utilise resorting functionality in the ranked list stage	320
6.15	Descriptive statistics for document full-text view, pop-up transparency and projection dimension control factors.	323
6.16	Descriptive statistics for document full-text view factor in analysis two.	323
6.17	Descriptive statistics for projection dimension control factor in analysis two.	323
6.18	Descriptive statistics for pop-up transparency factor in analysis two. . .	330
6.19	Descriptive statistics for pop-up transparency and document full-text view factors in analysis two.	330
6.20	Subjective workload response data for Theme Cloud control for Pop-up Transparency and Full-Text Integration factors	332
6.21	Subjective workload response data for heme list control for pop-up transparency and full-text integration factors	332
6.22	Dependent variable outcome observations for manipulated factors and levels in analysis one	336
6.23	Dependent variable outcome observations for manipulated factors and levels in analysis two	341
7.1	A comparison of online and offline methodological factors influencing search user interface evaluation	367
H.1	Objective performance and behavioural metrics in use across experimental chapters	409
H.2	Statistical techniques in use across experimental chapters	413

ABSTRACT

The research presented in this dissertation centres on the search user interface. The search user interface is the graphical user interface between where a human searcher interacts with a set of search results that a search engine serves in response to a request by the searcher.

We are accustomed to linear, ranked-list interfaces that support information search across pages upon pages of search results. However, whilst ranked-list interfaces have a number of useful and usable characteristics - that for the most part, have served our search activities well - some search is not well supported by such interfaces. Future designers should focus efforts on provisioning an appropriate level of information in appropriate forms to searchers.

Three human-based experiments are proposed and reported; each experiment tackles a different aspect of information display. Two experiments investigate ways that information can be presented in graphical form in an information visualisation tradition. In contrast, a third experiment investigates interface configuration with the intention to optimise the way textual information is presented to the user.

Together, the results form a picture of where future search interface design should move. By nature of the textual documents we search for, our interfaces must provision textual cues to the searcher. However and where possible, attributes of and relationships between documents should be expressed in graphical and spatial forms to facilitate quick and effortless comparison between documents.

Search user interfaces connect digital and cognitive worlds. It is increasingly apparent that building such interfaces necessitates a concerted, interdisciplinary effort of research and development. Accordingly, future search tools will be reliant on both an understanding of the human perceptual-cognitive system, as much as the bits and bytes that make up our search engine tools. Accordingly, perceptual-cognitive systems and phenomena have played a major role in the experimental work presented herein.

CERTIFICATION

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Signed

Dated

Kenneth R.F. Treharne

ACKNOWLEDGEMENTS

There are several people that I wish to acknowledge for their indirect though vital contribution to the development and production of this body of research. Foremost, I wish to thank my supervisors Professor David Powers and Dr. Richard Leibbrandt. Professor David Powers has bestowed on to me what it is to do research and I am most richer for this; many number of hours were devoted to a multitude of an idea's permutations and I have come to appreciate that this is a vital part of the process. Moreover, Professor Powers' diverse research interests have at long last provided direction for my own ambition and I can say with certainty that I have settled into my own area of computing where I have complete interest. Dr. Richard Leibbrandt has always been at an arm's length for professorial contribution to the task at hand. Dr. Leibbrandt's collegial and compassionate dedication to his supervision is exemplary.

The AiLab at Flinders University has been a second home for the much of the last decade and as such I would like to acknowledge in particular, Darius Pfiztner, Tom Anderson, Adham Atyabi, and Taskeshi Matsumoto. Darius, in parallel with Professor Powers, was a great source of academic inspiration early on and brought a refreshing collegial feel to the academic environment while Tommy, Adham and Takeshi were highly supportive, cooperative and engaging, and I am particularly grateful for our time together on the robotic competition. AiLab is not alone composed of the above people; accordingly, I extend my appreciation to all AiLab members for their friendly and engaging contribution to this academic life.

Finally, I wish to acknowledge friends and family, past and present who I will endeavour to reconnect with after so long withdrawn. To Tha'is, who has more or less, been there for the entirety of this thesis, thank you. Your support has taken on a multitude of forms and I am very fortunate and grateful to have had you by my side throughout this process, thank you again. To Archer, who has only been here for the finale, yet has strengthened my resolve to finish, thank you. To my parents, who have provided this opportunity, thank you. And finally, to my extended family and close friends who have maintained continued interest in my progress, thank you.

1. INTRODUCTION

1.1 *Statement of Problem*

Search facilities are increasingly prevalent in many computing applications - both online and offline. Whenever we engage in complex search activities, we are likely to endure a significant number of completely off-topic, or on-topic but irrelevant search results - among those that are actually suitable for our information need. This is a widespread problem in even contemporary search interfaces and more research should be devoted to facilitating the identification of useful results as well as indicating where similarly interesting results will be found subsequently.

When we do find a useful search result, the current methodology of presenting search results - the ranked-list - provides little guarantee that the next search result below will be relevant as well. Many information visualisation-based, alternative presentation techniques have been proposed to make it easier to identify and locate the subset of relevant documents in search result sets. Yet none has received sustained and widespread adoption. While these systems offer some unique advantage for text search, many take an overly literal interpretation of graphical visualisation of search results; subsequently, many search result presentation techniques are incompatible with contemporary search behaviours such as fast text skimming.

We need research that perceives alternative visualisation-based techniques as a specific type of information visualisation that does not engage the full set of analytical facilities that traditional information visualisation affords. More research is required to strike a balance between the visual representation of data and information, such as document metadata and inter-document relationships, and the inclusion of appropriate - but not dominating - textual information like document surrogates that searchers at present, make heavy use of in contemporary search interfaces. This thesis strongly contends that previous alternative search result presentation techniques have leaned more toward the visual aspect and have taken too much text out of the display. This imbalance is evidenced by prior systems in which it may be seen that a document shares a relationship with another document, yet it is difficult to perceive what each document is about and therefore what the semantic relationship actually is; consequently, this impacts on a decision of whether that relationship is useful to the current information need.

Furthermore and generally, search for information using a web-based search engine is not analytical, in the same sense of analytical information visualisation for multi-dimensional datasets. However, there has been an implicit assumption that if we visualise search results, somehow search for information will be improved, given the successes of visualisation-supported data analysis in other application domains. Accordingly, we need a big-picture or holistic approach to search result visualisation that takes into consideration the user's abilities and the tasks and behaviours they engage in, when searching for information.

The envisioned research approach and the approach taken in this thesis, is that of a balanced and holistic approach to search result presentation; it focuses on the point, the space and the interface. The point represents the search results and their attributes, the space represents the semantic space those points are arranged and presented within, and the interface that contains the controls and views that we use to interact and engage with search results.

Although the above perspectives are shared (Dong, 2008; Hoeber, 2012), largely, the field continues to show off visually-impressive though broadly unusable or unsustainable tools for everyday search activities - and few attempts are made by large commercial search companies to offer promising facilities to the mainstream. Therefore, the overall intention of this thesis is to support and facilitate the design of future search tools that are more effective. More effective tools will enable searchers to satisfy their information needs more efficiently and completely and furthermore, will instil greater confidence in the searcher's decision to terminate their search, having found information that satisfies their need.

However, there is no one solution to the problem of search, since our search needs are diverse in nature. Accordingly, improvements to search tools can be made in a range of areas both at the back-end of the search engine, the front-end of the search engine and through greater education of producers and consumers of digitised information. Specifically, the contributions of the present work are limited to improvement of the front-end or interface of the search engine.

1.2 Contributions of Thesis

Core aspects of search user interfaces include: the representation of individual search results, depiction of the relationships between search results, and the interactive capacities afforded to searchers. Succinctly, this is the point, space and interface of search result visualisation. Our search user interfaces contain each of these core aspects in some capacity.

Whilst there is no set formula for the design of search user interfaces, this thesis will attempt a bounding for such a formula - based on the point, space and interface

of search result visualisation. An initial outline and review goes to some lengths, to contextualise and establish each of the core aspects, while a survey of systems highlights a great diversity in prior research. Later and importantly, a series of experiments are reported, with the outcomes anticipated to partially contribute to an inner substance of the prescribed bounding.

The experiments presented in this thesis make a contribution to each core area of the search user interface. Specifically, they investigate the role of motion - vis. animation - to encode metadata attributes of individual search results; they evaluate the idea of using a user's pre-existing ideas, experiences and intuitions to define the data encoding process; and they contrast and compare the usability of user interface configurations for spatialisation-based search tools.

The reported experiments are a concerted effort toward improvement of search user interfaces in a holistic sense, such that all outcomes could feature simultaneously in a subsequent search user interface. They investigate novel but natural ways to convey information to a searcher; they investigate how to convey semantic relationships by way of visually-defined, spatial relationships; and they investigate how best to support exploration of information through usable and optimal information display and user interface design. The anticipated contributions of this research are key for future search user interface design and evaluation.

1.3 Outline of Chapters

The next chapter will establish a context for search user interface research that investigates visualisation-based techniques for the display and organisation of search results. This multifaceted discussion offers a broad snapshot of the user and system issues and includes a small survey of the various approaches evident across the literature.

The third chapter will motivate and report an experiment that investigates the use of motion-based graphical attributes to encode metadata - just as colour, shape or size can encode data. This experiment and subsequently, the experiment of chapter four are conducted online, so significant efforts are taken to explicate the benefits of performing experimental research of this kind over the Internet.

In the fourth chapter, it is argued that a data-encoding paradigm must take into account the affordances of the data. When data is encoded in a way consistent with our perceptual and cognitive expectations, the resulting interfaces should be easier and more natural to interact with, thus improving search outcomes. Later, results are reported for an experiment that explores this idea empirically.

The fifth and six chapters shift the focus away from document representation and toward inter-document relationship visualisation. Namely, the fifth chapter introduces document spatialisation and information spaces as a way to present search results to

searchers. In addition, the latter part of chapter five will define and investigate a set of key criteria to be used in the selection of algorithms for the construction of such spaces. Building on this introduction, the sixth chapter will detail and report an experiment that evaluates interface design components that are specific to search tools which feature spatially-organised search results. Three experimental hypotheses centre on the way searchers access and view document full-text, the way searchers access and view document surrogate information, and the way searchers control the layout of spatialised documents.

Finally, the seventh and final chapter summarises the findings of each experiment, reiterates the significance of the findings and reviews how each of the experimental findings will contribute to future search user interface design.

The studies presented herein are pilot studies of sorts. What is common among these studies is a tendency to be highly and broadly exploratory - in both the subject matter at the heart of the experimental hypotheses, and in the pragmatic methodological aspects, such as how to conduct research in a balanced way. From an original research perspective, the contributions of these studies are two fold; they introduce and develop new and previously unexplored research questions and ideas within the immediate field, whilst also contributing to a pool of methodological considerations made by many researchers situated within the broader human-computer interaction and usability discipline.

2. REVIEW OF AREA

2.1 *Introduction*

We indeed live in an age of information abundance; yet, despite having more data available to us, our ability to deal with it efficiently has not yet eventuated. In fact, paradoxically, despite often having all the data available that a situation demands, frequently we cannot deal with it in a timely manner to act on (Woods et al., 2002). We thus turn to tools that filter, integrate and organise information into manageable chunks and streams that relax and optimise information flow and streamline cognition.

As the volume of information increases, research suggests that we experience decreased satisfaction and confidence in our analysis of the information (Oppenheim, 1997; Schwartz, 2005). The underlying techniques that are employed to select and present information to the user are inconsistent with the way the user would operate given sufficient time to process the huge amounts of information (Russell et al., 2006). Accordingly, more effort should be directed to the development of tools that manage this situation better and in particular, tool developers should target their design toward the capabilities of the user.

Understanding the human factor in search and more widely, computer interfaces, alongside advances in the supporting technology i.e. algorithms, will contribute greatly toward better search in the future. However, at present, we have only a limited understanding of how humans interact with and extract information from an interface. We must extend our knowledge of the human in the human-computer interaction to motivate better design choices.

Engineering information retrieval - or just 'search' - is characterised as a 90/10 problem, whereby ninety percent of the problem has taken ten percent of time and effort, while the remaining ten percent will take ninety percent of time and effort (Mayer, 2008). At present, we are working on the final ten percent of the solution to make information retrieval almost perfect. Arguably, a significant portion of the remaining time spent on the search problem should be devoted to better understanding how people interact, understand and interpret data and information via the graphical user interfaces that separate the digital and cognitive worlds.

It is difficult to define today what the perfect search engine will look like in the future, given the diversity of present information needs. However, it is unlikely, at

this stage, that our information tools will become all-knowing, answering machines as depicted in science fiction. Furthermore, it is unreasonable to predict that information tools will be able to supply simple answers to complex information needs or be able to satisfy information needs where the user has a limited understanding of what they are searching for. Thus, whilst future tools will have greater access to and understanding of data and information and the ability to better interpret human language, there will remain the need for optimal display of information and interactive controls, to support and augment cognitive processes active in information processing and thus knowledge acquisition, consumption and production.

Future interface design will be founded on an understanding of the human factor, but at present, there remain many gaps in our understanding. Toward this, we should inform ourselves with an understanding of human cognitive processes and apply this knowledge in the design of new interfaces, and then evaluate those designs with respect to current state of the art in order to motivate design choices in the next iteration of information tools.

Till now, reference to an information tool has been broadly non-specific. The focus of this course of research will be on information tools that we engage for satisfaction of complex information needs, which typically feed much larger information integration tasks, such as analysis of an industry or field of research. In the following sections, a working definition of an information tool is derived for use throughout the remainder of this work.

2.1.1 Information Tools

Information Tools Characterised

In a generic sense, the tools we employ to manage our volumes of information, largely consist of the same components. Many modern information tools are Internet or Intranet-based and serve a large number of users and information needs. However, they are increasingly prominent in desktop environments where they typically serve a single user to assist with the management of personal information repositories. Nominally, we refer to these types of tools as Search Engines.

Broadly, the search engine consists of a user interface and a back-end. The back-end holds an index containing a processed form of the document collection or corpus that is available for searching. The back-end also contains the mechanisms for receiving and processing input from the user; selecting appropriate documents that match the request; and formatting results to facilitate review and judgement by the user at the interface.

The interface is the medium upon which the user communicates their information need and upon which the search engine presents its answer to the interpretation of that

need. A user's information need motivates the use of an information tool. Routinely, a need evolves with use of the tool and as the user processes more information. Such evolutions are observed in search engine logs as sequences of incrementally refined queries, whereby terms are replaced, added or subtracted though more readily replaced or added in and to the initial prescription of the information need i.e. the query (Spink, Wolfram, et al., 2001). Often, a user cannot define easily and precisely what it is they need, and so adopt various strategies to get to their desired information. Such strategies are employed to deal with the search engine's misinterpretation of the user's intent (Bates, 1979).

Information need largely influences the type of search task carried out and the types of information that the user must view to influence a successful search. The case of searching for websites is trivial; the tool should ensure the presentation of a set of web addresses and key terms corresponding to the website's branding. However, in the case of textual information spread across several independent sources, the presentation technique is non-trivial but, classically, performed by presenting a *keyhole* view of the resource's content and implying that the information satisfying the need may be somewhere within. Routinely, this paradigm is far from optimal. It is imperative that we understand how to make it easier and more efficient for the user to extract the necessary information in order to satisfy their need or to select a set of documents that contain the information that will satisfy their need.

Historically, the search engine presents results in a linear, ranked-list and places the most promising candidate at the top of the list. For each search result, the search engine provides a summary of the resource to assist the user with judging whether it will satisfy their information need. Such a paradigm serves a large variety of search use-cases, but not all - particularly when the need is complex, unbounded or abstract. It is not inconsequential to ignore these edge-cases, as these often mark the frontier of knowledge generation and discovery; arguably, these edge-cases are better served by alternative result presentation techniques.

Incorporating Alternative Search Result Presentation Techniques

Alternative search result presentation techniques are those that fundamentally differ from linear, ranked-list search result presentation techniques. There are many alternative techniques at the disposal of the interface designer and a sample of these alternatives will feature prominently in a survey discussed later in this chapter in Section 2.5.3 on page 54. The unifying characteristic that alternative techniques share is the depiction of information that ranked-list techniques do not show and which is typically depicted by predominantly graphical rather than textual means.

Whilst a number of new techniques exist, the widespread and sustained adoption of such techniques has not eventuated. Of the few evaluative studies available (C.

Chen and Y. Yu, 2000; Morse, M. Lewis, and Olsen, 2002; Chung, H. Chen, and Nunamaker, 2005; Julien, Leide, and Bouthillier, 2008; Hoeber and Yang, 2008; Hoeber and Yang, 2009; Hoeber and Liu, 2010), early indications suggest that alternative techniques provide some benefit to search outcome, although seldom do results point to an impending overthrow of the ranked-list. These results in the least, inform us that a greater understanding of how search is conducted is required to improve these tools, as to abandon the pursuit of alternative techniques altogether relegates us to our existing tools that incorporate ranked-list techniques which are sub-optimal for a range of important search use-cases. In order to define these use-cases it will be useful to establish a relevant context.

Context of Search

This section outlines a context of search that is relevant for the research reported in later chapters. This context encapsulates information need, query intent, relevance judgement and a model of information seeking behaviour. A notion of information need was earlier alluded to; however, the following subsection provides a more focused examination of this concept.

Information Need When the user has a problem and the solution is a piece of information, the user has an ‘anomalous state of knowledge’ (Belkin, Oddy, and Brooks, 1982; Veerasamy and Heikes, 1997). However, such a definition of information need does not specify how to arrive at the solution. Often, filling the gap between problem and solution is exacerbated by unknown topic knowledge, not knowing it’s ultimate breadth, and not knowing it’s ultimate depth (Kelly, 2008). Unknown topic knowledge complicates the formulation of a query in order to get to a location that will likely contain the answer. In contrast, unknown breadth and unknown depth pertain to the user not knowing how wide to cast their search i.e. generalisation, and not knowing how detailed to take their search i.e. specificity, respectively.

Beyond raw exploration of the topic domain, search user interface enhancements like query suggestion and completion, and query building interfaces like Quintura <http://www.quintura.com> see Alhenshiri, Watters, and Shepherd, 2011, Figure 1, offer various ways to deal with unknown topic knowledge. However, current search user interfaces do not adequately support estimations of coverage and detail for the user; it remains the task of the searcher to gauge whether they have covered all perspectives to sufficient detail. Conversely, alternative presentation techniques that make use of global and thematic overviews may provide an implicit way of representing the scope of the search result space.

Ultimately, motivating influences such as time pressures and analytical outcomes play a key role in the dispatch of resources devoted to generalisation and specificity.

Optimal information tools should support not just perceived satisfaction of an information need but provide cues to aid in the estimation of how adequately the information need is satisfied.

Different information needs warrant different approaches and information of different formats. A diversity of information need is reflected in studies of query intent. Intent is useful to ascertain where use of alternative interfaces may be appropriate as well as the type of information that should be made available.

Query Intent Searchers perform various types of information search and these typically fall into three categories: transactional, navigational and informational. Search type classification - alternatively, query intent classification - can take place by way of manual surveys and classification methods (Broder, 2002) or by way of automated methods that look for particular query word patterns (B. Jansen, D. Booth, and Spink, 2008). Research has attempted to gauge the spread of search intent in query logs and find that informational searches constitute a majority (Broder, 2002; B. Jansen, D. Booth, and Spink, 2008). However, for search conducted using mobile devices, informational search is the least contributory (Church, Smyth, et al., 2008) and is influenced by whether the searcher is mobile and on the go, or not (Church and Smyth, 2009). Predicting or identifying query intent can be used to tailor the search interface to the intent of the user.

Hearst (2009) suggests that these intent taxonomies do not make sufficient differentiation between *ad hoc* i.e. ephemeral and *standing* long term queries - note that while B. Jansen, D. Booth, and Spink (2008) make this distinction, they do not explicitly prescribe signals for ephemeral and standing queries in their classification algorithm. The searcher may be interested in a fact, thus encompassing a clear criterion when to terminate, or extended fact-finding, encompassing a broader investigation of longer duration and consisting of multiple facts across multiple sources (Shneiderman, 1998).

Search user interface designs, as applied by most commercial search engine companies, favour ephemeral and focused information needs; and consequently, neglect needs that encompass multiple facts, perspectives and sources or when the main need is of consensus or overview.

Regardless of intent, searchers have to make judgements regarding the likelihood that a result will contain one or several target facts within the source. The next section will outline how a user makes use of cues to decide whether to pursue a search result in greater detail, or to direct their attention at an alternative.

Relevance Judgement Satisfying information need by means of information retrieval is contingent on retrieving relevant information whilst ignoring non-relevant information (Borlund, 2005). Searchers make use of several different signals to judge relevance and

at various extents according to intent and information need (Matsuda et al., 2009; Balatsoukas and Ruthven, 2010).

Relevance is a multidimensional concept. Saracevic (1996) proposes five types of relevance: system and algorithmic, topical ‘aboutness’, pertinence or the match between information need and document content, situational relevance, and motivational or affective i.e. goal oriented.

The search engine generates its own relevance score for each document based on a mathematical function and reflects this in the ordering of the list. Ranking algorithms are opaque i.e. unclear, to the user and there is no information that shows how query words relate to the document or how each term is represented in the document (Hearst and Pedersen, 1996). Searchers heavily attend to results higher in the list (Granka, Joachims, and Gay, 2004; Nielsen, 2006) suggesting some inherent trust in the ranking by the search engine.

Judging relevance is a difficult task given that the user has a limited document ‘surrogate’ consisting of a title, a keyword-in-context snippet, a source URL and occasionally, some additional metadata. Although, a keyword-in-context snippet provides a keyhole view of the document, it does provide a significant source of information scent. Information scent is an imperfect perception of the relevance of information sources obtained via proximal cues (Pirulli and Card, 1999). In the case of search results, a distal stimulus is a relevant document and proximal cues are highlighted keywords, citation counts, or document rank. Proximal cues form a percept of relevance; the document is either relevant or irrelevant warranting further action, such as opening the document’s full text or shifting attention to a subsequent result.

The traditional ranked-list is suggestive that the higher-ranked documents are the most likely documents to contain the information that the searcher is interested in. However, rank position does not indicate the extent to which the information need will be satisfied by the information contained within. Search user interface enhancements such as TileBars (Hearst, 1995) or HotMap Hoeber and Yang, 2006 aim to show at least the relationship of query words to individual results. Later, it will be argued that alternative presentation techniques that make use of thematic cues provide an alternative form of relevance cue based on spatial relationships and an implicit organisation of results.

Balatsoukas and Ruthven (2010) identify a rich set of relevance cues that searchers make use of beyond simple topicality match. As well as topicality, other cues include quality, recency or age, format, tangibility, scope, type, affectiveness, user background, document characteristics, serendipity and ranking. Long and frequent eye-fixations for topicality and scope cues indicate that searchers do make heavy use of this type of information; however, searchers use alternative sources and in varying amounts for every search result as well. Authorship, time, and format are all observed to contribute

to relevance judgements and suggests that this metadata should be presented where possible.

The previous three subsections have outlined the notion of information need, query intent and relevance judgement. The next section considers each of these implicitly in the context of an information-seeking model, which will be called upon later in a survey of alternative result presentation techniques.

Information Seeking Model Researchers propose a number of models of search behaviour spanning the conception of a new information need and through termination of that need when the discovered information is utilised by the thought process that initially provoked the need.

Historically, models of information search are static, linear and procedural. Such models consist broadly of separate problem realisation, query formulation, evaluation and query refinement stages. However, despite having clear applicability in some situations, such models disregard behavioural detail present in not just the generic instance of search, but the more specific cases and intents of search. The notion that a searcher employs tactics (Bates, 1979) or that several related search sessions contribute to a problem solution (Bates, 1989), or the notion that a search engine might not even be consulted - instead a directory might be consulted for browsing - are glossed over by simplistic models of search.

A model of information seeking is adopted for four reasons. First, a survey of alternative result presentation techniques in a later section will seek to discuss alternative result presentation techniques according to system characteristics including the stage in the information seeking process to which the systems apply. Second, the ensuing empirical research into an alternative result presentation technique will assume a typical search tool use-case that follows a flow of information as modelled in Table 2.1 on the next page. Third, we should examine alternative presentation techniques within a realistic model of information search and for realistic result presentation problems. Forth, and finally, we should examine the provision of search support beyond that offered to searchers when formulating queries; initial stages such as query formulation and execution, influence a successful outcome and we should first consider problems that exist within contemporary interfaces, rather than problems with search prior to the advent of these interface enhancements. Accordingly, such a model assists with the identification and characterisation of the types of activities searchers would likely be engaging in, and therefore, the type of support that may be needed.

Table 2.1 on the following page presents the model of search that is assumed, in this research, to take place during the information seeking process. This model closely resembles that of (Marchionini and White, 2008), although this model is not proposed as a new model for research of this type. It simply draws together the essence of

Tab. 2.1: The information seeking model adopted in this research.

Information Seeking Model
Information Need
Tool Selection
Query Formulation
Query Execution
Exploration of Results
Review of Documents
Resolution

the standard models of search: need, query, execute, and review as well as some intermediary steps. One slight difference between this model and previously proposed models is the emphasis on the potential for backtracking possible in most stages. This ‘iterative reformulation’ is consistent with the knowledge state transformations model of Kerne and S. Smith, 2004 in which the flow of information in knowledge acquisition is not exclusively linear. In addition, the notion of cycles - repeated engagement and re-engagement of particular stages or combinations of - and feedback mechanisms are empirically observed by (Spink, 1997). Moreover, with an increasingly interactive search result interface, the *problem re-formulation* aspect of Marchionini and White’s model is overly vague since reformulation might entail re-thinking what the information need is, the choice of method in use, or the selection of query terms. Information that the user takes from the lower stages of the model, feeds back into the higher stages of the model when necessary, and thus restarting the otherwise linear process again.

The process begins with an information need - the ‘anomalous state of knowledge’. It is at this point or at some point in the future that the user will decide to rectify this anomaly in order to complete their current information-driven task. Upon making the decision to rectify the anomaly, the user then decides on their choice of tool that will allow them to satisfy their need. At this point, it is assumed the searcher has decided on a search engine as the tool of choice. The searcher must then formulate a query by articulating their information need. The searcher’s existing domain knowledge will influence this stage. If the domain of investigation is new, then this process is likely to be more difficult and the searcher may require help in devising an appropriate query. Conversely, if the domain is well known, then they may benefit by features that make query formulation more efficient. Query formulation does not necessarily have to involve words; searchers can input pictures, sounds or whole documents or website addresses and ask the search engine to identify or find related or similar articles or items.

A typical way of executing a query is by clicking a button; however, there are other ways of executing queries. Clicking links representing query suggestions, word tags, and related documents are some examples. Speaking and hand gestures on touch screen devices are additional ways of executing queries, but these are simply different modalities that achieve the same outcome. Furthermore, use of interactive controls

such as filters could be considered query building and execution, since applying filtering operations is really refining a set of results that meet a set of criteria and in combination with text querying, part of an interactive discourse between searcher and search engine.

We thus arrive at our set of search results. At this point, the searcher is looking at result surrogates as returned and presented by the search engine. In the ranked-list format, searchers typically see the document's title, a snippet containing keywords-in-context and a link to the document's full-text. They gauge the relevance of search results by way of information scent in the search result surrogates; if the scent is sufficiently high as to warrant further investigation of a specific resource, the searcher selects the resource for full-text view.

In the review stage, the searcher is looking at the document's full-text. Despite the increasing width of desktop monitors, viewing of documents usually takes place in different Internet browser windows or tabs. Notably, there is no connection between the full-text view and the result presentation technique.

Finally and hopefully, having satisfied the information need, the searcher incorporates the new piece of information into the information task that had originally initiated the need, and the anomalous state of knowledge is rectified.

This outline has assumed a linear and straightforward progression through each of the model's stages. However, frequently, information gathered from preview and full-text view, focuses the information need or suggests refinements to the searcher's query. Thus, consider potential pathways where backtracking may occur in the model; backtracking may occur at all levels with the exception of resolution. For instance, the query formulation process may indicate that the searcher's choice of tool is not appropriate and a new tool is required – the searcher's information need remains, while the searcher changes tool. Similarly, searchers will frequently backtrack between full-text view and exploration of results, while query re-formulation may occur several times as a result of exploration. Such backtracking is supportive of the - at times - non-linear nature of information search. This embodies the foraging models of (Bates, 1989) in that the searcher reviews a set of documents, but then goes back and refines the information need and starts the seeking process again, while still maintaining their intent to rectify the original information need. However, this model prescribes that the information seeking process does not necessarily have to restart; instead, it suggests that the search result set should evolve and update rather than be replaced.

The preceding sections sought to describe: how we act to rectify anomalies in states of knowledge that emerge as a result of our information driven lifestyles; how we engage information tools and direct our information needs and intent to these tools in order to have delivered a selection of resources that hopefully satisfy our need; and how we utilise cues and information scent to determine whether pursuing a suggested result will deliver a solution to our problem. Then, a model of information seeking behaviour

was described in order to systematise need, intent and relevance judgements. However, this context has assumed a straightforward transition between anomalous knowledge states through to resolution by way of an information tool. The next sections outline a series of situations where this straightforward process breaks down.

2.1.2 Three Problems

With a model of search and surrounding context in mind, this discussion will now address a series of problems and their ramifications that contemporary searchers face. These problems include information overload, information fatigue and the paradox of choice as applied to search. Clearly, not finding the information we need is an extreme case where our tools fail; however, these next sections also outline three situations that influence us in not finding the most optimal choice of information.

Information Overload

Information Overload is the point at which the volume of incoming information exceeds that which we are capable of attending to, processing, thinking about and utilising for whatever purpose and all within limited time constraints. However, in reality there are two perspectives to information overload; one is the huge availability of information, while the other relates to the consequences of that availability.

First, there is an increasingly large quantity of digital information in production. Internet search engines now index in the order of billions and spider trillions of web pages on the web, equating to a total size of the web in the billions of Gigabytes. Clearly, this is a slight embellishment of the potential overload faced by individual knowledge workers, yet the rate of cumulative growth across individual information domains means that there is more information available than ever before.

Second, information overload refers synonymously to the idea of cognitive or mental overload, but as applied to the information domain. Similarly, the same holds for the notion of data overload; though, the main differences pertain to the unprocessed, disorganised nature of raw data. For instance, consider the difference between a stream of sensor data and a set of search engine results. In either case, when the processing requirements of the incoming stimuli exceed that of a cognitive system's processing capacities, performance on the information processing task degrades.

Assessing workload, or the effort expended by the human cognitive system during task performance, is the subject of much research; (NASA, 2010) provide a rich summary of possible workload measures. However, due to the intangible and conceptual nature of cognitive processing, measurement is frequently consigned to indirect estimations, based on overall task performance, possibly in conjunction with a secondary

task paradigm, and possibly in conjunction with - to date, blunt - biological markers of workload (NASA, 2010).

Quantifying the level of workload is a necessary step toward building information tools that support and augment the limited capacities of the human cognitive system. However, the consequences of an overloaded cognition exceed that of mere task performance degradation: too much information impacts on our quality of life.

Information Fatigue

Information fatigue is a commentary on the effect that too much information has on its consumers. Information fatigue refers to the situation in which consumers are analytically paralysed by an inability to create order from large amounts of information bombarding them (Goulding, 2001). Goulding suggests that the fate of those overburdened by too much information is similar to that of those who are information poor or those that do not have access to information.

Whilst the underlying causes relate to excessive cognitive loading, information fatigue highlights the ongoing and far reaching consequences that too much information has on the consumer. Research discussed by Oppenheim (1997) is suggestive that information workers suffer degraded interpersonal relationships, degraded personal lives, poor health, loss of job satisfaction, and sleeplessness due to the sheer volume of information that they deal with in order to maximise their analysis. But, while this research observed people working in a highly information driven environment, it takes no stretch of the imagination to believe this happens periodically on smaller scales in academia, general industry and in the home.

A key factor in this appears to be the pressures and constraints placed upon workers to deliver robust analysis. Moreover, increasingly fast paced lifestyles mean that time constraints are limiting the amount of time and effort available to 'getting things done' (Poole, 2008). Our information tools, which are supposed to be making it easier to achieve these tasks, are lagging behind despite the burgeoning sources of information. We are literally, spoilt for choice when it comes to information and this has real consequences.

Paradox of Choice

The volume of information is increasingly only part of the problem; as search engines improve in terms of the sheer quantity of relevant information they can uncover, the number of potentially relevant documents for broad or popular topics will be very large, leaving users unable to optimise their decisions. Moreover, the paradox of choice as observed within a consumer context suggests that by providing many highly relevant

options in a situation where decision success is personally important, it will lead to poorer choice and degraded satisfaction (Schwartz, 2005).

The paradox of choice is typified by three phases. The first phase is characterised by inflated expectations and attraction to a large number of options. Paralysis of comparison follows, thus leading to suboptimal decision making, since the time taken to make pair wise comparisons is greater with more items. Then finally, the perceived disparity between opportunity and benefit elicits dissatisfaction and regret when it comes to evaluating the outcome.

Oulasvirta, Hukkinen, and Schwartz (2009) extend and apply this hypothesis to the information domain and find that it holds true. In their study, searchers self-rated higher satisfaction and confidence on task performance when using a search engine displaying six results per page in comparison to a display containing 24 results per page. The information tool in this case was a search engine which displayed search results in a ranked-list format. Participants are relegated to the information scent of each search result surrogate on which to base decisions about proceeding with a weighty full-text view of a document that may or may not contain the information they are after.

Thus, this research implies that our information tools do not support effective comparison of multiple and potentially useful alternatives, which leads us to sub-optimal selection or forced disregard for information. Reducing a large set of results down into a smaller set, in order to make it easier to make pair-wise comparisons, does not cater appropriately for complex informational search needs. It will be more beneficial to build tools that organise and structure results to facilitate comparison and facilitate exclusion of alternatives known to be irrelevant to the task.

But before we consider further why information tools struggle to support and augment cognitive processes, we should first consider some facets of communication and language that modern day tools struggle to deal with.

2.1.3 Three Explanations

Search engines support a range of search needs and user intent, but not all information need is met with a satisfactory or efficiently-achieved outcome. The next three sections outline reasons why in the current situation, we should not expect satisfactory outcomes in all cases. These reasons encompass the diversity and ambiguity of language, the inconsistencies between virtual and real worlds, and the inadequate presentation of information.

Diversity and Ambiguity of Word Usage

Conceptually, two measures of search result quality are recall which is the proportion of relevant documents selected from an a priori relevant set across the whole corpus, and precision which is the ratio of relevant and irrelevant documents in the result set. In an ideal situation, we want to maximise each of recall and precision, since we would like our result list to contain only relevant documents and to contain all of the relevant documents from the corpus.

In practise, search engines employ query expansion techniques by way of stemming or synonym thesauri to increase the likelihood that relevant documents will be returned but which are not described in terms provided by the searcher. This reflects the Vocabulary Problem in language use (Furnas et al., 1987) whereby users will draw on a range of words to describe the same concept. Theoretically, use of expansion techniques has an impact on both recall and precision in that potentially more relevant documents are returned even though they do not contain the original query terms; however the precision of the result set is likely to decrease due to polysemous words having different interpretations across different contexts.

The vocabulary problem influences the ambiguity of queries as well. The limited expressiveness of user queries - average query length around 2-3 terms - and lack of query operator use Spink and B.J. Jansen, 2006 contribute to the search engine's inability to interpret the precise information need of the searcher. In further support of this point, machine-learning research by Song et al. (2009) estimated that sixteen percent of queries in a sample real search engine log were ambiguous. Whilst B. Jansen, D. Booth, and Spink (2008) estimated that around twenty percent of queries had a vague or multifaceted interpretation i.e. those queries that cannot be assigned a single category of intent with confidence. If one considers the conservative figure of 10% as the real proportion of ambiguous queries that commercial search engine companies serve every day, then it is the case that web searchers are likely to notice confusing results as a matter of routine. Query diversification algorithms are proposed (e.g. Agrawal et al., 2009) to deal with ambiguous or underspecified queries by interleaving relevant but categorically diverse results throughout the list of results, in anticipation that the first page of results will reveal a keyword that can be used to reformulate the query.

Historically, building information tools to cope with ambiguity and diversity in word usage has focused primarily on mathematical or statistical models of language to predict the most similar documents. Such models remain imperfect and 'not human' meaning that they cannot produce comparable results to that by a human completing the same task given sufficient time.

Inconsistencies Between Information Processing of Humans and Computers

Historically, information retrieval has benefited from mathematically driven technologies: clustering, categorisation, and language modelling to name a few. All approaches perform fairly well in contrived scenarios but as we apply these approaches in the linguistic wilderness, most can only achieve a mediocre level of performance. While at this point, our simplistic text processing algorithmic techniques can exceed the processing rate of the human, they cannot achieve a proficiency or precision comparable to a human with sufficient time.

Pfitzner (2009) shows that highly weighted words calculated by TF-IDF - used to weight the importance of document terms - have little or no overlap with words selected by users in a document description task. Moreover, both human descriptive words and algorithmically derived words are different again to words that a user would use to query for the same documents. One way of calculating the relevance of a document to a search query involves considering a query word's frequency in a document relative to its frequency across the whole corpus, and then measuring the distance between document vectors and the query vector in a hyper-dimensional space. A human might apply a similar process, perhaps without strong conscious awareness, and look for combinations of unique or descriptive words in order to judge the similarity between several documents. The human does this with greater fineness; humans can more readily intertwine contextual and personal factors into their comparison, in order to achieve a sound outcome.

Russell et al. (2006) look at the paper sorting behaviour of human subjects when working with large document collections. They compare the physical sorting and retrieval behaviours against analogue behaviours in two electronic document management applications and find various trade-offs in play. The time to organise and access documents in electronic collections is overly long, in comparison to organisation and access of documents using tangible paper-based organisations. However, while the human can organise a collection of documents for quick access this organisation can take place over a small collection only. In contrast, the electronic manager can organise a great deal of information in limited time but it takes longer for the user to access that information as the organisation and the interaction with the electronic organisation and electronic organiser is inconsistent with how the human performs the comparable task in the physical world.

Cockburn and McKenzie (2001) demonstrate that interface designs that attempt to replicate real life in a virtual interface do not necessarily result in the expected benefits. Three-dimensional virtual interfaces do not facilitate optimal interaction with information despite the fact that we successfully interact with physical forms of information in everyday life.

Research comparing human and computer information processes, reveals interesting

insight on which to base modifications to existing paradigms. The combination of a predominantly mathematical approach and the human factor in dealing with our information glut has made for significant advances; however, research also demonstrates the paucity of some of the core underlying mathematical paradigms that are inconsistent with that of human behaviour.

The aforementioned examples do not convincingly support the idea that under performing information tools are wholly a direct result of ignoring the human factor. However, they do suggest that there are inconsistencies between the physical world and the virtual world and that these inconsistencies should inspire research that considers the human factor as important.

The complexity of modern day retrieval algorithms are testament to the integration of both mathematical and human factor inspired knowledge. Significant successes are evident in utilising attribution as a signal for relevance in information retrieval algorithms. Traditionally, this has translated to interpreting hyper link pattern structures across the web with the assumption that credible and reliable information may be available at ‘authorities’ frequently referred to by other websites on the Internet. More recently, these include analysis of social media attribution and social media participation by various sharing methods on social media websites.

Having evolved from a time when computing was more about bits and bytes and little about the affective user, the field of computing has increasingly diversified into systems research as well as human factors research after it became increasingly obvious that we could extend and improve on earlier work through a better understanding of the user and of the cognitive processes involved in human-computer interaction as exemplified in Carroll (1997), Hartson (1998), and Proctor and Vu (2006). Consequently, greater involvement of the user has warranted further investigation into better interfaces through which the user and computer communicate.

Inadequate Information Architecture and Presentation

The search user interface is a two-way channel; it facilitates a conversation between the searcher and the information tool: the searcher describes what they want and the tool prescribes what it calculates as the answer. For the most part, search engine interfaces convey this message or structure results in such a way that the searcher can get their desired information from the search result interface itself or from results within the first few positions of a ranked-list. However, when search becomes harder, when the topic of interest is not as prevalent within results, when the search outcome is reliant on several results or when the query has several interpretations, current information presentation and interaction techniques are inadequate.

The process of getting information to the user in any specific search scenario represents such a multidisciplinary mix of research and engineering that it is difficult to claim

that the way we represent information is the major unsolved problem in modern day information tools. It is however, a major part of perceiving and working with data in general and so without carefully engineered information presentation we would be worse off. In fact, we would relegate ourselves to reading large unstructured pages of text. Without ongoing research into information presentation, we may meet a comparable fate in the future as information volumes increase.

There are two perspectives on the problem of inadequate information presentation. The first perspective is that inadequate presentation is simply one aspect of the search tool problem. The second perspective is that the problem is acute in some areas of search - namely search verticals or search for a particular type of information - and that the future for open-domain search will simply be a refinement or variation on what we already have.

In the first instance, we can clearly see that information retrieval tools necessitate a blend of engineering feats and that currently unsolved and undiscovered problems exist in areas such as index construction and query processing which contribute directly or indirectly to the amount of time it takes to respond to a searcher's query. If information presentation was well understood but the underlying retrieval infrastructure was poor then information search tasks would ultimately fail. On the other end of the extreme, if we had perfect infrastructure and poor presentation, then perfect information would arrive to the user but with little assistance with finding the significant information in the search result set.

Woods et al. (2002) suggest that the main problems in data overload are high volume, high workload and being unable to recognise significance in data. They suggest that the possible solutions are filters, automation and better presentation; but even these taken together pose interesting consequences. They observe that with greater automation the user is increasingly confused by the current system state and left asking why or if something has taken place without explicit interaction on the part of the user. Woods et al. propose that solutions should entail context-sensitive and structured approaches that match the assumed context-sensitive and structured aspects of cognition. Yet, while their work is set in the context of situation awareness, parallels exist between these interfaces supporting situation awareness and those supporting information retrieval.

Also from this perspective, successful outcomes are dependent on optimizing speed, document understanding and information scent in the result presentation. The likelihood of the searcher arriving at their answer in a timely manner is greater if the search engine responds quickly, is able to rank the best results toward the first rank positions and provisions search result surrogates with sufficient cues to engage a percept of relevance. However, we should not have to accept that this paradigm is the only way to interact with search results, particularly for tasks that involve broad information needs, such as when novelty or serendipity is valued, or when the searcher could benefit from

an organisation of results that is facilitative of navigation through the result set.

Ranked-lists provide options and in some situations, those options are purposely diversified to maximise the likelihood that the searcher will arrive at their intended destination having provided an ambiguous query (Agrawal et al., 2009). However, having considered an irrelevant option offered by the search engine we should not have to consider further options of a similar nature further down the list - if it is not desirable to do so. Moreover, given that measures of attention plummet with increasing rank, the net effort of attending to irrelevant documents or partially relevant documents - yet not relevant enough to warrant further examination - is more overwhelming, and we seemingly become less resistant to noise with as we descend downwards, in comparison to noise higher up the ranked-list. At this point, offering a refinement of the query may assist the search engine with the delivery of a more precise result set; but the availability of a query refinement facility alone, offers no guarantee that the searcher is confident in their prescription of refinements. There are subtle anxieties introduced when there is a great diversity in the result list and when the searcher is forced to make a query refinement for a vaguely specified information need. Moreover, anxieties are also introduced when there are many partially relevant documents and no way of connecting documents rated as relevant at earlier positions in the list, to documents further down the list but obscured by an increasingly noisy signal. These two situations at least lead to outcomes that are symptomatic of the analytical paralyses proposed by (Goulding, 2001). Better information presentation paradigms, particularly those that depict semantic relationships between items may alleviate these situations.

In the second instance, (Hearst, 2006) expects that vertical search domains, or search about a specific class of information like automotive, travel or music search, stand to gain the most from advances in the search user interface, in comparison to search engines on the ‘open-domain’ web that serve vanilla search needs. Typically, vertical searchers rely on quality metadata to narrow a set of results down into a manageable size to find items that exactly match the search criteria. Metadata taxonomies and ontologies are well defined or easily definable for vertical domains, as the search space has well defined scope. In contrast, the scope of open domain taxonomies and ontologies has theoretically no bound and consequently are potentially granular, difficult to manage and maintain, and incomplete due to the sheer nature and volume of information.

Arguably, vertical search interfaces are responsible for much of the evangelism that surrounds search interfaces. There are many routinely cited and successful examples that propose and explore interactive search and browsing interfaces in vertical search domains for example Newsmap <http://www.newsmap.jp> and systems based on the ideas of Flamenco (Yee et al., 2003). In contrast, alternative information presentation paradigms for open domain search, whilst having received some research attention appear unable to stand the test of time. For instance, while Kartoo and Grokker (see

Koshman, 2006) were once the darlings of the visual search engine movement they have since closed for business. In addition, while there are interesting visual search engines in existence today, there are few if any that are comparable in the approach taken by Kartoo and Grokker.

The approach of Kartoo and systems of a similar nature that organise search results into a thematic space using a spatialisation algorithm, is not particularly wrong. Though in their final form, they were incompatible with generic search behaviours that we employ on ranked-list interfaces; this perhaps explains why successors have not eventuated. These systems have incorporated information visualisation too literally in their design and offer too little support for conventional search behaviours - and in particular, support for scanning of multiple text surrogates.

Empirical research suggests that presently, we cannot produce alternative result presentation techniques that greatly outperform ranked-list based interfaces for open-domain search (C. Chen and Y. Yu, 2000; Julien, Leide, and Bouthillier, 2008). Commercial search engines have mostly avoided the use of the experimental interface paradigms - instead choosing to optimise on other aspects such as speed of service, search enhancements (White, Jose, and Ruthven, 2003; Haas et al., 2011; Sandvig and Bajwa, 2011), and a focus on supporting searchers in completing simple but vital every day search tasks (Poole, 2008). These enhancements have involved predominantly textual evolutions to the ten-blue-link search result paradigm including query completion and suggestion, result pre-fetching, related search suggestions, deep search site links, and rich multimedia (Haas et al., 2011; Sandvig and Bajwa, 2011). These appreciated and vital-but-everyday search tasks encapsulate a large percentage of the search traffic in which users typically want to find the location of something, get a price of a stock, download a file, buy or get a review of a product or service, find a movie time, look up a recipe, find a specific fact, or catch up on their favourite celebrity or current event. For the most part, these searches are well served by the ranked-list paradigm and search engine companies can easily monetise these types of search through advertisements. Alternative result presentation techniques should play no role in supporting search of this nature. Rather alternative result presentation techniques should be a structural template that appears for particular search when intent detection indicates that searchers may benefit from an alternative presentation paradigm typically those that involve intense search over an unknown domain, or when novelty, serendipity, or consensus is valued.

Proposed solutions should not have to stand-alone, as many alternative result presentation techniques have done, nor should they be expected to wholly replace current result paradigms that already serve well the plethora of everyday-but-vital search tasks. There is no one single panacea to better information tools but a series of solutions that match the needs and intent of the searcher. For each need and intent, a suitable solution is one that considers the many facets of the whole system of humans interacting

with their information. This includes perspectives on the information, the human and humans interacting with information at the interface, the tasks and uses of information, the knowledge already learned and built upon, and the environment in which the search takes place. The main standpoint of this thesis contends that the human factor plays a great role in the future of the design of tools that deal with information overload into the future, and that understanding the human is paramount to understanding universal principles or at least guidelines for effective information presentation. This is not an innovative insight; rather, it is more part of an ongoing movement with a long history (Bates, 1979; Marchionini and Shneiderman, 1988; Pirolli and Card, 1999; Balatsoukas, O'Brien, and A. Morris, 2010) which has experienced improvement through iterative contribution.

2.1.4 A Proposed Solution

A proposed solution is to make improvements to search interfaces that are intended for use in scenarios where search is difficult - as outlined in the section prior. Such tools employ techniques that organise documents into semantic spaces, which depict relationships between documents and which have historically been unsupportive of conventional searching behaviours. This solution lies in the pursuit of an interface that supports a range of local and global information scent that the current ranked-list paradigm does not offer.

Landauer, Laham, and Derr (2004) appropriately temper this approach in their observation that decades of research have failed to isolate predominantly visualisation-based approaches for the representation of verbal information; but that the problem may lie in our lack of understanding of the human user - a direct consequence of, historically, a general disregard for human-based evaluation. Better understanding of information presentation is vital and will not be made redundant in the future despite the expected advances in artificial intelligence, as rarely will a sound bite or single sentence answer be wholly satisfactory for all forms of information need. Often the better answer is in the presentation and not exclusively a textual description.

For some information, an understanding and the motivation to understand is augmented and benefited by displaying that information in context, interactively, and in representations true to the underlying domain. This will also apply to future tools in which search engines could conceivably take on the role of teacher, using visualisation to explain concepts in real time (Hemmje, Kunkel, and Willet, 1994).

Further discussion of the solution will be divided into three distinct sections. In the first section, a common framework of visual perception is established - this will be relevant in subsequent sections. Then in the second section, discussion will focus on the representation of individual results. Finally, in the third section, discussion will focus on the representation of the result set as a whole and the depiction of relationships between

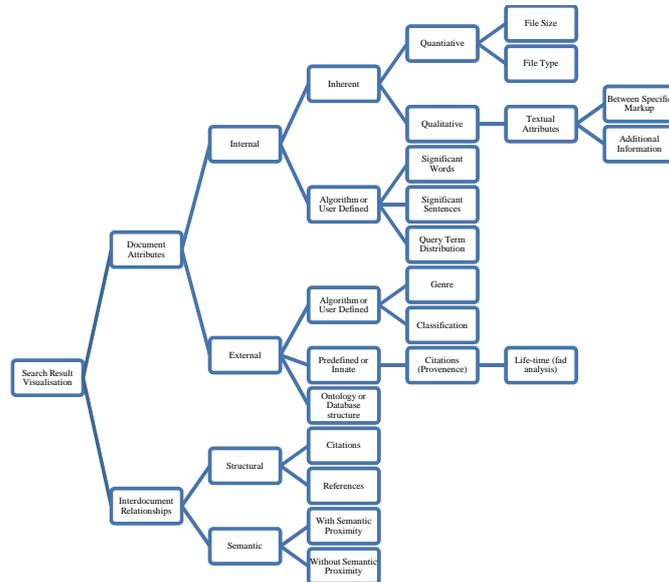


Fig. 2.1: A search result visualisation taxonomy derived from Bonnel, Cotarmanac'h, and Morin (2005) and Drori (2000).

individual results. This outlay closely conforms to the two main distinctions of search result visualisation - the visualisation of inter-document links and the visualisation of document properties (see Bonnel, Cotarmanac'h, and Morin, 2005) - and the set of search result display techniques of Drori (2000). Figure 2.1 depicts an integration of both works.

In the first section, the representation of document properties will take prominence. Two particular aspects to attribute representation are of interest. The first is the use of motion in representation while the second relates to an information-encoding paradigm based on the notion of ‘natural fit’ which may have previously been identified as encoding by intuition. In the second section, the focus will be on the visualisation of inter-document, or between-document, similarity using semantic proximity and the various ways we can visualise this. This outlay will also set the progression of later chapters, which will provide a basis upon which to report results on human factors and usability experiments that provision quantifiable results in support of the improvements that are needed in future search tools. Furthermore, while qualitative results are indeed important and certainly studies rely upon on qualitative assessment to shine further light on objective assessment (e.g. Rivadeneira and Bederson, 2003). The usability studies in latter chapters do invite participants to comment, stipulate and demonstrate their knowledge of, opinion of and preference for search user interface components under investigation.

A Holistic Approach to Solution

This research adopts a holistic approach to the problem of information presentation. The sum of each part's value to the paradigm exceeds that of the value of individual parts alone. An alternative approach is to reduce such systems down to fundamental and atomic parts as is done in the study of perception and to extrapolate the benefit in a wider system. Although in doing so, reductionism removes any essence of context of the representation, which may be important for the interpretation of those parts.

Moreover, it is not the aim of the present research to formulate psychologically pure experiments; and furthermore, this research takes notice of Kunar and Watson (2011) who suggest that in ecologically valid studies, literal interpretation of such pure theory may not hold. Nevertheless, it is intended that this research remain informed by perceptual and cognitive psychology, when investigating the application of theoretical knowledge to information tools in order to overcome practical challenges of building effective and efficient information tools.

Moreover again, context is paramount, as will be highlighted in the later discussion on natural data encoding paradigms. The information scent in information presentation takes many forms and can be guided by the current task demands. Consequently, the successful completion of an information task may rely heavily on multiple sources of information depending on the current strategy employed by the searcher.

Taking this standpoint of holism, as many have previously and evidenced by an abundance of systems employing alternative result paradigms, such a standpoint entails a particular risk of producing more of the same. Such a situation may be characterised by under performing systems due in part to sub-standard or experimental algorithms for keyword extraction or clustering, unmotivated information mapping paradigms leading to confusion in users, subjective results that favour novelty over functionality, and insignificant differences among competing presentation paradigms due in part to the aforementioned characteristics. Avoidance measures against more-of-the-same should focus on precisely defining the area of search the system applies to, the role of the systems in relation to everyday-but-vital search, if any, and a little scepticism of alternative approaches to information presentation in which proponents of such systems have shown overly inflated expectations and shallow motivations for visualisation-based interfaces for information retrieval.

The next section will make clear the importance of the human perceptual and cognitive systems. Inherently, information presentation paradigms are contingent on how we process rich visual information.

Modality and In-Modality Modalities

As was stated earlier, the interface is a two-way communication between a digital world and a human agent. This communication takes place via an increasingly rich multi-modal language of visual, auditory and tactile inputs and outputs. The widening of communication modalities is a means to support reinforcement of a message by redundancy, to spread system output over multiple channels as a way to manage cognitive load and avoid channel saturation, to manage attention to ensure the message is attended to, and to make it easier to convey a message in natural prose or by using rich gestural languages. This research will restrict itself to the visual modality only.

Each sensory channel has an operational capacity governed by cognitive resources devoted at processing time. Spreading or multiplexing information within and over sensory channels has efficiency gains to a point and is prominent in the early work on cognitive processing capacities as explored by Miller (1956). Capitalising on this idea, the field of visual analytics has concerned itself with building very highly dimensional visualisations that rely primarily on the visual modality to support understanding and decision-making on hyper-dimensional datasets, by combining multiple visual attributes to form complex iconic representations to reveal patterns and trends in datasets. The strongest distinguishing factor between visual analytics and traditional information visualisation is whether the tool supports data analysis capabilities that may be cast directly over the data and where a hypothesis about the data cannot be confirmed without a combination of automatic data analysis and human cognitive inputs such as background knowledge, intuition and decision-making (Keim et al., 2008). Moreover, an audience consulting traditional information visualisation has no hypothesis; rather they engage such a tool to facilitate insight and discovery which they could not otherwise achieve through inspection of the raw data. Such insights are dependent on the perception of patterns in the display and the capacity to establish what those patterns mean.

Data does not necessarily have to be nor is it regularly large, complex and hyper-dimensional (Chabot, 2009); yet, this is no reason to ignore an understanding of how information should be presented. The call for more understanding into the way information is presented is echoed by many proponents in the literature (e.g. Tory and Möller, 2004; Robertson, Czerwinski, et al., 2009) and specifically how information attributes are mapped to visual display dimensions (Nowell, Schulman, and Hix, 2002).

Studying perceptual-cognitive tasks in their most basic forms as a means to improve visualisation has been a priority area for information visualisation for quite some time (C. Chen, 2005). According to C. Chen, understanding what is perceptually salient versus what is cognitively salient is a key challenge for visualisation. Forsell and Johansson (2010) compress 63 previously published evaluation heuristics for information visualisation down to a set of ten heuristics that account for the majority of usability

issues in their test sample. They note that information coding consistently topped their compressed list of heuristics in all cases of analysis. This highlights how perception of information is tied to the way in which data and information is mapped or transformed to visualisation primitives and that getting it wrong will bring about usability problems. The next section will outline a framework in which to consider the role of perceptual and cognitive processes involved in the decoding of visually encoded or mapped information.

The Visual Expression Process

Before going further into the more focused aspects of this thesis, it is useful to outline an overall picture of the human perceptual system because visual perceptual processes are essential to recognising, decoding and interpreting the visually rich information in alternative result presentation paradigms. In information visualisation, we refer mostly to the perception of vision, but given the increasing multi-modality of interfaces, such consideration could be made for the other senses as well.

A general model of the perceptual system by itself does not readily lead to improvements in the design of information tools. Rather, in general, some individual principle taken from the literature provides the basis of a new direction or technique. The outcome of applying principles and knowledge of perception will be to build interfaces that show interesting patterns immediately and efficiently to the user without the user employing significant cognitive resources (Pickett and Grinstein, 1988). However, we should always contextualise that individual principle within the wider system. This section will rely mainly on the Visual Expression Process of Rodrigues et al. (2007) in order to address the relevant perceptual considerations that apply to alternative result presentation paradigms.

Ware (2004) writes that a general model of - visual - perceptual processing follows three stages: low-level extraction, then pattern perception and finally goal-directed processing. As stimuli pass through each stage, the brain processes, transforms and integrates the stimuli in a cumulative fashion and in parallel. Rodrigues et al. (2007) take this further and propose a practical and applied framework for use in relation to visualisation tools, whilst remaining informed by a general model of visual perception. They propose that acquisition of knowledge from a visualisation arrives out of three successive cognitive tasks that roughly match the stages outlined by Ware: conception, observation and reasoning.

In conception, pre-attentive features such as position, shape and colour are processed very quickly and without focused attention; here pre-attentive features are evident through mere observation but meaning does not necessarily follow. Next, in observation, the human perceives emergent features because of pre-attentive features in the conception stage. If emergent features are imperceptible, then the human can-

not reason with the visualisation. Finally, in reasoning, the user applies their domain knowledge to the perceptions in order to interpret and draw conclusions from the data. In contrast, the more sophisticated aspects of Ware's goal-directed stage, interface with higher cognitive processes such as decision-making and learning; however, there is no clear boundary between the two since both interact with each other.

It should be noted that despite an alignment with earlier taxonomic work (see Rodrigues et al., 2007, Table 1), there is no connection made to Norman's information design and gulf of evaluation (Norman, 2002) and such a connection could be made¹. Earlier work has suggested that Norman's gulfs and stages are starting points for any definition of an information visualisation taxonomic framework (Pfitzner, Hobbs, and Powers, 2001) and indeed, the three stages: conception, observation and reasoning, share resemblances to Norman's: perception, interpretation and making sense.

Conception In the first stage, conception, pre-attentive features are processed very quickly and without explicit attention. These features altogether form the visualisation; if these features cannot be detected the visualisation cannot be perceived (Rodrigues et al., 2007). This section outlines how this process takes place by way of perceptual psychology experiments that aim to understand the human perceptual system.

Visual search tasks are 'those where one looks for something' (Wolfe, 1998) and we perform these elementary tasks when interacting with information visualisations; routinely, we look for patterns in visually encoded data. In visual search experiments, human participants scan a display of visual stimuli for a single specific target amongst distractors. In some proportion of trials, the target is present, while in the other proportion of trials the target is absent. The participant's key press response is one of target 'is present' or 'not present' and the time taken to make this response is recorded. The number of visual stimuli on the display is a manipulated variable. With an increasing number of alternatives, finding the target should get harder and overall task efficiency should decrease.

An analysis of efficiency involves plotting average time against the number of visual stimuli termed the 'set size'. The slope of the graphed line indicates the efficiency of the visual search task; if the slope remains relatively flat for increasing set size the visual search task is considered efficient, in contrast, a steep slope indicates that the search is inefficient, as it takes longer to find a target amongst an increasing number of alternatives.

There are two types of visual search: feature search and conjunction search. In the former, search is for a unique target based on a uniquely identifying basic feature: classically, finding a red target amongst blue distractors. Efficient feature search generally coincides with the surprise of a unique feature popping out at the human

¹ I thank and acknowledge one of my examiners for raising this observation.

participant. In the latter, search is for a target consisting of a unique combination of two or more features: classically, finding a red circle target amongst red square and blue circle distractors. In conjunction search, while neither feature is enough to differentiate it from the distractors, the combination of features will differentiate the target from distractors.

There are four incontrovertible basic features: colour, motion, orientation, size, and several other features where the support is not as strong but of these others, only shape is relevant for this research (Wolfe and Horowitz, 2004). In feature search for basic features, response time is roughly constant for increasing set size and is thus considered efficient. In contrast, search for visual feature conjunctions can be either efficient, inefficient or somewhere in between. Historically, visual features were classified in a binary fashion - either efficient or inefficient. Efficient search was considered to take place in parallel and without explicit attention while inefficient search was considered to take place serially, with the searcher having to explicitly shift attention between alternatives until the target was found or the decision to terminate search was made. In contrast, Wolfe (1998) argues a theory of visual search incorporating a ‘continuum’ of visual search efficiency, along which situate graphical features and their combinations.

A robust theory of visual search still does not exist. All current theories are wrong in that no one theory can account for each of the observed experimental phenomena (Wolfe, 2007). These phenomena include: increasing time and difficulty for increasing set size, longer time on target absent trials, more efficient search with target-distractor dissimilarity, efficient search for distractor-distractor similarity, efficient search for linearly separable target and distractors, asymmetries where search for A among B is harder than B among A, efficient search for categorical uniqueness and finally, guidance of attention (Wolfe, 2007). In absence of a robust theory of visual search, we can at least draw on some of these initial phenomena to derive tentative encoding guidelines for information visualisation.

An obvious application of the feature search phenomena is incorporating the popping out perception into the design of graphical user interfaces to orient the human to areas of interest as quickly as possible using perhaps colour or motion. This assumes that either the system can determine from the data the most worthy-of-attention information or alternatively, can expect to receive from the user some indication of what they would like their attention drawn to. However, Rosenholtz et al. (2005) point out that for moderately complex interfaces it is unlikely that the user will experience the pop out effect due to the heterogeneity of most interfaces.

Observation In the second stage, observation, the user perceives compositional perceptions comprising the pre-attentive stimuli that were processed in the earlier stage. These perceptions include correspondence, differentiation, connectivity, arrangement and meaning and are observable without an understanding of the data (Rodrigues et

al., 2007). These observations might be thought of as preparing cognition to interpret the visualisation; perceiving emergent patterns of arrangement and connectedness, and assigning significance to patterns based on mental models for visualisation usage. Significant patterns will be interpreted relative to the actual task context in the next stage.

For differentiation, each pre-attentive stimulus type, colour, position and shape in the visualisation is perceived as an individual graphical item. For correspondence, a basis to interpret each graphical item is established between each graphical item and its referential map e.g. position is interpreted by spatial axes and colour is interpreted using a colour encoding legend. At this point, no decoding occurs; this process simply sets up a basis for decoding. For connectivity, an edge or line between two items establishes a basis for connection or relationship. For arrangement, the user perceives emergent Gestalts (Green, 1997): proximity, similarity, common enclosed region, connectedness, and figure-ground. Other Gestalts exist including: contiguity, closure, and symmetry, but a focus on the former is taken, as these more readily apply to the conception of the search presentation paradigm presented in this research and any influence of the latter is coincidental. Finally, for meaning, graphical items and patterns are perceived as significant if mental models in long-term memory provide a basis for significance; for example, prior knowledge suggests that an overly extreme or outlying value is significant and such exceptions may be important to detect. However, at this point, such meaning has no interpretation.

For this research, the perceptual Gestalts of proximity, similarity, common enclosed region, connectedness and figure-ground have direct relevance to the envisioned alternative result presentation paradigms and so further emphasis is forthcoming. The first examinations of the Gestalt were introduced by Wertheimer (Green, 1997) and are readily identified with the Berliner Schule (Horn, 2007). In this discussion of the relevant applicable Gestalt principles, the assumed visualisation paradigm depicts documents represented as graphical icons that are assigned a position in a semantic space in addition to document metadata attributes that are encoded by graphical and geometric attributes of those icons.

The principle of proximity suggests that icons or shapes that appear in close vicinity to each other will be similar in semantic content and can be grouped together based on a notion of ‘aboutness’. The principle of similarity suggests that icons of similar graphical or geometric appearance will be grouped together based on a common attribute or a non-spatially defined classification based on higher-dimensional clustering. The principle of connectedness while not prescribed in the assumed paradigm could be instantiated by drawing edges between items that share a common context as established through citation or link analyses assuming nodes represent web pages or academic documents containing hyper-links or citations. The principle of common enclosed region, also not prescribed in the assumed paradigm could be instantiated by drawing

a boundary around spatially disparate items, potentially overriding the principle of proximity, based on density clustering in the semantic space or based on an additional classification routine; items within the boundary are perceived as a group regardless of items in close proximity but within different boundaries. Finally, the principle of figure-ground, suggests that visual emphasis of items can cause un-highlighted items to recede into the background causing the foreground items to appear more prominent to attention.

As a final point, Rodrigues et al. identify but do not elaborate on the visual perception of textual labels. While such a lack of elaboration is unsurprising for an information visualisation context, the nature of the present research warrants further examination. They suggest that textual labels represent ‘compositions of shapes expressing perceptions of meaning or differentiation’. Yet, text labels are an important percept that applies to labelled semantic spaces. The role of labels in semantic spaces will be made more prominent in Chapter 5, however, for now it is simply emphasised that these are an important perception that are somewhat overlooked by the Visual Expression Process. In the least, text labels assist in establishing structure and interpretation of clusters in the next stage: reasoning.

Reasoning In the third and final stage, reasoning, cognitive resources are devoted to interpretation and decision-making with regards to each graphical percept, which had a significance of meaning attached to it in the observation stage. Reasoning will conclude whether a significant percept or pattern satisfies explicit queries the user has for the data e.g. ‘I suspect the existence of relevant documents based on a set of criteria which I will recognize if presented to me’.

Rodrigues et al. (2007) propose a list of interpretations including correlation, tendency, classification, relationship, order, summarisation, outliers, clusters, structure and reading. Beyond the Visual Expression Process, Valiati, Pimenta, and Freitas (2006) propose a visualisation task taxonomy which represents an annexe to the listed interpretations of Rodrigues et al. It should be noted that identification - analogous to counting in (Nowell, 1997) - is a key interpretation, which is listed by Valiati, Pimenta, and Freitas but not emphasised in the Visual Expression Process; but again, typical for a more analytical information visualisation flavour of work delivered by Rodrigues et al.

Moreover, not all interpretations are strictly relevant for interaction with search results. Identifying clusters, classifications, relationships and reading are important interpretations, but summarisation, correlation, tendency and order are not as important to the task of locating relevant documents. If however, the user desired to analyse the domain of interest as a whole, interpreting structures and outliers would perhaps be more important; however, this is atypical of tasks that a user engages a search engine for and prior research in this area has not explicitly acknowledged this observation.

A parallel exists with Nowell (1997) who argues that task, and not data and graphical characteristics exclusively, should guide the choice of graphical encoding features having observed a conflict between her results and earlier results of a similar intention but conducted within a different context. Furthermore, liberal generalisation of information visualisation principles across different task domains with an expectation of superior performance to an established baseline is misguided. This applies to the use of visualisation in result presentation paradigms and is reflected by evaluation work that finds little benefit to search outcomes using experimental interfaces. Simply put, in search result paradigms, we have too readily dismissed the reading interpretation of Rodrigues et al. and focused too greatly on the detail-on-demand aspect of the visual information seeking mantra of Shneiderman. We must ensure that visually encoded information beneficially augments current search behaviours rather than simply encoding information exclusively visually in the hope that it will be of benefit. With this view in mind, the present research restricts itself to the construction of result presentation paradigms that seek to support the identification of clusters and relationships and interpretations by the reading interpretation; and any facilitation of the other listed interpretations will be advantageous but incidental.

Search result paradigms incorporating information visualisations elicit visual perceptions due in part to the engagement of semantic layout algorithms and attribute encoding as a basis for construction. The framework of Rodrigues et al. provides the context for the remainder of this chapter and experimental interfaces presented in subsequent chapters. The next stage of this chapter will deal with the visual features composing graphical items and their impact on reasoning, after which the focus will turn to a survey of interfaces that facilitate a range of visual interpretations.

2.2 Document Attribute Visualisation

This section will introduce the two research foci presented in Chapter 3 and Chapter 4. The first discussions will focus on the role of motion as a way to encode data while in the second, a methodology for choosing the right graphical feature for the encoding of data is proposed. Chapter 3 and Chapter 4 will submit a quantitative measure of the effectiveness of each proposal obtained by human factors experiments. Both foci share a common context in that both involve encoding of icons or glyphs by way of data encoding paradigms, so this common context will be established first.

Glyph Attributes

Glyphs are small visual representations consisting of geometric and appearance attributes that permit identification of and interaction with the objects or entities they abstract M. Ward, 2002. Appearance and geometric attributes are the subject of much

exploration in a number of taxonomic papers (Shneiderman, 1996; Card and Mackinlay, 1997; Nesbitt, 2005; Pfitzner, Hobbs, and Powers, 2001; Ropinski and Preim, 2008; Brath, 2009). The palette is diverse; attributes include colour dimensions like hue, saturation, and brightness; and shape, size, orientation, shadow, texture, density and iconic pictures. Additionally, attributes can be dynamic: growing, expanding, flashing, pulsing, rotating, oscillating, vibrating, shuffling, and deforming and this list is by no means exhaustive.

Glyphs are widely used in the depiction of individual information entities and their properties in tools presenting results using alternative paradigms; for example Eye-Plorer <http://www.vionto.com/en>, LivePlasma <http://www.liveplasma.com>, Grokker (Foenix-Riou, 2006; Koshman, 2006), and Kartoo (Koshman, 2006), all make use of graphical properties to visualise one or more document attributes.

Data attributes and properties are mapped to geometric and appearance attributes by mapping or encoding rules during the visualisation construction pipeline (Wright, 2007). Data attributes or properties, which are transformed into appearance or geometric attributes, become mapped or encoded attributes. Mapping rules suit numerical data or data recoded into a small number of categories and not large sets of categorical or string-based attributes.

Co-located glyphs of similar appearance form emergent Similarity Gestalts, which allow the user to group items based on the common attribute. With prior knowledge of the mapping rules in play, a user may ascertain the value of a property of an information object by inspection, saving the need for a detail-on-demand request to the interface. The next section outlines the characteristics of data and the marrying of a glyph appearance attribute with a data attribute to form an encoded variable.

Data Attributes

There are three major data types as prescribed by Gowda and Diday (1991): quantitative, qualitative and structured. A quantitative variable can include continuous values, discrete absolute values, and interval values. Qualitative variables include nominal, ordinal and combinatorial types. Finally, structured variables include those with an inherent tree, graph or hierarchical structure based on relationships between nodes.

These three main data types describe the features selected or extracted for visual encoding in a data or information visualisation. Feature selection is the process of choosing data categories for representation while feature extraction is the process of calculating new features from the existing data set for representation (Jain, Murty, and Flynn, 1999). Transformation of a variable type into another data type such as the binning of continuous measures into ordinal range is one such instance of feature extraction. Having selected or extracted a set of features for graphical representation the next step is to decide exactly what graphical feature will encode the data features.

The Encoding Paradigm

Regardless of the level of automation i.e. performed by computer or by hand, a set of rules or conventions will guide the process of selecting appropriate graphical features for data encoding. Rules and conventions help in the construction of both production and experimental interfaces. However, there are many sources of data encoding rules including studies investigating data encoding for statistics and information visualisation (Mackinlay, 1986; Nowell, 1997), studies investigating visual search such as Healey, Amant, and Elhaddad (1999) show, or perhaps inspired from icon construction (Hofmann, 2008; Hofmann, 2009) or through use of metaphors (C. Harrison et al., 2011). Later, Chapter 4 will propose an alternative encoding paradigm based on natural fit between encoding and user.

Chapter 4, in particular, will routinely refer to the set of all encoded variables, all mappings between glyph appearance attributes and data attributes, as forming the encoding paradigm. On a cartographic map a legend or symbol key appears at bottom of the map allowing the user to lookup the meaning of symbols used in the map. Likewise, on an information visualisation a legend may be required to show the encoding paradigm in use. A legend alleviates any need for assumptions with regard to the meaning of encoded data. In time, a legend might not be needed although the user should determine the point at which the legend is no longer required.

Legends are visual dictionaries (Riche, B. Lee, and Plaisant, 2010) that depict data and graphical mappings present in a visualisation, thus permitting the user to look up a mapping at a glance (Tudoreanu and Hart, 2004). Historically, the legend, outside of cartography, has not been a focus of research Nowell, 1997 though seldom does one find a chart without an accompanying legend. Consequently, there are few guidelines for the design of legends for information visualisation. Of the legend guidelines available, Dykes, Wood, and Slingsby (2010) propose a series of guidelines for cartographic legends. Guidelines recommend that legend items should be arranged in a relational matter, graphical symbols in the legend should identically match those in the visualisation, foregoing any down-scaling of the legend to optimise screen space, to position the legend in an area that the user is likely to attend to first or less prominently if it is less likely the user will need to refer to the legend, and to use dynamics to re-order legend layout if necessary - following a filtering operation - or visually bias legend items - following a highlighting operation.

Investigations that are more recent propose the idea of interactive legends (Tudoreanu and Hart, 2004; Riche, B. Lee, and Plaisant, 2010) which take traditional static legends and incorporate interaction controls to facilitate filtering operations. Interactive legends support the navigation of the visualisation in a natural style (Tudoreanu and Hart, 2004). However, Riche, B. Lee, and Plaisant find that the data characteristics influence performance. In their study of performance on interactive legends, depicting

ordinal data was improved but performance on legends depicting numerical information was not, despite elevated participant confidence.

The discussion of encoding paradigm has assumed a one-to-one mapping between graphic and data. However, a combination of graphical codes can encode a single data attribute and this can have a beneficial influence on performance. The next section will outline the role of coding redundancy in many-to-one encoding paradigms.

Encoding Redundancy

Redundant encoding involves the use of two or more appearance or geometric attributes to encode a single data attribute. In contrast, complementary coding or non-redundant coding involves coding two or more data attributes with unique appearance of geometric glyph attributes. M. Ward (2008) suggests redundant coding is useful when dealing with low dimensional data sets, when it is important to convey a message reliably, and to support inter and intra object comparisons when data dimensions are encoded using the same appearance and geometric dimension.

Empirical support for redundant coding is variable. Christ (1975) showed strong support for a positive benefit of encoding colour redundantly. Weidenbacher and Barnes (1997) found that redundant coding was moderately better than an established baseline but that a simplified control of lower graphical complexity, without redundant coding, was superior to both. R. Simon and Overmeyer (1984) found that redundant shape and colour coding is faster than colour coding alone but not significant for shape alone. Jubis (1991) confirms the superior performance of redundant colour and shape coding but unlike in the case of R. Simon and Overmeyer, found colour coding performance to be faster than shape. Finally, Holten and vanWijk (2009) experimented with single and multi-cue features for graph edges and found that single features were better than multi-cues. The authors suggest more research is needed to ascertain the influence of reinforcement e.g. complementary feature pairing, or averaging occurring during the multi-cue conditions.

Whilst coding redundancy has some empirical support, redundant coding restricts the use of that dimension from encoding an additional data attribute. For high dimensional data sets, redundant coding may be hard to justify. However, there are additional ways to improve the extraction of data encoded visually such as the independence of graphical attributes.

Integral and Separable Dimensions

Integral and separable dimensions proposed by Garner (1974) describe the level of independence between two or more visual codes encoded into an object and therefore, how easily they may be extracted and decoded. An integral pairing is one where it is

harder - but not impossible - to perceive each dimension independently. A rectangle for instance is perceived as a whole shape and not really by width and height - perhaps we perceive the visual depiction of the width and height ratio. In contrast separable dimensions are those in which it is easier to perceive them as separate. Consider a red rectangle; the area of the rectangle is independent of the red hue dimension but the height and width of the rectangle is not.

Alternatively, Monahan and Lockhead (1977) offer a definition of integral stimuli based on relationships; if the removal of one attribute also removes additional attributes, then those attributes are considered integral. In their experiment, they investigated estimations of two adjacent vertical lines. If one vertical line is removed, it does not preclude the assignment of value to the remaining line; however, it does preclude the comparison from taking place. Furthermore, in the example above, if the width of the rectangle is unspecified i.e. set to 0, then this removes the height of the rectangle as well; therefore, based on this definition the width and height of a rectangle are integral and interpreted together.

Integral features interfere with each other when only one of the integral dimensions is the target, but when integral features are coded redundantly, there is a speed and accuracy gain expected. In contrast, non-integral or separable dimensions do not interfere with each other even if only one attribute is targeted, but when utilised together as a redundant code, no speed or accuracy gain is expected (Monahan and Lockhead, 1977).

This section has hinted at the scenario in which a user intends to extract two or more pieces of visually encoded data from a multi-attribute glyph. The next section will clarify the range of extraction tasks that the user may engage in order to complete their information activity.

Data Extraction Task Requirements

Single extraction tasks - referred to as non-integration tasks in Nowell (1997) and feature search in visual search literature - involve identifying data attributes encoded by a single appearance or geometric glyph attribute. In contrast, multiple extraction tasks - referred to as integration tasks in Nowell (1997) and conjunction search in visual search literature - involve identifying data attributes encoded by two or more appearance or geometric glyph attributes.

The specific role of integration tasks is contentious: should users engage with multi-dimensional icons composed of many features or should information be spread out over multiple views using the strongest visual codes? Carswell and Wickens (1987) found that when the task necessitates extraction of a single type of information, a multiple view paradigm is more suitable as it permits a user focus on a data set from one dimension at a time. However, when the task involves the integration of a number of different

data types, then a single integrated display results in better performance, as opposed to forcing the user to integrate information over several different views. In contrast, Yost and North (2005) observed that a view per data type and each view encoded using hue, elicits a superior performance than an integrated display. They recommend to split the set of data classes over multiple views and to encode each data type using the most efficient graphical attribute, namely hue.

The choice of multiple or integrated views largely comes down to the task and context of use and the screen real estate devoted to serving the task; context may not permit a large number of views displaying every different facet of the data set.

This concludes the discussion of the document attribute visualisation preliminaries. These preliminaries have set up a shared context for the following two sections. The next section will introduce the use of dynamic glyph appearance and geometric attributes as an encoder of data after which a methodology is proposed for choosing the right graphical attribute for each data attribute.

2.3 *Static and Non-Static Attribute Visualisation*

The terms animation, motion, and dynamics all share a connection with the notion of change. Each term could replace another synonymously with minimal loss of meaning, however the interchanging of one word for another introduces slightly inconsistent connotations depending on the context. We usually associate the term animation with entertainment on television and in cinema; in a generic sense, it is the imbuing of life, action, and movement to an inanimate object. Consequently, the animation process introduces a context in which the personality of the object evolves and interacts. In contrast, systems of power and force, and cause and effect are central to the context of dynamics and while one may draw weak analogies between the physical forces of nature with the intent of the animator, a bouncing ball from the perspective of the physicist is incomparable to that of the animator. In contrast again, the generic interpretation of motion can be reduced to a single increasing or decreasing variable regardless of context. Therefore, it is appropriate to use the term motion - also adopted by Bartram (2001) - in a human factors analysis of dynamic, animated or changing features in a graphical user interface. Context is both a differentiating and contributing factor to the goals of the designer; any model of user interface motion must unify each of the aesthetic, pragmatic and semantic aspects, with context in mind.

Taxonomic work in information visualisation and furthermore, visualisation for alternative presentation paradigms is concentrated on the use of predominantly static graphical devices for use in the mapping and encoding process. Dynamic devices however, i.e. motion, or more generally, animation, appear in computing interfaces with regular and increasing frequency. The nature of their contribution to the computing experience is variably semantic, pragmatic and aesthetic. The aesthetic appreciation

of motion is self-evident through our personal experiences with cartoon and cinema viewing; however, the semantic and pragmatic contributions remain largely under-explored empirically. Prominent exploratory work in user interface animation is offered by Baecker and Small (1990), Bartram (1997), Bartram (1998) and Bladh (2006). These works canvas the meaning of animation, where and how we should use animation in computing applications and how to describe animation when reporting evaluative research. Much of the pragmatic work on animation concerns affect and is generally less abstract; for instance, a lot of the early work focuses on the application of animation to the production of animated icons, and dynamic effects in operating system software much of which we take for granted in routine computer use - consider progress bar widgets for instance. In contrast, psychophysics-flavoured work on motion in information visualisation concerns attention guidance and control efficacies and was quite active around the turn of the millennium; though studies of motion for representation of data had appeared earlier (Limoges, Ware, and Knight, 1989; Ware and Limoges, 1994).

Since the turn of the millennium animation has taken off particularly due to the widespread adoption of affective and aesthetically-pleasing computing which has entailed animated icons with increasingly greater expressive capacity C. Harrison et al. (2011), directly manipulable interfaces (Thomas and Calder, 2001) and smooth view and context transitioning (Bederson and Klein, 2005). Simultaneously, the notions that ‘information is beautiful’ and ‘information visualisations tell a story’ (Segel and Heer, 2010), draw heavily on animation to play a major role in story telling through charts and graphs (Robertson, Fernandez, et al., 2008; Rosling, 2009). Moreover, our everyday applications like GPS navigators (P. Lee, Klippel, and Tappe, 2003) and peripheral reminders (Maglio and C. Campbell, 2000; McCrickard, Catrambone, and Stasko, 2001; Plaue and Stasko, 2007) employ animation for the benefit of everyday tasks. In every one of the aforementioned applications, overly judicious use of animation has drawbacks and consequences; however, it is found that an optimal combination of static and dynamic does have a positive benefit.

Animation is perceptually rich (Bartram, 1997; Bartram, 1998). These aforementioned examples are testament to the richness of animation and consequently, the more or less ubiquity of animation in computer applications. A subsequent section will pull-apart the perceptual richness of motion, however first we first examine how the human perceptual system detects motion.

2.3.1 The Perception of Motion

There are two sources of motion information. The first source originates in specialised motion sensitive neurons in the visual cortex that fire in response to motion and motion direction. A second source is proprioceptive signals based on feedback from the muscles in the eye that report whether movement of an object on the retina is due to the

movement of the eye or the object (Coren, Porac, and L. Ward, 1979). The qualitative experience of motion depends on a number of factors.

Perception of motion is termed the ‘apparent motion problem’. To see the apparent motion the observer must connect features present in successive frames in a process called Correspondence (Gershon, 1992). Note however, this is distinct from the notion of correspondence in the Visual Expression Process of Rodrigues et al. (2007). We perceive motion from jumps in location between static, multiple images presented in sequence.

Motion perception depends on three parameters: stimulus duration, inter-stimulus interval and the distance change (Griffin et al., 2006). Stimulus duration relates to the length of time the frame appears; the inter-stimulus duration is the frame rate or length of time between frames; and distance travelled relates to the change in value between frames; for example, the distance an object moves or grows between frames. As object movement distance increases, shorter inter-stimulus intervals are required for smooth apparent motion.

Beyond the perception of motion or simply change, we can differentiate or interpret motions based on a number of different dimensions; this is the topic of the following section.

2.3.2 *The Dimensions of Motion*

Every graphical encoding device has an inherent capacity to convey information (Huber and Healey, 2005); but we do not yet fully understand the limitations of these capacities nor fully understand the interplay between graphical attribute combinations. This situation is even more acute in the case of dynamic attributes, in contrast to static attributes like colour where many examinations have already taken place (Nowell, 1997).

Colour and shape are two visual attributes that designers readily employ to visualise information, due in part to their expressive capacity. Dimensions of colour include hue, saturation and brightness while shape has even greater dimensional capacity as exposed by Brath (2009). Motion, like colour and shape has great expressive potential as is decomposed by Bartram (1997) and recreated in Table 2.2 on the next page. The work of Bartram (2001) focused heavily on signalling and urgency aspects of motion as well as on phase as a way to perceptually group multiple items in a visualisation. The basic features of motion are discussed below and the interpretative aspects of motion in the following section. However, compound motions are not a focus of the present research and are only included for completeness.

A high school physics textbook should offer definitions for phase, frequency, and amplitude, so here a cursory definition is provided, followed by a generalisation of each definition to data encoding.

Tab. 2.2: Motion dimensions and expressive capacity for a single item and for multiple items; this table adapted from (Bartram, 1997).

	Basic	Interpretative	Compound
Single Item	Phase Frequency Speed Velocity Transformation, Direction, Trajectory, Smoothness, Continuity, Duration, Position, Amplitude, Shape, Temporal Continuity, Gestalt Continuity	Signalling, Active, Viewing, Jostling, Autonomy, Locomotive, Expressive, Exertion, Urgency	Not Applicable
Multiple Items	Per Single Item	Per Single Item	Relative Velocity, Relative Trajectory, Sequence, Transition, Filmic Techniques, Causation, Attraction, Repulsion

Amplitude is the peak deviation from an equilibrium or half-way location; the amplitude of a pulsing or flashing light is the distance between mid and full luminosity and mid to zero luminosity. Equally, a pulsing light might oscillate between a low luminosity state and a high luminosity state; in this case, the amplitude is the deviation from half luminosity.

Frequency is the number of occurrences of something over a unit of time, usually per second; the number of times a light flashes per second is the frequency of the flashing light. A light may flash every two seconds so the frequency is one half. Alternatively, the period is the inverse of frequency and is the time it takes for a single occurrence or cycle to happen. The period of a light flashing once per second is one, but the period of a light flashing once every two seconds is two as it takes two seconds to flash once. This distinction is important, because we might refer to a slow flashing light as one flash every two seconds and not a half cycle per second. In contrast, we might refer to a fast flashing light as two flashes per second but not one flash every half second. Another example is that it is more natural to consider a leap year as one event in four years and not ‘a quarter leap year per year’.

Phase is the fraction of a cycle that has elapsed relative to an arbitrary point like zero time. However, the interpretation of phase is relative; so phase is defined for two items under motion. Two items are in phase with each other if their cycle starts and finishes at the same time. Two items are exactly out of phase with each other if when a new cycle of one starts and a cycle of the other is half way complete. If one item is slightly out of phase with the other, then the event of one will appear to lag behind relative to the other. Two lights flashing at the same frequency and which started flashing at the same time are in phase. Two lights flashing at the same frequency which started together in the opposite state, that is, one started in the off position and one in the on position at start time, will appear exactly out of phase with each other;

Tab. 2.3: Interpolation functions used in the production of motion; t denotes time; $w=2\pi f$ denotes angular frequency; α denotes amplitude; and ϕ denotes phase

Name	Pseudo Function	Graphed	Motion Application
Sinusoidal	$f(t) = \alpha * \sin(\omega t + \phi)$		Non-Linear Oscillation
Saw Tooth	$f(t, a) = 2(\frac{t}{\alpha} - \text{floor}(\frac{t}{\alpha} - \frac{1}{2}))$		Linear Rotation
Triangle	$f(t, a) = \text{abs}(2(\frac{t}{\alpha} - \text{floor}(\frac{t}{\alpha} + \frac{1}{2})))$		Linear Oscillation
Square	$f(t) = \text{sgn}(\alpha * \sin(\omega t + \phi))$		Abrupt Flashing

when one light flashes the other will not and versa vice. When the lights are slightly out of phase with each other one light will turn on, then shortly after the other will turn on, shortly after which the first one will turn off and so on.

These definitions are laboured in order to illustrate the interplay of the basic features of motion: if two items of the same frequency start cycling together, they will exhibit similarity based on phase and frequency. Thus, when encoding data into frequency, it is important to randomise phase to ensure that similarity judgements are made for frequency and not a combination of frequency and phase. In any case, regardless of the type of graphical attribute undergoing change modifications to phase and frequency can influence the perceptual experience of the change. Moreover, the choice of interpolation function will also play a role in the perceptual experience.

A number of different algorithmic or computational techniques e.g. interpolation, kinematics, and kinetics, are used to produce the perception of motion; though the approach taken tends to identify with the domain of application. In essence, the change from one state to another is brought about by the interpolation between two states by way of interpolation algorithms, which add a small amount of change to a numerical quantity at each step in time. An exhaustive treatment of all approaches to computational motion generation is beyond the scope of this discussion. However, for the current treatment of motion, a focus is placed on the use of sinusoidal and non-sinusoidal functions - see Table 2.3 - for interpolation. The observer perceives a different pattern of motion for each function as illustrated by the graphics in the image column. Note that each function has a range of $(-1, 1)$ so for encoding purposes, offsets, shifts and multipliers are applied to constrain the range to $(0, 1)$.

This section has illustrated the dimensional richness of motion as evidenced by a variety of basic features and production algorithms that when manipulated, influence the observer's perceptual experience of motion. In the next section, the discussion will turn to the interpretive aspects of motion based on a series of application areas

identified by Bartram (1997).

2.3.3 *Studies of Motion*

Early research foci on animation were spurred on by the then recent advances in display technologies that made it possible to display colourful and vibrant animations. More recently, the widespread interest in motion and animation has likely been spurred on by better tool and library support and the burgeoning focus on user experience in computing. However, animation is not restricted to applications of affect. Bartram (1997) isolates six potential applications for animation. These include annunciation and signalling, grouping and integration, communication of data relationships, data display and coding, representing change and general visibility concerns. The following discussion is shaped by Bartram's six application domains and updated as necessary to reflect achievements and progress in the years subsequent to Bartram's survey.

The sensitivity of motion with respect to other visual features in the periphery of vision (McKee and Nakayama, 1984) has long since been a prime reason to use motion in user interfaces and accordingly has concrete applications for directing attention from a primary task to new events signalled outside the focal region (McCrickard, Catrambone, and Stasko, 2001; Bartram, Ware, and Calvert, 2001). There are trade-offs in play that the designer must consider. Design choices should promote awareness of incoming events, but without undue distraction and annoyance. In addition, Plaue and Stasko (2007) suggest that the physical orientation of peripheral animated displays plays a role in annunciation and signalling indicating a rich set of human factors in play.

Motion, like many other visual features including but not limited to colour, facilitates emergent and Gestalt features, and this has concrete applications for grouping and boundary detection (Bartram, 2001). For instance, Bartram and Ware (2002) demonstrate that coherent motion elicits emergent groups of like-motion, which the user can use to quickly and efficiently focus attention on target groups for further cognitive processing. Fast and easy visual segregation or filtering of groups of items in a display reduces load on cognitive resources that they would otherwise devote to efforts to maintain attention on a target group whilst conducting further cognitive analysis of the group in play. Ware and Bobrow (2004) also show how motion can be used to highlight a collection of related network nodes to this effect, while later Ware and Bobrow (2006) demonstrated a benefit of phase for highlighting clusters of points within dense scatter plots.

Communication of data relationships has received significant interest, particularly following the now seminal talk of Rosling (2009) who used the GapMinder Trendalyzer tool to show an animation of the progression of third and first world countries over time. However, Robertson, Fernandez, et al. (2008) suggest that although these types

of tools can be effective at telling a story they can be misleading and that static versions of the data sets may be more suitable for the initial analysis.

Traditionally, data display and coding has focused almost exclusively on static appearance and geometric features such as colour, and such research is typified by that of Nowell (1997). Corresponding research into the combination of static and dynamic features of motion is relatively light, though earlier work of Ware and Limoges (1994) suggests that encoding of motion features for data representation may be worthwhile. Bartram, Ware, and Calvert (2003) also investigate relevance coding in their grouping and distraction work and propose a number of motion guidelines including the idea that motion shape or type influences a user's perception of visual importance. Furthermore, Horn (2007) proposes the use of dynamic jellyfish and bacteria-inspired icons for the purposes of artist profile and email visualisation. Jellyfish mingle around similarity neighbourhoods in the artist visualisation while the motion of the email creatures related to whether the user has not read, read or responded to an email.

Representation of change as identified by Bartram (1997) means to indicate that change has occurred and to show how significant the change and how rapidly that change took place after the moment. A combination of dynamic and afterglow effects are implemented in the Phosphor system of Baudisch, Tan, et al. (2006).

Finally, visibility concerns raised by Bartram (1997) pertain to motion that is devoted to manipulating display configuration or the viewpoint of the user particularly to reveal hidden information to the user such as motion to disambiguate three dimensional structures by smooth transitions. Bartram also suggests the use of jostling windows or animated icons and this has been adopted in multiple contexts for example the bounce to alert animation in the task dock in Apple operating systems as well as flashing of minimised window bars on the task bar in Microsoft Windows operating systems. C. Harrison et al. (2011) propose an application of kinetic behaviours by way of dynamic transforms on static icons to establish a rich set of meaning through dynamic metaphors. Kineticons offer the pictorial representation of the icon plus an additional dimension of state information indicating that the task is starting, progressing, needs attention, or cannot complete. In addition they can convey a state change or replacement, initialisation or shut down of a task, and interactive affordances of an icon e.g. this icon is movable.

Both Bederson and Klein (2005) and Shanmugasundaram and Irani (2008) have found a benefit for using animation to smoothly transition between views for dense symbol displays and map reading respectively. Bederson and Klein found that the benefit was more pronounced for displays that are more symbolic while Shanmugasundaram and Irani found a significant improvement in time over a baseline that did not transition between zoom scales, but no difference for error was found.

This section has explored a wide range of animation and motion applications in

graphical user interfaces and information visualisation. Such a broad perspective provides a rich source of inspiration and ideas not only for the current research but for the overall design of alternative presentation paradigms as well.

It is the intention of the current research to focus only on using motion to encode data in information visualisation applications. The data display and coding section indicated a prevalence of research using motion phase to support visual clustering of items in high-density displays. Frequently, motion frequency is found to be among the worst features in supporting these types of tasks and similarly correlation tasks in Ware and Limoges (1994), but this is not cause to dismiss motion frequency altogether. The next section will outline a proposal for use of motion in the interface.

2.3.4 A Gap in Understanding - Encoding Data with Motion Frequency

We can encode motion attributes in much the same way as their static counterparts for the purposes of document attribute visualisation. However, this is something that is not widely prevalent. Motion attributes are subject to the same design choice considerations including data, task, perceptual and redundancy characteristics but our understanding of the ramifications of each consideration is comparatively poorer than what is known for static coding attributes.

Whilst significant efforts have been devoted to the study of motion phase and distractibility in user interface applications, there are other facets of motion like frequency that have not enjoyed the same acclaim. For example, Ware and Limoges (1994) examined statistical charts featuring static and dynamically encoded data to convey correlation. Motion phase encoded into points transitioning vertically within a histogram like graph was observed as an effective way of conveying correlation between two variables; however, frequency was highly ineffective. Ware and Bobrow (2006) found phase to be effective and frequency not effective for segregating groups of items in a high-density display. Bartram, Ware, and Calvert (2001) also showed a positive effect for motion grouping based on pattern and common frequency and phase.

The present proposal significantly differs from this earlier work in that the user does not use any class of appearance attribute for primary attention guidance as in grouping and integration work involving phase. In this formulation of a document metadata visualisation - one aspect of a result presentation technique - spatialisation algorithms allocate documents to spatial locations. High local densities produce theme neighbourhoods throughout the global visualisation space, which the user can navigate by way of provisioned semantic landmarks. Users base their search decisions on semantic or thematic content primarily. Then, and only then, do users focus attention to alternative relevance cues such as metadata. This formulation differs to earlier research (Bartram, Ware, and Calvert, 2001; Ware and Bobrow, 2006) in that appearance attributes are used to segment groups of items before interrogating specific and

perceptually defined clusters. However, this does not exclude from use, data encoded to appearance attributes of the glyph; these data are important following the establishment of thematic or semantic relevance but not before. Therefore, in this context, appearance attributes should be chosen such that they support efficient and effective identification of properties.

The course of research presented in Chapter 3 will propose to encode up to four data attributes using motion attributes. In this proposal, each glyph represents a single document, which has several metadata properties that could be represented visually. There are two main aspects to the current proposal of investigating motion in visualisation applications:

1. To investigate the effectiveness of multi-extraction tasks of up to four dimensions using a combination of static and dynamic attributes and;
2. To investigate the differences across different frequency-variant motion patterns: pulse, horizontal-shuffle, rotation and grow - grow is sometimes termed expansion-deflation or zooming.

In the first instance, this investigation seeks a better understanding of motion frequency in data encoding paradigms. This will enable designers to provide a richer representation of relevance cues beyond theme and topic in the result sets which Balatsoukas and Ruthven (2010) suggest searchers do use. This analysis seeks to compare the differences on performance between statically and dynamically coded objects for use in multiple extraction tasks of up to four dimensions.

In the second instance, a more thorough understanding of motion patterns is needed for the inclusion of frequency in encoding paradigms. Despite the earlier finding that phase-variant motion patterns were perceived equally well (Bartram, Ware, and Calvert, 2001), this research goal seeks to compare motion patterns in the suspicion that some frequency-variant motion patterns maybe more difficult to perceive than others. That is, the frequency of some motions such as flashing may be easier to process than say zooming motions.

These goals are significant as they seek to evaluate the utility of frequency in an application that does not necessarily favour motion phase. Previous research into motion attributes has favoured motion phase and not frequency and has cast an unfavourable light upon motion frequency. The outcome of this research will be a ranked set of data encoding rules that specify how to use motion frequency to encode data in alternative result presentation paradigms.

This section motivated a course of research that will indicate whether encoding by motion frequency is worthwhile. Nevertheless, this research will not specifically indicate which data attributes motion frequency should encode. The next section will offer a

discussion relating to a proposal of a methodology that should be adopted to influence the choice of encoding rule generation, based on a match between the data attribute and the appearance attribute.

2.4 *Natural Encoding Paradigms*

The previous section motivated the use of motion frequency in coding but it did not provide any guidance as to which data attribute it should encode. Yet, there is a defined point in the visualisation pipeline at which a search tool designer must select an geometric or appearance attribute to encode each data attribute in the visualisation.

There are several perspectives on how this decision-making process should be guided. These include the affordances of visual attributes to encode data of different characteristics, the perceptual efficiency of geometric and appearance attributes, use of metaphor, cultural aspects, and intuitions and conventions. It is the last perspective, intuitions and conventions that will be of most interest, as these have a natural fit to the user's prior expectations and beliefs. While Chapter 4 will examine further each of the competing perspectives, the following sections will introduce the notion of 'natural fit' and its role in guiding the derivation of encoding paradigms.

2.4.1 *The Principle of Best Foot Forward*

Several researchers (M. Ward, 2008; Bartram, 1997) suggest that some encoding rules are appropriate for data encoding based on intuitiveness. Intuition is the ability to understand something immediately without explicit reasoning. Intuition could be considered a convention in design - i.e an agreed way of doing something; but what is intuitive does not necessarily result in an optimal design. The ranked-list search presentation paradigm is a telling example as it is intuitive to put the best results at the top of the list, however, there are no guarantees that the best results will appear at the top of the list - and particularly so, when search is more complex.

Larkin and H. Simon (1987) noted earlier that in order to understand concepts through visual means, the visual representation needs to be explicit and analogous, or 'isomorphic' (Gurr, 1998) to that which it represents. A classic example is that it is intuitive to encode the magnitude of a variable to the size or area of a glyph; the difference between two magnitudes is represented by one glyph having a greater area than the other - assuming a comparable width to height ratio. Thus, it is logical to suggest that mapping magnitude is analogous to the appearance attribute of size.

In the depiction of dynamic processes, Narayanan and Hübscher (1997) suggest that when we externalise cognition by way of isomorphic representations, on annotated paper diagrams for instance, additional cognitive processing capacity is available to gain an understanding of these systems. Good learning outcomes are more likely to occur

because the diagram facilitates accurate resolution of visual hypotheses by inspection. Equally, if an encoding rule violates intuition, we should expect analytical disturbances that warrant an allocation of cognitive resource, in order to maintain those unintuitive mappings in short-term memory. In doing so, a detriment to task performance measured in time and quality of outcome is expected as the user must devote effort to overcoming something apparently unintuitive. If however, the graphical attribute explicitly or analogously represents the data attribute, such that the representation is isomorphic, then we should denote this encoding as exhibiting good naturalness or natural fit.

One can think of natural fit as an analogy of the Stroop Test and other forms of interference tests. In a Stroop test, words representing colour categories such as red, green and blue are presented in different font colours on a display. The participant's task, depending on the type of Stroop Test, is to call out either the colour of the text or the word itself. Task performance is measured by response time and accuracy and degrades when the ink colour is incongruous with the colour word e.g. red ink of the word 'green' is incongruous while red ink of the word 'red' is congruent. Participants are typically faster at calling out the colour red when the word red is coloured red, in contrast with calling out the colour red when the word green is coloured red. Under incongruous conditions, additional performance overheads are devoted to effortful inhibition of conflicting stimuli.

Like the Stroop Test, there are other instances of conflict tasks in the literature (Gevers and Lammertyn, 2005). Dehaene, Bossini, and Giraux (1993) observed that when making a parity judgement - odd or even even judgement - for a number, smaller numbers were processed faster when the response was made with the left hand, while larger numbers were processed faster when the response was made with the right hand. This is indicative of the Spatial-Numeric Association Response Effect or SNARC.

The SNARC effect is observed only to exist when numeric digits are presented on the apparatus display and not the word representation of the digit. Dehaene, Bossini, and Giraux showed that this effect might be due to left-to-right reading convention, as the reverse effect was observed for participants who adopt a right-to-left reading convention. However, as Ren et al. (2011) recount, a SNARC effect is also present for higher numbers up and lower numbers down and additionally report that a left/right preference for small and large circle sizes, and darker and lighter luminance (Ren et al., 2011) and references to past and future time (Santiago et al., 2007). They argue that this effect is due to a cognitive mechanism and not a cultural convention; we may be biologically geared to associate the left with small and the right with large and consequently, it is more natural to present the lower extreme on the left of the visualisation and the higher express on the right side.

Santiago et al. investigate the influence of spatial characteristics of stimuli and response on perception of time. Their results indicate that categorisation of words and sentences as past or future i.e. the response, happens faster when words referencing

the future appear on the right of a display and word referencing the past appear on the left of screen and when participants used their right hand to respond to future words and left hand to indicate past words.

These observations further support the notion of naturalness, or congruence between prior knowledge about some dimension and the way that dimension is visualised. These conflict tasks provide insight into how the brain processes magnitude information but, how does this translate to encoding paradigms for document attribute visualisation. It is natural and intuitive to seek out large glyphs for large magnitudes of size, but unnatural to seek small glyphs for large magnitudes of size - a Stroop Interference variant - or the steepest tilted lines or the darkest coloured glyphs. Therefore, it is intuitive that designers should make-prominent, aspects of the data set that the user is presumed to be interested in.

For instance, the font weighting of words in result surrogates has been manipulated in a number of studies (Woodruff et al., 2001; Olston and Chi, 2003; Aula, 2004). Additionally, Anderson (1990) (P. Hu, Ma, and Chau, 1999) reports that size influences arousal, suggesting that if arousal is key to triggering directed attention and an analysis of the stimuli then what is important should be made prominent and bigger. However, there are limitations and constraints on not only manipulating the magnitude of some visual variables but also the interpretation of those manipulated variables. Whilst we can assign a numerical magnitude to orientation or colour intensity, it assumes the magnitude is bounded, since there are upper and lower limits on line orientation and colour intensity. In contrast, there are no explicit bounding constraints for glyph size. Furthermore, while an ordinal meaning may be assigned to glyph size, the same intuitive ordering between hues cannot be established (Bertin, 2011)

The literature makes it clear however, that no graphical code works equally well for all users and that no single presentation works well for all purposes. Thus, Nowell (1997) states that it is the challenge of the designer to choose codes to support a range of information tasks, while supporting individual differences in the user population. Alternatively, the importance of encoding issues may be weighted according to the application domain. In visualisation-based multivariate analysis for example, a primary concern lies in expanding the dimensionality of the icon, while maintaining or improving analytical performance over high-dimensional data sets. Furthermore, from another perspective, design of icons for desktop applications focuses on communicating purpose and content (Setlur et al., 2005), brand identity, and making the interaction afforded more prominent (C. Harrison et al., 2011).

The difficulty in devising encoding paradigms for visualisation-based search tools is that a combination of information visualisation and ranked-list-based information retrieval theory entails consideration of both analytical aspects of information visualisation as well as raw visual - and textual - search tasks. Thus, on the one hand, information visualisation literature suggests encoding paradigms should support a range

of analytical tasks which may not all necessarily be engaged by searchers, while on the other hand, the visual search literature suggests encoding paradigms should ensure target identification is as fast and as accurate as possible. In yet another light, the visualisation literature typically adopts a task and data type perspective - i.e. matching visual features appropriate for the type of data - in contrast, visual search literature is void of any context and is primarily interested in raw recognition performance, thereby making any direct application to an information visualisation context, less than straightforward.

Accordingly, it is desirable to devise an encoding paradigm that favours fast recognition of targets - both perceptually and cognitively - since upon recognising targets within an information search context, it is unlikely that the user will want to perform tasks other than opening document full-text or finding targets of a similar appearance. Moreover, whilst an encoding paradigm for a typical information visualisation may favour encoding rules that support visual estimations and calculations across the whole data set, natural encoding paradigms favour encoding rules that support fast recognition - i.e. both fast perceptual recognition and critically, fast cognitive interpretation - of individual targets.

Table 2.4 on the next page collates a number of different alternatives, guidelines, suggestions, or perceived intuitive encodings that are seen across the various literatures. This collection applies directly to the search metadata under examination in Chapter 4; it is positively non-exhaustive; however, this small collection demonstrates that the potential scope of encoding rules is broad and, at the same time, conflicting in places.

Tab. 2.4: Intuitive and conventional encoding seen across literature.

Data	Graphic	Attribute	Meaning	Convention / Intuition	References
Age	Saturation	Vibrant Dull	New Old	Convention	Horn (2007)
	Position	Left Right	Past Future	Both	Santiago et al. (2007)
	Motion	Slow Fast	Old New	Convention	Horn (2007)
Popularity	Brightness	Bright Dull	Popular Unpopular	Both	Horn (2007)
	Position	High Low	Better Worse	Both	Hegarty (2011)
Importance	Colour	Red Blue Green	Relevant Irrelevant Neutral	Convention	P. Hu, Ma, and Chau (1999) and Nowell (1997)
	Distance	Near Far	Important Unimportant	Both	Baumgärtner et al. (2007)
	Motion	Slow Fast	Not Urgent Urgent	Convention	Ware, Bonner, et al. (1992)
Magnitude Cardinality	Size	Small Large	Small Big	Intuitive	Brath (1997)
	Shape Vertices	Few Many	Less More	Intuitive	Chapter 4
	Grey Scale	White/Light Dark/Black	Less More	Both	Hearst (1995)

2.4.2 *A Gap in Understanding Semantically Motivated Encoding Paradigms*

There exists an open research question centred around the role of natural encoding paradigms in search tool metadata visualisations. The question posed is whether a natural encoding paradigm is superior to encoding paradigms based on conventions, data type and task type, or visual search theories. However, before comparative analysis can be made, it must first be shown that encoding rules which are seen to be more natural or intuitive, result in superior performance to those which are unnatural.

Chapter 4 will argue that we should give preference to a user's prior expectations, experiences and beliefs when devising encoding paradigms for metadata visualisation. In doing so, natural encoding rules should limit the effect of cognitive interference - erroneous interpretations of graphical codes - when decoding metadata visualisations.

If, under a natural encoding configuration, an improvement on objective performance is observed, it will be unsurprising that participants are unable to accurately describe the characteristics of the encoding when polled. Good objective performance, but poor self-reported learning is expected, as under natural encoding conditions, there is little need to explicitly and consciously maintain a mental schema of the interface in immediate memory - meaning that under an natural encoding condition, participants may simply complete tasks without great thought.

In contrast, searchers under an unnatural mapping, must consciously inhibit their prior beliefs and expectations, in order to prevent those interfering with their processing of encoded data. At the point where a participant recognises a conflicting meaning, they must stop, refocus and reconfirm that all prior analyses were conducted under the actual unnatural encoding paradigm. Where encoding is counter to intuition, performance should degrade and learning will be required, thus forcing participants to spend conscious effort on rectifying their understanding of the encoding paradigm. Since having spent conscious effort to work with the unintuitive interface, participants may be more readily able to draw on the outcome of their learning.

2.5 *Visualisation of Inter-Document Relationships*

Prior sections have focused on the representation of document attributes using appearance and geometric attributes of glyphs. In this section, discussion will shift toward representation of between-document relationships that are based on thematic content.

It was earlier proposed that contemporary result presentation paradigms offer inadequate organisation of search results. This section will examine the ranked-list paradigm in greater depth, and in particular, focus on the advantages and disadvantages that this technique affords. Then a survey of alternative result presentation paradigms will be presented and a critique of these systems in the context of modern search behaviour

provided. This research, as several researchers have already before, reiterates that alternative result presentation techniques offer a more adequate organisation of search engine results than ranked-list based organisations. Furthermore, it will be argued that systems employing alternative result presentation techniques could benefit from further usability evaluation. A survey of systems presented below will motivate a set of competing design choices that a later comparative usability experiment will seek to evaluate.

This section will begin with an outline of what is meant by semantic relationship and the way spatial arrangements and spatial cues can be used to illustrate semantic relationships between search results.

2.5.1 *The Semantic Relationship*

Ranked-list result presentation techniques typically incorporate a sophisticated ranking algorithm that takes into consideration a range of ‘signals’ - such as word occurrences within topographically and structurally prominent areas in text documents - in order to match documents that share a similarity with a searcher’s query. However, list position only indicates mathematical distance from the user’s query; accordingly, inter-rank distance has no practical semantic interpretation. So what constitutes a semantic relationship and what semantic relationships do we look for when establishing relevance.

Chaffin and Herrmann (1984) investigate the notion of semantic relation in the context of human memory studies. In their study, they asked subjects to sort a set of words into groups based on semantic relatedness and, consistent with earlier studies, conclude that there are around five broad semantic similarity types. These are: contrasts, class inclusion, similars, case relations and part-wholes. Contrast relationships include pairs of words that contrast, oppose or contradict one another like hot and cold. Similars are words that overlap in meaning or connotation like ‘under’ and ‘below’. Class inclusion are words such as ‘game’ and ‘football’; in class inclusions, one word is either a generalisation or a hyponym of the other. Case relations are words pairs consisting of an agent and a function or instrument like ‘surgeon’ and ‘scalpel’. Finally, part-whole relationships include words like ‘flower’ and ‘garden’.

We can think of the work of Chaffin and Herrmann as relating to the semantic relations between documents on a micro scale or sentence level and draw a parallel to this work as a way to look at a document or set of documents on a macro scale. At the document level, there are a variety of potential semantic relationships documents can share based on the content discussed. The main themes of one document may contrast in opinion or approach to an issue or problem to another document; or discuss a specialisation of a topic discussed generally in another; or discuss a topic that forms a part whole-relation with another topic.

Macro scale semantics are typified or encapsulated by folksonomies or collaborative tagging of web resources on the Internet. Maguitman et al. (2005) discuss exploiting the semantic meaning of documents in the Open Directory Project <http://www.dmoz.org/> in order to calculate semantic similarities between documents. They highlight that the degree of similarity between documents involved in class inclusion or is-a relationships can be calculated by consideration of the degree of shared information and recount also that the meaning of topics can be ascertained by isolating a common immediate hypernym and then taking measurements of the degree of overlap between topics using information measures.

Given that we can - with adequate qualification, since calculations take place over intangible quanta - define a semantic relationship, how should this relationship be presented to the searcher? Despite sizeable breadth in semantic relationship types - as seen above - visualisation techniques are typically relationship agnostic. For example, node-edge diagrams that display a link between two related nodes suggests a relationship exists, but do not in the generic sense indicate what that type of relationship means i.e. *does this link represent a part-whole or class inclusion*. Similarly, use of spatial arrangement of documents reveals little about the type of relationship. However, these techniques provide a foundation upon which to provision additional graphical cues to convey the type of relationship.

2.5.2 Spatial Arrangement

The process of assigning a spatial position for abstract information objects, or visualisation of values without an inherent, physical analogue - such as found in scientific and geographical visualisation - is termed spatialisation or ordination. The result is a thematic landscape or arrangement of documents into a typically two-dimensional space, in which the spatial proximity corresponds roughly to the semantic similarity between documents (Granitzer et al., 2003)

Spatialisation has two interdependent characteristics. The first is semantic and specifies the basis for inter-document relationships, while the second is spatial and specifies how best to represent those relationships visually (Zhang, 2008). Semantic relationships depend on the analysis of a 'sampling unit' such as metadata units or semantic units (Börner, C. Chen, and Boyack, 2003), whereas spatial relationships depend on visual perceptions of correspondence, differentiation, connectivity and arrangement as Rodrigues et al. (2007) propose. These two characteristics are important to consider as Niemelä and Saariluoma (2003) indicate, high semantic coherence and spatial grouping influence the users' cognitive processing of the visualisation.

The mutual dependence of semantic and spatial also raises issues of dimensionality, use of metaphor and labelling of documents and labelling of the spatial substrate. Information search typically deals with highly dimensional documents - words, topics,

and themes are all dimensions - and so analytical techniques are used to extract the most representative dimensions. A critical aspect of this process is the degree of information loss during dimension reduction (Pérez and deAntonio, 2003).

Subsequently, the number of visualised dimensions becomes important. This research has deliberately chosen to focus only on two-dimensional visualisation at any one time, as evidence suggests that two dimensions are better than or as good as three dimensions (Sebrechts et al., 1999; Tory, Sprague, et al., 2007) - despite the back (Cockburn and McKenzie, 2001; Cockburn, 2004) and forth (Tavanti and Lind, 2001) often seen in research on the subject. However, omission of three-dimensional visualisation undervalues the use of some spatial metaphors that may benefit a user's navigation of information space. For instance, a cityscape metaphor, in which buildings represent documents, elucidates inter-relatedness through building proximity and neighbourhood proximity (Bonnell, Lemaire, et al., 2006) - concepts which we are familiar with in reality.

Alternative search result displays are motivated to promote better search outcomes by enhancing meaning, promoting recognition, supporting the discovery of patterns, concepts and similarities (Foenix-Riou, 2006) but research suggests that searchers engaging these search tools under-perform (Chung, H. Chen, and Nunamaker, 2005; Ostergren, S. Yu, and Eftthimiadis, 2010). However, these systems could be improved because a significant effort is made to design such interfaces overly unlike ranked-list interfaces, and as such are inconsistent with current, typical and ingrained information search behaviours. Spatial arrangement or position is a salient feature for the representation of document relatedness. Although relatedness is a subjective matter, a relationship may be made explicit e.g. an edge between nodes, or implicit e.g. based on point density and proximity. However, our knowledge remains imprecise about how spatialisation-based techniques can feature in systems that interweave both contemporary and alternative search behaviours.

An investigation of the rudimentary spatial component will expand on this knowledge. In pursuit of this, Figure 2.2 on page 59 collates examples of spatialisation paradigms to aide evaluation and comparison of past, present and future result visualisation. This de-featured survey harmonises with the work of Morse, M. Lewis, and Olsen (2002) and Butavicius and M. Lee (2007) who take a reduced form 'de-featured' approach for the evaluation and comparison of visualisation techniques.

2.5.3 Alternative Result Presentation Paradigms

Search user interface research is an active area of research. There are a number of recent exploratory publications made in the area of alternative result presentation paradigms and search user interfaces including: 'Search User Interfaces' by Hearst (2009), 'Search Patterns' by Morville and Callender (2010), an outline of exploratory search interfaces

by Kules, Wilson, and Shneiderman (2008) and finally an introduction to the technical fundamentals of search result visualisation offered by Zhang (2008).

After some review of the area, it is apparent that there are a very large number of alternative techniques. Mann (1999) points out that several stages of information retrieval may benefit from a visualisation technique, so the techniques visited shortly are not exclusively result presentation paradigms and may apply to earlier levels of the information-seeking model outlined above in Table 2.1 on page 12.

Figure 2.2 on page 59 depicts visually - and textually in Table 2.6 on page 58 - search result visualisation paradigms and systems organised by technique classification. Each row represents a class of technique and each column represents an instance exhibiting the essence of the technique. This collection deliberately omits non-essential aesthetics to facilitate comparison of spatial layout. It abstracts away from the inessential and is proposed as a reference for designers and in particular, those who wish to carry out comparative evaluation. In removing aesthetic debris, Figure 2.2 on page 59 emphasises inter-document relationships but not document attributes.

This survey is by no means exhaustive, but it certainly provides a rich overview of the types of techniques that exist that could serve as a basis for information tool design. It must be stressed, that this survey observes predominantly systems and techniques for the review of results; nevertheless, this collection is of sufficient diversity and richness as it covers many but not all systems that display search results spanning that presented by Hearst (1999) and more recently and comprehensively by Hearst (2009), the seven data types of Shneiderman (1996) - with the exception of systems depicting temporal data of search results which are uncommon; and finally, the data relationship structures presented in the framework of Pfitzner, Hobbs, and Powers (2001) - with the exception of lattice structures. This survey was conducted in 2009 and while there may be more recent systems in existence, it is nonetheless considered to be of sufficient diversity based on broad coverage of the aforementioned works. Moreover, it was not intended to survey systems in which a ranked-list could be manipulated by interactive controls; thus, in light of this, the survey does not consider faceted interfaces which are routinely prominent in present day online shopping sites.

The following discussion will outline each classification based on row order in the matrix of Figure 2.2 on page 59. A summary of the main characteristics of each system that have guided this examination is provided visually in Table 2.5 on page 57. This table provides a number of dimensions that could also be used to characterise each system. In addition to a classification of each system according to Figure 2.2 on page 59, other dimensions and attributes include: the name of the system where applicable, a representative reference, the type of corpus the system conducts search over, the potential insights a searcher may gain, where in the information-seeking process that a searcher may receive benefit, the way in which searchers access a surrogate or summary view of the results and finally the way in which searchers obtain a full-text view of the

search result. While knowledge of the context, expected insights, and information-seeking stage is useful to consider, the final two dimensions are significant here as they motivate a course of research to be presented in Chapter 6.

The surrogate or keyhole view property indicates how and what information is made available to searchers when they are initially interested in a specific document. For systems that do not show a list of results, where the document surrogate information is naturally and explicitly present, systems that employ alternative result presentation techniques utilise a range of typical techniques to show document surrogate information. These include, pop-up windows of various forms and appearances and containing various metadata; as well as the use of dedicated areas of the visualisation in which document surrogate information is presented; as well as linking a visualisation with a ranked-list of results.

Conversely, the full-text view property seeks to characterise systems on the way the searcher accesses the full-text view of specific results. In nearly all cases, authors specify that full-text is accessible through a new window or frame. Rarely, is there any graphical depiction - in academic reports - of this software frame. Generally, the full-text view is assumed to be unimportant to the reader, since, for the searcher, it is simply the means to an end of information need. Yet, our interaction with document full-text does not always mark the termination point of search; often it marks a turning point where we use some aspect of the full-text for a new search or a refinement of an existing search. Moreover, there are examples within the multiple coordinated views literature that do display document full-text in a dedicated view (Hubmann-Haidvogel, Scharl, and Weichselbraun, 2009); however, this is counter to the design of contemporary search user interfaces. Use of dedicated views for document full-text is suggestive of the potential role linking and coordination could have in our search tools. As a result, one aspect of Chapter 6 will examine differences in user behaviour and ultimately, search outcome, under interfaces with different ways of presenting a document full-text view.

Tab. 2.5: Feature definition and description for survey of search user interface systems that incorporate an information visualisation component.

Feature Type	Feature Name	Feature Description
Attribute	Type	Information visualisation component type
	Name	The system's name
	Reference	Reference to system
	Data	Corpus or collection supported by system
	Insight	Expected understanding gathered by user
	Stage	Applicable stage of search process system applies to
	Preview	Facility to preview document surrogate information
Preview Attributes	Full-text	Facility to display document full-text
	POP	A pop-up window
	DSL	Dedicated screen location
	LIST	Rank-ordered list
	THUMB	Thumbnail as Technique
	AUG	Rank-ordered list augmentation
	NS	Not specified
Surrogate Attributes	T	Document title
	S	In-context snippet
	M	Miscellaneous meta data
	U	Uniform resource locator
	Th	Thumbnail preview
Full-text Attributes	NW	Opens in new application window
	NWT	Opens in new browser window or browser tab
	NS	Not specified

Tab. 2.6: A survey of search user interface systems that incorporate an information visualisation component.

Type	Name	Source	Data	Insight	Preview	Full-text
Spring	VIBE Net	(Koshman, 2005)	News	Clustering	POP-T	NW
	Information Navigator	(Carey, Kriwaczek, and Ruger, 2000)	News	Clustering	POP-S	NW
	Bubble World	(Berendonck and Jacobs, 2003)	News	Clustering	NS	NW
Map	InfoSky	(Andrews et al., 2002)	Web	Clustering	LAB-T	NS
	NS	(Tory, Sprague, et al., 2007)	Spatial	Clustering	NA	NA
	WebRat	(Granitzer et al., 2003)	Multiple	Clustering	NS	NWT
	Envision	(Nowell, 1997)	Academic	Metadata	DSL-TM	NW
Nominal	Search Crystal	(Spoerri, 2004)	Meta Search	Overlap of Source	POP-TSU	NS
	Sparkler	(Havre et al., 2001)	Meta Search	Q-D Sim. Overlap	DSL-T	NWT
	DART	(Cho and Myaeng, 2000)	Web	QD-Sim.	POP-TS	NWT
	RankSpiral	(Spoerri, 2004)	Meta Search	Q-D Sim. Overlap	POP-TS	NS
	EyePlover	http://www.vionto.com	Wikipedia	Topic Overlap	POP-S	NWT
Set	Scatter Gather	(Hearst and Pedersen, 1996)	News	Clustering	LIST-T	NS
	Carrot2	REF http://www.carrot2.org	Web	Clustering	LIST-TSU	NWT
	Grokker	(in Koshman, 2005)	Web	Clustering	POP-T	NWT
	ResultMaps	(Clarkson, Desai, and Foley, n.d.)	Digital Library	Clustering	POP-T	NWT
Tree or Graph	Quintura	(see Alhenshiri, Watters, and Shepherd, 2011)	Web	Connection	LIST-TSU	NWT
	TouchGraph	http://www.touchgraph.com	Web	Connection	DSL-TSU	NWT
	Kartoo	(in Koshman, 2005)	Web	Clustering	DSL-ThTSU	NTT
	NS	(Kobayashi et al., 2006)	Web	Clustering	POP-TSU	NS
	Walk2Web	http://www.walk2web.com	Web	Connection	THUMB	NWT
Ranked-list augmentation	R-Wheel	(Grewal et al., 2000)	Web	Q-D Sim.	NS	NS
	TileBars	(Hearst, 1995)	News	QD-Sim.	AUG	NW
	HotMap	(Hoerber and Yang, 2006)	Web	QD-Sim.	LIST-TSU	NWT
	Ujiko	(in Foenix-Riou, 2006)	Web	Clustering	DSL-TSU	NWT
	Nextplore	http://www.nextplore.com	Web	Page Genre	DSL-TSU	NWT
	SearchMe	(Ostergren, S. Yu, and Efthimiadis, 2010)	Web	Page Genre	THUMB	NWT

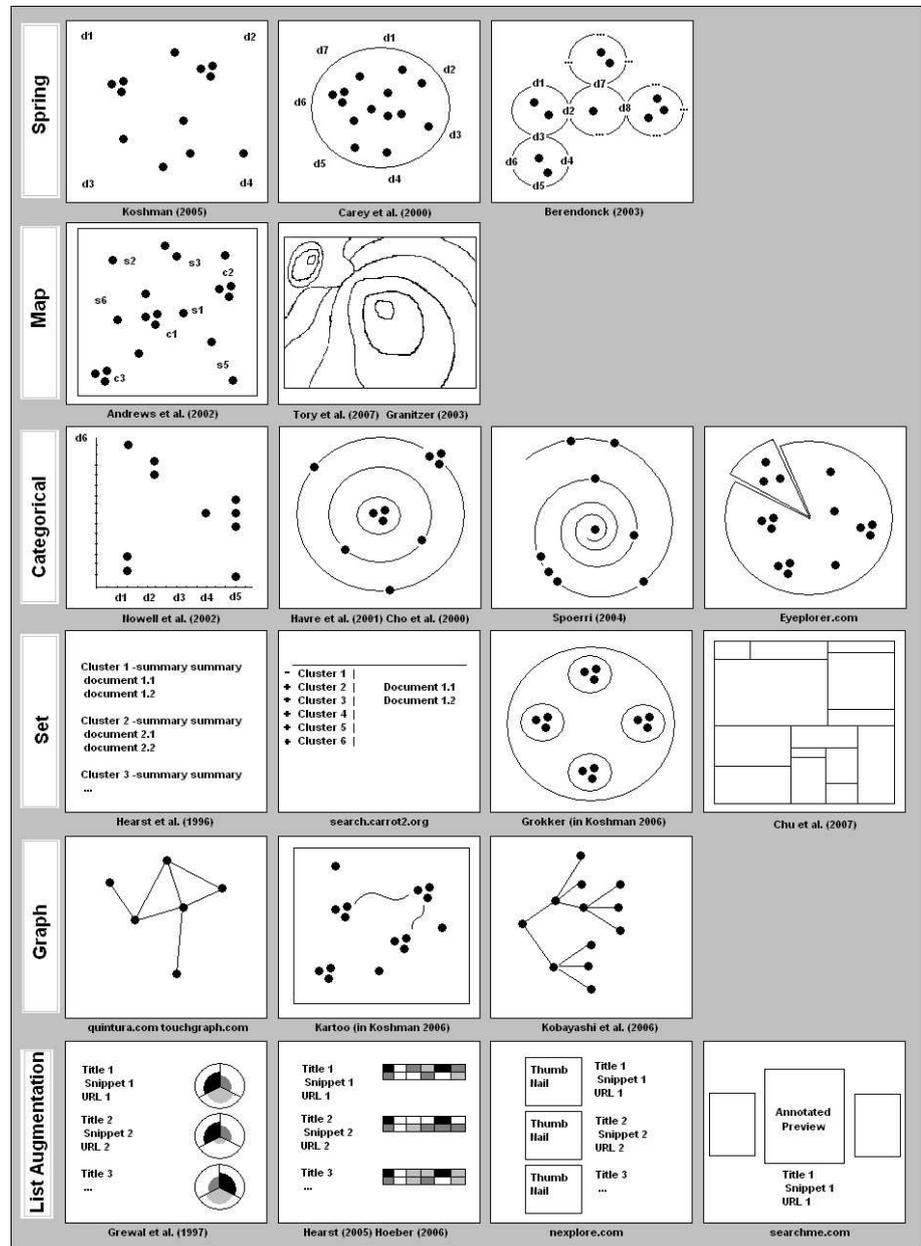


Fig. 2.2: A survey of search result interfaces incorporating a visualisation-based component; references below forward the reader to instances of the technique; note that in this survey, web-based systems have been favoured.

The following sections will break up discussion according to the matrix rows of Figure 2.2 on the preceding page. A brief critique is made for a selection of systems that were considered in this survey.

Spring

In the radial technique of Carey, Kriwaczek, and R uger (2000), dimensional anchors are placed around the edge of the circle and documents are positioned according to similarity to dimensions. Documents are connected to dimensions by imaginary springs. The force on each spring is interpreted as the similarity documents share with each dimension. Document position is calculated at the balance point of all pulling forces. A dimension can represent a query word, a theme or concept, or another metadata dimension that has a continuous scale; however, spring displays are not suitable for ordinal data.

Document location can be ambiguous given the large number of dimensions on the outer boundary, leading the user to question which of the dimensions contribute the most pulling force. The Information Navigator of Carey, Kriwaczek, and R uger facilitates highlighting of related dimensions by clicking on documents and conversely, highlighting related documents by clicking on dimensions. Interesting patterns to look for include clusters of documents that share common association with dimensions.

This layout is advantageous in that it provides the basis for proximity and similarity, leading to the perception of clusters; however, the Spring technique is not without its own disadvantages. Clustering is dependent on the positioning of terms and concepts around the perimeter of the outer boundary. Poor selection of terms and term positioning can result in poor clustering or even careful selection of terms can have unanticipated consequences as illustrated by Hemmje, Kunkel, and Willet (1994). Carey, Kriwaczek, and R uger suggest that highlighting terms influencing the position of document icons is a useful feature to counteract this usability problem.

In testing, Koshman (2006) finds that in general, novice groups have trouble decoding the visualisation. On the other hand, Morse, M. Lewis, and Olsen (2002) compare a ‘no-frills’ version of the spring technique within the context of a wider test that includes text-based equivalent tools. Their results indicate that as task complexity increases, so does the preference for visual or iconic based tools, such as a spring display.

Map

Scatter plots have strong connections with geographical visualisation and cartography, since horizontal and vertical dimensions on a map correspond to the orthogonal dimensions of latitude and longitude. Moreover, there is significant inspiration taken from natural formations in nature that influence the design of metaphor in cartographic-like

information visualisation (Wise, 1999). Examples include landscapes or ‘data mountains’ in which the mountain height or relief provides a encodable third spatial dimension.

However, in line with the earlier criticism of three-dimensional visualisation (Cockburn and McKenzie, 2001; Cockburn, 2004), Tory, Sprague, et al. (2007) compare two-dimensional point and landscape displays against three-dimensional landscape equivalents, and find that a simplified two-dimensional scatter plot to be routinely superior and the three-dimensional landscape to be routinely poor. Again this supports the notion that three dimensions, although looking pretty, is not a good idea.

Categorical

Categorical visualisations are among the most diverse class of techniques in terms of appearance. This class of techniques is devoted to the display of nominal or ordinal data. Document metadata is predominantly nominal and it is frequently the case that continuous or real valued metadata attributes like size or length are transformed into more manageable and useful intervals or categories.

Interface encoding legends will quickly become crowded with data categories of large size e.g. consider a topic metadata category encoded as colour; thus not all datasets are suited to such techniques. Clustering patterns may still be evident, but such patterns are interesting only if metadata carries important weight. In general, document metadata has only a small weight of relevance in comparison to keyword content, since we typically make relevance judgements over on a mass of keyword content initially, and over a few metadata subsequently - and seldom the other way around.

Nonetheless, metadata cues may contribute refinements. Loose heuristics may guide a search toward PDF documents e.g. for academic search; but if the search is conducted on an index of PDF documents then the type category is redundant. In contrast, the topic category may be too small or too large resulting in an overly weak code or overly heterogeneous display respectively. In addition, an automated topic extraction algorithm may not select the topics of interest and a constantly changing topic map scheme may disorient the user since they need to relearn the mapping every time their query changes. These techniques may be more appropriately suited for email systems or for document collections where highly weighted and reliable metadata is present.

Set

The class of set visualisations, despite their visual diversity, are among the more commercially popular ways to organise results in information tools. Furthermore, this class of technique characterises systems that are typically more successful and longer stand-

ing in commercial implementation; despite there being many such examples that are no longer active on the Internet.

The more text heavy set techniques, for example seen on Carrot2 <http://www.carrot2.org> are similar in nature to earlier directory style browsing interfaces, but are generated specifically for the result set. Such techniques aim to augment or replace long lists of items and to aid the searcher in attending to relevant subsets of the data. Each subset or cluster, has a list of sub clusters with each identified by a cluster label and a small set of documents. If the clustering technique employed produces a mixed membership clustering, meaning documents may appear in multiple clusters, then a user can visit each cluster in turn achieving serendipitous discovery. Interestingly, despite the presence of a comparatively large quantity of text, the text-based set techniques outperform the graphical based versions (Rivadeneira and Bederson, 2003) in both quantitative and subjective measures; although the differences were significant for subjective measures only.

Graph

Graph visualisation, or edge-node graphing, is among the most popular ways to demonstrate a relationship between objects in a dataset, particularly when tree and graph data are utilised for modelling.

Edges make explicit the idea that two objects are related somehow. The large degree of relatedness within a dataset quickly gives rise to emergent patterns like the perception of cliques of objects and the identification of weakest links. Such analyses are interesting, given the rise of social media and counter-terrorism. The interconnect- edness of the Internet, based on hyper-link structures, citation patterns in academic literature and the connections between words and themes in language, mean that the graph technique is invaluable for the presentation components of information tools. diGiacomo et al. (2008) compare multiple graph layouts specifically for search result visualisation. They find that the tree map style visualisation for hierarchical data outperforms that of the traditional tree visualisation techniques, whilst radial techniques lie somewhere in between.

Ranked-list Augmentation

Ranked-list augmentations are the final class of technique. Whilst all techniques thus far were considered an alternative presentation method, the set of ranked-list augmentation techniques are a descriptive or indicative tool to be used in combination with a ranked-list.

In this class, visual cues indicate the strength of query words to individual documents, the spread of keywords throughout the document - either by an abstract rep-

resentation of the document or via a reduced-form within context annotations - or a miniaturised screen shot of the result. These allow the user to estimate the distance between what they specified in the query and what the document contains. This allows the user to quickly scan for prominent features in the augmentation and focus attention on those particular results. Woodruff et al. (2001) argue that screen shot thumbnails allow searchers to identify genre and style very rapidly.

Survey Summary

It is apparent that some information domains are better accommodated by better search user interfaces, though the benefits afforded by what is good user interface design and what is search visualisation is not well defined. Each of the techniques in Figure 2.2 on page 59 show search results in a different way which may only be of benefit during the results review stage. In maintaining this observation, the ensuing chapters will focus heavily on techniques that show search results for review purposes.

In discussing the advantages of such systems, it is useful to choose one particular technique along with a set of tasks and contexts, and judge this combination against the advantages and disadvantages of the ranked-list. Thus, it is left to the reader, to select their own technique of interest, and to discuss and compare it against the advantages and disadvantages of the ranked-list.

2.5.4 The Ranked-list Result Presentation Paradigm

The de-facto standard for presenting a set of search results is a ranked-list, in cooperation with a ranking algorithm that endeavours to place the most relevant result at the top of the list. A ranked-list has several advantages. Its format is lean, ubiquitous and scalable, consistent, simple, usable and intrinsic, user inclusive and highly flexible. Such characteristics are easily measured and so early optimisations have realised in terms of speed of service in particular. Nevertheless, the ranked-list at present cannot support efficient and thorough information search for all search, given the high prevalence of partially relevant and irrelevant results, little indication of the relationships between documents (Kobayashi et al., 2006), and very limited controls that make use of semantic structure present in the result set. However, alternative result paradigms must possess the same lean, generic, consistent and inclusive characteristics, as well as a solution for each perceived shortcoming. A series of desirable characteristics is presented in Table 2.7 on the following page. This is an incomplete, yet still long list of advantages offered by de-facto ranked-list standard.

A parallel may be drawn between the ranked-list as a technique and notion that ‘good theories make predictions that go beyond the data they were meant to explain’ (Schick and Vaugn, 2002). If an agreed theory cannot explain or be applied to similar

Tab. 2.7: Desirable characteristics of a ranked-list interface.

Characteristic	Note
Clarity	A list interface has high clarity; text has a semi-predictable text-white space ratio; text does not overlap and optimisations for readability are possible.
Usability	A list interface has lower learning overheads; augmentations to list are evolution of the same fundamental idea.
Disability	A list interface is better suited users with visual disabilities whose experience is enabled through software like screen readers.
Fast	A list interface is predominantly hypertext; this is rendered fast - relative to graphics rendering that rely on additional software platforms or plug-in technologies.
Lean	Optimised mark-up and styling minimises bandwidth relative to plug-in enabled displays that require periodic download and update.
Ubiquitous	A list interface underpins search service of largest online search engines - Google, Bing/Microsoft, Yahoo etc.; population pool is familiar with using a list and this knowledge is transferable between services.
Scalability	A list interface can serve very large volumes of query traffic; constructing a query response i.e. the list interface can be optimised greatly.
Simplicity	A list interface is simple and straight forward; there are few extraneous functions and very few buttons and controls required for use.
Reliability	A list interface operates in much the same way across every search session.
Consistency	A list interface has a largely consistent look-and-feel between each result, each result page, each query, and each search session.
Trust	Searchers have an inherent trust that the search engine's first result will be the most useful.
Versatility	A list interface serves results for a range of query types including informational, transactional and navigational furthermore, a list interface supports and enables fast teleportation.

phenomena, then the theory is thrown out and a new one, most likely a refinement on the old, takes its place. By adopting this viewpoint, one may consider the ranked-list as the current theory or paradigm, the predictions or the outcome as the display of results, and the state of needing a paradigm shift as being a result of ranked-list interfaces, unable, in all cases, to display results in an optimal way.

A lack of optimal display is a complex, multi-faceted puzzle which reflects the multifaceted nature of search. Some facets of the puzzle may be solved by light-weight additions to the ranked-list (Hearst, 2006; Hoeber and Yang, 2006); however, alternative paradigms altogether are also one alternative *theory* to consider. Any paradigm shift will need not only to improve on the current sub-optimal cases, but also deliver the optimal cases offered by the ranked-list and at a consistent or superior level of service.

So why has the list format attracted so much dissent? Subsequent subsections present a series of undesirable characteristics, specific to the ranked-list paradigm. These include poor depiction of relationships between documents, poor understanding at the macro-level and micro-level, and unmet richness in interactive discourse. Furthermore, the cited disadvantages apply to particular search situations in which the search involves many documents and the information need is complex.

Poor Depiction of Inter-document Relationships

Poor depiction of inter-item relationships is a widely cited disadvantage of the ordered list (Kobayashi et al., 2006). The one-dimensional and ranked-list interface offers no strong textual or non-textual indication of semantic relationship between items at any two list positions. It is the searcher's duty to open each item that meets a minimum information scent, and to then make comparisons between each item.

Semantic relationships have several functions. They assist with locating items of a similar semantic content, assist with ignoring items of a similar semantic content, and may be used to infer semantic structures present in the topic area. The first two functions are practical in the sense that searchers use semantic relationships to guide attention away from irrelevant information and to focus attention to relevant information. In contrast, the third function is revealing. Semantic structures - composed of several items sharing semantic similarities - contribute to an overview perspective of the specific topic area. An overview tool promotes more efficient (Hornbäck, Bederson, and Plaisant, 2002) exploratory searching and may foster more thorough but incremental understanding of the topic space (Kules and Bederson, 2004), the information need, how to describe that information need in query terms, and in turn give direction to a searcher's exploration of the space.

Poor Understanding at the Macro Level

By macro-level, is meant regard given to the searcher's understanding of search results at a macro, global or big picture perspective. Macro level features depict the structures, relationships and patterns between documents. Such features are otherwise hidden by a keyhole view of the result set when served in paginated and ranked-list form.

Depicting relationships between ten items in an ordered list interface might not yield the functional and revealing benefits discussed in the section prior - there are simply too few results on display (Spoerri, 2004). For instance, a vague or ambiguous query may result in ten items of largely different semantic content a possible strategy adopted by search engine engineers to increase the chance that one item matches the intended meaning of the searcher (Agrawal et al., 2009). Alternatively, relationship depiction between ten partially-relevant or partially-related items may not provide a strong indication of the structures present in the topic space.

Different techniques deliver macro level understanding in different ways, though all depict or elucidate relationships between items. Such techniques rely on filtering actions and on emergent percepts. One may think of filtering actions as isolative, in that the searcher triggers the display of only items that have membership in semantic related groups. Alternatively, we might consider emergent percepts as partitioning, in that relationships are depicted visually with the aid of perceptual Gestalts among a background context. Accordingly, poor understanding at a macro level is a situation in which it is difficult to control the display of items meeting set criteria or an inability to recognise perceptually-defined structures through mere observation over global views.

Clustering interfaces - e.g. <http://carrot2.org> make up a majority of the longest commercially lasting alternative search user interfaces on the Internet. Searchers typically navigate clusters using a directory-tree visualisation that filters the result set out with the exception of those present in the selected cluster. Unlike the partitioning map-like interfaces, clustering interfaces for the most part do not support intersecting two or more clusters together. In contrast and inherently, partitioning interfaces allow this type of analysis; searchers may identify items that transcend multiple clusters by observation of items in proximity to each cluster centroid. Underpinning the demarcation of clusters in partitioning interfaces are perceptual Gestalts, which involve a collective of objects perceived together as a whole super object rather than individually. The perception of a perceptual Gestalt is practically instantaneous, consider for instance the pop-out effect discussed in Section 2.1.4 on page 28 - and is suggestive of a pattern or relationship in the data. However, it remains the searcher's task to determine the semantic nature of the Gestalt.

There is widespread subjective preference for global overviews and structured information as seen in both search tool interfaces (Wu, Fuller, and Wilkinson, 2001) and more widely, in focus plus context information visualisations (Hornbæk and Hertzum,

2011). Routinely however, what is novel, interesting or pretty is erroneously perceived as superior - despite significant objective measures indicating otherwise (deAngeli, Sutcliffe, and J. Hartmann, 2006).

Poor Understanding at the Micro Level

Conversely, by micro-level is meant the searcher's understanding of each individual search result. If the first page of results in a ranked-list interface provides a keyhole view into the search result set, then each document surrogate in the list provides a keyhole view into each document. Searchers base relevance judgements on the textual content present in document surrogates, until such time an interest threshold is exceeded and the document is opened for full-text view.

Document surrogate composition and configuration research (Drori and Alon, 2003; Haas et al., 2011; Aula, 2004) and eye-gaze fixation research (Balatsoukas and Ruthven, 2010; Granka, Joachims, and Gay, 2004) propose how document surrogates should be constructed and how searchers interact with them. It is well evident that searchers adopt various text scanning strategies depending on the circumstance. In some cases, item position and keyword highlighting fast track the search process; whilst, in other cases, layout and surrogate composition is critical.

Yet, an approach that adds additional descriptive text, keywords, or ontological labels to each item, only confounds fast text scanning strategies. Adding further text cues only increases the number of slow and serial decisions needed on one particular item. Moreover, a comparison between two items across pages, or items separated by great rank, is made more complicated by the increase of intermediate noise.

Conversely, additional yet appropriate information scent could be beneficial to searchers. Adding further relevance cues without text is possible by way of non-textual, visualisation-based cues that are effortlessly processed at a glance. Inherent in macro understanding and relationship depiction is that items in close proximity are likely to discuss the same semantic topics and themes. Accordingly, searchers may utilise these macro features at a glance to gain further insight into an item.

Example techniques, that largely fall under the list augmentations - see Section 2.5.3 on page 62 - include techniques such as TileBars (Hearst, 1995) and WordBars (Hoeber and Yang, 2006), which offer term distribution information expressed through visual codes. Both TileBars and WordBars in HotMap offer a hot and cold like code, which a searcher may leverage for fast evaluation. However, fast evaluation is operative; pattern recognition is important and instantaneous but, assigning meaning to those patterns is sometimes serial and effortful (C. Morris, Ebert, and Rheingans, 1999).

Such techniques are advantageous, a searcher need only make a loose or fuzzy comparison between an ideal TileBars or WordBars configuration, distribution pattern or

colour combination, and a the test item, in order to scan efficiently. In contrast, strict matching on colour or distribution patterns would likely slow the searcher down. Nevertheless, these techniques have an effect of widening the searcher's otherwise keyhole view into the document.

2.5.5 Perspectives on Alternative Result Presentation Paradigms

This section will outlay the advantages and disadvantages of alternative result presentation paradigms. However, this discussion will specifically refer to paradigms that make use of spatial information to convey inter-document relationships based on semantic or thematic content. Accordingly, this criterion excludes from discussion, systems in the list augmentation paradigm and the meta-search visualisations listed in Figure 2.2 on page 59.

Prominent advantages include the ability to organise search results by thematic and semantic similarity; to reveal patterns in metadata across the result set; to display more results at a glance and more than the 7-10 accessible at a glance in a ranked-list; and overall increases to the information scent available to the searcher.

Organisation by Thematic Content

The first notable disadvantage of the ranked-list is that a one-dimensional list consisting of documents containing a variety of genres and themes (Kobayashi et al., 2006) does not reveal any strong non-textual indication of the relationships between documents. The advantage of doing so is apparent when having found a relevant document, a ranked-list technique offers no indication as to where the next relevant result may be located. In contrast, techniques that gather related results into local regions of a visualisation offer such an advantage.

Whilst significant advances have been made toward visualisation-based techniques that reveal document and query similarity (Hearst, 2009; Hoeber and Yang, 2006), research suggests that searchers subjectively prefer alternative result presentation techniques that offer spatially structured results (Wu, Fuller, and Wilkinson, 2001). Yet, research also suggests that objective measures paint a different picture: what is novel, interesting or pretty is perceived to be a superior product, despite significant objective measures indicating otherwise (deAngeli, Sutcliffe, and J. Hartmann, 2006). This suggests that more work is required to build interfaces that searchers want to use, and which pose an objective advantage over ranked-list interfaces.

Organisation by Metadata

Secondly, a ranked-list does not allow the searcher to understand intuitively the features or dimensionality of the entire result set (Kobayashi et al., 2006) at a glance.

Whilst theme or topic may be considered a dimension of the dataset, this discussion separates thematic dimensionality from the broader idea of metadata dimensions. A metadata dimension includes file type, size, authors, the presence of images etc. but is highly dependent on the indexed media. Context determines the influence metadata has on its utility for search and not all metadata are necessarily as useful in different contexts.

However, when metadata are important to the searcher's task, a combination of both thematic and metadata search criteria is necessary. If a long set of results are presented in a list and pre-filtered for metadata the list must still be traversed in order to locate the correct result. In contrast, if the same list of results is organised according to thematic content and the searcher is able to orient their search to within a neighbourhood of the collection, the search is likely to terminate faster as the searcher is able to ignore large portions of the set in quick succession.

Global and Local Keyhole Views

A ranked-list displays a limited set of results at a time. In order to see more results the searcher must scroll down or click to the next page of results. In this sense, while a document surrogate provides a keyhole view into the thematic content of a specific document, a ranked-list which presents around 7-10 results on a screen at any one time, provides a keyhole view into the result collection

Such a keyhole view is further exacerbated by ranking algorithm strategies that favour diversity in result sets in order to increase the likelihood that a user will see additional important words in the first few results that they can then use to re-query. While it is often possible to configure one's preference for more results per page, the searcher must still scroll down the page to view those results. Scrolling between result frames makes it difficult to maintain earlier interesting surrogates in memory together with subsequently interesting surrogates - though this is potentially alleviated by adopting a window tabbing strategy as observed by Huang, Lin, and White (2012). This issue is referred to as information overlook by Carpineto, Osinski, et al. (2009).

In contrast, alternative presentation methods generally show all or a large proportion of the results in one area of the screen and upon finding one relevant document, the searcher is able to look around or branch out from the area surrounding that interesting result. Alternatively, if a result warrants the retrieval of an earlier seen result, Niemelä and Saariluoma (2003) suggest that the semantic organisation of the space will make the process of backtracking to that earlier relevant item more efficient.

Efforts to overcome poor global overview in ranked-lists have been directed to the provision of interactive controls and colour queues. For instance, WordBars (Hoerber and Yang, 2008) offers a sidebar term frequency histogram visualisation based on

filtered keywords extracted from snippet and title text. Both colour intensity and histogram bar length indicates term frequency; the histogram depicts only the top twenty or forty most frequent terms in the top 100 documents returned by the search engine. Clicking on keywords in the histogram re-sorts the ranked-list with documents containing selected keywords, appearing higher in the list. Double clicking words adds to the initial query and following the click of the search button a new set of results is returned.

WordBars offers both a global and local overview of a ranked-list of results, in that a view of each result is available to the searcher, while use of interactive controls enable the searcher an insight into different aspects of the information space according to different keyword choices. Accordingly, a searcher can enter unavoidably vague search query, and then re-sort the results until an appropriate aspect of that query surfaces. In this light, the selection of keywords for ranked-list re-sorting actions is similar to shifting eye-gaze from region to region within a spatialisation-based search visualisation, since the searcher's eye-gaze will typically move outwards from an important region to nearby spatial regions based on the thematic content queues as indicated by spatial annotations. In a spatialisation-based visualisation, eye-gaze shifting embodies the re-sorting action of the ranked-list as document icons present in the centre of visual field become the top ranked-documents and the region's thematic content and keyword annotations become the re-sorting criteria under examination by the searcher.

Limitations on Information Scent

Another disadvantage of the ranked-list relates to the availability of information scent. Like most search user interfaces, ranked-list based interfaces offer document surrogates that display a few key sentences and metadata. Although beyond document surrogates and rank information, this is the extent of information scent offered to searchers. Alternative result presentation techniques typically provide different and multiple sources of information scent in addition to document surrogates.

A core proposition embodied in the research presented in Chapter 6 will be that alternative presentation techniques should support ranked-list like text scanning strategies in combination with a variety of other information scents that the searcher can intermittently draw upon in order to focus their search. In support of this proposition, a greater emphasis should be placed on the implementation of pop-up windows in order to achieve this proposal.

Traditionally, details-on-demand (Shneiderman, 1996) has facilitated users with a way to obtain information about specific items in an information visualisation, by clicking or mouse hovering over an icon or glyph. Typically only one pop-up window is on display at any one time, and a user must serially visit each icon in a visualisation to see pop-up windows for each icon of interest. This facility suits analytical information visualisation tasks fine, in that users spend a majority of time looking for global patterns

involving many icons and might only need to visit a very small number of icons if at all.

In contrast, in a document search context, document surrogate information, if presented in pop-up windows, would necessitate the opening of a large number of pop-up windows in order to facilitate search. In this light, a detail-on-demand interaction paradigm would thus slow search down. Therefore, what is needed, is a way to show as many pop-ups as possible, without cluttering the visualisation interface, and without obstructing visual information otherwise hidden by a large number of pop-up windows on the display.

An open research question then, is to demonstrate the feasibility of such an approach, and to devise a pop-up implementation strategy that facilitates fast scanning of both document surrogates as well as structural cues offered by the visualisation. Such an approach is contingent on an adequate level of pop-up window transparency in order to facilitate interpretation of information in the foreground i.e. the pop-up window text, and interpretation of the background information i.e. the spatial structure.

Rich Interactive Discourse

Interactions between user and computer may be construed as dialogue; the user conveys what it is they need by way of input fields and controls, after which the computer responds by updating the interface appropriately. For the most part, the interactive dialogue between a searcher and a ranked-list interface is equivalent to alternative search user interfaces. However, alternatives afford greater richness in the interactivity; they are better positioned to leverage future gesture based interaction and afford a more direct and natural communication between searcher and computer.

Pike et al. (2009) describe analytic discourse as involving mutual feedback between the goal-oriented user and the interactive controls of a visual interface. Mutual feedback is the flow of a user's intent expressed by interactions and the flow of information from the interface to the observer because of those interactions. In this case, a discourse refers to the user's description of their need via the interactive capabilities of the interface; the tool's response is evident in the act of modifying the presented information.

However, unlike in analytics interfaces where the analytics interface does not necessarily ask questions of the user - it really only answers questions - contemporary search interfaces pose questions to the searcher in a more personified way. Thus, the discourse is not so one-sided; a searcher types their query and the search engine replies with a list of results and some further clarifying statements. Like humans, search engines are not mind readers and require clarification before they arrive at the best suggestions. Clarifying statements include, for example, did you mean ...? much alike how an expert would respond in regards to a request for resources on a particular topic.

Query suggestions e.g. *did you mean...?* and query auto-complete bring searchers closer to their goal by providing a convenient way to build, complete, or modify a query regardless of whether the searcher knows what to write. However, like filter and sort controls and user feedback mechanisms, these controls act as dialogue conjugates, down which searchers convey intentions. When this interaction is intuitive and almost unconscious the interaction is natural and the searchers cognitive burden devoted to the task rather than trying to deal with an uncooperative team member.

Filters, sorting, item selection, full-text view, and positive and negative feedback are interaction controls associated with search user interfaces. What searchers achieve interactively in the ranked-list interface is likewise achievable in alternative search user interfaces with the exception of sorting controls. Yet, alternative search user interfaces offer more efficient and natural interaction. For instance, selection of multiple, semantically related listed items necessitates a sequential traversal of the list; in contrast, techniques organising items semantically enables the searcher to select several documents in quick succession within the area identifying with the semantic content of interest.

Although searchers face an initial overhead of finding a semantically rich and relevant area or region, this process assembles the searchers model of the topic space along the way. In contrast, the searcher has little assistance with assembling such a model when using a ranked-list as list position has little or no semantic value. If the searcher needs or wants to build a model of the topic space while engaging with a ranked-list, they must maintain a large number of results in memory. In contrast, high memory load is not critical in a spatialised interface since the spatial layout is constant and always visible they need only shift gaze to reactivate portions of their model further. Evidence to suggest robust model building using spatial layout is fairly robust (Cockburn and McKenzie, 2001; Niemelä and Saariluoma, 2003).

The advent of touch and gesture is changing computing remarkably. A rich gestural language is possible (Elias, Westerman, and Haggerty, 2007), although one agreed language is not yet ratified. Clustered visualisation interfaces are well suited for gestural interaction leading to more natural interaction with results. Consider for example, selecting items from a list versus items from a visualisation. In the list, the searcher must find and select items in sequence perhaps separated by irrelevant results and across pages; in a visualisation, a searcher can select multiple items in a semantically relevant area of the space. Encircling items is not only more efficient, but the process of circling to form an active group that can furthermore have actions called upon it is far more natural. In the case of the list, some way is needed to track items selected for actioning in perhaps a dedicated screen location; once this list is formed a dedicated button or control is needed to enact further action. The two architectures are functionally equivalent; however, there is a disconnected between button-executed actions, versus gesture-executed actions on a group of items. The interaction on the list may be said

to be exhibiting low-stimulus response compatibility (F. Lee and Chan, 2005). The interaction not only results in a richer experience, but it also invites the chance of reducing expenditure of screen real estate. Furthermore, the searcher can be more efficient selecting whole groups of items for feedback purposes such as deleting items of little relevance or inviting a re-clustering of documents based on the selected relevant items.

2.5.6 *Meeting the Advantages of the Current Ranked-list Paradigm*

It was outlined how alternative result presentation paradigms overcome the identified disadvantages of the ranked-list. Though the one-size-fits-all philosophy of the ranked-list does not suit every type of search; there are a number of advantages that suggest why the list is so prominent and utilised by default. These advantages were presented in an earlier section - Section 2.5.4 on page 63 - and discussed relative to alternative result presentation paradigms. More widespread adoption of alternative techniques may be possible in future, but only if tool designers blend the advantages of the ranked-list paradigms together with the purported advantages of alternative techniques. A direct comparison to Table 2.7 on page 64 is included at the end of this section in Table 2.8 on page 76 and discussed more generally in this section as a whole.

Since the ranked-list is presently optimising on characteristics such as speed and general usability (Zamir and Etzioni, 1999), scalability (C. Chen, 2005), and support for individual abilities and disabilities (C. Chen and Börner, 2002; Leporini, Andonico, and Buzzi, 2004), alternative paradigms in the least, must address the full list of advantages presented in Table 2.7 on page 64 in order to offer a credible alternative for searchers. Regardless however, the ranked-list will likely remain the de-facto result presentation paradigm, at least over the medium term.

Instead of replacing the ranked-list paradigm altogether, alternative result paradigms could be triggered to serve a search detected and classified as overly complex. For instance, search engines already trigger customised search user interfaces in the case of searching for famous personalities, businesses and general topics and generic facts. Customised result templates augment traditional search results by providing a range of media including encyclopaedic references, images, videos, and related searches. By extension, queries that are more complicated could trigger alternative, non-linear presentation templates, which would better suit the needs of a searcher facing a complex information need. However, based on the low usage statistics for advanced search features in search engines (Spink, Wolfram, et al., 2001) the process of choosing a presentation method must be automated as it is misguided to assume that if given the option, users may ignore more sophisticated and helpful interfaces that are accessed at the click of an additional button or through the use of a specific search operator.

From the perspective of the searcher, among the most important system utility

metrics are speed and reliability. The simplicity of ranked-lists is due in part to the generic text-heavy HTML mark-up that drives the display of results. In combination with style sheets, the formatting can be changed quickly and easily by engineers. These generic technologies are fast, highly optimised, mostly standardised and reliable.

In contrast, prior to the current Java Script renaissance (Lerner, 2007) and prior to the HTML5 standards, drawing interactive graphics on a web page involved plug-ins like Flash and Java applets in order to run pre-compiled visualisation software. These technologies involve an inherent start-up and initialisation time, require additional system resources, and furthermore have been subject to a regular update cycle. However, the adoption of new open-standard technologies is experiencing a shift from propriety browser plug-ins (T. Lee, 2011) to more JavaScript driven technologies and updated HTML standards. At this point however, the jury remains out on whether client-side, native and plug-in agnostic technologies will completely replace propriety browser plug-in technology (J. Lee, 2008). It is assumed here that future visualisation applications will be programmed from this lightweight style of technology.

Assuming that browser-native technologies are adopted for search result visualisation systems, such systems will perpetually impose operating overheads over the searcher in order to deliver those simply more sophisticated techniques. It is by far more complex and resource intensive, to create code and to send it to the remote client in order to plot a visualisation, than it is to create a list of results using HTML mark-up. This will likely always be the case. However, as improvements in communications technology continue to enable faster network bandwidth, such problems should alleviate the communications delay, but the simplicity of the ranked-list will perpetually necessitate lower overheads at the client side. In addition, native visualisation technology i.e. HTML5-Canvas appears to be outperforming the plug-in based technologies.

D. Johnson and Jankun-Kelly, 2008 experiment with a variety of different browser based technologies and find that the new native JavaScript driven visualisation technology out performs plug-in technology in terms of set up and layout time under a number of configurations. They considered datasets of different sizes; the smallest dataset, a half a megabyte, which is more than what one would encounter with a text-based result set, while the medium and large data sets were four mega bytes and eighty megabytes respectively. The key message here is that technology and browsers are becoming increasingly powerful and capable of hosting sophisticated client-side processing which is both rich and interactive.

In regards to other advantages of the ranked-list, such as clarity of format, despite a multitude of techniques dealing with occlusion (Ellis and Dix, 2007), spatially-organised visualisations of large document sets tend to be quite busy and cluttered. Thus, in order to achieve result presentation that has high clarity, the processing and layout requires several layers of processing to ensure that the document set is easily scannable and navigable.

In contrast, ranked-list paradigms tend to have a uniform look, feel and layout and this makes it easier for searchers coming from different search engines. Consistency across different search engines makes it familiar while reducing the learning curve of a new tool to practically nothing, as the searcher will know how to operate the search engine based on prior experience. Furthermore, the ranked-list format is more appropriate for users with visual disabilities as text-based versions feed directly into reading aides. However, this does not preclude the possibility of new accessibility standards for alternative presentation formats. Furthermore, this motivates the need for the design of more natural visualisation systems and this will feature prominently in Chapter 4.

Tab. 2.8: To what degree do alternative result presentation techniques meet the desirable characteristics of a ranked-list interface.

Characteristic	Note
Clarity	Visualisation-based systems do not offer the same predictable clarity of presentation; however, occlusion techniques and strategies are available.
Usability	Visualisation-based systems have not been afforded the industry and mainstream attention as the ranked-list; accordingly, much of the research regarding usability of such systems remains.
Disability	There are seemingly few standards that assist a person with a visual disability during their interaction with visualisation-based search interfaces; this is likely to change quickly following mainstream adoption.
Fast	Current algorithm intensive visualisation processes and specialised plug-in technologies place visualisation-based systems at a disadvantage to fast text-based systems.
Lean	Visualisation-based systems are an emerging technology; the foundations upon which they are built remain incomplete; as new improvements are made, the delivery of search results via visualisation-based systems will likely achieve efficiency gains.
Ubiquitous	Visualisation-based systems are not ubiquitous and the few existing systems are heterogeneous in appearance; many of the skills from one system will not apply to another - though this will likely change with the emergence of a system that achieves the right balance between visual and textual.
Scalability	Visualisation-based systems that are heavily dependent on layout algorithms may not scale well; efficiencies are expected with the advent of new technology and algorithms.
Simplicity	Visualisation-based systems offer alternative interaction modalities that do not apply well to text; new modalities may simplify interaction.
Reliability	Expectations of visualisation-based system reliability, functionality and predictability are likely to improve with future development to the point where a visualisation-based system will work much in the same manner across search sessions; the current push to develop such systems means that a leading alternative standard - to the ranked-list - has not yet eventuated.
Consistency	Visualisation-based layouts can be vastly different between results depending on visual encoding, different perspectives of the information space may be different.
Trust	Visualisation-based systems are new; simply abandoning one tool that works for the most part in place of a new unknown tool is an unlikely proposal - trust in a search system takes time and positive search experiences.
Versatility	Visualisation-based systems are well-suited to specialised and exploratory search, they will always remain overkill for every day tasks.

2.5.7 *The Absence of Guidelines for Spatially-Organised Results*

Based on the survey of systems above, a consideration of the advantages and disadvantages of the ranked-list, and the spatialisation literature; there are three specific gaps in our understanding of search tool designs that engage alternative result presentation techniques.

Firstly, what is the optimal design choice for the way in which a searcher accesses and interacts with a full-text version of the search result; secondly, what are optimal design choices for pop-up windows that support exploration within information spaces; and thirdly, what are optimal design choices for the interactive controls that facilitate navigation through spatialised data and information spaces? Chapter 6 will present an experimental interface and the results of a human-based trial that seeks to investigate these gaps in our understanding of how to build more effective search tools.

These three specific gaps are exploratory in nature; these open research questions offer a contribution toward innovative search tool designs that combine how searchers engage contemporary search tools with search behaviours exhibited when using alternative result presentation systems; these research gaps are significant as they propose to investigate ideas that seemingly have not already been investigated at length.

Of the survey of systems proposed in Figure 2.2 on page 59 and Table 2.6 on page 58 above, no system has displayed - in a search context - the multi-pop-up window strategy as it was outlined, and nearly all systems utilise separate windows and frames for display of the full-text view of a search result which is disconnected from the search tool. In addition, little research has compared the suitability of information visualisation-based interactive components to support exploration of hyper-dimensional spaces; one major exception however is Elmqvist, Dragicevic, and Fekete (2008).

To argue that we should accept the current state of search engine result visualisation is to stifle innovation. Stronger criticisms pertain to proposing new search result visualisation techniques that stifle cognitively efficient presentation of information. At present, there are no alternative result presentation methods that can achieve the same level of versatility that the rank list methods have. However, a credible and sizeable niche does exist in the search community for these alternative interfaces to serve within.

Given that we can improve interfaces with more structured approaches, can we expect that users will use them to their full potential? Turetken and Sharda (2005) in their study draw on research by Todd and Benbasat (1994) to explain their neutral subjective agreement amongst alternative systems despite significant speed gains in visualisation based systems. Todd and Benbasat noticed that users did not spend additional time analysing problems when given tools that are sufficiently more powerful than others. Rather, they found that participants were using sufficiently more powerful tools to maintain a similar level of performance but over shorter periods. This poses

an interesting advantage for the ranked-list interface as speed ramps up, users will be no less satisfied as they are reaching the same level of performance in smaller time.

2.6 Summary

This chapter provides a foundation for designing, implementing and evaluating alternative result presentation paradigms. This foundation consists of three main aspects: a perceptual framework based on Rodrigues et al. (2007), visualisation of document attributes using appearance and geometric attributes of glyphs - in particular the use of motion to encode data and the role of natural representation of data attribute and visual attribute - and finally a spatial arrangement of documents that reveals inter-document relationships. New result presentation paradigms, or in the least, improvements to existing experimental interfaces are needed to account for the deficiencies in contemporary information tools such as web-based search engines.

Early on, discussion characterised the tools we utilise as part of our every-day information-driven lives and furthermore, the role information visualisation plays in the design of these tools. Discussion then established a context of use for these tools covering our information needs and intents that drive the engagement of tools, an idea of relevance as an intermediary between need and satisfaction of need, and established a model of information-seeking behaviour to systematise our search behaviours spanning: conception of information need, the engagement of information tools, and resolution of the need. In this context of search, it was suggested that we face a significant overabundance of information and that our tools cannot completely and adequately deal with this problem. Following this, three reasons were suggested as to why our tools cannot deal with this situation but emphasised that the way we present search results to searchers is among the most significant.

It was proposed that a holistic solution for better result presentation should entail document attribute visualisation as well as the representation of inter-document relationships. Consequently, this structured the discussion for the remainder of this chapter in three subsequent sections: the need for a perceptual framework, document attribute visualisation, and visualisation of inter-document relationships. The Visual Expression Process of Rodrigues et al. (2007) was adopted to explain the perceptual processing that underpins our interaction with alternative presentation paradigms. This framework stipulated the perception of information visualisation in three stages: Conception, Observation and Interpretation; however, not all interpretations were applicable for interaction with alternative result presentation paradigms - as the task-goals for information analysis by information visualisation and information search are fundamentally different. This framework outlined the perceptual processing of individual document attributes encoded into appearance and geometric attributes of glyphs as well as the perception of relationships between a collective of glyphs.

Subsequently, discussion examined document attribute visualisation by means of encoding data attributes into appearance and geometric properties of glyphs. A set of preliminary topics provided context for two research foci: the role of motion frequency as an encoding dimension; and a proposal to optimise data encoding rules based on a cognitive congruency between the data attribute and the graphical attribute.

In relation to motion, it was noted that despite earlier research into the use of motion in visualisation applications, most of this has been directed toward encoding with the phase dimension of motion and little work specifically to do with the frequency dimension of motion. In relation to natural encoding, little research exists on whether encoding paradigms result in superior task outcomes, if encoding rules are based on the degree of natural fit between - or intuitiveness of - data attributes and their representing graphical attribute.

These foci will be developed further in Chapter 3 and Chapter 4 respectively. Both chapters will report the results of human-factors experiments that intend to gather empirical evidence in support of these proposals. The discussion of these proposals was spatial-layout-agnostic, as this would feature prominently in a subsequent section.

The remaining discussion focused on techniques and systems that depict inter-document relationships based on primarily semantic or thematic content. First, the current result presentation paradigm, the ranked-list paradigm, was discussed and some advantages and disadvantages were outlined. Following this, a survey of systems revealed diversity in existing alternative result presentation paradigms and techniques; this survey is intended to encourage future design and evaluative investigations; however, it also motivated a discussion of the advantages of alternative result presentation paradigms. Such discussion suggested that all alternative paradigms, in their present form, are inconsistent with the way users conduct search. This may explain the lack of widespread adoption of such systems, in part due to an overly literal interpretation of information visualisation - since an overly literal interpretation disregards contemporary search behaviours. These observations will be developed further in Chapter 6.

The aim of this research is to improve current and future search tools. This dissertation strongly contends that in order to do so, we need to involve the human searcher interacting with alternative result presentation paradigms to complete search-tasks that current search tools do not sufficiently support. Toward this aim, later chapters will report empirical findings from three experiments that support - or otherwise - the proposals outlined here. If supported, the outcomes will indicate:

- The use of motion frequency for encoding data attributes in appearance and geometric attributes of glyphs should be considered for future encoding paradigms;
- Cognitive congruency between graphical attribute and data attribute should be considered when motivating encoding rules for document attribute visualisation;

- Integrating the full-text of the resource into the result interface is an important design consideration;
- Integrating document surrogates into thematic maps is an important design consideration; and,
- Interactive controls are needed to facilitate result set filtering based on dimensional rotations of multidimensional semantic spaces.

3. ON THE ROLE OF MOTION IN ATTRIBUTE VISUALISATION

3.1 Introduction

This chapter reports the results of an experiment that had two main aims. The primary aim was to investigate the use of motion frequency to encode data in a document metadata visualisation. The second aim was to evaluate the utility of a web-based data collection methodology. Such a methodology was expected to deliver a diverse participant demographic and to reduce time, resource usage and financial costs that are associated with conducting like research in a laboratory setting.

Chapter 2 outlined a process of encoding data attributes using appearance and geometric attributes of glyphs. The discussion noted that a rich palette of graphical attributes is available for encoding, but these attributes are typically static in nature. Above all, this chapter asks whether there is any benefit in expanding this palette to include dynamic or motion attributes and in particular using motion frequency to encode data attributes. Empirical investigations into document attribute visualisation are present in the literature (e.g. Nowell, 1997); but these only hint to the potential of motion as a device to encode data. Furthermore, the empirical investigations that examine motion frequency (e.g. Ware and Bobrow, 2004; Ware and Limoges, 1994) have perhaps unfairly dismissed frequency as an encoder of data, due to the nature of their respective experimental tasks.

Participants took part in this experiment over the Internet at a web site featuring an embedded Java applet. Yet, this is but one of many online delivery mechanisms through which to target a broad demographic of online users (Kittur, Chi, and Suh, 2008; Ross et al., 2010). In recent years, the explosion of social networking application platforms, an interest in utilising games for the purposes of research and data collection (vonAhn, Kedia, and Blum, 2006), and use of games for the improvement of cognitive function (Kearney, 2005; Gamberini et al., 2008) have opened up new recruitment channels through which to conduct research. Moreover, web based experiment hubs (Krantz, 2012) and research by crowd sourcing (Kittur, Chi, and Suh, 2008; Paolacci, Chandler, and Ipeirotis, 2010) have provided an additional gateway to a large and potentially diverse participant demographic. These platforms and hubs, which do come with both benefits (e.g. Reips, 2001) and weaknesses (e.g. Paolacci, Chandler, and Ipeirotis, 2010) have promoted but perhaps not yet persuaded the idea of web based experimentation to the academic mainstream.

This chapter begins with a discussion of the motivations for conducting research into motion and under what circumstances this research applies. The discussion will then turn to methodological considerations for conducting research online. The design of the experiment will follow in addition to a report of results. Following this, a discussion will interpret the results and discuss their significance. If the results favour the use of motion frequency, the conclusion will propose guidelines regarding motion frequency in a data encoding capacity. Similarly, a reflection on experiences and lessons learned regarding web based platforms for experimentation, will also be offered.

3.2 *Motivation for Research*

The main application providing impetus for this research centres on software tools that specialise in the representation of information-bearing entities projected into a spatial area. Utilising these displays, human users extract data from one or more visually specified criteria and subsequently use it to direct a search activity. These displays are found in search tools that incorporate a document metadata visualisation.

In the envisioned application, a spatialisation algorithm arranges each document onto a two-dimensional map according to dominant semantic themes. Each document has a set of metadata, which are encoded by appearance and geometric attributes of a glyph; document metadata can include file type, genre, format, source and age.

There is a need to build tools that present spatially arranged entities characterised by high dimensional metadata sets and to support a range of search tasks of varying complexity. The expected task sets span search for simple through to complex criteria. For example, search for simple criteria include locating all documents of a specific theme; search for complex criteria include locating all recent, academic documents about a topic with specific contextual keywords.

The importance of and presentation of metadata varies across domains of search. In open-domain search for instance, there is a greater, but not total reliance on title and keyword-in-context snippet text, rather than byte size, word length, source or file type (Balatsoukas and Ruthven, 2010). But, when metadata is important to full-text searching, secondary to establishing thematic content, the search typically warrants categorical or generalised values like file size as small or large size, topic cluster size as small or large cluster, age as old or new, multimedia as present or not-present, domain source as commercial or government, and rarely precise values such as file size in kilobytes, or exact word length. Such categories can be neatly depicted in a document metadata visualisation.

More research is required to devise perceptually efficient and usable visualisations given the potentially large pool of metadata available. Whilst Chapter 4 will address the idea of decoding efficiency, this chapter will address the size of the metadata pool

and specifically investigate ways to ensure all available metadata may be assigned a visual representation. By using dynamic graphical attributes, we can potentially double the breadth of the graphical encoding palette; however, at present, we have only a limited understanding of the use of dynamic attributes in information visualisation. This research will attempt to understand further the role of motion frequency as an encoding technique.

The next section will establish a basis for how far we should extend the breadth of this encoding. This basis refers to the number of graphical features we can manipulate in a glyph and how many variations of a single graphical feature we can manipulate - theoretically - before we should expect degraded task performance.

3.3 *The Breadth of Reliable Encoding*

A set of data attributes visualised as an integral visual object i.e. the glyph, defines a visual chunk of information. In the case of 3 visually-encoded data types e.g. size, age, format, each of two alternatives big or small, old or new, PDF or HTML, according to an interpretation of (Miller, 1956) by proponents Powers and Pfitzner (2003), the human cognitive system can - without overwhelming itself - reliably identify a target chunk from a set of 2^3 or 8 unique attribute combinations quite easily.

Consider a target glyph of three properties: small area, red and square shaped. The area, colour and shape are visual attributes of the glyph and the configurations small, red and square are instances of the visual attributes. When interpreting the limits of reliable judgement, the exponent should be taken to mean the number of visual attributes and the base number to mean the number of choices or instances for each attribute. Therefore, 2^3 should be taken to mean three graphical attributes, each varying by two instances e.g. area: small or large; colour: red or blue; and shape: square or circular. Furthermore, the result of $2^3 = 8$, gives the number of unique combinations or unique looking glyphs in the visualisation area. When looking at a display for a specific configuration or target, the user should be able to recognise reliably, a target amongst distractors when the number of possible alternatives is small.

By generalisation, 3^2 or 9 unique attribute combinations, or $3^2 \times 2^4$ chunks, i.e. a chunk defined by three graphical attribute classes each containing two attribute instances and three other graphical attribute classes each consisting of two attribute instances, could be handled by the cognitive system without too greater overload up to the point of 150 distinct chunks. However, reaching such lofty thresholds entails a large class of attributes with many attribute instances that cannot be identified reliably.

In order to reach a level of 150 identifiable categories, the experimental data supporting Miller's claim were taken from an experiment investigating 5000 possible combinations, leaving the remaining 99% as confusable chunks. It was seemingly Miller's

intention to argue that the limit of cognitive processing might be around $2^{7\pm 2}$ binary decisions. This research suggests that if we encode data attributes with graphical attributes adequately, such that items are uniquely identifiable from each other, then we should maximise the chances of making reliable judgements about the presence or absence of items in a visualisation.

Miller emphasized that we can reliably identify more chunks when there are more distinguishing graphical attribute classes and attribute instances, but that we should not expect to hit the theoretical limits prescribed by multiplying the capacities of single attribute classes. For instance, if the capacity of reliable hue detection is $2^{3.1}$ bits and another attribute class like size has a reliable detection of $2^{2.8}$ bits we should not expect to detect without error all 60 i.e. $2^{2.8} \times 2^{3.1}$ bits, distinguishable chunks composed of around 8-9 colours and 6-7 sizes. With the addition of a third dimension of screen location, - Miller reports $2^{4.3}$ - a perfect channel throughput would result in a reliable identification of over one thousand chunks without error i.e. $2^{2.8} \times 2^{3.1} \times 2^{4.3}$; yet this is not the case.

While it is accepted that all perceptual feature processing has a capacity limitation of some form (Shiffrin and Nosofsky, 1994), reasons beyond ‘bit processing’ capacity can explain why such limitations are present. In the case of grey scale perception for instance, the human perceptual system is prone to the interference of surrounding colour context as demonstrated by the Checker Shadow Illusion (Adelson, 1995) and judgements of size can also be biased based on environmental context as is typified by size constancy illusions (see Weiten, 2001, pg. 154-155).

To apply this idea to a realistic context, consider visual search on a display for a complex target icon, defined by some hue and size at some screen location. In this example, the participant has pre-encoded the target to memory, perhaps through extended use, such that the encoding paradigm is meaningful. The participant’s response is usually to: label the stimulus - as is typical in experiments discussed by Miller (1956); to reproduce precisely the stimulus set (Klemmer and F. Frick, 1953); or to report the presence or absence of the target (Wolfe, 1998). In a metadata visualisation, the response is generally a target present or not-present response that triggers further action relevant to that target or a continuation of the visual search. If the participant subconsciously labels the target as a non-target, inferred by a target-not-present response in a target-present trial, then according to an interpretation of Miller, the confusion is reflective of an overload of processing capacity. To increase the likelihood of reliable detection, Miller argues that we should increase the number of distinguishing features or dimensions; however, we should expect not to do better than around 150 chunks.

Redundant coding of data, whereby two or more features encode a single data value is consistent with this approach; by providing an additional way to distinguish a target we can increase the likelihood of reaching maximum channel throughput.

More recently, commentators and research are critical of the prevalence of the 7 ± 2 capacity of working memory, stating that people have misinterpreted the message of Miller's work in areas such as menu design and power point presentation design (Doumont, 2002; Farrington, 2011). For instance, it is a misconception that a menu should contain no more than 7 ± 2 items. Instead, cues to promote faster reading of the menu items such as sort order are more applicable to such design. Shiffrin and Nosofsky (1994) recount that while 5-9 alternatives is about right for limitations on uni-dimensional decisions, it does not hold for multi-dimensional decisions in the way as claimed by Miller. Shiffrin and Nosofsky advocate the view of Monahan and Lockhead (1977), and suggest that an increase in the number of identifiable objects is due to distance in psychological similarity space. Nevertheless, this view is consistent with the idea of reducing the number of ways a glyph can be differentiated on a single dimension and furthermore, increasing the number of dimensions on which a glyph can be differentiated.

Psychology similarity refers to the supposition that each stimulus can be assigned a unique location in a coordinate space, based on its characteristic dimensions or features. The distance between points indicates the similarity of those two stimuli. A similar approach is taken to spatialise documents within a coordinate system and can rely on the same analytical techniques such as Multidimensional Scaling (Borg and Geoenen, 2005; Buja et al., 2008) to do so. Conceptually, objects that are similar and near to each other will be confused, while distinct objects will be separable and lay far apart in the coordinate space.

The contemporary perspective of all of this is exemplified by visual search experiments as was introduced in Chapter 2. Where as Miller focuses on reliability of judgement or accuracy, visual search experiments are interested in both accuracy and time taken i.e. search efficiency, to investigate the limitations of human visual attention and perception. However, presently, while theories of human perception remain incomplete (Wolfe, 2007) they do advocate a number of scenarios where search for a known or unknown target will be most efficient and these can motivate the role of design choices in visualisation based search tools.

3.3.1 *Use of Perception Research for Encoding*

Using the findings from efficient search scenarios, designers advocate the use of the pop-out effect - i.e. target-distractor heterogeneity - to highlight outliers, perceptual grouping - i.e. target-target homogeneity - to focus attention to subsets of a data set, and the use of perceptually efficient features to encode the most important information.

Often, the pop-out effect is explicitly manufactured in visualisations by encoding the most important information with the most efficiently processed graphical features to avoid the potential for inefficient visual search. By highlighting items according to

thresholds, visually biased items will be the most readily interpreted information, since they stand out or pop-out prominently. This leads to faster insight and easier visual inspection and manipulation of the data. However, in the case of search tools that we utilise routinely, the type of information scent that we rely upon to find documents that satisfy information need, periodically changes depending on the information activity around the time of tool engagement. Furthermore, search criteria may routinely make use of complex multi-dimensional relevance judgements and search criteria may change mid way through the review of a search result set. Since, it is not possible to determine reliably what the moment-to-moment interests of the user are - without explicit interaction from the user - an unsolicited pop-out effect may not always benefit the searcher. For instance, a searcher may not always rely on the search engine's notion of importance if a searcher is restricting themselves to a particular thematic neighbourhood.

On the other hand, every time we interact with a tool we expect it to work in a more or less consistent fashion. Moreover, it is a violation of user interface guidelines to change the consistency of interaction across multiple interactions with the same interface (Galitz, 2007). Thus, if the tool decides to encode data differently across sessions based on a static, system-derived idea of importance - which is not flexible with the user's dynamically changing idea of importance - then each time, the user must re-orientate to the session's encoding rule set. Even so, research has investigated the use of perceptually motivated designs where the user provides an explicit notion of importance to the system in advance of use of the tool.

Healey, Amant, and Elhaddad (1999) propose the ViA perceptual assistant to optimize visualisation construction based on an interview of the user; the interview queries the user about the data set and their expected analytical outcomes and this feeds into decision-making about the selection of encoding rules for a visualisation under construction. Additionally, Yost and North (2005) suggest a work around for slow conjunctive search scenarios, showing that it is sometimes better to present multiple views of the same information entities differing on data feature in each view, with each view utilising the demonstrably superior colour hue feature to encode data.

Different again, Bartram and Ware (2002) demonstrate that motion phase could provide another work around for inefficient search on integrated visual chunks. By utilising motion phase as a way to visually group similar items in a display, the searcher can efficiently search each group of complex visual chunks for a target chunk, thereby avoiding the need to separate each chunk into constituent features across multiple views. Resultantly, the user perceives emergent groups of objects based on different motion phase characteristics. More recently, Hulleman and McWilliams (2011) indicate that it may be more appropriate to describe this phenomenon in terms of motion-based depth stratification, analogous to the parallax effect, in which near objects move fast and far objects move slowly. Under this interpretation, motion groups appear to move quickly

in the foreground ‘in front’ of static items perceived as forming the background. This interpretation appears consistent with the subjective responses outlined by Bartram including the unanimous report that static items fell away - i.e. into the background - from motion defined groups. While the interpretation between authors is different, the outcome appears the same.

Further work by Ware and Bobrow (2004) has utilised motion highlighting in a graph navigation application and found superior results for motion over more traditional static colour coding for localising areas within a large graph link-node visualisation.

The work of Healey, Amant, and Elhaddad, Yost and North, Bartram and Ware and research of a similar nature that incorporates perceptual cues to link data over a primarily global view of a data set - i.e. linking multiple entities by way of visual patterns over spatially distant areas - does not apply to the present research. Moreover, it is excessive, particularly in the case of (Bartram and Ware, 2002), to set filtering constraints explicitly, by way of interactive controls, for a limited number of items in a local area of a visualisation. Instead, such techniques are better suited to tasks, as was explored in the aforementioned research, in which there is a need to link spatially distant items. Because in search we generally and primarily use semantic information first - which here we assume such information will be depicted by spatial relationships - and we do not rely upon graphical features that depict metadata in order to narrow down a search space. Rather, having narrowed search to a local area of the visualisation, subsequent search may then be devoted to the interpretation and comparison of a small group of alternatives for a target icon, perhaps using a metadata attribute.

This research is inspired by old and guided by new. Miller suggests that there is a processing capacity limit to perceptual-cognitive processes and that to optimise throughput we should increase the number of dimensions on which to make judgements. Monahan and Lockhead suggest that limitations relate to the individuality or uniqueness of visual chunks and not a capacity limitation as such i.e. what is different is more easily recognised. Furthermore, more recent visual search literature seeks to explain successful search outcomes based on the role of low-level feature detection and controlling or guiding attention processes. Some features ‘guide’ attention (Wolfe, 2007) more so than other features allowing us to reach our target more quickly than a combination of others - a sort of divide and conquer facilitator. In the case of insufficient guidance to a target consisting of a conjunction of attributes, potentially due to target and distractor similarity, an extended and serial search must take place to decode each answer.

Miller suggests that regardless of time and given sufficient numbers of compositional features and consequently large separation in psychological space per Monahan and Lockhead, a number of targets are identifiable without confusion up to the point he believes is the capacity limitation of the human cognitive system. The key point here is the reference to time and error. Miller intended to observe the point at which

participants get confused - even when having an unlimited amount of time to make a decision - and so favours the confusion rate or error rate when the target is present. In contrast, the visual search literature predominantly favours both time and error for either target present or target absent trials particularly with increasing stimuli set size. Indeed the application of the visual search literature to the information visualisation field seeks to favour time as well, such that a selection of encoding features should enable fast and efficient detection of patterns, while promoting easy visual manipulation of visual information for operations such as visual filtering and querying. In the present application however, there are no guarantees that perceptual patterns will assist with guiding search among a few alternatives in a local area of a visualisation, and so time and accuracy will be reliant on how easily targets or their constituent attributes are committed to short term memory and then compared against alternatives in the stimulus set.

As raised by Powers and Pfitzner (2003), extending chunk composition to include dynamic graphical attributes is a natural extension to Miller's idea; with more ways to differentiate a chunk, the greater the number of attributes that maybe correctly identified - and consequently, the greater the separation in psychology similarity space.

In this light, more graphical attribute classes are required - influencing target-distractor heterogeneity - with fewer attribute instances - influencing target-distractor homogeneity - in order to maximise separability without exceeding the capacity limitations of a single feature. Enter, dynamic graphical attribute classes.

Even if Miller's analysis of absolute judgements of multi-dimensional stimuli is not applicable to a contemporary understanding (Shiffrin and Nosofsky, 1994), it still stands as a motivation for empirical research to ascertain graphical chunks that are reliably identifiable and to form composition guidelines to this effect. Extending chunk composition to include dynamic graphical attributes may be one way to increase the number of reliably detected encoded attributes. An experiment is needed to test this motivation.

3.4 *Motion in Encoding*

Predominantly, motion based cues in information visualisation have either been non-existent or utilised for purposes restricted to maintaining awareness between state transitions. Bartram's (Bartram, 1997; Bartram, 1998; Bartram, 2001) treatment of the richness of motion encompasses: basic properties e.g. phase, frequency, direction, amplitude, duration; interpretive properties e.g. signalling; and compound properties of motion e.g. sequences. However, the present research is only concerned with the basic dimensions of motion that include phase, frequency, direction, amplitude, duration, smoothness and shape.

More precisely, the experiment below seeks to explore the coding of data utilising frequency. How frequency is encoded into a static feature is specific to the feature type,

so further discussion of the expressiveness of motion for feature types - e.g. flashing, zooming, shuffling - is left to a subsequent section - see Section 3.6.1 on page 107. Next, however, discussion will seek to elaborate on the few research studies that have looked at motion frequency as an encoder of data.

Frequency coding has received minimal research attention in the information visualisation field, yet the human factors field has investigated the use of motion for quite some time now. In the information visualisation field, five prominent investigations are Limoges, Ware, and Knight (1989), Bartram, Ware, and Calvert (2003), Ware and Bobrow (2006), Weigmann et al. (2004) and Huber and Healey (2005).

The work of Limoges, Ware, and Knight (1989) is used as evidence to support the claim - e.g. by Bartram, Ware, and Calvert (2003) - that the perception of a range of frequencies is poor. However, in the case of Limoges, Ware, and Knight, the experiment task necessitates consideration and comparison of a group of items in order to make an estimation of correlation between groups. It may be that frequency does not facilitate perceptual Gestalts in the way that phase, duration, and amplitude do, which is integral for the perception of correlation of multiple items. This should not discount the use of frequency as an encoder of data in which the task necessitates extraction of data regarding a single item, since - as it will be seen shortly - human-factors research reveals that we can make judgements about frequency quite effectively. Furthermore, later research (Bartram, Ware, and Calvert, 2003; Ware and Bobrow, 2006) reiterates the suitability of motion phase for perceptual grouping purposes.

In regards to motion frequency, studies of the Just Noticeable Difference between frequencies, address the perceptibility of frequency more directly. Data for flashing stimuli is among the most prevalent in investigations on frequency judgement, perhaps given the widespread notion that flashing lights are suited to alarm states, and so one must be careful to make generalisations regarding the perceptibility of motion frequency across all motion types.

Two practical examples in which flashing light frequency is used to encode information includes that by Laxar and Luria (1990) in which frequency is encoded to proximity or distance to a channel boundary in a ship navigation task while the second relates to encoding of vehicle deceleration rate in brake lights on passenger vehicles (J. Lee, 2008). Tolin and Ryen (1986) reaffirmed earlier research findings that the Just Noticeable Difference in flashing lights is around 5% on a range between 1Hz and 6Hz. Similarly, Laxar and Luria (1990) find 24 noticeable frequencies - 4% separation over 6.7Hz - of flashing when encoding proximity to a channel boundary in a ship navigation task.

However, Tolin and Ryen (1986) indicate that the design of such experiments investigating frequency discrimination has a large influence on the outcome, which is important to mention here. They suggest that if judgements are made between two

side-by-side stimuli, then the noticeable difference is smaller and around 5%. However, and relevant to the present experiment, if the participant is comparing alternatives in an unordered sequence with a target frequency stored in short-term memory then targets and non targets need a minimum of approximately 20% separation (Mortimer & Kupec 1983 in Tolin and Ryen (1986)). These results are reported in terms of Weber ratios. Weber's law states that the just noticeable difference between two stimuli is constant. Therefore, if the JND for frequency is 21% per Mortimer and Kupec, the target flash rate is 1Hz and the flash rate of a stimulus under consideration is 1.25Hz and should be detected greater than 50% of the time but not for say 1.10Hz. Similarly, if the target flash rate is 2.0Hz then a flashing stimulus at 2.42Hz should be detected more readily than chance. If we adopt a conservative value of 50% then four flash rates 0.25Hz, 0.5Hz, 1.0Hz, 2.0Hz could be reliably detected - with much better than chance likelihood - and if 0Hz is adopted, this could extend the palette to five flashing instances to encode data. This number is in line with the present intentions to extend the number of graphical attributes consisting of only a few graphical attribute instances over which to make comparisons in line with Miller's original proposal.

These empirical data suggest that the intuition that motion frequency is a poor encoder of data may be hasty and that the perceptibility of frequency has been underestimated in favour of motion phase, direction and shape and routinely from a perspective of perceptual grouping and linking over large spatial distances. However, the influence of task plays a key role in perceptibility, thus this is a tentative assertion and is worthy of further investigation.

Prior work in animation use recommends judicious use of animation so is there a risk of encoding data with too much motion? Risks can be evaluated from multiple aspects including distraction efficacy - i.e. control of attention - irritation, and conveyance of a message and weighed against the consequences of an observer not receiving the message or the irritation suffered by too many interruptions.

The use of motion to control attention and benefit search for targets has a lengthy history in human factors research for a range of applications including warning systems or alarm state systems such as brake light warning systems (J. Lee, 2008), navigation lights (Laxar and Luria, 1990), flight control displays (Thackray and Touchstone, 1990), and map displays (vanOrden, Divita, and Shim, 1993). These examples utilise motion cues when normal operating characteristics are violated because motion is perceptually dominating. However, Woods (1995) illustrates the futility of alarm panels consisting of perceptually dominating signals that occur under high cognitive task load due to task pressures e.g. danger to life and equipment, that do not carry informative messages about the exceeded threshold or indicate why thresholds were exceeded. Thus, alarm states are not just about the control of attention but also the conveyance of a message.

Lowe (2003) explores the perceptual domination issue at length from a different perspective, finding that dominant static and dynamic features impedes learning and

mental-model building of a system such as features in weather maps. Novice observers focus attention on perceptually-dominating features, at the expense of equally important but less salient features. Additionally, situational awareness and supervisory control interfaces tend now to incorporate spatial and pictorial representation over which alarm visualisations are spatially located (Bennet, 1993; Weigmann et al., 2004) rather than large arrays of conceptually-grouped, backlit light panels as is discussed by Woods. Such pictorial displays are referred to as mimic displays (Bennet, 1993) and improve situation awareness of state and causal factors, and facilitate better error resolution by providing an overview of the system suggesting alternative resources to restore services and by providing immediate feedback in response to interaction (Bennet, 1993). Furthermore, the application of motion to represent the flow of information or resources through mimic displays demonstrably improves task performance (Bennet, 1993; Weigmann et al., 2004).

For the present research there is no causality implied between objects. The region of the interface that an icon is assigned to is by nature of the document content and application of spatialisation algorithms. The relationship between nearby articles is based on similarity to a central theme. Inter-regional comparisons are of no great consequence although items situated between two regions may indicate a similarity between two themes, but interpreting precise pixel distance as a reliable measure of proportionality of each theme is of little practical help to the search effort. In such displays, there is little realistic advantage of reasoning about two densely populated regions on opposite sides of a theme map, other than to speculate that the dominant keywords in one region are not present in the opposite region - and likewise the reverse.

In general, search for textual documents involves semantic information first and metadata second; thus, localising search to a spatial region occurs first, before selections on metadata may occur second. The use of screen location first and metadata second is simulated in the study of vanOrden, Divita, and Shim (1993) who asked participants to estimate the quadrant containing the largest proportion of target items. They found that there was no interference by flashing distractors during search for static targets despite the intuition that flashing motions are distracting Thackray and Touchstone, 1990 and generally motion in peripheral dual-task performance configurations (Ware, Bonner, et al., 1992; Maglio and C. Campbell, 2000; McCrickard, Catrambone, and Stasko, 2001; Bartram, Ware, and Calvert, 2003).

Interestingly, Bartram, Ware, and Calvert (2003) find that not all motions are equally distracting and importantly, irritating. Their guidelines indicate that motions that transition across the screen as opposed to moving about a central location - except for zooming motions - result in more negative subjective ratings of irritability. Thus, perhaps irritation could be reduced by judicious choice of motion type for encoding data. Critically, the motion types under exploration in the present experiment will be centralised or anchored.

This discussion has reasoned that since earlier investigations into motion in information visualisation applications have focused on perceptual grouping of information, this has painted motion frequency coding in an unfavourable way. It was discussed how in other application domains and as investigated through human factors research, the human visual perceptual system can perceive with reliability, a number of different motion frequencies in the form of flashing lights. An investigation into motion frequency use, in a task that does not favour an experiment task that relies on perceptual grouping to guide attention, is needed to explore how suitable motion frequency is for data encoding.

The discussion now turns to the methodological considerations for web-based experiment delivery.

3.5 *Web-Based Experimentation*

Undoubtedly, the Internet is an important tool for experimental research and there have been and are presently, many examples of Internet-based experimentation (Reips, 2001; Krantz, 2012). However, anecdotally, a large proportion of this experimentation is survey-based in nature, involving relatively generic collection methods such as web forms.

This section will discuss a motivation for conducting web-based experiments that call for web-technologies beyond simple web form methods. Later, in Chapter 7, a summative discussion will be provided, from the experiences, observations, and outcomes, which have arisen throughout this course of research.

3.5.1 *Experimenter Motivations*

There are several motivating reasons to adopt a web-based experimentation methodology. Firstly, to reach diverse participant demography, beyond exclusively computer science students; secondly, to make efficient use of limited resources; and thirdly, earlier research (e.g. Pfitzner, 2009) demonstrates the feasibility and appeal of doing research by a web-based experimentation methodology.

Due to the ubiquitous nature of information search, diverse participant demography is desirable to ensure that search tool design is generalised to the wider community engaged in search and to ensure that these types of tools are not simply specialised to a computer science or engineering student. Whilst this situation could be improved by extending recruiting efforts beyond computer science and engineering disciplines of the university, it may still exclude searchers who engage similar search behaviours but in different contexts. Examples include business, strategic, legal, patent, and defence analysts at government and corporate levels but also includes those who have personal

research interests for general subject matters. For an extreme commentary on this matter, Henrich, Heine, and Norenzayan (2010) propose that a large portion of cognitive psychology research based on experiment populations in Western, Educated, Industrialized, Rich and Democratic society, may not necessarily generalise to the full global village.

Admittedly, however, whilst the present research more thoroughly embraces the above philosophy to research, it was initially assumed due to organisational constraints. Unlike in some university faculties, there were no course credit for participation arrangements in place for undergraduate computer science and engineering topics in the department where this research was conducted.

In absence of course credit, an alternative attractor for research participation is monetary reward. However, obtaining financial grants through external collaborations using industry partners is not always possible - particularly for small-scale research projects. Even limited financial support can go some distance through use of prize draws as a reward model. These include the chance of winning a tangible prize, or shopping voucher, or a small amount of money. Whilst perhaps better than no tangible offering, a prize draw does imply that the participant will have to wait until the experiment's conclusion to receive their reward which may be unappealing.

Participant motivation may not necessarily be solely determined by monetary reward. A subsequent section will look at non-monetary rewards that promote an immediately rewarding experience in return for experiment participation; these types of reward models do not require monetary resources as such and thus may appeal to experimenters in a tight financial position.

A lack of monetary aspect is a primarily motivator for the lean online approach, as it is reported in Treharne, Pfitzner, et al. (2008). However, even with unlimited resources to spend on participation rewards, there are additional motivations as to why research should be conducted outside of closed-door laboratories, using an online methodology. In the least, an online approach is more accessible, portable, and convenient, and may capture interactions that are more natural.

In terms of accessibility, the participant does not have to get himself or herself to a location designated by the experimenter; and if using generic web technologies like embedded applets and JavaScript and other ubiquitous technologies, particularly those that are used to perform every day applications like internet banking or to play videos and online games embedded in web pages, participants do not have to download and install additional software. In addition to accessibility, online is more convenient. Not only do participants not have to arrive at a specific physical location, they can participate when it is most convenient for them, since participants are not bound to a specific time slot in which the experimenter is available.

From another perspective, an online approach enables the experimenter to approach

the participant, as opposed to the participant approaching the experimenter, such as in the case where an experimenter sets up a participation booth at an expo or in a public space. In addition to novelty value for the public, such a demonstration provides a unique way to attract interest in the research program and more widely the academic discipline, and provides a discussion point on the source organisation. Furthermore, public installations attempt to capture participants who may not be exposed to other recruitment methods such as noticeboard advertisements.

Variations on the travelling display include unattended information booths but also kiosk style installations. The kiosk technology driving such displays is simply an Internet connected device. A leading benefit of this approach is that any data collected during public display is sent securely to a central location, thereby negating the need for manual handling of the data when returning to the office.

Finally, a study by Reilly and Inkpen (2007) finds that making use of a highly controlled experiment setting can influence an experimental outcome. In this study, they compare the results of an experiment conducted in a ‘white room’ laboratory where all factors are as best controlled for against the same research conducted in a public space where they believe participants interact with an apparatus more casually. While the motivations to conduct research in a controlled environment, seek to remove external influences that can influence the observation of a variable, they suggest that a combination of social context issues and arousal, due to background interferences, are responsible for the differences between the lab and the public space. In the public space, their results suggest that performance is better for interfaces that require less carefully focused attention than the interfaces that require more focused attention. The casually interacted interfaces, result in significantly better performance in a public space. In contrast, performance on the focused interface is better in the laboratory, though it is not significantly better.

3.5.2 *Online Laboratories*

There are a range of online services at the disposal of the experimenter; these include tools to build online surveys and services through which to crowd-source research participation (see Kittur, Chi, and Suh, 2008; Paolacci, Chandler, and Ipeirotis, 2010). In addition, there have been several efforts to build online, web-based experiment laboratories. Reips (2001) discusses the Web Experimental Laboratory now at <http://www.wexlist.net/>, which was first conceived around 1994. Similarly, Krantz (2012) maintains a list of online experiments at <http://psych.hanover.edu/Research/exponnet.html> for a variety of disciplines and topics; much of the international participation for the experiment reported in this chapter and that in Chapter 4, may be attributed to this website.

A popular crowd-sourcing resource is The Mechanical Turk <https://www.mturk.com>.

com/. Mechanical Turk is one example of crowd-sourcing, which engages the help of a large community of people to solve many small sub-problems of a much larger problem that cannot be completed by computers alone. ‘Requesters’, qualifying for the Mechanical Turk service, break up a large problem into several small sub problems such as image or product classification, speech transcription, survey collection or general data entry which are unable to be completed using current computing technology or which requires the opinions of a human. These tasks are floated onto a market place consisting of ‘Workers’ who complete these sub tasks for a micro-reward of typically a few cents. With around a million workers available, a large dataset may be collected very quickly and very cheaply - relative to research conducted in a laboratory. Yet, despite the benefits to research, online crowd-sourcing has received significant criticism from multiple perspectives including: underpayment of workers (Ross et al., 2010), unethical use of micro-payment manipulations in order to entice work (Horton and Chilton, 2010), a homogeneous worker population (Ross et al., 2010), the threat of dubious workers pocketing payments in return for bogus data (Kittur, Chi, and Suh, 2008) and or inattentive workers (Paolacci, Chandler, and Ipeirotis, 2010).

While Requesters have the facility to restrict participation only to reputable Mechanical Turk workers who have earlier provided quality data and received reputation points for doing so, procedural controls are also possible to counteract suspect participation. ‘Trap Events’ or ‘Catch Events’ (Paolacci, Chandler, and Ipeirotis, 2010) are experiment trials that are used to gauge whether participants are paying close attention to the experiment. Oppenheimer, Meyvis, and Davidenko (2009) explore the utility of catch events masked as reading tasks. In their catch event test, a participant is asked to read a long text on ‘Sports Participation’ and then answer a question below the text, which asks which of a list of sports the participant likes to engage in regularly. Inattentive participants are caught out when they respond to the question incorrectly had they read the entire blurb, they would have noticed the final few sentences instructing them not to reply to the question and to move on to the next stage of the experiment.

From an altogether different perspective, Harward et al. (2008) discuss the iLab, which is used as an online laboratory - not for experimentation on a population, but as a way for students to engage in learning from a distance. The iLab connects students with expensive or specialised equipment that their education institution would not otherwise be able to provide. A parallel is drawn here between online laboratories for research and those for pedagogy, in the sense that technology is connecting people at remote locations to engage in sophisticated activities all in the comfort of home, library or a computer laboratory on the other side of the globe.

3.5.3 Apparatus and Delivery Mechanisms

Traditionally, psychological experiments and even computer-based usability experiments take place in controlled, closed-door laboratories in the presence of a research assistant who guides the participant through the session. Clearly, some experiments, such as those that require sophisticated measurement instruments like electroencephalograms, are not appropriate for online experiments. However, many experiments investigating a range of usability ideas involving portable experiment software that measure quantitative performance metrics, which can be embedded in web pages are potentially suited to web-based delivery.

Two remote delivery methodologies were considered in this series of research: the first is web-based, while the second is kiosk-based. However, both rely on client-server architectures and therefore share a great deal of similarity with each other.

The first mode of delivery considered is the web-based model. Initially, a simplified version of this framework started with the survey-based work of the nWords, rWords and inFields research by Pfitzner (2009), but later evolved to applet-based and HTML5 technologies as reported in Treharne, Pfitzner, et al. (2008).

The web-based delivery model involves a website containing an introductory page and the necessary information to ensure informed consent is acquired, training materials, demographics forms, a Java applet containing the experiment apparatus, an exit questionnaire form and finally a debriefing page. Web form data and experiment data is recorded to flat file via a dynamic scripting language over web sockets. Alternatives including a database management system were considered; however for experiments of limited scope, the effort expended through setting up, defining database structures and writing code to communicate with databases was overly excessive and flat file recording of data was sufficient.

The second mode of delivery was a kiosk style facility, which consisted of precisely the same set up described by the web-based delivery mechanism above. The main difference between the web-delivered and kiosk delivered was the targeting of a specific physical location at which participants could complete the experiment. A computer and monitor containing a version of the web-based experiment were placed in the corner of a university computing laboratory with a sign inviting the public to participate in a research experiment. On the wall above this kiosk hung a board-mounted poster discussing the research thereby providing additional information. Anecdotally, this second mode of delivery was found to be trying - participation rate is typically patchy, full completion rate also rare, and there is high potential for participants to leave the kiosk in a state not fit for a subsequent participant.

3.5.4 *Data Integrity*

This discussion has mostly considered the advantages of the web-based methodology; however, such a paradigm attracts a number of challenges and in particular, the reliability and integrity of the collected data.

With a disconnect between participant and experimenter, researchers Pfitzner, Treharne, and Powers (2008) notice that an incidence of bogus, offensive, comical, and unintentional data, is much higher in experiments conducted over the Internet.

In regards to environmental conditions at the time of data capture, there is no reliable way to measure and characterise interruptions that a participant experiences throughout the course of an online experiment. Further, there are only blunt ways to measure why a participant perseveres to the end of an experiment, knowing that they would not achieve a financial reward. Moreover, again, there are few ways to understand the reasons why participants exit online experiments prematurely. Clearly, we could add additional questions at the end of the experiment, requesting an estimation of the user's level of interruption, and reasons for participating - though this is far from a truly objective measure. In support of early exit detection and request for reason, a browser close event could trigger a request for response.

In contrast, when conducting research under supervised, closed-door and paid reward conditions - such as that in Chapter 6, there is very little evidence to suggest an incidence of participants 'gaming' the experiment session by providing bogus answers to achieve a quick financial reward without great effort. Researchers (see McCormac et al., 2012) have utilised general knowledge questionnaires and comprehension tests at the end of experiment trials in order to catch out bogus trials. However, but anecdotally, not one spurious result has been detected using these measures. Such a proposal is not suitable for an online experiment however; as it is an additional stage in the experiment, the participant must endure particularly if the experiment participant is not financially rewarded.

3.5.5 *Participation Rewards*

There is no such thing as a free lunch or so goes the adage. Efforts to attract and maintain participation in the present research were directed at selling a rewarding experience in place of a financial reward. From this perspective, there are at least three options: to thoroughly implicate participants in the outcome of the research; to provide a satisfying experience as a result of their participation by providing a game-like feel to the apparatus; or to provide a usable tool that participants retain at the end of the experiment for ongoing use.

In the first case, while a few sentences and statements of the aims, goals and intentions can hint at the ramifications of the experiment in relation to the participant, it

is challenging to tailor to circumstances and typical search activities of each individual participant. In addition, based on a researcher's own desire to participate in other online experiments out of interest and moral support, some participants may also be participating as a result of interest in web-based experiments. Thus, selling the web-based methodology is also important as is providing a capable and interesting website and experiment.

In the second case, vonAhn and Dabbish (2004) and vonAhn, Kedia, and Blum (2006) are proponents of using fun, challenging or competitive games for the purposes of collecting research data. vonAhn, Kedia, and Blum propose a game to collect simple and generic facts suitable for ontology building using sentence templates like X is a kind of Y where X and Y are replaced with real words during a trial. The participant's response is to answer true or false; for every correct response, the player receives a serving of points and for every incorrect response, a player has points taken away. The game is played with two players remotely or against a computer, and thus elicits fun through competitiveness. Their evaluation captured 267 people playing the game for an average of 23.58 minutes each session. In another game, developed with the same underlying philosophy, 13,630 people were captured playing the game over a period of four months and 80% of participants returned to play again. In the end, an astounding 1.3 million pieces of data were produced for the research.

Game-based learning in educational settings is increasingly prevalent. Research is driven by the presumed motivating factors that games lend to students in classroom situations including making learning fun and increasing enthusiasm to learn. For instance, Rosas et al. (2003) find that educational games increase attention, concentration and self-regulation of learning, and promote improved classroom dynamics. On the other extreme however, the use of computer games in education is not without controversy; researchers raise concerns with negative social impacts stemming from gaming, including addiction, violence and social isolation. Though, Griffiths (2002) argues that negative consequences found in student usage of computer games, occur in a minority of game players who make excessive use of computer games. In addition, Rosas et al. find that students maintain normal play activities during scheduled lunch breaks, hence alleviating isolation and addiction concerns. Furthermore, it is widely apparent that the use of violent games has little or no educational benefit whatsoever.

As the primary focus of this online experiment in this chapter is on the improvement of data encoding - and the development of an online methodology a secondary focus - not every methodological factor including participant reward can be subjected to comparative analysis. The reward scheme under examination in this chapter is based on performance monitoring and gaming/entertainment. While there are many financial reward schemes at the disposal of researchers - such as monetary payments, micro-payments in the case of Mechanical Turk or CrowdFlower, vouchers, and prize draws - no financial assistance was secured in this instance. However, had financial support

been forth coming, it would have been accepted.

3.5.6 *Ethical Considerations*

Like any other experiment involving human subjects, an ethical approach to experimentation must be embraced. Fortunately, research of this nature reported in these chapters is of a relatively benign nature with little perceived advantage for a deception device in the experiment design, beyond restricting the participant from seeing the competing display configurations under examination.

Azar (2000), Keller and S. Lee (2003), and Skitka and Sargis (2006) discuss a range of ethical issues applied to an online research context. Most relevant to all experiments and therefore this course of research is solicitation of informed consent, debriefing, and anonymity. Whilst further ethical complications may arise when a research topic moves toward socially and personally sensitive topics, such complications are beyond the scope of the present discussion.

Any participant in any experiment must always consent to their participation with full knowledge of the experimental methods, time inconveniences, personal demands, discomforts, and risks of the experiment. However, there is no guarantee that any participant, in any location around the world, completely understands the experiment's description and introductory information. For example, it is conceivable that a novice English speaker could participate in the experiment without fully understanding the methods and risks. This has implications for both the safety of the participant, the validity of the data, and the ongoing belief system of the participant at the conclusion of the experiment. Ultimately, however, if the participant feels uncomfortable, as Azar (2000) highlights, the participant may exit the experiment at any time by closing the Internet browser and switching the computer off.

Keller and S. Lee (2003) note the practical implications of obtaining informed consent and reiterate a parallel drawn between agreement to participation and the agreement to software end-user license agreements. Going one-step further, the home pages that host web-based experiments for this research, include a check list of items that the user must confirm before clicking a button that displays a confirmation dialogue box. These check boxes provide as practically as possible, a way for the participant to confirm they are above the minimum age, have read the participant information sheet, and have read an electronic letter officially introducing the researchers. These information sheets are an obligatory requirement per the ethics committee overseeing the research. An analysis of server logs would be one way to confirm participants have read the experiment's participant information sheet, though linking the start of the experiment from a participant information sheet could at least provide participants every opportunity to make an informed consent.

Another issue impacting on ethical standards relates to a participant's privacy and right to anonymity. Privacy issues are ever-present throughout the entire Internet with widespread distrust of web sites, largely attributed to malicious activities by a minority who trade in the sensitive data of people. Email addresses and other uniquely identifying information have little specific use in online web-based research of the nature reported in these chapters and so is never requested.

Where participants are external to the university, participants are asked to give a random word if they wish to see debriefing information at the end of the experiment - see Section 3.6.3 on page 116. Participants internal to the university are suggested to provide their student number with assurances that this would not be used to access any further personal information. In reality, the provision of student numbers provides one additional cue that they were internal and not external participants coming in from recruitment methods beyond noticeboard advertisements. In any case, during the analysis all identifier words are scrubbed and recoded to anonymous numbers for archival purposes.

Primarily, this research relies on the ethics committee overseeing research at Flinders University to provide an unbiased assessment of ethical soundness. The experiments reported here are largely benign in that they do not collect sensitive information, do not focus on sensitive topics, and do not use deception in experimental design beyond excluding some user interface designs between treatment groups. Generally, this is the case for many other usability experiments. However, an understanding of such issues is important to ensure appropriate design and implementation of the web pages hosting online experiments.

3.5.7 Data Security in Storage and Transmission

Online experiments pose additional concerns relating to the security of stored data. For the uninitiated, web server security can pose a concern for a range of reasons. University computing policies go a long way to ensure that data stored on internal networks are protected behind multiple layers of network security. However, as Reips (2002) advises, site visitors should not be allowed to view the contents of web server directories, web servers should be updated to protect against malicious attack, and sensitive data should as a rule, be written to inaccessible folders on the host server.

Other recommendations by Reips, which are particularly useful, include careful naming of folder directories that could invite curious participants to make changes to the URL of the site to look inside folders and interfaces configured for other participant groups. A random or nonsense nomenclature for files, scripts and folders reduces the likelihood that information will be revealed to unauthorised users. Furthermore, the URL should not be used to store personally identifying information such as that obtained in forms and sent over networks, which could be recorded by software mon-

itoring wireless networks for example. Additional assurances can be made by using various encryption techniques both for transferral of information between the participant's computer and web server as well as for server side storage of information.

The need for technological sophistication, including encryption, should be motivated on a case-by-case basis and while not discounting the fact that it could become standard procedure in the future, it was not considered necessary for the present series of experiments. In this work, a lack of encryption was justified since there were no personally identifiable data captured. For instance, there are no names, email addresses, dates of birth, or licenses numbers collected. Moreover, the performance and qualitative responses selected by participants are of a benign nature and do not involve topics of a sensitive nature such as health, sexuality, criminal behaviour, racism, or sexism. In contrast, should future experimentation collect such information then this would indeed justify that communications be encrypted throughout the experimental procedure.

3.5.8 *Weighing Up Web-based Experiment Methodologies*

Earlier sections have argued that a web-based methodology is a capable way to collect data for use in experimental research. The leading advantages primarily focused around efficient use of limited resources including time, financial cost and convenience. However, leading draw backs of a web-based methodology include data reliability, the threat of participants not taking the study seriously and a lack of observational data. Such draw backs all arise out of a weaker degree of control that a researcher would otherwise tighten within a laboratory-based experiment.

Control is a fundamental facet of the experimental method. As much in web-based settings as the laboratory setting, efforts are directed toward the establishment of protocols and procedures that minimise the incidence of uncontrolled factors unduly influencing one participant performance relative to another participant who is not subject to the same conditions. However, it is undeniably difficult to ensure the effectiveness of those controls in a web-based experiment when the researcher is not present alongside the remote participant.

3.6 *A Web-Based Evaluation of Motion Frequency Encoding*

Discussion will now turn an experiment that aimed to investigate the role of motion frequency encoding in a metadata visualisation context. This experiment was delivered remotely via a web-based methodology. Participants were not paid for participation; so to promote participation, a game-like feel allowing participants to track their performance over sessions was adopted. Both static and dynamic attributes of icons were

used to encode data; therefore, a discussion will first describe and motivate the use of specific graphical attributes in this experiment.

3.6.1 *Graphical Features for Encoding Paradigm*

This section outlines static and dynamic glyph attributes in separate subsections. Initially, discussion will consider prior guidelines for use of static features; the static encoding features under consideration are size, orientation, hue and grey scale. Conversely, comprehensive guidelines do not exist for dynamic features. Thus, a subsequent section discusses dynamic features under consideration. A description is provided of each motion type covering the expected motion pattern based on choice of interpolation function, amplitude, frequency, period and phase. The dynamic features under consideration are zooming or grow and deflate, rotating, pulsing and shuffling.

Static Visual Features

Size The prominent and widespread use of size as a data encoding technique makes glyph size an obvious technique to investigate. Ware (2004) recommends that the minimum size should not drop below 1° of visual angle, particularly if an icon's colour is concurrently encoding data and in particular, if green or blue is in use. Furthermore, an empirical result of Lindberg and Näsänen (2003) suggests a minimum icon size of 0.7° of visual angle and a minimum icon spacing of 0.35° to 0.7° of visual angle.

In the present experiment, four sizes encode for four data attributes. The differences in size are prominent and follow a linear relationship; each step in size is twenty-five percent larger than previous. The minimum size is eight pixels - well above the minimum recommendation of Ware; the maximum size is twenty pixels and the intermediate levels are twelve and sixteen pixels. Figure 3.1 on page 106 presents a visual depiction of size parametrisations. The default size under conditions that do not manipulate size is 12 pixels.

This size set considers only relative size. However, a participant's display or monitor size and quality will have an impact on absolute size encoding - the size of the icon cast upon the participant's eye or the icon's size measured by a physical ruler - since the size encoding set is denoted by a pixel scale and not degrees of visual angle or some other ruler metric. Screen resolution and pixel density as well as screen scale will have an influence on the size of the icon cast upon the eye of the participant and although not controlled here, can be.

With all other things held constant, two participants completing the same task on two different monitor configurations may have different performances if the pixel densities of their screens cause differences in absolute size. In the extreme case, one participant may have difficulty viewing the smallest icons - due to high pixel density

- while the other participant may have difficulty seeing the full set of icons - due to low pixel density - if the experiment's visualisation area exceeds that of the screen's dimensions. This might be controlled by detecting the pixel density of each participant's screen and ensuring a consistent absolute size by scaling up or down the size encoding set at experiment start up time so that all participants see exactly the same sized icons on the screen, as measured by a ruler.

However, very much uncontrollable factors do also apply such as the distance at which the participant is located from the screen. As a rule of thumb, we sit approximately 150% of the monitor width away from a computer screen. While a participant can be told to maintain the same monitor-to-head distance throughout the course of the experiment, it is unlikely that this will be exactly the same, and it is further unlikely that remote participants will subject themselves to a restraining headrest in order to do so.

Zooming and scaling a browser window is a different situation but can introduce differences between participants that choose to zoom and participants who do not choose to zoom. Zooming has the effect of changing the absolute size of the icon set making the perception of icons different to participants not zooming. Like lower pixel density, scaling above 100% will make the icons appear larger while scale below 100% will make icons appear smaller. Subtle variations may be introduced by subtle differences in absolute size. However, like controls for differences in pixel density, it may be possible to capture scaling events within the experiment apparatus and warn the participant about the consequences of doing so.

Clearly, a homogeneous computing platform, such as that offered in some laboratory-based experiments and the experiment of Chapter 6, can alleviate these absolute size considerations.

Orientation Icon orientation is the next feature under consideration, due to the human perceptual system's sensitivity to orientation approximately 1° (Olzak and Thomas 1986 in Wolfe (2007)). Healey, K. Booth, and Enns (1993) found that orientation was an effective encoding technique to facilitate extraction of patterns in large data sets, particularly given that it did not interfere with the perception of patterns encoded by hue, which is important for this study since hue and orientation are a feature combination of interest. In a study of orientation in scalar field visualisation, Weigle et al. (2000) find that a minimum separation of 15° between target and distractor orientation ensures fast and accurate target detection. Later Huber and Healey (2005) extend this to angular orientation of motion, suggesting 20° should be the minimum distance between two encoding orientations; similarly Bartram and Ware (2002) recommend a minimum of 16° . As the default shape for each object is a square, angles greater than ninety degrees appear the same. Thus, this experiment adopts a range of zero to sixty - 0° , 20° , 40° , and 60° . Figure 3.2 on page 106 indicates visually, the orientation-encoding

palette. The default orientation under trials that did not manipulate orientation was 0°.

Hue Hue or Colour is arguably the most prominent encoding feature in use. Yost and North (2005) and Nowell (1997) corroborate earlier findings that colour encoding facilitates the fastest and most accurate task performance of the features under study in this experiment. In addition to wide spread usage across information visualisation applications, colour is among the most researched graphical feature of the six in this experiment (Nowell, 1997).

Initially, two leading theories on colour vision were the opponent colour theory of Hering (1878) and the earlier tri-chromatic theory of Helmholtz (1852). Under tri-chromatic colour theory, a combination of red, green and blue light forms the spectrum of observable colours. In contrast, under opponent colour theory, colour is defined as some hue value along a red-green continuum, plus some hue value along a blue-yellow continuum, plus some brightness or luminance value along a white-black continuum. Contemporary colour theory reconciles both theories into a single theory of colour vision and a biological basis for both are found in the brain.

At the interface between the external world and the brain, light receptors in the eye process light energy according to tri-chromatic theory whilst later stages of processing in the visual cortex of the brain process light according to opponent colour theory Weiten, 2001. Tri-chromatic theory is relevant at the point of light energy absorption, where photosensitive cone receptors in the eye - each receptor sensitive to light in either a red, green or blue wavelength - whilst opponent colour theory is relevant further along the visual processing pathway where neurons, representing the red-green, blue-yellow and white-black continua, fire in response to signals arriving from cone receptors. Specifically, some proportion of red-green neurons are excited on the input of red and inhibited by input of green, while some proportion of neurons are excited by green input and inhibited by red input - likewise, for the blue-yellow and white-black continua. As in tri-chromatic theory where all observable colours are based on a mixture of red, green and blue, all observable colours in opponent colour theory are a mixture of hue values lying on red-green, blue-yellow and white-black continua.

An oft cited reason in support of opponent colour theory includes the paucity of tri-chromatic theory to explain the persistence of after-images in complementary colours - i.e. perceiving a green after-image while staring at a white canvas having previously stared at a red canvas. In addition, when describing colours, a 'reddish-green' or 'yellowish-blue' colour is seemingly not ever described; moreover, given only three hues red, green and blue, we are unable to describe all observable colours without the use of yellow (Ware, 2004).

Use of hue in a colour coding faculty is common, and accordingly, guidelines are widely available. Since red, green, blue and yellow demarcate the ends of opponent

continua, use of these base colours are appropriate for representation. There are four colours in the hue condition - red, green, blue and orange. Black is the default colour for combinations that do not manipulate hue. Figure 3.3 on page 106 indicates visually, the hue encoding set. The set of colours under consideration forms a subset of recommended colours for use in colour coding - the full set being: red, green, yellow, blue, black, white, pink, cyan, grey, orange, brown and purple; this recommendation follows that these colours are easily recognisable, have widely agreed-upon category names, and have reasonable separation in colour space (Ware, 2004).

Colour set guidelines refer to qualities of the data: is there a logical maximum, central and minimum point; are the data ordered; or are the data nominal. In this experiment, the data are nominal, thus, a hue set consisting of contrasting, separable hues is favoured. For ordered data, a colour set consisting of equally-spaced luminance intervals for a single hue is appropriate. Finally, for values along a continuum, the use of a red-green or blue-yellow is appropriate.

A selection of hues well separated in hue space is important for foreground colour contrast; however, foreground and background contrast is integral also. To ensure that foreground colours of the same colour are recognised more easily, a background colour that sufficiently contrasts with foreground colours should be selected (Few, 2008). In this case, a white background is selected. Use of gradient colours as background colours - often incorporated in the depiction of 3D graphics as a depth cue - can influence erroneous interpretations of the same coloured objects; for instance, a consistent grey swatch may be perceived as lighter or darker depending on a light-dark background gradient.

Harrower and Brewer (2003) report that their online colour-choosing tool can provide a suitable set of up to 12 colours for map visualisations but that this figure is smaller if additional constraints are placed upon the designers such as considering visual disabilities of the user population.

Saturation Four grey scales are chosen to maximise distance between alternatives in steps of 25% saturation; therefore, icons encoded by grey scale are 25%, 50%, 75% and 100% saturation. The default grey scale for non-grey-scale trials was 100% or full saturation. Figure 3.4 on the next page depicts the grey scale encoding set.

Gabriel-Petit (2006) reiterates that humans cannot distinguish easily between two greys that differ by less than 15%, and that the minimum distance between values should be at least 15%-25%. To encode additional data attributes beyond four categories, based on this recommendation up to six items could be encoded with grey scale. Alternatively, eleven equidistant values between white and black form the neutral axis of the Munsell colour system suggesting that up to 11, counting fully white, could be used to encode data.

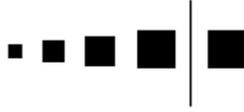


Fig. 3.1: The size palette. The divider separates the size trial palette on the left, from the non-size trial palette on the right.



Fig. 3.2: The orientation palette. The divider separates the orientation trial palette on the left, from the non-orientation trial palette on the right.

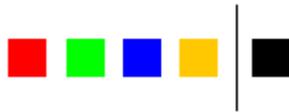


Fig. 3.3: The hue palette. The divider separates the hue trial palette on the left, from the non-hue trial palette on the right.



Fig. 3.4: The saturation palette. The divider separates the saturation trial palette on the left, from the non-saturation trial palette on the right.

Dynamic Visual Features

The previous section illustrates a sample of the rich understanding we have for static encoding features. In contrast, our understanding of motion features is reasonably lacking. This section describes the motion features and parameters for this examination.

As this experiment will investigate the role of frequency in encoding paradigms, up to four frequencies will be encoded; these frequencies are 0Hz, 0.5Hz, 1Hz and 2Hz. Zero hertz constitutes an encoding attribute as much as slow-motion and fast-motion do.

An alternative perspective on frequency is motion period. Motion periods are 0, 2, 1 and 0.5 seconds respectively per every cycle or reaching peak deviation or amplitude every 0, 1, 0.5 and 0.25 seconds respectively.

Table 3.1 and Table 3.2 below outlay motion type and parametrisation for four motions: zoom, rotate, shuffle and pulse. Figure 3.5 on the following page to Figure 3.8 on page 109 below illustrate each motion type using a static analogue representation. Different interpolation functions - see Table 2.4 on page 50 in Chapter 2 - are used to calculate each glyph's size, orientation and saturation value for every new step in time. A step in time is approximately 33 milliseconds in order to achieve a frame rate beyond that distinguishable by the human eye and to convey a sense of smooth motion.

Tab. 3.1: Description of the motion under investigation.

Motion	Description
Zoom	Width and height oscillate between minimum and maximum size
Rotate	Shape rotates about central point in a clock-wise direction
Pulse	Saturation oscillates between minimum and maximum
Shuffle	Position oscillates between minimum and maximum

Tab. 3.2: Parametrisation of the motion under investigation.

Motion	Attribute	Function	Min	Max	Amp.	Phase	Freq.
Zoom	Area	Triangle	15	35	10		0Hz
Rotate	Angle	Saw	0	2π	π	Rand.	0.5Hz
Pulse	Saturation	Sinusoidal	0.2	1.0	0.4	0- 2π	1Hz
Shuffle	Position	Sinusoidal	x-10	x+10	10		2Hz

Zoom motions utilise a triangle wave function to linearly increase and decrease the glyph's size. A sinusoidal function could serve the same purpose although the perception of motion would be slightly different. Rotation motions utilise a saw tooth wave function such that when maximum peak deviation is reached, the function returns to zero,

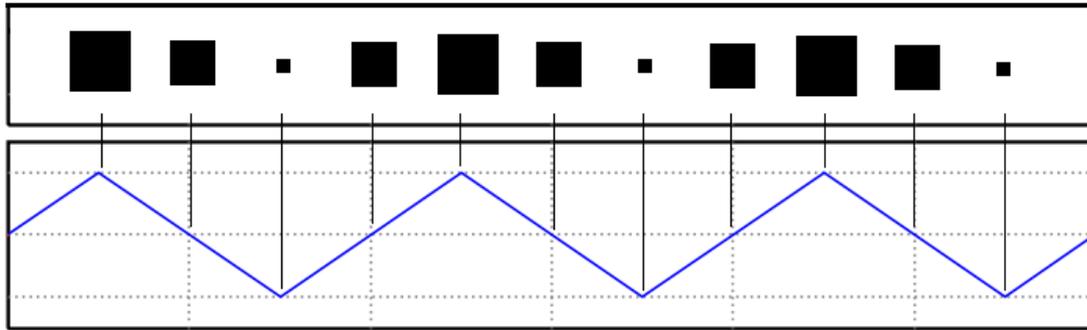


Fig. 3.5: The grow palette utilising a triangle wave function.

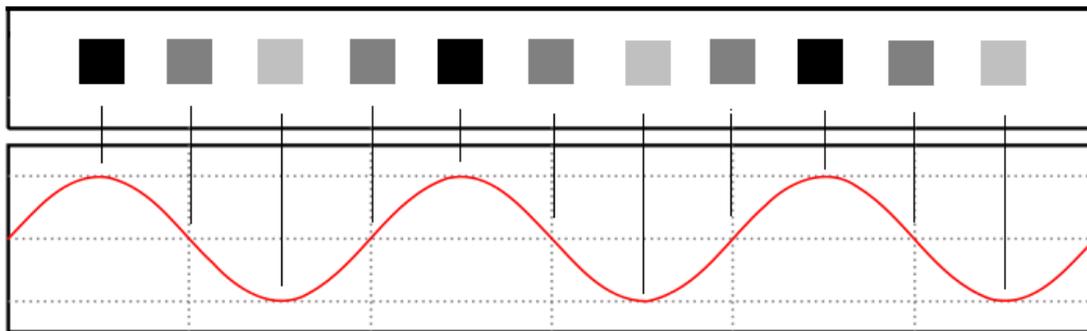


Fig. 3.6: The pulse palette utilising a sinusoidal wave function.

giving the impression of continuously circling around a central location. In this case, utilising a triangle or sinusoidal wave would result in a different perception of motion in that the glyph would rotate back and forth between 0 and 2π radians. Finally, pulse motions utilise a sinusoidal interpolation function to smoothly increase and decrease the saturation for glyph colour. This motion is perceived as a pulsation as the rate of change over the course of each cycle is different and rate slows as the interpolation approaches peak deviation. In contrast, a square wave changes the intensity of the saturation abruptly, giving the perception of something flashing on and off.

A pulsing light is selected over abrupt flashing lights in that flashing lights imply an alarm state or in the least have been investigated with the aim of drawing attention to alarm states (Thackray and Touchstone, 1990); something manipulating saturation, but not as abrupt, is desirable. Furthermore, Bartram, Ware, and Calvert (2003) observe that at the periphery of view, slow flashing cues were one of the least distracting and annoying while fast flashing lights were among the most distracting and annoying. Interestingly, they also find that zooming motions are among the most distracting and annoying in the periphery of view.

Each of the motions under consideration are termed anchored motions. Anchored motions are those that move about a fixed and central coordinate, location, or neighbourhood; in contrast, travelling motions involve the translation of a glyph across a large portion of the viewer's screen (Bartram, 2001). Bartram finds that the latter,

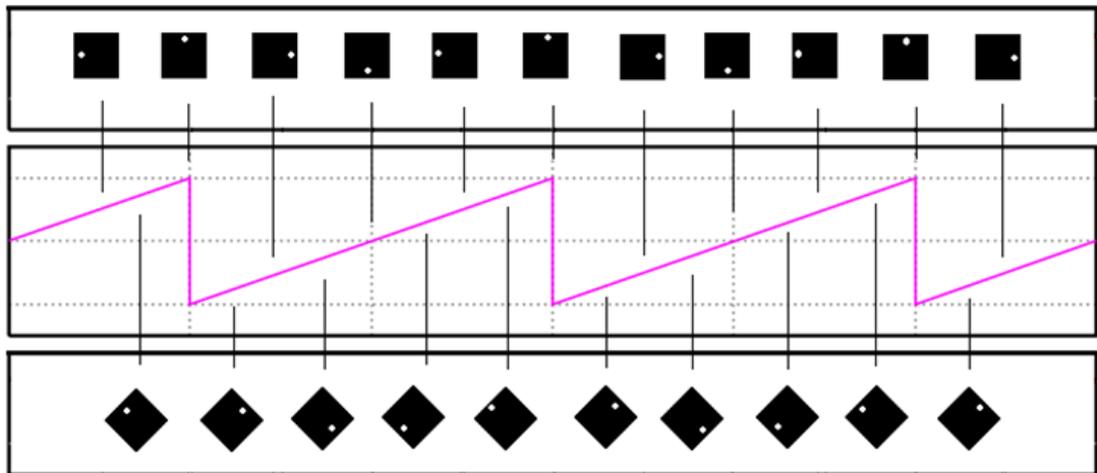


Fig. 3.7: The rotate palette utilising a saw-tooth wave function.

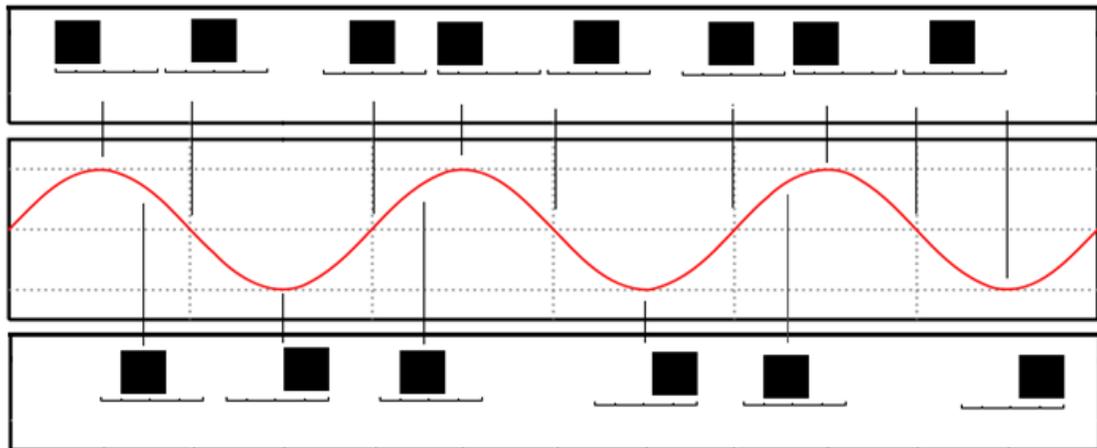


Fig. 3.8: The shuffle palette utilising a sinusoidal wave function.

travelling motions are more distracting than anchored motions when used as response cues in a secondary monitoring task. Anchored motions are of importance in the current context since the intention is to situate items in a generally fixed semantic space in order to depict relationships between items co-locating nearby and doing so in a way that does not overly distract or annoy the user.

Phase and peak deviation or amplitudes are also outlined in Table 3.2 on page 107 below along with minimum and maximum values for each attribute.

For all trials, all motion phase is randomised between 0 and 2π radians to control for effects of perceptual grouping by phase. For zoom motion, the glyph size oscillates between 15 and 35 pixels, glyph rotation cycles through 0 and 2π radians and pulse oscillates between 20% and 100% saturation. In the case of zoom and pulse, minimum offsets ensure the glyph does not disappear briefly when the interpolation function reaches zero making the glyph too small or too faint to perceive.

This completes the discussion of the graphical features for encoding; discussion will

now turn to the data features that the graphical features encode.

3.6.2 *Data Features for Encoding Paradigm*

Four data features consisting of four data attributes each are selected for encoding, allowing for complex search criteria of up to four target attributes per trial. The four data features are file type; file size, file age and file source.

File type instances are portable document format PDF, word document DOC, web page HTML and text TXT; file size instances are 64KB, 128KB, 256KB and 512KB; age instances are years 1996, 2000, 2004 and 2008; and source instances are commercial COM, governmental GOV, organisational ORG and educational EDU.

Alternative data features are conceivable ranging from other multimedia metadata to more general abstract features such as letters or numbers in which case the task statement would consist of a series of letters referring to the letters adjacent to the key items. Another possibility is to utilise generic adjectives reflecting a recoding or superlative categorisation of metadata instances. For example, file size instances could be coded as tiny, small, big and huge; while age could be coded as archive, old, new, and current.

The metadata feature selection took place before the development of the isomorphic encoding idea presented later in Chapter 4 and so as a result, the use of the generic adjectives that share a superlative relation were not considered by the design. Furthermore, the use of abstract letters and symbols for encoding was avoided due to an envisioned loss of search engine feel.

To recapitulate, both static and dynamic graphical attributes encode data attributes to form an encoding paradigm. This encoding paradigm appears in a legend on the experiment apparatus - outlined subsequently.

Classically, experiments in perceptual psychology show the target stimulus to the participant prior to the commencement of a block of trials. Practise trials maximise the chance that the target is coded to short-term memory for efficient and reliable access for comparison to the search set when real trials commence, thus eliminating the additional step of integrating several disconnected visual features into a mental chunk. Yet, orienting one's self with the encoding paradigm of a cartographic map is natural for a user not familiar with the encoding. It is the task of integrating disconnected features into the mental chunk that poses the greatest learning overhead for this experiment, while it is the ease of which a user may encode and decode the feature quickly that influences task performance.

We now arrive at the conclusion of the introductory discussion having established a scenario for a search tool incorporating a metadata visualisation component and a proposal for further investigation into the use of motion frequency to expand the

encoding capacity of the visualisation. In the proposed scenario, a set of search results are allocated a spatial location on screen revealing emergent theme neighbourhoods. Search takes place first and primarily on thematic context localising search to particular areas perhaps with the aid of textual annotations or perhaps a pre-attentive feature like colour (Wolfe and Horowitz, 2004) that encodes a rough ranking score. Secondly, search takes place over a rich set of metadata to ascertain further interactions with items in a localised area. For specific search over metadata criteria, a searcher should be able to evaluate each alternative item in the area against target criteria stored in short-term memory.

One way to increase the number of encoded data features is to utilise dynamic features to encode information. Motion features will expand the number of uniquely identifying features of an icon that could otherwise cause a searcher to confuse one icon with another based on a perceived similarity. Motion frequency, although previously suggested to be a relatively poor encoder of data has some support from the human factors literature and should be investigated further.

The narrow focus of this experimental context is realistic and advantageous since full-text information search is a large problem in itself. This context is advantageous as this experiment can ignore the more illustrative and expressive capacities of motion as outlined by Bartram (1997) and focus on the raw encoding power of the basic components of motion. Visualisations for full-text search do not need to support the whole gamut of data and information hypothesis-driven visual operations but do need to support fast irrefutable recognition of targets matching a search criteria in the least. The discussion will now report on the design and outcome of an experiment that aimed to investigate motion frequency further.

3.6.3 Method

Participants

83 unpaid participants started the experiment, 43 participants did not complete the experiment while 40 participants successfully completed the experiment in full. An analysis of drop out will examine the entire participant pool further, but the remainder of the results presentation and ensuing discussion will consider those from the set of successful submissions only. A successfully completing participant was considered to be a participant who submitted a demographics questionnaire, a response to each experiment trial and finally an exit questionnaire response.

Of the completing participants (21 male, 19 female), 60% reported their age within the 20-29 range (<20 15%, 30-39 17.5%, >40 7.5%). The data indicate that participants engage in a weekly average of 40.3 hours (10-110 hours) of computer-based work of which an average of 5 hours (1-50 hours) is spent using search engines and an average

of 6 hours (0-70 hours) is spent gaming. All participants reported using search engines every week. However, a third of participants report either zero or very few hours in a week playing computer games. All participants report having spoken English for their whole life or for a long time; however, one completing participant reported having spoken English for only five years. Furthermore, approximately one third indicated some experience with computer graphics, visualisation or design and approximately three quarters watched animated cinema or cartoons at least on a monthly basis while the majority report either weekly or daily viewing. Two report left-hand mouse use for the experiment while the remainder report right-hand use. Finally, twenty-two participants indicated that they were students spanning a range of subject foci including psychology, accounting, biological science, computer science, and mathematics. The remaining participants consisted of skilled and unskilled workers including cashiers, graphic designers, information security consultants, teachers, children's services, real estate appraisers and public servants. One participant indicated that they were unemployed and cited 110 hours of computer usage per week.

Analysis of the participant demographic data and by IP-Lookup reveals that the participant pool is geographically diverse. Predominantly, 70% - or 28 of 40 - of participants originated from within Australia, while the remainder originated from other countries. International participants were predominantly North American; two participants originated from the United Kingdom and one participant from Spain. Of the Australian participants, the majority 53% - or 15 of 28 - were internal to Flinders University; however, IP-Lookup indicates that there were interstate participants that perhaps learned of the experiment through handouts and informal presentations at national conferences and workshops.

While it is disciplined to conduct such tests, there were no pre-tests for colour blindness or general visual acuity.

Apparatus

Figure 3.9 on the facing page depicts a screen shot of the Java-based, browser-embedded experiment apparatus. The Applet is 613x693 pixels in dimension and is centred in the middle of the screen on a plain white background. The Applet communicates trial result data at the end of each trial to a remote server through a reliable network socket. Server-side scripting manages the recording of experiment data to a flat file. At the conclusion of the experiment trials, the Applet forwards the user to a web-based questionnaire via a JavaScript call. Similarly, demographics and questionnaire data utilise server-side scripting and flat file data recording.

The experiment apparatus consists of four components: visualisation, encoding legend, task panel and button panel. The visualisation - centre of Figure 3.9 on the next page - contains the target and distractor stimuli, the legend - at left of Figure 3.9

- depicts the trial's encoding rules, the task panel - at right of Figure 3.9 - displays the trial task statement and experiment progress label and finally, the button panel - at the bottom-left of Figure 3.9 contains two buttons that control the flow of the experiment. The Skip button cancels or skips the current trial and moves to the next trial while the ready button progresses the participant between the first and second stages of each trial. The two-stage trial design is described in Section 3.6.3 on page 115.

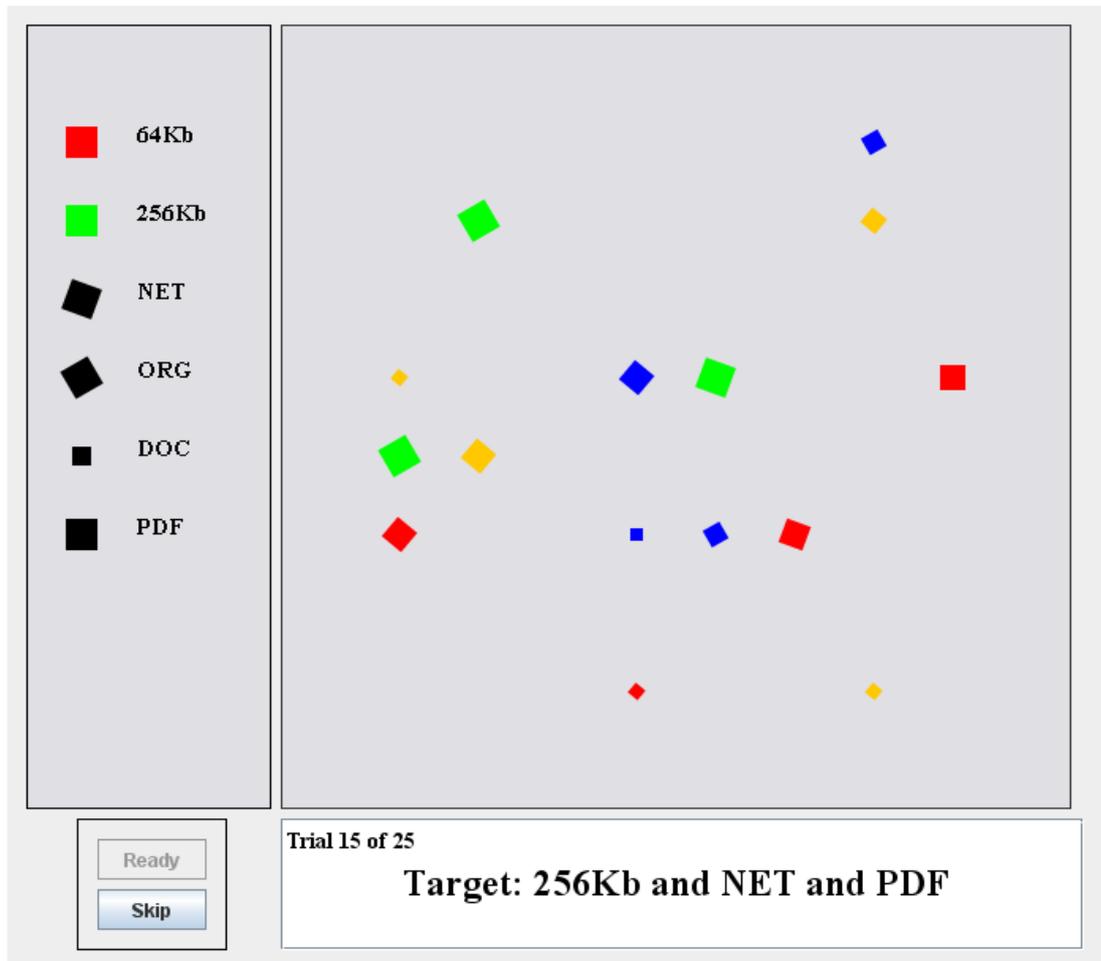


Fig. 3.9: A screen shot of the motion experiment.

Target and distractor glyphs are centrally positioned within randomly assigned grid cells of a 9x9-grid layout in order to mitigate any effect of occlusion. Glyph positioning is random for all trials and participants. Furthermore, the spatial layout and spatial distance or closeness has no intended semantic or thematic interpretation.

The legend depicts the trial's encoding paradigm. For each encoding rule, a glyph representing the encoded data attribute appears adjacent to the referent data attribute. Encoding legend entries are grouped by data attribute, but within-group as well as data group positioning is random i.e. the order of items in the legend was randomised across trials. Per the guidelines of Dykes, Wood, and Slingsby (2010) grouping arranges the legend in a relational manner, the legend appears in a prominent position, and graphical

features match exactly those in the visualisation.

Only two out of four features per data type appear in the legend at a time due to spatial constraints. Target features are guaranteed to appear in the legend, and examples representing the full set of encoding rules, whether present in the legend - or not - are equally likely to appear in the visualisation area. The full set of encoding rules for any trial type requires double the space on the legend leading to an excessively crowded legend with increasing trial complexity. In the interests of consistency, this legend design constraint applies to all trials regardless of trial complexity.

In a visual search experiment, a target is prescribed as a whole a picture of the target - like a red square. This also happens in applied research as in the case of Bartram (2001). Alternatively, this target may be expressed as a statement 'red square'. Following the target display, the trial begins and the task of the participant is to locate the target among the set of distractors. During target exposure, a participant is not privy to the impending trial's full graphical palette as is embodied by the design of this experiment. This design choice made more economical use of screen real estate for a four dimensional trial. A four dimensional trial would necessitate 16 individual icons each depicting a particular graphical feature configuration. Making the legend items smaller would be appropriate but the present design has opted to ensure the text is legible. Feature configurations unseen in the legend but present in the distractor set are always distractors.

The task area displays the target data attributes that participants will need to decode using the legend. Additionally, the task area displays the number of tasks they have attempted. When necessary, this panel displays visual feedback that informs the participant that they have selected an incorrect answer. The task panel background is highlighted in pink #FFFAFA for 750 milliseconds immediately following an incorrect selection. This highlighting is intended to reorient the participant's attention to the task statement in preparation for a subsequent attempt.

Procedure

Participants navigate to the experiment web site. A browser technology-check takes place and participants are required to confirm that they see the result of a JavaScript and Java Run Time Environment technology check as indicated by the presence of two green tick icons - that are updated programmatically during the check - to indicate JavaScript and Java Applet technology are operational.

A participant provides informed consent by clicking yes to the terms and conditions presented in a confirmation window that appears after clicking the begin button.

Next, participants provide demographics information via web form including a random uniquely identifying word, which they are told to remember if they want to view their performance report at the end of the experiment.

The participant then reads a page of training material presented as a series of apparatus screen shots and descriptive text; this material is included in Appendix A on page 377. The participant does not complete any practice trials. However, a detailed example is described in the instruction material.

Having completed the training phase, participants start the experiment. There are 25 experiment trials in total. Each trial starts at the end of the last and trial progression is reported in the task panel. There is no set opportunity for rest breaks during the 25 trials.

At the conclusion of experiment trials, participants respond to a short exit questionnaire. The questionnaire asks participants to provide an overall difficulty rating of the experiment task and to indicate if they perceived any flickering during the trials. There was a suspicion that participants could experience occasional discontinuous motion if the experiment was running in parallel to several other applications or browser windows or the participant's computer was otherwise slow or overloaded. On submission of questionnaire responses, participants are thanked for their contribution and invited to view a personalised performance report that provides a summary of their average trial performance - see Section 3.6.3 on the following page.

Task

For each trial, the participant's task is to find and left-mouse-click on the unique target glyph on the screen consisting of graphical attributes that encode the data attributes specified by the task statement.

A trial involves two stages. In the first stage, participants determine the composition of the target glyph based on the encoding rules displayed in the key. In this stage, the visualisation panel remains empty. For example, in left of Figure 3.10 on the next page, the task statement reads '256Kb and NET and PDF' and in the legend in Figure 3.10 on the following page at left, the attribute 256Kb is encoded by green, the attribute NET is encoded by 15° orientation and the attribute PDF is encoded by moderate glyph size. Therefore, the correct answer will be moderately sized, tilted to the right, and green.

Once the participant knows the intended target, they click the 'ready' button to proceed to stage two. In the second stage, target and distractor glyphs appear in the visualisation panel. The participant must find the single target amongst distractors and left mouse click on it. In this trial, the target glyph appears in the centre of the visualisation, immediately right to a smaller blue glyph. The target glyph shares similar features with other distractors; however, the combination of visual attributes is unique to the target and differentiates the target from all other distractors.

If the participant cannot find the target, they have the opportunity to skip the task by pressing the 'skip' button. A participant has three attempts at locating the

correct answer and the target is always present. Visual feedback will indicate that the previous selection was incorrect, where necessary. Moreover, if after three attempts the participant cannot find the correct answer, the current trial ends and the next trial begins. Figure 3.10 below depicts stage one at left and stage two at right.

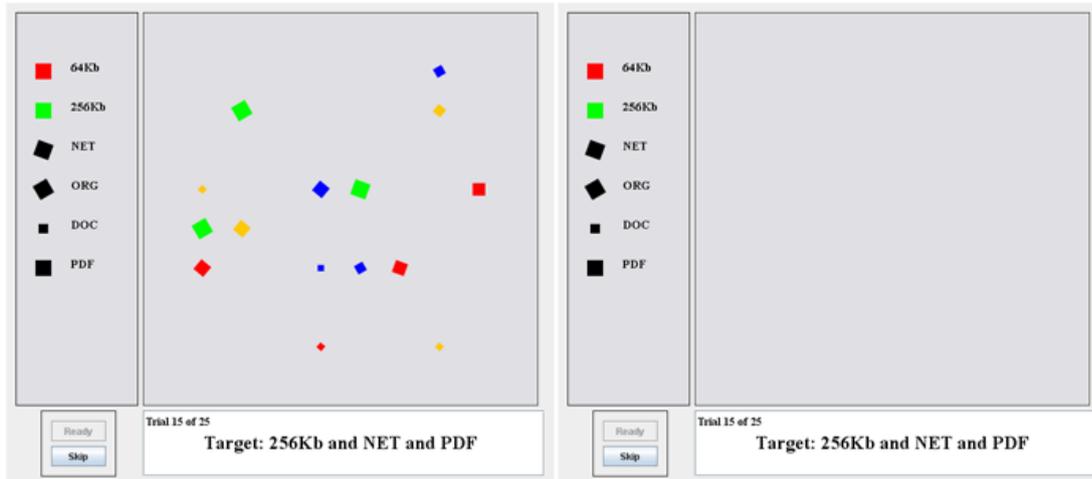


Fig. 3.10: Left - stage one shows the task statement and coding rules. Right - having clicked ‘ready’ and moving in to stage two, the target and distractors are now visible in the centre of screen. In this trial, the top two rules encode File Size with hue, the middle two rule encode File Source with Orientation and the bottom rule encode File Type with size.

Experiment Debrief and Performance Report

At the conclusion of the experiment, participants are thanked for their contribution and requested to participate again if they so desire. Additionally, they receive an invitation to review a personalised performance report; but in order to access the report they need to provide their random word user name that they specified in the demographics form.

The performance report is intended to promote a game like feel; participants can observe their average performance relative to the average of the participant pool. Given the widespread fascination with web based and mobile social media applications of seemingly little substance other than raw entertainment, this experiment and performance report has attempted to draw together similar elements of game, personal performance tracking and competition to entice participation based on word of mouth referral from participants who achieved a positive experience during their participation.

The personalised report shows a leader scoreboard and two graphs displaying the user’s average time and average error rate, relative to the average performance of the participant pool. The leader board displays the top 10 leading scores and user names - identified by their own random word. The scoring system considers only time performance, and allocates ten points for faster than average performance, 5 points for

average performance, and 2 points for below average performance. If the participant does not complete all tasks, a penalty multiplier mitigates any attempt at ‘gaming’ the leader board. On review, the scoring system was overly simplistic and prone to rewarding lower time in the absence of lower error rate. One can imagine the scenario where the leader board score is worse for participants taking longer time to be sure about their choice versus several fast but uncertain decisions.

The report page provides a link for users to participate again - without the overhead of the demographics form - to lift their score and average performance. This leader board did not appear on the welcome page of the experiment web site and was only accessible by providing a valid user name.

Figure 3.11 illustrates a typical performance report; in this report, the participant - as the blue line - performed on average slower than the rest of the pool - as the red line. However, the average error rate for this participant was lower than that of the pool average, indicating they may have completed more trials on the first attempt.

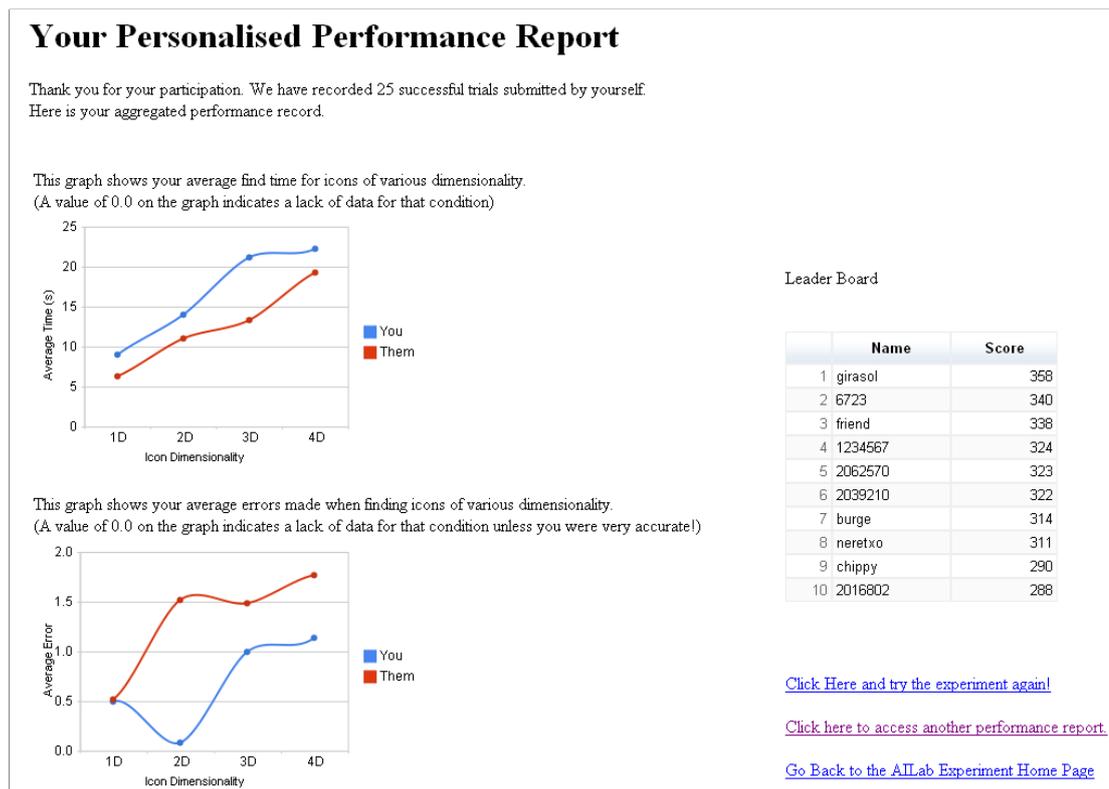


Fig. 3.11: Personalised performance report.

Design

The experiment had a within-subjects design with a motion presence independent variable of three levels: static, dynamic - motion - and mixed and the graphical feature type variable of four levels colour, size, and orientation. This design observes the influence

of feature type with increasing task complexity or the number of graphical features encoded into a glyph on time and error dependent response variables. Feature type is included as a variable as earlier research (Bartram, Ware, and Calvert, 2003) shows that motion type is a determinant of detection and distraction - not all motions are the same regardless of frequency.

Four static and four dynamic features are investigated. There are 162 possible conditions for up to four choices, without replacement with order not important: $C(8,1) + C(8,2) + C(8,3) + C(8,4)$. Consequently, constraints are added to reduce the potential set to 71. Trials involving both hue and saturation, hue and pulse, saturation and pulse, angle and rotate, grow and size are excluded from examination. There are 11 exclusively static trials, but only up to 3 dimensions, 15 exclusively dynamic trials up to 4 dimensions and 45 mixed trial types. Table 3.3 on the next page outlines the valid static combinations and motion combinations. Table 3.4 on page 120 below outlines the valid mixed static and motion combinations.

The dependent response variables recorded are time on task and number of attempts. Number of attempts is considered a dependent variable since participants are permitted a maximum of three attempts at completing a trial. Timing is not recorded for trials where the maximum attempt threshold of 3 attempts is exceeded.

Time is recorded for the two stages of the experiment task as outlined in Section 3.6.3 on page 115. In stage one, the participant determines the configuration of the target object based on the mapping rules and the task statement this time is referred to as the preparation time. In stage two, the participant is searching for the target - this time is referred to as the answer time. Preparation time starts at the commencement of the trial and ends when the participant clicks the ready button. This point marks the beginning of answer time, which ends at the selection of the correct answer.

This experiment does not solicit subjective responses and opinions on aspects of individual animated or static glyphs. In such an abstract context, it is of little use to ask participants what they think the most effective encoding paradigm is. However, a four-question exit questionnaire asks whether they understood the task and whether they perceived any flicker or visual artefact during the experiment.

The 25 experiment trials are built randomly at apparatus run time such that no two exact targets are searched for per participant session. The trial dimensionality, encoding rules, search set layout and key item layout is randomised. Trials are not blocked or ordered according to dimensionality, visual feature or data feature.

For each trial, there are 15 alternative glyphs to choose from and only one is encoded with the unique set of target attributes. The remaining 14 distractors are encoded by the graphical feature types present in the encoding legend; however, distractors can differ by up to four ways per graphical feature. For example in Figure 3.10 on page 116,

Tab. 3.3: Trial conditions from 4 static and 4 motion feature combinations; motion and static features unmixed. The ‘•’ symbol denotes that the feature is active in the combination.

Pattern	Graphical Attributes								Feature(s)
	Static				Dynamic				
	H	Z	O	S	G	R	P	T	
S	•								Hue
		•							Size
			•						Angle
				•					Saturation
SS	•	•							Hue Size
	•		•						Hue Angle
		•		•					Saturation Size
			•	•					Saturation Angle
SSS		•	•	•					Saturation Size Angle
	•	•	•						Hue Size Angle
D					•				Grow
						•			Rotate
							•		Pulse
								•	Shuffle
DD						•	•		Pulse Rotate
					•		•		Pulse Grow
							•	•	Pulse Shuffle
					•	•			Rotate Grow
						•		•	Rotate Shuffle
					•			•	Grow Shuffle
DDD					•	•	•		Pulse Rotate Grow
						•	•	•	Pulse Rotate Shuffle
					•		•	•	Pulse Grow Shuffle
					•	•		•	Grow Rotate Shuffle
DDDD					•	•	•	•	Pulse Rotate Grow Shuffle

Tab. 3.4: Trial conditions from 4 static and 4 motion feature combinations; motion and static features mixed. The ‘•’ symbol denotes that the feature is active in the combination.

Pattern	Graphical Attributes								Feature(s)
	Static				Dynamic				
	H	Z	O	S	G	R	P	T	
SD	•					•			Hue Rotate
	•				•				Hue Grow
	•							•	Hue Shuffle
				•		•			Saturation Rotate
				•	•				Saturation Grow
				•				•	Saturation Shuffle
		•					•		Size Pulse
		•					•		Size Rotate
		•						•	Size Shuffle
			•					•	Angle Pulse
			•			•			Angle Grow
			•					•	Angle Shuffle
SDD	•				•	•			Hue Rotate Grow
	•					•		•	Hue Rotate Shuffle
	•				•			•	Hue Grow Shuffle
				•	•	•			Saturation Rotate Grow
				•		•		•	Saturation Rotate Shuffle
				•	•			•	Saturation Grow Shuffle
		•				•	•		Size Pulse Rotate
		•					•	•	Size Pulse Shuffle
		•					•	•	Size Rotate Shuffle
			•			•		•	Angle Pulse Grow
			•				•	•	Angle Pulse Shuffle
			•			•		•	Angle Grow Shuffle

Tab. 3.5: Trial conditions from 4 static and 4 motion feature combinations; motion and static features mixed. The ‘•’ symbol denotes that the feature is active in the combination.

Pattern	Graphical Attributes								Feature(s)
	Static				Dynamic				
	H	Z	O	S	G	R	P	T	
SSD		•	•				•		Size Angle Pulse
	•	•					•		Hue Size Pulse
		•		•		•			Saturation Size Rotate
	•		•		•				Hue Angle Grow
			•	•	•				Sat. Angle Grow
	•	•						•	Hue Size Shuffle
	•		•					•	Hue Angle Shuffle
		•		•				•	Sat. Size Shuffle
			•	•				•	Sat. Angle Shuffle
		•	•					•	Size Angle Shuffle
SDDD	•				•	•		•	Hue Grow Rotate Shuffle
				•	•	•		•	Sat. Grow Rotate Shuffle
		•				•	•	•	Size Pulse Rotate Shuffle
			•		•		•	•	Angle Pulse Grow Shuffle
SSDD		•	•				•	•	Size Angle Pulse Shuffle
	•	•				•		•	Hue Size Rotate Shuffle
		•		•	•			•	Sat. Size Rotate Shuffle
	•		•		•			•	Hue Angle Grow Shuffle
			•	•	•			•	Sat. Angle Grow Shuffle
SSSD		•	•	•				•	Sat. Size Angle Shuffle
	•	•	•					•	Hue Size Angle Shuffle

the legend includes red and green features for the hue dimension and the target hue is green; however, as Figure 3.10 on page 116 at right shows, distractors can also be blue or orange even though not present in the legend. Furthermore, distractors can share similar aspects of the target glyph such that they are not targets themselves. For instance, there are a number of green coloured glyphs in the set of alternatives but only one of these alternatives is the correct answer.

Ethics Review

A social and behavioural research ethics committee reviewed the experiment design and granted approval for the experiment to proceed. There were no ethical concerns raised throughout the course of this research including either adverse health effects or breaches to personal privacy. Additionally, there were no other correspondences from any participant, unsolicited or otherwise other than that collected by the experimental apparatus. Participants were regarded as having provided informed consent by agreeing to the terms and conditions presented on the experiment welcome screen in a modal pop-up box.

3.6.4 Results

Analytical Procedure

Results are reported for glyph dimensionality - number of data features encoded to a target glyph - and then by the number of motion features in the encoding, and finally by graphical feature type.

In addition to time and error dependent variables, trials are divided into four groups based on completion status: successful completion on first attempt, successful completion within attempt threshold, failed attempts and the number of trials that participants opted to skip.

A number of outliers in the data set were identified and excluded based on exceeding mean answer time and or preparation time by more than 2.5 standard deviations. The prevalence of outliers was likely due to the online design of the experiment and was anticipated; either in the comfort of one's home or in a public space, interruptions are a facet of everyday life. In addition, the first trial for all participants, regardless of icon complexity, was excluded from the analysis as a graph of average performance against trial number revealed a distinct learning effect present.

Based on a data set of 1000 trials, having excluded the first trial of each participant, the remaining data set consisted of 123 static-only, 161 motion-only, and 490 combined trials which were successfully answered, thus totalling 774 trials for the analysis. Of

the unaccounted-for trials, 48 were skipped, 92 were failed attempts, 46 were dropped as outliers and 40 were first trial responses.

Where reported, mean results are reported with 95% Confidence Intervals. A list of the statistical procedures adopted for this analysis are included in Appendix H; for significance testing, the maximum rate of Type 1 error was set at $\alpha = 0.05$.

Dimensionality

Dimensionality refers to the number of graphical attributes encoding data and therefore glyph geometric and appearance attributes used to encode data in a trial. As encodings increased in dimensionality, preparation time increased markedly: beginning with one-dimensional encodings (M=4.15 seconds, SD=3.17), then two-dimensional (M=8.34 seconds, SD=6.63), then three-dimensional (M=11.47 seconds, SD=7.62) and finally four-dimensional encodings (M=17.60 seconds, SD=11.39). A One way ANOVA was conducted for dimensionality as the independent variable and time as the dependent variable. Levene's test revealed a violation to equal variance assumption. Therefore, a Brown-Forsythe statistic was adopted to deal with this violation. The effect of dimensionality on preparation time was significant $F(3,382.7)=70.75, p<0.001$. Post-hoc analyses using Games-Howell, due to a violation of equal variance assumption, indicated that all differences were significant at the $p<0.001$ level. These results are presented graphically in Figure 3.12.

Answer time followed a similar pattern with the fastest being one-dimensional encodings (M=4.36 seconds, SD=3.61), then two dimensional-encodings (M=6.83 seconds, SD=6.73), then three-dimensional encodings (M=9.52 seconds, SD=9.41) and finally the slowest being four-dimensional encodings (M=13.02 seconds, SD=11.95). Answer time was typically faster than preparation time, however despite the long preparation times, answer times were overly longer for high-dimensional targets. A Brown-Forsythe statistic indicated the effect of dimensionality on answer time was significant: $F(3,424.9)=27.0, p<0.001$. Post-hoc analyses using Games-Howell indicated that all differences were significant. These results are presented graphically in Figure 3.13.

In relation to error, participants made more errors as dimensionality increased. One-dimensional encodings resulted in the least number of errors (M=0.57 errors, SD=1.14), then two-dimensional encodings (M=0.75 errors, SD=1.16), then three-dimensional encodings (M=0.80 errors, SD=1.22) and finally four-dimensional encodings (M=1.16 errors, SD=1.36). A Brown-Forsythe statistic indicated the effect of dimensionality on attempts was significant $F(3,588.9)=3.95, p<0.01$. Post-hoc analyses using Games-Howell indicated a significant difference in errors for one- and four-dimensional encodings ($p<0.01$) only. These results are presented graphically in Figure 3.14.

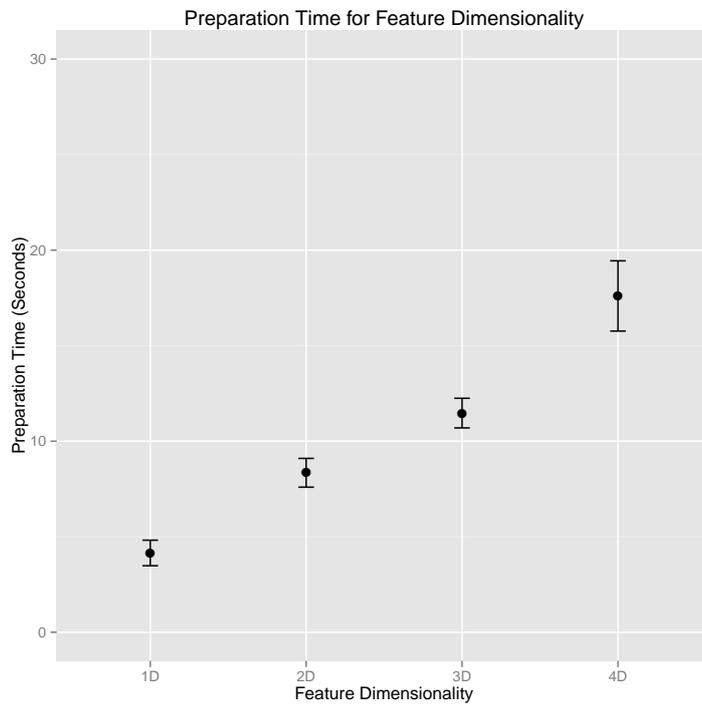


Fig. 3.12: A graph of preparation time for dimensionality; error bars are 95% Confidence Intervals.

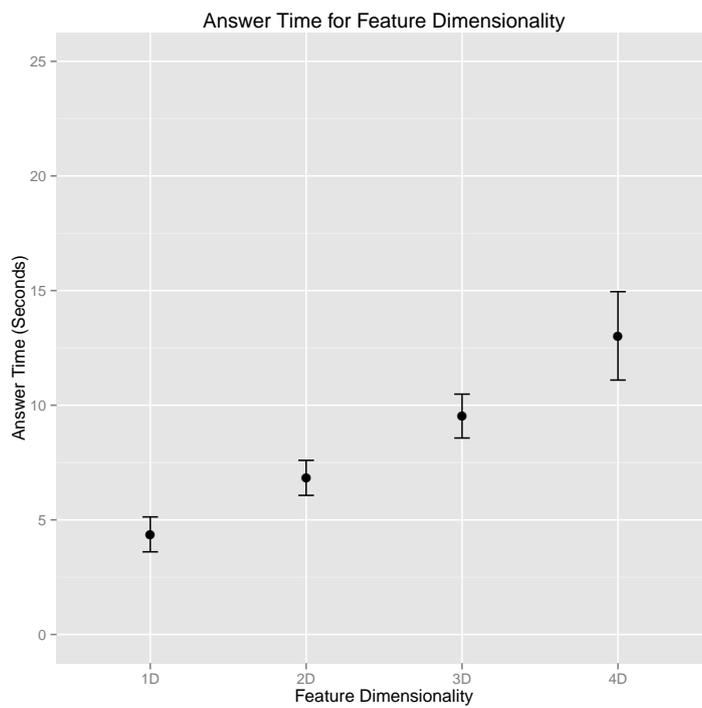


Fig. 3.13: A graph of answer time for dimensionality; error bars are 95% Confidence Intervals.

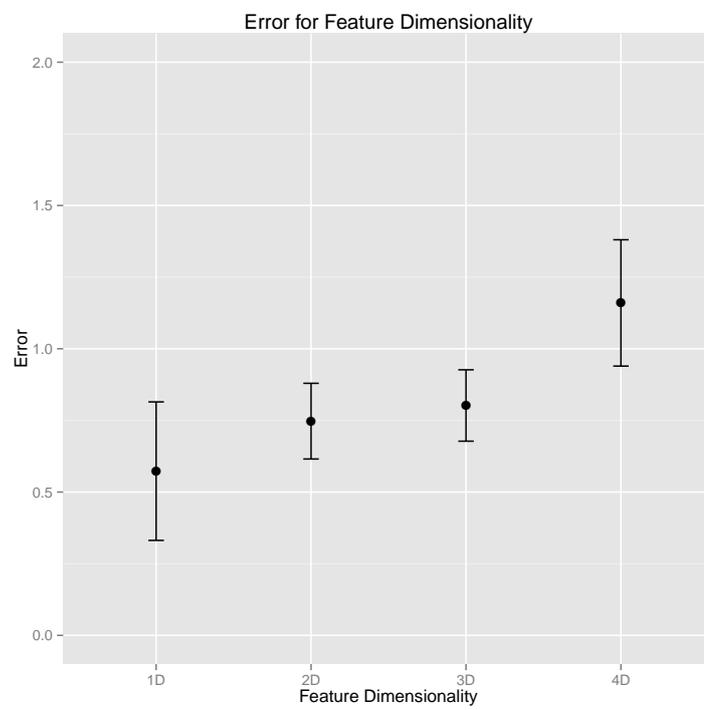


Fig. 3.14: A graph of error for dimensionality; error bars are 95% Confidence Intervals.

Static versus Dynamic versus Combined

While the earlier results relate to the effect of an increasing number of features, the next results relate to the effect of an increasing number of motion features. In addition to the overhead of additional dimensions, an effect of motion attributes was evident as well. Figure 3.15 on page 128 to Figure 3.17 on page 129 show the influence of dimensionality as well as the influence of motion attributes for preparation time, answer time and errors.

Overall, for 1-dimensional encodings, preparation time, answer time and error for static and motion features were very similar and no significant differences were indicated. In contrast, for 2-dimensional encodings, answer time and preparation time increased as additional motion attributes were incorporated into the encoding. Using the Brown-Forsythe statistic due to a violation of equal variance according to Levene's statistic, the effect of motion feature count on preparation time was significant $F(2,147.17)=7.58$, $p<0.05$. Post-hoc analyses using Games-Howell indicated a significant difference in preparation time between encodings consisting of two static features and two dynamic features only ($p<0.01$). An effect of motion feature count on answer time was also significant: $F(2,208.29)=3.73$, $p<0.05$ and post-hoc tests indicated that the difference of answer time was significantly different between two motion features and one motion feature present in the encoding. Against the trend, errors dip when a single motion feature is introduced to a 2-dimensional encoding but return to trend with the addition of a further motion feature. The effect of motion feature count on errors was significant $F(2,201.4)=3.78$, $p<0.05$ though post-hoc tests indicate that the difference was significant for the one feature and two feature encoding only.

Beyond two dimensions, an increasingly evident trend of time and error with additional motion features was evident; however, the variation between trials also increased. In relation to 3-dimensional encodings, preparation time climbed slowly and consistently as additional motion features were introduced, though these differences were not significant $F(3,155.80)=2.38$, $p=0.71$. In contrast, answer time grew more rapidly with increasing motion feature count. This effect on answer time was significant $F(3,177.89)=4.48$, $p<0.01$. Post-hoc tests indicated a significant difference between three motion feature and one motion feature encodings ($p<0.05$) and three motion feature and two motion feature encodings ($p<0.05$) only.

Finally, for 4-dimensional encodings, a clear trend was evident as depicted in the graphs of Figure 3.15 on page 128, Figure 3.16 on page 128 and Figure 3.17 on page 129. Preparation time increased with increasing motion feature count and this effect of motion was significant $F(3,43.31)=5.20$, $p<0.05$. However, despite the trend evident for answer time, a Brown-Forsythe statistic due to a violation of equal variance according to Levene's statistic, only approached significance $F(3, 53.05)=2.23$, $p=0.95$. Similarly, errors increased with increasing motion feature count, although these differences were

not flagged as significant. Four feature static-only trials are not shown on this graph as the encoding rules precluded four feature static-only tasks from experimentation.

The two feature static-motion combination deviates strongly from trend; in contrast, as dimensionality and motion feature count increases - so too does error rate. A clear explanation for this inconsistency is elusive. However, on review of data tables for two dimensional trials - see Table 3.9 on page 135 and Table 3.10 on page 138 - 2D zero motion feature trials, 30% of trials include a hue feature and these trials alone account for 22% of the error. In contrast, in 2D one motion feature trials, 38% of trials include a hue feature yet these trials account for only 5% of the error. Whilst further inspection reveals that error rate is lower in the latter, it is unclear why a combination of hue and a motion feature involves a superior error rate as compared to a hue and static feature combination. Ultimately, future research is needed with larger sample sizes to make more robust predictions.

In addition to the results reported for errors, Table 3.6 presents the proportion of trials completed on the first attempt, trials successfully completed on a subsequent attempt, trials considered failed due to excessive attempts, and finally the proportion of skipped trials. Within dimensions, the results reveal a tendency for an increase in attempts and proportion of skipped trials with an increasing motion feature count.

Separate exact Chi-Square tests for each encoding dimensionality were conducted to observe a relationship between the number of motion features and the trial outcome. However, no relationship was observed for the trial outcome and the number of motion features in 1-dimensional encodings $\chi(3, N=89)=0.40, p=0.93$; 2-dimensional encodings $\chi(6, N=301)=11.67, p=0.07$; 3-dimensional encodings $\chi(9, N=374)=16.09, p=0.07$; or 4-dimensional encodings $\chi(9, N=150)=13.80, p=0.13$.

Additionally, there appeared to be an effect of dimensionality on trial outcome that was not explored in the earlier section. Dimensionality of an encoding appeared to influence the likelihood that a trial was skipped or required additional attempts to locate the correct answer. For completeness, a Chi-Square test for dimensionality was also carried out to explore the relationship of dimensionality and trial outcome. There was a relationship observed for dimensionality and trial outcome $\chi(9, N=914)=18.51, p=0.03$.

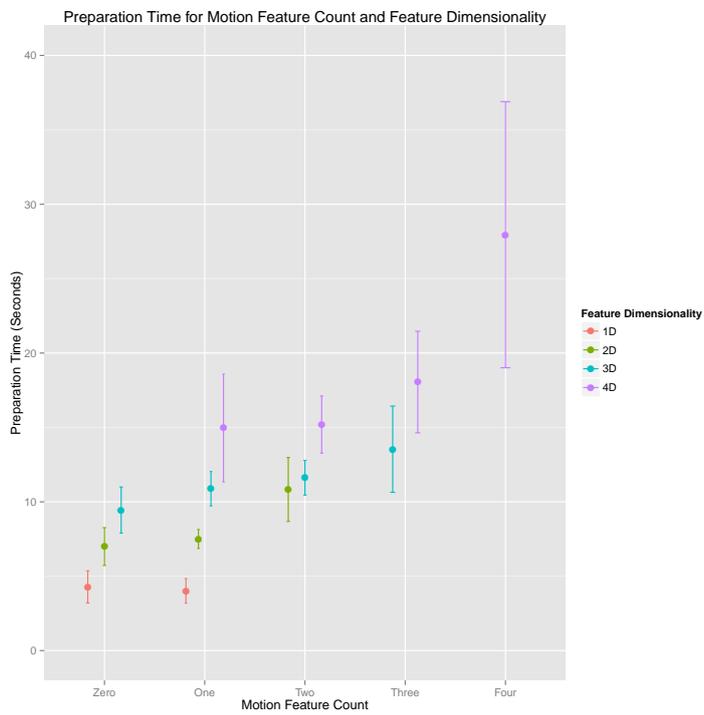


Fig. 3.15: A graph of preparation time (in seconds) for number of motion features and dimensionality; error bars are 95% Confidence Intervals.

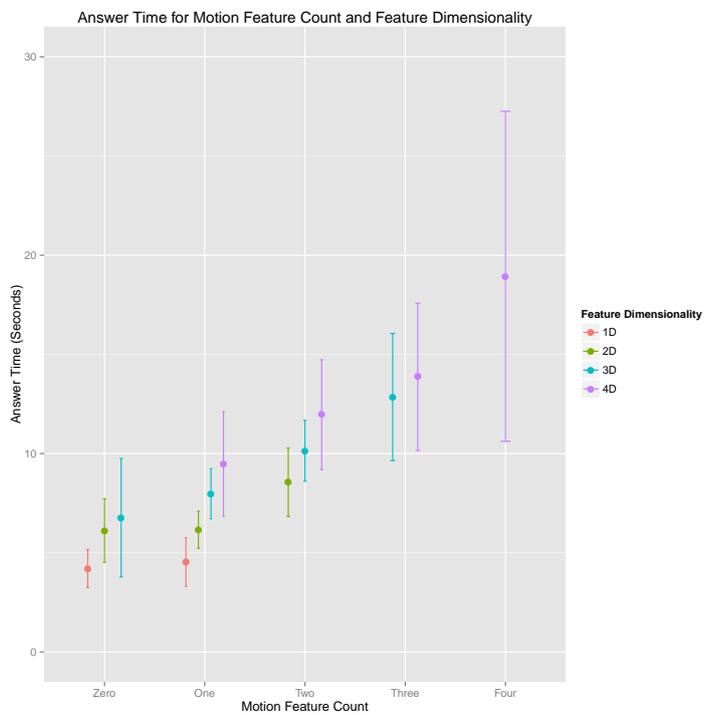


Fig. 3.16: A graph of answer time (in seconds) for number of motion features and dimensionality; error bars are 95% Confidence Intervals.

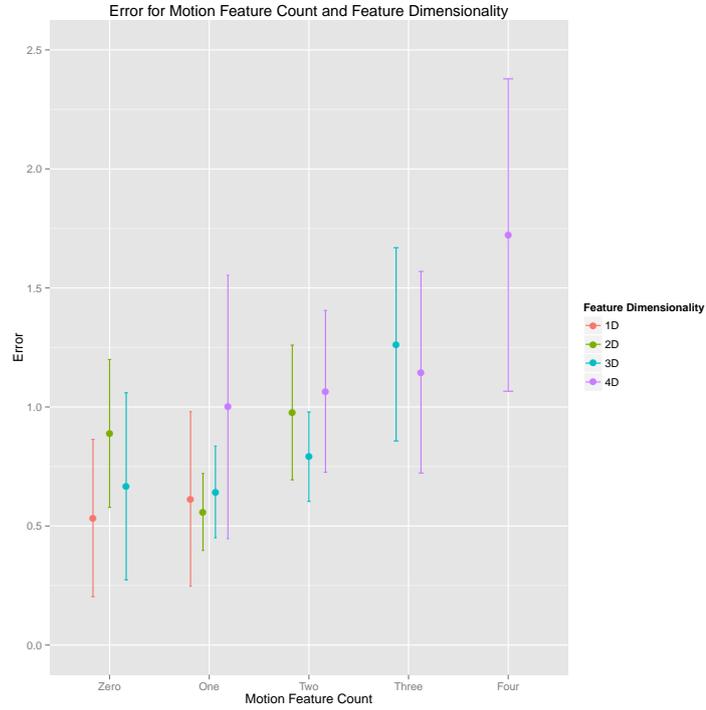


Fig. 3.17: A graph of error for number of motion features and dimensionality; error bars are 95% Confidence Intervals.

Tab. 3.6: Trial outcome by success category: successful on first attempt, successful on subsequent attempt, unsuccessful and exceeded attempt threshold and skipped trial.

Features	Motion Features	Trial Outcome			
		First	Subsequent	Threshold	Skipped
1	One	72.7%	15.9%	4.5%	6.8%
	Zero	73.3%	15.6%	6.7%	4.4%
	Two	54.7%	26.7%	11.6%	7.0%
2	One	69.7%	20.4%	8.6%	1.3%
	Zero	54.0%	31.7%	7.9%	6.3%
	Three	52.6%	15.8%	21.1%	10.5%
3	Two	59.5%	26.6%	8.9%	5.1%
	One	66.7%	23.0%	7.1%	3.2%
	Zero	63.6%	24.2%	9.1%	3.0%
4	Four	22.2%	50.0%	16.7%	11.1%
	Three	50.0%	27.1%	10.4%	12.5%
	Two	54.1%	23.0%	19.7%	3.3%
	One	47.8%	39.1%	4.3%	8.7%

Graphical Features

1 Dimensional The next results report time and error for the graphical features: size, angle, saturation, hue, grow, rotate, flash and shuffle and for select combinations as outlined in Section 3.6.1 on page 102. It is apparent that a low sample size was affecting the analysis of results at this level.

Figure 3.18 on page 132, Figure 3.19 on page 132 and Figure 3.20 on page 133 show preparation time, answer time and error for single feature encodings. Hue trials were fastest to complete and attracted the least error and the confidence intervals were slightly tighter about the mean. Interestingly, the second fastest feature by answer time was the shuffle motion feature. In contrast, the two slowest features were both motion features, grow and flash, and furthermore, the confidence intervals were wider about their mean indicating greater variation within trials. Participants completing grow and flash feature trials were also slower to prepare; however, participants undertaking the angle feature type also required a lengthy preparation process suggesting that interpretation of orientation was difficult. Separate 1-Way Analyses of variance were conducted for 1 dimensional feature types and dependent variables: preparation time, answer time and error. The effect of feature type on preparation time was not significant $F(7,71)=2.10$, $p=0.05$. Additionally, the effect of feature type was not significant for answer time: $F(7,71)=0.95$, $p=0.47$. The effect of feature type on error was not significant also, with $F(7,71)=1.43$, $p=0.20$.

Tab. 3.7: Success outcome for 1 dimensional trials; S1 denotes success first attempt; S23 denotes success subsequent attempts; F denotes failed trails; Sk denotes skipped trials.

Condition	Count	Success			
		S1	S23	F	Sk
Flash	11	8	1	0	2
Grow	12	9	2	0	1
Rotate	11	6	3	2	0
Shuffle	10	9	1	0	0
Angle	12	6	1	3	2
Hue	11	11	0	0	0
Saturation	11	9	2	0	0
Size	11	7	4	0	0

Tab. 3.8: Dependent variables by feature type for 1 dimensional trials; CI denotes 95% confidence intervals.

Condition	N	Preparation Time	CI	Answer Time	CI	Error	CI
Angle	7.00	6.72	1.39-12.06	4.42	2.68-6.16	0.14	-0.21-0.49
Flash	9.00	4.68	2.91-6.45	5.10	1.76-8.43	0.11	-0.15-0.37
Grow	11.00	4.50	3.17-5.83	5.59	2.52-8.67	0.18	-0.09-0.45
Hue	11.00	3.45	2.38-4.53	2.55	1.03-4.06	0.00	0-0
Rotate	9.00	3.40	2.57-4.22	4.38	2.11-6.65	0.56	-0.12-1.23
Saturation	11.00	4.22	2.32-6.12	4.96	2.77-7.15	0.18	-0.09-0.45
Shuffle	10.00	3.03	2.11-3.95	3.27	1.01-5.53	0.10	-0.13-0.33
Size	11.00	2.88	1.94-3.82	4.58	2.56-6.59	0.36	0.02-0.7

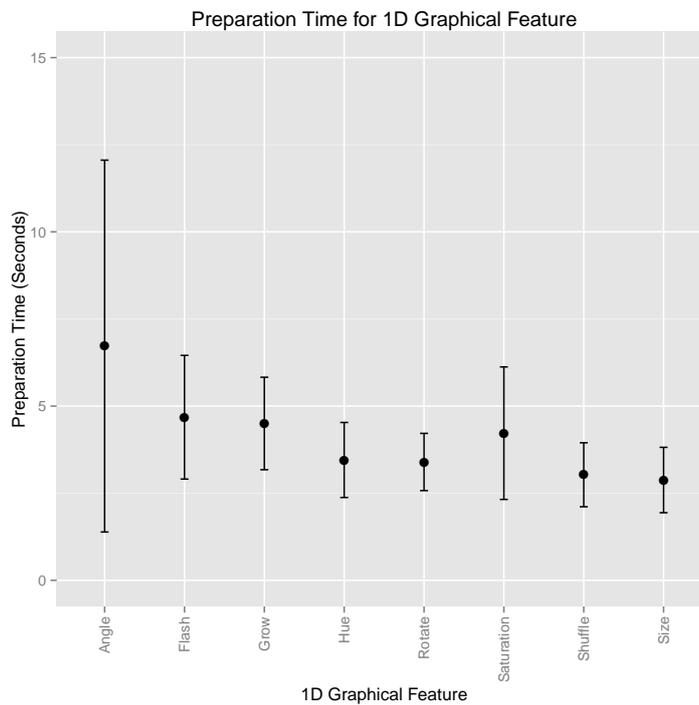


Fig. 3.18: A graph of preparation time (in seconds) for one dimensional trials; error bars are 95% Confidence Intervals.

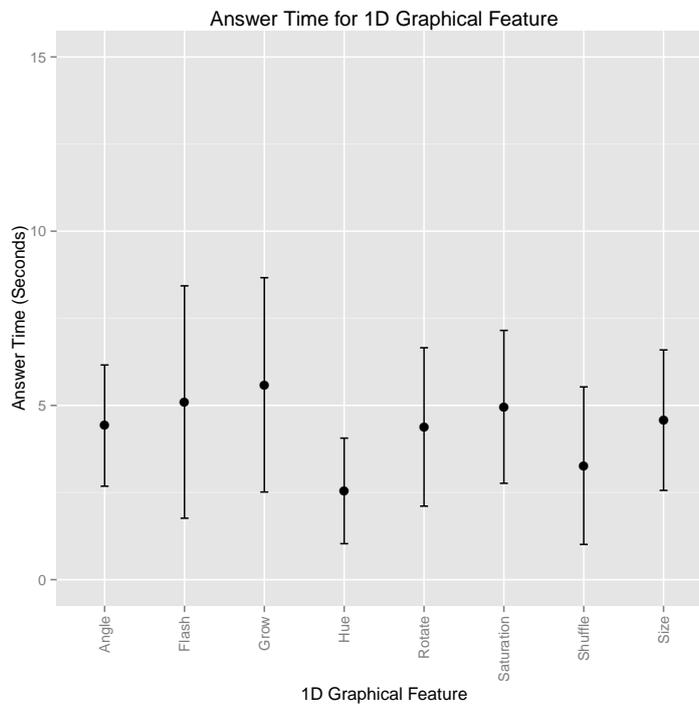


Fig. 3.19: A graph of answer time (in seconds) for one dimensional trials; error bars are 95% Confidence Intervals.

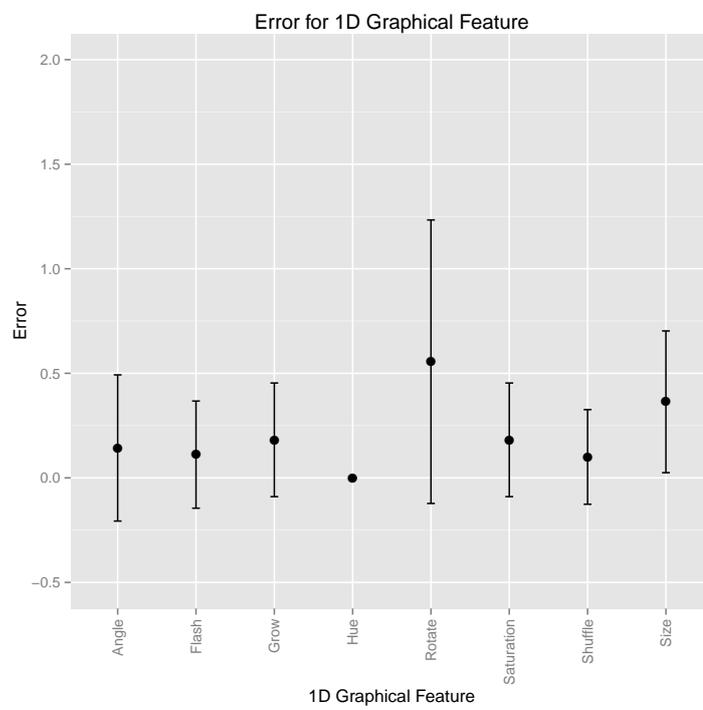


Fig. 3.20: A graph of error for one dimensional trials; error bars are 95% Confidence Intervals.

2 Dimensional Separate 1-way analyses of variance were conducted for 2 dimensional feature types and dependent variables: preparation time, answer time and error. The effect of feature type on preparation time was significant $F(22,238)=2.19$, $p=0.002$. Likewise, the effect of feature type was significant for answer time: $F(22,238)=1.81$, $p=0.01$. However, the effect of feature type on error was not significant $F(22,238)=1.40$, $p=0.11$.

TukeyHSD post-hoc tests were conducted for answer time in 2 dimensional trials. This test indicated significant differences between Size-Hue/Angle-Flash ($p=0.04$), Size-Saturation/Angle-Flash ($p<0.001$), Size-Angle/Angle-Flash ($p=0.01$), Saturation-Shuffle/Angle-Flash ($p=0.03$), Rotate-Saturation/Angle-Flash ($p=0.03$), Rotate-Hue/Angle-Flash ($p=0.004$), Hue-Shuffle/Angle-Flash ($p=0.001$), Grow-Saturation/Angle-Flash ($p=0.01$), Grow-Hue/Angle-Flash ($p<0.001$), Angle-Shuffle/Angle-Flash ($p=0.002$) and Angle-Hue/Angle-Flash ($p<0.001$). From this analysis, it is clear that the Angle-Flash combination invites significant time overheads to decode during a trial.

Tab. 3.9: Success outcome for 2 dimensional trials; S1 denotes success first attempt; S23 denotes success subsequent attempts; F denotes failed trails; Sk denotes skipped trials.

Condition	N	Success			
		S1	S23	F	Sk
Flash Shuffle	11	6	3	2	0
Grow Flash	16	7	4	2	3
Grow Rotate	14	6	6	2	0
Grow Shuffle	9	5	2	0	2
Rotate Flash	13	7	2	4	0
Rotate Shuffle	23	16	6	0	1
Angle Flash	8	3	2	2	1
Angle Shuffle	17	13	3	1	0
Grow Angle	14	9	1	4	0
Grow Hue	21	20	1	0	0
Grow Saturation	13	7	5	1	0
Hue Shuffle	12	10	2	0	0
Rotate Hue	13	12	1	0	0
Rotate Saturation	10	8	2	0	0
Saturation Shuffle	11	6	4	1	0
Size Flash	10	5	3	1	1
Size Rotate	11	5	4	2	0
Size Shuffle	12	8	3	1	0
Angle Hue	12	9	2	0	1
Angle Saturation	17	7	8	1	1
Size Angle	12	5	5	2	0
Size Hue	12	9	2	0	1
Size Saturation	10	4	3	2	1

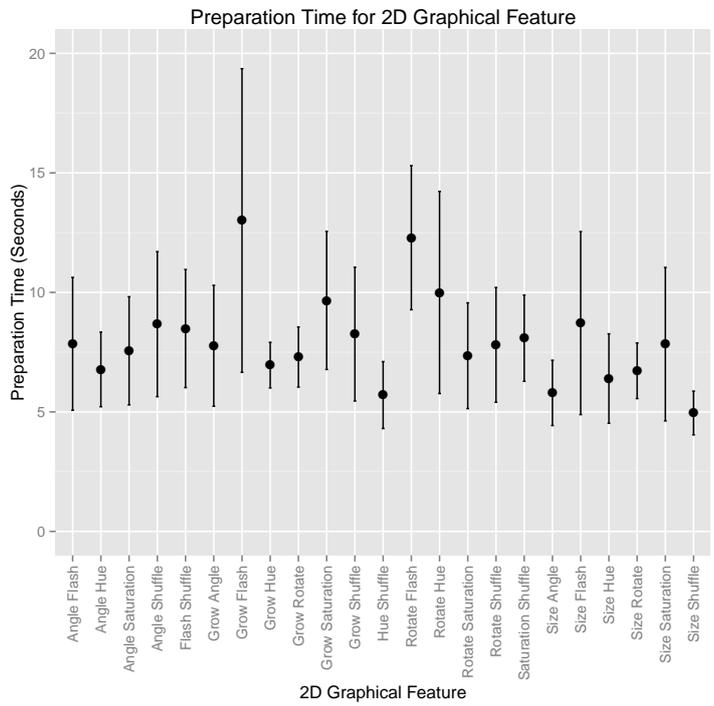


Fig. 3.21: A graph of preparation time (in seconds) for two dimensional trials; error bars are 95% Confidence Intervals.

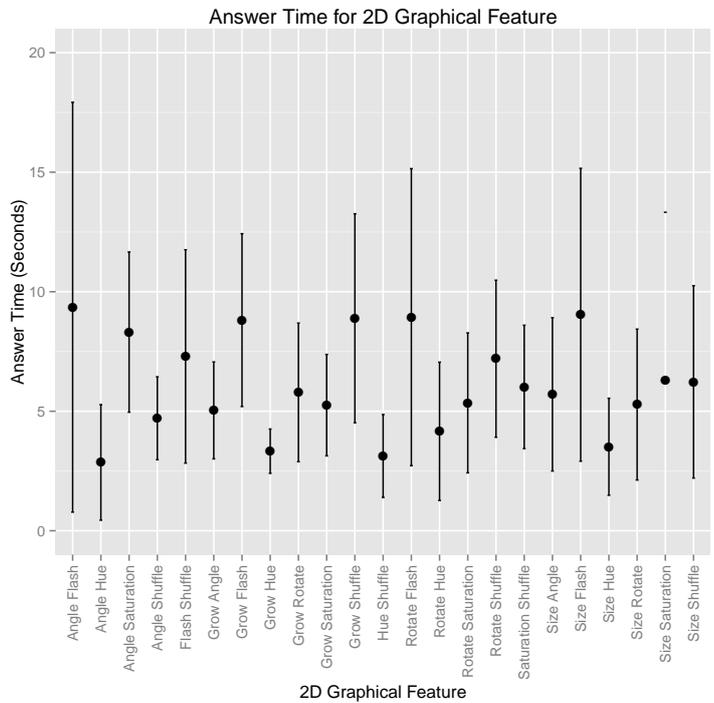


Fig. 3.22: A graph of answer time (in seconds) for two dimensional trials; error bars are 95% Confidence Intervals.

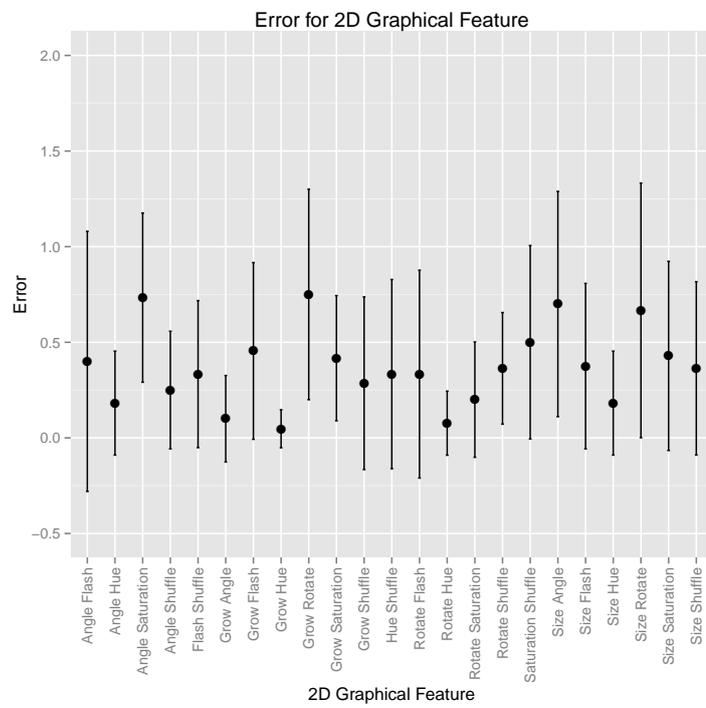


Fig. 3.23: A graph of error for two dimensional trials; error bars are 95% Confidence Intervals.

Tab. 3.10: Dependent variables by feature type for 2 dimensional trials; Sat. denotes saturation; CI denotes 95% confidence intervals.

Condition	N	Preparation Time	CI	Answer Time	CI	Error	CI
Angle Flash	5.00	7.85	5.07-10.63	9.35	0.78-17.92	0.40	-0.28-1.08
Angle Hue	11.00	6.78	5.22-8.34	2.86	0.44-5.27	0.18	-0.09-0.45
Angle Sat.	15.00	7.56	5.29-9.82	8.31	4.96-11.66	0.73	0.29-1.18
Angle Shuffle	16.00	8.67	5.64-11.7	4.71	2.97-6.44	0.25	-0.06-0.56
Flash Shuffle	9.00	8.49	6.02-10.96	7.29	2.83-11.75	0.33	-0.05-0.72
Grow Angle	10.00	7.77	5.24-10.29	5.03	3.01-7.06	0.10	-0.13-0.33
Grow Flash	11.00	13.01	6.66-19.36	8.81	5.2-12.43	0.45	-0.01-0.92
Grow Hue	21.00	6.96	6-7.91	3.33	2.4-4.25	0.05	-0.05-0.15
Grow Rotate	12.00	7.30	6.04-8.55	5.79	2.89-8.69	0.75	0.2-1.3
Grow Sat.	12.00	9.66	6.77-12.55	5.25	3.13-7.38	0.42	0.09-0.74
Grow Shuffle	7.00	8.25	5.46-11.05	8.89	4.52-13.26	0.29	-0.17-0.74
Hue Shuffle	12.00	5.70	4.31-7.1	3.13	1.4-4.86	0.33	-0.16-0.83
Rotate Flash	9.00	12.29	9.27-15.3	8.93	2.72-15.15	0.33	-0.21-0.88
Rotate Hue	13.00	9.99	5.77-14.22	4.16	1.27-7.05	0.08	-0.09-0.24
Rotate Sat.	10.00	7.35	5.14-9.56	5.35	2.42-8.28	0.20	-0.1-0.5
Rotate Shuffle	22.00	7.80	5.4-10.2	7.20	3.91-10.48	0.36	0.07-0.66
Sat. Shuffle	10.00	8.08	6.28-9.89	6.02	3.44-8.6	0.50	-0.01-1.01
Size Angle	10.00	5.79	4.43-7.16	5.71	2.5-8.91	0.70	0.11-1.29
Size Flash	8.00	8.71	4.89-12.54	9.04	2.91-15.16	0.38	-0.06-0.81
Size Hue	11.00	6.40	4.53-8.27	3.51	1.49-5.54	0.18	-0.09-0.45
Size Rotate	9.00	6.72	5.56-7.89	5.28	2.12-8.44	0.67	0-1.33
Size Sat.	7.00	7.83	4.62-11.04	6.31	-0.71-13.32	0.43	-0.07-0.92
Size Shuffle	11.00	4.95	4.04-5.87	6.23	2.2-10.25	0.36	-0.09-0.82

3 Dimensional Separate 1-way analyses of variance were conducted for 3 dimensional feature types and dependent variables: preparation time, answer time and error. The effect of feature type on preparation time was not significant $F(27,289)=1.46$, $p=0.06$. In contrast, the effect of feature type was significant for answer time: $F(27,289)=1.89$, $p=0.005$. However, the effect of feature type on error was not significant $F(27,289)=0.82$, $p=0.71$.

TukeyHSD post-hoc tests were conducted for answer time in 3 dimensional trials. This test indicated significant differences between Size-Rotate-Hue / Grow-Angle-Shuffle ($p=0.01$), Size-Hue-Shuffle/Grow-Angle-Shuffle ($p=0.01$), Size-Angle-Hue/Grow-Angle-Shuffle ($p=0.002$), Grow-Rotate-Hue/Grow-Angle-Shuffle ($p=0.01$), Grow-Angle-Shuffle/Angle-Hue-Shuffle ($p=0.01$).

Tab. 3.11: Success outcome for 4 dimensional trials; S1 denotes success first attempt; S23 denotes success subsequent attempts; F denotes failed trails; Sk denotes skipped trials.

Condition	N	Success			
		S1	S23	F	Sk
Grow Flash Shuffle	21	12	5	3	1
Grow Rotate Flash	13	4	3	4	2
Grow Rotate Shuffle	11	8	0	2	1
Rotate Flash Shuffle	12	6	1	3	2
Angle Flash Shuffle	11	7	3	1	0
Grow Angle Flash	13	6	5	2	0
Grow Angle Shuffle	10	2	4	2	2
Grow Hue Shuffle	10	6	4	0	0
Grow Rotate Hue	21	17	3	1	0
Grow Rotate Saturation	17	10	6	0	1
Grow Saturation Shuffle	11	8	2	1	0
Rotate Hue Shuffle	12	9	2	1	0
Rotate Saturation Shuffle	12	5	6	1	0
Size Flash Shuffle	11	6	3	2	0
Size Rotate Flash	18	8	2	3	5
Size Rotate Shuffle	12	10	2	0	0
Angle Hue Shuffle	11	8	3	0	0
Angle Saturation Shuffle	11	7	3	0	1
Grow Angle Hue	17	11	3	3	0
Grow Angle Saturation	12	6	5	0	1
Size Angle Flash	13	5	3	3	2
Size Angle Shuffle	12	9	3	0	0
Size Hue Shuffle	12	10	1	1	0
Size Rotate Hue	18	14	3	1	0
Size Rotate Saturation	9	7	2	0	0
Size Saturation Shuffle	11	7	3	1	0
Size Angle Hue	19	15	3	0	1
Size Angle Saturation	14	6	5	3	0

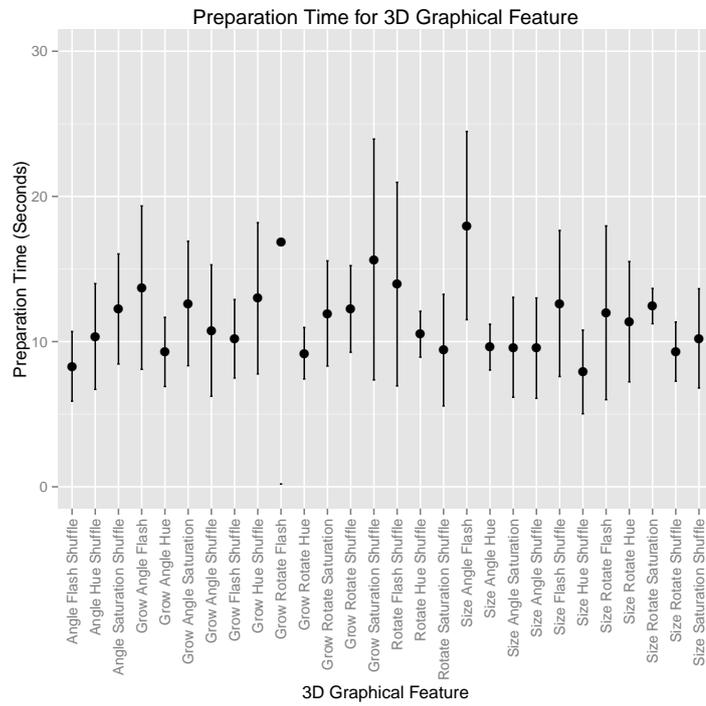


Fig. 3.24: A graph of preparation time (in seconds) for three dimensional trials; error bars are 95% Confidence Intervals.

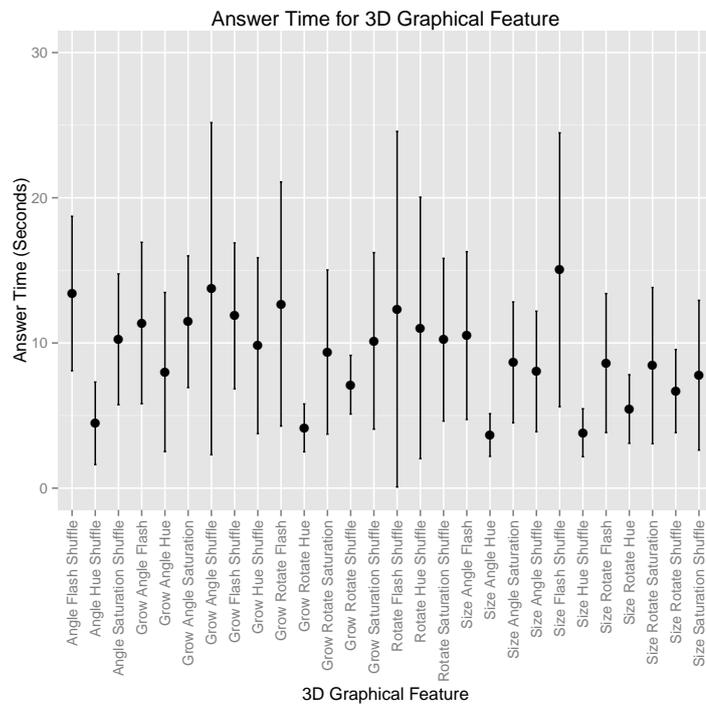


Fig. 3.25: A graph of answer time (in seconds) for three dimensional trials; error bars are 95% Confidence Intervals.

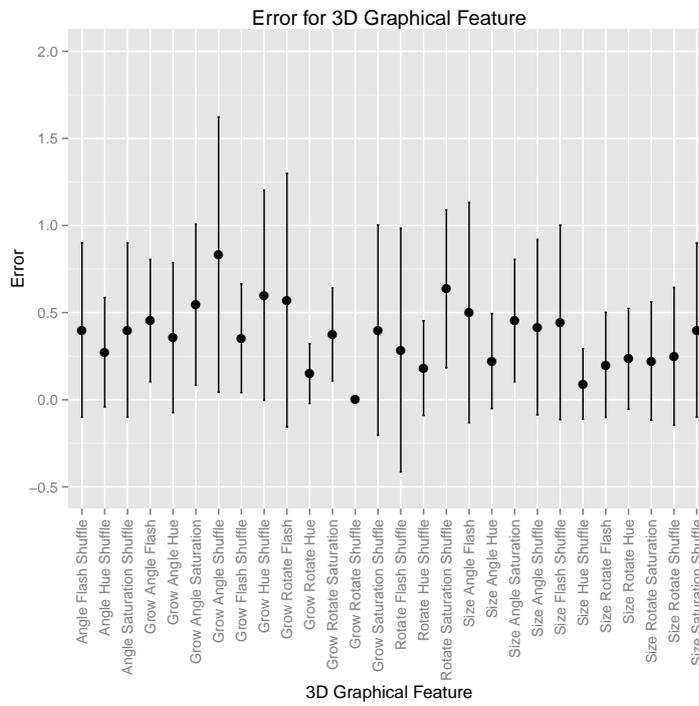


Fig. 3.26: A graph of error for three dimensional trials; error bars are 95% Confidence Intervals.

Tab. 3.12: Dependent variables by feature type for 3 dimensional trials; CI denotes 95% confidence intervals.

Condition	N	Preparation Time	CI	Answer Time	CI	Error	CI
Angle Flash Shuffle	10.00	8.30	5.9-10.69	13.40	8.09-18.72	0.40	-0.1-0.9
Angle Hue Shuffle	11.00	10.35	6.71-13.99	4.47	1.63-7.31	0.27	-0.04-0.59
Angle Sat. Shuffle	10.00	12.25	8.46-16.04	10.25	5.75-14.75	0.40	-0.1-0.9
Grow Angle Flash	11.00	13.72	8.09-19.34	11.37	5.82-16.93	0.45	0.1-0.81
Grow Angle Hue	14.00	9.29	6.91-11.67	8.00	2.52-13.48	0.36	-0.07-0.79
Grow Angle Sat.	11.00	12.63	8.34-16.91	11.46	6.93-16	0.55	0.08-1.01
Grow Angle Shuffle	6.00	10.76	6.23-15.29	13.74	2.3-25.17	0.83	0.04-1.62
Grow Flash Shuffle	17.00	10.20	7.5-12.89	11.86	6.84-16.89	0.35	0.04-0.66
Grow Hue Shuffle	10.00	12.98	7.78-18.19	9.82	3.76-15.87	0.60	0-1.2
Grow Rotate Flash	7.00	16.89	0.19-33.58	12.68	4.28-21.09	0.57	-0.16-1.3
Grow Rotate Hue	20.00	9.20	7.43-10.97	4.15	2.5-5.8	0.15	-0.02-0.32
Grow Rotate Sat.	16.00	11.94	8.32-15.56	9.37	3.72-15.03	0.38	0.11-0.64
Grow Rotate Shuffle	8.00	12.25	9.27-15.23	7.12	5.1-9.15	0.00	0-0
Grow Sat. Shuffle	10.00	15.66	7.36-23.95	10.14	4.06-16.22	0.40	-0.2-1
Rotate Flash Shuffle	7.00	13.96	6.95-20.97	12.32	0.08-24.57	0.29	-0.41-0.98
Rotate Hue Shuffle	11.00	10.51	8.93-12.09	11.04	2.03-20.04	0.18	-0.09-0.45
Rotate Sat. Shuffle	11.00	9.41	5.57-13.26	10.22	4.62-15.82	0.64	0.18-1.09
Size Angle Flash	8.00	17.99	11.51-24.47	10.51	4.73-16.29	0.50	-0.13-1.13
Size Angle Hue	18.00	9.62	8.04-11.2	3.66	2.19-5.13	0.22	-0.05-0.49
Size Angle Sat.	11.00	9.61	6.17-13.05	8.66	4.5-12.82	0.45	0.1-0.81
Size Angle Shuffle	12.00	9.55	6.09-13.01	8.03	3.88-12.19	0.42	-0.09-0.92
Size Flash Shuffle	9.00	12.62	7.6-17.65	15.04	5.61-24.47	0.44	-0.11-1
Size Hue Shuffle	11.00	7.91	5.03-10.79	3.82	2.17-5.47	0.09	-0.11-0.29
Size Rotate Flash	10.00	11.99	6-17.97	8.61	3.83-13.4	0.20	-0.1-0.5
Size Rotate Hue	17.00	11.37	7.23-15.51	5.45	3.09-7.81	0.24	-0.05-0.52
Size Rotate Sat.	9.00	12.45	11.23-13.66	8.44	3.07-13.82	0.22	-0.12-0.56
Size Rotate Shuffle	12.00	9.31	7.28-11.35	6.69	3.83-9.55	0.25	-0.14-0.64
Size Sat. Shuffle	10.00	10.22	6.81-13.64	7.78	2.62-12.93	0.40	-0.1-0.9

4 Dimensional Figure 3.27 on the next page, Figure 3.28 on page 146 and Figure 3.29 on page 146 show preparation time, answer time and errors for four-dimensional trials. Trials involving hue - and one or two other static features - account for the first five out of six fastest trials by answer time. In contrast, the slowest trials involve the largest motion feature counts and few static features. By inspection of the graphs few other combinations appear to differ greatly from one another. Separate 1-Way Analyses of variance were conducted for 4 dimensional feature types and dependent variables: preparation time, answer time and error. The effect of feature type on preparation time was significant $F(11,105)=1.96$, $p=0.03$. In contrast, the effect of feature type was not significant for answer time: $F(11,105)=1.78$, $p=0.06$. Furthermore, the effect of feature type on error was not significant $F(11,105)=1.26$, $p=0.25$.

Post-hoc tests were conducted for preparation time in 4 dimensional trials. This test indicated significant differences between Grow-Rotate-Flash-Shuffle/Grow-Angle-Saturation-Shuffle ($p=0.01$) only.

Tab. 3.13: Success outcome for 4 dimensional trials; S1 denotes success first attempt; S23 denotes success subsequent attempts; F denotes failed trails; Sk denotes skipped trials.

Condition	N	Success			
		S1	S23	F	Sk
Grow Rotate Flash Shuffle	18	4	9	3	2
Grow Angle Flash Shuffle	14	6	3	1	4
Grow Rotate Hue Shuffle	11	7	4	0	0
Grow Rotate Saturation Shuffle	11	6	2	2	1
Size Rotate Flash Shuffle	12	5	4	2	1
Grow Angle Hue Shuffle	11	8	1	2	0
Grow Angle Saturation Shuffle	16	7	3	6	0
Size Angle Flash Shuffle	12	5	4	2	1
Size Rotate Hue Shuffle	13	10	3	0	0
Size Rotate Saturation Shuffle	9	3	3	2	1
Size Angle Hue Shuffle	12	7	4	1	0
Size Angle Saturation Shuffle	11	4	5	0	2

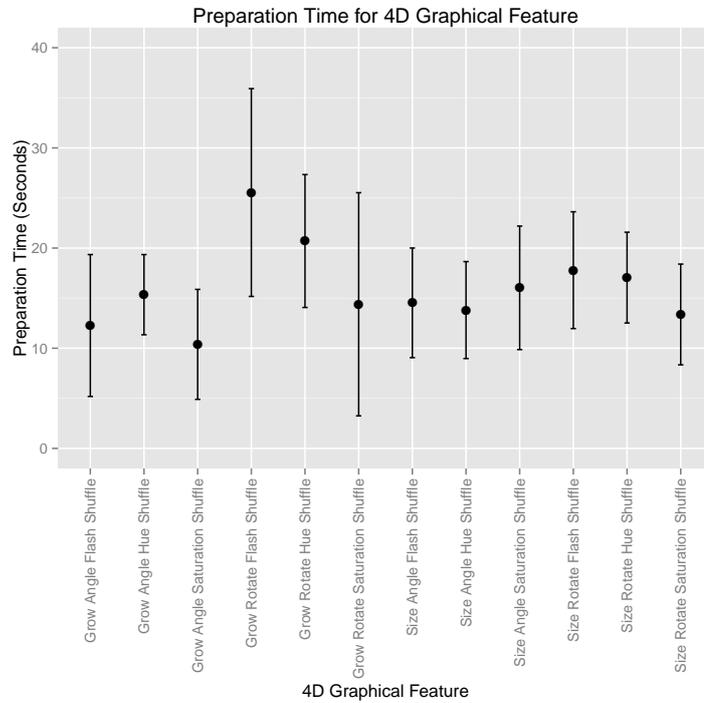


Fig. 3.27: A graph of preparation time (in seconds) for four dimensional trials; error bars are 95% Confidence Intervals.

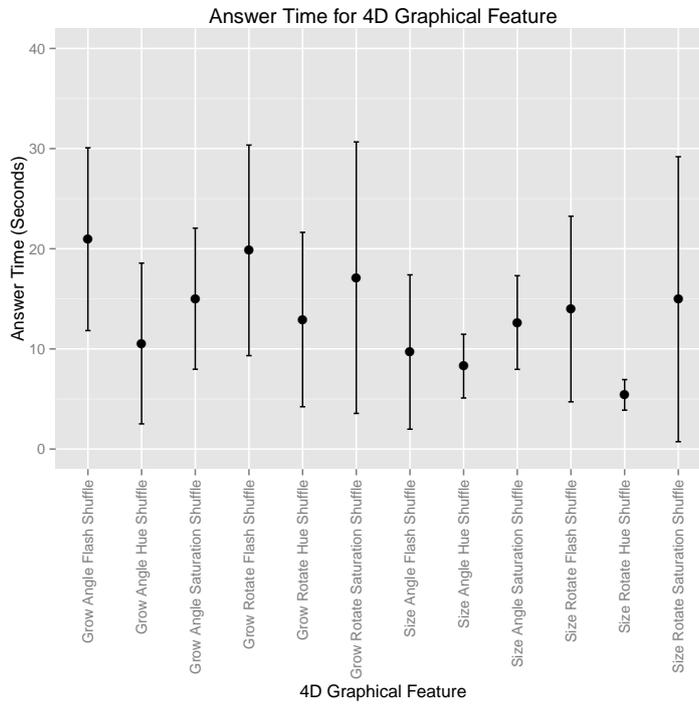


Fig. 3.28: A graph of answer time (in seconds) for four dimensional trials; error bars are 95% Confidence Intervals.

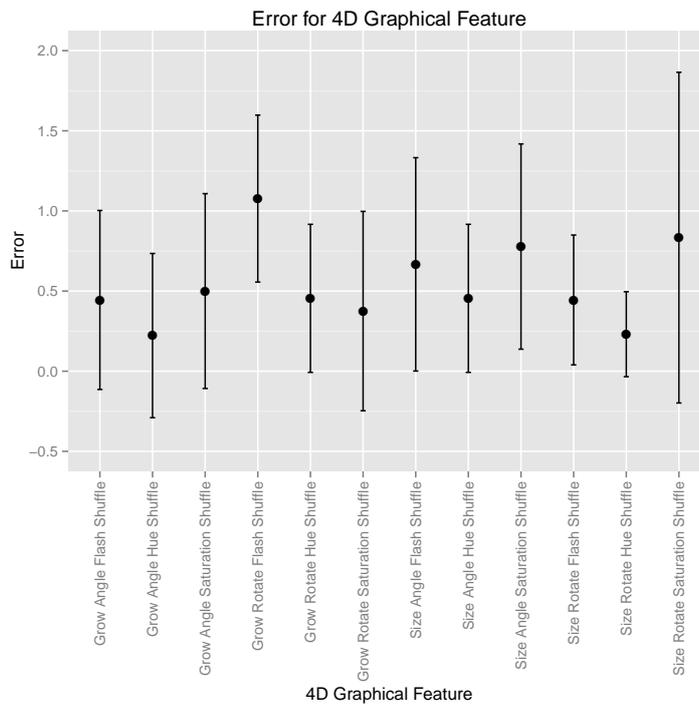


Fig. 3.29: A graph of error for four dimensional trials; error bars are 95% Confidence Intervals.

Tab. 3.14: Dependent variables by feature type for 4 dimensional trials; Sat. denotes saturation; CI denotes 95% confidence intervals.

Condition	N	Preparation Time	CI	Answer Time	CI	Error	CI
Grow Angle Flash Shuffle	9.00	12.27	5.18-19.35	20.95	11.83-30.08	0.44	-0.11-1
Grow Angle Hue Shuffle	9.00	15.35	11.34-19.35	10.54	2.52-18.57	0.22	-0.29-0.73
Grow Angle Sat. Shuffle	10.00	10.39	4.9-15.88	15.01	7.97-22.06	0.50	-0.11-1.11
Grow Rotate Flash Shuffle	13.00	25.54	15.17-35.92	19.84	9.33-30.35	1.08	0.56-1.6
Grow Rotate Hue Shuffle	11.00	20.70	14.07-27.34	12.93	4.23-21.64	0.45	-0.01-0.92
Grow Rotate Sat. Shuffle	8.00	14.39	3.24-25.53	17.11	3.56-30.66	0.38	-0.25-1
Size Angle Flash Shuffle	9.00	14.53	9.06-20.01	9.69	1.99-17.39	0.67	0-1.33
Size Angle Hue Shuffle	11.00	13.81	8.97-18.65	8.28	5.1-11.46	0.45	-0.01-0.92
Size Angle Sat. Shuffle	9.00	16.03	9.87-22.2	12.64	7.96-17.31	0.78	0.14-1.42
Size Rotate Flash Shuffle	9.00	17.79	11.96-23.62	13.98	4.72-23.25	0.44	0.04-0.85
Size Rotate Hue Shuffle	13.00	17.05	12.52-21.58	5.41	3.89-6.94	0.23	-0.03-0.5
Size Rotate Sat. Shuffle	6.00	13.37	8.34-18.4	14.96	0.74-29.19	0.83	-0.2-1.87

Exit Questionnaire

Participants responded to an exit questionnaire using a seven point Likert scale, ranging from ‘strongly agree’ to ‘strongly disagree’. Subjective results are tabulated below in Table 3.15. Responses were compressed into a three-point scale for analysis. Responses that were neutral in the expanded scale remained neutral in the compressed scale and responses on either side of neutral were compressed into a single positive and negative category.

Tab. 3.15: Subjective questionnaire data. Participants responded on a 7 point Likert scale; however, analysis transformed data to a 3 point scale.

	Agree	Neutral	Disagree
Task easy to perform	32.3%	29.0%	38.7%
Task was annoyance	54.8%	29.0%	16.1%
Animation was annoying	51.6%	35.5%	12.0%
Task was easy to understand	45.2%	25.8%	29.0%
Animation was jerky	33.3%	23.8%	42.8%

Subjective results did not indicate a clear message. Participants were split on the ease of the task though generally, participants reported that they found the task easy to understand. However, it is unclear whether participants were reporting that the task was easy to understand initially, or to what level they still did not understand the task at the conclusion of the experiment. The strongest, yet still largely inconclusive indications come from annoyance ratings. Participants indicated that both the task and the animation were annoying. Over half suggested that the task was annoying and very few thought that it was not annoying. A trend suggested that the animation was not jerky, though a third of respondents maintained that it was jerky indicating some hint at why participants believed the animation was annoying.

Subjective responses for task ease, task annoyance and animation annoyance are inconsistent with the view that might be held by a designer who is looking to install motion in an information visualisation. It is desirable for the task to be easy and pleasing and the animation also pleasing; yet participants largely reported the opposite. Moreover, whilst it is desirable for the task to be easy to understand and the animation smooth, the results have only partially indicated this to be the case.

Drop Out Analysis

On the recommendation of Reips (2002), this section presents an analysis of the prevalence of drop out in the experiment. This analysis seeks to isolate the conditions under which participants did not want to proceed or could not proceed further in the experi-

ment. For this analysis, the total participant pool size was considered to be the number of users who submitted demographics information.

The drop out graph in Figure 3.30 depicts the proportion of participants at each stage of the experiment. There are 29 stages in total encompassing the demographics form, training material, experiment initialisation, twenty-five trials and finally, the exit questionnaire. This analysis does not account for visitors to the site that chose not to participate in the experiment.

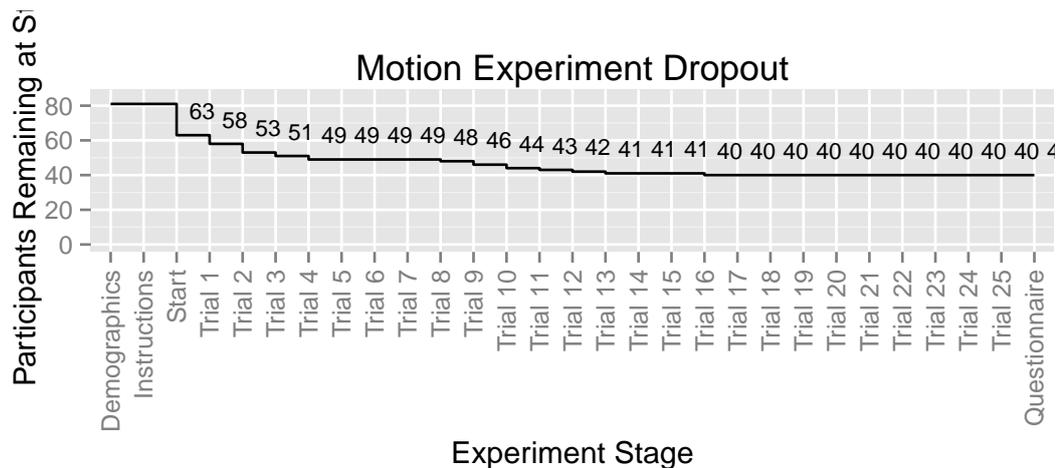


Fig. 3.30: A graph of participant drop out for experiment stage. Values indicate the number of participants remaining at that stage.

In total, over half of all starting participants did not complete the experiment. This level of in-completion is in line with results highlighted by Reips, for research that does not offer a tangible reward.

Of the whole participant pool, twenty-three participants submitted their demographic information by web form but did not begin an interaction with the experiment's apparatus. Of this number, 18 did not proceed beyond the instruction material and five participants started but did not submit a response to a single trial.

The first drop in participants occurs on presentation of training material, suggesting that a large 'wall of text' is a barrier to quick progression. This may have been compounded by a lengthy demographics form. However, A. Frick, Bächtiger, and Reips (1999) found that when the personal information request came at the beginning of the experiment, participants more readily and completely answered, in contrast to when the request came at the end of the experiment. Therefore, there are sound reasons to place the demographics form at the beginning of the experiment.

In regards to the drop out of 10 participants following the instructional material, there are two possible explanations. The first is that having completed the demographics form and completing only a cursory skim-read of the instructional materials, upon seeing the interface participants may have felt overwhelmed or unprepared and

abandoned any further participation. As there was no incentive to remain on task, the abandonment of the experiment is a marked reality. This would also account for those who submitted one trial only; it may have been the case that they had clicked the correct answer - for their first trial response - but still did not understand why it was the correct answer - due to not having read the instructional material properly.

An explanation for the participants that requested the experiment apparatus but did not submit a single trial response is that a technical glitch may have prevented the applet loading on the participant's browser. A significant effort was made to alert participants to the minimum requirements of the experiment software; however, participants were not restricted from proceeding through to the demographics stage if the technical requirements were not met. Participants were asked to check that two images of a green tick were visible before clicking 'begin experiment'. To acknowledge that they saw two green tick images, they had to tick on a form check box before pressing the begin button. If there was no green tick, participants were directed to websites where they could learn how to update and configure their browser appropriately.

A review of demographics data was conducted to observe whether any variables could be used to explain some of the drop out. Such an analysis could in the future be used to motivate additional design constraints for the experiment apparatus. These include instruction time, whether the participant was external or internal to the university and the number of hours spent gaming per week.

The results of this review indicate that mean instruction time - time spent looking at instructions - was remarkably lower and average gaming was slightly lower in the drop out group, suggesting that these participants were impatient and perhaps prefer learning-by-doing and or participants who do not appreciate game like interfaces. Furthermore, participants whom were internal to the university were more likely to complete the experiment. This may be attributed to recruitment methods such as personal invitation and this may have introduced a bias toward this user group, since these participants may have experienced a greater duty to complete the experiment in full.

These observations could be used in similar future research such as by providing an upper bound to the estimated average reading time of instructional materials. A figure for average reading time of an unpaid online participant complements the dimensions of clarity and conciseness of instruction material development.

Tab. 3.16: Demographics and drop-out incidence. ‘*’ denotes value was approximated from the host server log. Average instruction times come from 51% and 65% of participants who completed and did not complete respectively. One outlier from the incomplete group was excluded.

Completed	Instruction Time (Sec.)	External	Internal	Gaming (Hr.)
Yes	179.29*	25	15	5.55
No	95.92*	40	1	4.65

3.6.5 Discussion

Should we consider further the use of motion in encoding paradigms? This web-based experiment set out to investigate the role of motion frequency in a document attribute visualisation task. Eighty-one participants started the experiment, though only half completed the experiment in full. This section will discuss the results of this experiment as well as reflecting on conducting research by way of a web-based data collection methodology.

Overall, prominent trends in the results suggest that as additional attributes were used to encode data, the dependent variables time and error, also increased. Furthermore and importantly, as additional motion features were incorporated into the encoding paradigm, time and errors also increased significantly. However, the results remain inconclusive regarding individual motion features due to the overall small sample size; though an emerging trend reiterated the power of static hue features in encoding paradigms.

As additional data were encoded into glyphs, the time it took to complete the trial task increased also. This finding is evident in Figure 3.12 on page 124, Figure 3.13 on page 124 and Figure 3.14 on page 125. This increase was approximately linear for both preparation time and answer time. For every additional feature, a further 4.48 seconds was necessary to prepare for a target trial and an additional 2.8 seconds was needed to locate the correct answer.

However, this result simply reflects the notion that with a more complex task, longer time to arrive at the answer is required, and potentially, it is more so unlikely that participants can arrive at the correct answer without error.

In addition to the results seen with the number of graphical features, the results were reported according to the number of motion features present in the encoding. Focusing on purely static and purely motion based trials, a summary of average preparation and answer time is extracted and presented in Table 3.17 below.

Tab. 3.17: Preparation time and answer time for unmixed static and dynamic feature trials.

Trial Configuration		Preparation Time	Answer Time
Static	S	4.28	4.21
	SS	6.12	6.99
	SSS	6.77	9.44
Dynamic	D	4.01	4.53
	DD	10.83	8.55
	DDD	13.57	12.85
	DDDD	27.95	19.93

Although further results are required, initial indications suggest that overall, for

every additional motion feature in an encoding, an additional 7.97 seconds is required to prepare for the trial and an additional 4.79 seconds are needed to locate the correct answer. In contrast, for every additional static motion feature, 1.24 seconds is required to prepare for the trial and an additional 2.62 seconds is need to locate the correct answer. There were significant effects observed for an increasing number of motion features in an encoding, but these effects were significant for trials involving 3 or 4 features only and between full motion and one motion feature or less. These rough estimations do not take into consideration variation between graphical feature types since currently, the collected data is not sufficiently large to make reliable judgements at the individual feature level. However, the results do indicate that Angle-Flash and Grow-Angle-Shuffle combinations are particularly bad. The lack of significant findings in the single dimension features means that it is not possible to determine which feature is likely to blame; individually, both size and flash features did not result in the longest trial times in the 1 dimensional group.

However, while a clear trend is evident by the graphs, the high degree of variation across trials as indicated by the wide confidence intervals suggests that differences between motion features may be influencing these observations. This might also be because of variation due to demographic differences such as computer gaming experience.

Broadly, the small sample size collected made it difficult to detect reliably, differences between individual features. However, the strongest trend apparent from these analyses comes from the observation that trials involving the static hue feature have a clear advantage in reducing the amount of time required to answer a trial question. This is to be expected as earlier studies have shown hue to be an extremely efficient encoding device (Yost and North, 2005; Nowell, 1997).

Taken at face value, the results indicate that introducing motion features to the palette of graphical encoding features may not be appropriate. This was evidenced by the graphs of Figure 3.15 on page 128, Figure 3.16 on page 128 and Figure 3.17 on page 129 that showed increasing the number of motion features, resulted in an increase in preparation time, answer time and errors made; however, clear and robust statistical significance could not be established for all encoding dimensionalities. Furthermore, trends were evident in Table 3.6 on page 129 which typically showed that the proportion of trials answered correctly on the first attempt increased as the number of motion features in the encoding decreased; though statistical analyses did not reveal a significant relationship between outcome - first time success, success, fail or skip - and motion feature count.

Table 3.18 on the following page and Table 3.19 on the next page offer rankings for features based on answer time and error rate only - an assessment of rankings based on preparation time could be constructed based on tabulated in the results sections. These rankings highlight the earlier data that place hue at the lower end of answer times - regardless of feature dimensionality.

1 Dimensional	2 Dimensional	3 Dimensional	4 Dimensional
Hue	Angle Hue	Size Angle Hue	5 seconds
Shuffle	Hue Shuffle	Size Hue Shuffle	Size Rotate Hue Shuffle
Rotate	Grow Hue	Grow Rotate Hue	Size Angle Hue Shuffle
Angle	Size Hue	Angle Hue Shuffle	Size Angle Flash Shuffle
Size	Rotate Hue	5 seconds	10 seconds
Saturation	Angle Shuffle	Size Rotate Hue	Grow Angle Hue Shuffle
5 Seconds	5 Seconds	Size Rotate Shuffle	Size Angle Saturation Shuffle
Flash	Grow Angle	Grow Rotate Shuffle	Grow Rotate Hue Shuffle
Grow	Grow Saturation	Size Saturation Shuffle	Size Rotate Flash Shuffle
	Size Rotate	Grow Angle Hue	Size Rotate Saturation Shuffle
	Rotate Saturation	Size Angle Shuffle	15 seconds
	Size Angle	Size Rotate Saturation	Grow Angle Saturation Shuffle
	Grow Rotate	Size Rotate Flash	Grow Rotate Saturation Shuffle
	Saturation Shuffle	Size Angle Saturation	Grow Rotate Flash Shuffle
	Size Shuffle	Grow Rotate Saturation	20 seconds
	Size Saturation	Grow Hue Shuffle	Grow Angle Flash Shuffle
	Rotate Shuffle	10 seconds	
	Flash Shuffle	Grow Saturation Shuffle	
	Angle Saturation	Rotate Saturation Shuffle	
	Grow Flash	Angle Saturation Shuffle	
	Grow Shuffle	Size Angle Flash	
	Rotate Flash	Rotate Hue Shuffle	
	Size Flash	Grow Angle Flash	
	Angle Flash	Grow Angle Saturation	
		Grow Flash Shuffle	
		Rotate Flash Shuffle	
		Grow Rotate Flash	
		Angle Flash Shuffle	
		Grow Angle Shuffle	
		15 seconds	
		Size Flash Shuffle	

Tab. 3.18: Feature ranking based on average answer time.

1 Dimensional	2 Dimensional	3 Dimensional	4 Dimensional
Hue	Grow Hue	Grow Rotate Shuffle	Grow Angle Hue Shuffle
Shuffle	Rotate Hue	Size Hue Shuffle	Size Rotate Hue Shuffle
Flash	Grow Angle	Grow Rotate Hue	Grow Rotate Saturation Shuffle
Angle	Angle Hue	Rotate Hue Shuffle	Size Rotate Flash Shuffle
Grow	Size Hue	Size Rotate Flash	Grow Angle Flash Shuffle
Saturation	Rotate Saturation	Size Angle Hue	Size Angle Hue Shuffle
Size	Angle Shuffle	Size Rotate Saturation	Grow Rotate Hue Shuffle
0.5 Errors	Grow Shuffle	Size Rotate Hue	0.5 Errors
Rotate	Hue Shuffle	Size Rotate Shuffle	Grow Angle Saturation Shuffle
	Flash Shuffle	Angle Hue Shuffle	Size Angle Flash Shuffle
	Rotate Flash	Rotate Flash Shuffle	Size Angle Saturation Shuffle
	Size Shuffle	Grow Flash Shuffle	Size Rotate Saturation Shuffle
	Rotate Shuffle	Grow Angle Hue	1.0 Errors
	Size Flash	Grow Rotate Saturation	Grow Rotate Flash Shuffle
	Angle Flash	Size Saturation Shuffle	
	Grow Saturation	Grow Saturation Shuffle	
	Size Saturation	Angle Saturation Shuffle	
	Grow Flash	Angle Flash Shuffle	
	0.5 Errors	Size Angle Shuffle	
	Saturation Shuffle	Size Flash Shuffle	
	Size Rotate	Size Angle Saturation	
	Size Angle	Grow Angle Flash	
	Angle Saturation	0.5 Errors	
	Grow Rotate	Size Angle Flash	
		Grow Angle Saturation	
		Grow Rotate Flash	
		Grow Hue Shuffle	
		Rotate Saturation Shuffle	
		Grow Angle Shuffle	

Tab. 3.19: Feature ranking based on average error rate.

While earlier research has indicated the proficiency of motion cues to attract and guide the attention of the human perceptual system and the earlier human factors studies demonstrating the use of motion frequency to encode data (e.g. Tolin and Ryen, 1986; Laxar and Luria, 1990; vanOrden, Divita, and Shim, 1993; J. Lee, 2008), these results suggest that interpreting motion frequency may be more challenging than was argued in the introduction to this chapter. Yet, there are a few factors at play that prevent more robust conclusions being drawn from these data; the first relates to the small sample size, the second relates to the actual task, and the third factor relates to the variation introduced by participants.

While 81 participants initiated the experiment process, over half of all respondents did not complete the experiment. With the full contingent of participant data, a more robust signal may have been viable. Approximately 150 participants would be necessary for a future experiment given an exact replication.

However, such a large number of participants may not be necessary if each participant completed double or triple the number of experiment trials per session, though, clearly, each participant has a maximum level of effort they are willing to expense, given that there was little tangible reward on offer.

General task influences and general coding of the stimuli may also have an effect on the results. The qualitative responses, although lacking clear strength in the direction of agreement, indicated that participants found the motion annoying and perceptually uncomfortable. In relation to the annoyance factor, these qualitative results do not go far enough to explore which of the graphical features participants were most annoyed by. The experimental design under-emphasised the need for more specific qualitative response to the graphical feature types and in the least should have provided a free-text open response facility for participants to express their comments relating to suggested aspects of the interface. Furthermore, a future experiment would seek to pilot graphical features for annoyance or to poll participants at experiment time given the lack of qualitative results in this experiment's data set, which is attributed to the earlier underestimation of the utility of such responses within such an abstract context.

Based on informal observations, a potential area of improvement along the annoyance dimensions could be to reduce the amplitude of zooming and shuffle motions and to explore the role of vibration, which may result in a reduction of perceptual dominance. Furthermore, and ultimately, a future experiment should utilise a pilot experiment or ask participants to rate an ensemble of motion types on several qualitative dimensions, prior to the commencement of an experiment and to utilise the least annoying and distracting features arising out of the test routine.

In relation to the experience of motion, ultimately, it is difficult to ensure a consistent or homogeneous computing platform for all web-based participants. In future, this could be controlled for by way of richer data collection using a Java Applet and

JavaScript to interrogate the participant's computer system properties. However, in doing so, additional challenges may be introduced to the analysis due to the number of different ways system configuration can vary. In addition, rough estimates could also be made by querying participants in the exit questionnaire regarding the number of programs opened in the background or the number of computer-based tasks the participant was conducting in parallel prior to - or during - the experiment session.

The subjective data collected in this experiment were regrettably vague and there are several reasons why this is the case.

First, the exploratory design of the experiment - which attempted to deal with the combinatorial explosion problem by maximisation of trial condition diversity - meant that a participant did not have extended exposure to any one particular graphical feature or feature combination. Remembering what they did like and did not like over the course of the experiment would be difficult. Furthermore, pair-wise comparisons for each trial combination on a range on subjective dimensions would be prohibitively expensive as would require ranking of combinations on several subjective continua.

Secondly, there is the question of which subjective dimensions. Bartram (2001) makes use of two subjective dimensions: irritation and distraction - these subjective ratings turned out very similar; these are two dimensions in which ratings may be made without a context i.e. the feature may be displayed and evaluated. By context, is meant a situation in which the graphical feature under subjective examination is encoded with data and the participant judges the appropriateness of that feature for the decoding task. In contrast, dimensions such as perceived satisfaction, accuracy, confidence and effort require a context for a participant to judge how accurate they believe they are in extracting information from a particular graphical feature. It would be difficult for the participant to retain such subjective records over the course of 25 heterogeneous trials. Furthermore, it may be a disruptive and drawn out process to request that participants judge each trial on a few subjective dimensions before proceeding to the next trial.

Thirdly, concern regarding the altruistic affordance of the unpaid participant weighed heavily in the design of subjective responses, particularly early on during data collection when repeat participation was anticipated. Early on, design efforts were devoted to ensuring that the task was well understood and whether the perception of motion was smooth, as it was not possible to predict on any one particular computing platform how the experiment apparatus would run or how well participants would absorb the instructional material. Accordingly, with the above factors absorbing most of the design effort, comparatively less effort was available to dedicate to the development of sound subjective data collection, as it was feared that overburdening the unpaid participant would be met with a low experiment completion rate.

In a future experiment, the experiment trial design could be extended into three stages: preparation, answer and subjective assessment. Subjective dimensions might

include accuracy, confidence, effort, irritation and/or distraction and such judgements could be made immediately following an individual trial. Alternatively, a repeated-measures design may increase experience with and exposure to a small set of trial types and subjective assessment could follow batches of say ten homogeneous trials.

Variation across participants may also be a factor contributing to the large error margins expressed in the reported results. This experiment attempted to investigate general human factors at play rather than to evaluate the usability of a fully functional interface. However, in attempting to maintain the application domain i.e. an information retrieval feel, in the design of the experiment apparatus, the design diverted from a more traditional data collection design. One of the leading differences was the mouse-based response mechanism as opposed to the keyboard press response as is more typical with visual search experiments in psychology.

The discussion will now address further this potential source of variation introduced by the response mechanism but, ultimately this will push the discussion toward a broader consideration of the merits of online web experimentation.

The experiment's design attempted to combine aspects of a visual search experiment more commonly associated with perceptual psychology, with that of an information visualisation experiment. In the former, investigations into perception tend to be more clinical and produce results that motivate hypotheses and theories regarding the human perceptual system. In the interests of robustness, such trials are often repeated several hundred times in order to ascertain reliable and controlled measures across treatment groups. These pure results motivate how under perfect conditions, the human perceptual system may be operating. This experiment should not be considered an investigation into the human perceptual system; but rather the results should motivate a possible case for or against the use of motion for the encoding of data.

In this experiment, participants had to first encode their target to short-term memory and upon finding it, utilise their mouse to indicate the correct answer. In contrast, visual search experiments utilise visual targets presented at the centre of screen prior to the commencement of a trial - or batch of trials - thus making it easy for the searcher to encode the target as opposed to integrating several disjointed visual features into a single target as was done for the present experiment. However, in practise, use of graphs, maps or information visualisation tools necessitates an orienting with the encoding paradigms applicable to the relevant visualisation. This step was simulated by first asking users to figure out, construct and store the target chunk in short-term memory based on a textual description. Accordingly, this simulates the realistic situation where the user's information need triggers the task statement, which triggers the user's activity of decoding from the legend.

The mouse-based response methodology may have introduced an influence to answer time, though it should not have introduced an effect on the preparation time or

accuracy. The demographics data suggested that participants have sufficient weekly computer use to suggest that mouse use skill should be good but the high prevalence of computer gamers may have introduced an extra level of variability. To counter this influence, a keyboard response could have been used though at the expense of realism.

However, despite these findings, why was a mouse-based response favoured. A mouse-based response replicates the typical interaction mode undertaken by data visualisation software; but doing so in a human factors experiment comes at the expense of fidelity and the potentially confounding variable of the participant's motor skills and capacity to operate a mouse. Age related differences have been found for mouse usage, for instance by Chaparro et al. (1999); in contrast simple response times, although also diminishing with age (Der and Deary, 2006), have been recorded with a comparatively effortless single button press. Despite these findings, mouse interaction is a common way of interacting with visualisation tool interfaces and accordingly, a mouse based response was favoured due to the added level of realism.

The usability of motion features depends both on the ability to locate the target and to pick the target interactively for the wider information task. In information visualisation generally, the identification of the target chunk is primarily important since such presentations facilitate visual interaction. In contrast, for tools employing visualisation for searching rather than hypothesising, identification and interaction are equally important since search will more readily involve further interaction with items meeting set criteria. In the case of the hypothesis testing, the end-result is more readily the visual perceptions - e.g. trend or proportion estimation - and intuitions and insights acquired from the refined data set; whereas for information retrieval, opening documents and judging relevance via reading and analysis is more typical. Moreover, as visualisation tools roll out onto touch activated displays, the case of clicking by pointing will also become relevant to experiment design. Accordingly, future experiment design should consider the role of touch interaction as a response mechanism.

Furthermore, mouse based gestures are important for research conducted online. Criticism is directed at online experiments using crowd-sourcing websites due to the prevalence of spurious users 'gaming the system'. A True/False response design is prone to gaming such that it is easy for a remote user to press either of two keyboard keys to skip through the experiment trial and collect their payment as quickly as possible. Locating the correct answer to proceed using a mouse click provides an additional level of confidence that participants are not cheating, because they must indicate that the target is present and furthermore, specify the actual location of the target as indicated by the mouse cursor at the point of clicking. If the participant is constrained to finding the target before progressing, then the time spent clicking randomly for a target amongst a set of distractors, particularly with large set size, should exceed that of locating the correct answer.

Having discussed sources of variation introduced by the apparatus, variation intro-

duced as a result of the participant demographic will now be considered. Experiment participants originated from a broad demographic; a broad demographic in terms of cultural and educational background is fairer in that general computer and search engine users originate from a diverse background. Furthermore, the setting in which participants partake in research is relevant (Reilly and Inkpen, 2007), as too is task motivation such as incentive (A. Frick, Bächtiger, and Reips, 1999) or voluntarism (Reips, 2002) and consequential factors such as engagement and perseverance.

In relation to motivation, despite the significant effort devoted to implementation of a scoring system, and despite only 46% of completing participants looking at their personalised performance report, there was little indication that this spurred competitive behaviour among participants and there was no clear indication - based on user prescribed random words - that participants returned for subsequent participation.

Whilst the experiment software did not analyse the referring URL, it is presumed that international participation was the result of listing with the Hanover College online experiment page <http://psych.hanover.edu/research/exponnet.html>. In future experiments, it would be of use to sign up to a web analytics service that annotates this type of information with little overhead to the researcher. Furthermore, this would make it easier and more reliable to record information such as site behaviour and timing data. As was earlier noted, mean-time spent reviewing instructional material was estimated from the server logs. This process is unreliable as it is difficult to match participants to entries in the server logs with entries in the data set. While more reliability could be achieved through better data collection, an analytics service could provide an additional richness to the data set to expedite unexpected calculations during analyses. For instance and interestingly, of the completing participants approximately 50% looked at the ethics information before starting the experiment while approximately 23% of participants who did not complete the experiment looked at the ethics information. This information was not a requirement for the analysis but it provides additional insights that can influence future experiment design. The above factors may not be noticed during the development of the main experiment apparatus software, and many peripheral issues only arise after deeper analysis of the results - by which time it is too late.

A greater understanding of the origin of participants reveals greater insight into the diversity of the participant group. Demographic diversity does come with its drawbacks including the provision of adequate training material. All participants indicated a lengthy history of spoken English language yet there was an indication that the delivery of training could have been better. Participants did not conclusively report that the task was easy to understand. It may have been the case that the task was difficult yet still easy to understand.

The marked learning effect observed and consequently accounted for during the analysis underscores the case that long-winded, 'great walls' of text-based instruction

are a poor choice for pre-test instructional and training material. In a laboratory-based experiment, a research assistant would likely provide an explanation to the participant while pointing out the relevant interface components. In contrast, when online, the remote participant must make the connection between the verbal instruction and the interface component themselves. If the vocabulary of the instruction is foreign to the participant, then this may make the task even more cognitively demanding; the remote participant requires the research assistant to be there.

The advent of video streaming services such as <http://www.youtube.com> and faster broadband internet have enabled the potential for video based instructional material as a medium to convey simulated demonstrations co-annotated with voice over instruction. Yet, while this medium still lacks the degree of interactivity an in-lab participant may engage in - e.g. asking clarification questions - it does provide reinforcement of an idea - i.e. the task description - by verbal and active demonstration.

Discussion has canvassed several issues impacting on this experiment. As a result, there are two alternative designs for a future experiment which would seek to improve the current research methodology. The first is an improvement on the current web-delivered experiment. The second alternative is to port the web-based experiment to a social network platform or mobile application development platform. The modifications and their expected impacts are now discussed and where possible the benefit of porting to other platforms is highlighted.

One way to combat drop out due to technical reasons before the commencement of trials is to utilise strict pre-test technical requirements - rather than a voluntary confirmation - at the welcome page. An extension to this approach is to encapsulate each experiment stage within the plug-in applet so that participants cannot start the experiment processes without first getting the technology to work. Thus, rather than a web page for each of welcome, demographics, training, trials and exit questionnaire the applet could display web pages or user interface based forms instead of the web pages. The experiment site would then consist of one page and embedded within that site a single applet. A failure to load the experiment applet could be detected by the browser and the participant forwarded to an appropriate help page.

In this case, a web page for each aspect was favoured due to the extra development overheads for creating an all inclusive applet. However, the all inclusive applet approach is more consistent with the development of mobile applications and could potentially port straight over to a mobile-based application platform or integrate with a social networking platform. Having access to birthday, gender and location information may render the use of a demographics questionnaire practically unnecessary. Moreover, return participation is likely to be higher for such an application as the re-entry requirements are potentially lower and participants do not have to re-register if the application is already installed. In the current web-based set up, this was not possible; thus a way to cater for re-participation should have been easier. One way to

have achieved this could have been to upload a cookie to each participant's browser cache.

A browser cookie could be stored on the participant's computer to indicate that the user has accessed the experiment site. The cookie would contain a unique identifier that could be used to bypass demographics and training material for subsequent visits. The cookie could also be used to automate the retrieval of the performance report rather than needing a user name to log-in to the performance reporting function. There are alternatives to this including calculating a browser signature based on the currently installed fonts, plug-ins, browser version, time zone and cookie state through HTTP Request. Eckersley (2010) experimented with this idea and found that 94% of their large sample resulted in a unique browser signature. This could be another way to combat the scenario where cookies are not enabled or expunged at the completion of browsing sessions. While the browser signature does not guarantee perfect reliability as in the case of cookies, it does mitigate the need for participants needing to remember unique words as their user name. In the case where one user is mistaken for another a fall back is to provide prominent links to 'First Time User' information which could then poll user's for demographics information and provide instructional material.

3.7 Summary

This chapter proposed to investigate the role motion frequency could play in encoding paradigms for metadata visualisation. Previously, the application of investigation has not favoured encoding data in frequency. Human factors experiments suggested a human capacity for frequency interpretation and this provided a case to investigate motion frequency coding further. The results of this experiment indicated that as further motion features are used to encode data, time and error rates increase. Although significant findings could not provide the statistical robustness in support of these trends, these initial findings appear to indicate that it may not be wise to effectively double the number of features available to data visualisation designers by incorporating motion features into the data encoding palette.

This experiment was conducted over the Internet. Consequently, a number of methodological issues faced by researchers performing similar research over the Internet were discussed. Web-based methodologies introduce a number of benefits to research but they also present practical challenges that require additional thinking in comparison to the same experiments completed in closed-door laboratories. Many of these challenges can be overcome but this is an ongoing and learning experience.

4. ON THE ROLE OF NATURALNESS IN ATTRIBUTE VISUALISATION

4.1 *Introduction*

This chapter reports on an experiment that aimed to observe the effect of natural data encoding rules on task performance. There are a multitude of suggested ways for encoding data; however, this chapter will emphasise and investigate a set of encoding rules that exhibit the characteristic of naturalness i.e. the most natural way to encode data. Naturalness characterises an encoding rule that maps a data instance to a graphical appearance or geometric attribute of a glyph, which is obvious, innate, analogous, consistent or which preserves or reinforces the message of the datum instance.

Frequently, Metadata attributes are encoded into appearance and geometric attributes of glyphs. Metadata is an important aspect of search and searchers do make use of information scent in the form of document metadata (Balatsoukas and Ruthven, 2010). This chapter will devise and motivate a natural metadata encoding paradigm that is easier for a searcher to learn and use. Ultimately, this research will improve our search tools by making them easier to learn and use, by allowing faster and more confident decision-making.

Natural encoding should result in lower learning overheads, while promoting interaction in a relatively unconscious fashion; namely, interaction without explicit and significant conscious thought. By analogy, consider novice versus expert problem solving. An explicit, conscious and stepwise interaction - through a problem which might entail interaction with a user interface - by typically a novice rather than an expert user, takes comparatively longer time to complete, is more burdensome on working memory, and decision-making is made with lower confidence and accuracy (Hardin, 2002).

A prime example in support of this is that of a person learning to drive a car for the first time versus the same person in five years time. The seasoned driver operates the car naturally, with very little of the explicit, inner-vocalised, conscious mental-check listing relied upon by the learner driver. The same analogy applies to interaction with computer interfaces; by way of experience, the expert has the experience to make confident interpretations of the data presented in the interface and does not have to rely on moment-to-moment validation of the encoded datum, through use of a decoding legend.

It is not argued that devising natural data encoding paradigms will abolish the

learning curve altogether, nor turn novice users into experts instantaneously. Rather, a natural data encoding as opposed to an unnatural encoding, should at least flatten the learning curve by allowing users to draw upon general, existing, and generic prior experiences and expectations. As a result, overall task performance should improve markedly and more rapidly than in comparison to a novice user utilising an unnaturally presented set of data.

Designers of new, visualisation-based information tools should not anticipate that searchers will endure a confusing and unnatural way of presenting data, particularly given the plethora of available alternatives that while maybe warranting additional effort, get the job done without confusion. However, search tool designers have only the data encoding recommendations from several fringe disciplines to draw upon. While conceptually similar, the guidelines that statistical data visualisation designers or pedagogical diagram designers utilise are not an exact fit for search tool design. For instance, in statistical data visualisation, encoding rules are based on the type of data under visualisation and the affordances of graphical features.

Empirical work has sought to validate data encoding recommendations (Mackinlay, 1986; Nowell, 1997); although, as Nowell concludes, the type of task the user is engaging the visualisation for, should also guide the selection of encoding rules. From another perspective, in relation to the pictorial representation literature, proponents indicate that good representations are those that ‘...are similar or analogous to the represented; and which preserve the properties of the represented in an explicit way’ (Narayanan and Hübscher, 1997).

This chapter will argue that further investigation is required to ascertain if naturalness in encoding meets with superior task performance, relative to encoding paradigms that are based solely on earlier guidelines that focus on data type alone. Discussion will focus first on the analogous concept as it appears in pictorial representation. It then highlights this as applied to particular forms of information visualisation, particularly in graph visualisation and geographic visualisation. Following this, discussion will turn to the role of metaphor as applied to computer-based icon design and the impact that social norms and conventions have on our interpretations of these representations. After which, discussion will focus on the recommended encoding guidelines originating in analytical data visualisation contexts; following this, an extrapolation of the recommendations based on human-perception research will be presented. Finally, discussion will draw on the idea of interference tasks to illustrate the situation where unnatural data encodings interrupt learning and processing of encoded data in visualisations. This discussion will then lead to a new proposal for encoding data for a metadata visualisation. This proposal will be compared and contrasted to the earlier review of guidelines.

Following this discussion, the results of an experiment are presented; this experiment provided an empirical examination of natural encoding. It was anticipated that

objective task performance would improve because of natural encoding of data in the interface; however, peculiarly, it was expected that participants who perform better objectively, should not necessarily be able to more readily and accurately self-report their learning of the interface. Instead, it would not have been unexpected to observe superior self-reported learning outcomes from those who struggled with the interface under unnatural encoding conditions. Users under unnatural encoding conditions are comparatively burdened by the interface, thereby taking longer time and effort to understand how the interface works and therefore spending greater time consciously adjusting their prior expectations to fit in with the way the interface works. Conversely, participants under the natural condition were expected to complete tasks with ease and success, but without conscious awareness of how the interface worked. Moreover, participants under the unnatural condition were expected to do objectively worse, but more readily recall a freshly modified mental model of the interface when successfully reporting how the interface worked and furthermore, how bad it was.

4.2 *Encoding Paradigms in Attribute Visualisation*

4.2.1 *Diagrams and Pictorial Representation*

A picture is worth 10000 words or so goes the adage. However, it is widely held that if a designer wishes to reap the benefits of pictorial representation, a picture or diagram should capture the essence of or ‘share a similarity of structure with’ the system it is representing (Gurr, 1998). Intuitively, pictorial representations can instantaneously convey a concept or idea utilising graphical means, which may otherwise take a large number of words to describe with adequate contextualisation. However, as Larkin and H. Simon (1987) argue, diagrams are only powerful if the observer knows how to use them effectively and as Levie and Lentz (1982) illustrate, pictorial representation does not guarantee that the full detail in an equivalent text passage will necessarily be made available through the illustration.

Progress to formalize pictorial representation and its benefit to learning and understanding has persisted for quite some time (Larkin and H. Simon, 1987; Narayanan and Hübscher, 1997; Gurr, 1998; Morrison, Tversky, and Bétrancourt, 2000; Tversky, Morrison, and Bétrancourt, 2002); with the advent of modern computing facilities, the role of animation in pictorial representation, offers a modern aspect to the study of such designs (Lowe, 2003). Hegarty (2011) collates a number of the assumed benefits that pictorial representations afford including the externalisation of information, the organization of information and the ability to offload cognitive processes to perceptual processes. By externalising the entity or system into short-term memory, for the purposes of: hypothesis testing, understanding, and decomposing a system into its constituents, the observer is free to spend additional cognitive resources on learning or gaining new insight.

Gurr (1998) argues that despite the differences between a great number of cognitive models that hypothesise why good pictorial representations are useful, all models reference the observation that good pictorial representations are: analogies of, similar to, morphological to, homo-morphological to, or iso-morphological to that which they represent. Thus, in the absence of a single robust model of pictorial representation, diagrams should in the least, be designed in such a way as to reflect the system under consideration.

Diagrams promote mental simulation; which is something that might typically be difficult to do when information is in textual form, thus, forcing the user to first build a mental visualisation, and then to simulate and hypothesise about the system all in short-term memory (Narayanan and Hübscher, 1997). Contemporary computer based diagrams have the benefit of dynamic and animated components to make those simulations more concrete. Intuitively, animation should improve comprehensibility when the systems under consideration are naturally dynamic. A dynamically changing diagram should exhibit a higher degree of isomorphism. However, as Lowe (2003) suggests, by adding additional cues such as animation in an effort to benefit comprehensibility, particularly for dynamic systems such as natural or mechanical systems, perceptual domination is possible, therefore leading to unnoticed information in the diagram and as a consequence, a detriment to learning.

It follows then that encoding paradigms for visualisation should ensure that the graphical attributes share a similarity to the data they encode; such an encoding would be considered more natural and perhaps less abstract. A notion of naturalness already exists in the pictorial representation literature (e.g. Hegarty, 2011; Tversky, 2011) though it appears under a variety of guises including one by Kosslyn, 2007, who proposes the principle of compatibility.

The principle of compatibility states that a message is easiest to understand if its form is compatible with its meaning. There are four aspects to the principle of compatibility. The first aspect is appearance-meaning correspondence; for instance, matching the magnitude of something to the size of an icon. Therefore, greater quantities should be represented by greater perceptible quantities such as assigning a bigger number to a bigger graphical quantity like size or colour brightness. The second aspect is that interpretation should correspond to cultural conventions such as what is good appears at the top of an importance-ranked list. In addition, the third aspect relates to compatibility with perception; for instance, is the form and shape of graphical attributes and their meaning consistent with perception, or does explicit effort need to be expended to overcome intuitive meaning. These aspects of Kosslyn's compatibility will feature prominently in this experiment as applied to a search visualisation context.

This section has discussed representation as it applies to the design of pictorial representations for pedagogical purposes, which offer a different understanding to that conveyed by pictorial representations of abstract data such as search results. In the

next section, a discussion of the syntax of node-edge diagrams and maps from Ware (2004) provides as a partial overlap between pictorial representation and information visualisation.

4.2.2 *Semantics of Charts and Graphics*

Whereas pictorial representations may be designed to promote learning and understanding through spatial reasoning, Ware (2004) writes of a loose visual syntax present in node-link diagrams and maps. The visual grammar of node-link diagrams is roughly separable into attributes of entities and relationships and relationship attributes between entities. Conversely, a visual grammar of maps is distinguished by entities, points, lines and regions on the map.

With regard to node-link diagrams, Ware suggests nodes, entities and objects - represented as closed contours - use colour and shape to denote entity type and size to denote magnitude. Additionally, relationships between nodes are denoted by grouping, including node compositions and attachments, edges, closed contours enclosed by other closed contours including partitioning lines typically found in Tree Map visualisations, and or spatial orderings. Furthermore, relationship types can be denoted by manipulating the line thickness or stroke pattern of edges.

In contrast, entities on maps are denoted by points or lines to indicate features like roads and rivers. Geographical regions are denoted by closed contours, which may be coloured or textured to indicate a type of geographical region. The combination of region and feature leads to a number of relationships such as containment and termination. For instance, containment indicates the depiction of geographical regions encapsulating a town while termination indicates the beginning or end of a river.

Ware suggests that the above graph and map features have a perceptual basis rather than conventional basis for interpretation. This perceptual basis interpretation is consistent with the isomorphic view of pictorial representation: that which closely resembles the data or shares a similarity of structure will result in the best representation. Furthermore, it is natural to think of points on maps as small towns and lines as rivers and roads because if we looked at the physical landscape from a sufficiently high vantage point, our view of the geographical area below would resemble the scale and type of features present on a map.

The map metaphor as applied to information visualisation, places greater emphasis on the point and region features rather than line features. Accordingly, the remainder of this chapter will focus on attributes of points rather than lines. In contrast to Ware's perceptual rather than conventional basis for interpretation, Hofmann (2009) recounts the cultural influences in the interpretation of icons marking significant features on web-based map services. The next section discusses the role of social norms and conventions.

4.2.3 *Encoding by Social Norms and Conventions*

Whereas basic appearance and geometric attributes can be used to represent quantitative data, pictorial representations can be used to depict qualitative aspects of data. These representations convey meaning that can be misinterpreted if the wrong symbol is utilised. Often a choice of symbol is governed by convention and social norms.

Social norms and customs influence the way we interpret pictorial representations and symbols (Pappachan and Ziefle, 2008) such that their influence cannot be ignored in this discussion. McDougall and Curry (2004) argue that icon interpretation takes place within a broad and complex context, encompassing complexity, workload, the nature of the display or task, time of day, and user skills, preferences and knowledge. This discussion will seek to address some of these issues in this section and the following section. First, a brief outline is provided on the role of background knowledge, familiarity and experience in icon interpretation.

Experience is governed by age. Age differences are increasingly important as suggested by Hofmann (2008). In his informal observations, Hoffman suggests that children of a younger generation cannot identify the symbols for television and telephones that are familiar to an older generation. Instead, having grown up in a modern age in which televisions no longer have their aerials on top, are flat panels and not boxes, and furthermore in which telephones fit into shirt pockets and are not affixed to a wall, children more readily identify with their contemporary representations.

Experience is governed by culture. A prime example of this is the different meanings cultures assign to colour. For example, western cultures associate red and orange with stop and warning. However, in eastern cultures, the colour red is associated with one's fortune, yellow is considered a sacred colour and orange is considered a colour associated with happiness.

If the user is unfamiliar with antiquated social conventions because of their age and because a more recent convention has replaced an old one, then users are liable to make erroneous interpretations of those unfamiliar representations, or they may go unnoticed altogether - to the detriment of the user's interaction. Early versions of Touch Graph <http://www.touchgraph.com> utilise a pictorial approach to depict file page source i.e. using the 'favicon', however, these pictorial cues are only relevant if the searcher recognises the source site branding. Experience dictates how many of the icons have been seen before, and therefore how useful they will be for search. In addition, many companies, technology or otherwise, seemingly make a habit of re-branding themselves every few years in order to revitalise or redirect their brands. Accordingly, the colours and symbols we associate with companies, changes with time.

When encoding data, particularly categorical data, the benefit to interpretation has to be seriously considered. The next section will outline the role of metaphor in encoding.

4.2.4 *Encoding by Metaphor*

Metaphors are linguistic devices used to enrich linguistic prose; computing has a substantial history of incorporating metaphor in user interface design to enrich usability and user experience. While metaphor use spans many contexts (e.g. Benford et al., 1999), the focus of this section will be on metaphor use in icon design.

Metaphor devices are incorporated into user interface components to improve recall, and to convey or demonstrate the functionality of buttons (Baecker, Small, and Mander, 1991) and other interactive facilities of an interface. Empirical data reveal the influence of task abstraction and task complexity (Bodner and MacKenzie, 1997), icon detail and complexity (Byrne, 1993), and the use of animation (Schwalm, Shaviv, and Goldschmidt, 2000; C. Harrison et al., 2011) - indicating a broad range of issues to consider for design.

C. Harrison et al. (2011) differentiate between graphical icons in which an icon pictorially resembles a physical icon, and ideographic in which the icon is a visual representation of a concept. They offer the idea of ‘kineticons’, which apply kinetic behaviours by dynamic transforms to static icons to establish a rich set of meaning through dynamic metaphors. Kineticons offer the pictorial representation of the icon plus an additional dimension of state information such as ‘the task is starting’, ‘the task is progressing’, ‘the task needs attention’, or ‘the task cannot complete’. In addition, kineticons can convey a state change or state replacement, e.g. the initialisation or completion of the task, and indicate affordances of the icon e.g. ‘this icon is movable’.

Setlur et al. (2005) propose a process to extract semantic information from file names on a computer hard drive to drive the selection of clip art pictographs for the basis of icon composition; they deem the resulting icons as semantic icons. Empirical data of Setlur et al. suggest that users more readily and easily identify and recall from short-term memory, a majority of semantic icons in comparison to traditional system icons. The benefit of semantic icons is that document type is preserved since the clip art pictures are superimposed over the existing icons to reflect the semantic content of the document.

On the other hand, Byrne (1993) suggests that high icon complexity results in a decline in efficiency for visual search, where as a simple icon design results in a more efficient search and the effect is markedly evident with increasing numbers of distractors. Furthermore, experience does not counteract the decline in efficiency for complex icons. Thus, while a pictographic representation is favoured over no pictographic cue, there is a simplicity caveat that a designer must meet.

It is conceivable that information retrieval tools could utilise pictorial, kinetic, or symbolic representations to convey information and identity to the searcher. Branding by way of the favicon has appeared in a number of systems e.g. <http://www.touchgraph.com>, as well as overly simplistic pictorial representations of documents or

collections of documents - e.g. Groker and Kartoo (see Foenix-Riou, 2006; Koshman, 2005). An extension to the semanticons of Setlur et al. (2005) could utilise title or keywords in documents to build semantic representations for icons presented on metadata visualisation tools.

Predictably, as search becomes increasingly interactive and respondent of the user's moment-to-moment interactions, kinetic behaviours like those explored by C. Harrison et al. (2011) could have concrete applications. Kinetic behaviours could provide visual feedback regarding the outcomes of user feedback or interactions such as searcher-initiated filters and refinements.

This discussion has focused on the multitude of design factors influencing the depiction of data and concepts through pictographs and symbols. The main theme has been that better pictorial representations will be those that appear morphologically similar to that which they are representing and those that meet the experiences and expectations of the users. Typically, in information visualisation however, data is represented by simple appearance or geometric attributes of glyphs rather than pictorial representations. The discussion now turns to encoding guidelines suggested for the representation of data through these more rudimentary graphical features.

4.2.5 *Encoding by Data Type*

A task agnostic approach to choosing encoding rules is to look only at the type of data under visualisation i.e. whether the data are ordinal, nominal or continuous. Mackinlay (1986) suggests a ranking of graphical features in terms of suitability for quantitative data types. This ranking is provided in the fourth column of Table 4.2 below. Position is consistently favoured as the best graphical feature across all data types, while the suitability of the many dimensions of colour and area - shape and size - shift in rank depending on the data type. Furthermore and pragmatically, a ranking of graphical features naturally leads to Mackinlay's principle of importance ordering: 'encode more important information more effectively'.

Nowell (1997) sought to validate empirically a subset of Mackinlay's ranking of graphical features. The features chosen were hue, shape and size for both quantitative and qualitative data types in a metadata interface experiment. Her findings were inconsistent with that of Mackinlay and two other ranking proponents Christ (1975) and Cleveland and McGill (1984). However, Nowell resolves the discrepancy by suggesting it is not exclusively the data type that is important, but the exact nature of the task. Her empirical data reflect performance on counting and identification tasks, which are relevant to the current investigation and more generally in search engine tools.

In contrast, Christ (1975) and Cleveland and McGill (1984) base their findings on graphical perception tasks such as the extraction of single values through perceptual comparisons and calculations, like proportion estimations tasks that are traditionally

associated with data visualisation for statistics. In counting tasks, the user is searching through a set of alternatives and counting the number of times a particular target occurs based on a set of visual criteria. In a document search task, instead of counting, the searcher opens each document that meets a set of criteria, generally in order of discovery. The document search is terminated - like the counting task - once each alternative is accounted for, or when the required information is found. The perceptual processes driving such tasks are the subject of much perceptual psychology research and there exists a movement to utilise the insight gathered from such research to devise perceptually efficient data encoding (e.g. Healey, Amant, and Elhaddad, 1999).

4.2.6 *Encoding by Perceptual-Cognitive Psychology*

The previous section focused on meaning and making a connection between the graphical attribute and the data variable. However, this all assumes the observer can perceive the target in the first place.

Clearly, if the encoding rules are easily stored in short-term memory, the search will be improved, since the overhead of saccading between the visualisation and the encoding legend should be lower. However, if the designer chooses graphical attributes to encode data that are not conducive to efficient search, then the task will largely suffer regardless of whether the searcher can decode it correctly. Thus, on the one hand, we need guidelines regarding cognitive efficiency based on the searcher's understanding and intuition of meaning - thereby making it easier to learn and recall encoding rules - while on the other hand, we need guidelines based on how fast our perceptual system can locate and recognise the raw graphical features.

In the previous section, Mackinlay (1986) wanted to produce a ranking of graphical features by data type such that these could provide the basis for automatic graphical presentation construction. Similarly, Healey, Amant, and Elhaddad (1999) propose to utilise empirical data from perceptual experiments to devise a set of guidelines in order to drive automatic construction of data visualisations that are perceptually optimal, thereby fostering rapid and accurate visualisation tasks. While the goal of both sets of authors appear the same, the emphasis Healey, Amant, and Elhaddad place on the role of both empirical research results and contemporary artificial intelligence, sets the two pursuits apart. They propose that a system should be able to interpret the analytical intentions of the user by way of an interview, before taking that information and generating an appropriate visualisation with data encoded efficiently depending on the interpreted goals.

Empirical perceptual research of the variety that Healey, Amant, and Elhaddad pursue is typically that obtained through a visual search task. Visual search tasks were outlined in Chapter 2 and specifically during the discussion on the perceptual framework adopted for this thesis. Another potential source of encoding guidelines

derives from the study of human visual perception by way of visual search tasks. Much of this research has concentrated on one or two basic features and their conjunctions; so at this stage, we have only small insights into using perceptual research as the basis for encoding rule sets. However, there have been perceptual investigations into more than a conjunction of two features, namely a search for triple conjunctions (Humphrey and Kramer, 1997; Williams and Reingold, 2001). Interestingly, as a third feature is incorporated into the target the search efficiency is at times superior, depending on the combination of features; interestingly, search for a conjunction of three features can be faster than a conjunction of two similar features.

It was noted earlier, that Wolfe (2007) has claimed that all theories of visual search are wrong and that future theories of visual search should account for a set of eight phenomena, one of which related to the role of categorically-defined targets and their influence on efficient visual search. For the present experiment, the eighth characteristic of Wolfe, the influence of category, is of most importance. Wolfe (1998) and Wolfe and Bose in Wolfe (1998) suggest that there exists a limited vocabulary that may influence visual search, by control of attention in a top-down manner.

Bottom-up attention control is characterised by stimuli driving the control of attention as evidenced by efficient search on tasks in which the target is a priori unknown - rather the task is to spot the odd one out or to find the singleton among distractors. In contrast, top-down attention control is user driven and based on directing attention to targets deemed as important based on a trial criteria statement. Top down processes are evidenced by experiments in which the user is looking for a specific coloured target amongst a multicoloured sea of objects. For a metadata search, top-down guidance of attention is important given search for a known target criteria are well defined i.e. find all documents of type A. In contrast, bottom-up processing is important for cueing the observer to an outlier or trend in the data that they had not suspected to see or that they were not primed to look for.

Wolfe, Friedman-Hill, et al. (1992) found that search for a basic feature was better when a top-down categorically-defined search task was used e.g. 'find steepest tilted target' or 'find the biggest target', rather than a bottom-up 'spot-the-odd-one-out' task was used. However, Wolfe (1998) suggests that there is no reason to assume that there exists a rich vocabulary beyond that of the titled or flat, left or right, steep or shallow and big or small categories. Smilek, Dixon, and Merikle (2006) suggest that this effect may more readily reveal itself when the user is passively looking for a target - i.e. by allowing the target to pop into [one's] mind - rather than actively searching for the target and actively directing attention. This remains consistent with the top-down observation of Wolfe, Friedman-Hill, et al. since Smilek, Dixon, and Merikle displayed both the category label and the graphical target before the distractor set. In addition, the study of Smilek, Dixon, and Merikle shows that category-symbol relationships do not necessarily have to be based on a close relationship between label and symbolic repre-

sentation; through training, participants categorised simple line strokes with seemingly unrelated category labels such as animal types and stationary items. This shows that provided the user is familiar with the pairing based on a prior experience, the mapping of category label to the graphical symbol results should result in a superior visual search and decoding performance.

It may be the case that these are coincidental observations. Accordingly, perceptual psychologists remain divided on the issue (Wolfe, Friedman-Hill, et al., 1992; Wolfe, 1998; Mack et al., 2002; Smilek, Dixon, and Merikle, 2006). However, the effect of category on a user's search task may facilitate a user's experience with the interface in other ways, in addition to a potential improvement to search efficiency. One possible way is that visual target configurations that can be referred to by category e.g. biggest, smallest, could also ease the learning overhead of the mapping between data and graphical feature and provide ongoing reinforcement that a larger item indicates a larger magnitude, without necessitating an eye gaze shift - i.e. saccade - to an encoding legend to re-confirm the encoding paradigm in use.

At the present, there remains a need for a significant portion of applied research in naturalistic visualisation tasks to ascertain the applicability of these perceptual insights. It is unclear whether the results from pure and highly controlled experimental examinations transfer smoothly into tangible performance benefits in applied settings. In information retrieval tools for instance, the great heterogeneity of the data set exacerbates the likelihood that target-distractor differences and similarities will clash and make it harder to process efficiently. Furthermore, of the applied visual search research, Kunar and Watson (2011) hint that a smooth transition between theory and practice may not always apply. On the other hand, Bartram and Ware (2002) and Ware and Bobrow (2006) demonstrate that perceptually motivated design choices can result in tangible performance benefits in naturalistic or ecologically valid settings. Furthermore, Healey, K. Booth, and Enns (1993) also demonstrate tangible benefit when applying visual search insights to information visualisation; but do concede that the dimensionality of a data set is potentially large and only a small number of data features are on show at any one time. Such limiting displays do not account for analyses that are more complex and suggest that a rebuild of the visualisation is necessary when new hypotheses necessitate further data variables for confirmation.

At present, no single theory of visual search can account for all phenomena observed during visual search experiments (Wolfe, 1998; Wolfe, 2007). Consequently, this situation suggests that encoding rules motivated from a visual perception perspective will continue to evolve. If however, a clear set of encoding rules does eventuate, a connection should be made between encoding rules based on both data type and models of perception. To this end and in the present context, the emphasis that Nowell (1997) places on the importance of task type is of key importance.

We can view visual search for a single multi-feature target as corresponding to

single counting tasks as explored in Nowell's experiment, and thus we could define an encoding paradigm based on data type and perceptual efficiency. Alternatively, we could take a list of data attributes, then arrange each entry by importance based on the search context per Mackinlay's importance principle, and match those to a perceptual efficiency ordered set of graphical attributes.

This sorting implies that there exists some ranking based on perceptual efficiency. However, presently, we can only estimate such a ranking based on the literature, since with no robust model of visual search, the results from such experiments are exploratory in nature. It is apparent however, that colour is likely the most efficiently perceived graphical feature - observed in the applied sense by Yost and North (2005) - followed by in no certain order: orientation, motion direction, motion phase, size, curvature, and shape; and only then, assuming the right conditions are met as outlined by Wolfe's eight criteria.

Search for a target gets harder with increasing target-distractor similarity and easier with distractor similarity (Duncan and Humphreys, 1989). Thus, the choice of features in use must be constrained if it is important that features are to guide the user to a single target. This alone is explanatory of the majority of applied research already; much research has sought to utilise a single efficient feature to guide attention to a specific subset of the data (Bartram and Ware, 2002; Ware and Bobrow, 2006) or a limited set of feature instances (Healey, K. Booth, and Enns, 1993) in order to maximise search efficiency. These applications assume some kind of filtering or brushing action has defined a subdivision of the data set, rather than a user's cognitive shift and re-focus of attention to different graphical features.

In terms of a metadata visualisation application, an encoding paradigm based on perceptual efficiency should support visual search for at least 1-4 conjunctions and should be efficient in the presence of distractor features. Consider colour, size, orientation and density with each of four levels. A visual search should be efficient for any instance of colour, size, orientation and density; visual search for any combination of two to four feature instances should also be efficient. A visual search should also be efficient for search on any combination of features and a combination of a single feature's instances. In the latter case, it is possible to imagine document search criteria consisting of all documents of type A but not type B encoded as different hues and with large size encoded by icon size. However, not all conjunctions are efficiently processed.

Kristjánsson, Wang, and Nakayama (2002) offer some perspective on conjunctive search and the role of priming. They suggest that organisms are not typically on the look out for novelty in the visual field and that targets tend to be relatively stable over time. They show how search efficiency improves when the target has appeared under similar configurations in earlier trials and show how efficiency degrades when the target they are searching for swaps from something familiar from earlier trials to something entirely new.

A parallel exists in search for information based on the context of document search sessions. Typical search sessions may be characterised by the model the searcher applies to their document search. For searchers engaged in research, the user model typically consists of document search across a single field or set of sub fields, perhaps within a date range, using a set of key themes and topics or a set of key authors. The document search targets are likely to be defined by similar attributes and represented quite similarly. Thus, the searcher is unlikely to witness significant novelty throughout their search session, as the appearance of distractors and targets is likely to be relatively stable.

The notion that categorical information, under some circumstances can improve visual search efficiency is important because the idea of naturalness of encoding depends on the cohesion between categorisation of the graphic encoding and categorisation of the data. If naturalness of encoding can be of benefit to search and benefit learning, then efficiency and usability should improve as a whole. In the following sections, discussion will turn toward sources of categorisation.

4.2.7 *Encoding by Natural Encoding*

The breadth of the above overview signifies an overwhelming number of factors that a visualisation designer might consider when conceiving graphical encoding rules. However, overall task effectiveness depends on whether the user can assign meaning to the encoding. Typically, this understanding is made possible by observation of a map legend indicating the graphical features and their propositional meaning. With ongoing use, users should become less reliant on the legend having committed the rules to memory. Furthermore, the process of committing and maintaining encoding rules to memory should be easier if the encoding rules do not conflict with the established beliefs, experiences and expectations of the user.

If however, the encoding rules violate established beliefs and expectations, the user will be forced to commit temporary exceptions to their established experiences, in order to work with a visualisation. Since additional load is devoted to this explicit and conscious maintenance process, task performance should suffer as a result. A natural encoding paradigm will be one which is consistent with the expectations and beliefs of the user such that encodings are intuitive, effortless and require little conscious effort to extract encoded data from a glyph.

Encoding naturalness may be developed further if the data set can be re-coded into labels that exhibit a superlative relationship and therefore a basis for potential categorisation. For instance, continuous variables such as file size, file age, file word count and rank can be re-coded into a small number of corpus-appropriate cases. By corpus appropriate, is meant recoding can take into consideration the distribution of file sizes, file ages, and ranks across the corpus, when calculating an appropriate bin

Tab. 4.1: A natural encoding scheme for use in metadata visualisation.

Object	Data	Type	Value	Label	Graphic
Document	Rank	O	1, 2-3, 4-9, 10-27	Best, Better, Worse, Worst	Colour Scale, Hue Metaphor
	File Type	N	PDF, HTML	Formal, Informal	Hue, Iconic
	File Age	O	2008, 2009	Archive, Old, New, Latest	Vibrancy, Movement
	File Source	N	EDU, GOV	Academic, Government	Iconic
	File Size	Q	64Kb, 128Kb	Tiny, Small, Big, Huge	Size
	Word Count	Q	.1K, 1K	Least, Little, Lots, Most	Texture, Density
Cluster	Cardinality	O	1, 2-3, 4-9, 10-27	Tiny, Small, Big, Huge	Size
	Rank	O	1, 2-3, 4-9, 10-27	Best, Better, Worse, Worst	Colour Scale, Hue Metaphor

size. With an appropriate textual adjective for each bin that can form a superlative relationship, e.g. smallest, smaller, small, big, bigger, and biggest, searchers should not have to remember specific data numbers - only the labels or categories. Thus, when searching for a target, the searcher need only to look for targets defined for short labels like ‘best-small-latest’. In contrast, search for targets of top ranked, 5kb or less, and authored today - while more specific - are more challenging to hold in immediate memory.

Table 4.1 below proposes an encoding paradigm for a set of metadata that are relevant to the scope of a web search visualisation. This proposal exhibits the greatest naturalness because the ordinal data features can be recoded into a small number of bins and labelled or categorised by adjectives that are more easily associated with the data. A similar encoding proposal could be devised for a cluster visualisation where each glyph represents one or more results. The tabulated graphical features are suggested as exhibiting the best natural fit. Further work toward motivating the best natural fit is presented below in Table 4.2.

Table 4.2 compares a set of encoding rules for the six data features present in Table 4.1, utilising each of the different sources of encoding paradigms visited in the above discussion; that is, values in Table 4.2 are extrapolated from the literature review in the first part of this chapter.

In this table, the fourth column is devoted to encoding by data types under different assumptions per the work of Mackinlay (1986). Application of Mackinlay’s guidelines involves a systematic application of the best graphical features to the most important metadata feature - followed by the second best graphical feature and the second most

important metadata feature and so on.

The isomorphic column is comparatively empty; only file size has a direct correspondence to icon size. An argument could be formulated for or against the use of icon size to encode rank; rank is a magnitude and size should encode magnitudes, and bigger size solicits greater arousal in the observer (Anderson 1990 in P. Hu, Ma, and Chau (1999)); however, strictly speaking the most important rank is rank one - the smallest magnitude. Similarly, time could be considered a magnitude of units from an epoch and word count could be considered a magnitude as well. However, with increasing word count, we are increasing the size of the file. Furthermore, per convention, we refer to the number of bytes of a file as the file size and the mere presence of the word 'size' provides a direct correspondence; this engenders the notion that a bigger logical file size would take up more physical space in reality.

In the next column, encoding by pre-attentive features, the disparity between encoding by data type guidelines is apparent. Hue, a top of this column, is typically the most efficiently processed graphical attribute, yet this is not reflected in the data type guidelines. Next, motion - which is ambiguous as the previous chapter can attest to - can be taken to mean one or more of flashing, moving, rotating and pulsing. Healey and Enns (2011) present three empirically-supported examples of pre-attentive motion: flicker, motion velocity and direction of velocity. In contrast with the paradigm obtained by Mackinlay's work, the visual search literature do not offer clear feature rankings; instead it recommends a set of features that are conducive to efficient visual search (Wolfe and Horowitz, 2004).

Social norms and conventions are more or less the same and like the 'By Isomorphic' - diagrams and pictographs - column, are extrapolated from the literature, which likewise offers no established ranking of how one way of encoding is necessarily better than another. With no existing rankings in place, the mappings are substantiated by earlier examples in the literature, and thus the order of these would likely undergo change in the future.

Finally, the last column lists a natural encoding paradigm for web search visualisations. The naturalness proposal favours the conventions and metaphor columns. Interestingly, if rank is encoded as a quantitative variable the encoding by data type paradigm matches more closely that of the naturalness paradigm. However, it is of little use to encode rank quantitatively as small differences in rank mean very little in a pragmatic sense and the task of locating the say, top five documents, could be made harder if perceiving differences between the top ten documents were difficult. Accordingly, if considering rank as ordinal, for this metadata set, encoding by data type would not encode file size to icon size in stark disagreement with each of the other encoding paradigms.

The metadata present in Table 4.1 and Table 4.2 provide non-textual information

scent during document search; in line with Mackinlay's principle of importance ordering, the set of metadata have been ranked. Not all metadata are equally useful at all times, for all search tasks; but, the order of items in Table 4.2 reflects the importance of each metadata - according to the author - in this imaginary search task. A sorting of the metadata are important, in order to establish a mapping that pairs the most important metadata with the most effective graphical features.

Position is not shown here, but is assumed to be coded to position along the projection axes of the visualisation and therefore assigned to semantic interpretation - still consistent with Mackinlay's principle of importance, since semantic information is important first and foremost, with metadata of secondary importance. The next most important features are file type and query similarity or rank which could be re-coded into a small set of relevance categories. The order of these is debatable. Ultimately, having directed visual search activity to one particular semantic region of the visualisation, in lieu of additional semantic information like textual labels, the user is likely to fall back to selecting results based on query similarity or some other heuristic such as file type to guide decision-making. Similarly, a searcher could preference more recent documents or documents from a particular source. However, file size and word count offer very limited semantic scent and are consequently ranked last. In contrast, if glyphs represented clusters, cardinality could indicate singleton outliers and overly confused and bloated clusters.

For this metadata set, which is appropriate for a search result visualisation, encoding by naturalness violates the least number of expectations a typical user may have. Such a claim is clearly violated if the observer does not possess the same expectations that is argued are necessary for natural encoding. However, with the exception of perhaps the iconic representation of file source, the interpretations and expectations of the other attributes are generic and natural. File size and magnitude is culturally unspecific, hue and file type relates to the branding of the software companies that developed those ubiquitous file formats and therefore is not subject to cultural interpretations; brightness implies order, and magnitude, and fast motion denotes fresh and latest in contrast to old and slow. The role of shape or density to encode word count is debatable. With more words, a page appears denser; however, we count the sides of shapes i.e. 'more sides' as we do words 'more words' and so a weak correspondence exists in either case.

Perhaps the most important message to note from Table 4.2 is the difference between the 'By Type' and 'By Naturalness' columns. All other columns summarise the potential encoding palette available; however, with no available ranking information, it is not yet possible to determine if these orderings are correct nor if the actual features are widely agreed upon. Mackinlay's rankings are among the pre-eminent work that a data encoding practitioner may make use of when constructing charts and graphs. However, the naturalness hypothesis envisions that in search applications, search outcomes

Tab. 4.2: Encoding paradigms based on approaches discussed in chapter.

Feature	Type	By Type	By Iso-morphic	By Pre-Attentive	By Social Norms	By Metaphor	By Naturalness
Rank	O	Density	.	Hue	High & Low	List	Brightness
File Type	N	Hue	.		Iconic	Iconic	Hue
File Age	O	Saturation	.	Motion	Left & Right	Motion	Motion
File Source	N	Texture	.	Orientation	Iconic	Iconic Symbolic	Iconic
File Size	Q	Orientation	Size		Size	Size	Size
Word Count	Q	Size	.		.	Density	Shape

will be superior when encoding is more natural to the searcher. A major discrepancy in Mackinlay's ranking - which is at odds with the 'Social norms' and 'metaphor' columns - is the encoding of file size.

This chapter submits that metadata visualisations encoding data using other data encoding guidelines are suboptimal because the encoding paradigms present data in ways contrary to expectations and established beliefs. As a result, we need to consciously and explicitly override our predispositions in order to interpret data encoded according to the dictating guidelines, which may not even be suited for the types of tasks users engage with metadata visualisations. As our previous beliefs and expectations interfere, our performance on task suffers because explicit effort is required to overcome the interfering information. In order to further illustrate the breakdown of processing, consider interference tasks in psychology, such as the Stroop test.

In a Stroop test, task performance on calling out the ink colour of a colour word degrades when the ink colour is incongruent with the colour word - e.g. red ink of the word 'green' is incongruent while red ink of the word 'red' is congruent. The Stroop test is among the most seminal works in the field of psychology and MacLeod (1991) provides a comprehensive survey and overview of literature emanating from the original experiment. The original purpose was to investigate interference effects of the visual and linguistic channels in the perceptual-cognitive system. For completeness, the original experiment consisted of three main experimental stages. In the first, two experimental and two control experiments investigated how performance differed on reading out colour words printed in incongruent ink colours. The control condition provided a base line such that words were printed in black ink. In the second experiment, the task was to read aloud the ink colours of the same experimental stimuli as in experiment one except that the control cards were squares of colour instead of words. An interference effect - evident by a large increase in task time performance - was noticed in experiment two. In experiment three, the effects of training of ink naming on colour word naming performance were examined.

Over the almost century that has now passed since Stroop's experiment, various explanations have been proposed to explain the interference effect. Initially, models assumed a sequential processing of information in contrast to more recent parallel

models of the human brain. Such explanations include relative speed of processing and automaticity (MacLeod, 1991). Relative speed of processing proposed that we read words faster than colours. When naming the colour and naming the ink colour, separate brain processes compete to be the response produced; the interference causes a time cost - two responses that race to control the final output. Automaticity on the other hand, is based on the idea that processing of one dimension - i.e. colour - requires more attention than does processing of the other - i.e. a word. Naming the ink colour requires more attention compared with reading a word since reading a word is said to be automatic - it is something we do on a daily basis in comparison to naming colours.

For the present discussion, the Stroop effect and the other interference tasks such as the SNARC effect that was earlier alluded to in Chapter 2 are discussed solely to draw an analogy between naturalness in encoding and the detriment to task performance when say a colour word does not meet with the text of the word. The analogy drawn is that for natural mappings, which pair a data value - having an inherent spatial, graphical or verbal cognitive interpretation such as magnitude, intensity, or age, etc. - with a graphical value - which preserves the property present in the data attribute such as using size, texture density, or vibrancy respectively - no incapacitating interference or interruption will be present. Therefore, when one channel interferes with another, task performance will degrade since the user has to overcome the interference in order to make the correct response.

The notion of interference task has similarly been commented by Dormal and Pesenti (2007) to predict superior performance on tasks and also by Kosslyn (2007) who utilises the Stroop Test and other conflict tasks to illustrate how we are confused when messages ‘from the form itself and from the meaning - conflict’.

While participants can improve with practice (MacLeod, 1991), additional, subtle effects could compound a negative experience during interactions with interfaces sporting interfering presentations and encodings. As participants struggle to overcome their prior beliefs and expectations, more effort and time is expensed and directed away from the task. Participants engaging such interfaces may experience those interfaces as subtly unusable and result in negative impressions and experiences. More research is required to devise encoding paradigms that are more tuned to the context of data encoding, the user, and the typical tasks they perform with tools. The second half of this chapter will report on the outcome of an experiment that attempted to validate a subset of this natural encoding paradigm: specifically, that it is natural to match file size to icon size and word count to shape ‘arity’ i.e. the number of sides of a shape. A successful outcome of this experiment will motivate a more thorough investigation and validation in the future.

4.3 *A Web-Based Evaluation of Natural Encoding Paradigms*

The opening half of this chapter has discussed alternative perspectives on how we might obtain guidelines, direction or inspiration when devising metadata encoding paradigms for use in search tools. A unifying theme of each perspective has been the degree of association the represented data has with the encoding graphic attribute. This association is expected to influence task performance when used in encoding paradigms. Interfaces should be easier to learn and use, and as a consequence, result in faster task completion time and superior accuracy. Searchers should not have to store and restore an unintuitive mapping in short-term memory, while they interact with an interface; this situation is posited to occur when the user cannot draw a link between the graphical code and the label, thus requiring the user to repeatedly saccade back and forth between an encoded object and a legend displaying the encoding paradigm in use.

4.3.1 *Method*

Participants

59 unpaid participants commenced the experiment. 29 participants did not complete the experiment, while 30 participants successfully completed the experiment. An analysis of drop out will later examine the entire participant pool in full, but the remainder of the results and ensuing discussion will consider only the results from the set of successful completions only.

Of the completing participants (17 male, 13 female), 38% and 29% reported their age as within the 20-29 and 30-39 ranges respectively, while the remainder all spread across the under twenty (10%) and over 40 age brackets (23%). The data show that participants engage in a weekly average of 41 hours (4 min, 110 max) of computer-based work of which an average of 12.5 hours (1 min, 40 max) is spent using search engines. All participants reported using search engines every week. Participants indicated spending an average of 6.3 hours (0 min, 60 max) per week playing computer games however over half of respondents report either zero or very few hours of gaming, with just a small minority of heavy gamers responsible for this elevated figure. The same cannot be said for average computer usage or search engine usage. Furthermore, all but one participant reporting having spoken English for their whole life or for a very large part of their life; the exception reported speaking English for three years only. Finally, 35% of participants reported some experience with graphic design and three reported left-handed use of the computer mouse.

Participants did not undertake a test for colour blindness. This experiment intended to capitalise on normal human vision; therefore, participants unable to complete the experiment due to poor colour perception, were assumed to have abandoned the experiment.

This experiment was conducted over the Internet and participation was not restricted to any particular demographic. An analysis of the geographic origin for each participant shows an approximate 1:1 ratio of international to Australian participants. 23% of completing participants were internal to Flinders University of South Australia. These students span a range of disciplines including health sciences, social sciences, science and engineering. External participants report a number of different backgrounds including students from unreported disciplines, occupations in nursing, information technology, education, hospital administration, and building, while one reported being a home maker and one reported unemployment at the time of participation. Analysis of the participant pool based on IP-Lookup estimates that the pool is geographically diverse. Of participants external to Flinders University, 26% originated from within Australia while 51% came from outside of Australia in countries including Canada, New Zealand, the United Kingdom and the United States of America.

Materials

This section outlays the process of construction of two task-sets consisting of 6 questions each and 125-150 text documents. The construction of each task set followed the same process, so where applicable examples are illustrated for one task set only.

Documents and Document Source Search results selected for use in this experiment were retrieved from the Yahoo Search API service. Three factors motivated a commercial search API service for the source of search results.

First, a medium-term goal for this course of research involves improving or augmenting ranked-list interfaces, by intercepting results on the client side, in order to experiment with the format of result presentation and furthermore to provide additional interactive control over the set of search results. Thus, a source of web-based search engine results was desired.

Second, Julien, Leide, and Bouthillier (2008) argue that information retrieval experiments should utilise known baselines for experiments including web search engines or library catalogues and not private or inaccessible document collections. The advent of tools such as Lucene <http://www.apache.lucene.org> makes it extremely easy to set up fully functional search engines for many purposes; accordingly, this experiment could have easily utilised such a set up to search through a personal collection of documents for instance, which would not be publicly available.

We should use a publicly available test collection in place of a private corpus for information retrieval experiments in an effort to foster replication and standardisation. Julien, Leide, and Bouthillier (2008) suggest a visualisation-based system should be compared to a baseline system i.e. a text equivalent and this system should be publicly accessible. Borlund (2005) also raises that generic test collections should be employed

so that expert knowledge is not a pre-requisite to participation. Like all commercial web search engines, what is known about the Yahoo! API search system is limited and result lists may change from time to time; but, reporting retrieval performance over a collection restricted from the public domain is tantamount to asking readers to simply entrust their faith.

Third and finally, a commercial search API provides a significant quantity of meta-data with each search result, which has the effect of shortening development time and task set preparation. For this experiment, only the data provided by the Yahoo Search API service were incorporated into task sets, thereby discounting any need to download and parse the full text of each result a tedious process that adds additional development overheads and theoretical considerations.

While there are several API services available to obtain search results programmatically, Yahoo's API was selected because it provided the richest source of metadata and therefore lowered development time. However, several researchers have analysed search APIs to determine their applicability for research purposes. McCown and Nelson (2007) provide a detailed analysis of three popular web search APIs over a period of six months. They found a marked difference between the results delivered at the search engine's main site and the results delivered via search API. While both web based and API indexes were periodically updated - as evidenced by the changes to the top ranked results - the re-index time can be significantly longer for the API service.

The observed update disparity was not expected to influence the outcome, as all participants utilised the same set of results and could not submit their own search queries ad hoc. In addition, Kumar and Kang (2008) performed an analysis of web search APIs concluding at the time that Yahoo's API service was superior on factors such as quality of service and of product, and of the level of restrictions and limitations that would otherwise hamper efforts by researchers whom were intending to build test result sets or experimental search interfaces.

At the time of this experiment's design, the Yahoo search service offered a limited quantity of free search results per day and a rich set of metadata including title, keyword-in-context snippets, URL, file size, date of last update and 20 descriptive keywords. The current version of Yahoo's search service offers most of these but the provision of keywords for each result do not appear to be offered as part of the search API service any more. Moreover, since the aforementioned surveys and analyses of Kumar and Kang (2008) and McCown and Nelson (2007), the major search engine companies have sought to monetize these services by introducing a small cost per submitted query. While the cost is very low, the effort outweighs the benefit of subscribing to such services. Furthermore, the per-day free query allowance is sufficient for development or task set preparation purposes.

Two conjunctive queries: 'dog + train + security' and 'music + festivals + Aus-

tralia' were submitted to Yahoo and the first 150 hits were downloaded for each. Each result set - i.e. document set - contained a number of duplicate documents. Using the source URL and file size metadata, lower-ranked duplicate documents were detected by visual inspection and excluded from the analysis. In addition, several small documents that did not have a full set of 20 keywords were removed from the set. The eventual document set sizes for the two queries were 125 and 146 documents respectively. The following sections describe the processing steps involved for preparing a set of search results into a task set for the experiment.

Task Set Construction Participants completed a set of questions for each of two task sets. Each task set consisted of a collection of documents, a clustered spatial model of the documents, a set of keyword clusters, and key metadata. The pre-processing of document sets took place in two stages.

The first stage involves the creation of a document spatialisation and clustering; the second stage involves the organisation of document keywords into a clustering that will facilitate navigation of the document set.

Document Spatialisation In the first stage of pre-processing, documents are processed to produce a spatial model of the set of documents. Each document's full-text is not downloaded; instead, each document's title and snippet text from the search result set are utilised for the purposes of spatialisation. Use of document title and snippet text, in place of full-text content has the benefit of reducing parsing overheads and processing time, while achieving reasonably comparable performance with a like experiment that utilises the full-text (Zamir and Etzioni, 1998). Furthermore, (Kriegel, Kröger, and Zimek, 2009) argue that employers of spatialisation methods must strike a balance between too many variables i.e. words, which make pattern finding more difficult and which make interpretation of proximity and distance less meaningful, and too few variables, which increases the risk of missing new insights into the data. These problems typify the more general notion of the 'curse of dimensionality' whereby each unique word in a document corresponds to a dimension, and too many dimensions make it difficult for pattern recognition. For the present experiment, it may be reasoned that the use of words in the title, snippet and keywords, is one way to reduce the number of variables down such that they are representative of the documents. This reasoning however is founded on the assumption that the search API service provides a set document metadata that is representative of each document's full-text.

In preparation for spatialisation, the document's title and snippet text are stripped of stop words and special characters, transformed to lower case, tokenized and collated into a term frequency matrix. However, keywords accompanying each result are processed differently. Keywords are not tokenized and are not subject to stop word

filtering. Keyword n-grams are inserted into the term frequency matrix as multi-word phrases.

The term frequency matrix is row and column demeaned before undergoing a Single Value Decomposition - here on SVD - by way of the Java Matrix library <http://www.nist.gov/numerics/jama>. SVD is employed in order to reduce the very high dimensional word space down into a few representative themes and topics made up of patterns of co-occurring words.

A similarity matrix is then constructed from the SVD (document x theme) output matrix, using a dot product as the similarity measure. Each cell of a similarity matrix represents the similarity between every other document including itself. The value of each cell is then subtracted from one, in order to produce a dissimilarity matrix, which is passed as input to an implementation of the Multidimensional Scaling - here on MDS - algorithm (Group, 2009). MDS projects each document into a reduced space such that the distances between documents in the input matrix - i.e. similarities - are preserved as faithfully as possible in the projection (Buja et al., 2008). In this case, MDS is used to reduce the space to two dimensions, suitable for visualisation purposes.

Finally, a k-means clustering algorithm, utilising the first two MDS projection dimensions, with k=11 chosen arbitrarily, clusters documents into a set of eleven clusters. A cluster centroid is calculated for each cluster and recorded. This centroid location marks the final position of the cluster icon appearing in the visualisation.

To assist the participant with navigation of clusters, document keyword sets are processed into a shallow tree structure and visualised adjacent to the cluster visualisation. The construction of this hierarchy is described in the next section.

Keyword Tree Construction In the second stage of pre-processing, a keyword tree - Figure 4.1 - is constructed to facilitate cluster exploration. Use of a similar structure for the presentation of cluster labels is discussed by Carpineto, Osinski, et al. (2009) and is acclaimed for its simplicity, ubiquity, shallow learning curve and computational efficiency. In this apparatus, the keyword tree is linked to a document cluster visualisation; as the searcher selects interesting keywords from the keyword tree, cluster icons change colour to reflect the spread of selected keywords in the visualisation. The keyword tree structure is constructed by first gathering a set of prominent keyword phrases for each cluster, generating a sorted list of unique keyword phrases and then merging similar keywords and phrases together into keyword clusters.

To form each cluster's prominent keyword set, twenty unique, top-ranked keywords are extracted from cluster member keyword sets. Starting from the highest ranked document in the cluster, the top-ranked keyword is selected and inserted into the cluster's prominent keyword set. The top-ranked keyword from the second highest ranked document is then selected and inserted into the cluster's prominent keyword set, provided

this keyword is not already present in the prominent set. In the event that the keyword is already present, the next top-ranked keyword is not selected for insertion in order to ensure only the top ranked keywords are present in the prominent set. This process continues in order of document rank until twenty prominent keywords are selected for each cluster. At the end of this process, each prominent keyword set is recorded for each cluster.

For example, a prominent keyword set for a singleton cluster is the set of twenty keywords provided by the search API. Alternatively, for a cluster of two members, the prominent keyword set consists of half of the top ranked keywords from the first document and half of the top ranked keywords from the second document. Alternatively again, a cluster of 20 members has a prominent keyword set consisting of each document's top-ranked keyword, if each top ranked keyword is unique. However, if each top-ranked keyword in a cluster is unique with the exception of two, the prominent keyword set consists of each of the unique top-ranked keywords and the second highest ranked keyword from the top-ranked document and the second highest ranked keyword from the second top-ranked document, provided the second ranked keywords are also unique.

Next, the prominent keyword sets are merged into a single unique set of keywords which is then pruned to remove pluralised word forms, provided that the singular word form is also present e.g. 'dogs' is removed if 'dog' is present. The pruned list is sorted alphabetically using the first non-stop word token in each keyword phrase as the sort key.

Then, starting at the top of the list, keyword clusters are formed by merging keywords that share a derivationally similar sort-key term; for example, 'training' and 'training your dog' merge together, but 'training your dog' and 'trick' do not merge. The set of keyword clusters and keyword singletons is sorted again, using the second non-stop word in keyword phrases as the sort-key and merging follows in a similar fashion; for example, 'clicker training' merges into the set of words beginning with 'training'. This process continues until no further merges are possible. At the end of merging, each keyword cluster is labelled by selecting the top-ranked keyword from the top-ranked document in each.

All remaining singletons are merged into a single 'others' set. The final 'others' set is typically long, yet contains potentially relevant keywords. However, where possible, all keywords should be assigned to a conceptually related set, since there is no information scent offered by a keyword set labelled 'others'. A more sophisticated clustering algorithm, based on semantic similarity, might place 'aggression' into the 'behaviour' keyword cluster; however, in this experiment, no further processing of the 'others' set takes place.

The main advantage of this approach is that it is quick and has relatively low pro-

cessing overheads in comparison to more sophisticated approaches to concept clustering - such as utilising the term space (terms x themes matrix) of an SVD for instance. The main drawback to this approach is that - at development time - there was no obvious simple method of dealing with the overly large and typically bloated 'others' set; however, an average rank distance approach, utilising the ranked keyword sets, could have been taken in order to find the keyword groups most appropriate for each item in the 'others' set.

For this experiment, quick and dirty methods are favoured in anticipation that similar approaches could be used in browser plug-in software (Treharne, Pfitzner, et al., 2008) to provide JavaScript-driven, real-time augmentations to web based search services without needing computationally intensive processing. An alternative may have been to arrange this keyword tree by hand, but this was not preferable in the interests of replication and again, in anticipation that similar techniques could form the basis for future lightweight implementations.

The keyword tree provisions a way for participants to interact with the visualisation, but not necessarily a way to locate documents based on relevance to a query; the keyword tree facilitates navigation of clusters based on prominent keywords. However, experiment trial questions also require participants to think about additional cluster metadata such as the size of the cluster and the number of words present in the cluster. The extraction of this metadata is outlined in the next section.

Document and Cluster Metadata Participants interact with the clustered results only, as the apparatus purposely lacks an interactive control to facilitate inspection of individual documents. In order to assist understanding of each cluster's semantic content, a set of metadata are devised for each cluster, consisting of the cluster's cardinality - i.e. the number of documents in the cluster - a set of prominent keywords - outlined in the prior section - and the number of words in each cluster.

While word count is usually an unimportant metadata attribute a case can be mounted for word count particularly if the search domain involves lengthy documents such as documents that cover a broad range of topics or for a situation where one does recall something about the length of a document one is looking to re-find - cluster size is not unimportant. Cluster size indicates whether a point representation depicts a singleton document or a collection of documents. Cluster word count on the other hand is somewhat less important. This experiment called for one metadata feature reflecting the size of some item such as a document or cluster and one metadata feature which was numerical but not necessarily reflecting the size of some quantity. The cluster's mean rank mean rank of member documents - was considered, but given that participants answer six different questions using the same document set, encoding rank may have been deceptive, since participants are not privy to the original query and the question topic changes across questions. Moreover, no other universal metadata for a document

search application were thought of - beyond metadata that are unique to a specific flavour of document search such as search for emails, news articles, or academic articles that could form the basis of this experiment. This meant that word count, as an ordinal data type, was adopted and coded to shape as a natural encoding, since a greater side count should be associated with a greater number of words.

The cluster's total word count is calculated by summing each document's word count in the cluster. A document's word count is estimated as a function of the file's byte size. While this does not guarantee a precise number, a rough estimate was sufficient for this experiment. Since the full-text document is not made available to participants, an accurate value of word count can not be confirmed by the participant. This saves the need to download and parse each search result only to provide an accurate indication of the number of words in the text.

Each cluster's metadata is presented in a pop-up window triggered when the participant mouse-hovers over the cluster icon. The discussion of the apparatus in the next section provides a more thorough examination of the cluster visualisation, the keyword tree and the pop-up windows.

Apparatus

The apparatus in this experiment is JavaScript driven and browser-based, which utilises the HTML-5 canvas element for drawing. The apparatus downloads each task set at experiment time and populates each of the four main components of the apparatus. The four main components of the apparatus are the keyword tree, the task bar, the cluster visualisation and the keyword colour-coding display. A screen shot of the apparatus is depicted in Figure 4.1 configured for the control condition. The only difference between the control configuration and the natural or unnatural encoding configuration is that pop-up windows contain the number of words in each cluster. This information appears adjacent to where the cluster cardinality is displayed; there is no metadata value coded to the glyph size in the control condition.

The cluster visualisation depicts the eleven cluster icons. The cluster icon's shape, size, border colour, inner colour and location are configurable parameters depending on the trial condition and the participant's interactions. A mouse-hover event triggers a pop-up window for the requested cluster and this pop-up displays the cluster's metadata consisting of the number of words in the cluster, the number of documents, and the prominent keyword set. The pop-up is removed from display, when the mouse cursor exits the bounds of the icon.

Participants select answers by left-mouse clicking on icons; the red bar below selected icons provides visual feedback to indicate that the icon is selected.

The concept tree visualises the shallow keyword hierarchy. Clicking on the '+' sign next to a branch node opens the keyword cluster revealing the set of related keywords

for that node. Clicking on a keyword node highlights every cluster icon containing that keyword. The selected keywords and the colour assignment appear at the right of the visualisation.

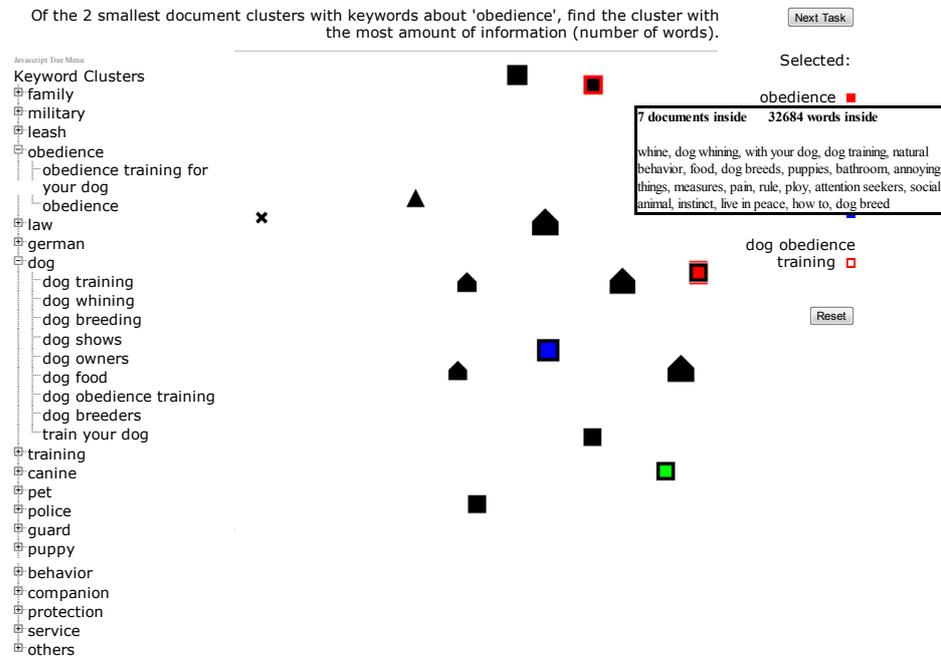


Fig. 4.1: A screen shot of the apparatus for naturalness experiment; the control interface is depicted - under natural and unnatural conditions the pop-up window displays the number of words adjacent to the number of documents.

The keyword colour assignment depends on the number of keywords already selected. Cluster colour coding depends on the pool of keywords selected and the degree of keyword overlap in the cluster's keyword set. The selected keywords appear in the keyword colour-coding widget; up to six keywords can be selected at a time. On selection of a seventh keyword, the keyword selection display resets and the seventh selected keyword becomes the only keyword selected for colour coding.

The first three selected keywords control the level of red, green and blue in a cluster's inner area colour. The fourth, fifth and sixth keywords control the level of red, green and blue in the cluster's border colour. The resulting colour coding of a cluster is based on an additive red, blue and green colour scheme. For example, if three keywords are selected, and a cluster contains only the third selected keyword, the inner area of the cluster icon is painted blue. If the cluster contains the first and third keywords, the result is the additive outcome of red and blue i.e. magenta. If the cluster contains all three keywords, the result is white. If the cluster contains all six keywords, a thin outline of a cluster maintains the icon's appearance. All possible combinations are depicted in Figure 4.2 for a square shape.

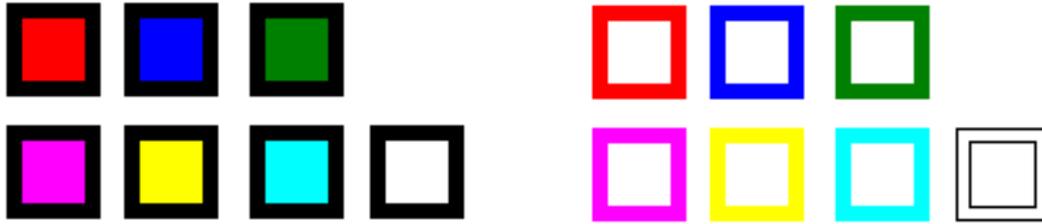


Fig. 4.2: Hue palette for naturalness experiment; icons on the left depict coding of inner area while icons on the right depict coding of outer area - an icon can have both inner and outer areas coded concurrently.

Finally, the task statement that appears at the top of the interface defines the trial task. The ‘next task’ button appears adjacent to the task statement. Participants click this button to proceed through each experiment trial. If the participant has not made an answer for a trial, a modal error pop-up is displayed informing them that they have to select at least one icon in order to proceed to the next trial. As participants move through each trial, the results of the previous trial are sent to a web server.

Procedure

Participants navigated to the experiment web site - at a physical location of their own choosing - after having been personally invited, received a handout advertisement, seen a noticeboard advertisement or seen a listing on the Hanover College Psychology Study Website (Krantz, 2012). The starting procedure involved a technology check and provision of written materials to foster informed consent. A participant is assumed to have provided informed consent by clicking ‘yes’ to the terms and conditions presented in a modal confirmation window that appeared after clicking the begin button. Participants then provided responses to a demographic questionnaire, followed by a training module.

Each participant undertook a training module consisting of two stages. In the first stage of training - see Appendix B - a series of screen shots accompany a descriptive text explaining how the interface works. In the second training stage - see Appendix C - the material offers a walk-through sample question with screen shots. Critically, participants are told how the colour coding works and how they can utilise the keyword tree to locate clusters containing the words that they are interested in finding; however, participants are not told how the shape or size encoding works.

At the end of training, participants started the experiment. There were two task sets and six questions per task set, totalling 12 experimental tasks. Each trial started at the end of the last, and there was no set opportunity for rest breaks between task sets or individual trials.

At the conclusion of experimental trials, participants responded to an exit ques-

tionnaire. Three questions of the exit questionnaire require participants to explain how aspects of the interface work; specifically, how did the colour coding work, what do the different icon shapes indicate, and what do the different sized icons indicate. As size encoding was not manipulated under the control condition, the size encoding question did not appear for those participants in the control group. The remaining questions relate to usability aspects of the tool and participant opinions regarding their experiences with the experimental apparatus and their experiences with search engines in general.

Task

Each task involved finding and selecting icons that satisfy metadata criteria specified in the task statement. There are four task types: cluster topic, cluster cardinality, cluster word count, and combined cluster cardinality and word count. For each task set, participants completed two topic questions, one cardinality question, one word count question and two combined cardinality and word count questions. Cardinality, word count and combined questions also include keyword criteria in order to promote an added level of realism in the sense that search for a cluster, based on the number of documents or number of words, without first establishing a semantic basis, is an artificial task.

Once the participant finalised their selection, they clicked the next task button to proceed on to the next task. The six questions asked for the ‘Dog Train Security’ task set are tabulated in Table 4.3 while the ‘Australian Music Festival’ task set questions are tabulated in Table 4.4.

Tab. 4.3: Experiment tasks for the Dog Train Security (DTS) task set; tasks differ by target topic, keyword, information (word count) and/or cluster cardinality.

Trial	Task Statement	Keyword	Cardinality	Word Count	Topic	Type
1	Of the 2 smallest document clusters with keywords about ‘obedience’, find the cluster with the most information (words).	X	X	X		Combined
2	Find document clusters about training dogs for law enforcement and military.				X	Topic
3	Find the document cluster with most information (words) mentioning ‘Training Tips’.	X		X		Words
4	Find the smallest document cluster mentioning ‘training’ and ‘dog training’.	X	X			Cardinality
5	Of the 2 biggest document clusters mentioning ‘dog training’, find the one with the least amount of information (words).	X	X	X		Combined
6	Find document clusters about training dogs for people with disabilities.				X	Topic

Tab. 4.4: Experiment tasks for the Australian Music Festival (AMF) task set; tasks differ by target topic, keyword, information (word count) and/or cluster cardinality.

Trial	Task Statement	Keyword	Cardinality	Word Count	Topic	Type
7	Find document clusters that list music festivals held in Australia.				X	Topic
8	Out of the 2 smallest clusters that mention the keyword ‘music’, select the one with the least information (words).	X	X	X		Combined
9	Find the document cluster containing the most information (words) about music festivals held in Australia.	X		X		Words
10	Find the largest document cluster with the most information (words) mentioning ‘music’.	X	X	X		Combined
11	Find document clusters that are likely to contain documents about classical music.				X	Topic
12	Find the smallest document clustering mentioning ‘Womadelaide’ music festival.	X	X			Cardinality

Design

The experiment had a between-subjects design with one independent variable: encoding naturalness of three levels: natural, unnatural and control. Under the control condition, encoding was natural but trial tasks did not involve decoding of graphically encoded data. Participants were randomly allocated either natural, unnatural or control encoding at experiment time.

Design configuration is outlined in Table 4.5. The first condition is a control condition in which only answers based on topic-based criteria were solicited. In the unnatural and natural conditions, cardinality and word count are task relevant in that tasks specifically requested answers based on cluster size and number of words in the cluster. In contrast, for the control condition, cardinality and word count are irrelevant to the task because participants looked for clusters matching keyword criteria only.

Tab. 4.5: Experiment condition configuration for cluster icon shape and size; conditions marked whether the configuration is natural or unnatural.

Condition	Icon Shape	Naturalness	Icons Size	Naturalness
Control	Cardinality	Unnatural	-	-
Unnatural	Cardinality	Unnatural	Word Count	Unnatural
Natural	Word Count	Natural	Cardinality	Natural

Participants were never cued to the size-encoding and shape-encoding paradigm in use. It was expected that participants would be able to determine the encoding after exposure. In addition, even if participants did not explicitly report being aware of the size and shape coding, it was expected that the efficiency of their searching would be enhanced by the use of natural encoding and thus evident in largely superior behavioural measures. In contrast, participants were told how the colour coding works; consequently, self-reported learning regarding colour coding was expected to be excellent.

Accordingly, the primary response variables i.e. behavioural measures, were time on task, task accuracy; pop-up window count i.e. pop-up window triggers, and self-reported learning outcome.

Task set order was randomised to balance any effect of learning and task set difficulty. However, order of task questions was not randomised for each participant, unlike question type order across task sets. It was important that the participant was given the opportunity to explore the document set prior to answering questions regarding specific metadata. This is more consistent with search behaviour as we are unlikely to search on metadata first, before consideration of semantic content. Therefore, a question that was purely topic-based was asked first to give participants the opportunity to familiarise themselves with the result set. Although the experiment's design would

be more robust by fully randomising the order of questions 2-6 and 8-12, ensuring adequate coverage for each level of task set ordering was preferred since the number of participants that were expected to complete the task could not be predicted ahead of time.

Three main outcomes were hypothesised. First, task performance, measured as time on task, task accuracy and pop-up windows triggered or opened, would be superior under the natural encoding condition compared to the unnatural condition. It is reasoned that participants perform better under the natural condition due to unconscious inclinations toward cluster icons that match the envisioned target icon and unconscious inclinations away from those icons that do not match. Consequently, attention is weighted toward a set of promising candidates that when considered explicitly, result in streamlined decision-making, since no violation of expectations takes place. In contrast, under the unnatural condition, having inclined toward the wrong subset of alternatives, detected violations force an explicit and conscious re-evaluation of how the interface actually works. For instance, if searching for clusters of the largest size, it is natural to incline to large icons, though if cluster size is encoded as shape, such a natural inclination would be wrong. This process should take time more and effort to maintain an exception to a participant's prior expectations and beliefs.

The second hypothesis was that since participants in the unnatural encoding condition devote additional effort to understanding the interface, self-reported learning outcomes should be the same or better than participants under the natural encoding paradigm. In line with the reasoning of hypothesis one, participants in the natural encoding condition are more likely to complete tasks without significant additional learning. In contrast, participants in the unnatural condition spend more time and effort constructing and maintaining their mental representation of how the interface works. Consequently, this manipulated representation is more highly activated, more readily accessible in immediate memory, and more likely to be reported by the participant when polled. This hypothesis assumes that participants clue to the encoding paradigm in use, through experimentation with the interface - across multiple tasks - and accordingly, formulate an internal representation of the interface.

The third hypothesis was that subjective responses should reflect a more positive attitude toward the interface under the natural encoding paradigm since a negative attitude is expected when the interface does not present information in a way that the participant feels is natural and or is difficult to use.

Stimuli

Cluster icon size, shape and position were set at experiment run-time according to the experiment condition and task set. The encoding scheme for icon shape is outlined in Table 4.6 and the encoding scheme for icon size is presented in Table 4.7. A singleton

cluster or cluster with lowest word count is represented as a cross shape or smallest icon size, while the largest cluster or cluster with highest word count is represented as two concentric circles or largest icon size. Based on an earlier result by Li, vanWijk, and Martens (2009) who found that a hexagon could not reliably be distinguished from a circle when small, this encoding scheme limits the maximum number of sides to five before moving to the circular shapes. Circular shapes are chosen to represent the largest cardinalities and word counts, as we perceive shapes as more circular with increasing number of sides. Cluster cardinalities and word counts are recoded according to a logarithmic function.

The definition of shape is critical; in this context shape is taken to mean the set of regular polygon shapes triangle, square, pentagon, hexagon, heptagon, octagon and so on. For regular polygons, it is obvious that the ordering of shape corresponds to increasing vertex count and this can be encoded to a value. In contrast, a notion of shape provided by Brath (2009) is vastly more complex as demonstrated in his exploration of the shape attribute palette. Brath's shapes are clearly not conducive to an obvious ordering. Furthermore, symbolic shapes such as stars, crescents, hearts, crosses, and diamonds are commonly recognised shapes, but these are not conducive to ordering. Bertin (2011) and Mackinlay (1986) appear to adopt a more general perspective on shape e.g. regular polygons, semi-circles, trapezoids, stars and crescents - but which is far from the extreme attribute palette that Brath describes.

The concentric circle shape of this experiment might be conceived as entirely unrelated to others in the set; since it is a shape within a shape and since it like the single circle is a shape that is not composed of straight line edges. A further critique of this shape set is that the cross is the only unenclosed shape, and furthermore, the cross - like the circle shape - is not a polygon. As a consequence of the factors in the aforementioned critique, and ordered shape set may not always be possible, especially when the number of data attributes that are to be encoded by shape, is large.

There is seemingly no specific, empirical evidence on the difficulties participants face when making a connection between regular polygon shape and a numerical value. Nowell (1997) found that polygon shape is superior to size to convey quantitative data, but not nominal data. However, Nowell observed that participants applied a metaphoric interpretation to irregular shape. When representing three intervals of relevance a variable considered quantitative by Nowell participants erroneously interpreted an upward pointing triangle as increasing relevance, offered no interpretation for the diamond shape and interpreted the star shape as superior relevance. Thus, while showing that shape can effectively encode a quantitative variable, it may be due to the intuitive interpretation of shape, more so than an increasing number of sides.

Efforts to test a participant's own connection between shape vertex count and numerical value were not piloted ahead of this experiment. Assumptions drawn about the anticipated analytical efforts by participants should not be seen to ignore the recom-

Tab. 4.6: Encoding of icon shape for word count and cardinality; cardinality values are mapped to shape and icon size according to the function $\text{floor}(\log_2(c))+1$ where c denotes the cardinality value.

Icon Shape	Vertices	Word Count	Cardinality
Cross	2	1-5000	1
Triangle	3	5001-10000	2-3
Square	4	10001-25000	4-7
Pentagon	5	25001-50000	8-15
Circle	∞	50001-100000	16-31
Concentric Circles	∞	100000+	32-64+

Tab. 4.7: Encoding of icon size for word count and cardinality; cardinality values are mapped to shape and icon size according to the function $\text{floor}(\log_2(c))+1$ where c denotes the cardinality value.

Icon Size	Word Count	Cardinality
13	1-5000	1
23	5001-10000	2-3
33	10001-25000	4-7
43	25001-50000	8-15
53	50001-100000	16-31
63	100000+	32-64+

recommendations of Bertin (2011) since in this case, the order of regular polygons is seemingly more intuitive than irregular polygons or symbols.

There is no particular reason as to why the icon size mapping is non-linear, as it was so in the motion experiment. Cardinality values are mapped to shape and icon size according to the function $\text{floor}(\log(c)) + 1$ where c denotes the cardinality value. A separate function is used to assign word count to shape and icon size, but for the purposes of this experiment, the mapping outcome is massaged slightly to fit the word count distribution. This may be justified in the fact that participants do not need to extract specific numerical values from the visualisation. While this introduces a slight artificiality, it nonetheless ensures that all shape icons and shape sizes are depicted at least once in the visualisation.

Colour coding of keywords as described above utilises red, green and blue. The colour combinations possible are depicted in Figure 4.2 and made concrete in Table 4.8; The RGB values are Red #FF0000, Blue #0000FF and Green #00FF00. As additional

Tab. 4.8: Complete cluster colour-coding scheme and interpretation.

Keyword One Colour	Keyword Two Colour	Keyword Three Colour	Final Colour	Keyword Interpretation
Red	-	-	Red	One
-	Green	-	Green	Two
-	-	Blue	Blue	Three
Red	Green	-	Yellow	One, Two
Red	-	Blue	Magenta	One, Three
-	Green	Blue	Cyan	Two, Three
Red	Green	Blue	White	One, Two, Three

keywords are selected, the resulting icon colour is based on additive interaction of keyword colour codes, which are present in the selected keyword set. When no keywords are selected, including at the beginning of each trial, cluster icons are coloured black #000000.

Colours of maximum saturation were selected for basic colour matching and combinations; it is desirable for participants to make judgements about colours based on pure colour mixing, rather than mixing of colours that might be labelled something unfamiliar to all participants or shades of colour. For example, ‘what is the result of mixing a salmon pink and Prussian blue’ was to be avoided in preference for ‘what is the result of mixing red and blue’.

Colour combination and mixing is taught to school children using primary colours. Anecdotally, intimate colour knowledge is not common to the mainstream; in contrast, colour experts such as visual artists are more likely to hold the ability to make such judgements of salmon pinks and Prussian blues. Furthermore, in this experiment, the depiction of keywords in a cluster is a binary representation. Cluster icon colour changes to reflect whether a keyword is present or not present, rather than reflecting a fuzzy indication of keyword presence, which might be reflected in a manipulation of the saturation level of the colour.

Ethics Review

A social and behavioural research ethics committee reviewed the experimental design and granted approval for the experiment to proceed. There were no ethical concerns raised throughout the course of this research including neither adverse health effects or breaches to personal privacy. Additionally, there were no other correspondences from any participant, unsolicited or otherwise, other than that collected by the experimental

apparatus. Participants were assumed to provide informed consent having agreed to the terms at the experiment's home page.

4.3.2 Results

The analysis revealed an intermittent recording error for trial 6 and 12. Participants were assumed to have completed all tasks, since the apparatus ensured that at least one answer was selected for each task before the participant could progress further. Further analysis of the apparatus code suggested that a race condition between an asynchronous networking request and the experiment's management code, was responsible for this error. However, this was not expected to introduce bias in either task performance measures or the subjective measures; accordingly, affected trials were marked as missing data and the remaining results from affected participants were retained in the analysis. A consequence of this error was a disproportionate number of cardinality and word count question types in the analysis.

Furthermore, there were no substantial grounds to exclude outliers from the result set. An outlier analysis - by way of box plot inspection - revealed the presence of several modest outliers based on trial time but only for the task set presented first. For these cases, there were high numbers of pop-up windows triggered as well. This suggested that for these cases, participants were actively engaged with the experiment and were not distracted by a parallel task. In view of this, no trials were excluded from the analysis.

Where reported, mean results are reported with 95% Confidence Intervals. A list of the statistical procedures adopted for this analysis are included in Appendix H; for significance testing, the maximum rate of Type 1 error was set at $\alpha = 0.05$.

Task Performance - Time and Interaction Cost

In this section, only results for natural and unnatural conditions are reported. The first objective results relate to the influence of task and task set order to isolate whether tasks in one set were simply easier and therefore faster to complete. There is an expected influence of learning since once participants grasp how the interface works, the tasks should become easier to complete.

Table 4.9 presents tabulated results for mean trial time, pop-up window count, task accuracy score and the number of trials collected for each question type. The unequal number of trials for question types across conditions is attributed to the race condition explained in the introduction to this section. Results for accuracy follow in a subsequent section. Results are reported in detail below.

Participants completed tasks more slowly in the Dog-Train-Security (DTS) task set ($M=71.74$ seconds, $SD=58.77$) than they did in the Australian-Music-Festival (AMF)

Tab. 4.9: Objective performance measures time (in seconds) and number of pop-ups triggered for condition and question type.

Condition	Question Type	μ Time	μ Pop-ups	Correct	N
Control	Topic	84.83	8.90	-	88
	Combined	-	-	-	0
	Words	-	-	-	0
	Cardinality	-	-	-	0
	(Total)	84.83	8.90	(-)	(88)
Unnatural	Topic	92.68	13.58	-	34
	Combined	62.00	18.15	18	40
	Words	50.83	13.35	10	20
	Cardinality	77.42	26.95	3	17
	(Total)	70.73	18.00	(31)	(111)
Natural	Topic	104.87	12.04	-	33
	Combined	54.62	14.12	21	40
	Words	52.52	11.35	11	20
	Cardinality	46.55	13.75	5	15
	(Total)	64.64	12.81	(37)	(108)

task set ($M=67.74$ seconds, $SD=28.31$). In addition, there were more pop-up windows triggered in the DTS set ($M=17.71$ pop-ups, $SD=14.87$) than in the AMF set ($M=12.10$ pop-ups, $SD=5.38$). These results are depicted below in Figure 4.3 and Figure 4.4. Separate one-way analyses of variance ANOVA were conducted for both time and pop-ups as the dependent variables, and task set as the independent variable. A significant effect was found for pop-ups $F(1,19)=4.47$, $p=0.04$ but not for time $F(1,19)=0.14$, $p=0.0.71$.

Further to an effect of task set, a learning effect was observed for task set order. Tasks were completed faster for the task set completed second ($M=59.78$ seconds, $SD=29.31$) compared to the task set completed first ($M=79.69$ seconds, $SD=56.53$). In addition, there were more pop-ups in the task set completed first ($M=17.97$ pop-ups, $SD=14.84$) in comparison to the task set completed second ($M=11.84$ pop-ups, $SD=5.16$). These results are depicted in Figure 4.5 and Figure 4.6. Separate one-way analyses of variance ANOVA were conducted for time and pop-ups for the dependent variables and task set order for the independent variable. A significant effect was observed for pop-ups $F(1,19)=5.60$, $p=0.02$; but not for time $F(1,19)=4.27$, $p=0.05$.

Participants in the natural condition completed the Dog-Train-Security (DTS) task set faster ($M=65.63$ seconds, $SD=40.84$) than participants in the unnatural condition ($M=77.86$ seconds, $SD=74.86$). However, participants in the unnatural condition completed the Australian-Music-Festivals (AMF) task set faster ($M=63.92$ seconds,

SD=28.35) than participants in the natural condition (M=71.55 seconds, SD=29.26). For the DTS task set, participants in the unnatural condition triggered more pop-up windows (M=21.14 pop-ups, SD=19.82) than participants in the natural condition (M=14.27 pop-ups, SD=6.92). Similarly, for task set AMF, participants in the unnatural condition triggered a greater number of pop-up windows (M=12.57 pop-ups, SD=5.71) than participants in the natural condition (M=11.64 pop-ups, SD=5.29). These results are depicted below in Figure 4.7 and Figure 4.8; the red bar corresponds to the unnatural encoding group and the blue corresponds to the natural encoding group. Separate 2x2 mixed factorial analyses of variance ANOVA were conducted with document set as the within subjects factor, encoding condition as the between subjects factor and time and pop-ups as the dependent variables. There were no interaction effects observed, nor any main effect of document set $F(1,18)=0.14$, $p=0.71$ or encoding condition $F(1,18)=0.01$, $p=0.90$ on time. There were no interaction effects observed for pop-ups, no main effect of encoding condition on pop-ups $F(1,18)=0.83$, $p=0.37$; however, a main effect of document set on pop-ups was observed $F(1,18)=4.53$, $p=0.47$.

Participants completing the first task set, regardless of topic, completed tasks faster under natural conditions (M=72.33 seconds, SD=39.89) than under unnatural conditions (M=87.06 seconds, SD=70.96). Conversely, regardless of task set, participants completed tasks faster under unnatural conditions in the task set presented second (M=54.71 seconds, SD=28.90) in comparison to tasks completed under natural conditions (M=64.86 seconds, SD=30.36). In addition, for the second task set, under unnatural conditions, there were fewer pop-up windows triggered (M=11.68 pop-ups, SD=5.60) compared to the natural condition (M=12.01 pop-ups, SD=4.98). However, for the task set completed first, the number of pop-up windows triggered under unnatural conditions was markedly greater (M=22.04 pop-ups, SD=19.38) than those triggered under natural conditions (M=13.90 pop-ups, SD=7.27). These results are presented graphically in Figure 4.9 and Figure 4.10. Separate 2x2 mixed factorial analyses of variance ANOVA were conducted with task set attempt as the within subjects factor, encoding condition as the between subjects factor and time and pop-ups as the dependent variables. There were no interaction effects observed, nor any main effect for encoding condition on time $F(1,18)=0.01$, $p=0.90$; however, a main effect of task set attempt on time was observed $F(1,18)=4.44$, $p=0.04$. Likewise, there were no interaction effects observed for pop-ups and no main effect of encoding condition on pop-ups $F(1,18)=0.83$, $p=0.37$; however, there was a main effect of task set attempt $F(1,18)=6.18$, $p=0.02$.

Focusing on naturalness of encoding overall, the results indicated that participants were practically no faster to complete trials in the natural condition (M=68.61 seconds, SD=31.95) than in the unnatural condition (M=70.32 seconds, SD=46.16). Though participants trigger slightly fewer pop-up windows in the natural condition (M=12.93 pop-ups, SD=5.78) compared to participants in the unnatural condition (M=16.68 pop-

ups, $SD=11.81$). These results are depicted in Figure 4.11 and Figure 4.12. Separate one-way analyses of variance ANOVA were conducted for encoding condition as the between subjects factor and time and pop-ups as the dependent variables. There was no significant effect of encoding condition on either time $F(1,18)=0.009$, $p=0.92$, nor pop-ups $F(1,18)=0.81$, $p=0.37$.

Turning to question type, for topic based questions, under natural conditions, answer time was markedly slower ($M=104.87$ seconds, $SD=39.56$), but triggered pop-up windows were fewer ($M=12.04$ pop-ups, $SD=7.55$), than under unnatural conditions in which answer time was faster ($M=92.68$ seconds, $SD=54.58$) and triggered pop-up windows more numerous ($M=13.58$ pop-ups, $SD=7.24$). For cardinality question types, under natural conditions, answer time was faster ($M=46.55$ seconds, $SD=29.03$), and there were fewer pop-up windows triggered ($M=13.75$ pop-ups, $SD=6.42$) in comparison to unnatural conditions in which answer time was slower ($M=77.42$ seconds, $SD=115.75$) although more pop-up windows triggered ($M=26.95$ pop-ups, $SD=41.77$). For word count question types, under natural conditions, answer time was longer ($M=52.52$ seconds, $SD=47.41$) and pop-up windows were fewer ($M=11.35$ pop-ups, $SD=7.99$) than under unnatural conditions where participants answered in shorter answer time ($M=50.83$ seconds, $SD=42.74$) and with more pop-up windows ($M=13.35$ pop-ups, $SD=11.32$). Finally, for combined word count and cardinality questions under natural encoding conditions, answer time was faster ($M=54.62$ seconds, $SD=38.04$) and fewer pop-up windows were triggered ($M=14.12$ pop-ups, $SD=6.32$), in contrast, in the unnatural encoding condition, answer time was slower ($M=62.00$ seconds, $SD=48.32$) and pop-up windows were more numerous ($M=18.15$ pop-ups, $SD=13.70$). There were no significant differences indicated for combined, word count and cardinality question types.

Topic based questions were not considered further in this analysis, as answering did not involve decoding of visually encoded data. Figure 4.13 and Figure 4.14 depict these results graphically; whilst the time based measure is mixed, under the unnatural encoding condition, participants require more pop-up windows to complete tasks.

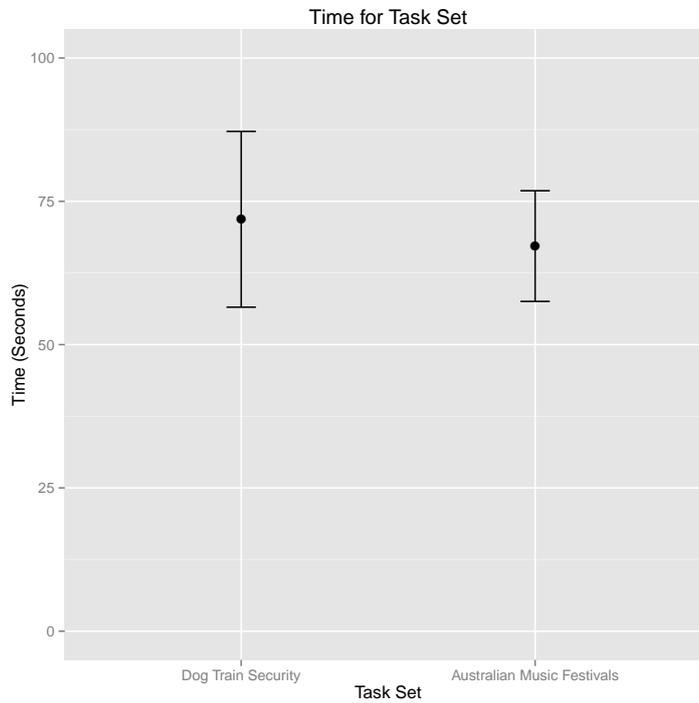


Fig. 4.3: A graph of time (in seconds) for task set; error bars are 95% Confidence Intervals.

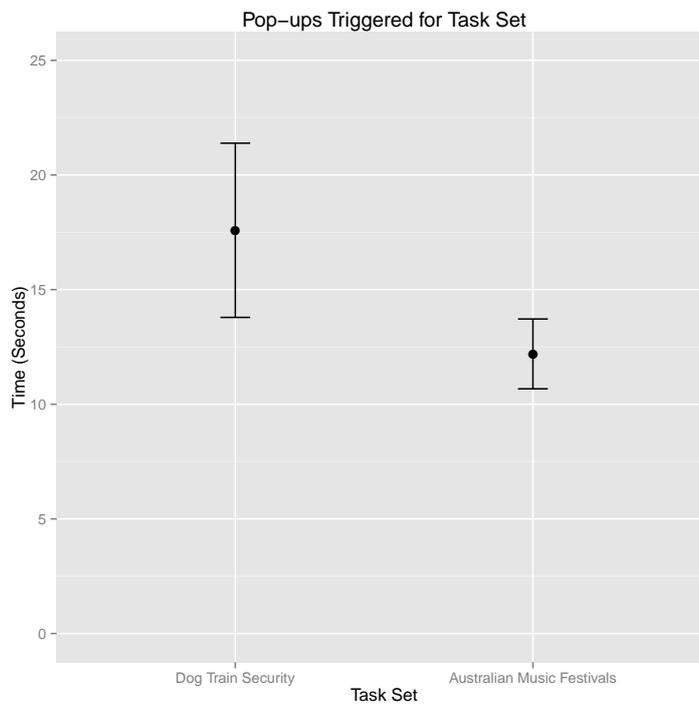


Fig. 4.4: A graph of pop-ups triggered for task set; error bars are 95% Confidence Intervals.

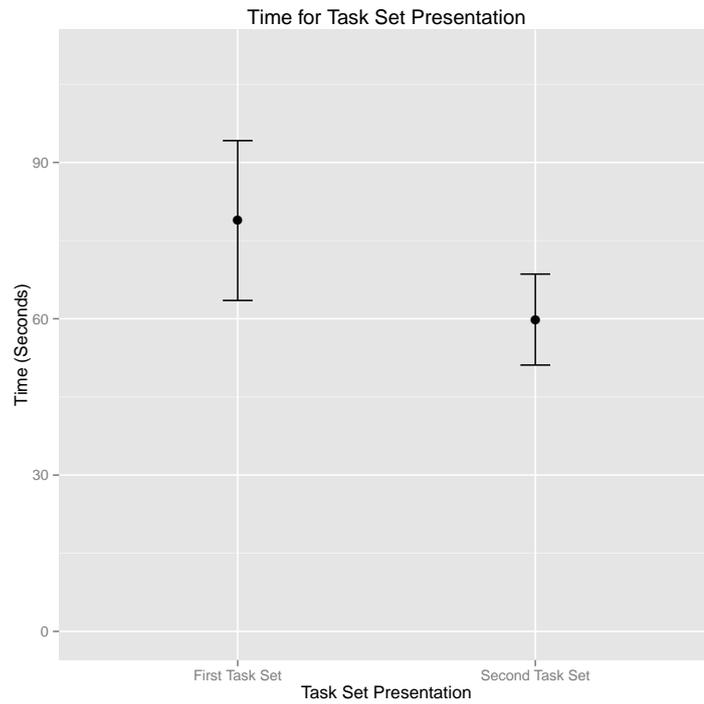


Fig. 4.5: A graph of time (in seconds) for task set presentation; ‘First Task Set’ corresponds to the first task set attempted, regardless of the topic of the task set; error bars are 95% Confidence Intervals.

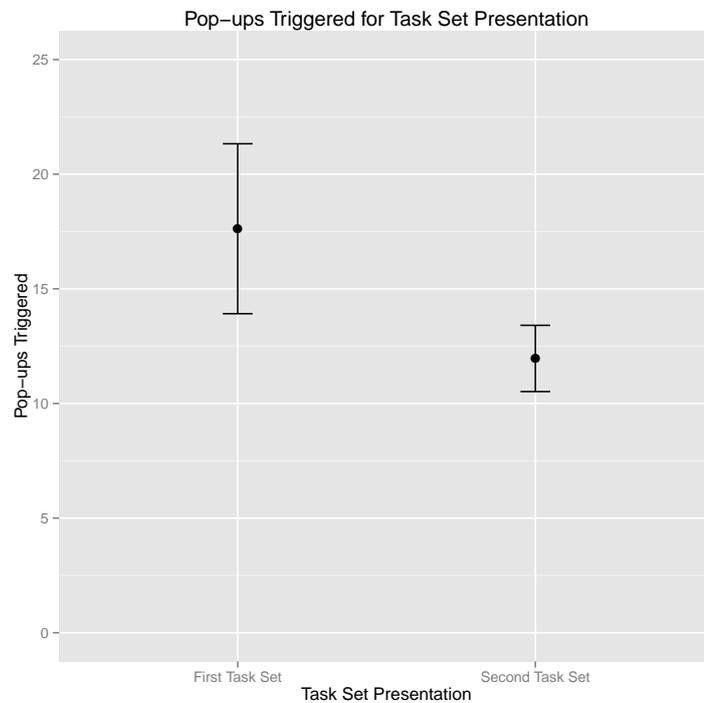


Fig. 4.6: A graph of pop-ups triggered for task set presentation; ‘First Task Set’ corresponds to the first task set attempted, regardless of the topic of the task set; error bars are 95% Confidence Intervals.

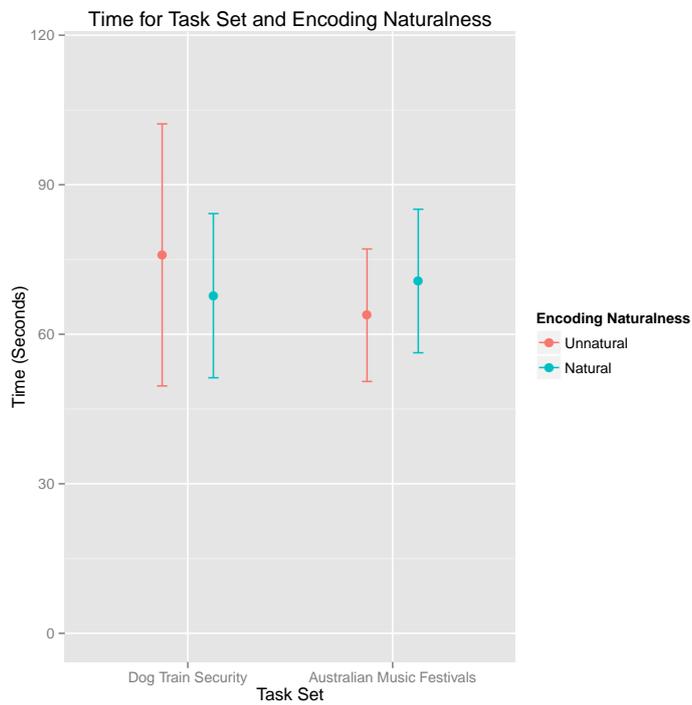


Fig. 4.7: A graph of time (in seconds) for task set and naturalness of encoding; error bars are 95% Confidence Intervals.

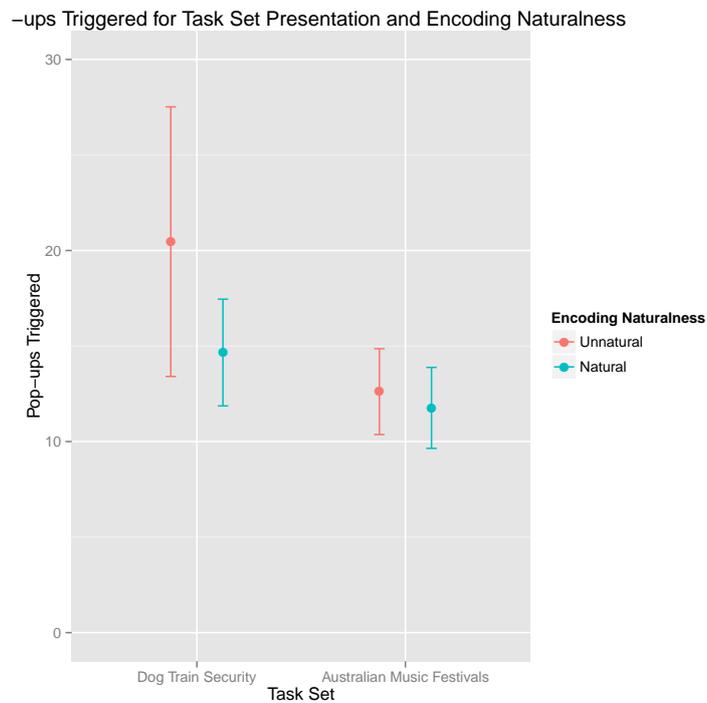


Fig. 4.8: A graph of pop-ups triggered for task set and naturalness of encoding; error bars are 95% Confidence Intervals.

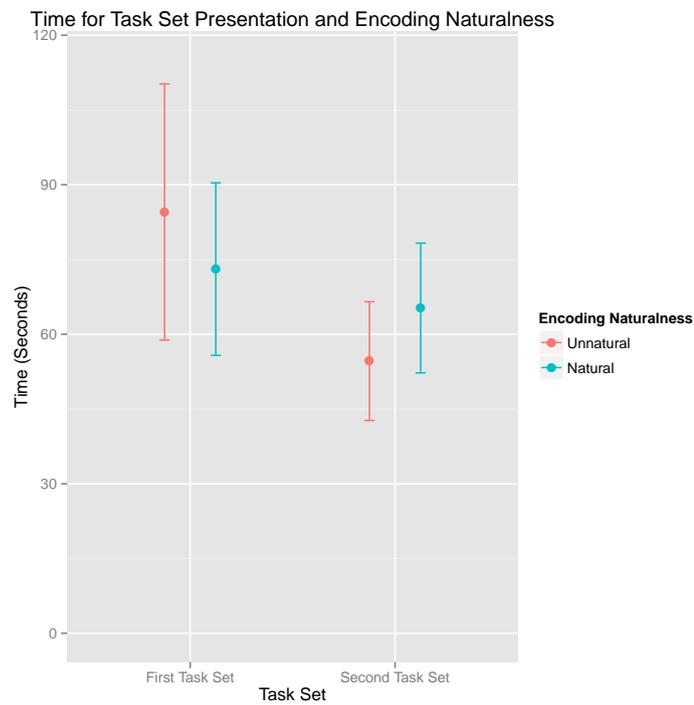


Fig. 4.9: A graph of time (in seconds) for task set presentation and naturalness of encoding; error bars are 95% Confidence Intervals.

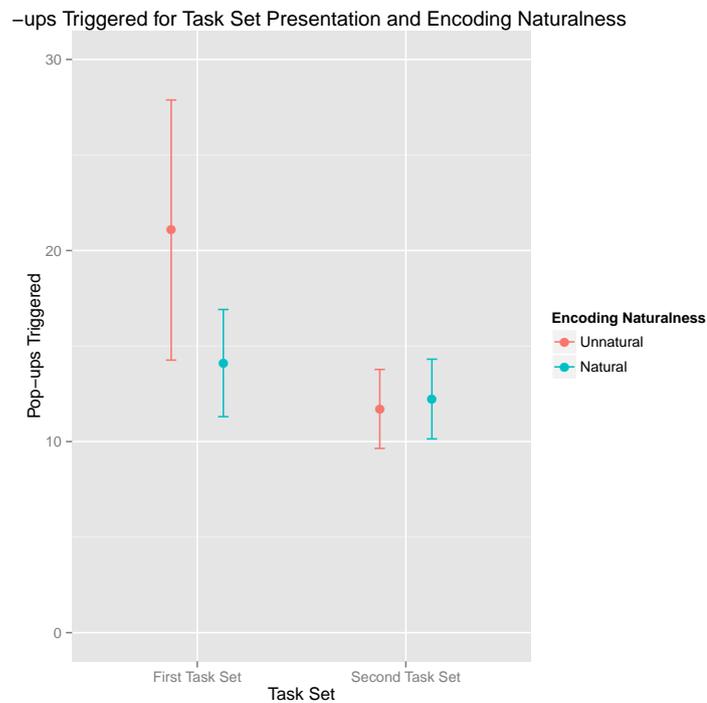


Fig. 4.10: A graph of pop-ups triggered for task set presentation and naturalness of encoding; error bars are 95% Confidence Intervals.

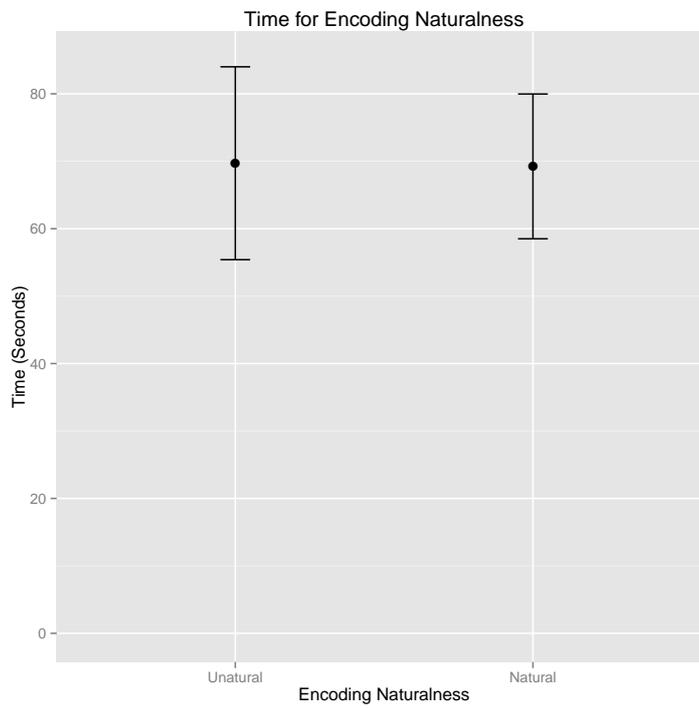


Fig. 4.11: A graph of time (in seconds) for naturalness of encoding; error bars are 95% Confidence Intervals.

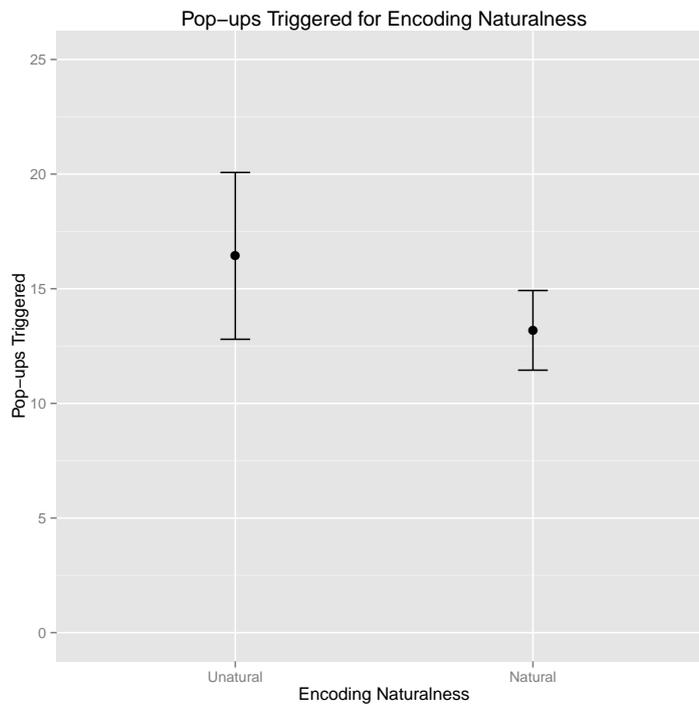


Fig. 4.12: A graph of pop-ups triggered for naturalness of encoding; error bars are 95% Confidence Intervals.

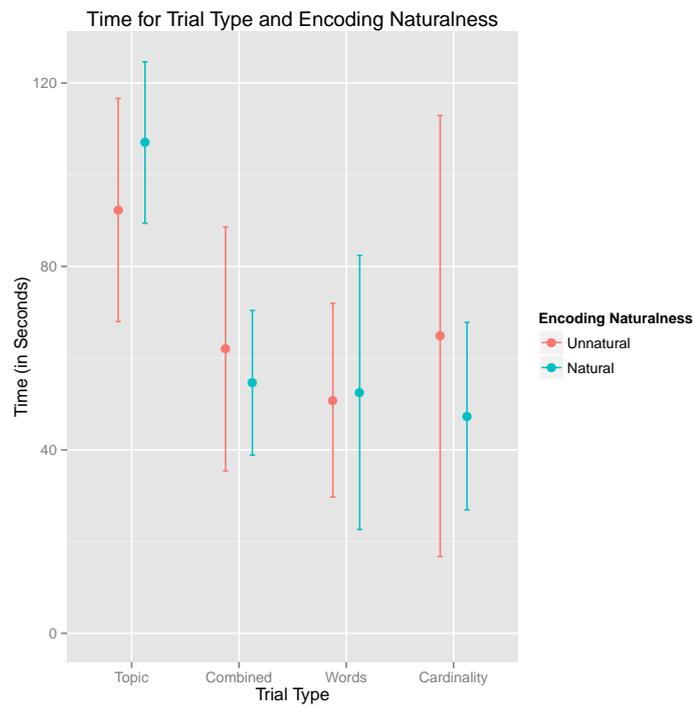


Fig. 4.13: A graph of time (in seconds) for trial type and naturalness of encoding; error bars are 95% Confidence Intervals.

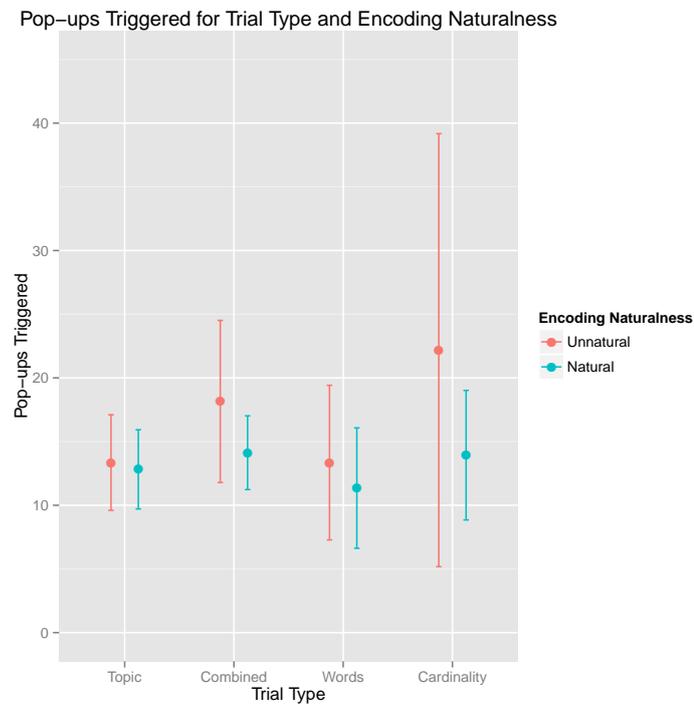


Fig. 4.14: A graph of pop-ups triggered for trial type and naturalness of encoding; error bars are 95% Confidence Intervals.

Task Performance - Accuracy

The next results relate to participant accuracy answering the trial questions. For this analysis, only question types relating to words, cardinality or a combination were relevant. Accuracy of topic type questions is of limited relevance for the current research hypotheses and so are not reported. As depicted in Table 4.10, overall accuracy was quite poor. In almost all cases, the number of incorrect answers and correct answers was approximately the same, although generally, correct responses outnumbered incorrect responses by a slim margin. Furthermore, cardinality-based questions under natural conditions were answered markedly better in comparison to the same questions under the unnatural condition. Furthermore, the margin between incorrect and correct responses was moderately larger under natural conditions, suggesting that participants were slightly more successful overall under the natural condition.

Tab. 4.10: Accuracy over condition and question type.

Condition	Question Type	Incorrect	Correct	Total
Unnatural	Combined	21	19	40
	Words	10	10	20
	Cardinality	8	9	17
	(Total)	(39)	(38)	(77)
Natural	Combined	18	22	40
	Words	9	11	20
	Cardinality	3	12	15
	(Total)	(30)	(45)	(75)

To drill down further, correct and incorrect responses are depicted in Table 4.11. There is a stark difference of task accuracy for combined question types across each data set and this pattern was maintained across encoding conditions. Under unnatural conditions, 4/20 responses were correct for combined metadata questions in the Dog-Train-Security (DTS) task set, in contrast, 15/20 responses were correct for the combined metadata questions in the Australian-Music-Festivals (AMF) task set. A similar pattern was evident under natural conditions; however, participants answered combined questions more successfully across both task sets: 6/20 in the DTS task set and 16/20 in the AMF task set. A similar disparity though in reverse, was observed for the word count question types. Under unnatural conditions, 8/10 questions in the DTS task set were answered correctly, while only 2/10 were answered correctly in the AMF task set. This pattern was more or less the same under the natural condition: 3/10 trials were correct in the AMF task set while 8/10 were recorded correct for the DTS task set. Finally, for cardinality questions, under unnatural conditions, 6/10 were answered correctly in the DTS task set but 3/7 answered correctly for the AMF task set. Under natural conditions, the pattern was broken and the same proportion of

Tab. 4.11: Accuracy for condition, task set and task type. Low totals for cardinality in the AMF task set is reflective of an intermittent data recording issue.

Condition	Set	Question Type	Incorrect	Correct	Total
Unnatural	DTS	Combined	16	4	20
		Words	2	8	10
		Cardinality	4	6	10
	AMF	Combined	5	15	20
		Words	8	2	10
		Cardinality	4	3	7
Natural	DTS	Combined	14	6	20
		Words	2	8	10
		Cardinality	2	8	10
	AMF	Combined	4	16	20
		Words	7	3	10
		Cardinality	1	4	5

correct answers was observed: 8/10 for the DTS task set and 4/5 for the MFA task set. The smaller number of cardinality questions for the AMF set is attributed to the intermittent data recording issue.

Overall, 45/75 questions (60%) were answered correctly under natural conditions, while 38/77 questions (49%) were answered correctly under unnatural conditions. One participant answered all questions incorrectly under unnatural conditions and no participant answered all questions correctly. Only one participant in the natural condition achieved the high score of 7/8 correct answers.

A Fisher's Exact Test was conducted for results in Table 4.10. There were three contingency tables analysed: a condition contingency, a set contingency and a question type contingency. Fisher's exact test offered no evidence to suggest that the naturalness hypothesis had an effect on accuracy ($p=0.26$) and the same possibly so for question type ($p=0.06$); however, there was evidence to support the notion that document set had an effect on accuracy score ($p=0.001$).

Learning Outcome

Learning outcome was measured by open-ended qualitative responses. There were three enquiries of the participant's understanding of colour, shape and size coding and each are outlined in Table 4.12. Despite an open-ended answer style, there were only two marginally correct or ambiguous answers; answers were either clearly incorrect or correct, otherwise. Table 4.13 illustrates the diversity in incorrect responses. Significance testing was not conducted due to the sparseness of the observations.

Tab. 4.12: Correct encoding recall for condition; this reflects a user’s understanding of colour, shape and size encoding in the interface.

Condition	N	Colour	Shape	Size
Control	10	7	0	-
Unnatural	10	7	2	4
Natural	10	8	0	3

Tab. 4.13: Diversity and frequency of incorrect responses to recall question for condition.

Condition	Response	Colour	Shape	Size
Control	No Idea Didn’t Understand	3	8	-
	Outline	-	-	-
	Didn’t Use or Notice	-	-	-
	Category	-	2	-
Unnatural	No Idea Didn’t Understand	1	2	1
	Outline	2	1	-
	Didn’t Use or Notice	1	-	1
	Category	-	2	1
	Word Count	-	2	(Correct)
	Cardinality	-	(Correct)	3
Natural	Thought Random	-	1	-
	No Idea Didn’t Understand	1	8	4
	Outline	1	-	-
	Didn’t Use or Notice	-	-	-
	Category	-	2	-
	Word Count	-	(Correct)	3
	Cardinality	-	-	(Correct)
Thought Random	-	-	-	

Generally, participants understood how the colour coding scheme worked but, were less certain about size encoding and quite uncertain about shape encoding. Colour coding was recalled well but, not perfectly recalled as was expected. Two participants mistook the visual feedback - a red outline indicating icons that were selected as answers - as the only colour coding in play. This also reflected the fact that those participants did not utilise the keyword tree at all, suggesting that they accessed all of the required information by looking at pop-up window content and a brute force approach to answering questions. The remainder of incorrect responses were *no idea* or *did not use*. With regard to task accuracy, the average number of questions answered correctly was higher for participants who reported the colour coding incorrectly (task accuracy 58%)

in comparison to those who answered the colour coding correctly (task accuracy 51%).

In contrast, shape was very much unnoticed; only two participants in the unnatural encoding condition could report the relationship between cardinality and shape and no participant could determine the relationship between word count and shape under the natural encoding condition. Furthermore, exactly two participants per condition reported shape pertaining to category or thematic content. In relation to average participant score, the average number of questions answered correctly was higher for those who answered the shape question correctly (task accuracy 54%) than those who answered it incorrectly; however, to reiterate, only two participants reported the shape encoding correctly.

Participants were not required to report the encoding of size in the control condition. Size was reported more frequently in the unnatural encoding condition; however, in both cases, an ambiguous response precluded any clear determination of the superior case. Furthermore, with regard to average participant score, participants who reported size encoding accurately, had a markedly higher average score (task accuracy 66%) than those who reported size incorrectly (task accuracy 47%).

Additionally, eight responses, five in the unnatural condition and three in the natural encoded condition, assigned the opposite encoding rule i.e. attributed word count to icon size, when in fact icon size encoded cardinality. Accordingly, due to such diversity of incorrect responses, incorrect responses have been categorised and reported in Table 4.13.

Exit Questionnaire

For this analysis, questionnaire responses were reorganised into three categories: usability questions, task understanding questions, and opinion-based questions. For many of the questions, participants were divided in agreement and cohesive responses were only evident in a handful of the questions. The subjective response results are presented in Table 4.14 and Table 4.15.

Responses were re-coded to a three level scale: agree, neutral and disagree; re-coding isolated three main groups those that agree regardless of strength, those that disagree regardless of strength and those that are undecided; participants who make a parity judgement provide a stronger signal from which to draw conclusions. A position that is clearly negative or positive is more insightful than a position of indecision. Therefore, if the mode value of the combined agree responses or disagree outnumbers the undecided, then this is an important finding.

A non-parametric independent samples Mann-Whitney U test of significance using naturalness condition as the grouping factor, found no significant differences of distribution between groups or median response value except in the case of *Q5* *There*

were keywords missing that would have been good to answer the task/questions which reached significance $U(18)=23.5$, $Z=-2.091$, $p<0.05$. In this question, participants in the control reported resoundingly that there were additional keywords - that were not present - which they wanted to use to answer the task questions. Conversely, participant responses in the natural/unnatural condition indicated no clear agreement between participants i.e. almost equal numbers of participants agreed, disagreed or were neutral toward Q5. This finding is understandable, given the greater prevalence of topic-based question types - but no emphasis will be placed on the significance of this finding.

Given the non-significant differences between experimental groups and the remaining questions, results are reported as a whole and not from the perspective of specific groups. The results presented in Table 4.14 and Table 4.15 are organised by question type for reporting purposes with opinion based questions appearing in Table 4.14 and usability questions appearing in Table 4.15. Questions were presented to the participant in the order specified by numbers in the second column. For each question, the experimenter's preferred response, the observed mode response and a breakdown across response groups is tabulated. A conflict between preferred response and observed mode response is taken to mean an undesirable problem with the interface; there were six conflicts observed and these are highlighted in the discussion.

The first set of questions relate to the participants' opinion of search and of the hypothetical role that this experimental interface could play in the participants' future search activities. While participants largely favour looking through lists of search results (Q8), they thought clustered search results (Q9) would help with information search and indicated that the experiment apparatus would not be as useful if it displayed exclusively documents, and not clusters (Q13). Participants were averse to using a version of the experiment apparatus for their day-to-day search (Q2) and did not find the design appealing (Q17). However, if the interface was improved aesthetically, most indicated that they would use a tool similar to this apparatus for their everyday search (Q22). In addition, most participants suggested that they would trust results more in a list format than in the format presented by the experiment apparatus (Q21). Finally, very few participants indicated that they would *not* refer the experiment to a friend and a clear majority indicated that they would (Q23).

The second set of questions related to the task, and were asked in order to ascertain whether they understood the task. Most participants indicated that they thought the apparatus was simple and straight forward (Q1) however; the breakdown suggests that a sizeable number thought the opposite. In favour of those who did not think the interface was straight forward, a majority reported that they did not understand how the experiment worked and that they were not confident using the apparatus (Q3). Again, the divide is evident in responses relating to how easy the interface was to use (Q15) but overall, a majority did report that it was easy to use. Slightly more participants indicated that there was not sufficient information available to complete

Tab. 4.14: Subjective response data sorted by question type; desired response denotes the mode response achieved by a superior experiment and apparatus design, while observed response denotes the observed mode response - a disagreement between desired and observed response is indicative of poor apparatus or experiment design

Type	N	Question	Desirable Response	Observed Response	(A)gree	(N)utral	(D)isagree
Opinion Based	8	I prefer to look through lists of search results to find my documents	D	A	46.67%	30.0%	23.33%
	9	Search engines that present results as clusters would help me with my information search	A	A	66.7%	20.0%	13.3%
	13	This tool would be better if it only showed documents and not document clusters	D	D	16.67%	36.67%	46.67%
	2	I would use a tool like this for my day-to-day information search	A	D	16.67%	36.67%	46.67%
	17	The design of the interface was appealing	A	D	30.00%	16.67%	53.33%
	21	I would trust results in a list format more than in the visualisation format in this experiment	D	A	50.0%	26.67%	23.33%
	22	If improved aesthetically, I would use a tool such as this for my day-to-day search	A	A	53.33%	16.67%	30.00%
	23	I would recommend this experiment to a friend	A	A	50.00%	46.67%	3.33%

the tasks (Q18) but responses reporting on the provision of good keywords (Q5) may explain this.

The third set of questions related to the usability of the interface. These were asked to make sure participants' performance and opinions regarding functionality were not adversely biased by poor usability of the interface. The subjective responses indicated that overall usability was sound. Participants responded that the keyword tree was a suitable way to organise keywords (Q4) but that there were some keywords missing that would have been useful (Q5). The results for (Q4) are echoed by those in (Q6), which asked the same question regarding the appropriateness of the keyword arrangement - keywords were presented in a way conducive to completing experiment tasks.

For the remaining usability questions, responses were encouraging: mouse interaction was not overly burdensome, the size of the text was adequate for reading, the level of information presented was not too great and the keywords were representative of the search result set. However, participants were unsure of the trustworthiness of the keywords presented.

Drop Out Analysis

Following Reips, 2002, this section reports the prevalence of experiment drop out. This analysis sought to isolate the conditions under which participants did not want to proceed or could not proceed further in the experiment. For this analysis, the total participant pool consisted of all participants who submitted demographics information. Each point on the curve in Figure 4.15 represents the number of participants attempting each stage of the experiment.

All participants completed the demographics questionnaire; however, ten participants left the experiment shortly after, at the instruction stage. Immediately, this suggests that the *big wall* of instruction text may have exceeded the effort allowance of some participants.

Further remarkable drop out occurred at the first few experiment tasks - three participants did not complete the first task and five dropped out without completing a second task. By the end of the first set of six questions, most participants who dropped out, had already done so. One participant abandoned the experiment at the beginning of the second task set, while two participants dropped out right at the end when presented with the questionnaire - the server logs indicated that the questionnaire page was requested. At this point, these participants either had a technical problem or had extinguished their altruistic allowances, precluding them from completing the exit questionnaire.

The descending drop out curve indicates where participants drop out and one can only speculate as to why they left the experiment at each point. Ultimately, this

Tab. 4.15: Subjective response data sorted by question type; desired response denotes the mode response achieved by a superior experiment and apparatus design, while observed response denotes the observed mode response - a disagreement between desired and observed response is indicative of poor apparatus or experiment design.

Type	N	Question	Desired Response	Observed Response	(A)gree	(N)utral	(D)isagree
Task	1	The experiment interface I have just used is simple and straightforward	A	D	43.33%	10.00%	46.67%
	3	I clearly understood how the experiment worked and was confident using the apparatus	A	D	33.33%	20.00%	46.67%
	15	The interface was easy to use	A	A	43.33%	13.33%	43.33%
	18	All the information was present to complete my tasks	A	D	30.00%	30.00%	40.00%
Usability	4	The keywords were arranged in a way that made it easy to find desirable keywords	A	A	53.33%	6.67%	40.00%
	5	There were keywords missing that would have been good to answer the task	D	A	56.67%	23.33%	20.00%
	6	The Explorer Tree was not useful for finding document clusters of requested topics	D	D	30.00%	13.33%	56.67%
	7	The Explorer Tree could be useful for my day-to-day information search	A	A	60.00%	30.00%	20.00%
	10	The mouse over functionality was annoying and time consuming	D	D	26.67%	36.67%	36.67%
	14	The size of the text made it easy to read	A	A	73.33%	13.33%	13.33%
	16	There was too much information displayed	D	N	26.67%	36.67%	36.67%
	19	The keywords were not reliable or representative of the search result set	D	D	16.67%	33.33%	50.00%
20	The keywords were trustworthy	A	N	43.33%	50.00%	6.67%	

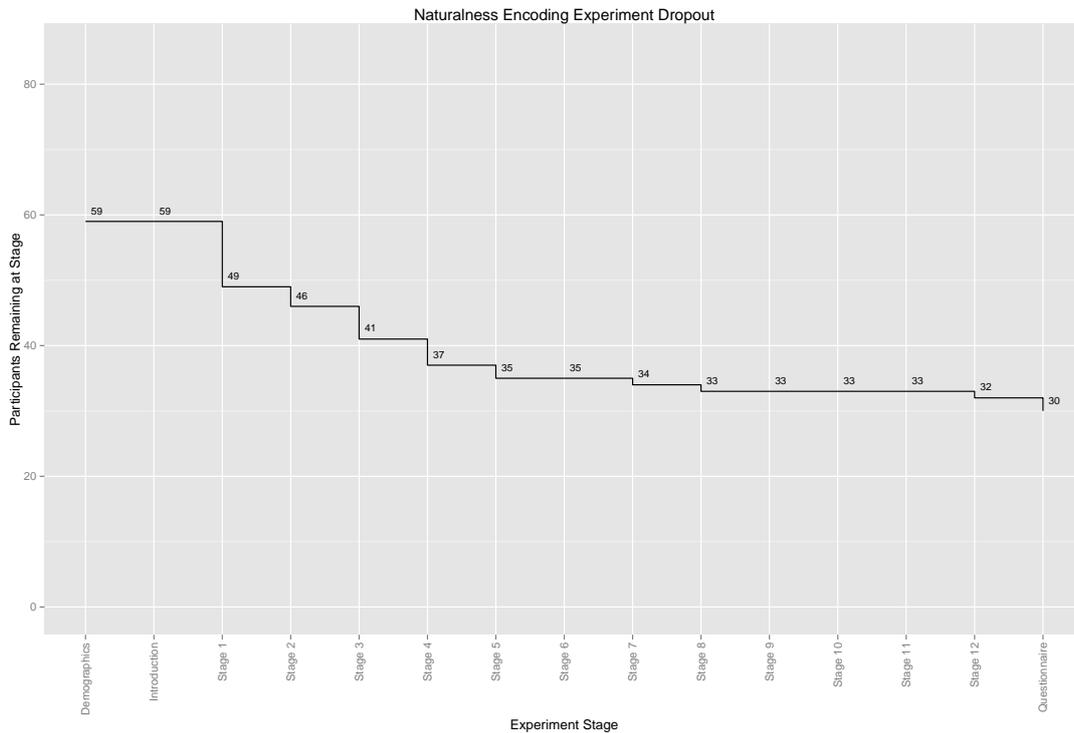


Fig. 4.15: A graph of participant drop out for experiment stage. Values indicate the number of participants remaining at that stage.

graph offers little certainty as to why participants drop out, therefore, the full set of demographics data were analysed for potential insight into the subset of participants that did not complete the experiment.

An analysis of demographics indicated very little difference between those that drop out and those who complete the experiment in full. All participants except one, cited having spoken English for their entire life or for a very large portion of their life. One participant indicated that they had only spoken English for one year, and while that participant did drop out, a language barrier cannot account for the remaining withdrawals. In addition, experience with graphics design, mouse handedness, age, gender or geographical location does not show any obvious pattern.

4.3.3 Discussion

Overall, the results are starkly inconclusive with regard to the role of naturalness of encoding paradigms in visualisation-based search result tools; however, the results are exploratory and limited by a small sample size. Whilst the experimental data do not yield the anticipated result, it is envisioned that they will, in the least, trigger further examination of natural encoding paradigms in the near future

While not statistically significant, the results indicated that under natural encoding, objective measures like task accuracy and number of interactions were marginally

better, although self-reported learning outcome indicated that participants may not have been conscious of the fact. In all, an unconscious understanding of an interface might not necessarily be such a bad thing if we can perform tasks quickly, accurately and with optimal levels of effort.

An analysis of task set revealed that although participants spent the same amount of time answering questions for either task set, there was a significant effect found for the Australian-Music-Festivals (AMF) task set and pop-ups. Participants triggered significantly more pop-up windows in the Dog-Train-Security (DTS) set than the AMF task set; indicating that questions may have been more challenging in the DTS task set. Furthermore, a significant effect was observed for presentation order. Participants triggered significantly more pop-ups when answering questions on the task set completed first in comparison to the task set they complete second. This is expected; on presentation of a new interface, we naturally take time to explore it and to figure out how it works. In addition, the initial performance cost is likely due to the purposely-inconsistent training information and the appearance of the interface; participants were not informed about the size and shape encoding in the training material and on presentation of the interface in the experiment, participants may have immediately thought to establish the connection of the encoding paradigm in use. This trend is only present in the interface presented first. Nevertheless, what is most important to note is that interaction measures are higher when the encoding is unnatural. This suggests that the initial learning phase of the interface is made more difficult if encoding is not natural. Furthermore, in the second task set, the time to complete is more or less the same; by this point, participants under unnatural conditions may have been successful in holding their unconscious inclinations at bay, in order to complete tasks. Above all, the results in Figure 4.10, while not statistically significant, suggest that on trend, natural encoding may have a positive benefit to the number of interrogations of a visualisation - measured as pop-up window count.

Whilst future experimentation should remain opposed to cueing participants to the encoding scheme in use, a future experiment should take into consideration the influence of training material on the outcome. This experiment demonstrates that even if the training material explains some aspects of the interface, such as how the colour coding works, there are no guarantees that participants will accurately recall it as demonstrated by the less than perfect scores on participant responses to the question: *please explain how the colour coding works*.

With regard to hypothesis one, the analysis did not reveal a significant overall influence of encoding naturalness on time to complete tasks; on average, participants completed tasks in more or less the same amount of time regardless of encoding naturalness. In contrast, there was an apparent trend of encoding naturalness on the number of pop-up windows triggered; typically, fewer pop-up windows were triggered under the natural encoding paradigm and moreover, there was greater variation between partici-

pants under the unnatural condition as indicated by the wider error bars. In addition, despite the markedly poor accuracy utilising this interface under either condition, on tasks necessitating a consideration of cluster cardinality and word count, participants were more accurate under the natural encoding condition (45/75) in comparison to participants under the unnatural encoding condition (38/77). Most of this advantage can be attributed to the markedly superior performance on cardinality questions, however participants were consistently, albeit modestly, more accurate under the natural encoding condition.

With regard to hypothesis two, the influence of encoding naturalness may be confounded. Fewer participants correctly reported how colour coding worked under the unnatural condition, in comparison to reporting performance under the natural encoding condition - and there is no clear explanation for this. Colour coding is unrelated to shape and size encoding and furthermore, colour coding was explained in the training material. Thus, one would expect a near perfect score for the role of colour encoding. But, even under the control condition, where participants were expected to rely on the colour coding even more so, perfect recall of the colour encoding was not observed. Moreover, participants under unnatural encoding more readily and correctly reported the role of shape coding and reported size encoding commensurate with participants under the natural encoding. Under the natural encoding condition, participants more readily and accurately reported the role of colour coding, but like the control condition, in which there were no questions cueing participants to the encoding in use, there were no participants that could correctly report the meaning of shape.

Therefore, has poor general knowledge of colour mixing theory impacted on a participant's ability to focus on a subset of the cluster set and upon which to cast metadata judgements? If so, this seemingly innocuous facility, to highlight the presence of key terms in a cluster, may have been used incorrectly, thereby setting the participant up for task failure. For instance, in tasks specifying two semantic criteria - e.g. find the smallest cluster with *dog* and *dog training* - participants may have considered clusters containing the term *dog* only as denoted by say red encoding; and clusters containing *dog training* as denoted by say blue encoding - and not clusters containing both terms as denoted by the combination of red and blue resulting in a magenta encoding.

It is of concern that some participants could not even recall the presence of a colour coding scheme; specifically, the 4/30 participants who responded *Outline* or *Did not Use/Notice*. This suggests that they did not use the keyword tree as instructed to do so, which may have been a result of poor engagement or a lack of instruction material absorption. On further review of the results, with the exception of two, participants who did not provide a valid response to colour, also did not provide a response to shape or size. Of the 2 participants who did not provide a valid response to the colour, but did provide a valid response to shape and/or size, one appears not to have interacted with the keyword tree at all and only reported the red colour bar below the answer icon,

while the second responded cryptically with *it does not*, despite giving a valid response for size. The latter suggests either something went wrong with data recording or the participant lost their train of thought or they were being deliberately facetious. There is however, an issue with the 6/30 participants who *had no idea/did not understand*. This equates to either a lack of training support or poor colour-mixing general knowledge in the participant pool - or both.

In a future experiment, a refresher course and a pre-test of colour mixing theory should be included in the experiment introduction phase and the legend design could be overhauled to reflect the proposal in Figure 4.16. Such a legend would not only show which colour corresponds to which word, but also reinforce the colour combinations formed with different words. Alternatively, simply including a colour mixing guide such as that depicted in Figure 4.17 may help. Subsequently, with the aforementioned additional training and assistance, participants being unable to answer a question about how the colour coding works would provide sufficient reason to exclude them from the analysis i.e. a trap event.

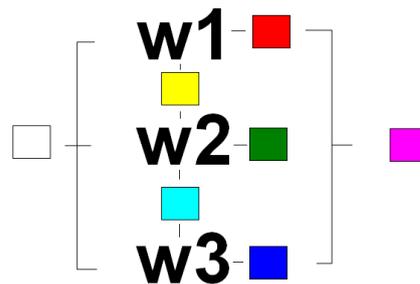


Fig. 4.16: A proposal for future encoding legend; this design incorporates the colour mapping and colour mixing outcomes.

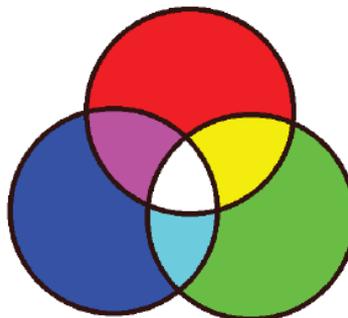


Fig. 4.17: A colour mixing guide.

Presently, the colour coding legend in this experiment, provides only colour coding information per selected word. It remains the task of the participant to inspect each cluster icon and discern whether or not each colour code is composed of a proportion

of the target colour. It is not immediately apparent - at a glance - which of the colour coded cluster icons that contain a proportion of a target colour. For instance, it is not immediately apparent that a proportion of red is present in yellow, since we perceive yellow holistically as a combination of red and green Ware (2004). Moreover, with an increasing number of colour codes in use, and an increasing number of cluster icons present, colour heterogeneity is increased thereby decreasing the likelihood of identifying any one target colour at a glance (Wolfe, 2007).

One possible reason for the poor performance may be rooted in the observation by Ware (2004) that hues cannot be separated into constituent primary colours pre-attentively and in parallel, however, it may not necessarily factor into consideration here. Such a claim would perhaps be valid if all participants could show mention of the general idea of colour encoding. Regardless of the efficiency, participants should have been able to gauge that there was a relationship between the colour coding and the keyword content contained within. Poor efficiency may well be observed if the participant can describe what it is the colour coding means; however, this claim cannot alone explain why the accuracy on self reported learning was low; it is a barrier to the true estimation of the experiment's hypotheses and it is appealing to believe that it is reflective of the participant's poor absorption of training material.

One final point regarding colour coding can be made in regards to colouring of both inner and outside icon regions. In 6 out of 49 cases of inner colour and outer colour coding - e.g. when the inner is coded red and outer is coded red - a participant is unable to notice a difference between inner and outer colour coding - without a particularly careful focus - and may in fact overlook a cluster containing the first and fourth selected words for highlighting. Clusters containing all six selected words are the exception to this issue; since a black border is drawn around the cluster's inner and outer region as initially it was clear that a pure white icon would be impossible to perceive - given the white background. A future experiment would seek to demarcate the inner and outer region of the colour coding more comprehensively.

Remaining on discussion with hypothesis two - though leaving aside the difficulties introduced by colour coding - ambiguous self-report answers for size encoding, meant that no clear advantage for naturalness of encoding could be determined. One participant reported size and shape to mean the same thing under the unnatural condition - they were marked incorrect for the shape encoding, however, they were marked correct for the size encoding. Furthermore, whilst one participant attributed size encoding to *more content* under the natural condition, it is unclear whether the participant meant by *content* as: more *words* or *documents*. In any case, if the numbers in the size column of Table 4.12 for the unnatural and natural conditions were swapped around, the advantage for natural encoding would remain unconvincing i.e. a difference of one. Interestingly however, there was greater diversity in incorrect answers provided for the explanation of shape under the unnatural condition. Under the natural condition,

8/10 responses were simply *no idea* or *did not understand*, while the remaining two responses attributed shape to category. Similarly, under the unnatural condition, 2/10 participants attributed shape to category, but 2 attributed shape to word count when in fact they should have attributed shape to cardinality; the remaining four incorrect responses indicated they did not understand or did not assign significance to shape. In relation to incorrect responses for size, three participants attributed size to word count or size to cardinality when they should have reported size to cardinality and size to word count for natural and unnatural conditions respectively. However, under the natural condition, participants either answered correctly or did not understand; in contrast, under the unnatural condition, participant responses were more mixed including one response that attributed size to category. In future, a multiple choice questionnaire format may alleviate some of the ambiguities encountered during analysis of the participant's understanding of the encoding, however, in this particular case, the open response format has captured responses that may not have been collected by a multiple choice questionnaire.

If the objective measures somewhat favoured the natural encoding, beyond simply a favourable trend, why do participants perform badly on self-reported learning? Moreover, since objective performance measures were advantageous, is it such a bad thing that participants cannot report how the interface works? A possible explanation for the first question is that participants may not be consciously aware of the choices they make when choosing icons in the interface to interrogate. If without awareness, a participant orientates their attention to the biggest icon when cued to locate the biggest cluster - a natural mapping - only later to locate a smaller icon representing larger cluster cardinality based on an observation of the cluster's pop-up, then it would take a conscious effort to re-evaluate their understanding of the interface. Furthermore, on needing to re-evaluate their expectations and understanding of the interface, when polled to explain how the interface works, they would have their consciously activated mental-model in recent memory. In contrast, for those participants that completed the experiment without having their expectations and assumptions violated, largely no additional learning to counter a violation of assumption was required and no conscious use of the interface took place; consequently, when polled to explain how the interface worked this group does not have a freshly activated mental-model because there was never a need to consciously modify one.

With regard to hypothesis three, there were no significant differences detected between natural and unnatural encoding groups and subjective response. However, there were strong opinions voiced by participants overall. While usability aspects of the interface were sound, participants were unsure of keyword utility and trustworthiness - although, in this particular case, perfect utility and trustworthiness is largely unattainable without offering a facility for participants to express their own queries. Furthermore, participants did not find the interface appealing and signalled that they would

not utilise a similar interface for their everyday search. However, this is expected, as there was no facility to access individual documents or to submit queries ad hoc.

Furthermore, nearly half of all participants indicated that they did not understand how the experiment worked and were not confident using the apparatus. A third of participants indicated that they understood how the experiment worked while the remainder were undecided; this suggests that the instruction and training materials were insufficient. Moreover, these observations are complemented by the analysis of drop out which showed that over ten percent of the potential participant pool exited at the training stage of the experiment. Consequently, it is possible that a proportion of participants devoted less than sufficient effort toward absorbing the training material, even though still managing to finish the experiment; the less than perfect self-reported learning results for colour coding may be an indicator of this situation.

More broadly, an analysis of drop out indicated three main areas for improvement: simplify the instructions and make it easier to convey training material, go to greater lengths to describe the importance of the research and finally, to reconsider or reformat - overly long - subjective response questionnaires at the end of the experiment. Furthermore, and in the least, it is likely that a tangible reward may have encouraged more enthusiastic and complete participation.

An immediate fix for the *big wall of text* training format may be to provide video-based instruction to reduce cognitive overhead. However, research indicates that the assumed benefits of video-based instruction do not necessarily realise. Studies of Choi and S. Johnson (2005) and Breimer, Cotler, and Yoder (2012) find no significant differences on task performance between groups trained on text-based material versus groups trained on video-based material for a web-based learning module, even though Choi and S. Johnson (2005) observes significantly more attentiveness to the video-based instruction. These findings suggest that making it easier to attend to inherently difficult or cryptic training material will not improve participant understanding, rather either mode of instruction will be sufficient if the material is inherently easy to understand.

The inconclusive results may also suggest the presence of confounding factors. One such factor may be the use of the keyword tree since the use of the keyword tree may reduce the number of alternatives down significantly, leaving the user to trigger a small number of pop-ups and arrive at the answer; from this perspective, there is no incentive for participants to decode the visual variables. However, the keyword tree invites a degree of realism to the experiment task. In document search, we restrict ourselves to a subset of the search result set, since often - anecdotally - we can easily recognise the blatantly irrelevant documents with ease unlike the effort devoted to recognising partially-relevant from relevant documents. Labelled clusters can help us in this respect since irrelevant labels stand out as plainly irrelevant. In this experiment, the participant made keyword tree interactions that set the focus points for search, effectively drawing attention to a subset of the result set. Whilst descriptive labels could have been affixed

to clusters, the keyword tree in effect provided a similar outcome whilst allowing the participant to be more expressive in the way they identify relevant clusters.

The main aim of the keyword tree was to stimulate a targeted search of the clusters rather than asking participants to look at clusters at random. It is not a natural search task to pick clusters in a random fashion; rather, in clustered interfaces, a semantic textual cue of sufficient information scent invites the participant to look inside a cluster. Without a keyword tree, participants could choose to visit each cluster, open the cluster's pop-up, make a judgement on whether it is topic relevant, check to see if the cardinality and or word count criteria are suitable candidates and then either submit it as an answer or use it in subsequent comparisons against other clusters.

This experiment has attempted to validate an approach to data encoding for document metadata visualisation. The scope of this research was well bounded to a few metadata types and for document clusters composed of documents obtained from a search engine. A discordantly obvious question asks where this research can go next, since this experiment considers only two types of metadata for visualisation: cluster word count and cluster cardinality. Admittedly, this experiment uses the strongest candidate in order to support the naturalness-encoding hypothesis, namely size. Although the results are preliminary, and are obtained from a small sample, a natural encoding paradigm would generalise to other metadata types if the selected graphical attribute bears a resemblance or semantic connection to the metadata type it is representing. Table 2.7 presents additional potentially natural encoders that could be explored in a future experiment.

A future experiment should make it more challenging for the participant to obtain the cluster cardinality and word count information presented in the pop-up. This could be achieved by necessitating *click to reveal information* functionality in the pop-up window. This functionality would challenge participants to find an easier way to make judgements about the metadata without having to click a button every time i.e. by recognising and using the visual encoding. Moreover, this interaction would provide a further indication of how often the participant draws on the textual form of the metadata and not the visual cue. A future experiment should also ask more direct questions in regards to how the participant went about answering tasks that relied on the metadata. Typically, we look for ways to short-cut effort and decoding a visual encoding will ultimately be a short-cut way around illusory effortless mouse-triggered pop-up navigation; a future experiment design must guarantee that the effort spent decoding the visual encoding will far outweigh the effort of working out the answer manually via physical interaction.

4.4 Summary

This chapter has highlighted a rich source of potential guidelines for devising metadata-encoding paradigms and proposed that an encoding paradigm might involve selecting graphical attributes that closely resemble their data attributes. There is little clear guidance as to how data encoding rules should be devised for a metadata visualisation approach. This chapter has noted the conflict between guidelines that recommend how to represent data from a data type perspective, and the indirect guidelines that recommend how to represent data for fast and accurate visual search.

A web based experiment was conducted to test the idea that icon size encoding cluster cardinality and icon shape encoding cluster word count results in better task performance outcomes because it is a more natural way to encode data, compared to the opposite way around. Thirty participants from a range of backgrounds participated in the experiment. The results have not indicated a clear benefit in favour of natural encoding since task time metrics were not significantly different across naturally and unnaturally encoded interfaces. However, a weak trend indicated that initial interaction costs may be reduced and task accuracy higher under naturally encoded interfaces. Furthermore, participants could not readily report the way in which data was encoded in the interface and several participants made erroneous assumptions regarding the encoding rules in place and particularly for shape encoding. Future research is required to build on and confirm these findings particularly given a concerning finding that participants did not understand the interface nor were they confident with their use of the interface despite the provision of seemingly detailed training material.

5. ON THE ROLE OF SPACE AND INTERFACE

5.1 *Introduction*

The topic of this chapter is spatially-organised search results. Specifically, this chapter will seek to elucidate the theoretical foundations underpinning such organisation. The next chapter will build on this foundation and seek to improve user interaction with such organisations to achieve better search outcomes.

While earlier chapters have concentrated on the representation of information objects - i.e. documents and their attributes or metadata - the next two chapters are concerned with the representation of semantic relationships. Moreover, this and the following chapter will focus on methods for the construction of usable interfaces through which to interact with these relationships.

A large volume of research regarding information spatialisation does exist, and this chapter will give an overview of key research. Yet, information space techniques have not received widespread adoption; consequently, they remain experimental. Nevertheless, the main contention of this next phase of research will be that our interactions with such information spaces can improve, by addressing fundamental usability problems that influence our search behaviour.

Spatial arrangement featured prominently in the survey of search tools presented in Chapter 2 and in Treharne and Powers (2009). The survey highlighted a general absence of accompanying evaluation work, though some exceptions indicated that more work is needed before the touted theoretical benefits of non-linear presentation are realised. Furthermore, it was suggested that there are core information-carrying elements such as pop-up window document surrogates, which are at present, under-explored, and that we may see more widespread adoption if we build search interfaces that account for the disadvantages of current search tools and capitalise on the affordances that experimental techniques offer.

On the proposition that visualisation techniques, based on spatial organisation, can improve search, a comment of Hearst (2009, pg. 274) is that such interfaces, while abundant, do not live up to the challenge of improving search outcomes. She suggests that these interfaces are too restrictive in the algorithmic sense of organisation and place too much emphasis on spatialisation, while lacking the textual cues that we routinely make use of during interactions with ranked-lists.

The next phase of research will attempt to build information spaces that are consistent with both the observations of Hearst and of Chapter 2. It will address a small but fundamental set of usability factors that may play a role in the production of better visualisation interfaces for search result visualisation. These factors include how best to show document surrogates, how best to show document full-text, and how best to control the layout of spatially arranged documents. Whilst these three investigation points relate to user interface configuration primarily, they are all equally reliant on an underlying information space.

Accordingly, we must first familiarise ourselves with how this integral part of the search tool is constructed. Such familiarisation will have two objectives: first, to solicit a theoretical basis for information space; and second, to motivate a choice of algorithmic approach to build information spaces.

In relation to the first objective, we will explore the pre-existing notion of information spaces that adopt spatial metaphors, which necessitate spatial reasoning and cognition as the basis for effective information presentation. Such spaces in the digital information realm afford cognition the same spatial relationships that we employ in reality. These relationships facilitate judgements of similarity and dissimilarity and identification and recognition of emergent patterns.

While the domain of information visualisation has inherent conceptions of using space to represent information, the information space literature is more specific about the semantics of space. In the field of information visualisation, research generally centres on understanding the interplay and benefit of multiple views or techniques for information processing tasks, and any benefit of spatial relationships is assumed. However, the first objective will deviate from an information visualisation context as is presented in Chapter 2, and instead favour a more fundamental and theoretical perspective on *information spaces*.

The construction of information spaces is exceedingly dependent on algorithms that map non-spatial information from documents into a spatial layout. Thus, in relation to the second objective, we will visit a range of these algorithms for the purposes of investigation, since little immediate guidance is evident in the literature regarding the best algorithms to use in such applications. Such naivety is fair for emerging interface experts, since the available resources for developing one's craft in user interface design, should not be prioritised to solving the problems associated with suboptimal layout algorithms.

The status-quo in guidance toward a leading approach to spatialisation is best illustrated by Kriegel, Kröger, and Zimek (2009) who outline four problems with spatialisation. These problems primarily relate to the number of variables - in this case words in a document - that determine similarities between observations, i.e. documents, in a corpus.

The first problem is that with many variables, it is hard to detect patterns across the corpus. The second problem is that as more variables are added, proximity, distance, and neighbourhood are less meaningful. Similarly, Cribbin and C. Chen (2001) believe that ‘as (keyword) dimensionality increases, representing the thematic diversity of documents using spatial proximity alone becomes less and less effective...although is highly task dependent’. The third problem is that many irrelevant variables are likely to interfere with other more useful variables so there is a need to find a subset of attributes to describe the similarity of objects belonging to similar groups, which could possibly involve different document subsets for different groups of objects. The fourth and final problem is that reducing down the number of attributes risks missing new insights.

Accordingly, there exists a continuum of too many and too few variables and there is no strong guidance beyond using one’s discretion. These four problems typify the complexity of issues surrounding the selection of an algorithm and ensuring a suitable input.

There are of course, exceptions (e.g. Pérez and deAntonio, 2003; Stefaner, 2005; Newman et al., 2009), though such guidance takes its shape in exploratory investigations, and none offer a definitive formula for the production of an information space. Seemingly, the current approach is to experiment with a small number of techniques and to find something that meets a set of desirable criteria and that functions according to expectation.

Throughout this chapter, words and phrases such as ‘information space’, ‘spatialisation’, and ‘spatialised document’ are used interchangeably. Each are taken to mean the process or result of taking a textual document and assigning it a spatial coordinate, according to a measure of semantic similarity between itself and every other document in the corpus.

This chapter will outline a theoretical basis for information space, and such discussion will lead into a more focused examination of the spatialisation process. In addition, a brief review of empirical research will be provided and some open research questions identified. The next section will entail a consideration of what information spaces are, how they can be used and what they look like.

5.2 *The Concept of Information Space*

In a generic sense, an information space is comprised of an underlying spatial arrangement of information, perhaps augmented with a spatial metaphor that provides a basis for navigation or exploration of the space, which changes with time as the space is updated or left to degrade - and which should change to reflect the real-time, dynamic and evolving needs of the user (Benyon and Höök, 1997). Information spaces are ubiquitous; our experiences with the world are defined by discovery, exchange, organisation

and manipulation of information, regardless of context and not necessarily through electronic mediums (Benyon and Höök, 1997). Yet, while the context of exchange and discovery may differ - consider for example a street map, a public transport timetable, or a newspaper - the general notion of information presented on some medium and interpreted by a goal-oriented human cognitive system, remains. Thus, conceptually, interaction with a collection of documents is no less different to working out the departure time for the next express city train.

Benyon (2001) argues that people utilise a variety of information artefacts within an activity space for everyday actions. Interaction with such artefacts involves decision-making; for each artefact there exists an information space that helps us to elucidate the necessary information on which to base decisions and carry out actions. He argues that we need to design for easier navigation of information space to promote achievement of actions and activities. This idea is relevant to the research reported in this and the following chapter, in both the conceptual sense according to Benyon, and in the literal sense. Namely, we may conceptualise information tools as information spaces, because such tools are contributory to broader information activities in which we search for digitised information to assimilate into knowledge and/or for completing everyday actions.

According to Benyon's reasoning, a ranked-list of search results is an information space and so too is a non-linear, visualisation-based arrangement of search results. Given these result presentation paradigms serve a variety of need, we should seek to design for easier navigation to help users achieve their activities regardless of complexity, importance or triviality. However, in the case of visualisation-based paradigms, the anticipated type of search leans more toward complex tasks in which the search is less about look-up and more about look-around to find the answer.

5.3 *Formalising Information Space*

Significant research effort has sought to formalise the idea of spatialisation of information and information space, in which the semantics of documents are represented in terms of location on screen (e.g. Skupin and Fabrikant, 2003; Zhang, 2008). Skupin and Fabrikant (2003) advocate a manifesto entailing a need to discover and convey high-dimensional structures, determine appropriate dimension reduction techniques and utilise cognitively relevant visualisation techniques for the depiction of document corpora. In addition, Zhang (2008) maintains that developers will synthesize future information tools by integrating multiple existing techniques into multiple coordinated views of data or alternatively, by taking the best components of a selection of existing techniques to produce new tools and techniques. Yet, regardless of the approach, Zhang believes that a successful formalisation process or synthesis, if perceptually, semantically and topologically smooth, will mitigate the cognitive burden experienced by

users when interacting with large data sets. This observation of Zhang is consistent with that of Skupin and Fabrikant (2003), and more generally that of the wider user interface community, which in the previous decade has tended to focus more intensely on the human user and human capacities, relative to research conducted prior.

Equally, in this course of research, understanding the capacities of the human are as important as the algorithmic tools underpinning the construction of semantic spaces. Clearly, one such aspect understands how we explore and navigate ourselves around in these spaces.

5.4 *Models of Navigation in Information Space*

Dourish and Chalmers (1994) discuss two distinct spatial models for information navigation. The first pertains to information that is inherently spatial, while the second relates to information that is non-spatial, but which is mapped to a metaphor for the purpose of visualisation.

A similar distinction differentiates that of scientific visualisation and abstract information visualisation even though both are applicable in geographical visualisation and geographical information systems. In the conception of information space by Benyon and Höök (1997), geographic visualisation and search result visualisation are equivalent in terms of the way a user extracts information from the display. However, in practice, they serve different purposes, and consequently, have different and specific research needs. While various fields of visualisation make clear distinctions, the two spatial models as applied within information space, do not make such distinctions; rather the two models share the following three information extraction tasks as suggested by Benyon and Höök: navigation, exploration and identification.

During navigation, a user makes use of way point following to arrive at a destination. Establishing landmarks or way points, in order to go from start to destination, involves a process of orienting one's self in the environment, choosing the correct route, monitoring progress down the route and recognising the arrival at the intended destination. In contrast, in exploration the user is not expecting to reach any particular destination; instead, they want to wander around the information space looking at interesting things. In exploration, the user is concerned with identifying where they are, but their current location is not necessarily important for the next stage of their journey. Finally, in identification, the user looks to identify and recognise categories and types of information, spatial clusters and configurations of objects and object metadata.

Dourish and Chalmers (1994) make a point of separating spatial models into two distinct components, firstly the layout and secondly the navigation or movement between different pieces of information. If the information is inherently spatial then the user is more aptly engaged in navigation, in contrast, if the information is non-spatial

then they consider this information to be organised and the user exploring a semantic space. It is the latter conception that is valid here in this research context. As the corpus used in the ensuing investigation will consist of news articles - consisting of no inherent spatial property - the spatial model that is adopted here describes user behaviour occurring when exploring rather than navigating a semantic space. The experiment of Chapter 6 will support exploration tasks rather than navigation tasks since it is often the case that searchers do not know ahead of time where their information will be found; this applies particularly in situations where the solution to information need is vaguely understood, at best, by the searcher. Moreover, the notions of destination and way point are abstract in this context, since the intended destination is the arrival at a state of knowledge while the journey of getting there is marked by the discovery of pieces of information that direct further movements. Thus, by navigation - of information space - is meant exploring.

Exploring, according to Benyon and Höök (1997) may require a searcher to identify where they are in information space, but it does not necessarily require the searcher to use their current location to predict subsequent linear movements between locations. Establishing where they are based on the region's topic influences the user to continue searching in the local area or forces them to make a non-deterministic jump to another region.

Although navigation is not directly relevant in this context, exploration does produce a historic sequence of interaction within information space. A user may make use of landmarks to undo an exploration pathway in order to reach an earlier location i.e. to re-find information. In addition to semantic landmarks and 'visual breadcrumbs', adoption of spatial metaphors may provide a means to establish landmarks in addition to supporting exploration within information space.

5.5 *Spatial Metaphor*

Use of metaphor in user interface design is based on the intuition that background knowledge or analogy can alleviate learning overheads and make systems more easy and natural to use (Barr, Biddle, and Noble, 2002). A spatial metaphor may benefit interaction and manipulation of search results, since metaphors can offer structural and functional entailments not otherwise available to searchers using ranked-lists.

In the survey of techniques in Chapter 2, there was specific exclusion of systems that make heavy use of metaphor - although some systems such as the bullseye/target inspired system of Cho and Myaeng (2000) were included due to the simplicity of the visualisation. The survey could easily have been extended to include more sophisticated - and most likely three dimensional - visualisation systems, including for example, those surveyed by Benford et al. (1999), the NIRVE system as evaluated by Sebrechts et al. (1999), a townscape inspired system of Bonnel, Lemaire,

et al. (2006), data mountain inspired ideas as evaluated by Cockburn and McKenzie (2001), mountain range inspired <http://www.search.tianamo.com>, geometrically inspired <http://www.search-cube.com> and forest based metaphors as explored by Akhavi, Rahmati, and Amini (2007). This small sample is based on a set of spatialisation metaphor classes identified by Benking and Judge (1994) which include geometric forms like cubes, spheres, and polyhedra; artificial forms like townscapes, houses or rooms; natural forms like landscapes and forests; systemic structures like road highway systems, pathways or flow systems; dynamic systems like atomic, molecular, planetary, or galactic systems; and traditional symbol systems like mandalas - inspired by Buddhist art work depicting the universe or cosmos - and sand paintings.

More recently, MacLennan (2005) thought to interview metaphor end-users and queried how they believe spatially motivated layouts should look, given his observation that the preferences of such users had not previously been considered in new system design. Interestingly, MacLennan recorded adaptations on library e.g. bookshelves, forest, cityscape, and galactic solar system metaphors, suggesting that these types of metaphors are intuitively understood and readily transferable to an abstract context like document visualisation. This may be explained by a reiteration of Barr, Biddle, and Noble (2002) that the aforementioned structural metaphors - e.g. cityscapes, forests, galactic solar systems - have metaphoric entailments including properties and functions.

A cityscape metaphor entails properties such as neighbourhoods and city regions that are connected by roads that the user may travel on by way of virtual transport i.e. functions. Furthermore, orientational metaphors, such as the bullseye system of Cho and Myaeng (2000) incorporate our understanding that the game of darts or archery is won by propelling the dart or arrow into the centre of the bullseye. Thus, the items in the centre of the visualisation are more likely to be highly sought after in contrast with those situated outside the central location. Different again, but widespread in many computing user interface metaphors, are ontological metaphors which quantify otherwise abstract concepts such as documents and files stored electronically. A sequence of bytes stored electronically, quantified and labelled as a file, can have attached to it, a set of descriptive data such as name, and virtual location and a file size. Instead of acting on a sequence of bytes in memory, we execute operations on labels and icons.

However, before a notion of semantic neighbourhood may be established, there must be some way of assigning a thematic label to a particular region of information space and subsequently, to relevant documents in that region. That is, in order for regions or neighbourhoods of houses or mountains and valleys to be drawn and visualised, a layout algorithm must first allocate related information objects to a substrate. Then, once formed, the layout remains fixed and the user spends the remaining time exploring the information space utilising a set of interactive mechanisms that are consistent with the selected metaphor - e.g. flying, driving or walking an avatar around the space.

This research does not adopt a spatial metaphor of the types described here; there

are at least three reasons for this decision. The first is centred around an aversion to the use of three-dimensional graphics for visualisation; this was highlighted earlier, based on empirical evidence suggesting that two dimensions are sufficient for visualisation (e.g. Tory, Sprague, et al., 2007; Cockburn and McKenzie, 2001). The second is based on the observation of Benyon and Höök (1997) that information spaces should be dynamic and responsive to the user's continually changing information need; this is consistent with the information-seeking model that is adopted for this research and furthermore, there are no known town or landscape metaphors that smoothly change in layout in response to user interactions. Finally, the third reason is that documents can contain multiple topics as is emphasized by Blei, Ng, and Jordan (2003) and techniques that categorise or decompose corpora into two or a few representative dimensions stop short of supporting a user with making finer distinctions between documents more quickly and easily. Reconfiguring the searcher's perspective of information space is a useful way to convey these distinctions; however, changing the layout - by rotating projection dimensions or thematic dimensions of information space - necessitates that users understand the utility of doing so as well as having an interaction control that supports informed decision-making regarding dimension selection that will benefit search. It is this third reason that has practical implications for the design of the experiment in Chapter 6.

5.6 *Spatial Configuration*

There are inherent difficulties in visualising large quantities of text. The human cognitive system processes text in a serial and linear fashion, thus prolonging judgement of the main themes and topics within large passages of text. Use of mathematical tools that leverage the user's ability to spatially reason, provide ways to focus a searcher's attention; however, effective display and use of the output of these tools is critical to success.

Frequently, themes and topics are buried amongst a sizeable volume of qualifying linguistic filler. In order to extract the main themes and topics automatically, and in order to build summarised and perceptually efficient corpora visualisations, a dimension reduction or analysis algorithm is cast over the corpus. Dimension reduction techniques reduce the very highly dimensional space, where each word is a dimension, down into a smaller dimensional space, where each theme consisting of a combination of words, is a dimension.

Visualising any more than two or three dimensions is technically difficult to do and interpret; further reduction is often necessary. A projection algorithm considers the distance between documents in a high dimensional space and projects documents into a two or three-dimensional coordinate system; the distances between documents in the high dimensional space are preserved as closely as possible in the low dimensional space.

In practise, themes take the form of highly dimensional orthogonal vectors. Each cell of each theme vector corresponds to a document's 'association' - or to what degree a topic is contained within the specific document. A theme's vector size is equal to the corpus size, though individual documents may have zero or close to zero - due to noise - association with a theme. By taking one theme vector as the horizontal coordinate axis and another theme vector as the vertical coordinate axis, we arrive at our visualisation of information space.

Theme vectors are characterised by different keyword patterns and multiple documents can share relationships or associations with multiple theme vectors. Rotating the searcher's perspective of information space, by selecting an alternative theme for the vertical - or horizontal - dimension, will likely change the layout of documents. Observing the layout of documents across rotations can indicate interactions between documents that were not obvious in the prior view. Tracking the movement of document icons across changes in perspective might be made easier by the provision of perceptual cues like path trails (Baudisch, Tan, et al., 2006) or colour coding. However, deciding which dimensions to rotate in, in the first place, has its own challenges.

An elegant solution to rotation control is found in the Scatter Dice proposal of Elmqvist, Dragicevic, and Fekete (2008). Scatter Dice incorporates interactive multiple scatter plots that preview the configuration of each dimension pair in a micro-view, as well as providing a facility to select a specific projection dimension configuration. Clicking on a micro-plot rotates the searcher's perspective of information space; pre-selecting several micro-plots of interest provides the basis for an automated 'grand tour' routine. Smooth icon animation ensures the transitions between perspectives are not abrupt, thereby attempting to preserve the knowledge of where document icons have come from and where they have arrived at. The philosophy of the rotation control research in Chapter 6 is similar to that of the scatter dice idea; however, in the research of Chapter 6, instead of metadata dimensions for a consumer product catalogue, the dimensions that searchers will rotate are thematic.

5.7 *Evaluation Results*

Hearst (2009) writes that systems that utilise spatialisation are yet to show that they are useful and understandable, due to a static organisation of information space. Instead, a more flexible arrangement should be better for showing large overlap between themes in documents. Yet, there are many factors impacting usability and comprehension of information spaces in search tools. These factors include in the least: fundamental spatial comprehension skills; an ability to infer semantic relationships based on spatial arrangement, and the role of spatialisation interfaces as one particular view within a number of coordinated views each devoted to some aspect of information-seeking. The available research is sometimes conflicting on these aspects.

Visual-spatial displays augment cognition by externalising information storage, organising information, offloading cognition to perception, and offloading cognition to action (Hegarty, 2011). A direct example that illustrates this point, in the ranked-list versus visualisation context, is that upon finding a relevant document in a ranked-list there is no reliable basis to determine where the next relevant result will likely be. In contrast, in a spatially organised presentation, close spatial proximity provides that basis to look for the next relevant document.

In relation to comprehension of information space, studies reveal that users understand the fundamental metaphor that close spatial proximity entails semantic similarity, while greater distance entails dissimilarity. Skupin and Fabrikant (2003) reiterate that participants can identify that distance between documents indicates relatedness, participants can identify that clusters reflect structure within the corpus, and participants can recognise that different scale or zoom corresponds to level of corpus detail. Furthermore, Niemelä and Saariluoma (2003) show that semantically organised spatial layouts can influence recall and learning of items within the space, thus, making it more challenging to recreate the spatial layout in memory in order to interrogate them. Therefore, it is necessary that information spaces are constructed such that spatial groupings are evident upon inspection, and furthermore that items within those groups share a common semantic meaning.

However, as Fabrikant, Ruocco, et al. (2002) observe, the use of node-edge network diagrams can override similarity judgements based on ‘as the crow flies’ distance and so too visual containment - e.g. by using a colour to define a topic region within which points are located. For distance between two nodes, users more readily think that similarity is based on the total sum of edge lengths on the path from one point to another (Fabrikant, Montello, et al., 2004). While this suggests that a range of spatial cues could be employed to indicate semantic relationships between items in an information space, only the strongest semantic relationships of specific classes should be made explicit. This avoids the situation whereby an overly large number of explicit network connections clutter the interface making it difficult to use.

In terms of the users’ spatial aptitude, Skupin and Fabrikant (2003) reiterate that specialised expertise and background knowledge are not necessary for the understanding of spatial metaphor and spatial perceptions, meaning that use of these techniques is natural for users. However, by definition, information spatialisation is visual in nature and Benyon and Höök (1997) recount that not all users prefer a visual representation of their information - rather they prefer verbal instructions or prose. In addition, they reiterate that task performance correlates with measures of spatial ability, experience, technical aptitude, and learning style.

The role of visualisation-based information spatialisation in multiple coordinate view paradigms is investigated by McCormac et al. (2012). They find that a spatialisation approach, in combination with a number of other arrangements of the same

data, result in superior user performance compared to each individual way to display an aspect of the data. On the other hand, while Hornbæk and Hertzum (2011) observe widespread subjective satisfaction with *overview and detail* or *focus and context* interfaces that provide multiple views of a data set, objective performance measures are unclear.

Usability factors influence our interactions with visualisation-based information spaces. In support of this view, in the context of a geographical visualisation, Hornbæk, Bederson, and Plaisant (2002) demonstrated the need for research dealing with the improvement of usability of visual displays indicating a difference between the interaction style of browsing supported by *focus and context* interfaces versus zoomable interfaces.

Another factor is illustrated by a comparison of Vivisimo (see Koshman, Spink, and B. Jansen, 2006) - a text clustering interface - and Grokker (see Koshman, 2006) - a text and visualisation clustering interface - in which Rivadeneira and Bederson (2003) identified that the immediately accessible results in Vivisimo are advantageous in contrast to other visualisation-based interfaces that do not offer immediately accessible results. Result surrogates are a leading way to provide information to users as this provides the basis for query refinement. If the user has to drill into several categories before seeing results, which turn out to be wrong, then this will be frustrating. A details-on-demand (Shneiderman, 1996) inspired interface that facilitates search of a document collection is likely to engender frustration in its users since a mouse interaction is required to integrate each potentially interesting result in a sequential and slow fashion.

These evaluation results suggest that users understand the role of spatial cues, whether implicit or explicit, in representing a semantic relationship between two or more spatially arranged items. Furthermore, spatialised information should exhibit clear spatial grouping as well as clear semantic meaning in order to improve interrogation of spaces over time - or across dimension rotations. However, while the fundamental idea is comprehended, the way systems incorporate spatialisation into functional tools should be subjected to further research. Studies indicate that in isolation, a single method of display such as a list or a single spatial organisation, may not as much engender positive search outcomes as multiple coordinated views (Hubmann-Haidvogel, Scharl, and Weichselbraun, 2009; McCormac et al., 2012)

Systems that employ such displays should be subjected to further usability testing; more attention should be devoted to evaluating the integration of spatialisation into functional information tools, further to improving our interactions with these types of displays. There has been too great a presumption of the power of information visualisation when applied to information retrieval tools. The presumption that if we make everything graphical we should expect to gain an increase in performance, is wrong; a too literal interpretation of information visualisation applied to search tool design overlooks the typical way users do search and designers should not continue

to design tools with the expectation that purely graphical tools will be a panacea for information overload.

5.8 *Open Research Questions*

So far, this chapter has established the notion of an information space and our potential explorations within these spaces as a means to find and interact with information. In addition, the opening chapter indicated the need for spatial organisations of information that depict semantic relationships between documents. However, for the successful marrying of spatialisation and search engine results, there are at least three open research questions that pertain to inefficiencies and suboptimal interactions with these spaces. These include the role of configurable information spaces in search tools, the integration of multiple document surrogates in spatialisation, and finally, the attachment and linking of the information space to the full-text view of the documents.

Information spaces and overviews generally remain static and fixed but do not necessarily have to remain so (Hornbæk and Hertzum, 2011; Sabol et al., 2009; Benyon and Höök, 1997). Moreover, one could consider configurable information spaces as a method to reformulate an initial ambiguous query by changing the visualisation's perspective of projection. Given the observation that users understand and can work with semantic spaces (Skupin and Fabrikant, 2003) and if users can understand the notion that information spaces are configurable, one open research question is to explore whether the navigation controls influence those interactions. How do we facilitate manipulation of information spaces and furthermore, how should we offer semantic cues to indicate the expected advantages gained if a change of the layout of an information space is enacted.

Selecting an appropriate spatial layout does not guarantee a perfect outcome. A searcher must subsequently identify the thematic content of the configured space - perhaps with the aid of labels - and after having located a region of interest, prior to opening specific documents, the user must identify what the documents are likely to discuss. This process is traditionally serial, in line with a details-on-demand paradigm per the visual information seeking mantra of Shneiderman (1996). In list-based interfaces, searchers make heavy use of document surrogates to estimate document content, and they do so quickly. Serial, detail-on-demand paradigms violate this interaction behaviour because they are slow and laborious. We must take further steps to integrate both surrogate information and thematic information together if we are to take advantages of the benefits afforded by both information visualisation and contemporary search tools. Hornbæk and Hertzum (2011) indicate that for overview and detail interfaces, to be more useful, there should be some improvement in the support for current search practices and more powerful search behaviours. Consequently, this too should include fast and optimal document surrogate scanning behaviours.

Hornbæk and Hertzum note the differences in philosophy between that of the visual information seeking mantra (Shneiderman, 1996) - *overview first, zoom and filter, then details-on-demand* - and of Tufte (1990) and Tufte (2001) *detail at all times*. Many interfaces proposing alternatives to search result visualisation - e.g. (Spoerri, 2004), Kartoo (see Koshman, 2006), Grokker (see Rivadeneira and Bederson, 2003) and Ujiko (see Foenix-Riou, 2006) - follow closely that of Shneiderman's visual information seeking mantra by offering a details-on-demand facility to view document surrogates. This situation is in conflict with current search result lists. Therefore, how do we create an interface that offers both a global overview and local document detail - something that is necessary if we are to see greater evolution beyond traditional ranked-list interfaces. We need both automated and manual detail-on-demand facilities in the foreground that do not interfere with a global overview in the background. This combination will be developed in the next chapter Chapter 6 when a description of an experiment apparatus is offered.

While the first and second research questions relate to earlier sections in this chapter and more generally to the latter part of Chapter 2, the detachment of full-text document view from the search results page has come about by observation of systems presented in the survey of Chapter 2. The default mode for searching for documents using web search engines is one that separates the search results from the full-text of the document. However, in information tools - like desktop email clients - we are used to interacting with a list of emails, a folder structure and the full-text of a selected email all within an integrated view. This suggests that it may be useful for full-text views to be integrated into our search tools to provide a foundation to facilitate the execution of subsequent searches in our ongoing search processes. For example, a searcher should be able to initiate a search from the full-text of the document, to select parts of a document's text and to look at other documents in the result set that are similar. In this light, the search result set evolves in response to the user's ongoing interactions - something information spaces should offer according to Benyon (2001).

Generally, the searcher will open potentially interesting documents in new browser tabs, windows or even in their current tab - thereby superseding the result list altogether. A final open research question relates to the observation that while these behaviours are routine in web search contexts, many experimental interfaces provide integrated interfaces consisting of overviews and document previews in the same window. The open research question is to explore and compare the impact on search performance and behavioural differences among users under these different full-text view scenarios.

The three open research questions will be addressed in an experiment reported subsequently in Chapter 6. A description of the experimental apparatus will focus on the interface features directly applicable to the aforementioned open research questions. For the remainder of this chapter however, discussion will outline the construction of

an information space utilising a combination of dimension reduction and projection algorithms, as this will directly contribute to the construction of the information space presented by the experimental apparatus in the next chapter.

5.9 *Deciding on a Spatialisation Construction Approach*

Whereas the previous sections set a foundation for information space, this section will concentrate on how information spaces are realised. The aim of this section is to motivate, by qualitative and quantitative assessment and measures, a suitable approach for the construction of an information space; this information space will be present in the experiment apparatus in Chapter 6.

Typically, selection of a spatialisation methodology rests on the outcome of an objective assessment; in this particular case, selection will be foremost based on a set of qualitative criteria and subjective assessment, since such consideration is rarely seen in the literature. On the other hand, Morse, M. Lewis, and Olsen (2002) note that algorithm comparison is often lacking and implicitly carried out as part of a wider system evaluation. Accordingly, given a set of alternative spatialisation algorithms under consideration, some attempt will be made to provision objective evidence as well. Through considering both the qualitative and objective support, a combination of dimension reduction, projection and pre-processing algorithms will be selected for use in constructing spatialised document sets.

When attempting to design and build interfaces for the improvement of information retrieval tools, much time can be directed to wrangling suboptimal algorithms in order to produce a tool that works consistently with human expectation. One might suggest a collaboration could divert such wrangling to a suitable expert, though such collaboration is not always possible. Alternatively, one could also suggest that usability research in the vein of this experiment could be conducted on hand-crafted and specifically engineered datasets that specify the layout and clustering attributes for each document. However, this runs the risk of artificially inflating the capability of an interface. Real tools operate over imperfect data, and to do otherwise would be unrealistic and potentially unhelpful.

There are significant time and resource overheads required when building even moderately sophisticated experimental search tools that cater for large and diverse corpora - such as web pages from the Internet. To make better use of such resources, we need algorithms and frameworks that are known to work suitably, to act as pluggable black boxes in top-down style usability research, thereby leaving the researcher more effort to devote to the actual research. This would reflect the current directions in information visualisation, in which libraries and tools are making it easier to transform a data set into a usable visualisation with little overhead. Moreover, in the context of tool design,

it ensures that the researcher is usability expert first and document clustering algorithm expert second rather than the other way around. For the experiment of Chapter 6 specifically, the research focus lays in the usability aspects of the information space and not the space itself, even though the space is vital to the successful running of the experiment.

The main goal of the experimental apparatus in Chapter 6 is to support the exploration of a hyper-dimensional information space. The apparatus will display an information space for a set of search results; the space will be labelled to assist with the searcher's exploration of the space. Interactive controls will facilitate both exploration of documents and exploration of the space itself. Relationships between documents and document clusters will be made clear by way of rotating around the user's perspective of the information space.

This main goal and set of requirements invites a set of qualitative criteria that a potential spatialisation construction approach must meet. Discussion of the relevant qualitative criteria will come later and after an outline of the techniques under consideration, as well as a discussion of the document corpus and of its pre-processing. The next section describes the document corpus in use.

5.9.1 *Document Corpus and Topics*

The document corpus in use for this evaluation and the experiment in Chapter 6 is a subset of news articles taken from the Reuters RCV-1 collection (D. Lewis et al., 2004) spanning the years 1996-1997. Topics chosen for each task set are inspired by the 1996 and 1997 year of events page on <http://en.wikipedia.org/wiki/1996> and topics from the Text Retrieval Conference TREC ad hoc and novelty search track topic questions - see <http://www.trec.nist.gov/data.html>. A set of queries were devised - see Table 5.1 on the following page - involving the task set topic words and a set of related words prescribed by the online version of WordNet - see <http://wordnetweb.princeton.edu/perl/webwn>. While a searcher might not explicitly construct such queries, it is often the case that queries are expanded automatically in the back end of commercial search engines like Google to increase recall.

By query expansion is meant the addition of words by the search engine to the searcher's query to improve recall of relevant documents. Carpineto and Romano (2012) propose a survey of query expansion techniques that are used to generate additional terms to expand search queries. This collection encapsulates - but is not limited to - linguistic analyses such as using term stemming, corpus specific techniques such as term clustering, query log analyses such as looking at the content of top ranked documents obtained by related queries, query log analyses to identify similar search queries - and therefore additional query terms - that have been incrementally refined over a search session, and finally by using web data such as Wikipedia and anchor text. Carpineto and

Tab. 5.1: Topics and queries that form the basis for task sets in experiment.

Task Set	N	Topic	Search Query to Lucene
RM	109	Recyclable Materials	recycl* + 'resource recovery' + reuse + reprocess
HK	134	UK Hand Over of Hong Kong to China	uk + 'united kingdom' + china + 'hong kong' + sovereignty + handover
NP	156	News Paper Circulation Decline	'newspaper' — 'news paper' — broadsheet — newsprint — tabloid — magazine — circulation — decline — reported — material — subscription -reported

Romano suggest that query expansion is not yet widespread in commercial web based search engines; although query expansion techniques are more widespread in Intranet search, desktop search and in systems based on Lucene <http://lucene.apache.org>, the latter of which provide search technology for a large number of popular websites - although it is not clear which of these sites actually incorporates such technology. The limited uptake of query expansion techniques in commercial web-based search engines is due to several reasons including server processing time, the suitability of expansion to some query types, and finally, searcher confusion when results appear to offer terms that the searcher did not actually use to search. Despite this claim, Google indicates on their search help page that 'synonyms might replace some words in [the] original query...' (Google, 2013) while Baker (2013) provides a discussion about the quality of query expansion by synonyms at Google, thereby implying that query expansion is a feature of some Google searches. It can be verified that this is the case on a popular engine Google with a few test queries. Spelling corrections are another example where queries are modified for the searcher, for instance, typing in 'FCBook' turns up zero results in the first page containing 'FCBook' - results only relate to the <http://www.facebook.com> website. These queries have the effect of simulating diversity in the result set, as happens in web search, due to the size of the Internet. Prior to widespread prevalence of query term expansion at Internet-based search engines, Muramatsu and Pratt (2001) showed that searchers not only intuitively understand query expansion, using morphologically related words, they in fact expect it to happen.

By observation, each result set elicits appropriate ambiguity and a number of different themes. For example, the training task set contains documents about rockets in warfare, satellite launches, the Mars rovers, space technology, space industry and NASA's space shuttle program. Additionally, the Hong Kong Handover task set includes articles about the United Kingdom's handover of sovereignty over Hong Kong to China; however, this set also includes other sovereignty crises that were occurring around the same period - though involving regional countries and China. Finally, the

newspaper circulation topic has an extra restriction in that the word ‘reported’ cannot appear in the article, in order to deal with the overly abundant number of hits containing the phrase ‘newspaper reported’ - articles containing this phrase are generally reporting a story using an another news article as the source. Without this exception, the prevalence of irrelevant documents is too severe; nonetheless, the word ‘newspaper’ would be an appropriate keyword if wanting to find documents about the reasons for a decline in newspaper circulation around the world.

Each of the four result sets has a size of 100-160 articles. Pre-processing steps taken to prepare each document set for dimension reduction, and eventual spatialisation, are described subsequently.

5.9.2 Pre-Processing

In this evaluation, pre-processing occurs at query time but could could happen at indexing time thereby opening up a range of optimisations. Pre-processing steps include filtering for long articles, junk - i.e. semantically empty - articles and duplicate or near duplicate documents. These filtering steps attempt to control for variation in reading time in the ensuing experiment. Long articles are considered to be any document having five hundred words or more and are excluded from analysis. Additionally, specific article types like stock market reports and sporting event reports are also excluded based on their lack of content.

Duplicate and approximately duplicate documents are detected using a simple and rough algorithm, which is sufficiently effective by observation. Two documents are considered duplicate or near duplicate, if after the exclusive disjunction XOR of their term sets, a set of 10 unshared terms remain.

A threshold of 10 terms was set by trial and error; in this case, such a threshold results in better performance than a threshold of 15 or greater - particularly for short documents. To reiterate, this is a trial and error process operating on the content words only; there has been no formal verification of the method, nor has any formal tuning been performed. Throughout informal tuning, it was noted that too large a number results in many false positives, while too small a threshold results in many false negatives.

While this simple technique is quite effective for the detection of exactly similar documents and documents that have been edited slightly during the editorial process it is however, overly strict for documents that have a consistent reporting format like sports results in which there will be one or two words of difference e.g. a day of the week and a difference of word order. Since this method considers only words containing alphabetical characters and hyphens only, two sports result articles reporting the result of a game between the same two teams but on different days, potentially with different score results may be incorrectly detected as duplicate documents. The same is observed

for finance market reports. These cases provide an additional motivation to exclude sports results and some finance reports from task set construction. The RCV-1 news article category code metadata provides a reliable way to filter out these types of articles rather than relying on thematic or genre detection.

A list of standard stop words are removed from articles including a number of words relating to days of the week and words like ‘Reuters’ and ‘news room’ that appear in most documents, but do not contribute greatly to a document’s theme set. Furthermore, named entities are extracted from each document, using the Stanford named entity extractor (Finkel, Grenager, and Manning, 2005) and each entity is considered as a whole, rather than individual terms when tokenisation takes place. The remaining terms along with named entities are collated into a term frequency matrix, which serves as the input, along with appropriate term and source look up structures, to the algorithms discussed in the next section. These algorithms will seek to reduce the very high dimensional space, whereby each word is a dimension down into a smaller number of dimensions or themes by looking at the pattern of words within the term frequency matrix.

5.9.3 Algorithms

There are five techniques under consideration, including three dimension reduction techniques and two projection techniques.

The dimension reduction techniques under examination include Single Value Decomposition (SVD), Latent Dirichlet Allocation (LDA), and Fuzzy Hierarchical Clustering (FHC). The projection techniques under examination include Multidimensional Scaling (MDS), and Correspondence Analysis (CA). For this analysis, SVD will also be considered as a projection algorithm in the sense that orthogonal SVD vectors can serve as the basis for a 2D visualisation.

A qualitative treatment of each algorithm will be presented, rather than an in depth mathematical treatment regarding the inner workings of each technique. For the specific details of each algorithm, the reader will be referred to the relevant literature where appropriate.

Dimension Reduction Algorithms

In the following sections, a sample of dimension reduction and compression algorithms that are adopted for this case study are presented. These algorithms are adopted to reduce highly dimensional documents - where each keyword is a dimension - down into a few representative themes.

Latent Semantic Analysis via Singular Value Decomposition Latent Semantic Analysis (LSA) is a technique first proposed for an information retrieval application by Deerwester et al. (1990). LSA applies a matrix decomposition technique, Singular Value Decomposition (SVD), to break down a term frequency matrix into a term space matrix U , a document space matrix V and a scaling matrix S ; a subsequent multiplication allows the recovery of the original matrix by the equation $A=U*S*V^T$. The decomposition operation of the SVD considers each document as a linear equation, and the decomposition analogous to the process of solving very large sets of simultaneous linear equations (Landauer, Laham, and Derr, 2004).

LSA's application to information retrieval is advantageous, as the rank of each decomposed matrix may be cropped while still permitting a near complete recovery of the original term frequency matrix, thereby reducing storage overheads. Moreover, once built, new documents may be classified according to the model by first boosting a document vector into the term space and then calculating the similarity between the new document and existing documents in the model, resulting in a similarity score upon which to sort documents.

Each column of U and V represents an orthogonal theme, which contains association scores for documents and terms that typify the meaning of the theme. Highly associated documents and terms can be employed to describe each theme vector in order to assist with navigation of the hyper-space.

One advantage of incorporating a Latent Semantic Analysis into a search tool is that although the decomposition algorithm is slow to build and increasingly slower for increasingly large number of variables i.e. words, the model can be built offline and polled quickly in real time. The technique can support ad hoc searching by taking the user's query as a pseudo document and calculating the cosine similarity between it and document vectors in the V matrix. The similarity provides a sortable measure that could be encoded by a graphical attribute in a document visualisation e.g. using relevance bands assigned to hue or colour saturation, for example.

This technique provides the basis for a document query score as well as a visualisation of a document space, while columns of the U matrix provide a source of term scores for visualisation. In principle, using the highest activated words in the U matrix - terms by themes - is a straightforward approach to labelling the latent themes of the document space. One possible way of choosing terms is identifying the top terms that exceed the noise level of all terms in the dimension. However, informal investigations suggest that the most extremely activated words do not form a cohesive picture or common theme for dimensions - the set of activated words do not necessarily meet human expectation. Although those words do appear in documents that are highly weighted for that theme, they do not adequately characterise the set of documents overall. Theme words tend to be overly specific to a few documents only.

For this analysis, the first orthogonal dimension is ignored from consideration, as this dimension does not offer interesting themes; the first dimension is considered the dimension, which characterises a word's overall frequency across the corpus (X. Hu et al., 2003). Thus describing the first theme would result in labelling the theme with the top terms across the corpus. Accordingly, in this use of LSA, only themes two through twelve will be considered.

Fuzzy Hierarchical Clustering Hierarchical Clustering is a traditional unsupervised cluster analysis technique with essentially two flavours: bottom-up agglomerative and top-down divisive. Whether agglomerative or divisive, hierarchical clustering algorithms iteratively merge or divide clusters together based on distance measures until a single cluster remains, all clusters are atomic objects, or some termination condition is met (Han and Kamber, 2001).

In addition, regardless of flavour, the resultant clusters are typically hard, such that each object viz. document is assigned to exactly one cluster only. Typically, once a document is assigned to a cluster, it does not ever shift to another cluster even if a more appropriate cluster exists.

However, documents can discuss multiple themes within the same document, in addition to documents discussing predominantly one theme. Therefore, multi-themed documents should be placed within or associated with multiple clusters for each theme discussed by a multi-themed document.

Out of convenience, a fuzzy hierarchical clustering algorithm was adapted from the parsimony driven co-clustering research of Leibbrandt (2009), since this work had similar intentions of the present application, albeit using different notions of objects i.e. documents. In the case of Leibbrandt, the document - termed a 'frame' - is a sentence template and the outcome of clustering to form categories or clusters of frames in addition to representative words.

As will be outlined below, this co-clustering idea is relevant to the present context such that documents can be assigned into multiple categories, as can words resulting in a model of the corpus, which can then be projected into a coordinate space by way of the association scores that words and documents have with the emergent clusters.

Latent Dirichlet Allocation Latent Dirichlet Allocation (LDA) represents the more recent approach to probabilistic computational topic model algorithms for use in topic modelling. A decomposition of the technical mathematics underpinning the Latent Dirichlet Allocation can be found in the paper by Blei, Ng, and Jordan (2003).

Similar to the Latent Semantic Analysis approach incorporating an SVD algorithm, documents are a mixture of topics and topics are based on patterns of word

co-occurrence. However, unlike in SVD, topics in LDA are not orthogonal dimensions, in the sense that latent themes are in a SVD.

Consequently, documents and terms are not positioned within a hyper-dimensional space. Thus, the mixed-membership assumption and availability of association scores will be interpreted as a forming topic vocabulary in order to establish a basis for projection into a lower dimensional space by way of a projection algorithm.

Projection Algorithms

In the following sections, a sample of projection algorithms adopted for this case study are presented. Two algorithms are discussed that take as input the topic models produced by way of dimension reduction algorithms presented in the previous section. This reduction further reduces the number of dimensions that a user may need to consider during search, and furthermore, further reduces the number of dimensions such that it becomes easier to visualise a high dimensional space in a two dimensional visualisation.

Multidimensional Scaling Multidimensional Scaling (MDS) is a projection algorithm that moves points around in a similarity space until the typically Euclidean distances between each document are approximately equal to the distances between documents in the dissimilarity input matrix. This process continues until a stress function that captures the degree of error between the dissimilarity space and the map space minimises (Buja et al., 2008) or a termination condition whereby a maximum iteration or processing time threshold is exceeded.

MDS is advantageous as the process of obtaining a spatialisation of a topic space is straightforward. A dissimilarity matrix provided to MDS requires only a small modification to the output of each of the topic models. Furthermore, in prior collaborative investigations (e.g. McCormac et al., 2012) MDS was successfully used to build spatialisation to serve search experiments using email corpora and spoken dialogue corpora.

One drawback of the MDS approach is that unlike in SVD and Correspondence Analysis, it is challenging to assign labels to the MDS space. Furthermore, the projection dimensions produced by this technique have no basis for interpretation i.e. we cannot say dimension one is typically described by a set of keywords as can be done for SVD.

One approach to labelling is to insert dummy label documents in the MDS, but this course of research has not objectively verified the suitability of this approach. Aside from comparing common keywords within a set radius of each spatialised document, the general lack of guidance on this use of MDS suggests that there is no agreed way of labelling, and it does not appear that MDS is ever utilised in the fashion that is prescribed here.

Correspondence Analysis While not a projection algorithm per se, informal experiments throughout this research have shown Correspondence Analysis (CA) to produce visually clean and coherent topic model spatialisation. Murtagh (2005) provides a solid introduction and examination of the correspondence analysis technique.

Correspondence analysis fills the labelling of space shortcoming that was identified in the MDS section above. In CA, observations i.e. documents, and variables i.e. topic labels, are plotted in a high dimensional space; furthermore, weightings of variable association to projection dimensions can be used to characterise the semantic content of a specific projection dimension in addition to regions of the spatial area.

Java code for a correspondence analysis was obtained from a related website - see <http://classification-society.org/csna/mda-sw/correspondances/>. No definitive conclusion can be made regarding the maximum number of observations that this software can handle, as the document sets in this research have been small.

5.9.4 Evaluation

Having introduced a suite of dimension reduction and projection algorithms, this section will report the results of an evaluation of each algorithm. The result and objective of this evaluation was to select one dimension reduction algorithm and one projection algorithm to use in the building of an information space for a subsequent experimental investigation. In discussing the results of this evaluation, reference is made to the projection and reduction algorithm selection as simply the ‘spatialisation approach’.

There were two stages to this evaluation. In the first stage, a qualitative evaluation, based on five criteria, ranked the suitability of each algorithm for the current research context. In the second stage, an objective evaluation compared performance of algorithms using an objective measure. The superior algorithm was that which met each of the five qualitative criteria and which performed best according to the objective measure.

The following section introduces the set of criteria that were used in the qualitative evaluation. Following this, discussion turns to the results of the qualitative evaluation and then to the objective evaluation.

Qualitative Evaluation Criteria

For the present work, an appropriate spatialisation approach must meet five criteria to qualify for selection. Originally, this research had not intended to prove which of the dimension reduction and spatialisation algorithms were objectively superior; rather the intention was to find algorithms that would achieve a main goal and set of functional requirements as discussed in Section 5.9 on page 238.

Criterion One - Mixed Membership Topic Model The reduction and projection approach must produce a mixed membership model of the document corpus. This reflects the assumption that documents discuss more than one theme or topic and have connections to other documents based on these themes, regardless of the strength of each theme. This criterion necessitates or permits document membership in one or more clusters and an association measure to indicate the strength of the association with each cluster. This allows a comparison of two or more clusters or themes against each other, to isolate interactions between to the two.

A mixed membership model facilitates the provision of a set of topic projection dimensions, which through perspective rotation operations, facilitates the movement of documents around the spatialisation; this process visualises the inter-twinning of documents and topics.

Criterion Two - Representative Topic Descriptors The reduction and projection approach must produce or facilitate the production of a useful set of descriptor labels that uniquely and adequately identify the nature of documents in the topic. Descriptor labels are used to annotate the spatial substrate and to annotate projection dimensions, thereby provisioning searchers an understanding of the meaning of regions in the spatialisation, as well as facilitating an informed choice of projection dimension selection.

Criterion Three - Document-Topic Association Scores The reduction and projection approach must provide document association scores per topic for a specified number of dimensions, to facilitate the visualisation of documents in a document space. This criterion is some what dependent on the first; however, it is possible to achieve the first criterion without satisfying this third criterion. For example, a clustering algorithm could simply output the raw multi-membership cluster identifiers - by criterion one - without providing any association to those clusters - by criterion three. Therefore, this criterion requires the actual association scores so that they may be visualised; it is not simply enough that document A belongs to cluster 1, 2 and 3; an association score between each document and each cluster is a necessary requirement in order to visualise those associations graphically.

Criterion Four - Descriptor-Topic Association Scores The reduction and projection approach must provide word association scores per cluster for a fixed number of dimensions, to facilitate the visualisation of terms, thus providing information scent within regions of the document space. This will also provision the ability to plot descriptor keywords in the spatialisation and to provide descriptor keywords that best describe the projection dimensions selected for display.

Criterion Five - Feasible Provision of Ad hoc Search Support The reduction and projection approach should provide workable answers for a document set size of between 100 and 1000 documents. Topics should not be so large as to be unmanageable and equally, topics should not consist of word frequency patterns so vague as to not reveal any obvious pattern. In addition, algorithms should be picked with a vision of serving ad hoc search in real-time applications in the future and should process input quickly.

Qualitative Evaluation

The five qualitative criteria were used in an evaluation of the algorithms discussed in Section 5.9.3 on page 242. The results of the qualitative evaluation are presented in Table 5.2. \oplus , \circ and \ominus in the last columns denote whether the qualitative criteria were either: fully, partially or not at all met respectively. Note that SVD is listed twice because SVD is considered both as a tool for dimension compression as well as projection. LDA and the fuzzy hierarchical clustering algorithm were not utilised in this fashion since the projection is poor - refer to graphics in Appendix G on page 393.

Tab. 5.2: A summary of the qualitative evaluation: \circ denotes criteria is marginally satisfied, \oplus denotes criteria is satisfied and \ominus denotes criteria is not satisfied; note that SVD is listed twice because SVD is considered both as a tool for dimension compression as well as projection.

Purpose	Algorithm	Subjective Criteria				
		One	Two	Three	Four	Five
Reduction	SVD	\oplus	\circ	\oplus	\oplus	\circ
	FHC	\oplus	\circ	\ominus	\oplus	\oplus
	LDA	\oplus	\circ	\ominus	\oplus	\circ
Projection	MDS	\oplus	\ominus	\oplus	\ominus	\oplus
	CA	\oplus	\circ	\oplus	\oplus	\circ
	SVD	\oplus	\circ	\oplus	\oplus	\circ

Qualitatively, each algorithm under consideration meets the mixed membership model criterion in some form or another. By itself, hard hierarchical clustering does not meet this criterion, since all documents are assigned membership to one cluster only. Thus, the *fuzziness* of the selected fuzzy hierarchical clustering algorithm ensures that the hierarchical clustering approach is not excluded.

In contrast, provision of descriptor label candidates is variable across algorithms. Subjectively, LDA topic descriptors tend to form cohesive sets; conversely, the above-noise threshold selection technique for LSA dimensions produces sets that do not form cohesive sets.

Utilising the most highly activated words above-noise in the LSA term space matrix, tends to bias toward unexpected keywords that carry very limited indication of context.

A review of the top twenty to thirty LSA descriptor candidates confirms that while there are useful words present, there is no obvious pattern suggestive of where appropriate mid and upper threshold values should be set to avoid terms that are too unique or too general. Furthermore, descriptors produced by the hierarchical fuzzy clustering algorithm, also tend to be unique and overly specific, whereas even some appropriate descriptors like 'NASA' are lost during pre-processing - if the recommended frequency filtering pre-processing steps for the HFC algorithm per Leibbrandt (2009) are followed. In general, variations in pre-processing greatly influence the outcome of topic modelling and therefore descriptor generation. This situation is further complicated by the observation that there are seemingly no concrete guidelines for pre-processing in document spatialisation construction.

A good descriptor label set is one which defines the essence of the topic or latent theme - i.e. a portion of context words - and which at the same time differentiates the theme from other slightly related themes - i.e. a portion of specific words. The process of descriptor generation is made harder when the candidate set size is restricted to say, ten descriptors. A small set is desirable both theoretically in line with the work of Polowinski (2009) and Pfitzner, Treharne, and Powers (2008) - who recommend around 7-9 descriptor words or 2-3 times the number of words typically used to search for a document, and pragmatically, based on screen real estate constraints effecting the design of human-computer interfaces.

In regards to criterion three, each algorithm provides document association scores that may be visualised directly as in the case of SVD, or can be visualised by way of a projection algorithm such as MDS or CA. Therefore, since each reduction and projection approach meets criterion three, several additional qualitative sub-criteria were introduced in order to optimise spatial layout. These additional criteria included the degree of spread, and the degree of noisy data about the origin point. An ensemble of graphs depicting a spatialisation for each reduction and projection approach on which these subjective assessments were made is included in Appendix G on page 393.

Correspondence analysis is favoured based on the additional aesthetic criteria as CA utilises white space more effectively than SVD, but less so than MDS. Moreover, spatial annotation is more intuitive than that by SVD while annotating at all is not obviously possible by MDS. MDS has an optimal spread of points in the available area, as does CA, though SVD tends to produce inconsistently sparse layouts in which points compress about the origin point. In a practical situation, an observer would need to make several zooming or panning operations before interacting more closely with an SVD; in contrast, an MDS provides a better global overview with typically lower incidences of point compression.

In regards to criterion four, projection and reduction approaches - except those involving MDS - provide topic-descriptor association scores that could be used to locate descriptor labels to the spatialisation substrate. While both SVD and CA provide

topic-descriptor association scores, subjectively, CA meets this criterion more comprehensively than does SVD, due to superior point spread about the coordinate space.

Descriptor annotations on the spatial substrate provide an immediate cue to the thematic content of a particular region and such cues are important for supporting navigation and exploration of the space. While SVD does support labelling of the spatial area, informal experiments have shown that the SVD document space is more readily compressed about the origin and furthermore, several outlier clusters and singletons receive little semantic annotation. A labelling approach has not been found for the MDS projection algorithm; thus, in place of incorporating dummy label documents into the input matrix, there is - seemingly - no known procedure that prescribes how to label the spatialisation produced by the MDS algorithm.

Finally, in regards to criterion five, the outcome is multifaceted. On the one hand, algorithm selection should favour fast processing such that a future system could utilise the selected approach to respond to ad hoc search queries in real time. On the other hand, the selected approach should not necessarily produce poor, bloated arrangements in order to optimise on processing time; document spatialisation should capture the dissimilarities between groups of documents and differentiate within groups to direct attention to particular facets of a specific document cluster.

Ultimately, algorithm selection based on the final criteria cannot be adequately justified without an optimised implementation of each algorithm. In this case, very few optimisations have been installed in an effort to stimulate exploration of the algorithm space. Advantageously, in this case, a restriction on the maximum number of documents i.e. smaller than 1000 given that similar restrictions are in place on commercial web-based search engines, means that we do not have to devote massive processing times in order to deal with huge corpora. If serving realistic search needs, sub second processing times are needed to compete with existing search engines, as searchers are simply not prepared to wait for their results. Whilst LDA and LSA can produce topic models offline, a projection of a subset of the data may have to be built in real-time to service ad hoc search.

Objective Evaluation

The qualitative evaluation in the previous section involved a level of subjectivity. Therefore, this section attempts to offer an objective comparison from a perspective that was initially motivated by the methodology of Newman et al. (2009). However, they raise that there is no single metric to analyse objectively a visualised topic model. Therefore, the approach taken here is that of average Spearman's Rank Coefficient.

The methodology of Newman et al. (2009) proposes to compare for each document, the distance distributions of a subset of the closest documents on the spatialisation, appearing in the subset of topically closest documents in the topic space; and for each

document, the distance distribution of a subset of the closest documents in the topic space appearing in the subset of closest documents in the spatialisation. Accordingly, if documents are close in the topic space, then they should be close in the spatialisation; and if documents are close in the spatialisation then they should be close in the topic space. Topically closest and spatially closest subsets correspond to 10% of the corpus size relative to each point.

The notion of a neighbourhood of points around each candidate document is important since one may liken this to a user's visual attention radius. However, if starting at a particular useful document and working outwards, we would expect to visit the documents closest to the central document first on our spiral outwards since our spatial reasoning would suggest that these are the documents that share the closest semantic resemblance, given that they are separated by the least distance. Thus, this evaluation will measure how well the map space preserves the information in the topic space from the perspective of pair-wise ranked distances.

The analysis procedure is as follows. Pair wise comparisons are made for distances between each document in both the topic and map spaces. Each distance is ranked and ties handled by assigning the mean rank of the tied documents. Spearman Rank Coefficient is taken for the rank of distances around each document in topic and map space.

There are two ways to calculate Spearman's rank coefficient and the correct way depends on whether ties are present in the rankings. In this case, as ties are known to exist, the alternative approach is taken:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

As the measure approaches 1.0, we can consider the map space to be faithfully preserving the layout of documents as denoted by the topic model. Where the measure approaches zero, the map space bears little resemblance to the topic space. As correlation approaches -1.0, we can imagine a layout opposite to the topic space, which would also be of interest.

Results presented in Figure 5.1 on page 253, Figure 5.2 on page 253, Figure 5.3 on page 254 and Figure 5.4 on page 254 show box plots for each task set and collection of algorithms. Each box plot provides a distribution of Spearman's rank coefficient values for all pair wise comparisons in the topic space and map space while Table 5.3 on page 255 summarises the average Spearman's rank coefficient for projection approaches of the topic space generated by a Latent Dirichlet Allocation approach, across different task sets. The results are averaged over 50 runs for each task set due to random initialisation in both the LDA and MDS algorithms.

A single run consists of topic space construction and spatialisation construction and then an analysis of correlation of the rank of document distances from the perspective

of each document in the topic space versus the same document in the spatialisation. The projection approach that preserves a similar ordering in the topic space will be the best candidate for a projection of the topic space and have a score that approaches 1.0.

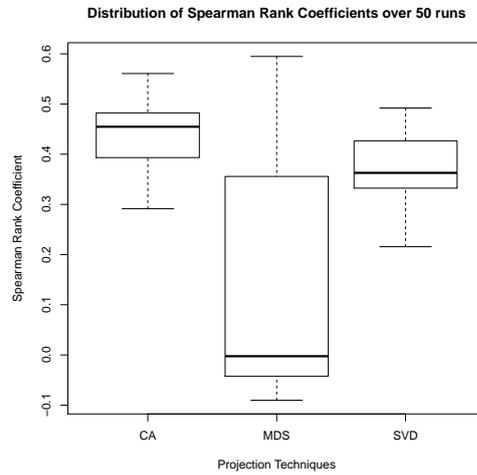


Fig. 5.1: A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Hong Kong Hand Over Task Set

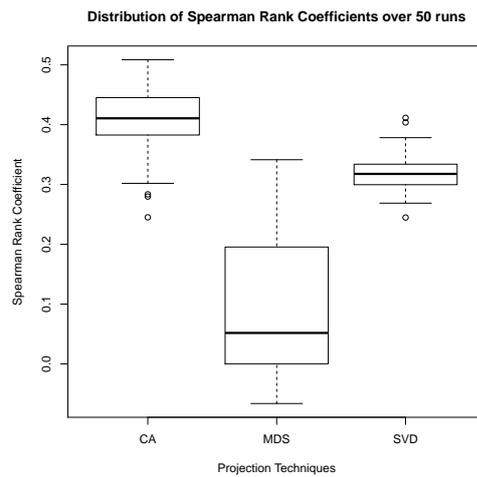


Fig. 5.2: A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Newspaper Circulation Task Set

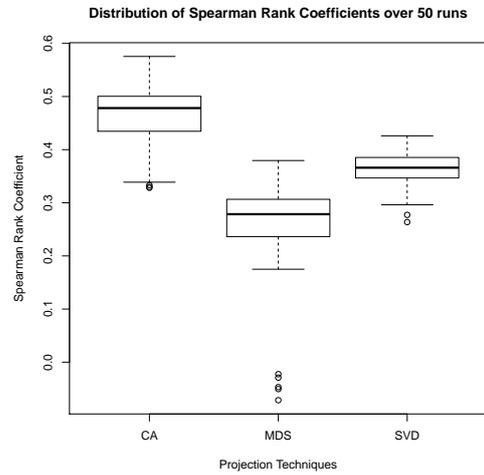


Fig. 5.3: A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Recycled Materials Task Set

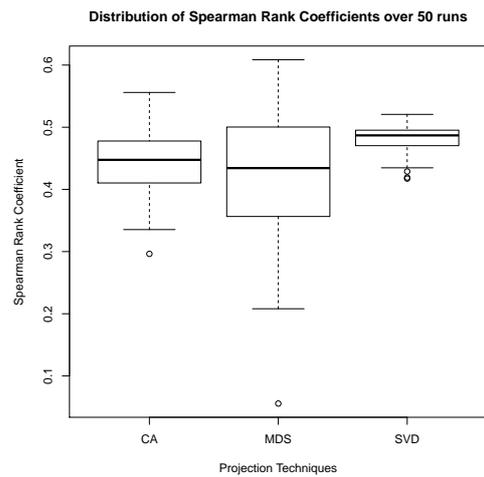


Fig. 5.4: A box plot graph of Spearman Rank Coefficient distributions for three projection techniques for the Space Shuttle Task Set used for training purposes.

Tab. 5.3: Mean Spearman Rank Coefficient for document distance ranks in LDA topic space versus document distance in map/projected space for four task sets; values are averaged over fifty runs due to randomisation in techniques; RM denotes the recycled materials task set, HK denotes the Hong Kong sovereignty task set, NP denotes the decline of newspaper circulation task set and TR denotes the scientific instruments on board space craft task set which is utilised for training purposes in Chapter 6.

Task Set	Spearman Rank		
	CA	MDS	SVD
RM	0.461	0.250	0.362
HK	0.444	0.148	0.372
NP	0.406	0.090	0.319
TR	0.443	0.421	0.480

The results indicate that while all algorithms are performing poorly, according to the Spearman Rank Coefficient measure, the Single Value Decomposition and the Correspondence Analysis approaches that both utilise the LDA topic space matrix as input, result in a similar outcome; Correspondence Analysis routinely achieves a slightly better performance. In contrast, the multidimensional scaling approach is routinely last. Furthermore, the outcome of the MDS is more sensitive to the task set whereas the Correspondence Analysis and Single Value Decomposition approaches are more stable as evidenced by the shifting median in the box plots below.

This objective evaluation approach favours a comparison of the CA and SVD projection approaches but possibly not the MDS due to the difference of input. Whereas the CA and SVD take as input the topic space matrix consisting of topic association scores for each document, the MDS takes as input a dis-similarity matrix of inter-document association.

5.9.5 Selected Approach for Information Space Construction

The intended outcome of this evaluation process was a selection of a reduction and projection approach for constructing an information space to organise text documents into a spatial arrangement. The projection technique chosen is correspondence analysis and the reduction algorithm chosen is Latent Dirichlet Allocation. A brief summary will now follow which argues as to why this approach to information space construction was selected and how these techniques will be utilised to construct an information space for the subsequent experiment in Chapter 6.

First and foremost, the LDA/CA combination meets each of the qualitative criteria outlined above - see Section 5.9.4 on page 246 - in addition to achieving a perceptible cleaner arrangement of documents. Objectively, Spearman Rank Correlation indicates

that correspondence analysis preserves topic-space similarity in the spatialisation in a comparable way to SVD, and potentially, more faithfully; and the qualitative measures are sufficiently better as well.

The pre-processing technique can be summarised as follows. A collection of full-text news articles are filtered to remove duplicates, overly long documents and documents classified as belonging to specific categories. Each article is then entered into a Lucene search index - see <http://apache.lucene.net>. A set of queries consisting of main topic words and qualifying words are used to extract a pool of documents for processing. A set of named entities for each document are devised from the document's full-text by way of the Stanford Named Entity Extractor (Finkel, Grenager, and Manning, 2005). The full-text is tokenised and delimited by punctuation marks - excluding hyphens. Then with the exception of stop words, all tokens and named entities are collated into a raw term frequency matrix.

The Latent Dirichlet Allocation library (Phan and Nguyen, 2008; Phan, Nguyen, and Horiguchi, 2008) takes as input, a file consisting of each document's tokens in string form. Each line of this file represents a document and each space separated token represents each word contained within the document. This file is generated from the term-frequency matrix.

The LDA process produces 15 topics and 15 descriptor words for each topic, document-topic association scores and term-topic association scores. Algorithm parameters are set according to the author's recommendations. Next, the LDA document-topic association matrix is passed to the correspondence analysis software - see <http://classification-society.org/csna/mda-sw/correspondances/>. The correspondence analysis process produces document and topic ordinations and a hard hierarchical clustering for documents and topics. The dominant topics for each projection dimension are identified by the topic-projection association. The topic having the greatest absolute association with a projection dimension is considered to uniquely identify the projection dimension, regardless of other strong topic-projection associations.

Next, a Latent Semantic Indexing model, based on a decomposition of a log-transformed and row demeaned version of the term-frequency matrix is generated. This model will serve the ranked-list baseline condition in the experiment reported in Chapter 6.

Lastly, each of the outputs are processed into an experiment task set structure containing the projection dimensions, text documents, topics, and descriptors. Each document contains a set of metadata including title, query snippet, document full-text, cluster membership data and coordinates for the spatialisation. Similarly, each topic's spatial coordinates and set of descriptor words are stored in this structure.

5.10 *Summary*

This concludes the discussion on theoretical aspects relating to information space and document spatialisation. This chapter has focused on what an information space is and the major components that make up an information space. Furthermore, a description and motivation for an approach to constructing an information space for a small set of search results was proposed.

The next chapter will build on this chapter directly; it will present an evaluation of an apparatus that incorporates an information spatialisation as the basis for a search result visualisation. The evaluation will not focus on the utility of spatialisation as such; rather, it will explore factors that may improve interaction with and use of information spaces.

Nevertheless, this chapter has provided the foundation over which usability research may take place. Without such preliminaries, attempting to improve of the usability of information spaces cannot happen.

6. A LABORATORY-BASED EVALUATION OF INFORMATION SPACE USABILITY

6.1 Introduction

This chapter presents an experiment which has the aim of investigating three usability factors impacting search behaviours observed during interactions with a document spatialisation. It is not the intention of this experiment to confirm the utility of document spatialisation as a basis for search tools; rather it is intended to improve the usability of search tools incorporating spatialisation.

The first factor relates to how and where a searcher gains access to a search result's full-text. There are two full-text view levels: integrated - adjacent to the search results - and modal - in a separate modal frame isolated from the search results. The second factor relates to the level of transparency that a pop-up window has when overlaid on a document spatialisation. There are two transparency levels: semi-transparent and non-transparent. Semi-transparent windows reveal content otherwise obscured by the pop-up window background, in contrast, non-transparent pop-ups obscure content below the pop-up window background. Finally, the third factor relates to the interactive control, which facilitates changes to the participant's perspective of the information space. By changing the perspective of the information space and therefore document layout on the spatialisation, participants are able to see different clustering patterns across the search result set. There are two interactive control types: a theme cloud control and a theme list control; the theme cloud looks similar to tag cloud visualisations while the theme list control looks similar to a faceted list.

A data analysis will seek to identify differences in user performance and behaviours based on measures including: time to complete each task and answer set quality, as well as interaction behaviours like the number of documents opened, pop-up window usage, and the number of perspective rotations made. An optimal search performance is defined to be that which maximises answer set quality, given a limited time and effort allowance.

The next section outlays the main aspects of the experiment apparatus that is devised to test the aforementioned usability factors. Each factor is discussed at length, as it is appropriate to introduce them with adequate context. In contrast to the nomenclature of the previous chapter, participants were exposed to the phrase *theme map*

in place of the phrases *information space* and *document spatialisation*, when learning about the experiment apparatus. Accordingly, here on in, *theme map* and *information space* may be used inter-changeably.

6.2 Apparatus

The apparatus for this experiment is complex and so effort is devoted to explain each component in detail with appropriate graphics. Discussion is broken into roughly each of the three main stages a participant completes over the course of the experiment. Within each stage, the design motivations underpinning the manipulated experiment variables are highlighted. In the first stage, participants interact with a ranked-list of search results; in subsequent stages they interact with a theme map visualisation. However, whilst the method of search result visualisation changes across stages, the manner in which a participant accesses document full-text remains consistent across stages, and so is discussed first.

Since the document full-text view impacts greatly on the interface layout, screen shots of the apparatus are introduced in the following section; these screen shots will be referred to in later sections.

6.2.1 Document Full-text View

Reading a search result's full-text is critical for the satisfaction of an information need. The apparatus displays a document's full-text in the document view.

In many of the systems surveyed in Chapter 2, there is very little indication of how a full-text article is accessed; but it is usually assumed that the full-text appears in a separate window, frame, or tab. In browser-based Internet search engines, a ranked-list provides a collection of URLs that searchers click on to open documents - either in place of the ranked-list, in a new browser window or in a new browser tab. Seemingly, this is done to support a diversity of screen sizes; though this creates a disconnect between the set of search results, and documents that the searcher selects for full-text view.

Moreover, at present, there is little sophisticated support for interaction with the full-text view, that enacts a change in the order of, or in the content of, a search result set. Yet, interactions such as selecting words, paragraphs of text, images, links and concepts in the document full-text view, could provide a basis for the manipulation of result order or content of the result set, in real time.

More recent web browser versions do support text selection in documents to initiate search sessions, although these actions do not have an influence on an existing search result set. This experiment is a first step in investigating behavioural changes for different full-text view configurations, with a vision to increase the degree of control over a search result list via interactions with a document's full-text.

In experimental interface research, the document view is often integrated with the result visualisation (e.g. Cribbin and C. Chen, 2001; Wu, Fuller, and Wilkinson, 2001; Fortuna, Grobelnik, and Mladenic, 2005), superimposed under or over the result visualisation (e.g. Leuski and Allan, 2000) also see <http://search.tianamo.com> - although these two examples only show surrogate information and not full-text - thus retaining some context of the result set, or modal or separate in a new tab or window (e.g. Nowell, Schulman, and Hix, 2002; Sutcliffe, Ennis, and J. Hu, 2000). There is - seemingly - little research that has evaluated alternative document full-text view configurations. Nevertheless, a proportion of research has investigated the role of multi-tabbed browsers and their role in the information seek process and more generally, the role of multiple-coordinated windows has been investigated from the perspective of multiple coordinated views visualisation. Thus, it would be complementary and interesting to investigate search behaviour and a coordinated document full-text view.

Dubroy and Balakrishnan (2010) investigate the role of tabs in browser windows and observe a clear preference for tabbed browsing, as opposed to browsing in separate windows. A resounding majority of participants in their longitudinal study, make heavy use of tabbed browsing - with only one participant opening web pages in new windows - and nearly half of all participants reporting that tabbed browsing results in a cleaner, more organized and less cluttered user experience. The authors observe: the use of tabbed windows to manage web pages according to category - like work related and personal interest tabs; the use of tabbed browsing for multi-tasking and tabs for frequent tasks or ongoing, short-term tasks; and the use of tabs to queue page downloads in parallel with sequential evaluation of a ranked-list of search results.

Downloading web pages in the background, while the searcher evaluates other search results, may save time for the searcher. By the time a searcher has opened each promising link from a page of results, it is likely that the first-opened result has downloaded at least partially; by the time the searcher turns their attention to the last-opened page, it is likely that there will be no wait time at all.

However, this use of tabs has important implications for search engine evaluation, particularly if time between clicking on successive links has an impact on any implied searcher feedback (Huang, Lin, and White, 2012). If a document is not relevant, then the searcher is likely to click the back button shortly after opening the irrelevant result; in future search sessions for the same query, the search engine may opt to rank that irrelevant result further down the list. In contrast, under tabbed browsing, it is observed that the searcher opens a series of results, reads them in no particular order, and closes each tab if not relevant. The equivalent of the back button under this model is a reorientation to the tab containing the search result page. In this scenario, the search engine has no clear indication of the searcher's reading order or the time spent reading each page.

Huang, Lin, and White characterise multi-tab searching as *branching* and propose

that the searcher's task time is reduced - at least partially - and because the searcher can be more thorough in their evaluation, since they do not have to re-find results higher up the list, for comparative analysis against those lower down the list. In the latter case, leaving promising tabs open and closing off irrelevant tabs builds a set of high quality results over the course of a search session.

A number of different uses for browser tabs are evident. This suggests that a more functionally-rich document full-text view may appeal to searchers, since searchers already leverage these views for more efficient search. One possible offering is to facilitate manipulation of an existing result set using the document full-text view. However, a disconnect between the document's full-text and a search result set - a result of showing document full-text in a new tab thereby hiding the view of the search result set - may make for a blind interaction, in which the searcher is unable to see changes to the user interface as their interactions take place.

Thus, this experiment explores two document full-text view configurations: integrated in Figure 6.2 on page 264 and modal in Figure 6.3 on page 265. If this nomenclature is unclear to the reader, another way of thinking about integrated full-text view is to consider the full-text presented in a separate frame and adjacent to the search results, whilst another way of thinking about a modal full-text view is to consider the full-text presented in a separate window or browser tab. The expected main benefit of the integrated view is that changes to the result set are observed by the searcher as text selections are made through the document full-text view. In contrast, observing change to the result set, following interactions in the modal view, is only possible after having first closed the modal full-text view.

Under the integrated condition, a document's full-text appears adjacent to the search result presentation technique. Figure 6.1 on the facing page depicts a ranked-list interface configured for integrated full-text view; the full-text view appears adjacent to a ranked-list of results. Figure 6.3 on page 265 depicts a modal full-text view of a document's text. When a document is selected for viewing, the document view takes up the whole screen and the user must explicitly close the modal view in order to return to their search results. In contrast, in the integrated view, the searcher can open a document's full-text, without needing to close an already open full-text view. Figure 6.2 on page 264 depicts a ranked-list interface configured for modal full-text view configuration; in this figure no document full-text is open.

In browser-based search, we see a full-text view in what might be considered a modal view in that the searcher must explicitly close the window or tab, reselect the results tab, or click the back button in order to return to the results view.

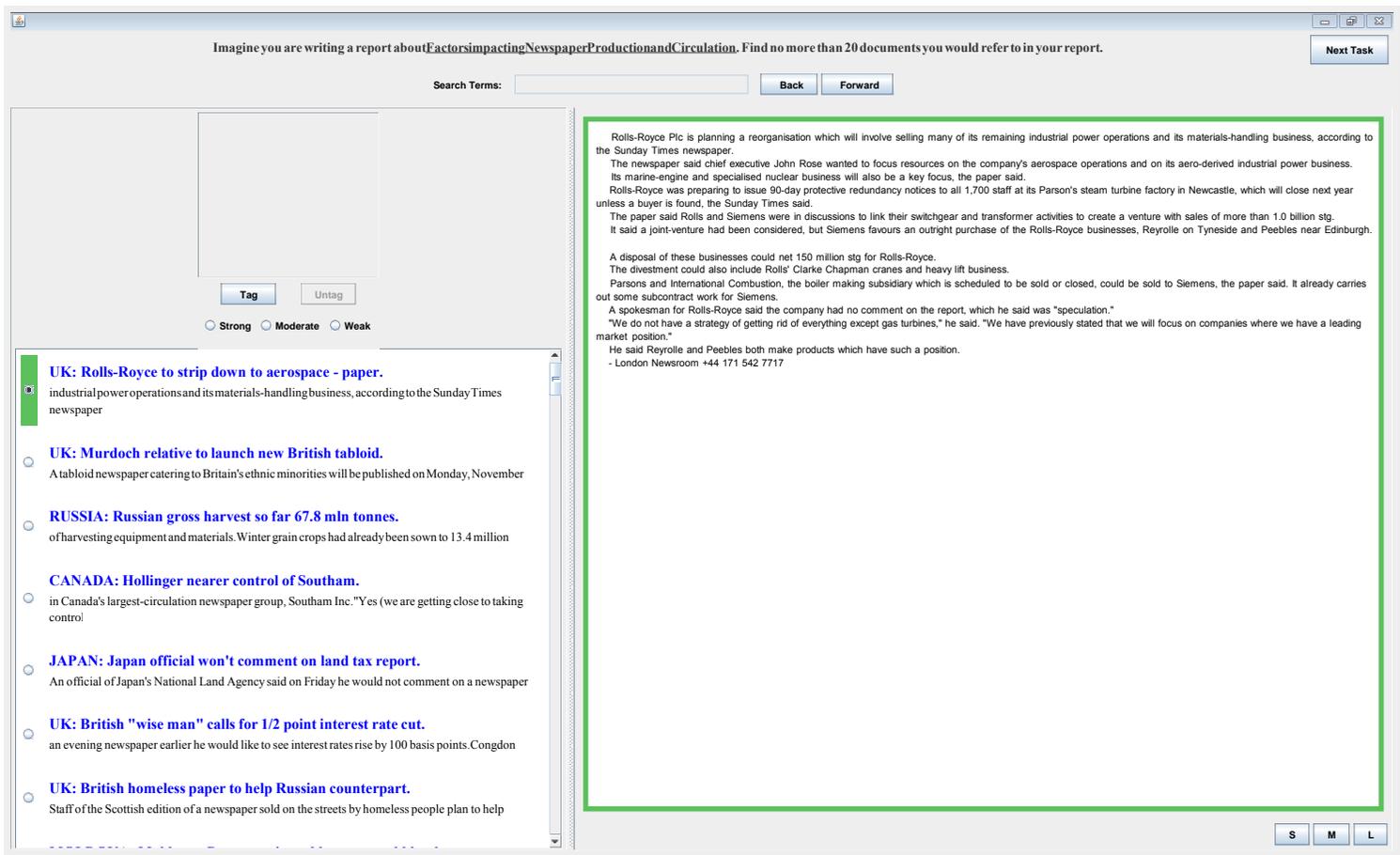


Fig. 6.1: A screen shot of the ranked-list interface with integrated full-text.

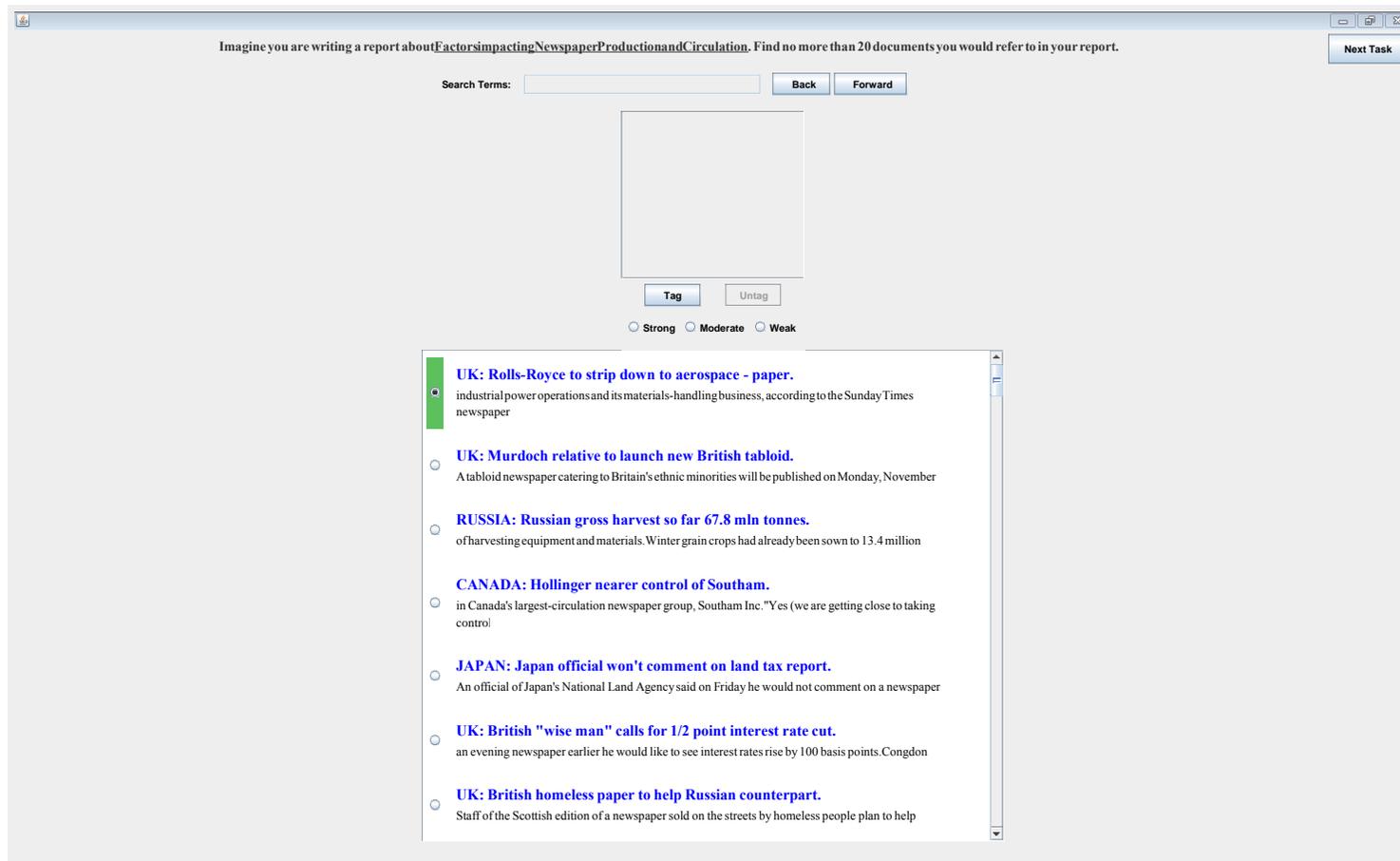


Fig. 6.2: A screen shot of the ranked-list interface with modal full-text.

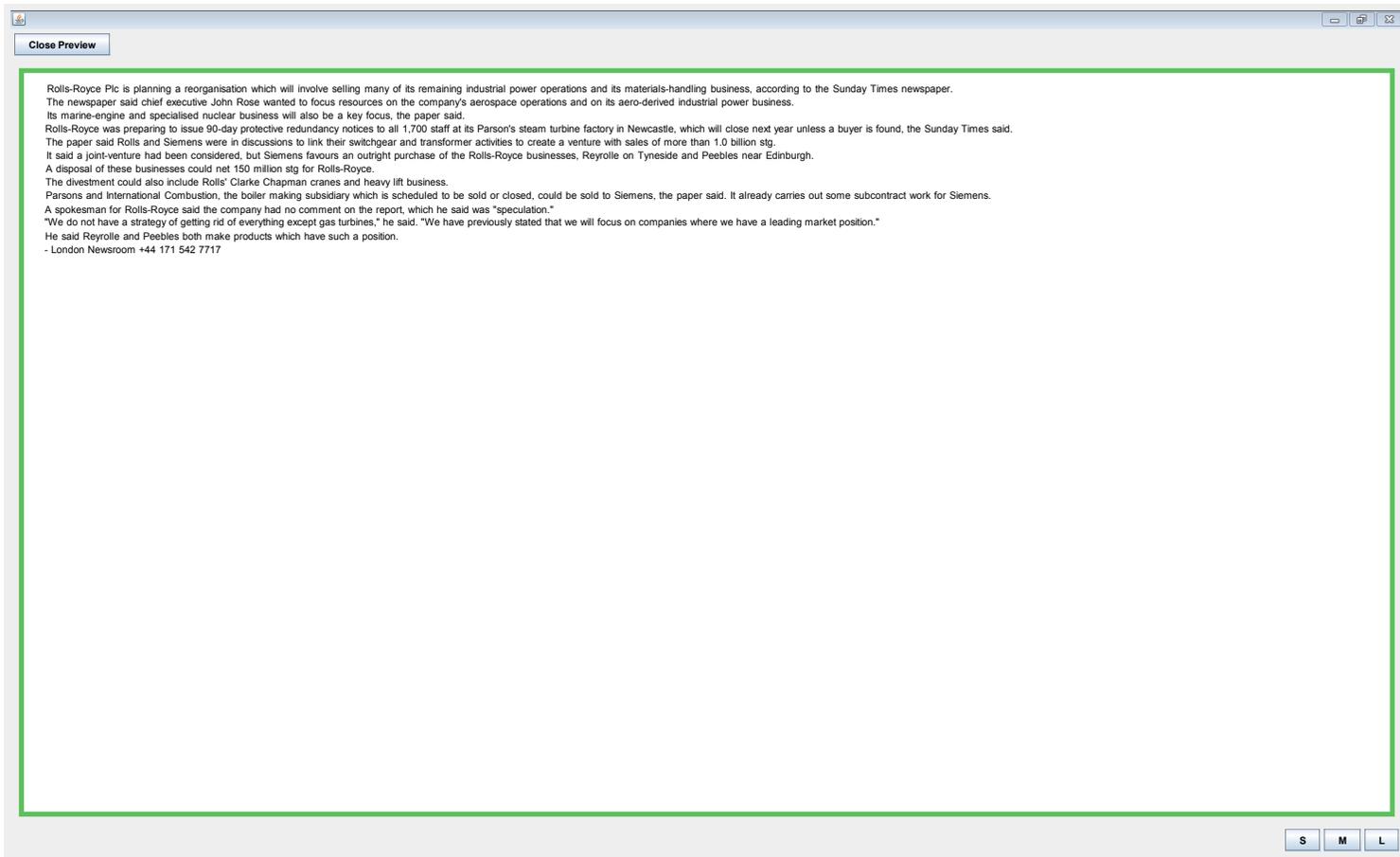


Fig. 6.3: A screen shot of the modal full-text view.

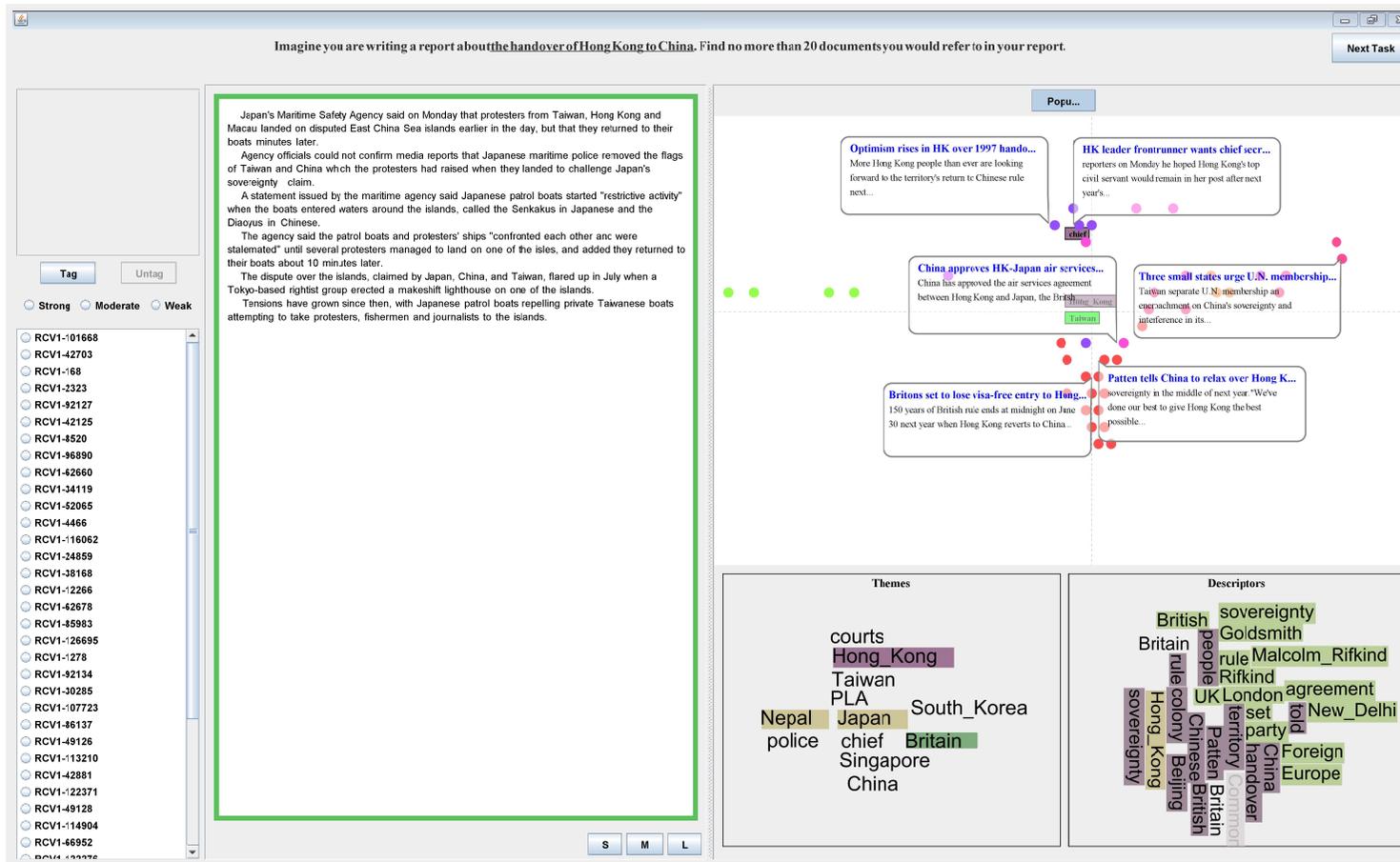


Fig. 6.4: A screen shot of the theme map interface with integrated full-text, transparent pop-up windows and theme cloud control.

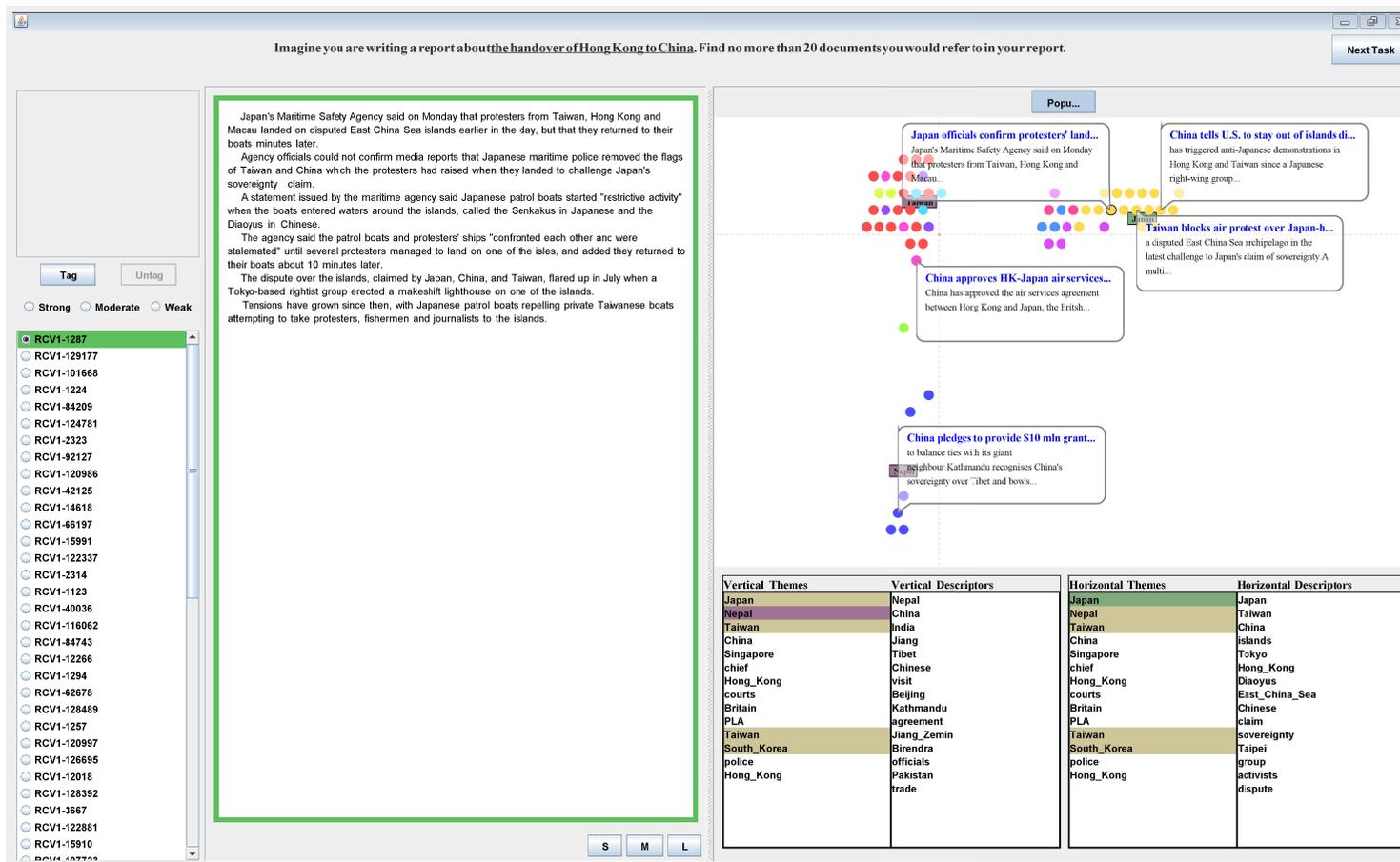


Fig. 6.5: A screen shot of the theme map interface with integrated full-text, transparent pop-up windows and theme list control.

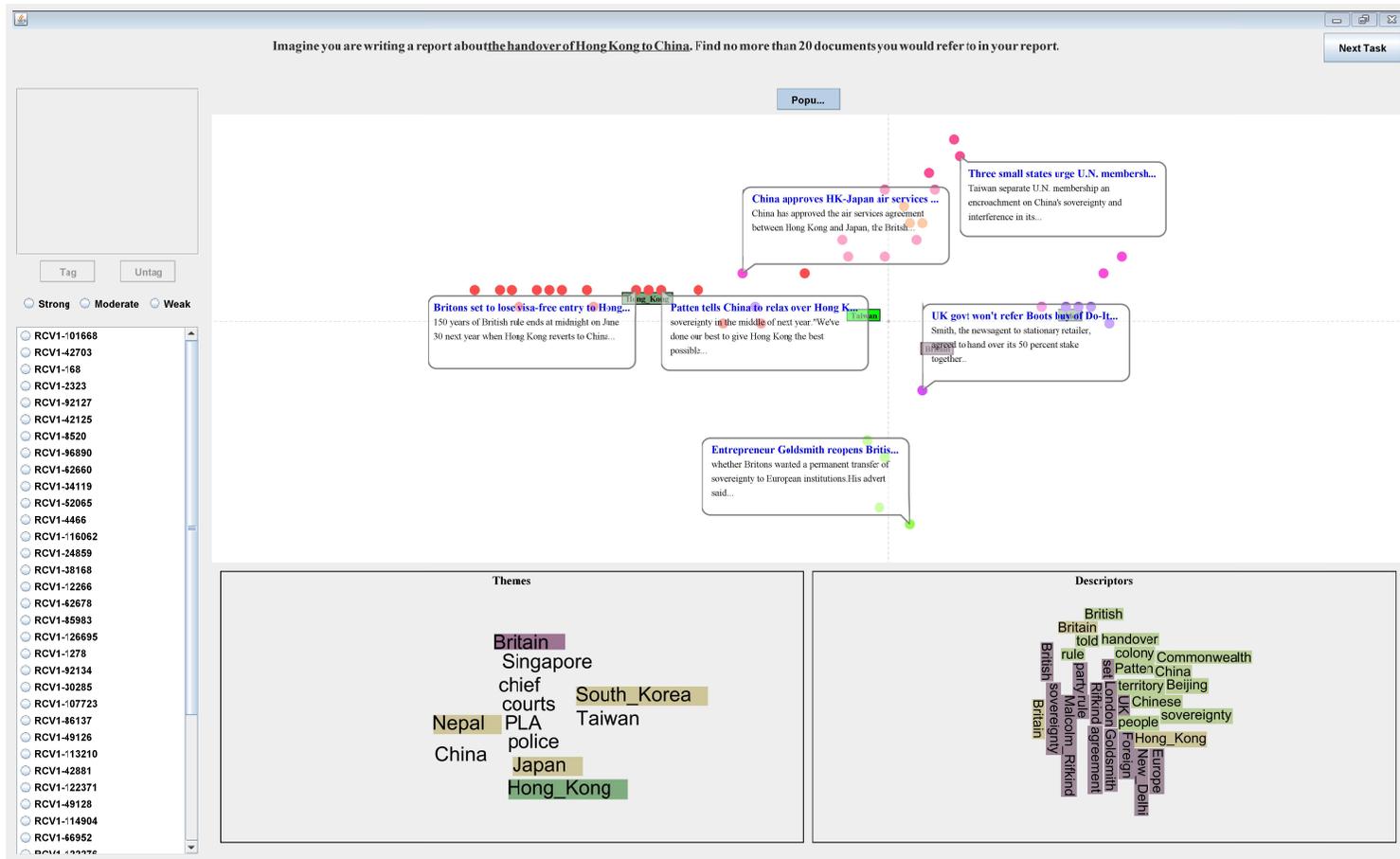


Fig. 6.6: A screen shot of the theme map interface with modal full-text, transparent pop-up windows and theme cloud control.

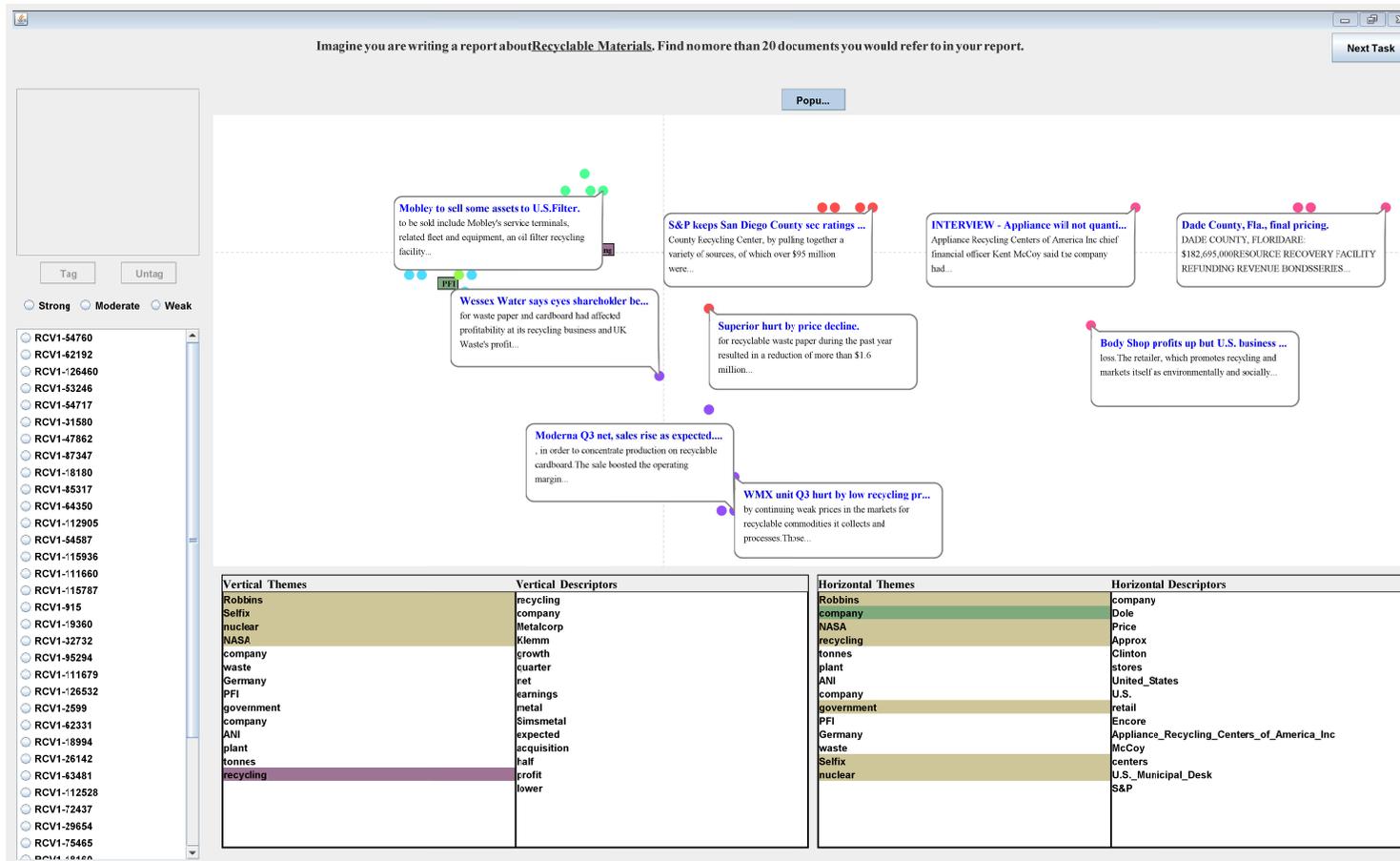


Fig. 6.7: A screen shot of the theme map interface with modal full-text, non-transparent pop-up windows and theme list control.

Figure 6.4 on page 266 and Figure 6.5 on page 267 depict the integrated full-text configuration for the theme map with theme cloud and theme map with theme list interfaces respectively; theme cloud and theme list refer to the type of perspective rotation control that accompanies the theme map visualisation - these will be discussed in a later section. Figure 6.6 on page 268 and Figure 6.7 on the preceding page depict the modal full-text configuration for the theme map with theme cloud and theme map with theme list interfaces respectively. In the modal configuration, the theme map width is greater than the width of the theme map in the integrated configuration. In the integrated configuration, the available screen real estate has to accommodate the document full-text window as well as the theme map. Full-text window size and consequently theme map size may be adjusted by moving the divider left or right.

There are three main methods for opening a document's full-text. On the ranked-list interface, the full-text is opened by clicking on the blue title text of result surrogates in the result list. On theme map interfaces, a user may click on entries in the de-featured list on the left hand side of the interface, however, there is no title text or snippet text to judge document relevance and furthermore, the order of entries is random. In addition, the user may click on document icons in the theme map visualisation to trigger a full-text view.

For every document icon in the theme map there is a corresponding entry in the de-featured list. The de-featured list is only present in the theme map interfaces, to foster a consistent answer submission process; theoretically participants should not need to interact with the list at all. Primarily, a list appears in the theme map condition as a fall back if participants absolutely do not want to interact with the visualisation.

Finally, regardless of view type, the document full-text view is formatted according to the original news article's paragraph structure - to retain the author's intended flow. The full-text view's font size is configurable by clicking one of three buttons at the bottom right corner of the document view window and this setting is recorded by the software and remembered for subsequent article views.

6.2.2 *Ranked-list Interface*

Participants initially interact with a ranked-list interface. An example of the ranked-list is presented in Figure 6.8 on the next page though screen shots in Figure 6.1 on page 263 and Figure 6.2 on page 264 show the ranked-list in full context. Participants interact with this list much like they do at web-based search engines.

The design of the ranked-list closely matches that found at web-based search engines in terms of colour coding, keyword highlighting, the number of snippet text words, the number of results visible on screen at once, and the overall width of each list item. Keyword highlighting of the snippet content is intended to support visual scanning of

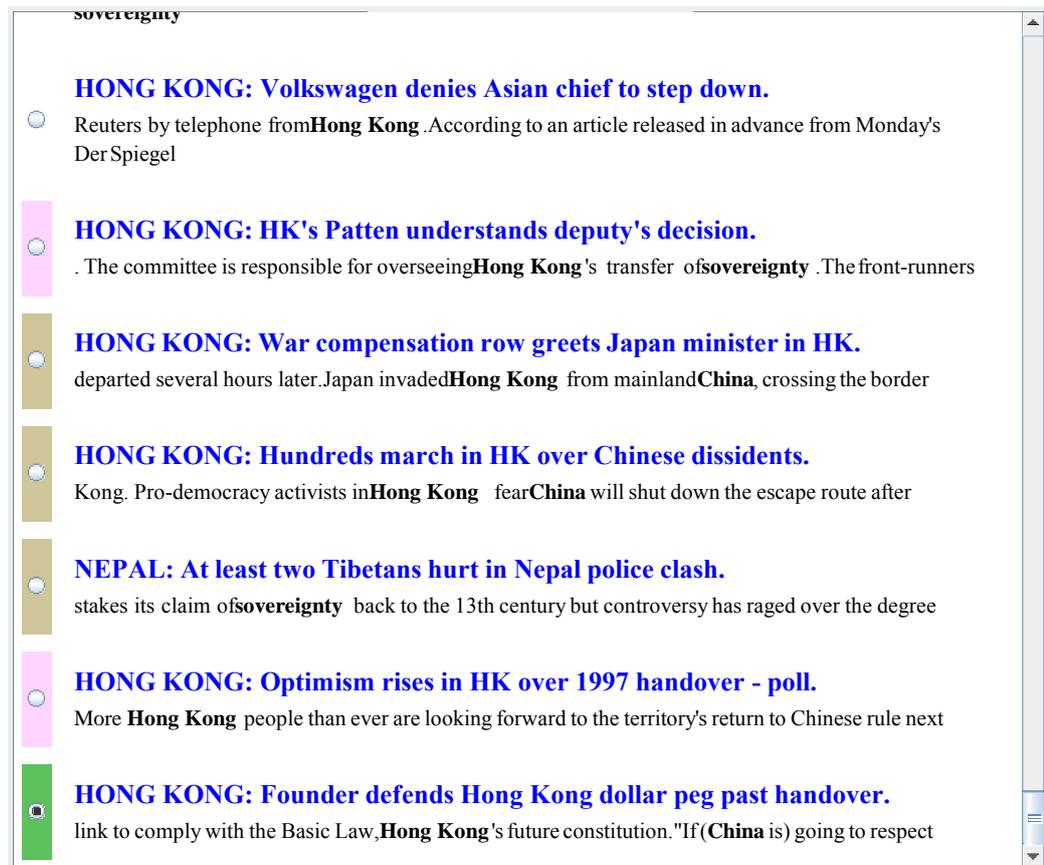


Fig. 6.8: A screen shot of the ranked-list; each document has a title and snippet text - search terms are emphasised in bold face; highlighting at left of item indicates currently open as green, previously opened as khaki and selected for answer as pink.

the list, while colour coding at the left of each result surrogate indicates whether the searcher has previously read or submitted the result as an answer.

The major differences between web-based search engine interfaces and this interface are that results do not include a URL address and each result has a radio button to facilitate the submission of results into an answer box. Furthermore, there is no result pagination; instead, participants scroll down the list to see additional results.

List order is calculated by way of Latent Semantic Indexing, which was briefly introduced in Chapter 5. Participants can manipulate the sort order of the list by selecting words from document full-text. When a word or sequence of words is selected, a re-ranking vector is generated and compared against each document in the semantic model. Cosine similarity distance between document vectors and the re-ranking vector prescribes a score for each document; documents that are similar to the ranking vector appear at the top of the list. Initially, list order is calculated using keywords present in the task statement.

6.2.3 Theme Map Interface

Participants subsequently interact with each of two theme map interfaces; in these stages, a theme map provides the primary mechanism through which to interact with search results. There are several components that appear on or are linked to the theme map visualisation. These include document icons, topic and descriptor labels, document pop-ups, and the coordinate axes. Each component is discussed in relevant subsections below.

Figure 6.9 on the facing page is a screen shot of a theme map for the *Recyclable Materials* task set. Each coloured circle represents a news article and the rectangular shapes represent topics. In this example, the selected projection dimensions are labelled as ‘waste’ and ‘company’. Topics appear in the theme map area; the ‘waste’ topic appears at the top right of the figure adjacent to the orange article icons and the ‘company’ topic appears at the bottom right of Figure 6.9 on the next page amongst the pink article icons. These topics have the greatest weight or association to the selected projection dimensions.

Other topic labels appear on the theme map visualisation though these are not as dominant or highly associated with the selected projection dimensions. The searcher can judge the dominance of topics by inspection of the theme map. Labels on the outer extremes or edges of the theme map correspond to the most dominant topics. Labels that are positioned further toward the centre of the visualisation are less dominating or descriptive of the projection axis.

Figure 6.10 on page 274 is a screen shot of the training set theme map for the *Scientific Instruments on board Space Craft* task test.

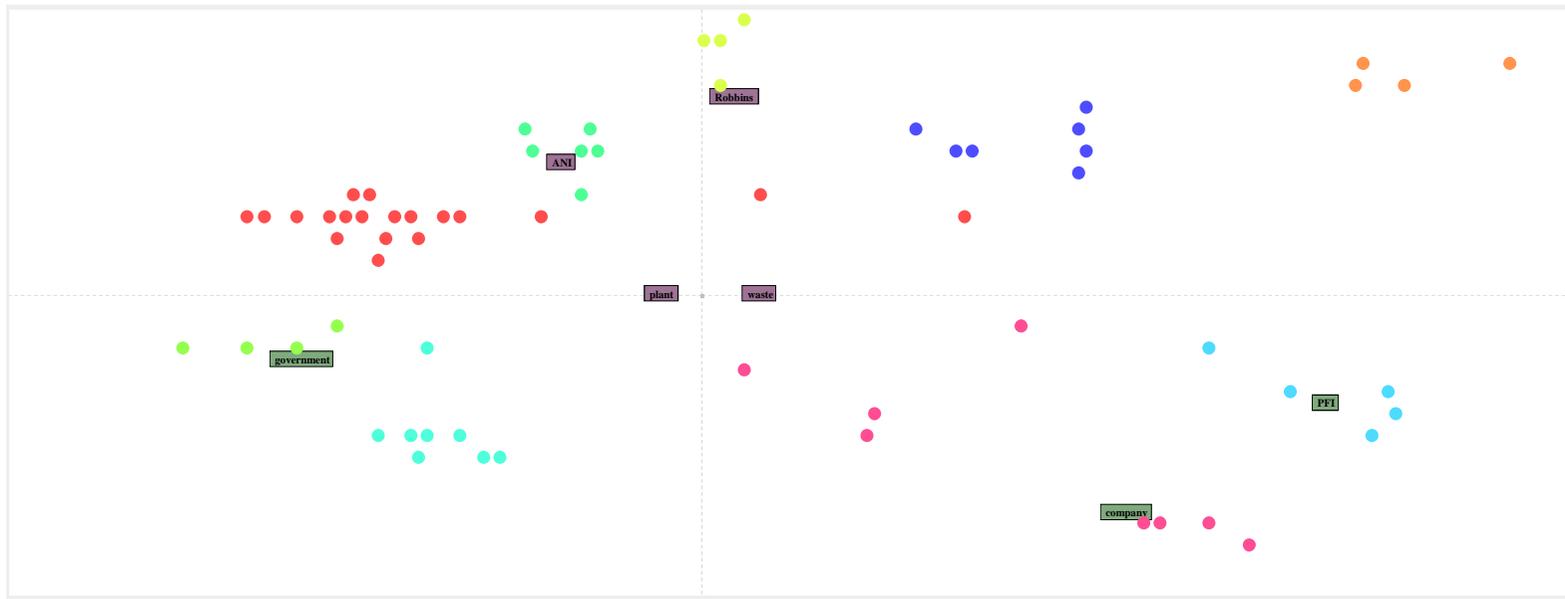


Fig. 6.9: A theme map of the *Recyclable Materials* task set; in this theme map, the layout of icons appears less grid-like relative to Figure 6.10 on the following page in which the icons are not offset.

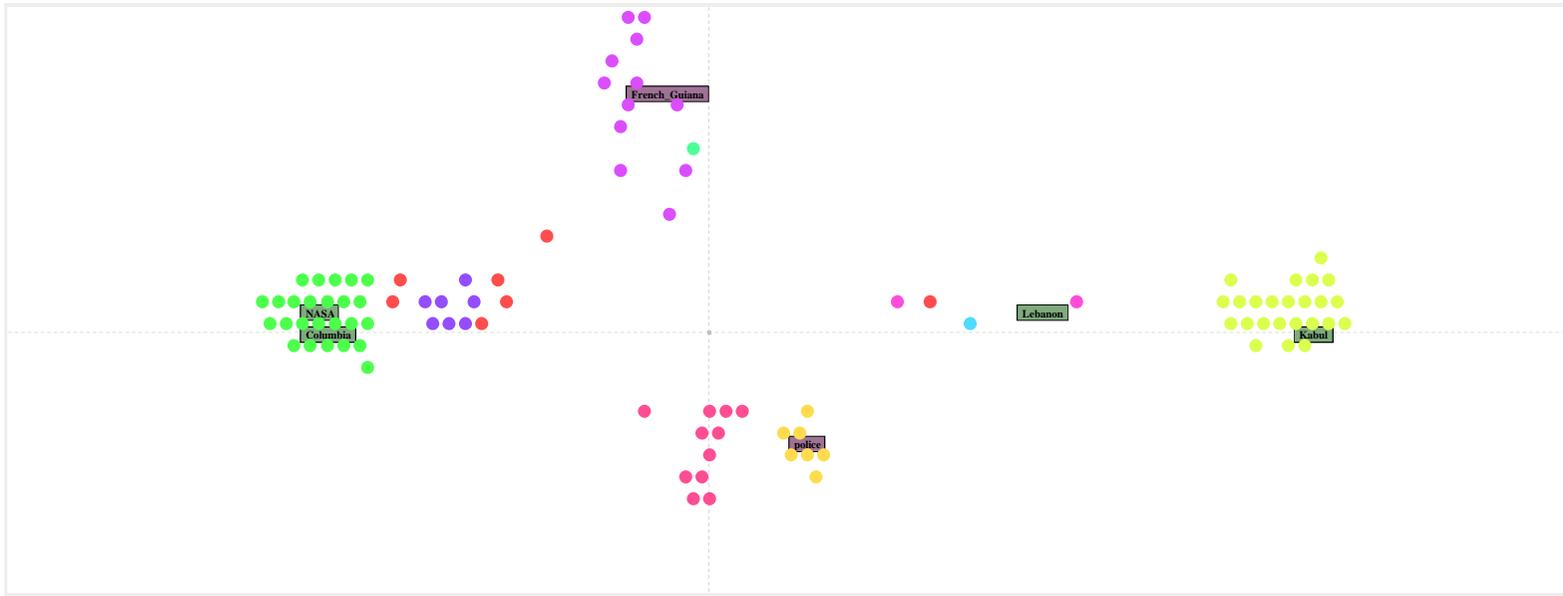


Fig. 6.10: A theme map of the training task set; in this theme map, the layout of icons appears more grid-like relative to Figure 6.9 on the preceding page in which the icons are offset.

Document Representation - Colour and Occlusion

Notwithstanding the earlier investigations into motion and naturalness encoding in Chapter 3 and Chapter 4, the representations of individual documents in this experiment appear in a reduced form. It is intended here to focus on document relationships, defined by semantic content, rather than metadata similarity. As such, the interface represents each document as a circular shape. If a document is tagged as an answer, a '*' is drawn in the centre of the shape. If the document has been opened but not tagged as an answer an 'o' is drawn in the centre of the document icon.

Colour coding reveals the result of a post hoc hierarchical clustering of articles. Using Figure 6.9 on page 273 and Figure 6.10 on the facing page as examples, articles clustered by the hierarchical clustering in the correspondence analysis procedure, tend to reinforce the spatial arrangement - i.e. colour coding is approximately the same in local regions of the spatialisation. However, there are projection configurations in which spatially defined groups are coloured heterogeneously - i.e. reflecting differences in similarity according to the spatialisation algorithm and the hierarchical clustering.

This is evident in Figure 6.10 on the preceding page around the 'NASA' topic. Here, heterogeneity in colour coding is particularly interesting, since if results of the hierarchical clustering were presented as a list of clusters in the fashion of Hearst (2006) the interaction of these three clusters - depicted as dark red, yellow and dark green - would be hard to identify. Thus, from one perspective, a correspondence analysis and spatialisation indicates the presence of cohesive groups, but a rotation of perspective changes the layout of document icons, such that clusters can become more homogenised by colour - thereby indicating interactions between groups. This potentially benefits the searcher by fostering serendipitous search.

Initially, <http://www.colorbrewer2.org> was consulted for a theoretically motivated colour set for colour coding of document icons. ColorBrewer (Harrower and Brewer, 2003) intends to provision colour sets for cartographic visualisation purposes, such that background colour coding provides appropriate contrast with foreground textual labels and geographic overlays - thereby optimising foreground legibility.

A twelve colour, qualitative colour set obtained from ColorBrewer was trialled in the experiment interface. However, the white background did not contrast well and not all document glyphs were equally distinguishable or identifiable against the white background. Therefore, on the recommendation of Harrower and Brewer (2003) - that qualitative colour sets are most optimal when hue is varied in combination with constant lightness or brightness and neither high nor low in constant saturation - a new colour set of 15 colours was calculated with each hue having a one-fifteenth separation in colour space - corresponding to a 6% hue separation, 100% lightness/brightness and 70% saturation. The Just Noticeable Difference or JND of hue is around 3% (Lubin and Fibush, 1997, pg. 25 the average of the model is approximately 3%); therefore

use of 6% hue separation can be seen as a separation of 2 JNDs. This culminated in a colour set of 15 colours that were distinguishable and identifiable against a white background i.e. the colour set provides a suitable contrast between document glyphs and background. The colour set is depicted in Figure 6.11. Colour of spatialisation annotations were allocated markedly lower saturations in an effort to differentiate more vibrant coloured document glyphs and annotations.

An alternative method to devise a colour set is to consult colour processing theories. The colour opponency process discussed in a previous chapter, is suggestive that a viewer will be most likely to disambiguate items coloured if the colour sets include those of the opponent pairs - therefore blue-yellow and red-green and black and white channels. These colours are recognisable, are known by a familiar label and are generally not distinguished as a mixture of any two colours (Brewer, 1999).



Fig. 6.11: Colour set for document icons; colour denotes differences in cluster membership; this colour set is made up of 15 hues of constant full brightness and constant 70% saturation.

Overlapping or occluding icons make icon selection difficult. A trade off exists between the creation of icons that are ‘pickable’ such that the user can, with ease, mouse-click on the icon and the sizing of icons such that they do not occlude one another.

Ellis and Dix (2007) propose a taxonomy of clutter reduction techniques which offer strategies for dealing with occluded displays. Their taxonomy spans sampling, filtering, icon size modification, use of transparency, clustering techniques, point displacement, topological displacement and animation driven techniques.

Sampling and filtering techniques involve only the display of representative candidates or candidates that meet criteria; icon size modification reduces the likelihood that two icons will overlap; and icon transparency can reveal wholly obscured icons - but only if icon outlines are non-transparent. However, such strategies make no guarantee that occlusion will be wholly dealt with - nor guarantee easy picking. Conversely, clustering techniques can amalgamate overlapping icons into a single icon that might be ‘exploded’ or opened to reveal content in another view; point displacement techniques shift icons around to nearby locations such that they are non-overlapping; topological distortions - driven by interactive controls - make temporary zoom or stretch distortions of the spatial coordinate system to reveal differences between clumps of overlapping icons; and finally, animation driven techniques cyclically rotate a series of ‘stacked’ icons through the foreground view.

Primarily, this apparatus deals with occlusion by way of point displacement. An invisible grid is defined over the space of the theme map. Each document is fixed to the closest cell location in the grid. Every second line is offset by a half-cell width in the horizontal direction to give a honeycomb-pattern appearance. Figure 6.9 on page 273 shows the result of the horizontal offset while Figure 6.10 on page 274 shows icon layout without the offset - the appearance is more grid-like. Furthermore, the icon spacing is more in line with the recommendation of a minimum of one-half icon spacing, in order to benefit visual search (Lindberg and Näsänen, 2003).

Ties, or two or more overlapping icons in the same cell, are dealt with by selecting the least relevant document and moving the tying document to the next available cell, along an outward spiralling path. Selecting the least relevant document as opposed to a random selection ensures the process is deterministic across rotations. For small thematic neighbourhoods this approach is adequate; however, for very large and dense visualisations topological distortion techniques may be more appropriate. In dense visualisations, the likelihood of free adjacent cells is diminished. Thus, a tying candidate could potentially end up far from its original location.

Furthermore, this apparatus does not guarantee that all documents are visible from any one perspective of the information space. Documents that do not highly associate with display dimensions, tend to cluster about the origin point; accordingly, such points are filtered from the display. This makes for a cleaner and less cluttered theme map visualisation and prevents a misguided but intuitive orient of attention to the centre of the screen, in the anticipation that relevant documents will be found there.

Further to a visual abstraction of a search result, icons provide a point of interaction at which users carry out actions specific to each icon. For instance, document icons listen for mouse events; a mouse-click event will open document full-text whilst mouse-hover events trigger pop-up windows that display document surrogate information. Further discussion of these actions will be left for Section 6.2.5 on page 295 although the design of pop-up windows is discussed in the next section.

Spatialisation Annotation and Document Pop-ups

The previous chapter highlighted the importance of landmarks in information space for navigation, exploration and identification. Text annotation is one method of provisioning such landmarks. This apparatus incorporates two annotation types: topic descriptors and pop-up windows. Earlier visualisation design has predominantly made use of topic descriptor annotations, and only pop-up windows on demand. This apparatus will display both topic descriptor annotations and as many document pop-ups as possible - without occlusion - in an effort to blend the benefits of ranked-list based search tools and visualisation paradigms together.

Topic descriptors, in the form of textual annotations are affixed to locations of information space during document spatialisation. In principle, topic descriptors should, in one to three words, provide a basic outline of a concept or theme, and accordingly, suggest what nearby documents are likely to share - semantically - in common. In contrast, pop-up windows show individual document metadata just like document surrogates in a ranked-list search result interface. Since topic descriptors and documents co-exist in the same space, document pop-ups act as secondary annotations. With multiple pop-up windows on display at any one time, a searcher may ascertain the meaning of regions of information space, by way of the themes evident in the document surrogate information.

Careful consideration of annotations and pop-up windows is rarely seen across the information visualisation literature and furthermore, in search user interface literature. In contrast, there is a long history of annotation and labelling research in cartography (Plaisant and Fekete, 1999) and diagram design. Of specific research findings in visualisation, with regard to labels, Pirolli, Card, and Wege (2000) find that strong information scent - depicted by annotations and labels - is a leading factor in the successful use of a hyperbolic tree visualisation. While the underlying connection to spatialisation is weak, the key message is that despite the push away from visualising large amounts of text, there remains a need for annotation for attributing meaning to regions of information space.

Regarding actual evaluation of pop-ups windows, Caro (1997) and Bétrancourt and Caro (1998) find that it is advantageous to offer supplementary text in pop-up windows - triggered by mouse hovering over activating words and phrases - over equivalent and bracketed text presented in-line, in text articles. In these investigations, supplementary text facilitates deeper understanding and clarification of concepts.

The results show that task performance on a text understanding task with the aid of bracketed supplementary text, is worse than an equivalent task that incorporates supplementary text in pop-up windows; however, this result only reaches experimental significance for task questions which draw on text at the end of the article.

A possible reason for this finding is that the pop-up method shortens the article, in comparison to the bracketed method. While the authors suggest an influence of task difficulty, more generally, by the end of the article, the user is perhaps increasingly fatigued. When fatigued, irrelevant information - not all supplementary information is relevant - takes more effort to ignore and skip over, in comparison to a text with all supplementary text hidden by inactivated pop-up windows. As a consequence, through having to read both the article's text and all supplementary text - regardless of relevance - the likelihood of arriving more quickly at the required information, diminishes.

These investigations show that pop-ups are an appropriate way to alleviate the influence of irrelevant text on reading since users have the option to view supplementary

text ad hoc and only when required. The above findings are consistent with the use of pop-ups on demand in information visualisation applications; removing text from a display allows the user to focus on information encoded visually.

However, the use of information visualisation in search user interfaces demands a balance between text and visual information. The study of Bétrancourt and Caro highlights the utility of pop-up windows to carry text; however, their findings motivate the use of pop-ups in a completely opposite fashion to this research. Instead of restricting a searcher to a single pop-up window on demand, the apparatus will permit the display of multiple pop-up windows at once, in order to blend the benefits of information visualisation with the fast scanning strategies searchers adopt for interaction with ranked-list result interfaces.

Very little other consideration is devoted to pop-ups in visualisation-based search tools. However, Rivadeneira and Bederson (2003) notes that truncated labels can make search difficult when interacting with the Grotzer interface. This observation reveals a trade-off: due to the limited size of labels and pop-up windows, text size must be made smaller in order to fit more content. Clearly, a consequence of maximising content at the expense of text size is that text legibility will surely degrade.

Consideration of label and pop-up construction is perhaps taken for granted, due to their ubiquity in many every-day software applications. Yet, regardless of their ubiquity, a number of factors do exist that influence interaction with interfaces via pop-up windows. Some of these factors are addressed subsequently, in order to devise labelling for the apparatus. First, a list of usability and aesthetic criteria are noted; following this, a classification of pop-up and label types is provided. Later, discussion focuses on label layout methodology and the problem of label occlusion; this will lead into a discussion on the pop-up layout approach. Finally, these considerations are brought together and a summary is provided on the actual implementation of labels and pop-ups in this apparatus.

Label Aesthetics

K. Hartmann et al. (2005) and Plaisant and Fekete (1999) suggest a list of usability and aesthetic criteria for labels. Usability criteria include appropriate length, clarity, content, and expressiveness, while aesthetic criteria include occlusion, distribution and consistency, readability and unambiguous referral. Additionally, Rosenholtz et al. (2005) suggest local and global clutter should be adequately managed.

The experiment apparatus design has attempted to optimise on aesthetics and functionality as best as possible. For semantic landmark labels, colour coding is present to establish unambiguously the theme vector the label most closely describes. Global and local clutter is implicitly optimal since theoretically, only 15 theme labels are possible for display at any one time.

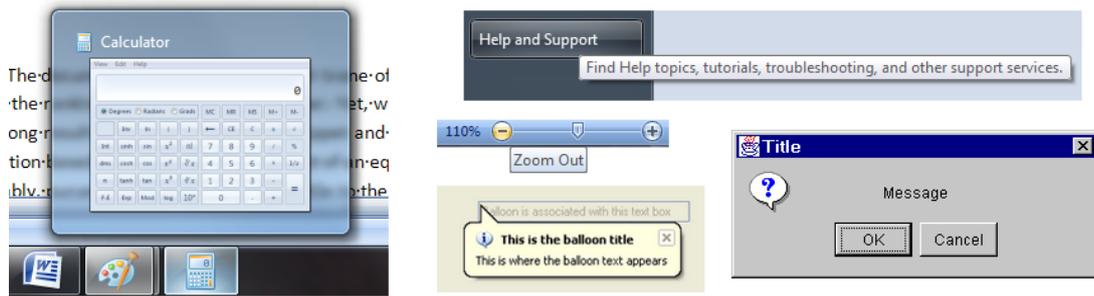


Fig. 6.12: Five techniques to display message text in a desktop computing application; A) toast; B) infotip; C) tooltip; D) help balloon; and E) modal dialogue pop-up.

Label and Pop-up Form

Pop-ups, tool tips, help balloons, and toast are all terms to describe small, concise messages that reveal underlying functionality in graphical user interfaces. While the main purpose is the same for each label or pop-up type - that is, to convey a message - the application, appearance, and location of the message is different. This apparatus incorporates both tool-tip like labels as well as small pop-up windows to display spatialisation annotations and document surrogates, respectively.

Despite the ubiquity of pop-up windows and labels in modern software applications, there is seemingly no known pop-up type taxonomy; accordingly, each of above pop-up types has been observed in the Microsoft Windows operating system. A graphical depiction of each of the pop-up types appears in Figure 6.12; many other pop-up examples can roughly identify as one of the types shown here.

Pop-ups (Figure 6.12 E) may refer to different user interface message types. A pop-up may be considered a context menu that appears when a user right-clicks on an area of the software interface; or an informative, confirmatory modal dialogue box demanding interaction before it disappears; or historically, a small advertisement-containing Internet browser window.

Tooltips (Figure 6.12 C) are small overlays that label normally unlabelled user interface controls, thus revealing a name and other relevant information e.g. a keyboard short cut. In contrast, infotips (Figure 6.12 B), a variation on tooltips, typically contain more text used to explain the functionality of user interface controls rather than to identify them. In desktop software applications, tooltips and infotips tend to reflect native tool tip support and so are mostly uniform in appearance.

Help balloons (Figure 6.12 D), unlike tool tips, tend to contain more textual content as well as icons and images. The other main visual difference is a tail or needle that points to the specific interface component to which the message refers i.e. promoting unambiguous referral. Like tooltips, help balloons in desktop software applications tend to engage native help balloon design and so are uniform in appearance. In contrast, help balloons in web-based interfaces are highly heterogeneous and take on a variety

of different appearances and styles; but, most often retain a rectangular or rounded-rectangular shape and characteristic needle or tail.

Toast messages (Figure 6.12 on the preceding page A) are aptly named due to their square shape and animated transition in and out of view of the desktop task bar giving the perception of toast rising out of a toaster. This type of pop-up is readily associated with the early versions of Microsoft Messenger chat software in which a new message notification transitioned up and out from the software's task bar icon before disappearing or transitioning back down after a time out period. Despite stylistic diversification of the shape and appearance, the characteristic animated transition remains a defining feature of toast messages. In Windows 7, a variation on the toast idea displays a micro-screen shot of the software interface when the user mouse-hovers over the opened application label on the task bar (e.g. Figure 6.12 on the facing page A)

Tooltips, infotips, help balloons and toaster messages all share some similarities in terms of their appearance and disappearance. Tooltips and infotips appear in response to a user's mouse hover action on an icon or a region of the interface. Help balloons and toaster messages tend to be system initiated in response to some system event - e.g. your hard drive is nearly full - or at the failure of some system-performed user input check - e.g. illegal form input - the completion of some background process or the arrival of a new message. Equivalently, tooltips and infotips exemplify a pull model, whilst help balloons and toaster messages typify a push model of information.

Finally, the system initiated modal and user initiated menu pop-ups (Figure 6.12 on the preceding page E) differ from tips, balloon and toaster messages in that pop-ups necessitate an additional interaction from the user in order for them to disappear. Modal pop-ups do serve an important purpose; particularly, when the system is set in an anomalous state and in need of input from the user to continue. However, timing of modal pop-up display is critical, particularly when modal pop-ups interrupt a primary task (B. Bailey and Konstan, 2006); if modal interruptions are presented at the boundary of primary task completion - i.e. once the user has completed their primary task - time of task, annoyance and error rate, is greatly less than modal dialogues presented throughout the execution of a primary task.

Pop-up Placement

The diagram and map labelling literatures suggest several approaches to affixing labels to diagrams without incurring significant overlap. Plaisant and Fekete (1999) offer a taxonomy of labelling paradigms, which in many ways bears resemblance to the clutter reduction approaches offered by Ellis and Dix (2007). Plaisant and Fekete divide their taxonomy into two parts: static layout and dynamic layout.

In the static case, a label appears with or without filtering of the display. However, excluding particular items from holding a label offers no guarantee to ensure labels

do not overlap with each other. Optimising label layout with placement algorithms that calculate the best label layouts without overlap is also possible; but with large numbers of labels, this process may be computationally excessive and slow and provides no guarantee that all important labels will be on display. Bekos et al., 2007 discuss boundary-labelling paradigms in which labels appear around the diagram or map like a border of labels. Their contribution focuses on algorithms that minimise the length and number of bends of referring leader lines connecting a label and a referent.

Dynamic labelling techniques invite interaction from the user. These include mouse-hovers over items to trigger pop-up windows adjacent to the triggering item or in a dedicated interface location. Overall, the most common approach observed in the list of surveyed systems in Chapter 2 was the single adjacent pop-up - i.e. the tooltip or help balloon style - in response to a mouse-hover event. However, the dedicated interface location was observed in the Kartoo interface (see Koshman, 2006) and in Touchgraph - see <http://www.touchgraph.com>. What is more, Grokker (see Koshman, 2006; Rivadeneira and Bederson, 2003) achieves a similar effect by controlling the scroll of a ranked-list by way of interaction with a visualisation; in effect, Grokker displays a document surrogate in a dedicated interface panel which is triggered by a mouse-hover interaction.

In a single pop-up window paradigm, occlusion is effectively dealt with, since only one pop-up is open at a time; however, this may be impractical. A variation of the adjacent pop-up involves triggering multiple pop-ups in sequence without closure. In this case, occlusion is still a problem but only the most recently triggered pop-up is presented in the foreground on top of the stack; however, this could make comparison with underlying pop-ups troublesome.

Alternative techniques include filtering, topological distortions, and zooming as is similarly proposed for clutter reduction by Ellis and Dix. These approaches involve removing information to make room for more important pop-ups and labels or changing the topology of the space in order to fit in more labels.

Plaisant and Fekete introduce a diagram labelling technique called Excentric Labelling. Excentric labelling is an extension of mouse-hover triggered pop-up labelling, but instead of showing just one label, labels for all points within a radius of activation around the mouse cursor are displayed. Where multiple labels are triggered, leader lines are drawn from each label to its referent object, thereby maintaining unambiguous referral. Such an approach is applicable for spatialisation interfaces, in that once the searcher has directed their eye gaze to a region of information space, insight into the meaning of that region may be gained by scanning multiple document surrogates. It follows then that speed of insight will be in some capacity, reliant on how efficiently the searcher can access document surrogate text for each of a number of tightly packed documents in a region in information space.

Excentric labelling necessitates that the searcher's mouse act as a cross hair for the user's focused attention. In contrast, the apparatus of this chapter will necessitate a pop-up display technique that can open as many document pop-ups as possible across the whole space and not necessarily at the demand of the searcher. Furthermore, a pop-up display technique should also react to the searcher's requests to see pop-ups for a specific document on demand.

Historically, non-overlapping, real-time, dynamic pop-up layout was considered a computationally expensive problem though Mote (2007) offers a suitable solution that meets each of the above pop-up display and placement criteria. The net result of applying this placement approach is evident in Figure 6.4 on page 266 through Figure 6.7 on page 269.

Mote (2007) presents an approach for annotating cartographic maps with a large set of small and uniformly sized labels, such that no two labels overlap and that where possible, only the most important labels are displayed out of a set of overlapping candidates. The approach is generalised here, to display pop-up windows in a similar fashion. A brief description of the approach is as follows:

- For each document icon, four uniformly-sized pop-up candidates are created; each candidate differs on where its leader tail is positioned i.e. top-right, bottom-right, top-left or bottom-left corner of the pop-up - and therefore, how it will be positioned relative to the document icon.
- The visualisation area is divided into a grid or *trellis* with each cell dimension equal to that of one quarter the size of each pop-up. Each grid cell contains a reference to each pop-up that fully or partially overlaps the cell.
- Then, processing takes place in three stages: conflict detection, expense calculation and finally candidate selection:
 1. The outcome of conflict detection is the creation of a set of overlapping pop-ups. By dividing the visualisation into a grid, the number of comparisons that need to be made for each pop-up is greatly reduced. There are only a few different ways a pop-up A can overlap with another pop-up B if their centre points are within a four-cell radius of each other. If the centre points of pop-up A and pop-up B are greater than four cells apart, they are guaranteed not to overlap. Consequently, a test for overlapping pop-ups i.e. those pop-ups in conflict, are implemented using conditional statements and atomic operations. The result of each test determines if two pop-ups overlap, in which case they are considered a conflict pair.
 2. In cost analysis, conflicting pop-ups are reconciled, based on a pre-determined priority; a pop-up's priority is based on the referent document's similarity

to a pseudo document vector containing keywords from the task statement. In the simplest case, when two or more pop-ups conflict, the pop-up with highest priority is selected for view. However, there are four pop-up configurations, so selection takes into consideration, up to four occluding sets for each document icon. The selected pop-up from a conflict set is the one which occludes the least important pop-ups based on a weighted sum for each occluded set. Conflicted pop-ups that are trumped by more important pop-ups are removed from future consideration.

3. Finally, for each document icon, the remaining pop-up candidates are compared with each other. The candidate with the highest priority based on cartographic preference is selected i.e. in order of preference: top-right, bottom-right, top-left, and bottom-right corner leader position on the document icon. Pop-up position has an inherent preferential score attached to it in the aforementioned order. It is preferable that the pop-up appear in the top-right position as we are accustomed to reading top-down and left-to-right.

The above approach is affective; a user's experience is near uninterrupted by layout computation in real time. However, many pop-ups on display at once introduces a new problem: occlusion of underlying document icons and labels. Moreover, a trade-off exists between the number of document pop-ups open and access to semantic information depicted in document layout and spatialisation annotation. To counteract this occlusion, use of semi-transparent pop-up backgrounds may permit fast scanning strategies over information in the foreground yet, also permit interaction with spatial relationships and annotations in the background.

Pop-up Transparency

The use of transparent pop-ups provides a balance between textual information in the foreground i.e. in pop-up windows, and visio-spatial information, depicted in the layout of document icons and spatialisation annotations, in the background. However, the level of transparency should not impact on either of foreground or background legibility.

Two examples are depicted in Figure 6.13 on page 287 and Figure 6.14 on page 287. With reference to the pop-up at right in Figure 6.13 on page 287, the circular document icons are visible through the pop-up window; in this case, the pop-up's background transparency level is set to 50%. In contrast, in Figure 6.14 on page 287, the background of the left hand pop-up is non-transparent and consequently, content behind the pop-up is obscured. Without no change in pop-up layout, under non-transparent conditions, the user may miss information that is available in obscured documents.

Use of transparency in software design is not uncommon. Moreover, there is comparatively limited literature on the advantages and configuration of transparent windows.

Nevertheless, exceptions do exist and do offer some guidance toward selection of a level of pop-up transparency for a document spatialisation application.

B. Harrison, Kurtenbach, and Vicente (1995) investigate the use of semi-transparent tool palettes, menus and dialogue boxes. They highlight a need to optimise on the level of performance and level of transparency. With increasing transparency of the foreground window it is harder for the user to focus attention on the foreground object. Conversely, as foreground transparency increases it is easier to focus attention on background objects. In a document spatialisation application, it is important to preserve text legibility in the foreground i.e. the document surrogate, while revealing background spatialisation annotations and document icons. Their investigations reveal that beyond 50% transparency, performance deteriorates significantly on tasks involving the identification of icons on a foreground tool palette for a range of icon types and background types. Response time increases and legibility errors increase, when foreground windows are highly transparent.

Ishak and Feiner (2004) discuss a paradigm to reveal obscured content in multiple overlapping software frames. However, instead of rendering an entire window frame transparent, their approach ‘free space transparency’ renders only parts of the interface that are deemed unimportant. Unimportant regions are those that contain empty white space, colours and textures or regions set explicitly by the designer or learned based on the eye-gaze patterns of the observer. Unimportant, transparent regions of the foreground window allow obstructed background content into view; however, important, non-transparent regions in the foreground interface remain non-transparent for maximum legibility.

Baudisch and Gutwin (2004) argue that full window transparency has not received widespread adoption, since transparency impacts on readability of content and since foreground and background colours wash out under transparent conditions. They propose multi-blending, which preserves the fidelity of visual features in the foreground transparent panels, that are deemed important for the foreground task. For example, colour chips in a foreground palette are displayed in full fidelity despite a transparent foreground window. This technique is notable for information visualisation applications, in that colour features in the background - e.g. hue and saturation as encoders of information - are more conducive to accurate visual manipulation when a user does not need to account for a mix of pure and washed out colour features.

Pop-up Implementation

Earlier sections discuss a number of factors pertaining to pop-up construction and configuration; such discussion is comparatively absent in earlier research. In this research, carefully crafted pop-ups are an integral strategy toward balancing the merits

of information visualisation and existing search behaviours. Accordingly, this section centralises the considerations made for pop-up windows featuring in this apparatus.

Pop-up windows in this apparatus most closely resemble balloon help pop-ups in 6.12 on page 280; each pop-up depicts a single document surrogate consisting of a document title and snippet text. The pop-up's leader or tail points to the referent document icon in an effort to promote unambiguous referral.

The pop-up layout algorithm closely resembles the approach of Mote (2007), although, earlier label layout research dealing with multiple label layouts instigated the need for a multi pop-up display. Any one particular pop-up may be opened - and is guaranteed to open - at the user's request; however, the interface should also attempt to open as many surrounding pop-up windows at the same time. This multiple pop-up approach stands in opposition to a singular pop-up on demand paradigm, but such a position is necessary to achieve a balance between visual and textual information.

However, increasingly numerous pop-up windows in the foreground quickly obstruct background information and tip the balance from overtly visual to primarily textual. In an effort to further balance visual and textual information, pop-up window transparency is intended to retain background information and context in addition to foreground information. However, an optimal transparency is such that a user can, with ease, switch between foreground and background information, as required. Whilst this apparatus will only manipulate transparency on two levels, it is done so to establish whether this effect on dependent variables, may be observed at all. If there is no effect, it may be the case that legibility of text in pop-up windows is sufficient for reading. Moreover, whilst accuracy and time are key performance measures, other interactive measures are equally revealing. If participants opt to turn the multi-pop-up facility off - thus reverting to a single pop-up at a time - under non-transparent conditions more readily than participants under transparent conditions, then this result would suggest that participants under the non-transparent condition are finding the search task too difficult to complete with a large quantity of obstructed information.

The pop-up background is opaque white #FFFFFF in the non-transparent treatment group or 50% transparent white in the transparent treatment group. All other colours are opaque: the title font is blue #0000FF, bold and 13pt while the snippet text is black #000000 plain font in 11pt. Each pop-up has a black #000000 rounded corner border and has a 'tail' in one of four corner locations. The position of the tail depends on the label placement algorithm. Non-transparent pop-ups are depicted in Figure 6.13 on the next page and transparent pop-ups are depicted in Figure 6.14 on the facing page.

In devising an appropriate level of transparency, consideration of an adequate contrast between foreground text over a noisy background is important. Luminance contrast is the perceived lightness or brightness difference between two colour stimuli (Zuffi

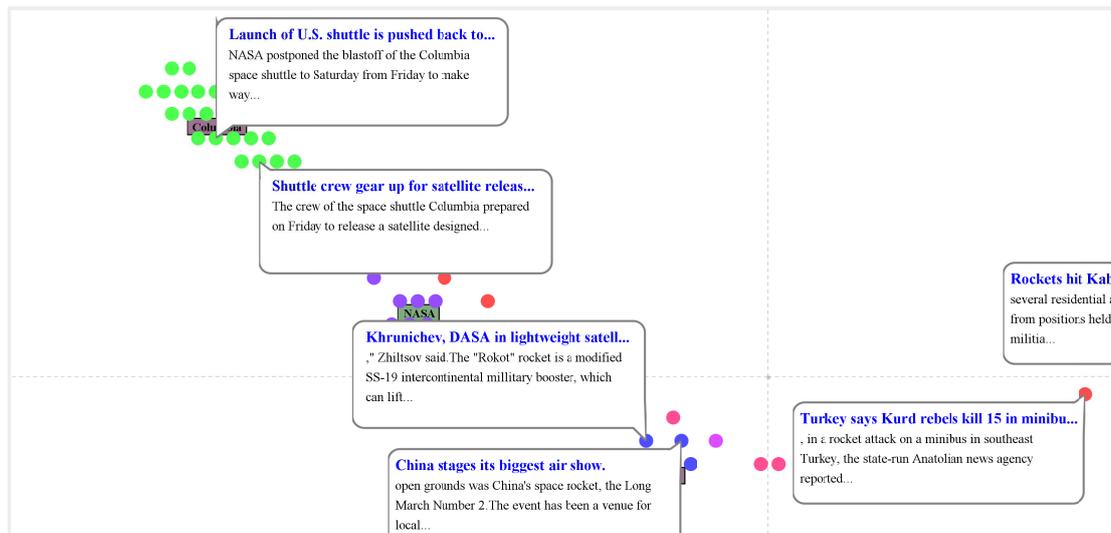


Fig. 6.13: Non-transparent pop-up windows.

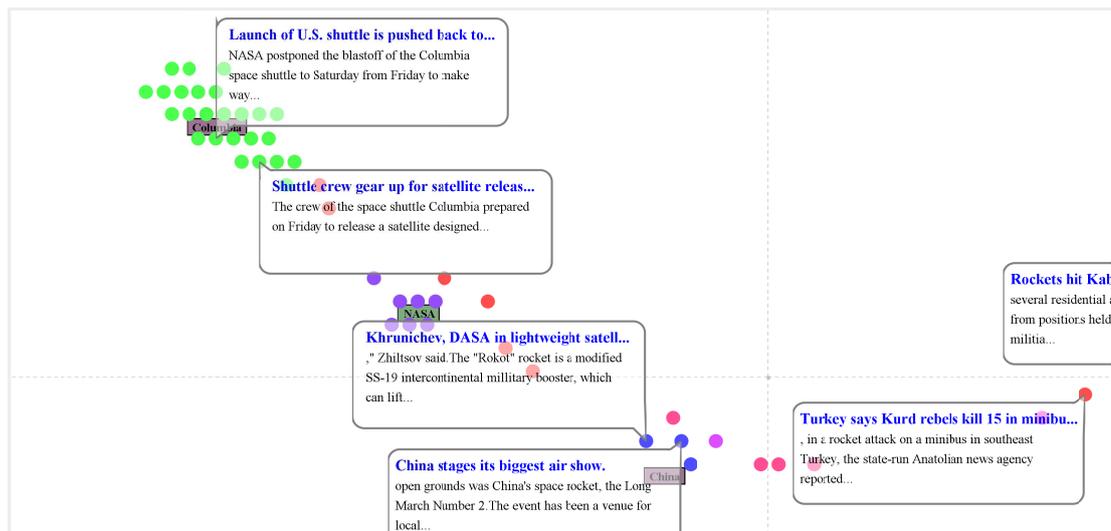


Fig. 6.14: Transparent pop-up windows.

et al., 2007). Typically, high luminance contrast enhances legibility and while minimum contrast ratios for text are often quoted at 3:1 provided the text is large; however, the world wide web standards authority suggests that upwards of 4.5:1 and up to 7:1 to ensure that persons with a visual impairment can read text with ease (*Web Content Accessibility Guidelines - WCAG 2.0 - Section 1.4.3 Contrast Minimum* 2008).

Luminance contrast factors into the consideration of a level of pop-up transparency, since background information below a semi-transparent window will bleed through and impact on legibility of foreground text. Background noise can include document icons, spatial annotations and dimension axes. Though, since the background is sparse, it is not always the case that background noise will interfere with text legibility and thus maximum contrast - i.e. black text on a white background - is achieved.

As semi-transparent windows in the foreground are superimposed over a background, background colours are blended with the semi-transparent foreground colour, in order to strengthen the illusion of semi-transparency; this process is commonly Alpha Blending, though alternative techniques such as (Ishak and Feiner, 2004; Baudisch and Gutwin, 2004) are proposed. Alpha blended colours can be calculated ahead of the experiment and contrast ratios between black text upon these blended colours can be evaluated using online tools.

Black foreground text upon each swatch in the colour set - in figure 6.11 on page 276 - meets minimum luminance contrast. However, since title text is encoded in blue, to match the same look and feel as contemporary search result pages this colour must also be tested. Unlike black text, luminance contrast fails in several cases posing a problem for the use of blue link text on a noisy background.

To deal with this blue on noisy background issue, there are potential options including making a strip of non-translucent text behind the title text, adding strong contrasting outlines, and adding contrasting shadows (Paley, 2003). However, while increasing pop-up transparency above 50% is an alternative way to deal with this, a leading trade-off is that contextual information is lost. By increasing the transparency further and below 50%, the contrast ratio is exacerbated further and blue text and even black will be harder to read. Since legibility of text is critical in this task, it may be necessary to sacrifice the familiarity participants have with the colour scheme of a search result, for a colour scheme more conducive to legibility on a background of contextually relevant information.

Annotation Implementation

To this point, spatialisation annotation has focused primarily on pop-up windows and has largely ignored the role of traditional spatial annotation labels. Most commonly, short, skinny, text labels are affixed to a spatialisation to cue a user to the semantic or thematic meaning of spatial regions. In addition to pop-up windows, this apparatus includes textual annotations that are positioned during spatialisation. Each annotation corresponds to a topic variable introduced by the correspondence analysis. In the expected use-case, annotations are utilised first to establish the meaning of different regions of a spatialisation, before attention is devoted to specific and relevant regions.

Each correspondence analysis variable or topic is assigned a position on each projection dimension; however, in this apparatus, weakly associated topics - below a noise level - are filtered from display, leaving only dominant topics to appear as annotations. Filtering is necessary, as it is useful to retain annotations that contribute greatly to the explanatory power of the projection dimension and which do not lead to an occluded noise of labels about the origin.

Annotations are colour coded according to the projection dimension they describe. Annotations for a horizontal projection dimension are coded in dark green #7FAA7E while annotations for a vertical projection dimension are coded in purple #9E7495. Colour coding of projection dimensions is made explicit on the layout configuration controls and will be discussed further in a subsequent section - see Section 6.2.3. In an effort to avoid document points occluding spatialisation annotations, annotations are offset so they appear in the vertical gaps between document icons. Figure 6.14 on page 287 illustrates this point; the annotation 'Columbia' appears between document icons and between the title and snippet text. This has an additional benefit of not significantly interfering with the pop-up's textual content.

To observe different document layout - corresponding to an alternative perspective of information space - and to observe other themes and topics in the corpus, a user must engage an interactive control. The next section will discuss two layout controls for this purpose.

Layout Control

When a spatialisation layout is re-configured, a user has changed their perspective of information space. This section will describe an interactive control that a user engages to enact changes in document layout.

Layout manipulation permits a searcher to see documents that are otherwise obscured or filtered from view; to observe relationships between documents and topics as evidenced by the movement of documents between projection perspectives; and furthermore, restricts search for documents to only those that have an association to a subset of the topic space that the searcher is interested in.

There are - at least - three considerations for layout control design. Firstly, since a layout control provides a medium at which a user expresses their intention to manipulate layout, a layout control should provide an intuitive abstraction of a set of projection dimensions. A control that obstructs the fact that there are multiple projection dimensions for configuration - e.g. the use of drop down boxes for each projection dimension - relegates the user to stepping through projection dimensions in sequence. Secondly, a control should facilitate streamlined selection of projection dimensions; this is important, as a user may utilise a rapid and or repetitive sequence of calculated configurations as a strategy for the observation of document positioning in more than two dimensions. Finally, a third consideration is that a layout control should provide clear and useful labelling of projection dimensions, such that a searcher may evaluate the relevance of those dimensions and thus, avoid or select as required. Importantly, labelling of projection dimensions should exceed that offered by a single word or phrase topic labels.

This apparatus examines two different layout controls. The first control - here on *theme cloud* - adapts the Wordle word cloud layout algorithm of Feinburg (2010). The second control - here on in *theme list*, is motivated by faceted search (Hearst, 2006).

Each control has an advantage over the other, specifically in the intuitiveness of representation, degree of selection efficiency, and or perceived effectiveness of dimension labelling. However, each control is functionally equivalent; each control: allows the re-configuration of the current document layout by rotating in a new projection dimension; each provides a set of annotations for each projection dimension; and each facilitates plotting of projection annotations in the vicinity of related documents.

Upon selection of a projection dimension, document icons move smoothly, from their existing coordinates to their new coordinates. If a document icon's final coordinates lie within a threshold radius about the origin point, document icons will fade smoothly out of view at their initial location. Likewise, document icons that are filtered from view in the initial projection, but appear in the updated projection, will fade smoothly into view at their final location.

A user may observe a descriptor that is relevant for their search, and may opt to plot descriptors on the spatialisation to aid semantic interpretation of a region in space. When selected and plotted, descriptors appear on the spatialisation nearby their projection dimension. Descriptor annotations have no specific coordinate in the spatialisation; instead, descriptors initially take the coordinate of their projection dimension topic, and are jittered until they do not overlap other annotations. While not a feature of this apparatus - do due time constraints - there is scope to calculate descriptor-document associations and thus, to calculate descriptor coordinates relative to documents and projection dimensions.

Each of the two layout control candidates and their expected advantages and disadvantages are discussed directly.

6.2.4 *Theme Cloud Layout Control*

The theme cloud control - depicted in Figure 6.15 on page 293 - incorporates two word cloud visualisations. Word cloud or tag cloud visualisations are popular in the visualisation community, and are one of the few and more recent visualisation techniques to have penetrated a mainstream audience.

Typically, word clouds present a set of user-defined descriptive word tag sets or Folksonomies that describe or reference a collection of documents (Sinclair and Cardew-Hall, 2008), while another prominent application is the presentation of results of word frequency analyses of literature (e.g. Feinburg, 2010). As a consequence of such popularity, existing evaluative research proposes a set of advantages and disadvantages for word cloud usage.

A word cloud is typically composed of horizontally-oriented word glyphs, and are placed within a space such that no two word glyphs overlap, despite a close proximity with each other. The emergent perception of the set of glyphs is that of a cloud, due to the perceived randomness of an invisible boundary line around the set of words, in combination with a dense inner area.

When presented on web pages, cloud glyphs are typically hyper-linked to a collection of documents or resources that a community has designated to the glyph's word or phrase. This facilitates query execution and simplistic query formulation and is consistent with a browsing mode of searching. In addition, Lohmann, Ziegler, and Tetzlaff (2009) lend further support for tag clouds as browsing tools, by showing that tag clouds are not conducive to optimal visual search for specific word tags.

When presenting the results of a literary analysis, cloud glyph font size typically encodes a word's frequency across the corpus, while visual features such as colour, style, or position, may be assigned to word class, semantic category, or any other variable of the analysis. However, encoding data in visual and positional attributes of word tags has performance implications. Lohmann, Ziegler, and Tetzlaff (2009) confirm earlier findings that important tags should be made larger than unimportant tags because larger tags are located more rapidly; in addition they suggest that string width, cloud position, and semantic neighbourhood position all impact on search performance. Moreover, they confirm that a user utilises scanning patterns rather than reading patterns and that more attention is devoted to tags in the centre of the tag cloud.

Lohmann, Ziegler, and Tetzlaff (2009) suggest that there is no best way to represent tags that have importance ratings, particularly if the interface is likely to support multiple browsing tasks. For three different tasks: finding specific tags, finding popular tags and finding tags that belong to a topic, three different layouts are possible: sequential layouts using alphabetical sorting, circular layouts with decreasing popularity working outwards, and thematically clustered tags i.e. assigning tags to a semantically-defined spatial region of the word cloud.

Hoeber and Liu (2010) investigate tag clouds, term histograms and list widgets in combination with a ranked-list search result interface. Words appearing in tag clouds, term histograms and tag lists are extracted from documents that the system and searcher flags as relevant. Tag clouds, histograms and lists are interactive; clicking on word tags re-ranks the ranked-list of search results. Tag cloud words are arranged alphabetically and font size encodes term relevance; for term histograms, tags are arranged vertically in a list and a histogram bar adjacent to each term encodes term relevance; for term lists, terms are arranged in alphabetical order but there is no indication of term relevance. The results indicate that although there is no overall improvement to average precision on a document search task, the results are variable depending on the participant. Average time on task is generally better utilising the term histogram, followed by the tag cloud and then the term list. In addition, subjective responses

favour the term histogram approach over the tag cloud.

Kuo et al. (2007) evaluate the utility of tag clouds to summarise search results taken from a biomedical corpus; they compare a search engine with a tag cloud against the same search engine without a tag cloud. Their tag cloud has a similar appearance to that in Hoeber and Liu, however font size corresponds to each word's frequency across the corpus of abstracts, while font colour corresponds to the average size of publication data for abstracts containing corresponding word tags. A mouse hover event triggers a pop-up containing related terms for each tag; a searcher can access relevant abstracts by clicking on a cloud tag. Their usability experiment indicates participants are more effective with the tag cloud supported search engine. when answering simple fact-finding tasks, but less effective when attempting to answer relational questions. This reiterates an influence of task specificity on task performance seen in earlier investigations (Lohmann, Ziegler, and Tetzlaff, 2009). Moreover, subjective results indicate that while participants rate the cloud-supported search engine as less helpful than the control, they indicate higher satisfaction. Finally, regardless of condition, participants report a similar level of understanding.

One of two layout controls in this apparatus incorporates a word cloud visualisation approach. Since a layout control depends on the selection of one vertical and one horizontal projection dimension, which are themselves labelled and have dimension-specific annotations, a word cloud visualisation in this context has excellent encoding potential. The Wordle algorithm of Feinburg (2010) has inspired the theme cloud control in favour of alternative algorithms and frameworks (e.g. Seifert et al., 2008), in part due to the possibility of incorporating vertically oriented word tags - something that is seldom observed in web-based tag clouds.

Figure 6.15 on the next page depicts a theme cloud control for the *Recyclable Materials* task set. As is seen in Figure 6.15 on the facing page, the theme cloud control is composed of two word clouds. The first word cloud depicts rotatable projection dimensions. Each word tag represents a single projection dimension and is annotated by the most highly-weighted topic descriptor, generated during document pre-processing - see Chapter 5. Clicking on a word tag enacts a change in projection dimensions and thus, a change in document layout. The second word cloud depicts horizontally and vertically-oriented dimension-specific descriptor annotations. Fifteen of the most highly-weighted topic descriptors appear in the descriptor word cloud; each descriptor is oriented in accordance with the projection dimension it describes. If a descriptor annotates the vertical dimension it is oriented vertically and likewise, if a descriptor annotates the horizontal dimension it is oriented horizontally.

The word cloud layout process generates a glyph for each projection dimension and descriptor. Glyphs in the descriptor cloud are rotated $90deg$ counter-clockwise - if annotating a vertical dimension, or remain horizontally oriented otherwise; glyphs in the projection dimension cloud are not rotated. Then, starting from the centre of

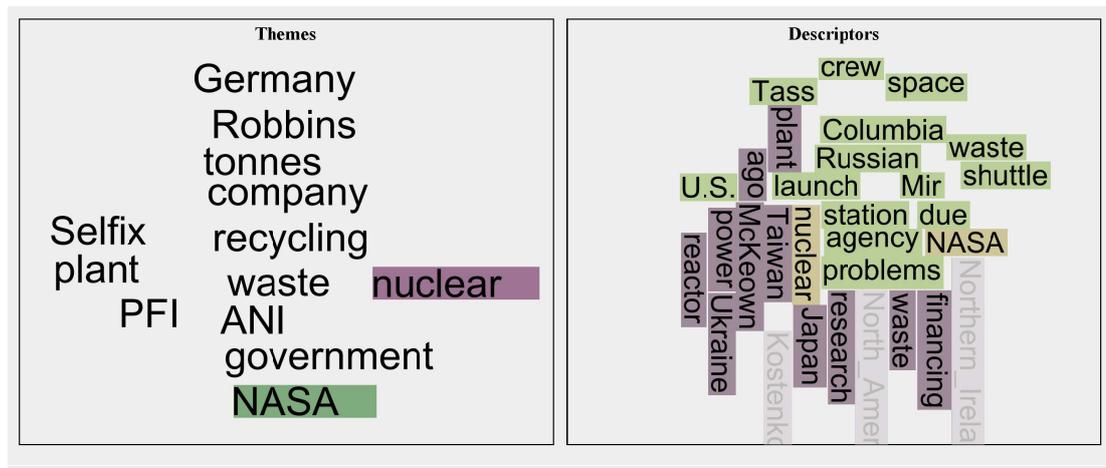


Fig. 6.15: The theme cloud control consisting of theme tags which facilitate layout configuration, and descriptor tags which facilitate labelling of the theme map.

the word cloud visualisation area and working outward in a spiral pattern, the layout algorithm attempts to place each glyph such that it does not overlap any previously placed glyphs and such that it is contained within the bounds of the visualisation.

A change in document layout is enacted by clicking on a word tag in the projection dimension control; however, since each projection dimension tag appears just once - as opposed to a set horizontally oriented and a set vertically oriented - a user must specify either the vertical or horizontal screen dimension to assign the selected projection dimension. To assign a projection dimension to the vertical screen axis, a mouse click must occur on the left hand side of the word tag, conversely, a mouse click on the right hand side of the word tag assigns the dimension to the horizontal screen axis. The mouse cursor icon depicts a horizontally oriented double-ended arrow when the mouse cursor is positioned on the left hand side of the word tag and a vertically oriented double-ended arrow when positioned on the right hand side of the word tag.

If choosing to plot specific descriptors, a mouse click over any part of the descriptor word tag will plot the descriptor in the spatialisation.

Following each dimension configuration, descriptor words appearing in the descriptor cloud seen at right of Figure 6.15 change in response to the selection of a new projection dimension. As descriptor word orientation corresponds to the orientation of its referent projection dimension, it is an easy visual search task, to identify tags corresponding to a particular projection dimension. Moreover, redundant colour coding of each projection dimension reinforces each tag's screen dimension configuration.

Theme List Control

The theme list control - depicted in Figure 6.16 on page 296 incorporates two sets of two word lists. This control has its origins in both faceted search interfaces and in the observed shortcomings of an explorer tree - similar to a file system tree view - when used as a projection dimension control.

An explorer tree control can incorporate all of the functional aspects of a theme cloud control - i.e. representing a projection dimension as a direct child of the root and by representing each descriptor as direct leaf node of a projection dimension node. However, screen real estate limitations impact on the user's exploration of many projection dimensions in succession. Such exploration is mouse intensive, involving much repeated scrolling of the view, opening of nodes, closing of nodes, and selection of leaf descriptors.

Faceted search systems incorporate interactive interface components and layouts to browse document sets, which are organised into multi-membership, broad and shallow hierarchical sets (Hearst, 2006). By utilising facets, a searcher can progressively filter a large set of search results based on a selection of categorical criteria (G. Smith et al., 2006). A prominent example of faceted browsing is that in Apple's iTunes media player; selecting genre in the first list refines the second list containing artists of that genre, selecting an artist populates the third list with albums by that artist, and selecting an album populates the bottom list - initially, the library's entire collection - with songs from the selected album. However, faceted browsing is by no means limited to the entertainment domain. For instance, Hearst (2006) examine on a number of text-based faceted systems that feature in e-commerce sites; G. Smith et al. (2006) propose an aesthetically pleasing bubble-like visualisation as a front end to a facet browsing application; Wilson, André, and Schraefel (2008) propose an evolution on the multiple interactive list widget that visualises alternative pathways to to the same search result; and finally, Polowinski (2009) surveys a range of facet controls and interfaces for both specific and generic domains.

Projection dimensions provide a shallow hierarchy; one may view filtering by two dimensions, somewhat analogous to faceted search. In the present context, a manipulation of two facets - i.e. dimensions - limits the scope of refinement. However, this approach would quite easily extend to three dimensions, with the addition of two extra lists - one for the projection dimensions and one for descriptors since, as in the case of the theme cloud control, each theme list has a corresponding descriptor list that annotates the currently selected projection dimension. The main difference though is that a projection dimension list and descriptor list is needed for both horizontal and vertical screen axes, thus necessitating four lists in total.

Figure 6.16 on page 296 shows the same task set example as is presented in Figure 6.15 on the preceding page, though now using the theme list control. There are two

lists per projection dimension axis - one for the projection dimension selection - labelled 'Vertical Themes' and 'Horizontal Themes' in Figure 6.16 on page 296 - and one for the descriptor selection - labelled 'Vertical Descriptors' and 'Horizontal Descriptors' in Figure 6.16 on page 296 - giving four lists in the projection dimension control panel.

While functionally equivalent to the theme cloud control, the process of dimension configuration, with the aid of theme list, is slightly different. To manipulate the horizontal projection axis, a user clicks on entries in the horizontal projection dimension list; similarly, to manipulate the vertical axis, the user clicks on entries in the vertical projection dimension list. Selection of projection dimensions in either the horizontal or vertical list will enact a change in document layout in the spatialisation and furthermore, will update the list of descriptors annotating the selected projection dimension. In addition to rotating in dimensions, plotting descriptors on the spatialisation is achieved by clicking on descriptors in the descriptor lists.

Functional equivalence aside, there are clear structural and aesthetic differences between a theme list layout control and the theme cloud. Although colour coding is incorporated into the theme list, only descriptors that the user has selected for plotting are colour coded; this contrasts with colour coding in theme cloud in which all descriptors are coloured redundantly in combination with word tag orientation information. A concomitant level of encoding redundancy is not evident in the theme list; while descriptors are colour coded and appear in an exclusively horizontal or vertical descriptors list, the encoding is not as intuitive as the orientation information. However, colour coding of selected projection dimensions occurs in the same manner, in that the active projection dimensions are colour coded uniquely.

As a final point, although not explicitly examined in both the theme list and theme cloud, but worthy of mention here is the potential for further encoding. Research listed in the previous section made mention of the usage of word tag size and font style manipulations for encoding of importance metrics. While there are tighter restrictions on the capacity to manipulate font style in the theme list, it would be possible. Furthermore, use of list order - an intuitive indicator of importance generally (Tversky, 2011) - is a likely reciprocal of font size in the theme list.

This concludes most of the relevant discussion for the experiment apparatus. Thus far, the discussion has considered document full-text view, a ranked-list interface and a spatialisation interface. The discussion on spatialisation interfaces entailed an overview of document icons, spatialisation annotations, document pop-up windows, and interactive controls to rotate the user's perspective of the information space. The next section will outline how a user may incorporate each element into a real search task.

Vertical Themes	Vertical Descriptors	Horizontal Themes	Horizontal Descriptors
Japan	Nepal	Japan	Japan
Nepal	China	Nepal	Taiwan
Taiwan	India	Taiwan	China
China	Jiang	China	islands
Singapore	Tibet	Singapore	Tokyo
chief	Chinese	chief	Hong_Kong
Hong_Kong	visit	Hong_Kong	Diaoyus
courts	Beijing	courts	East_China_Sea
Britain	Kathmandu	Britain	Chinese
PLA	agreement	PLA	claim
Taiwan	Jiang_Zemin	Taiwan	sovereignty
South_Korea	Birendra	South_Korea	Taipei
police	officials	police	group
Hong_Kong	Pakistan	Hong_Kong	activists
	trade		dispute

Fig. 6.16: The theme list control consisting of horizontal theme tags which facilitate layout configuration of the horizontal coordinate, and horizontal descriptor tags which facilitate labelling of the theme map relative to the horizontal coordinate; adjacent vertical theme tags and descriptors control configuration of the vertical coordinate.

6.2.5 Search Facilities of the Apparatus

Till now, discussion has offered a qualified description of the structural, functional and interactive aspects of the apparatus. However, there has been little mention of how each interface component contributes to a user's search task.

Foremost, this apparatus does not offer a search query input box as in the case of traditional search engines. Whilst unconventional to do so, a lack of query input box invites thinking on alternative ways to execute search queries, while dealing with the fact that this apparatus, in its current form, cannot respond in real-time to ad hoc search requests.

Logically, this apparatus supports a searcher midway through their information seeking process - after having submitted their initial query. Accordingly, this apparatus provisions a searcher with the ability to manipulate and filter a set of search results. For instance, when interacting with the ranked-list interface a searcher can manipulate the order of the list using an in-text highlighting feature in the document full-text view. In contrast, when interacting with the spatialisation interface, a searcher can manipulate the display of documents by way of a theme list or theme cloud layout control.

This apparatus purposely excludes a traditional search box for several reasons. First, per Käki and Aula (2008), the use of pre-formulated queries seeks to introduce some level of control over the variability of query terms submitted by participants. Second, a lack of query input box alleviates the need for ad hoc query processing algorithms and ultimately, opens up additional development resources for implementation of the search user interface. Third, while Gerken et al. (2009) advise that an interface should support multiple query formulation methods, when exploring new techniques

Tab. 6.1: Interactive capabilities of interface.

View	Interaction	Target	Action
Theme Map	Mouse Click	Document Icon	Open document full-text
	Mouse Hover	Document Icon	Open document snippet balloon
Theme Cloud	Mouse Click	Theme Tag	Rotate Theme Map layout
	Mouse Click	Descriptor Tag	Plot descriptor in Theme Map
Theme List	Mouse Click	Theme List Entry	Rotate Theme Map Layout
	Mouse Click	Descriptor List Entry	Plot descriptor in Theme Map
Ranked List	Mouse Click	List Entry	Open document full-text
Full-text View	Button Press	Font Size Selector	Modify text size
	Text highlight	Text	Select text for list resort operation (ranked-list only)

that are radically different to current practice, it may be beneficial to first evaluate new techniques in isolation, and not within a context of an existing search ecosystem. Resistance to change may mean that an investigator observes more about the user's predominant search behaviours and less about their performance on the new technique. Fourth and finally, taking away that which we most heavily rely upon in search, triggers an experiment in innovation. The use of in-text highlighting, as a means to submit search queries - as seen in (McCormac et al., 2012) - is increasingly prevalent in search-supported applications (Baird and Zollinger, 2007, e.g.). In contrast to the case of opening a search form and typing in a text query, a searcher need only select a snippet of text, thereby, leaving the application to forward on a request to a search engine - and subsequently, to present the search engine's response. Such an approach is advantageous as the query's context may be inferred from the document that initially triggers the query. Entire sentences, paragraphs or phrases, not necessarily highlighted in the query, might also be submitted by the application without direct input by the user.

An overview of each interface's interactive controls and their contribution to search task completion is provided below in Table 6.2 on the next page. Furthermore, interaction capabilities that a user may leverage to extract information from the interface are provided below in Table 6.2 on the following page. The interactions listed in Table 6.2 on the next page underpin the measured dependent variables in the experiment; in addition to answer set quality, these interactions will factor into the composition of search behaviour profiles of participants assigned to different experimental groups.

A visual scan strategy in Table 6.2 on the following page is a fundamental way to locate interesting documents, regardless of the interface in use; a visual scan supports actual review of keyword-in-context content. However, when solving search tasks with

Tab. 6.2: Search facilities and strategies afforded by the experiment apparatus.

Condition	View	Action	Search Strategy
Ranked List	List	Visual Scan	Use keyword search in snippets to locate interesting documents
Ranked List	Document	Mouse Drag	Select text to re-sort list based on similarity to selected text
Theme Map	Theme Selector	Rotate Theme	Configure Theme Map to contain documents related to selected themes
Theme Map	Descriptor Selector	Select Descriptor	Plot descriptors on Theme Map; observe documents in vicinity
Theme Map	List	Select item	Visit each document sequentially; read snippet balloon
Theme Map	Theme Map	Mouse Hover	Visit each document icon sequentially; read snippet balloon
Theme Map	Theme Map	Visual Scan	Observe patterns, read snippets, choose areas of local density and work outwards

a ranked-list interface, participants are restricted to just sequential visual scans and list re-ordering actions. In contrast, the spatialisation interfaces offer a richer set of search facilities, which the participant may choose to switch between ad hoc. For instance, a participant may reconfigure document layout if the projection dimension's descriptor label indicates relevance to the search task. Following layout configuration, they may observe interesting descriptor tags in the layout control and choose to plot those as annotations in the spatialisation. Subsequently, they may opt to scan already open document pop-up snippets near annotations, before mouse hovering over documents with no current pop-up on display. At any stage, they may wish to open documents for additional confirmation of relevance; however, they cannot use the text to submit result set manipulations as is possible in the ranked-list interface. Finally, participants can opt to configure projection dimensions that contain a large number of documents and then sequentially open each document by the list adjacent to the spatialisation. However, as the list in the spatialisation condition carries very little information - purposely, to discourage participants from adopting this strategy - it may indicate that the user simply does not find the spatialisation particularly helpful.

6.2.6 Summary of Apparatus Design

In summary, this apparatus, at any one time, can display a set of search results in one of two ways: using a ranked-list Figure 6.1 on page 263 or using a spatialisation visualisation Figure 6.4 on page 266. The ranked-list attempts to replicate the format

presented on most search engines, while a spatialisation interface necessitates a spatial arrangement of the search result set in order to present search results.

Also at any one time, a searcher can access a document's full-text view in one of two ways: in a modal, windowed view Figure 6.6 on page 268 or in an integrated, framed view Figure 6.5 on page 267. A modal, windowed view also replicates the type of full-text view presented by most search engines, while the integrated, framed full-text view does not detach the searcher from the search results after opening a document; the latter may have implications for coordinating interaction in the document full-text view with search result display.

Structure and aesthetics of semantic annotations and document pop-up windows were described in relation to the spatialisation and the role that they play in a user's understanding of spatial region was discussed. Furthermore, it was argued that we need more than one document pop-up visible at once, because gaining an understanding of a spatialisation by way of a one pop-up on demand paradigm, is slow, serial, tedious, and inconsistent with current ranked-list search behaviour. However, incorporating too many pop-up windows in the foreground more than likely obstructs information in the background; pop-up background transparency was proposed as a way to overcome this situation - see Figure 6.13 on page 287 and Figure 6.14 on page 287.

At any one time, the spatialisation interface depicts a two dimensional perspective of a hyper-dimensional information space. A user may change their perspective of the information space by selecting projection dimensions that are relevant for their information need. Doing so modifies the layout of document icons and reveals similarities and dissimilarities between documents.

Two interactive controls, the theme cloud Figure 6.15 on page 293 and theme list Figure 6.16 on page 296, were proposed to facilitate projection dimension selection. Both controls are functionally equivalent, permitting the assignment of a projection dimension to either the vertical or horizontal axis and furthermore, depicting a set of descriptors for each configured projection dimension.

Document full-text view, pop-up background transparency and projection dimension control are the subject of examination in the experiment reported directly.

6.3 *Exploratory Hypotheses*

Having outlined the experimental apparatus and competing design alternatives, this section will present a set of exploratory hypotheses. Three hypotheses predict outcomes for task performance under different levels of pop-up window transparency, full-text view integration and theme map layout control. Tabulations of the predicted outcomes will appear at the end of this section in Table 6.3 on page 302 and Table 6.4 on page 302.

6.3.1 Hypothesis One - Pop-up Transparency

Multiple pop-up windows cover a large proportion of the spatialisation area, and if non-transparent, are likely to wholly obstruct document icons and spatialisation annotations.

Semi-transparent pop-up windows allow otherwise obstructed content into view, thereby saving additional time and effort that a participant would otherwise devote to revealing obstructed content. If efforts are not taken to reveal obstructed content, the participant is more likely to miss relevant information.

Previous research (B. Harrison, Kurtenbach, and Vicente, 1995) suggests that foreground legibility degrades sharply when foreground window transparency is set beyond 50% alpha. While undertaking experiment tasks, participants will need to comprehend textual information in the foreground, while maintaining background context. If a level of 50% transparency is too transparent, such that foreground legibility is affected, task performance is likely to degrade.

In contrast, task performance is expected to be poor in the non-transparent condition because a lack of background context - hidden by non-transparent pop-up windows - is likely to complicate exploration activity. One indicator that non-transparency is undesirable will be a tendency of participants to opt for a single pop-up display. Such a strategy is offered to participants by way of an interface toggle button. When the button is toggled, only one pop-up window will appear and when this button is toggled again, multiple pop-up windows return to the spatialisation.

Participants are expected to take longer to complete tasks, suffer degraded answer set quality and revert more readily - and for longer periods - to single pop-up display in the non-transparent pop-up condition. When pop-up windows are transparent, participants are expected to complete tasks faster, with a better outcome, and will not readily revert - if only briefly - to a single pop-up display.

6.3.2 Hypothesis Two - Document Full-text Integration

A document surrogate provides a keyhole view of a document's content; to see a document's full-text, the user must click on the document's icon or ranked-list entry. When using Internet based search engines, a searcher clicks on a search result's hyper-linked title and the full-text appears in a new browser window, a new browser tab or takes the place of the search result page. In any case and broadly, this draws the searcher's attention away from the search result page and directly to the full-text. As a consequence, a searcher will have to re-orient themselves with the set of search results, on return from a document full-text view and this will impact on search efficiency.

Often a search does not terminate at the first opened result; searchers use keywords, concepts and snippets of information obtained from each full-text view to reformulate an

initial query. However, searchers should not have to create a completely new tangential set of results every time a full-text view reveals an interesting keyword. Instead, an existing set of search results could be iteratively refined and updated to temporarily highlight the influence of those keywords in the current result set. Moreover, a searcher should be permitted to observe each change to the information space as it happens, rather than having to figure out how the information space has changed on their return from a document full-text view.

Overall search performance is expected to deteriorate in the modal full-text view condition. A dedicated document full-text view adjacent to the results allows faster skimming of document content, improving the likelihood of locating relevant documents, and saves the need to reorientate to the results - as participants maintain the results in the periphery of view. Participants are expected to locate relevant documents faster as they are able to continue their search within the spatialisation, without needing a disorientating switch between spatialisation and document full-text view.

Furthermore, within the ranked-list interface only, the number of text-highlighting queries is anticipated to be greater in the integrated document full-text view condition because the searcher can watch the ranked-list update in real time making for a more intuitive interaction.

6.3.3 Hypothesis Three - Theme Map Layout Control

A searcher manipulates their perspective of the information space and thus, document icon layout, in an effort to observe interactions between sets of documents based on document themes. This manipulation is enacted by a projection dimension control; however, it is not clear how best to implement such a control. Two layout controls are proposed: the theme cloud control and the theme list control.

Search performance is expected to be better in the theme cloud condition since the interaction costs for extracting information from the theme list are considered greater. That is, a participant is expected to take longer to interpret information in the theme list than in the theme cloud condition because the information is spread over additional windows in the theme list and the orientation information is made more explicit in the theme cloud. The orientation information in the theme cloud is expected to assist the user with orienting themselves with the projection dimensions of the spatialisation, thereby, making for a more natural interaction with the layout control. With a more natural interaction and a perceived objective benefit, the theme cloud control is expected to receive a higher subjective rating score.

6.3.4 *Exploratory Hypotheses - Summary*

The analysis will take place in two stages as outlined in the experimental design section - see Section 6.4.4 on page 314. Table 6.3 on the following page and Table 6.4 on the next page present a set of predictions made for the various time, outcome quality and behavioural measures for each factor and analysis. Later, reference will be directed back to these tables when discussing the experiment's results. For all hypotheses, unless specifically stated, an overall measure of search performance consists of a trade-off between the time on task, the number of interactions made and the quality of the answer set.

Tab. 6.3: Dependent variable outcome predictions for the manipulated factor and levels in analysis one. ‘●’ denotes no prediction is made.

Design	Factor	Level	Time	Outcome Quality	Documents	List Re-Sort
Between	Full-Text Integration	Integrated	Fastest	Best	●	Most
		Modal	Slowest	Worst	●	Least

Tab. 6.4: Dependent variable outcome predictions for manipulated factors and levels in analysis two. ‘●’ denotes no prediction is made.

Design	Factor	Level	Time	Outcome Quality	Documents	Pop-Up Time	Layout Changes
Between	Full-text Integration	Integrated	Shorted	Best	●	●	●
		Modal	Longest	Worst	●	●	●
	Pop-up Transparency	Transparent	Shortest	Best	●	Longest	●
		Non-Transparent	Longest	Worst	●	Shortest	●
Within	Rotation Control	Theme Cloud	Shortest	Best	●	●	Most
		Theme List	Longest	Worst	●	●	Least

6.4 Method

6.4.1 Participants

In total 52 participants completed the experiment; of this number one participant was removed from the analysis for statistical reasons. The remaining 51 participants, (26 male, 25 female) had an average age of 23.1 years (18 min, 37 max) and most cited English as their primary language (39 primary, 12 secondary). Participants responded to notice board and handout-advertising recruitment methods that included a mention of the \$40 reimbursement, estimated 2-hour time commitment and the requirement of a strong competency in English.

Participants were predominantly university students from diverse academic backgrounds; however, there were two participants external to the university. Of the internal students, 12% were from Science and Engineering, 62% percent from Health Sciences, 8% Education, Law or Humanities and 2% from Social Science. Health Science students were not specifically targeted - the unbalanced proportion of participants from this discipline was a result of word of mouth within a large clique of students.

Of the internal university students, 25% were in their first year, 39%, 25% and 6% were in second, third, fourth year respectively, and 4% of participants indicated they were in their fifth or later. Finally, 47% of participants reported their highest attained education was a high school certificate, 35% reported undergraduate degree, while approximately equal proportions of participants reported TAFE/Technical College, Honours, Masters, or PhD.

The experimental procedure was reviewed by a research ethics committee and approved for use. In line with the ethical guidelines and requirements under which this experiment operated, the research assistant notified participants that they were free to leave at any time they wished to do so and that their informed consent was assumed by clicking the agree button on the electronic consent form on start up of the experiment apparatus.

6.4.2 Materials

Task Sets

Three experiment task sets and one training set were generated for the experiment using the pre-processing methodology described in Chapter 5. Each task set consisted of a set of search results, a spatial model of the results and a task statement. The full-text of each document was available to the searcher along with a keyword-in-context or snippet text generated by the Lucene search engine - see <http://lucene.apache.org>. Each task set topic and the number of news articles within are tabulated in Table 6.5 on the following page.

Tab. 6.5: Topics and queries that form the basis for task sets in experiment.

Task Set	N	Topic
RM	109	Recyclable Materials
HK	134	UK Hand Over of Hong Kong to China
NP	156	News Paper Circulation Decline

Task sets were produced offline as the combined time for indexing, searching, pre-processing and formatting was prohibitively long, thus precluding any possibility of real-time processing of participant queries. While there are many possible indexing and pre-processing optimisations that could be made in order to serve real-time requests, it was not a priority for the current research.

Relevance Judgements

The Lucene search engine API offers a ranking of search hits based on a combination of term frequency, term boosting and normalisations; but objective, mathematically driven relevance measures offer no guarantee of relevance from the perspective of the searcher. Nevertheless, to evaluate a participant's search performance, a reliable relevance score for each document - originating from a human judge - is necessary.

Typically, definitive relevance ratings for a pool of documents are made by two or more corpus experts, who rate each document independently - according to an agreed standard - and follow up on conflicting judgements with an in-group re-evaluation. An expert's judgement is arguably more reliable than that of the experimental subjects, as they have greater familiarity with the corpus and spend a longer period of time in deciding on a determination of relevance. However, in this experiment, only one corpus expert - the author - was available to make expert relevance judgements. Therefore, all experiment participants were asked to commit - separate to their main experiment task - relevance ratings as part of their experimental contribution. Whilst this method is unconventional, it is acknowledged that cross-validation with additional and equally expert judges would make for a superior relevance standard, though this was not possible at the time of design and analysis, due to resource constraints.

Accordingly, a combination of one expert judge's ratings and a pool of participant relevance ratings, formed the definitive set of document relevance ratings. Each document's rating was one of: irrelevant, weakly relevant, moderately relevant or strongly relevant.

This four-partite scale was adopted to strike a balance between simplicity and utility. The critique of Borlund (2005) suggests that a binary notion of relevance is broadly

intolerable, yet as to how and what information to capture (Fox et al., 2005) and on what scale (Tang, Shaw, and Vevea, 1999) has been the subject of research. For the present experiment, Tang, Shaw, and Vevea would perhaps recommend an extension of the four-level scale to a seven-level scale to ensure a judge is more confident in their ratings in comparison to the case where the scale is smaller. For relevance judgements however, if the participant judges a document as not relevant, assumed by opening the document but not tagging it, then it is of little use to know how irrelevant, either partially or fully irrelevant, the document was.

In determining a document's relevance, the expert judge adopted the same recommended procedure as was provided to participants during training:

The idea of relevance is up to you; you should think about how you write essays. If the news article contains many facts about the topic and you would definitely reference it in your essay, then mark it as strongly relevant. If the news article contains a fact that you might reference in your essay then mark it as moderately relevant. If the news article contains only background information, then mark the article as weakly relevant. If the news article contains nothing of use to your imaginary essay, then mark it as irrelevant.

To capture participant judgements, at the end of the experiment, participants were requested to make relevance judgements on a set of fifteen documents for one of the previously completed task sets. The participant's task set was randomly selected and presented on an interface composed of an unranked list and an answer box; and document full-text was presented in an integrated full-text view - a screen shot is depicted in Figure 6.17 on the next page. Participants opened, read each document, and gauged its relevance to the task statement - depicted at the top of the interface. If the document was relevant they undertook the same scoring procedure as was used in the experiment tasks - see Section 6.4.3 on page 312. Participants must have opened and read each document before completing the relevance judgement task.

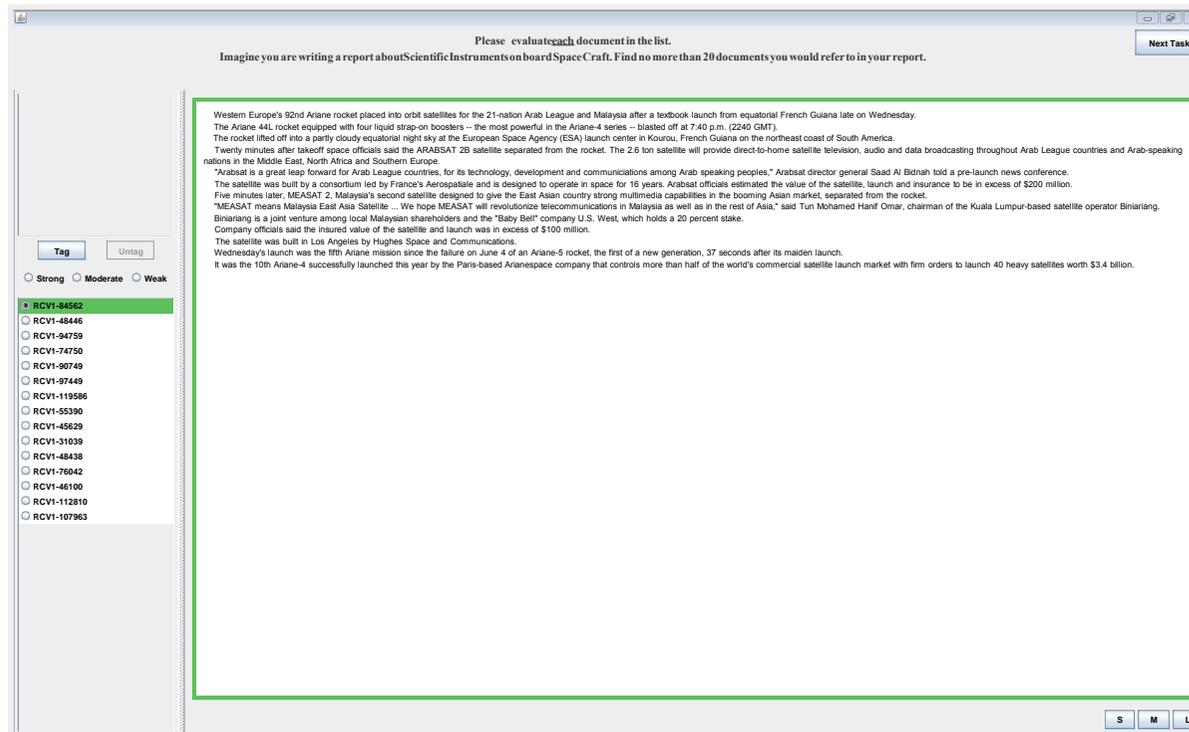


Fig. 6.17: A screen shot of the relevance judgement task interface; all participants completed this task using the same interface configuration.

The set of documents in the relevance judgements task set was periodically and programmatically compiled based on all prior relevance judgements; however, the initial relevance sets were randomly selected from each task set. Each set consisted of five previously relevant rated documents, five irrelevant rated documents and five documents not previously opened. Each participant was randomly assigned a relevance judgement set at experiment time and participants completed relevance judgements for one task set only.

On completion of the experiment, all judgements were collated together; the procedure for obtaining a definitive relevance rating for each document is as follows.

A score of 1-4 is assigned to irrelevant, weak, moderate and strong ratings respectively. For document d , its final relevance score R_d is equal to the sum of the average participant score R_p and the expert score R_E . The average participant score is given by the sum of all participant scores for a document divided by the number of participants scoring document d .

The final relevance calculation can be written in the following equation form:

$$R_d = \frac{\overline{R_P} + R_E}{2}$$

The result of this calculation falls on the range 1-4 inclusive and is seen to approach the average of the expert and participant judgements. This score is then binned according to the rules depicted below in Table 6.6 on the next page, which then assigns a final relevance category for each document.

Table 6.7 on the following page outlines the results of the crowd's average relevance and the expert judgement of relevance while Table 6.8 on the next page presents the corrected relevance scores. The corrected model results in a greater number of irrelevant documents and fewer strongly relevant documents.

Tab. 6.6: Relevance score ranges for relevance calculations.

Range	Bin
$1 \leq x < 1.5$	Irrelevant
$1.5 \leq x < 2.5$	Weak
$2.5 \leq x < 3.5$	Moderate
$3.5 \leq x \leq 4$	Strong

Tab. 6.7: Task set relevance ratings source from experiment population and one expert.

Set	N	Judge	Irrelevant	Weak	Moderate	Strong
Recycled Materials	108	Crowd	57	35	13	3
		Expert	79	14	10	5
Hong Kong Hand Over	134	Crowd	69	32	18	15
		Expert	82	14	23	15
News Paper Circulation	153	Crowd	93	31	25	4
		Expert	108	19	9	17

Tab. 6.8: Task set relevance ratings for corrected model.

Set	Corrected Model			
	Irrelevant	Weak	Moderate	Strong
Recycled Materials	71	23	12	2
Hong Kong Hand Over	74	24	27	9
News Paper Circulation	102	26	21	4

Tab. 6.9: Fictitious document ratings contributing to the production of Gold Standard; labels are determined by way of bin ranges presented in 6.6.

Expert Score	Participant Scores	R_E	$\overline{R_P}$	R_d	Mean Participant Label	Expert Label	Final Label
4	4,4,4	4	4	4	Strong	Strong	Strong
1	4,4,4	1	4	2.5	Strong	Irrelevant	Moderate
4	1,1,1	4	1	2.5	Irrelevant	Strong	Moderate
3	1,2,3	3	2	2.5	Weak	Moderate	Moderate
1	1,2,2	1	1.6	1.3	Weak	Irrelevant	Weak
1	1,1,4	1	2	1.5	Weak	Irrelevant	Weak
4	2,2,3	4	2.3	3.1	Weak	Strong	Moderate

Tab. 6.10: Agreement between expert and crowd.

Set	Corrected Judgements Agreement			
	Irrelevant	Weak	Moderate	Strong
Recycled Materials	86%	30%	25%	50%
Hong Kong Hand Over	77%	33%	30%	77%
News Paper Circulation	87%	27%	19%	50%

A set of fictitious examples presented in Table 6.9 on the facing page and the equation above shows that the overall relevance score for a document is obtained by calculating the average of the participant ratings, then adding that to the experts score, and then dividing by two. Thus, although the expert has, per capita, more weight in deciding the final relevance score of a document, the crowd ratings for a particular document can override the expert.

The crowd and expert routinely agree on what is a relevant document and typically agree on what is a strongly relevant document. In contrast, the expert and crowd disagree on what is a weak or moderately relevant document. By agreement, is meant, where the average participant rating is the same as the experts rating or R_E . For example, when $\overline{R_P}$ and R_E , the average participant rating and the expert rating receives the same bin label.

Calculating agreement follows no sophisticated process; it only requires counting of the number of instances where the average participant rating results in the same bin label as the expert. As Table 6.10 shows, there is quite marked agreement between participants and expert on what constitutes an irrelevant document and somewhat good agreement between what constitutes a relevant document. But since the irrelevant documents largely outweigh relevant documents it is not surprising that there is high agreement on the irrelevant documents there were more opportunities to agree on what was irrelevant compared to what was relevant¹.

However, it would be better to show an objective correlation analysis between expert and participant ratings; this analysis is shown below in Table 6.11 on the following page. Correlation is calculated using the experts score and the binned crowd score i.e. whole integer values. The results indicate that the final ratings more readily favour the experts notion of relevance; however it is apparent that the crowd does have influence over the final rating to some degree.

By observation, each task set reveals an approximately Zipfian scale (Zipf, 1949). However, if in perfect agreement with Zipf there would be approximately twice the number of irrelevant documents, three times the number of weakly relevant, four times the number of moderately relevant documents and five times the number of strongly relevant documents. Based on the average number of documents in irrelevant, weak,

Tab. 6.11: Correlation calculations for expert and crowd - overall and by each task set; R_E is the expert's score, R_P is the mean crowd score and R_d is the final corrected score. There were no relevance judgements made on the training task set.

Set	$R_E \cdot \overline{R_P}$	$\overline{R_P} \cdot R_d$	$R_d \cdot R_E$
Hong Kong	0.750	0.856	0.942
News Paper Circulation	0.604	0.796	0.914
Recycled Materials	0.637	0.801	0.877
Overall	0.674	0.820	0.910

moderate and strong categories, this pattern tends to hold; however, the number of strongly relevant documents is far fewer than a distribution consistent with Zipf.

Overall, the expert is overly strict regarding what constitutes a relevant document with many more documents cited irrelevant in comparison to the crowd. Similarly, the expert is overly strict about what constitutes a strongly relevant document. Consequently, there are far fewer weak relevant documents according to the expert in comparison to the crowd.

The end of this process sets in place a definitive relevance rating to compare against individual participant answer sets, which also enables the calculation of an average performance for each treatment group. After some deliberation, for the ensuing analysis, all documents rated weakly relevant were considered irrelevant while only moderately and strongly relevant documents were considered relevant. This simplified the analysis into two categories of relevance.

Subjective Response

Participants responded to three questionnaires throughout the experiment task - a demographics questionnaire, a subjective response questionnaire and a general knowledge questionnaire.

¹ Furthermore, in the case of the Hong Kong handover to China task set according to the expert there were more moderately relevant documents than weakly relevant documents but according to the crowd there were more weakly relevant documents. This may be attributed to the nature of some of the articles. There were several articles in this task set that reported a sovereignty dispute between China and a regional country. In some cases toward the end of these articles one or two sentences would be devoted to a general statement regarding the handover of Hong Kong from the United Kingdom to China. This is a fact, which could be considered relevant to someone writing an essay; but in the expert's view, this would only constitute a weak relevance rating if it contained a date or other piece of information beyond simply the fact that it was happening. For example, an article discussing the general performance of the Hong Kong economy but mentioning how they - the article's authors - think it's a "...very good result for the Hong Kong economy as it heads into the final stages of transition to Chinese sovereignty (next year)" would be considered irrelevant by the expert. However, an article mentioning "...to revert to Beijing sovereignty at midnight on June 30, 1997" as a filler fact in an article more or less unrelated to the actual event would be considered moderately relevant by the expert. Such a document should be considered moderate but not strong given the occurrence of other documents that not only contain that fact, but also report a number of other relevant facts regarding the hand over of Hong Kong to China.

Demographics Questionnaire A mostly multiple choice demographics questionnaire solicited demographics information. Participants reported their age in years, gender, academic faculty, highest completed academic level, current year of academic study, and whether English is their first spoken language.

Subjective Response A 5-point Likert questionnaire solicited subjective ratings on participant experiences with the experiment interfaces; subjective responses focused on the projection dimension rotation controls only - i.e. the theme cloud and theme list controls. Participants responded to the following questions:

- How successful do you think we were at achieving your goal?
- What level of frustration did you experience?
- How hard did you have to work (mentally and physically)?
- How much time pressure did you feel?
- How much physical activity was required?
- How much mental and perceptual demand was required?

The five point scale spanned very low, low, moderate, high, to very high. It is a defect that a more descriptive explanation regarding each question was not offered to participants. For instance, while the only physical activity the participant was likely to have engaged in would have been mouse activity, a description may have cued participants to think about how their head and eye movement.

No further subjective responses were sought; thus, there were no subjective ratings regarding the ranked-list as this experiment did not aim to build an interface that could out-perform a ranked list. Furthermore, there were no subjective ratings sought for the document full-text view or pop-up transparency.

General Knowledge Questionnaire A general knowledge questionnaire, masked a *Catch Event* task (Oppenheimer, Meyvis, and Davidenko, 2009; Paolacci, Chandler, and Ipeirotis, 2010), and was employed to flag suspicious participation - i.e. participants offering limited participation effort for fast reward. Multiple choice general knowledge questions consistent of questions obtained from the Internet and adaptations from prior user interface experiments. Questions were displayed sequentially with each set of four alternative answers selectable by a multiple choice drop down box. Initially, each question had an answer configured for the participant, with some answers blatantly incorrect. A suspect participant would make no changes to the questionnaire and simply move on with the next task.

6.4.3 Procedure

Participants arrived at an agreed computer laboratory location at their pre-arranged timeslot. Participants were welcomed by the research assistant and then directed to a pre-configured computer workstation. In any one session, there were up to five participants completing the experiment at the same time.

Participants undertook formalities to establish a basis for informed consent; both the participant and the research assistant retained a copy of a signed consent form. The research assistant advised participants that they could leave at any time and for any reason. Then, an overview of the experiment session was delivered, before the official training module commenced.

Participants undertook training with their software apparatus configured for the experiment session's assigned treatment group. Participants completed the training session in two stages and took approximately 30 minutes to complete.

In the first stage, using a wall projector and PowerPoint presentation, the research assistant gave an overview of the key concepts applicable to document spatialisation - a copy of the PowerPoint presentation is included in Appendix D on page 384. In the second stage, participants completed a practice question and were free to raise any issues. To ensure consistency, the research assistant followed a pre-established script - particular to the session's experiment condition - when delivering the training lecture and interface walk through. The research assistant stood at the front of the laboratory while talking, with the PowerPoint presentation projected on to an adjacent wall.

At the end of the lectured training, the participant had 10-15 minutes to practise with the apparatus software using a training task set - the TR task set in Table 6.5 on page 304. Participants were requested not to spend too much time reading text articles at this point; rather they were advised to familiarise themselves with the workings of each component of the interface. While participants were informed that the research assistant would take questions at any stage of the experiment, participants were also provided with paper based, annotated screen shot diagrams - included in Appendix E on page 389 and Appendix F on page 391 - to clarify the functionality of specific interface components.

A schematic of the participant's pathway through the training component of the experiment is shown in Figure 6.18 on the next page.

At the conclusion of training, the research assistant restarted the experiment apparatus. At this point, participants were reminded that the order of interfaces would likely appear in a different order than in training, that there would be previously unseen task sets for each interface, and that they should work quickly, but without sacrificing accuracy. Then, participants were invited to start the experiment at their convenience - taking a break first if they so desired.

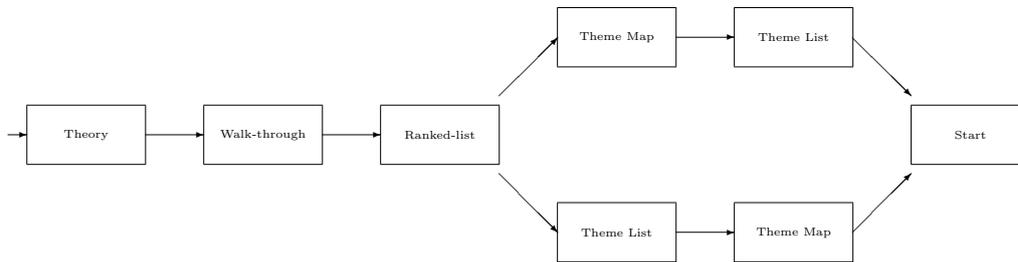


Fig. 6.18: Progression of participant through training stages for experiment; path way between theme list and theme cloud is dependent on random assignment to experiment group - see Table 6.12; document set assigned to ranked-list, theme list and theme cloud stages is the Training Task Set.

As shown in the schematic of the participant's pathway Figure 6.19 on page 314, participants completed the demographics questionnaire, followed by a search task using a ranked-list interface, followed by two separate search tasks using a spatialisation interface, followed by the definitive relevance judgement task, a multiple choice subjective response questionnaire and finally, a multiple choice general knowledge questionnaire.

The participant's search tasks involved reading and evaluating news articles relative to the experiment task statement. The participant was asked to imagine that they were writing a report about a particular topic - present at the top of the interface - and to find no more than twenty articles that they would refer to in their report. There was a minimum of one answer to proceed to the next stage and a recommended limit of 20. Participants were advised that 5-10 good answers would constitute an acceptable effort but they were also advised that it would not be too difficult to find relevant documents.

The idea of a document's relevance was left to the participant but a suggested scenario - see Section 6.4.2 on page 304 - was provided to assist with relevance assessment. To submit an article as an answer, they must have first opened the document, and presumably read it. Having decided that an article was relevant; the participant tagged it as an answer.

Tagging an answer involved a three-step process. In the first step, the participant selected the article ready for tagging. In the second step, the participant rated the relevance of the document by selecting one of the three relevance radio buttons labelled weak, moderate or strong. Finally, in the third step, the participant clicked the button labelled *Tag* which signalled to the software to copy the article into the answer box. The answer box contained a list of all the submitted answers for the task. The participant could review and change the significance rating of individual answers or remove a specific answer at any time throughout the task. Participants were encouraged to modify their relevance judgements over time in light of new discoveries during the experiment.

Participants followed the above procedure for each of three search tasks, with each task conducted on either a ranked-list or spatialisation interface. Likewise, participants followed roughly the same procedure when completing their definitive relevance judgement task.

At the conclusion of the experiment, each participant received their financial reimbursement and departed the laboratory.

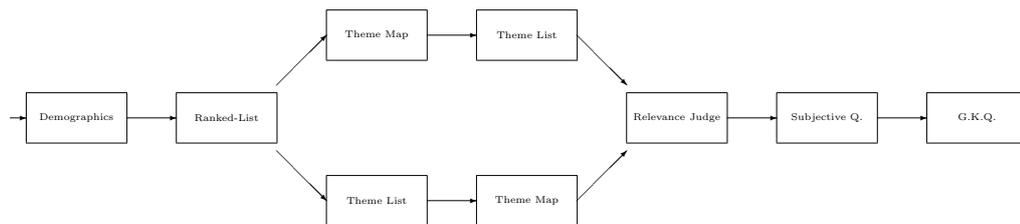


Fig. 6.19: Progression of participant through experiment stages; path way between theme list and theme map is dependent on random assignment to the experiment session group - see Table 6.12; document set assigned to ranked-list, theme list and theme map stages is dependent on random assignment to the participant's document presentation group.

6.4.4 Design

The experiment design was split into two parts for simplicity of analysis, since projection dimension rotation control and pop-up transparency do not factor into the ranked-list interface. Moreover, it was not a primary aim of this experiment to show whether or not a spatialisation-based interface is superior to that of a ranked-list interface. Accordingly, the first analysis was conducted as a one-way factorial analysis of variance and incorporated data from the ranked-list interface. Then, a second analysis was conducted as a 2x2x2 mixed factorial analysis of variance and incorporated data from the spatialisation interfaces.

The one-way analysis had one between subjects factor, document full-text view type of two levels, integrated and modal. The 2x2x2 mixed factorial analysis had one within subjects factor, projection dimension rotation control of two levels theme list and theme cloud, and had two between subjects factors, document full-text view type of two levels, integrated and modal, and pop-up transparency also of two levels, transparent and non-transparent. Each participant was randomly assigned a level of document full-text view and pop-up transparency for the entirety of their experiment session; participants in the same experiment session were assigned the same level of each factor to streamline training. Lastly, regardless of session, participants were randomly assigned a task set order, meaning that participants in the same session were likely to attempt each task in an different order even though their interfaces were structurally similar. The schedule of randomisation is depicted below in Table 6.12 on the facing page.

Tab. 6.12: Schedule of experimental factor randomisation; full-text integration and pop-up transparency are combined in addition to randomisation of the interface presentation order; under Interface Order, L denotes ranked-list, TC denotes theme cloud and TL denotes theme list.

Session Type	Full-Text	Pop-up Transparency	Interface Order
1	Integrated	Non-Transparent	L-TC-TL
2	Integrated	Non-Transparent	L-TL-TC
3	Integrated	Transparent	L-TC-TL
4	Integrated	Transparent	L-TL-TC
5	Modal	Non-Transparent	L-TC-TL
6	Modal	Non-Transparent	L-TL-TC
7	Modal	Transparent	L-TC-TL
8	Modal	Transparent	L-TL-TC

Dependent variables for the one-way analysis were task time in seconds, Bookmaker Informedness, number of documents opened and the number of re-ranking operations conducted. In contrast, dependent variables for the 2x2x2 mixed factorial analysis were task time, Bookmaker Informedness, number of documents opened, number of projection dimension configurations and the proportion of trial spent with the multi-pop-up facility toggled on.

Bookmaker Informedness was the main answer quality metric (Powers, 2003) and is an unbiased measure that rewards both successes and failures in a classification task such as rating the relevance of documents. Succinctly, Bookmaker Informedness specifies the probability that a prediction is informed by the condition relative to chance (Powers, 2003). Powers suggests that Bookmaker is more appropriate than traditional measures Recall and Precision as the former takes into consideration true negative classifications whereas Recall does not; furthermore, Bookmaker offers one measure instead of two.

To calculate the quality of an individual's answer set, the following process was adhered to. Each document opened and rated by the participant is compared against the definitive relevance rating. True positives are those documents rated either strongly or moderately relevant by the participants and likewise in the definitive relevance set; true negatives are those rated irrelevant or weakly relevant by the participant and in the definitive relevance set; false positives are those documents rated relevant by the participant but irrelevant in the definitive relevance set; and false negatives are those which are rated irrelevant by the participant but relevant in the definitive relevance set.

From a contingency table of these measures, Informedness is calculated (Powers, 2003). Informedness is calculated as $Recall + InverseRecall - 1$. Informedness may be graphed on a Receiver Operating Characteristic ROC graph by plotting a participant's

true positive rate against their false positive rate and observing the vertical distance between a point in the graph and a diagonal chance line from the origin point out to (1,1). Further discussion on the metrics of Bookmaker Informedness in information retrieval are available in Appendix H.

6.5 Results

This experiment sought to observe different search behaviour under three design factors in a search result interface. In spatialisation-based interfaces, should pop-up windows that display document surrogate information be transparent to make spatial data easier to see. Second, should documents selected for full-text view appear in a modal window or adjacent to the result visualisation. Third, is a theme cloud or a theme list a superior interface control for supporting exploration of thematic aspects of the result set and therefore manipulating theme map layout?

Where reported, mean results are reported with 95% Confidence Intervals. A list of the statistical procedures adopted for this analysis are included in Appendix H; for significance testing, the maximum rate of Type 1 error was set at $\alpha = 0.05$.

All participants attempted the *catch event* general knowledge questionnaire in good faith; therefore, on this basis, no participant was excluded. The average score was 12/19 (8 min, 18 max); the expected score by random selection was 5/19 and a score that flagged no submission - i.e. submission of the default answers - was 7/19.

6.5.1 Analysis One: Analysis of Document Full-text View in Ranked-list Interface

Multiple one-way analyses of variance was conducted to assess the effect of document full-text view on task time, Bookmaker Informedness, number of documents opened and the number of ranked-list re-ranking actions. These analyses considered only data from the ranked-list interface - the results of which are tabulated in Table 6.13.

An initial exploration of the data revealed the presence of one extreme outlier who had completely ignored the directive to work quickly and without sacrificing accuracy; this outlying participant had opened every document in the document set at least once.

Participants were slower to complete tasks with an integrated full-text view ($M=546.84$ seconds, $SD=274.95$) than with a modal full-text view ($M=437.41$ seconds, $SD=289.02$). An analysis of variance revealed that this difference was not significant $F(1,49)=1.89$, $p=0.17$; Figure 6.20 on the next page depicts these differences visually. In addition, participants, using an integrated full-text view achieved a lower Bookmaker Informedness score ($M=0.50$, $SD=0.25$) than did participants using a modal full-text view ($M=0.53$, $SD=0.35$). This difference was not deemed significant by an analysis of variance test $F(1,49)=0.14$, $p=0.70$; Figure 6.21 on page 318 depicts these differences visually.

Full-Text	N	$\mu_{Time}(\sigma)$	$\mu_{Bookmaker}(\sigma)$	$\mu_{Opened}(\sigma)$	$\mu_{ListSort}(\sigma)$	$\mu_{ ListSortVector }(\sigma)$
Integrated	23	546.84 (274.95)	0.50 (0.25)	37.57 (34.72)	6.30 (6.02)	8.49 (7.19)
Modal	28	437.41 (289.02)	0.53 (0.35)	15.39 (8.56)	4.39 (6.58)	7.46 (8.61)

Tab. 6.13: Descriptive statistics for document full-text view factor; mean values quoted with bracketed standard deviation and number of trials in analysis one.

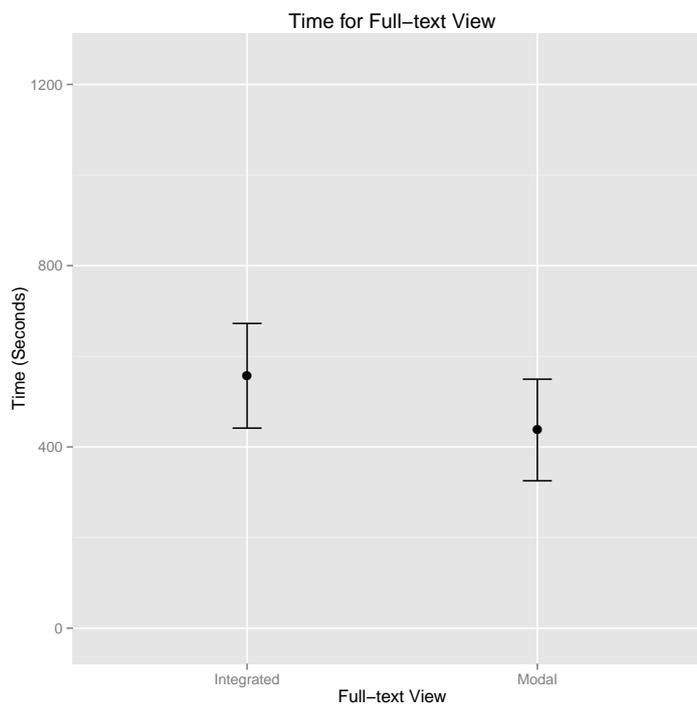


Fig. 6.20: A graph of time (in seconds) for full-text integration; error bars are 95% Confidence Intervals.

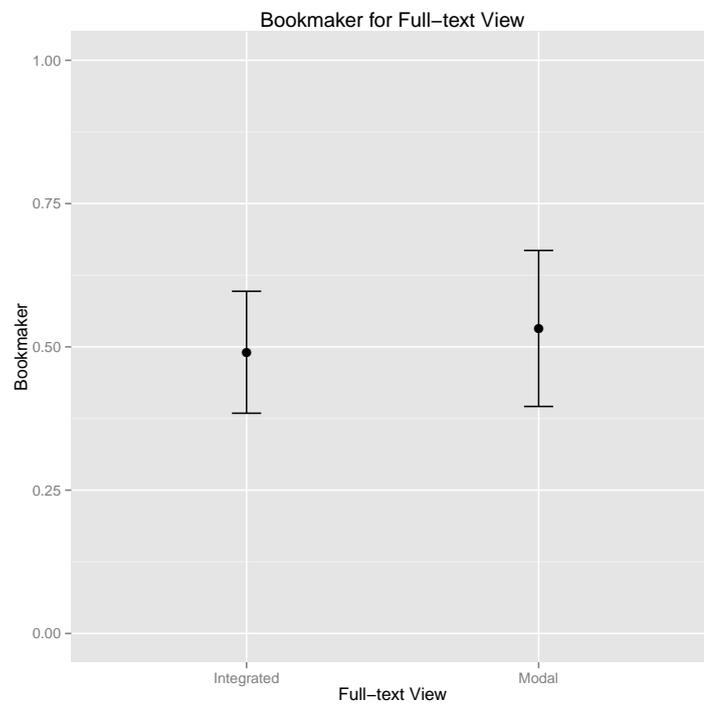


Fig. 6.21: A graph of Bookmaker for full-text integration; error bars are 95% Confidence Intervals.

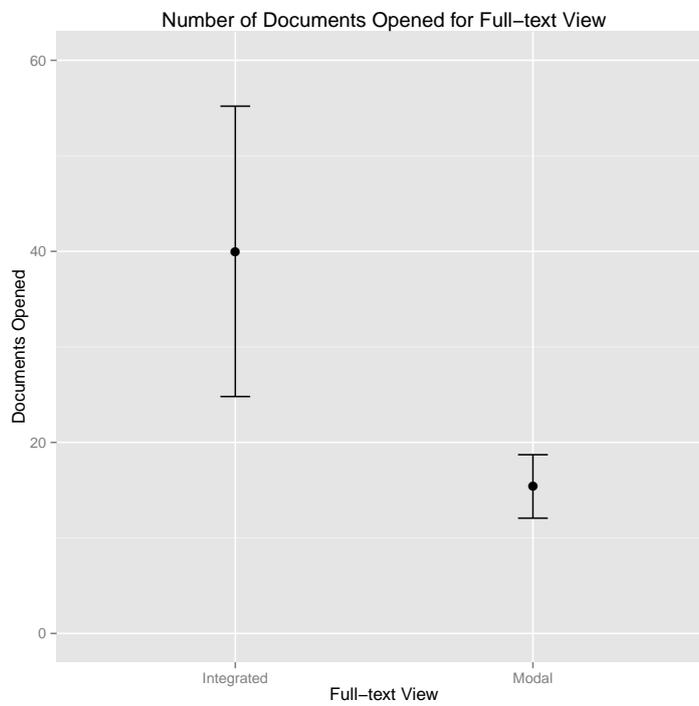


Fig. 6.22: A graph of the number of documents opened for full-text integration; error bars are 95% Confidence Intervals.

Moreover, regardless of full-text integration, on average, nearly half of all participants achieved an accuracy score of slightly worse than that expected by random selection of documents from the document set. Finally, participants opened markedly more documents under an integrated full-text view ($M=37.57$ documents, $SD=34.81$) than those under a modal full-text view ($M=15.39$ documents, $SD=8.56$). An analysis of variance indicated that this difference was significant $F(1,49)=10.67$, $p=0.001$. A conservative Bonferroni post hoc correction for multiple comparisons indicates that this finding is statistically significant i.e. $0.001 < 0.0125$. Figure 6.22 on the previous page depicts these differences visually.

Finally, an analysis of ranked-list re-ranking behaviour was conducted. There were 15/51 (29%) participants who opted not to make any re-ranking actions whatsoever and answered the trial question by utilising the results present in the ranked-list. Furthermore, 19/23 (82%) participants in the integrated full-text view utilised the re-ranking facility, while only 17/28 (60%) participants in the modal full-text view condition utilised the re-ranking facility. These results are available in Table 6.14 and Figure 6.23 on the next page.

Of participants whom did re-sort the ranked-list, there was practically no difference in the number of re-sorting actions made under the modal full-text view ($M=7.23$ re-sorting actions, $SD=7.16$) than the integrated full-text view ($M=7.6$ re-sorting actions, $SD=5.80$). Furthermore, the average number of unique words in the re-sorting vector under modal full-text view ($M=12.29$ words, $SD=7.88$) was higher than in the integrated full-text view ($M=10.27$ words, $SD=6.69$) and this is depicted visually in Figure 6.24 on the facing page. Separate one-way analyses of variance were conducted with the number of re-sorting actions and number of unique words incorporated into the re-sorting vector as the dependent variables and full-text view integration as the independent variable; differences between re-sorting actions $F(1,34)=0.03$, $p=0.85$ were not significant nor were re-sorting vector length differences significant $F(1,34)=0.69$, $p=0.41$. Due to the large proportion of participants opting not to re-sort the ranked-list, a Chi-Square analysis was conducted to investigate any effect of full-text view on whether the re-ranking facility was utilised. However, there was no evidence to support a relationship between full-text view integration and re-ranking behaviour $\chi(1,N=51)=2.915$, $p=0.08$.

Tab. 6.14: Proportion of participants who did and did not utilise resorting functionality in the ranked list stage.

Full-text Integration	Used Resort	
	No	Yes
Integrated	4	19
Modal	11	17
Total:	15	36

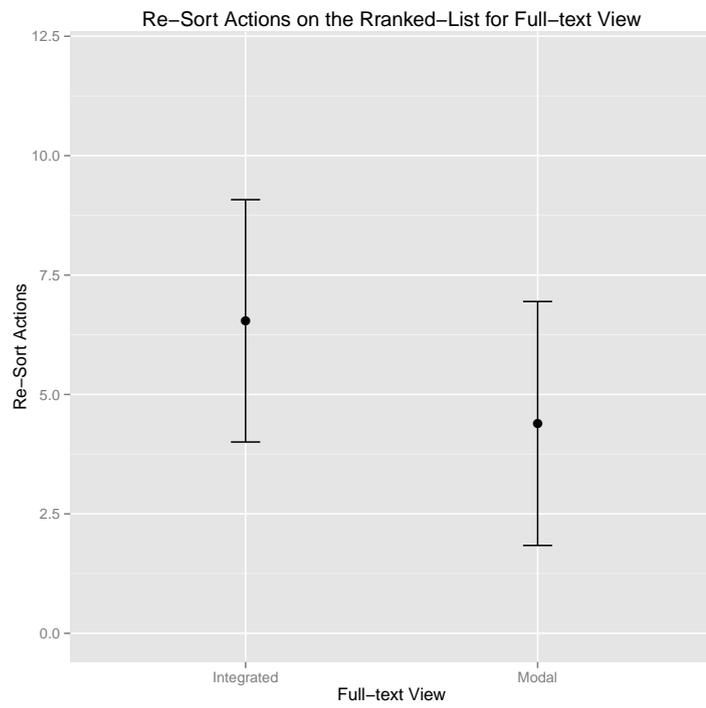


Fig. 6.23: A graph of the number of ranked-list resort actions for full-text integration; error bars are 95% Confidence Intervals.

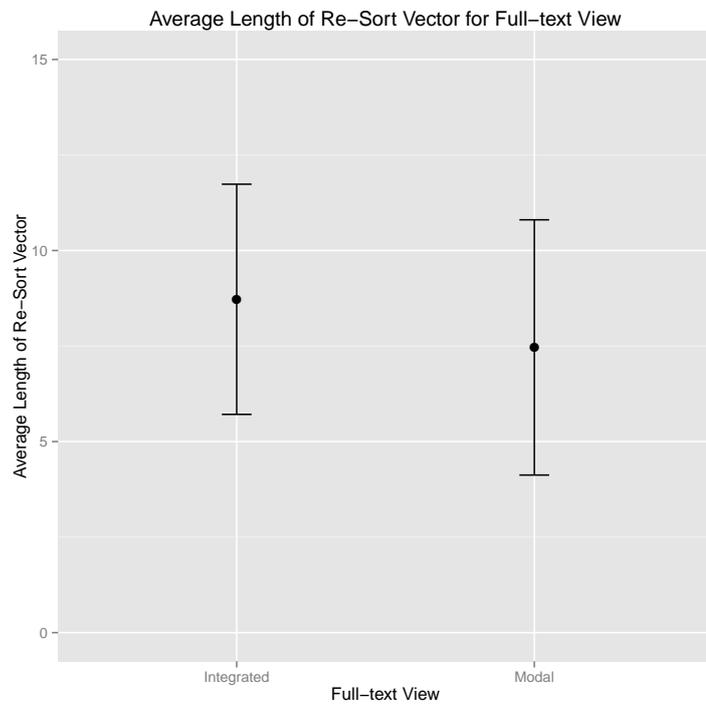


Fig. 6.24: A graph of the average length of the re-sort vector for full-text integration; error bars are 95% Confidence Intervals.

6.5.2 *Analysis Two: Analysis of Document Full-text View, Pop-up Transparency and Projection Dimension Control in Theme Map*

Multiple 2x2x2 mixed factorial analyses were conducted to assess the effect of document full-text view integration, pop-up transparency, and spatialisation projection dimension control on task time, Bookmaker Informedness, number of documents opened, pop-up usage and projection rotations. This analysis considered only data collected from the two spatialisation trials - demarcated as the theme cloud and theme list. In this analysis, one participant was excluded from the analysis as they opened every document in the document set; this participant was also excluded from the previous analysis.

Pair-wise correlation analyses of the dependent variables yielded little evidence to suggest moderate correlations between dependent variables with the exception of time and number of documents opened. Accordingly, these analyses motivated the use of multiple mixed factorial analyses of variance (ANOVA) and not a mixed multivariate analysis of variance (MANOVA). In analysing the data for significance, multiple 2x2x2 mixed factorial analyses were conducted, with trial type as the within subjects factor - theme cloud and theme list; and document full-text view integration - with two levels: integrated and modal - and pop-up transparency - with two levels: transparent and non-transparent - as the between subjects factor. An 2x2x2 mixed factorial ANOVA was conducted for each of the dependent variables for time, Bookmaker Informedness, number of documents opened, pop-up usage, and number of projection rotations i.e. spatialisation configurations.

Time, Bookmaker score, documents opened, projection dimension rotations and pop-up usage are presented in Table 6.16 on the next page for document full-text view overall, in Table 6.18 on page 330 for pop-up background transparency overall, in Table 6.17 on the next page for projection dimension control overall, in Table 6.19 on page 330 for document full-text view and pop-up background transparency and finally in Table 6.15 on the next page for full-text view, pop-up background transparency and projection dimension control. Results for time and Bookmaker score are described for the full complement of results tables while pop-up usage and rotations are described only for tables incorporating the pop-up factor and only tables incorporating the projection dimension control factor, respectively. In addition, Figure 6.20 on page 317 through Figure 6.24 on the previous page depict time, Bookmaker score, documents opened, projection dimension rotations and pop-up usage, by visual means.

Full-Text	Pop Trans.	Projection Control	N	$\mu_{Time}(\sigma)$	$\mu_{Bookmaker}(\sigma)$	$\mu_{Opened}(\sigma)$	$\mu_{PopupUsage}(\sigma)$	$\mu_{Rotations}(\sigma)$
Integrated	Non-Trans.	Theme Cloud	11	416.90 (135.52)	0.27 (0.36)	26.00 (14.36)	0.25 (0.32)	9.18 (7.41)
		Theme List	11	434.23 (158.03)	0.49 (0.25)	32.00 (19.66)	0.31 (0.29)	13.64 (11.81)
	Trans.	Theme Cloud	12	567.70 (268.31)	0.47 (0.30)	35.67 (19.80)	0.68 (0.46)	11.58 (10.41)
		Theme List	12	577.15 (273.36)	0.43 (0.17)	31.92 (13.41)	0.84 (0.36)	12.75 (13.30)
Modal	Non-Trans.	Theme Cloud	13	378.93 (234.26)	0.44 (0.47)	9.46 (2.70)	0.63 (0.42)	6.62 (9.79)
		Theme List	13	375.01 (139.67)	0.39 (0.52)	10.00 (3.49)	0.65 (0.39)	7.00 (3.67)
	Trans.	Theme Cloud	15	472.14 (238.38)	0.43 (0.29)	22.47 (16.70)	0.50 (0.44)	5.47 (4.96)
		Theme List	15	514.32 (324.64)	0.55 (0.30)	19.07 (9.90)	0.50 (0.48)	8.20 (4.69)

Tab. 6.15: Descriptive statistics for document full-text view, pop-up transparency and projection dimension control factors; mean values quoted with bracketed standard deviation and number of trials.

Full-Text	N	$\mu_{Time}(\sigma)$	$\mu_{Bookmaker}(\sigma)$	$\mu_{Opened}(\sigma)$	$\mu_{Rotations}(\sigma)$	$\mu_{Pop-upUsage}(\sigma)$
Integrated	23	502.19 (225.93)	0.42 (0.28)	31.50 (16.84)	11.80 (10.76)	53% (43%)
Modal	28	439.25 (247.36)	0.45 (0.39)	15.64 (11.52)	6.82 (6.06)	56% (43%)

Tab. 6.16: Descriptive statistics for document full-text view factor in analysis two; mean values quoted with bracketed standard deviation and number of trials.

Projection Control	N	$\mu_{Time}(\sigma)$	$\mu_{Bookmaker}(\sigma)$	$\mu_{Opened}(\sigma)$	$\mu_{Rotations}(\sigma)$	$\mu_{Pop-upUsage}(\sigma)$
Concept Cloud	51	458.95 (231.38)	0.41 (0.36)	23.02 (17.19)	8.00 (8.40)	52% (44%)
Concept List	51	476.32 (248.11)	0.47 (0.34)	22.57 (15.25)	10.14 (9.18)	58% (42%)

Tab. 6.17: Descriptive statistics for projection dimension control factor in analysis two; mean values quoted with bracketed standard deviation and number of trials.

Overall, participants were slower ($M=502.19$ seconds, $SD=225.93$) when full-text was integrated in contrast to when full-text was presented modally ($M=439.25$ seconds, $SD=247.36$). Likewise, participants were slower to complete tasks when pop-up backgrounds were transparent ($M=528.43$ seconds, $SD=274.12$) compared to participants completing tasks when they were non-transparent ($M=399.24$ seconds, $SD=169.85$). Overall, participants were fastest to complete tasks under modal full-text view and non-transparent pop-up backgrounds ($M=376.97$ seconds, $SD=188.97$) followed by participants with integrated full-text and non-transparent backgrounds ($M=425.56$ seconds, $SD=143.93$), modal full-text and transparent pop-up backgrounds ($M=493.23$ seconds, $SD=280.66$) and slowest to complete tasks with integrated full-text and transparent pop-up backgrounds ($M=572.43$ seconds, $SD=264.94$). In addition, overall, participants were slowest to complete tasks using the theme list ($M=476.32$ seconds, $SD=248.11$) than the theme cloud ($M=458.95$ seconds, $SD=231.38$). A mixed factorial ANOVA was conducted using document full-text view and pop-up transparency as the between subjects variables, projection dimension rotation control as the within subjects variable and time as the dependent variable. The analysis indicated a significant main effect for pop-up transparency $F(1,47)=5.79$, $p=0.02$ only.

Overall, participants had a higher Bookmaker score ($M=0.45$, $SD=0.39$) when the document full-text was modally presented in contrast to when the document full-text view was integrated ($M=0.42$, $SD=0.28$). In addition, when pop-up backgrounds were transparent Bookmaker score was higher ($M=0.47$, $SD=0.27$) compared to when pop-up backgrounds were non-transparent ($M=0.40$, $SD=0.42$). Accordingly, Bookmaker score was highest when document full-text view was modally presented and pop-up windows transparent ($M=0.49$, $SD=0.30$), followed by integrated full-text and transparent pop-ups ($M=0.45$, $SD=0.39$), then modal full-text view and non-transparent pop-up windows ($M=0.41$, $SD=0.49$) and lowest when document full-text was integrated and pop-up windows non-transparent ($M=0.38$, $SD=0.32$). In addition, on average, a higher Bookmaker score was achieved by participants completing a task with the theme list ($M=0.47$, $SD=0.34$) than with the theme cloud ($M=0.41$, $SD=0.36$). Consistent with the above results, participants completing tasks using the theme list with modal full-text and transparent pop-up backgrounds, achieved the highest Bookmaker score ($M=0.55$, $SD=0.30$) compared to the group completing tasks using the theme cloud with integrated full-text and non-transparent backgrounds ($M=0.27$, $SD=0.36$). A mixed factorial ANOVA was conducted using document full-text view and pop-up transparency as the between subjects variables, projection dimension rotation control as the within subjects variable and Bookmaker as the dependent variable. This analysis indicated no significant interaction effects nor main effects for the variables under investigation.

Participants opened markedly greater numbers of documents when document full-text view was integrated ($M=31.50$ documents, $SD=16.84$) than when document full-

text view was modally presented ($M=15.64$ documents, $SD=11.52$). Furthermore, notably more documents were opened when pop-up windows were transparent ($M=26.56$, $SD=16.25$) than when they were non-transparent ($M=18.56$, $SD=15.15$). In addition, projection rotation control offered no such marked differential between observations.

Participants opened markedly fewer documents when document full-text was presented modally and pop-up windows were non-transparent ($M=9.73$ documents, $SD=3.07$), followed by modal document full-text and transparent pop-up windows ($M=20.77$ documents, $SD=13.60$). The next highest document count was under integrated document full-text view conditions and non-transparent pop-up windows ($M=29.00$ documents, $SD=17.08$) and finally, highest with an integrated document full-text view and transparent windows ($M=33.79$ documents, $SD=16.65$). The group with the highest number of documents opened had high variation, while the group with the lowest number of documents opened had the least variation as indicated by the comparatively and markedly low standard deviation.

A mixed factorial ANOVA was conducted using document full-text view and pop-up transparency as the between subjects variables, projection dimension rotation control as the within subjects variable and number of documents opened as the dependent variable. The analysis indicated no significant interaction effects. In contrast, a significant main effect of pop-up transparency $F(1,47)=5.63$, $p=0.02$ was observed and a significant main effect of document full-text view $F(1,47)=23.47$, $p<0.0001$.

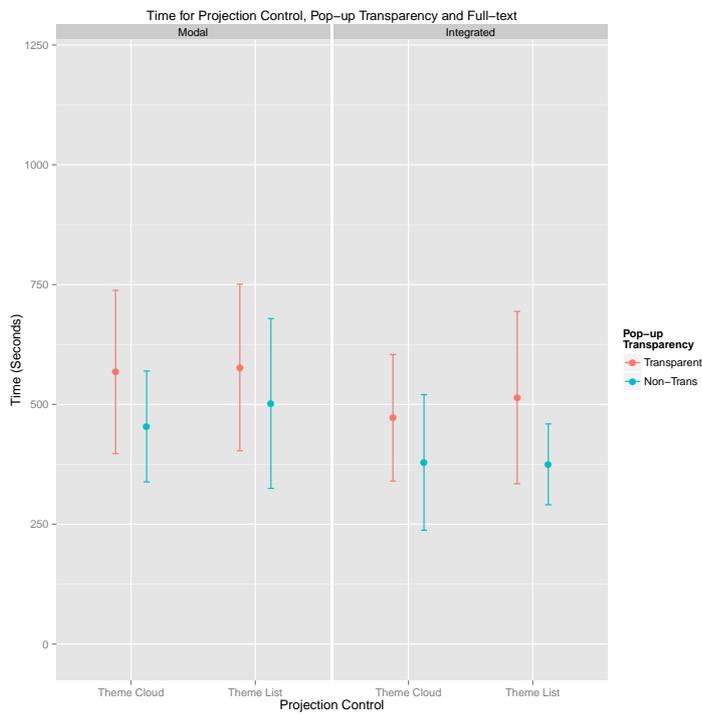


Fig. 6.25: A graph of time (in seconds) for projection control, pop-up transparency and full-text integration; error bars are 95% Confidence Intervals.

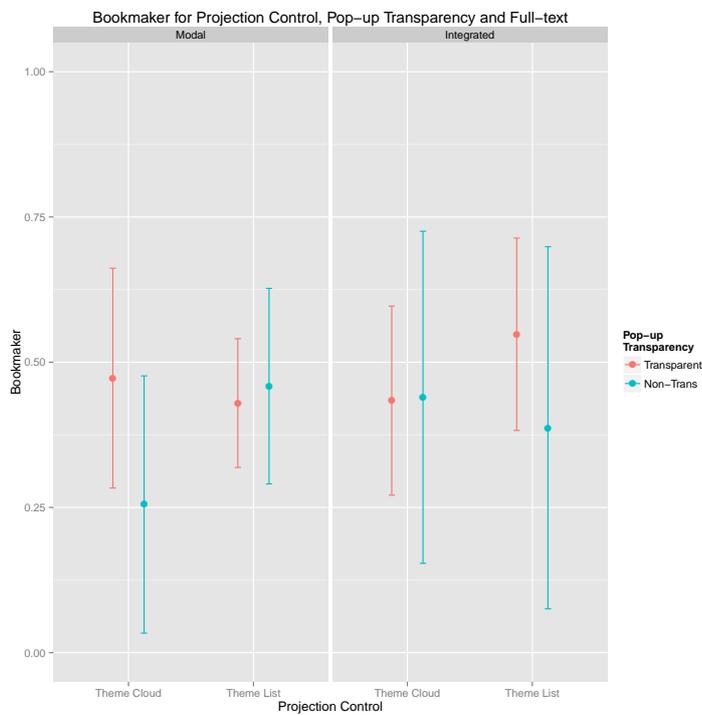


Fig. 6.26: A graph of Bookmaker score for projection control, pop-up transparency and full-text integration; error bars are 95% Confidence Intervals.

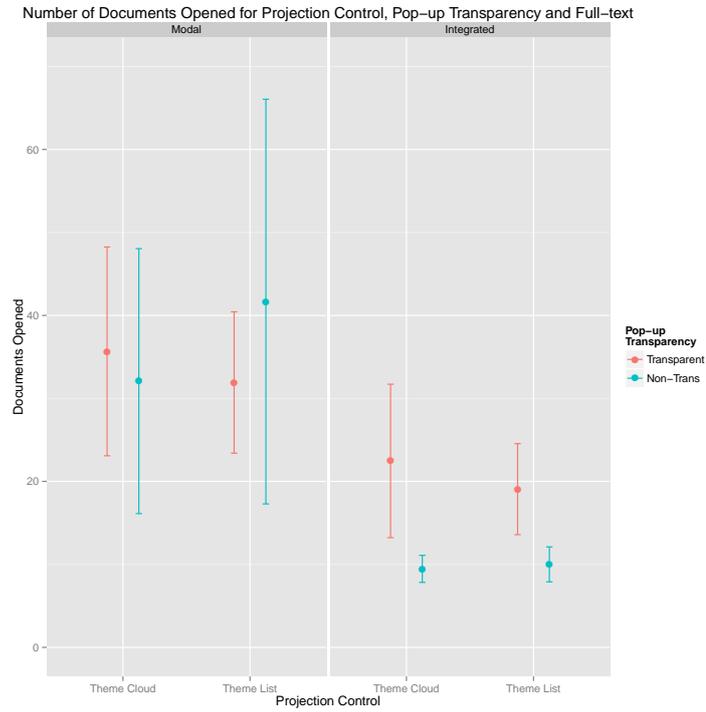


Fig. 6.27: A graph of the number of documents opened for projection control, pop-up transparency and full-text integration; error bars are 95% Confidence Intervals.

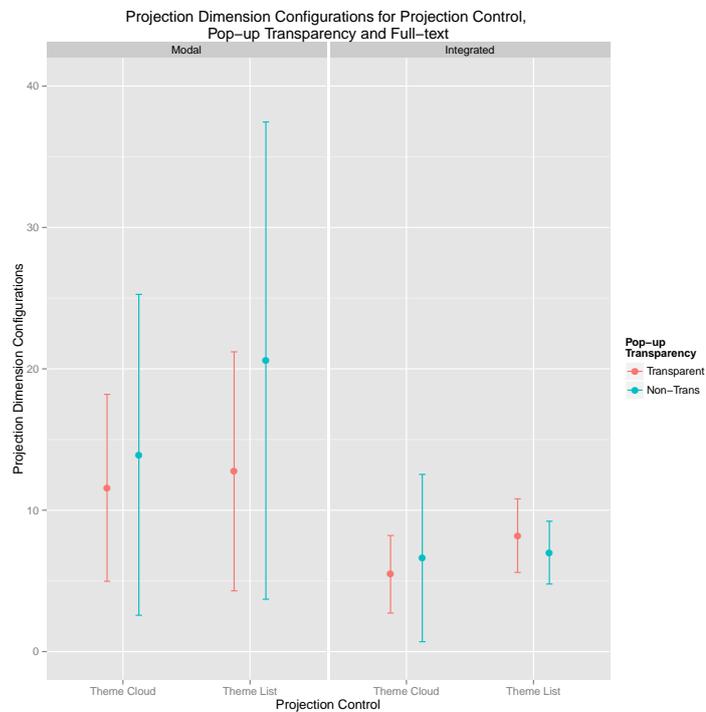


Fig. 6.28: A graph of the number of projection dimension configurations for projection control, pop-up transparency and full-text integration; error bars are 95% Confidence Intervals.

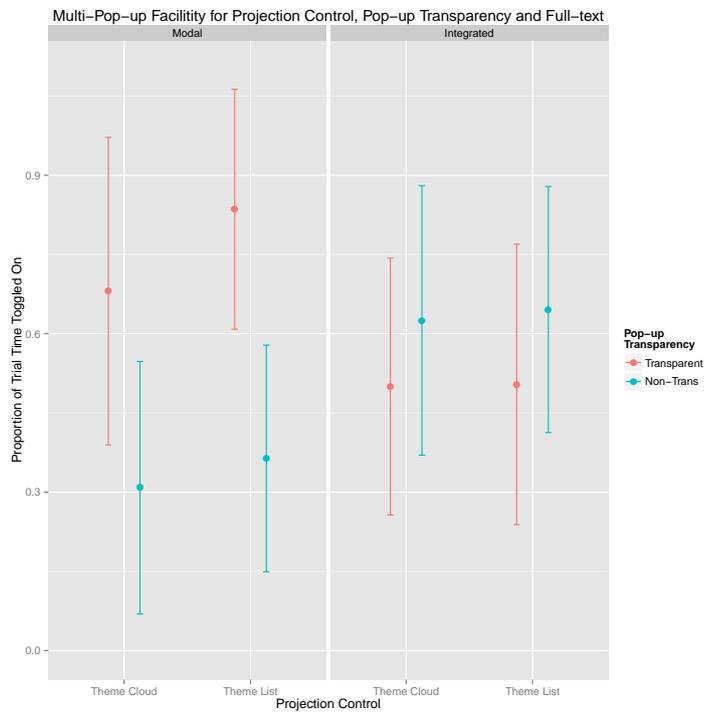


Fig. 6.29: A graph of the proportion of trial time spent with the multi pop-up facility active for projection control, pop-up transparency and full-text integration; error bars are 95% Confidence Intervals.

Document full-text view had a noteworthy influence on the number of projection dimension rotations. Overall, participants carried out more rotations when document full-text view was integrated (M=11.80 rotations, SD=10.76) compared with when document full-text view was modal (M=6.82 rotations, SD=6.06). In contrast and overall, rotation control had very little influence on the number of rotations made; participants completed marginally more rotations using the theme list (M=10.14 rotations, SD=9.18) than the theme cloud (M=8.00 rotations, SD=8.40).

A mixed factorial ANOVA was conducted using document full-text view and pop-up transparency as the between subjects variables, projection dimension rotation control as the within subjects variable and number of projection dimensions rotated as the dependent variable. The analysis indicated no significant interaction effects; however, a main effect for document full-text view was observed $F(1,47)=6.80$, $p=0.01$.

Turning to participants' usage of the multiple pop-up window facility, pop-up window transparency had a marked influence on whether or not participants opted to revert to a single pop-up window facility. Overall, the multiple pop-up facility was de-activated more readily by the non-transparent pop-up window group (M=47% trial time activation, SD=40%) compared to the transparent pop-up window group (M=62% trial time activation, SD=45%). In other words, participants in the transparent pop-up window group retained a multiple pop-up facility for a greater proportion of their trials than participants in the non-transparent pop-up group.

However, overall, document full-text view had little impact with participants in the integrated full-text view opting to deactivate the multiple pop-up facility for slightly longer periods (M=53% trial time activation, SD=43%) than participants in the modal full-text view group (M=56% trial time activation, SD=43%). Despite the comparatively small difference between document full-text groups, participants retained the multiple pop-up facility for the longest under integrated document full-text and transparent pop-up windows (M=76% trial time activation, SD=41%) whilst the integrated document full-text and non-transparent pop-up window group retained the facility for the shortest proportion of trial time (M=28%, SD=30%). Table 6.19 on the following page offers further insight to this observation.

A mixed factorial ANOVA was conducted using document full-text view and pop-up transparency as the between subjects variables, projection dimension rotation control as the within subjects variable and multiple pop-up facility trial activation as the dependent variable. The analysis indicated a significant interaction effect between document full-text view and pop-up transparency $F(1,47)=8.17$, $p=0.006$ but for no other interactions. Furthermore, there was no significant main effect observed for either pop-up transparency $F(1,47)=2.62$, $p=0.11$ nor document full-text view $F(1,47)=0.22$, $p=0.63$.

Pop-up Transparency	N	$\mu_{Time}(\sigma)$	$\mu_{Bookmaker}(\sigma)$	$\mu_{Opened}(\sigma)$	$\mu_{Rotations}(\sigma)$	$\mu_{Pop-upUsage}(\sigma)$
Non-Trans.	24	399.24 (169.85)	0.40 (0.42)	18.56 (15.15)	8.92 (8.78)	47% (40%)
Trans.	27	528.43 (247.12)	0.47 (0.27)	26.56 (16.25)	9.20 (8.94)	62% (45%)

Tab. 6.18: Descriptive statistics for pop-up transparency factor in analysis two; mean values quoted with bracketed standard deviation and number of trials.

Full-Text View	Pop-up Transparency	N	$\mu_{Time}(\sigma)$	$\mu_{Bookmaker}(\sigma)$	$\mu_{Opened}(\sigma)$	$\mu_{Rotations}(\sigma)$	$\mu_{Pop-upUsage}(\sigma)$
Integrated	Non Trans.	22	425.56 (143.93)	0.38 (0.32)	29.00 (17.08)	11.41 (9.89)	30% (29%)
	Trans.	24	572.43 (264.94)	0.45 (0.24)	33.79 (16.65)	12.17 (11.69)	41% (76%)
Modal	Non Trans.	26	376.97 (188.97)	0.41 (0.49)	9.73 (3.07)	6.81 (7.25)	40% (64%)
	Trans.	30	493.23 (280.66)	0.49 (0.30)	20.77 (13.60)	6.83 (4.94)	45% (50%)

Tab. 6.19: Descriptive statistics for pop-up transparency and document full-text view factors in analysis two; mean values quoted with bracketed standard deviation and number of trials.

6.5.3 Subjective Response

In addition to objective measurements, subjective responses were sought on estimated physical and mental workload, time pressure and success rate for both the theme cloud and theme list layout controls.

Looking at the subjective results in Table 6.21 on the next page, median and mode scores for the theme list were largely stable across treatment groups, in contrast to participants reporting for the theme cloud as evidenced by Table 6.20 on the following page - except for *overall workload*, in which participants were consistently undecided.

Middle level responding was common for participants offering subjective opinions on the theme cloud. Median scores for all but *physical workload* were at the moderate level, and similarly, mode scores for all but *physical workload* and *frustration* were moderate also. In contrast, frequency of overall moderate loads and exertions for the theme list were less frequent and mirrored that of theme cloud responses in the same category.

The subjective results for the theme list indicated that *frustration* and estimated *overall time* were lower, while *success* rating was higher; moreover, frustration and overall time was higher in ratings of the theme cloud. Nevertheless, participants reported positively when providing their estimation of task *success* in combination with a theme cloud - though less so than when completing tasks with the theme list.

Participants suggested a clear advantage of the theme list over the theme cloud on the *overall success* dimension - theme list more successful. In addition, some advantage is apparent on dimensions including *frustration* and estimations of *overall time* - despite similarities on mode score. A non-parametric Wilcoxon Signed-rank test offered evidence to suggest that *overall success* ($Z=-2.06$, $p=0.03$, $r=0.28$), *time* ($Z=2.09$, $p=0.03$, $r=0.29$) and *physical workload* were superior in the theme list condition. However, there was no strong evidence to suggest that *frustration* ($Z=1.95$, $p=0.05$, $r=0.27$) was superior in the theme list condition, nor *perceptual workload* ($Z=0.95$, $p=0.35$, $r=0.13$) or *overall workload* ($Z=1.73$, $p=0.08$, $r=0.24$).

Tab. 6.20: Subjective workload response data for Theme Cloud control for Pop-up Transparency and Full-Text Integration factors; Median and (Mode) score provided.

Theme Cloud							
Full-Text Integration	Pop-up Transparency	Perceptual Workload	Physical Workload	Overall Time	Overall Workload	Frustration	Success
Integrated	Non-Trans.	3 (3)	3 (3)	2 (2)	3 (3)	3 (2)	3 (3)
	Transparent	3.5 (4)	3 (2)	3 (4)	3 (3)	3 (2)	3 (3)
Modal	Non-Trans.	3 (3)	2 (1)	3 (3)	3 (3)	3 (3)	3 (3)
	Transparent	3 (3)	2 (2)	2 (2)	3 (3)	3 (4)	3 (4)
Overall		3 (3)	2 (2)	3 (2)	3 (3)	3 (2)	3 (3)

Tab. 6.21: Subjective workload response data for theme list control for pop-up transparency and full-text integration factors; Median and (Mode) score provided.

Theme List							
Full-Text Integration	Pop-up Transparency	Perceptual Workload	Physical Workload	Overall Time	Overall Workload	Frustration	Success
Integrated	Non-Trans.	3 (3)	2 (2)	2 (2)	3 (3)	2 (2)	4 (4)
	Transparent	3 (3)	2 (2)	2 (2)	3 (2)	2 (2)	4 (4)
Modal	Non-Trans.	3 (3)	2 (2)	2 (2)	3 (3)	2 (2)	3 (3)
	Transparent	3 (3)	2 (2)	2 (2)	3 (3)	2 (2)	4 (4)
Overall		3 (3)	2 (2)	2 (2)	3 (3)	2 (2)	4 (4)

6.6 Discussion

The aim of this experiment was to determine the superior option for three interface design choices and to examine how they affect a searcher's exploration of an information space. First, how does the integration of document full-text view influence task performance and outcome; second, how does the transparency of pop-up windows, superimposed over a document spatialisation, influence task performance and outcome; and thirdly, does the choice of interactive interface control used to navigate a multi-dimensional space influence task performance and search outcome. The analysis took place in two stages, and accordingly, this discussion will address each analysis individually and then discuss the experiment as a whole.

6.6.1 Relevance Assessment Methodology

Each reference to outcome quality, in the tables of results in the following sections, is multifaceted. Specifically, outcome quality refers to Bookmaker Score, although each table as a whole represents an overall measure of overall success. Overall success is contingent of finding relevant documents - as indicated by a higher Bookmaker score - however, the importance of speed and interaction cost cannot be ignored.

In this instance, Bookmaker Score acted as the quality metric on which to assess the relevance of documents obtained in each experiment condition. Although regardless of the chosen relevance metric, an accurate assessment of a system's capacity to meet a searcher's information need, is dependent on a repeatable and reliable authoritative standard of relevance. Traditionally, the production of such a standard has involved the use of a small number of topic experts; however, this process can be resource intensive, particularly with increasing task set size - i.e. corpus size.

The construction of this experiment's gold standard - as was outlined in 6.4.2 on page 304 - adopted a hybrid approach consisting of gold standard judgements - by a single topic expert - and a number of crowd-sourced bronze standard judgements (P. Bailey et al., 2008). This approach somewhat deviated from a more traditional method for relevance assessment, in favour for an emerging methodology in which relevance judgements are obtained via crowd-sourcing.

The selected relevance assessment methodology was motivated by many of the same factors listed by proponents of crowd-sourced relevance judgements. Primarily, these include savings in time and resource expenditure (Clough et al., 2013; Alonso, Rose, and Stewart, 2008), scalability (Blanco et al., 2011) and flexibility (Alonso, Rose, and Stewart, 2008). Nevertheless, however well-intentioned and advantageous this methodology may be, the reliability of crowd-sourced relevance judgements has not yet received widespread acceptance.

P. Bailey et al. (2008) presented analysis revealing lower levels of agreement between gold assessors - both topic experts and topic creators - and silver assessors - topic experts only - and also between gold assessors and bronze assessors - neither topic experts nor topic creators.

Similar disparities between experts and non-experts have been observed (Blanco et al., 2011; Clough et al., 2013); however, these studies - including P. Bailey et al. (2008) - also highlight that crowd-source judgements at least demarcate the best and worse performing systems and moreover, provide a comparable ranking of systems to that of rankings based on gold standard relevance assessments (Clough et al., 2013; Nowak and R uger, 2010; Alonso and Mizzaro, 2009). Moreover, gold standard judgements are themselves not without flaw, with anomalies and inconsistencies observed across a single expert's set of relevance judgements (Scholer, Turpin, and Sanderson, 2011) and evidence to show that even expert judges can introduce erroneous relevance judgements (Alonso and Mizzaro, 2009).

In this particular instance, the use of a hybrid assessment methodology, is acceptable given: the exploratory nature of the experiment; similar to earlier research, an observed pattern of negative ratings in which experts are more likely to rate documents as negative (Blanco et al., 2011) - also see 6.7 on page 308; the moderately high correlation between expert and crowd - see Table 6.11 on page 310 - and moreover, a high correlation between the expert and the definitive standard of relevance; and finally, the use of a concise, precise and clear statement about how to evaluate the relevance of a document (Blanco et al., 2011).

As evaluation is increasingly valued as an integral part of system construction and as growth in construction of task sets increases, we are likely to see a continued push toward crowd-sourced relevance assessment as the basis for gold standards of relevance. In the interim, future work will seek to ratify the definitive standard of relevance adopted by this experiment.

6.6.2 *Analysis One*

Analysis one focused solely on data collected from the ranked-list interface and the document full-text view factor of two levels: integrated and modal. Table 6.22 on page 336 presents the actual outcomes predicted by hypotheses in Table 6.3 on page 302 including a summary of significance testing. Where the results correspond to the predicted outcome, the cell entries are presented in bold type; and asterisks denote the level of significance where applicable (* $p < 0.05$, ** $p < 0.01$).

Participants completing tasks with an integrated full-text view took the longest to complete, opened a significantly greater number of documents and achieved a marginally worse outcome as measured by Bookmaker score. In contrast, participants in the modal

full-text view condition were fastest to complete, achieved a marginally better Bookmaker score and opened significantly fewer documents. Furthermore, participants more readily engaged the ranked-list re-sort facility when the document full-text view was integrated.

From a high level perspective, in the modal full-text condition, searchers are less interactive and faster to complete. These data suggest that a better search outcome is afforded to those who work harder for it. However, a goal of search is to help searchers find the best information efficiently - faster and with fewer clicks; these data suggest that since users achieve a non-significant Bookmaker difference, the modal full-text view is superior as these participants are needing fewer clicks to achieve their goal.

Tab. 6.22: Dependent variable outcome observations for manipulated factors and levels in analysis one. ‘*’ denotes $p < 0.05$ ‘**’ denotes $p < 0.01$; bold type denotes observations were as predicted; bracketed values indicates no initial prediction was made

Design	Factor	Level	Time	Outcome Quality	Documents	List Re-Sort
Between	Full-Text Integration	Integrated	Slowest	Best	Most**	Most Likely
		Modal	Fastest	Worst	Least**	Less Likely

The effect of document full-text view integration was unexpectedly strong. On average participants opened double the number of documents when the full-text view was integrated and adjacent to the spatialisation visualisation. An explanation for this difference may be attributed to the mandatory overhead of closing each full-text view under the modal full-text condition. In the integrated full-text condition, this mandatory step is not necessary and participants need only shift attention from the document view and back to the theme map visualisation.

With additional documents opened by participants, a longer trial time should have eventuated; yet despite the greater number of documents opened, trial time was not significantly longer for participants in the integrated full-text condition. This may suggest that under the modal full-text view condition, participants spend longer reading documents than those under the integrated full-text view condition. One explanation for this may be that participants are more selective regarding the document candidates they open for review and in addition, more thorough in the review of full-text content. Accordingly, the modal group may have attempted to maximise the likelihood that opening a document would be beneficial. Alternatively, they may have overestimated the likelihood that a valuable piece of information might be missed. In contrast, in the integrated view, where the overhead of opening a document's full-text was lower, given that it was easier, less laborious and less interruptive, participants appeared to have adopted a more rapid relevance judgement strategy and perhaps underestimated the probability of missing a relevant piece of information.

Whilst document view time can be calculated more reliably for the modal full-text view group, a reliable estimate for the integrated full-text view group cannot be made. There is no clear signal, such as a button press event, indicating exactly when the participant's attention is diverted from the full-text view. As a consequence, there may be additional factors that are contributing to the differences in the number of documents opened across full-text view factor levels. A future experiment incorporating, for example, eye-tracking measurements could be utilised to support estimations on the proportion of time spent between integrated document full-text and the spatialisation visualisation.

While this result is interesting, there are two further aspects regarding document access that influence the reliability of this result. The first aspect relates to the time taken to open each document. In a web-based application, there is an inherent time cost associated with opening a document for relevance judgement, due to the time taken to download the file from a remote server. Even though this overhead could be made approximately constant in a future experiment, it is anticipated that a reduction in document inspection count would be more severe for the modal full-text view condition, due to the mandatory interaction cost needed to close an opened document. In addition, in this experiment, a lack of interaction cost may have artificially confounded an underlying benefit of integrated full-text view; a pathway to faster task comple-

tion via fast scanning strategies, instead of moderately paced and conscious review of document textual content, perhaps ensures that a participant's momentum or flow is maintained by a more reactive interface.

The second aspect relates to the lack of browsing tabs. In a web-based search context, a searcher can optimise some of the download time associated with opening documents. A searcher may achieve this by opening interesting documents in sequence, as they traverse down a ranked-list of search results. On cessation of the traversal, the participant is then free to review tabbed documents that by this point have downloaded either completely or substantially (Huang, Lin, and White, 2012). To investigate this further, a change to the modal full-text apparatus would see it becoming increasingly alike an Internet browser with results in one browser tab and opened documents in subsequent tabs. In contrast, modifications to the integrated document full-text interface include displaying each open result in its own tab within the integrated full-text view component adjacent to the spatialisation.

One final discussion point, relates to the differences in re-sorting behaviour for the ranked-list interface; participants could re-sort the ranked-list, by highlighting words in the full-text view. The mean number of re-sorting actions was similar across groups; although, participants more readily re-sorted at least once when document full-text view was integrated. However, statistical tests offered little evidence to suggest that this difference was significant.

Although not observed in this above analysis, a large differential between mean re-sorting actions and document full-text view type may be attributed to more thorough reading strategies that modal full-text view participants were thought to have adopted. One can imagine the situation where a participant begins reading a document and locates a useful set of words to re-sort on, enacts the re-sort action but continues on reading the document, only later to find even better words to submit as a new query. A lower average re-sort action count would also be consistent with participants whom skim documents using the integrated full-text view, since participants would be more likely to gloss over important content and therefore less likely to trigger re-sorting actions. Moreover, participants in the modal full-text view group would perhaps be more keen to find adequate keywords, in order to maximise the likelihood that relevant documents will be at the top of the ranked-list and to reduce as much interaction expenditure as possible. In contrast, participants in the integrated full-text view condition may have been more inclined to click through results in sequence, given the perception that additional interaction allowance was available.

Whilst no significant difference was indicated for participants opting to re-sort, or not, a suggested explanation for the still marked difference relates to the potential disconnect participants have between the search result set and the full-text view of a document. When in a modal full-text view, participants are focused solely on the full-text of the document; in contrast, in the integrated view, participants are focused

on the document view, however, they are able to maintain the spatialisation in their peripheral view. If a particular result or document satisfies an information need, then the searcher can rightly devote full attention to the document and forget about the remaining results. If however, the searcher is intermittently switching between the result presentation and marginally relevant documents, in order to find a small subset of quality items, then those marginally relevant documents will not be receiving a great deal of attention.

In this case, there is an opportunity to more closely tie the search result presentation with the document full-text view, such that the searcher can use those marginally relevant documents to refine the subset of results down iteratively. However, linking the full-text view with the spatialisation was beyond the scope of this experiment. A future research focus could be to apply integrated and linked full-text document view to a spatialisation-based interface and to use the text highlighting interactions, which are a form of user feedback no less, to control the layout of documents in the spatialisation.

Ultimately, there was a large quantity of information packed into the experiment session and participants were expected to operate, with competence, three different search interfaces. While every effort was made to ensure each participant had an introduction to each interface, a set of practise trials, and opportunity to clarify any confusion, it may be the case that participants simply forgot that they could re-sort the result list. It could also have been possible that some participants opted to use the initial ranking of documents only and been satisfied with working sequentially through the list. However, the unbalance between integrated and modal full-text document views is marked and there is potential scope for future experimentation regarding this design factor.

6.6.3 Analysis Two

Analysis two examined factors: dimension control, document full-text view and pop-up transparency. Table 6.23 on page 341 presents the observed outcomes predicted by Table 6.4 on page 302 including a summary of significance testing. Where no prediction was made, the actual outcome is denoted in parentheses. Where the results correspond to the predicted outcome, the cell entries are presented in bold type; and asterisks denote the level of significance where applicable (* $p < 0.05$).

With regard to pop-up transparency, when pop-up backgrounds were non-transparent, the results indicated that participants completed tasks faster, opened fewer documents, but achieved a lower Bookmaker score. In contrast, participants interacting with transparent pop-up windows were slower to complete, opened significantly more documents and achieved a better Bookmaker score. Furthermore, participants in the non-transparent pop-up condition turned off the multi pop-up facility for the largest proportion of trial time.

With regard to document full-text view, participants undertaking tasks with a modal view were faster to complete tasks, achieved a marginally better Bookmaker score, opened significantly fewer documents and on average made four additional projection dimension configurations. In contrast, participants undertaking tasks with an integrated full-text view, were slower to complete tasks, achieved a marginally lower Bookmaker score, opened significantly more documents, and made fewer projection dimension configurations.

With regard to projection dimension control, participants were marginally faster to complete tasks with the theme cloud, though achieved a lower Bookmaker score. Furthermore, participants made slightly more projection dimension configurations with the theme list, though this difference was approximately two extra rotations on average.

Tab. 6.23: Dependent variable outcome observations for manipulated factors and levels in analysis two. ‘*’ denotes $p < 0.05$; bold type denotes observations were as predicted; bracketed values indicates no initial prediction was made

Design	Factor	Level	Time	Outcome Quality	Documents	Pop-Up Time	Layout Changes
Between	Full-text	Integrated	Slowest	Worse	(Most*)	(Least)	(Most*)
		Modal	Fastest	Best	(Least*)	(Most)	(Least*)
	Pop-up	Transparent	Slowest*	Best	(Most*)	Most	(Most)
		Non-Transparent	Fastest*	Worst	(Least*)	Least	(Least)
Within	Rotation Control	Theme Cloud	Slowest	Worst	(Most)	(Least)	Least
		Theme List	Fastest	Best	(Least)	(Most)	Most

Despite devising a level of transparency, based on informal experimentation and guidance from the literature, the results have not shown a clear advantage for transparent pop-up windows. Whilst participants completing tasks with transparent pop-up windows achieved a marginally better Bookmaker score, they were significantly slower in doing so. A lower task time may be attributed to the significantly higher number of documents opened and the marginally greater number of projection dimension configurations. While this result alone does suggest that participants were more active in their use of the tool, it does not yet show that the experimental apparatus has successfully blended contemporary search behaviours such as text surrogate scanning with a spatialisation-based visualisation.

It is unclear as to why participants were more active in terms of their examination of documents and projection dimensions; a participant's insight into the benefit of pop-up transparency is a key element missing from the data set. However, one explanation is that participants opted against more calculated consideration of harder to read pop-up surrogate text and instead based relevance judgements on the document's full-text alone. In contrast, non-transparent pop-ups make it easier for participants to base relevance judgements on document surrogates and harder to navigate around the spatialisation. With a more precise selection of candidates, less time is spent dealing with irrelevant documents. Moreover, this explanation is consistent with a reduction in usage of the multi pop-up facility, since reverting to a single non-transparent pop-up window mitigates occlusion and renders navigation easier.

Further consideration of the data reveals an interaction between full-text view and pop-up transparency and moreover, the presence of an uncontrolled variable: the spatialisation's area. Spatialisation area was higher when document full-text was modal and lower when the participant's screen had to accommodate both an integrated document full-text view and a document spatialisation.

Inconsistent spatialisation width meant that in a modal full-text view, document icons were more widely spread with comparatively more pop-up windows open at any one time. In contrast, in an integrated full-text view, document icons were more tightly packed and fewer pop-up windows on display at any one time. Consequently, one would expect to see participants more readily turning off the multiple pop-up window facility when pop-ups were non-transparent and integrated document full-text; similarly, one would expect more time spent evaluating unseen documents hidden behind non-transparent pop-ups and packed into a tight space. It is reasoned that participants had no other choice but to revert to single pop-up mode in order to complete the trials as it became overly difficult to see obstructed content when pop-up backgrounds were non-transparent and the layout area small and compressed. In contrast, if the pop-up backgrounds were transparent, participants may have been less likely to revert to a single pop-up mode since obstructed information appeared through the pop-up window. In contrast again, under the modal full-text view, document icons were more

spread out thereby reducing the chance that document icons would be obstructed, this is supported by the observation that under a modal full-text view, pop-up transparency made little difference.

Turning now to the projection dimension control factor, based on data from the subjective response questionnaire, participants indicated that they were more successful, experienced less frustration, required less time and required less physical workload when using the theme list. Moreover, while not recipient to statistical support, the objective measures showed that participants completed tasks faster and obtained a higher Bookmaker score, when completing tasks with the theme list. Accordingly, additional research is required to determine a superior projection dimension control; toward this, further design considerations arising from observations in the data are discussed here for future experimentation.

Rotated words in the theme cloud stage are consistent with the naturalness hypothesis presented in Chapter 4; however the use of rotated text is subject to some cautionary research. Byrne (2002) earlier found that reading rotated text was slower. He found that the direction of rotation does not make a difference; however, presenting labels in a marquee has a negative impact on reading of text - in marquee text, text is vertically oriented and each character appears below the next. In light of Byrne's findings, participants reading a box of horizontally oriented descriptor words should be faster at gaining an understanding of the selected dimension, compared to reading those descriptor words vertically oriented. Theme lists solve this problem by presenting all text horizontally, however the theme list control requires double the number of visualisation components in comparison to the theme cloud control. Furthermore, the theme list carries no explicit graphical cues beyond textual labels that demarcate the box containing vertical or horizontal themes and descriptors.

Moreover, from a design perspective, the theme cloud presents the same information as four theme lists do, in a more compact way and offer an additional visual cue regarding the association of concept descriptors with projection axes. On the other hand, under the theme list condition, participants may have anchored on either the horizontal or the vertical dimension, having identified a potentially interesting dimension and then rotating through alternative dimensions in order to find a good layout of documents. A tendency to anchor on a vertical dimension could suggest that participants had a preference for horizontal descriptors.

Evidence for anchoring on a particular dimension is available to support this. Regardless of projection dimension control, participants made more configurations of the vertical axis ($M=5.38$ configurations, $SD=6.37$) than they did for the horizontal axis ($M=3.90$ configurations). A paired-samples t-test indicated this difference was significant: $t(101)=-2.93$, $p=0.004$. This effect was even more prominent in the theme list condition in which vertical axis configurations were higher ($M=6.27$ configurations, $SD=7.64$) than horizontal axis configurations ($M=3.86$ configurations, $SD=2.89$). A

paired t-test indicated that this effect was significant for theme list: $t(50)=-2.45$, $p=0.01$. For the theme cloud condition the difference was less prominent though vertical axis configurations still higher ($M=4.9$ configurations, $SD=4.69$) than horizontal axis configurations ($M=3.50$ configurations, $SD=4.73$). A paired t-test indicated that this difference was not significant $t(50)=-1.63$, $p=0.10$.

This is not necessarily a bad thing and can be attributed to the layout of the interface. Regardless of experiment treatment groups, participants followed the same procedure to rotate axes as every other participant did. In the theme cloud condition, participants clicked on the left side of the word to select the concept for the vertical axis, while in the theme list, participants used the left most list box to control the vertical axis. These design choices were purposely fixed rather than randomised to facilitate easier learning of the interface. By convention, the left edge of a graph is the location where the y-axis or vertical axis appears and the right edge of a graph is the furthest point on the x-axis or horizontal axis away from the y-axis.

With a propensity to anchor on one particular projection controller, in addition to the legibility of oriented text, brought to the forefront by the legibility of pop-up surrogates under different levels of transparency, in addition to the favourable subjective and objective data for the theme list, future experimentation should consider a theme list control placed at the left side of the interface.

6.7 Summary

The results of this experiment have hinted at the superior of two alternative designs for each of three user interface components installed within a document search tool. Whilst these results are tentative and should be supported with follow up experimental work, with a strong focus on collecting greater insights from the user with regard to pop-up transparency and even document full-text, the results are indicating several themes:

Multiple Pop-up Facility

If a multiple pop-up facility is provided to searchers, ensure a toggle button to revert to single pop-up mode is provided. Utilise a pop-up background transparency well below 50% as anything above may influence readability and ultimately search outcome.

Document Full-text View Integration

Careful consideration must be devoted to the integration of full-text document view as integration leads to more rapid document revision behaviour in participants; however, the additional screen real estate devoted to full-text integration may influence point density in the result visualisation and consequently search outcome.

Spatialisation Projection Dimension Control

Favour a theme list style control over a theme cloud control to support exploration of a multi-dimensional information space, and consider placing this control at the left side of the screen.

7. DISCUSSION AND CONCLUSIONS

7.1 *Introduction*

Searchers need better tools to satisfy complex information needs; the design of such tools must be based on user-centred research. This thesis has reported three human-based experiments to motivate three facets of such tools: natural encoding, motion frequency data encoding, and document spatialisation interface components. Broadly, the experimental hypotheses were centred on finding optimal data presentation in search tool interfaces.

This chapter will unify key experimental observations into a single perspective of search result visualisation tool design; doing so will highlight the significance of this program of research and will motivate the continuation of research in the area. This chapter will begin with a brief overview of each experiment, including the observed outcomes, then, discussion shows how in a holistic sense these experiments relate to and complement each other. Later, discussion will focus on the significance of these outcomes, will contrast the methodological differences between the online and lab-based experiments and will propose a set of recommendations for future work. Finally, a proposal is made for a conceptual organisation of the aspects of search tool design that were highlighted over the course of this thesis. This organisation unifies each chapter's main themes and concepts; in addition, it contextualises the contributions of the experimental work, and furthermore, it will provide a general overview of the design of search tools that feature a document spatialisation. It is envisioned that such a unified view will benefit future researchers in the area.

This chapter will begin with a recap of experimental work that focused on the way individual results are represented.

7.2 *On the Role of Motion and Natural Encoding in Attribute Visualisation*

The experimental observations arising from the motion and naturalness of encoding experiments - Chapter 3 and Chapter 4 - apply predominantly to the representation of individual search results in a search result interface; specifically, this research focused on the point, or glyph. A set of homogeneous points or glyphs, assigned positions in Cartesian space, implicitly convey relationships based on point density and proximity.

However, there are a number of geometric and appearance attributes of the point that can be manipulated to encode document metadata. Encoding data in graphical attributes involves at least two decisions: what graphical features should be used and which data features should be encoded by which graphical feature.

The graphical encoding palette is broad; there are a number of appearance and geometric attributes that can be used to encode data. However and predominantly, data encoding paradigms in information visualisation have made almost exclusive use of static attributes to encode data. In contrast, motion attributes do not readily appear in data encoding paradigms.

It was shown that earlier investigations have revealed that motion can be used effectively in visualisation-based interfaces, though typically, motion is used to visually bias or highlight aspects of the data set and to foster more efficient processing of an information visualisation. Resoundingly, motion frequency - in contrast to motion phase - is reported to be a poor graphical device to support this type of highlighting facility.

The motion experiment of Chapter 3 investigated the role of motion frequency in data encoding paradigms and was motivated by the observation that while motion frequency may not be suited to grouping related information over a whole interface, human factors literature supports the notion that the human visual system has a propensity to perceive different motion frequencies. Furthermore, it was argued that when interacting with a document spatialisation, the searcher's task requires first narrowing search activity to a subset of search result data, based on spatial information, before further narrowing search to targets meeting criteria specified in appearance or other geometric attributes. Having restricted their attention to a local area of a document spatialisation, the searcher's task is reduced to the comparison of a few alternatives, and checking each alternative to determine if particular search criteria are met.

The second aspect that was noted as relevant for encoding data at the point, related to the mapping of a graphical feature with an appropriate data feature - the set of mappings referred to as the encoding paradigm. It was argued that encoding paradigms should take into consideration the naturalness between an encoding graphical feature and the encoded data feature, in order to produce more natural to use interfaces. Interfaces that exhibit high naturalness are those where the encoding of data meet the prior beliefs and expectations of the user, such that the user does not have to expend explicit conscious effort to interpret and utilise the encoding paradigm for a data extraction task.

Whilst a similar notion of display naturalness appears within disciplines at the fringe of the information visualisation field, such as in pictorial representation and diagram design, the information visualisation discipline has no guidance on how to utilise naturalness in data encoding paradigms. Furthermore, no known, previous and

empirically supported research, wholly recommends the user's expectations and prior beliefs to guide data encoding rule production. Conversely, as reviewed in Chapter 2 and Chapter 4, there are encoding guidelines based on aspects like the type of data or the searcher's perceptual processing capabilities that are often the basis for encoding rule sets. However, depending on task requirements, such as accuracy, speed, and the type of extraction tasks - e.g. visual estimation, pattern recognition or counting tasks - a data set's encoding rules, based on guidelines from say data visualisation (Mackinlay, 1986; Nowell, 1997) are not necessarily the same encoding rules that would be devised from insights garnered from visual search experiments (Wolfe, 2007, e.g.).

The results of Chapter 4 indicated that choosing an encoding rule that is natural and consistent with the user's expectations and prior beliefs, may reduce the degree of interaction with an interface if the user can utilise graphical information more readily than an explicit numerical value presented in a pop-up window. Empirical results indicated that interactive behaviours measured as the number of pop-up windows triggered were on average lower when data was encoded naturally and that the variation between participants was less so than that of participants attempting tasks with an unnatural encoding. Nevertheless, statistical tests did not yield sufficient evidence to suggest that this difference was significant.

Moreover and again, while not flagged as statistically significant, participants in the natural condition were marginally more successful at answering tasks. In particular, participants were markedly more successful answering cardinality based questions when cardinality was encoded to icon size. This finding suggests that a user interface may be easier to learn and use if the data are presented in a way that is natural for the user.

Furthermore, no convincing difference was observed for self-reported learning but, on face value, the results suggest that participants in the unnatural encoding condition were slightly more accurate at recalling how data was encoded in the interface. If this difference were more apparent, a degraded self-reported learning outcome finding may still yet be interesting; participants who do not have their prior expectations and beliefs violated should not have to exercise comparatively more explicit and conscious effort to learn an encoding paradigm, in order to use a data visualisation. In contrast, participants in the unnatural encoding condition, in order to utilise a data visualisation tool, must spend additional time learning the encoding mappings in use and furthermore on an ongoing basis, spend additional effort overcoming mappings that confuse decoding of the visual data.

The naturalness of encoding experiment investigated how to choose graphical features that are most suited to the encoding of specific metadata features; in contrast, the motion frequency experiment investigated how to effectively double the size of the graphical feature palette that could be selected to encode data attributes. Clearly stated, the results indicated resoundingly that motion frequency is a poor choice for encoding data, as earlier studies have also hinted.

However, this investigation remains a valid and unique indication that motion frequency may not suit a result visualisation application, where, in addition to tasks which utilise motion frequency to link spatially disparate information, the search task necessitates consideration of motion encoded attributes within a local region of a spatialisation. Thus, this experiment sought to show how quickly and accurately we perceive motion frequency to extract data from, say a pulsing or flashing shape, in comparison to extraction of data from say a coloured shape.

Overall, time and errors increased with increasing trial complexity - i.e. relative to the number of graphical features defining the target icon. Time and error increases were approximately linear and occurred regardless of whether the encoding involved static or motion feature attributes. However and additionally, with increasing motion feature count, the average time and error rate increased consistently and more rapidly.

When icon complexity was greater, the average additional overhead was greater also, though if that additional feature was a motion feature, the severity of that additional overhead was significantly greater than compared to if that additional feature was static. Overall, on average an additional static dimension used to encode data adds 2.62 seconds to answer time; in contrast, an additional motion feature used to encode data necessitates 4.79 extra seconds to answer time. Furthermore, while an additional 1.24 seconds is added to preparation time with every additional static feature, an additional 7.97 seconds is added to preparation time with every additional motion feature.

7.3 *On the Role of Space and Interface*

The third phase of experimental work focused primarily on the space and the interface of search result visualisation. An experimental apparatus presented a set of search results using a document spatialisation; a searcher interacted with the spatialisation, in order to locate documents that satisfied their information need. There were significant development overheads devoted to the preparation for this investigation and this testifies to the enormity of potential scope of search result interface tools - and their evaluation. Overheads included selecting an appropriate compression and projection approach for building document spatialisations, choosing and preparing an appropriate test corpus, establishing a definitive relevance model for a set of three search tasks, actually building the interactive search tool, preparing participant training material, recruiting and conducting the evaluation and finally, analysing the collected data.

The first half of the investigation dealt with the development of the document spatialisation; this process was the subject of Chapter 5. An objective and qualitative assessment of several approaches to building spatialised search result sets was conducted. The spatialisation provided the basis of a tool; however, in isolation, a document spatialisation does not necessarily lead to a better search outcome. Interface

augmentations are necessary to make spatialisations more usable and useful; three interface components were evaluated and were the subject of Chapter 6. However, unlike the glyph experiments, this experiment took place in a laboratory setting.

The main outcome of evaluating several spatialisation approaches was the selection of a set of compression and projection algorithms that could take as input a collection of text documents and provide as output a spatial representation of those documents. A single arrangement of the document set was not desirable, since documents can share multiple similarities. Accordingly, a spatialisation technique was required that showed off different aspects of these semantic interactions, depending on the perspective from which the participant observed the document space.

There were two phases to spatialisation construction: compression and projection. In the compression stage, the highly dimensional document set - i.e. each unique word in the set a dimension - is reduced to a smaller set of representative dimensions or themes - i.e. a combination of words forming a theme - and each document is assigned an association to each theme. Then, in the projection stage, similarities between documents in the theme space are re-calculated in order to present them in a rotatable 2D visualisation, representing a further compression of the theme set.

In addition to an algorithmic approach that illuminates semantic relationships between documents, it was argued that a spatialisation construction approach should also meet a set of five qualitative criteria. These criteria are likely to represent the minimum requirements for a spatialised document set if such a technique is to be useful for a searcher. The first criteria dictate the need for coordinates - so as to depict relationships through spatial perceptions - as well as the means - i.e. textual labels - to convey the semantic meaning of spatial region. Furthermore, in order to support changes to the searcher's perspective of information space, further criteria dictate the need for labelling of projection dimensions, so as to facilitate informed changes to the searcher's perspective of information space.

The final outcome of spatialisation included a set of labelled projection dimensions, a set of coordinates for each document along each projection dimension, and a set of coordinates for each region label along each projection dimension. These components were then installed into an experimental apparatus to test several interface designs. A series of interface designs were considered and discussed and a collection of empirical data provided some evidence in favour of the superior designs.

Three interface designs were expected to influence a searcher's interactions with a spatialised set of search results. These design alternatives included document full-text view integration, pop-up window transparency and projection dimension selection controls. Historically, each design alternative has seldom been the focus of search result interface research, due possibly in part to the overly literal application of information visualisation to search result interfaces. For example, information visualisations have

typically made use of single pop-up windows to convey information regarding individual points in the space, which is consistent with the details-on-demand aspect of ‘overview first, zoom and filter then details-on-demand’, also known as the *visual information seeking mantra* (Shneiderman, 1996). However, details-on-demand is not consistent with the way we search for documents as searchers engage in liberal amounts of rapid text scanning during search in combination with a ranked list of search results.

Yet, maximising the number of pop-up windows visible at any one time, introduces an additional problem in that many pop-up windows then obscure the underlying spatial information. Accordingly, the first design choice evaluated the influence of pop-up transparency on search behaviour, in order to ascertain a compromise between pop-up window content in the foreground and insight from the spatial information in the background. Semi-transparent windows should have revealed obscured semantic landmarks i.e. textual labels indicating semantic meaning of regions, as well as other document icons.

Ultimately, obstructed document icons influence the searcher’s discovery of both irrelevant and relevant documents; yet, the cost of obstructing a relevant document icon is far worse. Accordingly, the transparency hypothesis intended to demonstrate that by improving the chances that the searcher would find relevant documents, the quality metrics on the participant’s answer set would improve. The results were complicated by the observation that participants opened significantly more documents when the pop-up was transparent, and while answer set quality was higher, participants spent longer completing tasks.

An explanation for this finding was that the legibility of the foreground text might have been affected by the background content. While efforts were made to ensure a balance between foreground and background legibility, the level of transparency may not have been sufficiently opaque, thereby making it harder to read document surrogate content in pop-up windows. In contrast, under non-transparent background conditions, it may have been more challenging to find and look at a greater number of documents even though it may have been easier to make a more reliable judgement of the pop-up textual content.

It may be the case that the frequency and impact of obstructed document icons and semantic labels was overestimated, whilst the impact of poor foreground legibility underestimated. Furthermore, the reliance that searchers have for spatial cues and landmarks for the specific experimental task sets may also have been overestimated. This view does not advocate that searchers do not benefit from this information; however, it suggests that the few spatial cues available to searchers could have been processed in alternative ways or very quickly during transitions between perspectives.

During a projection dimension rotation, document icons and semantic landmarks smoothly transitioned to a new coordinate location. Pop-up windows however, were

removed from view during this transition and only re-appeared after a 500ms delay, and after the finalisation of the rotation. The final frames of transition may have been sufficient to prime searchers with a basic idea of the density of points in a particular spatial region and a semantic interpretation of that particular spatial region based on semantic landmarks that were likewise transitioning to the same area.

Another interpretation is that pop-up window content may have provided enough information as to define the semantic region in place of semantic landmarks. Potentially, an initial first pass of pop-up window content across different areas might have been sufficient to indicate where the semantically relevant areas were and which were irrelevant. In this light, if the legibility of document surrogates was such that they could not be processed easily, due to difficulties in perceiving keywords in the text, then performance could have suffered as a result.

This study, in addition to a lack of power, is limited by the little subjective data collected, specifically targeted at the participant's opinion of the utility and usability of pop-up window transparency. Future research should seek to isolate what participants think of the legibility of the pop-up text and whether they find themselves relying on the semantic landmarks, if at all, or alternatively to see if pop-up windows could be used as a primary source of spatial annotation in addition to document surrogate information. If pop-up windows were acting as auxiliary semantic landmarks, then future research should also seek to improve the information scent of the pop-up window and research does provide some indication regarding how to achieve this through text mark-up (Aula, 2004).

Nevertheless, this does not account for the obstruction of document icons below pop-up windows. Consequently, even slight window transparency may be necessary to reveal obstructed icons - potentially only icons that have not yet been visited - to appear at some low level of fidelity through the pop-up window. Moreover, this does not negate the need for semantic labels altogether; in fact, semantic labels could themselves be pop-up windows, though visually distinguished. Semantic label pop-ups could include a set of words that more precisely describe the region of interest in contrast with the single words and phrases that were present in the current experiment. Utilising pop-up windows for semantic annotation such as that described here could be quite easily implemented in the current apparatus.

Two interface designs facilitated access to the full-text of individual documents in the result visualisation: modal and integrated; document full-text view was the second interface design choice under examination. For participants attempting tasks with an integrated document full-text view, the document full-text appeared in a frame adjacent to the spatialisation visualisation. In contrast, for participants completing tasks with a modal document full-text view, the document full-text appeared in a frame that replaced the document spatialisation totally for the duration of the full-text view; to return the spatialisation, the participant had to close the full-text view window by

clicking a button.

A modal full-text view is likened to that which participants experience when using a browser-based search engine, in that documents are opened in new tabs, windows or even replace the search result tab. In order to return to the result list, searchers have to click the back button or return to the dedicated search result page window or tab.

The results indicated that full-text view has an interesting effect on search behaviour. There were significantly more documents opened under the integrated full-text interface configuration. However, the reduction in effort expended to look at an integrated document full-text does not necessarily bring about an improvement to quality of search. With additional effort available, with savings gained from a lack of need for explicit button clicking to close document full-text, the searcher should have engaged in better processing of each document. Although lacking statistical support, time and Bookmaker score were worse under the integrated full-text view condition. This was likely due to the significantly higher number of documents opened.

Furthermore, despite no significant differences for task time for both full-text view levels, the extrapolated time spent reading individual documents - a ratio of task time to document count - under the integrated full-text view was far shorter than the modal full-text view. However, given that there was no explicit event to determine reliably when participants under integrated full-text view stopped reading and resumed searching, this extrapolation cannot be confirmed. In contrast, the total time spent reading under the modal full-text view might be determined reliably as the time difference between events signalling the open and close of a full-text view.

Under the integrated full-text condition, given the low overheads to open a document for full-text view, participants may have adopted an interaction strategy that favoured opening documents without careful consideration of the document pop-up content. This could explain the rapid and fleeting document view behaviour evident in the group subjected to the integrated full-text view. Since the modal full-text view condition involved greater interaction overhead, this strategy may not have been applicable given the additional interaction and reorienting overheads involved with closing the document view.

While strategic reasons go some way to explain the differences between the full-text view configurations, a problem with the experimental design may also be influencing behaviour under different full-text view conditions. Under the modal full-text view condition, the spatialisation area was greater, resulting in a greater spread and spacing of document icons. In contrast, under the integrated full-text view condition, the spatialisation area was smaller due to a restricted panel width, resulting in a compression of document icon layout. While the data provide a contrast between large and small spatialisation sizes, it reduces the power of the conclusions that can be made. The compression of icon layout may likely have influenced the density of icons, and therefore

the number of pop-up windows open and the degree of obstructed spatial cues; these all may have had an influence on the behavioural measures. However, compression of icon layout may not necessarily be a bad thing, in that the spatialisation area is within a few degrees of visual angle saving the need for head movement in order to see the whole visualisation.

Furthermore, a compressed layout could necessitate the use of an occlusion strategy and this was indicated by the multiple pop-up toggle state. On average, participants toggled the multiple pop-up facility off for longer periods, under the non-transparent condition in contrast to the integrated full-text view. However, this difference was not large enough to reach statistical significance. Interestingly, this was so more markedly apparent for the integrated document full-text view and non-transparent pop-up group. Under modal full-text view, this trend was not as apparent for either level of pop-up window transparency. Nevertheless, these results motivate the provision of a mode switching facility between single and multiple pop-ups in future search tools, particularly, if the area of a spatialisation view is limited.

In relation to the projection dimension control, no strong conclusions can be drawn from the results. Though on raw values alone, participants were more interactive with the theme list control, time to complete tasks was faster and overall answer quality measured by Bookmaker score was marginally better. Furthermore, the subjective ratings did not favour the theme cloud; participants indicated that they felt more comfortable and confident using the theme list, than they did using the theme cloud. Moreover, the literature on the readability of vertically oriented text strings adds an additional negative to the theme cloud control.

The results altogether indicated some significant behavioural differences but no strong differences in measures of search quality. Foremost, search is about finding documents quickly and that satisfy information need. Optimally satisfying need means that only the most relevant documents are utilised to satisfy that need. In the experiment of Chapter 6, there were no design components that significantly influenced Bookmaker score. Whilst asking participants to actually write the fictitious essay they were requested to research for in the experiment, may be the only optimal way to judge outcome under different search tool configurations, this would introduce additional between subjects variation that would require consideration in future experimental design. Furthermore, participants were requested to rate the relevance of a document on a four-point relevance scale with one point devoted to irrelevant and three points devoted to three degrees of relevance. A further improvement to the quality of outcome measure would be to ask participants to highlight important sentences in documents they rate as relevant. However, an initial priority for future evaluation would be to double the size of the participant pool before consideration of changes to the experiment's design.

7.4 *A Holistic Perspective on Search Result Visualisation*

Each chapter presents a variety of research, and altogether they present a roughly complete picture of search result visualisation tools that feature document spatialisation. This picture shows that search result visualisation tools require significant development overheads including the sourcing of search results, constructing search result visualisation and document spatialisations, implementing document full-text views and document surrogate views, providing interactive controls and facilities, implementing and evaluating language models, creating encoding legends, and coding and extracting document metadata to name but a few. Furthermore, the goal-driven, perceptual, cognitive and environmental influences of the anticipated human searchers must be taken into consideration throughout the design and implementation process.

Each chapter has discussed many of these aspects and now are collated here in the beginnings of a taxonomy of components, in the hope that it will make it easier for future designers and researchers throughout implementation of future research tools.

Furthermore, this collation provides an appropriate way to contextualise the contributions of the present research outcomes. Figure 7.1 on the facing page depicts a visual representation of the collation. Foremost, literature and concepts are collated according to their relevance to the point, the space or the interface of search result visualisation. However, due to the nature of search, several peripheral issues are included for completeness. Like the survey of systems that featured in Chapter 2, this organisation should serve as an introductory summary of search result visualisations that incorporate document spatialisation for a new designer or researcher in the area.

The grey box at top left of Figure 7.1 on the next page lists a series of features driven primarily by the back end and supporting operations. These will not be discussed further; they are simply listed as alternative pathways for sourcing a collection of search results, or they are an additional augmentation of the interface that may facilitate search in some way.

7.4.1 *The Point: Representing an Individual Search Result*

The point constitutes how one represents an individual document in a search result visualisation; though by extension, attributes of the point could also be used to represent individual collections or sets of points such as clusters - as was featured in the naturalness of encoding experiment in Chapter 4.

The potential scope of encoding, involving metadata and geometric and appearance attributes of the point, is broad. Choosing the most appropriate combination of metadata and graphical attributes, necessitates a set of encoding rules that define an individual or set of graphical attributes that are appropriate for encoding an individual or set of metadata attributes. While there are multiple partial sources of guidelines that

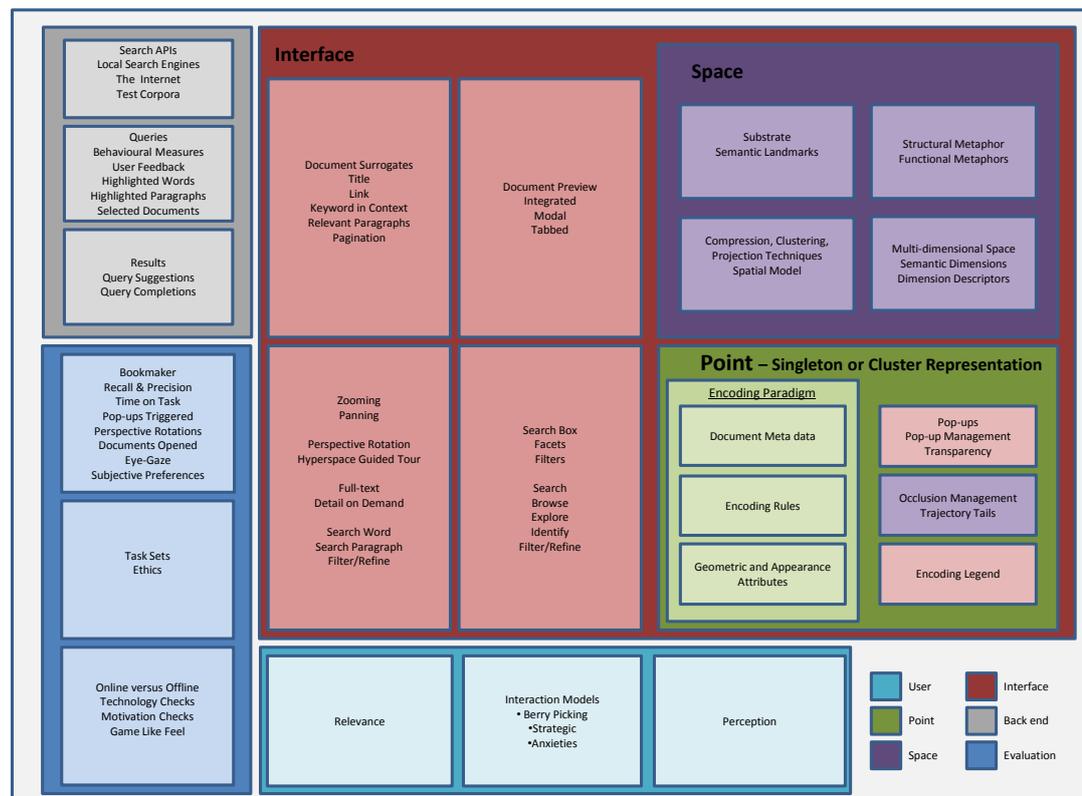


Fig. 7.1: The Point, Space and Interface of Search Result Visualisation presented in this thesis.

recommend the selection of one encoding combination over another, the naturalness of encoding experiment has offered preliminary evidence to suggest that an encoding rule should maximise naturalness between the graphical attribute and the data attribute such that the encoding is consistent with the prior expectations and beliefs of the user.

The eventual set of encoding rules forms the encoding paradigm of the visualisation, which should be displayed in a legend adjacent to the visualisation. Legends provide a look up to allow a user to ascertain to which data attribute the graphical attribute refers. While legend design has not received significant research attention outside of the study of cartographic visualisation, principles of good legend design can be adopted from fields peripheral to information visualisation.

Metadata can also take the form of textual data, which does not lend itself to simple or intuitive data encoding using graphical attributes of a point. In the motion and naturalness of encoding experiments, participants could access textual metadata such as a document's title, keyword in context or cluster keywords. This was made possible through pop-up windows that triggered on display when participants passed their mouse cursor over document icons. In the Space and Interface experiment Chapter 6, instead of one pop-up appearing at a time, as many pop-up windows that could be displayed without overlapping were displayed. In contrast, in Chapter 3, only one pop-up window was made available. A multi-pop-up facility involves additional overheads

and management routines that ensure pop-up windows do not overlap with each other.

However, with increasingly large numbers of pop-up windows available the chances of pop-up windows obstructing content below, increases. Window transparency is one way to deal with this problem; however, the results of Chapter 6 indicated that display density may determine whether this transparency level is or is not necessary.

7.4.2 *The Space: Representing Semantic Relationships*

The Space refers to the coordinate system that provides the substrate upon which documents are assigned positions, according to the spatialisation algorithm. This space may be multi-dimensional necessitating additional overheads necessary to support interaction and exploration of this space.

Building a document spatialisation involves a set of compression and projection algorithms that reduce the very highly dimensional documents down into a few representative dimensions that when used as coordinate axes for documents, reflect the thematic relationships between subsets of documents. Currently, no robust candidate for the algorithmic approach is obvious and even robust objective assessment methodology remains un-established. However, a combined qualitative and objective measurement methodology as was pursued in Chapter 5 may be appropriate. In terms of an approach to building multi-dimensional information spaces that both performs well according to objective measures and importantly, satisfies a series of qualitative measures, a combination of Latent Dirichlet Allocation and Correspondence Analysis may be used effectively.

If there are more representative dimensions than screen dimensions, the searcher must be provided with a way to change their view perspective of the document space. Searchers should be provisioned with semantic descriptors for each dimension, to facilitate informed choices regarding the next projection dimensions to select for viewing.

Additionally, a set of homogeneous points in a spatialisation do not offer any semantic interpretation other than the idea that related groups may be present. The space requires the provision of semantic landmarks or labels as a minimum in order to support this understanding and ultimately guide a user to and between regions of interest. Semantic landmarks need not necessarily be textual in nature in order to aid in the return of a searcher to an earlier location in the information space. Graphical landmarks evident in spatial metaphors may also provide suitable information to support exploration of the space.

Documents arranged on a labelled spatial substrate are just a picture however. Therefore, interactive capabilities are necessary to permit a searcher to interact with documents of interest and to manipulate the arrangement of the layout in order to refine a document set down into a manageable size.

7.4.3 *The Interface: Facilitating Interaction with Information Space*

The interface encapsulates both the space and the point. The interface provides the facilities with which searchers can enact a change to their view of the document space and gain further insight into individual documents.

Whilst encoding paradigms impact on the representation of individual search results, the accompanying encoding legend is mentioned here under the interface heading. Foremost, encoding legends provide a visual dictionary (Riche, B. Lee, and Plaisant, 2010) permitting a searcher a way to look up the relationship between geometric and appearance attributes of glyphs and document data features. However, as prior research indicates (Tudoreanu and Hart, 2004; Riche, B. Lee, and Plaisant, 2010), encoding legends may also double as interactive controls to enact modifications in a visualisation. Further to this, encoding legends should themselves change in response to the application of filters to the visualisation (Dykes, Wood, and Slingsby, 2010). Encoding data into appearance and geometric attributes of glyphs facilitate some forms of document search, though often we choose to focus our search by reviewing short textual summaries.

In many visualisation-based interfaces - such as TouchGraph <http://www.touchgraph.com>, Kartoo (see Koshman, 2006), and Grokker (see Rivadeneira and Bederson, 2003) - obtaining information about individual results is facilitated by a details-on-demand pop-up window. However, this is not consistent with text scanning behaviours adopted by searchers using contemporary web-based search engines. A blend of contemporary search behaviours and information visualisation may be possible through use of multiple pop-up windows displaying textual content and superimposed over non-textual, graphical information. However, multiple pop-up windows introduce further issues such as occlusion and this necessitates careful management of where and how multiple pop-up windows appear and whether pop-up windows should reveal obstructed content below.

Pop-up content only conveys a keyhole view of a document; a document full-text view offers the complete view of a document to the searcher. Critically, the way that this view is integrated into the search tool may have a significant influence on the strategy and behaviour of searchers. However, while currently there is a basic connection between full text and search engine (Baird and Zollinger, 2007) future search interfaces will likely provision additional interactive capacities to the full-text view to enable searchers to enact changes to the display of search results. Arguably, interaction with both the search result set and a document full-text view will be superior, if following a full-text view, effort expended reorienting to and with the search result set, is lower.

Nevertheless, document full-text integration introduces its own influences to an interface such as the impact of screen area - some devoted to the search result visualisation and some devoted to the full-text view - which in turn influences the density of presentation, which in turn necessitates careful consideration of the design of pop-up

window components.

The interface also includes the interaction controls that enable searchers to manipulate the layout and display of documents in a spatialisation. The main interactive facility investigated in this research was the projection dimension control component. Effectively this user interface filter moved documents in and out of view according to the semantic content of documents. This control relied entirely on the existence of multiple thematic dimensions generated by the compression and projection algorithms discussed in Chapter 5.

Lastly, there are a number of interface components and interactive facilities that were not considered in this research, but are included briefly for completeness. Prominent components include those under the interface enhancement category, such as query input facilities, query suggestions, query completion, and result pagination (e.g. Haas et al., 2011; Sandvig and Bajwa, 2011); furthermore, traditional faceted browsing interfaces despite their popularity, have not received great consideration in this work (e.g. Hearst, 2006; Polowinski, 2009). In addition, and specific to document spatialisation, interactive facilities such as zooming and panning support, whether semantic or topological (e.g. Hornbæk, Bederson, and Plaisant, 2002) have not been considered even though prominent in many system evaluations.

7.4.4 *The User*

The user will play a key role in the future of search result interface design. In all of the experimental work presented in this research, the user has featured centrally in the evaluation of search interface designs.

Driving the user's activity is the human perceptual and cognitive system. The Visual Expression Process of Rodrigues et al. (2007) provides a breakdown of perception of visual information in three stages: conception in which raw visual features are perceived; then observation in which raw visual features form perceptual Gestalts; and finally reasoning in which reasoning decides whether the emergent perceptions are practicable and useful for further, higher level cognitive processing and interpretation.

An understanding of this or a similarly descriptive breakdown of the human perceptual system provides a basis from which to tune and motivate changes to search interfaces. For instance, Skupin and Fabrikant (2003) provide validation for the notion that perceptual Gestalts lead to interpretations consistent with structural relationships between documents. Spatial distance is naturally understood to represent similarity, clustered points understood to indicate groups of related items, and furthermore explicit connections or edges between nodes strongly indicate the presence of a relationship. Consequently, a sound understanding of the conditions under which these perceptual interpretations will be made is important for tool design.

However, as the results of the motion experiment have revealed, simply because we can perceive motion frequency well - as was supported in the opening review of literature - the use of motion frequency in an interface does not necessarily result in optimal interface design. Thus, ongoing validation of the psychological and human factors literature applied to real world contexts is required. For instance, future evaluation should consider a wider number of cognitive dimensions such as perceptual comfort aspects and the degree of naturalness in information presentation, which was proposed in the naturalness of encoding experiment.

Deeper insight into the cognitive processing of the user in search interface usage and more broadly in search tasks has been comparatively shallow in the experimental work in this thesis. For instance, the main tools that were discussed or examined in this thesis included models of information seeking behaviour, which only hint at the cognitive processes occurring during interaction with search interfaces. Potentially, greater insight into interface components and information scent and their influence on progression through the information-seeking process is likely to motivate greater insight into how search interfaces should evolve. Furthermore, greater insight into assessing quality of outcome and understanding relevance could come about by requesting users to provide supporting evidence such as highlighting key sentences in documents in addition to their relevance rating scores.

Several examples in the literature testify to the notion that human behaviour can be used to motivate research and technical development (e.g. Bates, 1979; Bartram, 2001). Furthermore, there are examples where traditional mathematical approaches to information retrieval do not operate in the way that a human would operate given the same task and sufficient time (e.g. Polowinski, 2009; Russell et al., 2006) and we cannot yet build interfaces that replicate the same perceptual and cognitive experiences present in reality (Cockburn and McKenzie, 2001). Accordingly, more attention should be devoted to understanding how users approach search tasks and more generally human computer interfaces in order to build better search interfaces.

7.4.5 *The Evaluation*

Evaluation is central to ascertaining the superiority of one tool over another. However, methodological aspects of evaluation are extremely diverse. The experiments outlined in this thesis provide a good contrast of the type of experimental methodology that is possible when conducting research online and over the Internet, or within closed-door laboratory sessions.

The point-based experiments share an additional common thread in that both were conducted over the Internet and both offered no financial reward for participation. For this research, conducting online experiments originated out of a need to make efficient use of scarce resources in order to conduct small-scale research projects and to test ideas.

However, conducting research over the Internet introduces technological, experimental and ethical issues that the researcher could otherwise exercise sufficient control over, in an equivalent laboratory-based experiment.

Among the most significant methodological observations taken from the online experiments was the prevalence of drop out. Approximately half of all participants that started either experiment, did not complete the experiment. This immediately halved the quantity of data for analysis and consequently affected the power of the experiments and conclusions.

Both online experiments - Chapter 3 and Chapter 4 - did not offer a financial reward for participation. In contrast, the space and interface experiment - Chapter 6 - offered a monetary reward, which resulted in faster recruitment and data collection - and no incidence of drop out. In lieu of tangible reward, in exchange for participation, the online experiments had to offer alternative ways to motivate participants to complete the experiment in full.

The motion experiment attempted to promote a game-like feel, by offering participants a performance report at the conclusion of the experiment. The report provided an average performance score for all trials and difficulty levels and showed how participants performed relative to the rest of the participant pool. Furthermore, an overall score as a function of time, accuracy and difficulty level placed each participant in a top scores table in order to motivate a sense of competitiveness. In order to measure the game-like effect, it was expected that the experiment apparatus would capture a number of repeat data captures from the same participant. However, the results indicated very little repeat participation.

Subsequent consideration of the experiment design has indicated further steps to promote a game-like feel, such as making the top scores table more prominent on the experiment landing page and opting for a more puzzle-like game format similar to mobile phone or social media applications that provide raw entertainment more than anything - and not necessarily an obvious information retrieval experiment dressed up as a game. Additional elements to further a game-like feel include a cumulative or per-trial score on the experiment apparatus and to anthropomorphise the otherwise generic icons into small characters or aliens. For instance, the encoding legend and two-stage trial design could be preserved in a redesign by using the legend to encode body parts of the creatures that the participant puts together in order to form a target creature.

The naturalness of encoding experiment offered no game-like feel, though it framed the importance of the experiment in terms that should have appealed to Internet search engine users, and to a level comparable with the motion experiment. With little other source of reward, this experiment achieved a similar level of drop out. Furthermore, having achieved a level of drop out in both experiments that was as expected per the claim of Reips (2002), the additional measures taken by the motion experiment appear

to have been ineffective - very few participants utilised the personalised performance report.

However, the drop out analyses reveal further insight into where and perhaps why drop out occurred. Interestingly, the majority of incomplete participations occurred at the training stages and later during the first few experiment trials. Likely explanations for these observations include the instructions being overly long and intense, the experiment task particularly daunting, or an estimated effort to be expended too large. Additionally, reasons as to why drop out may occur after the first few trials include the participant losing interest, not understanding what to do, or an external interruption. Furthermore, two participants at the end of the naturalness of encoding experiment did not complete the exit survey, suggesting that an overly long questionnaire should be avoided in order to ensure participants do not extinguish their altruistic allowance right at the last stages of the experiment. Lastly, technical reasons may also account for why participants drop out before answering one question. Future experiment designs should address these issues in order to maximise the number of successful participant completions.

More information about the online participant should be gleaned during online investigations. A demographics questionnaire should seek information on the motivations to complete online experiments. This could be achieved by a set of options such as *entertainment purposes* or *an interest in the research area*, or *an interest for doing paid surveys* or *a friend or colleague of the researcher*, or *researchers doing online research themselves and looking for inspiration*.

Furthermore, future online experiment web pages should seek to install an online web analytics service to gain greater insight into how participants arrive at the experiment website such as whether they were referred to the site by another website or from what query words they entered into a search engine. Then, having reached the site, a greater understanding of how the user navigates around the website could also be interesting, how long they spend - if at all - reading the participant information sheet, the instruction material and how long they spend completing demographics and exit survey questionnaires. Online web analytics services should save on development and analysis time and provide an additional source of data to complement the data collected using the experiment apparatus.

Drop out could also be attributed to technical reasons, such as not meeting the minimum plug-in requirements. While attempts were made to ensure that the participant's browser was compatible, the results of these tests were not strictly enforced. A participant could proceed with the experiment without heeding the warning that the experiment apparatus might not run without updating their Java run time environment. Furthermore, at the time when the naturalness of encoding experiment was taking place, HTML5 and JavaScript graphics were not natively supported by browsers other than at least Firefox and Safari. Internet Explorer for instance, required an additional library

in order to support the canvas element. In the case of the motion experiment that necessitated a recent version of the Java run time environment, participants were provided with a link to instructions that detailed how to make the necessary upgrades. A future experiment would seek to use native browser-based technologies such as HTML5 and JavaScript for drawing, as this technology is now sufficiently mature and has already been shown to offer comparable performance to traditional plug-in based technologies for information visualisation purposes (D. Johnson and Jankun-Kelly, 2008).

A major difference between the offline/online studies was the financial reward; it is certainly appearing to attribute a lack of financial reward as the leading reason for the observed high dropout in the online studies. However, other reasons may include: degree of connection with research, researcher and research institution; social pressure of being seen to abandon a research season midway through - in the presence of peers and a research assistant; interruption by external factors that subsequently extinguish the participants altruistic allowances and thereby casting the experiment out of mind and closing down the researchers window of opportunity to capture a complete result; lack of training material absorption - a participant may have simply missed the point of the experiment if they had glossed over the training material - and being unable to rectify any anomaly of learning with a research assistant in combination with an unfavourable effort/reward ratio that would deem emailing for remote assistance unlikely. Ultimately however, exact reasoning in each case of dropout is not known as the dataset was not sufficiently rich; this is a drawback of the online methodology and further tools to capture this information should be adopted in future research.

In contrast, a laboratory-based methodology offers insight into dropout as the research assistance has a first hand account of the technical problem, the noisy ambience, building evacuation or emergency phone call. Moreover, a laboratory-based method offers the potential for useful observational data, eye-tracking and video recording that offers a hugely rich contribution to the dataset and that could not be collected by a remote participant.

Moreover, a laboratory-based methodology has the capacity to offer: a homogeneous computing platform; an interruption free environment; a research assistant to help with anomalies and questions; and a pool of participants who are more likely to hold a higher degree of connection - and therefore an innate enthusiasm or duty to participate - with the research, institution and potentially, the researcher. While the latter has the potential to introduce subtle selection bias into the data set, the above factors offer greater assurances and controls over influences in the data that would otherwise find themselves into datasets collected by online and remote methodologies.

These factors - immediately above - may all contribute in some degree to a higher completion rate; however, a lack of financial reward may still complicate a fast recruitment drive and offer further stimulus for a participant to work through a challenging experimental contribution. One way to test the power of a financial reward within an

online experiment methodology would be to recruit participants without mention of a financial reward, to detect incidences of dropout and to examine whether an offer for payment for participant to work through difficulty results in a higher completion rate i.e. *will you still proceed if we offer X dollars?*

Due to the developing nature of the experimental methodology adopted in this course of research and the exploratory nature of the research, the experiment findings are not robust; this is most attributable to low participation and completion rates and several overly broad and exploratory hypotheses. A clear message for future and similarly audacious research is that recruitment is key; it is ultimately necessary to take advantage of crowd sourcing resources, researchers must plan for high dropout and set out to gather more insight into dropout by way of analytics services and through the use of exit questionnaires. This information not only complements the analysis of the main data, it could provide moment-to-moment indication of how participants are engaging with the whole experiment transaction.

This thesis provides a significant overview of visualisation-based search tools in one location and should aid future researchers familiarising themselves with the area. It is hoped that the lessons learned in this thesis will into future research planning. It offers a proposal for the evaluation of spatial evaluation with a number of qualitative dimensions previously not highlighted. It also recommends that while this technology is sufficiently established, researchers should have a healthy scepticism when approaching this research and a realistic perspective that the ranked-list is very likely to stay in the future. We must maintain a healthy scepticism until usability research offers a streamlined way to build document spatialisations.

While drop out was high in the experiments of Chapter 3 and Chapter 4, in principle, online experiment methodologies open up the experiment to a very large and diverse participant pool. However, at least two barriers to this online pool were encountered during this program of research: drawing attention to the experiment and enticing participation. With regard to drawing attention to the research, listing on online experiment web sites is a sound approach; however, this alone is not sufficient to attract a high participation rate. The exception to this is that of crowd sourcing websites which surface a large number of potential participants who are willing to complete a number of small tasks in succession for a small monetary reward.

It is highly likely that Mechanical Turk would have solved the recurrent small population sample issues reported in this thesis. Moreover, each of the online experiments would require minimum modification to suit a crowd-sourcing application; the only modification would be the need to implement a page displaying a code at the end of the experiment for workers to enter into the referring crowd-sourcing page - so as to trigger dispersment of the monetary reward. However, there are a number of barriers for those researchers outside of the United States and even today, Australian researchers at least, must overcome the requirement of a US profile. It may be the case that Australian

researchers who are using Mechanical Turk either have experience with forwarding companies in the United States or have a collaborative research partner in or associated with the United States; this is a major barrier to obtaining a Mechanical Turk ‘Requester’ account. Nevertheless, whilst Mechanical Turk is beyond the reach of Australian researchers, there are alternatives; CrowdFlower <http://www.crowdfunder.com> is one such example.

Furthermore, there are additional factors relevant for evaluation beyond online or offline methodological considerations. These include sourcing a test collection and importantly a set of adequate experiment tasks that permit a user to utilise a tool to its full potential. Related to this, an appropriate set of metrics both behavioural, time and outcome based are necessary to establish a benchmark for comparison between competing design alternatives. Accordingly, involved with establishing a measure of outcome - whether Bookmaker score or more traditional Recall and Precision measures of information retrieval - a definitive relevance rating for each document is required. Establishing a definitive standard of relevance might include an expert/crowd methodology adopted in the space and interface experiment of Chapter 6.

Online experiments necessitate slightly more explicit thinking regarding the above factors; however, they are dealt with more easily and without great effort in an equivalent laboratory-based experiment. In a laboratory-based experiment, many of the listed considerations are managed by a research plan or carried out by a research assistant over the course of the experiment session. However, a research assistant is not always available to answer questions for participants in a different time zone and consequently these must be dealt with in a more creative fashion. Each of the above items is listed in Table 7.1 on the facing page and a comparison of strategies adopted between laboratory offline and online methodologies.

Factor	Offline Methodology	Online Methodology
Technological	Ensure homogeneous computing infrastructure in lab. Pre-install and pilot test set up ahead of time.	Include ‘pre-flight’ checks to confirm remote participant has appropriate software installed. Provide links to assist users with installation if necessary; alternatively, use browser native technologies and/or utilise a social media API or platform.
Ethical	Provide each participant with ethics material. Request informed consent by signature. Reassure participant before the start of experiment.	Link to ethics material at landing page of website or involve ethics page as part of the experiment path. Use a modal ‘confirm’ box to gather informed consent.
Drop Out	Ensure professional and knowledgeable research assistant; ensure session is stimulating.	Conceal a research agenda as: an enjoyable or challenging game, by providing a performance evaluation review, by framing importance of research in terms that would motivate person to participate. Avoid large walls of text for instructional material.
Reliability	Use test of effort such as questionnaires to check for gaming of session.	Per offline. Furthermore, ask additional questions regarding participant’s enthusiasm and reasons for participation. Use a web analytics service to complement dataset, to understand where from/how participant arrived at site and how long participant spent on each page.
Recruitment	Use word of mouth, email, e-newsletters, noticeboard advertisements and personal invitation.	Per offline. Furthermore, use website listings specialising in online experiments and social-media channels to reach a global audience.
Collection	Record incoming data to hard drive and move to alternative location as required. Archive data with participant anonymity in mind.	Implement server-side scripting to store incoming data to flat-file. Ensure data is held in a secure location of server with appropriate security measures in place to protect against unauthorised access. Archive data with participant anonymity in mind.

Tab. 7.1: A comparison of online and offline methodological factors influencing search user interface evaluation.

7.5 *Future Experimental Work*

At face value, the ensemble of empirical work suggests: that data encoding paradigms should remain static in nature; that the user's prior expectations and beliefs could yet still provide a basis for mapping graphical features and data features together; that document spatialisations should include a theme list style control to facilitate projection dimension selection, some level below 50% pop-up background transparency should be used, a choice of multiple pop-up window facility or detail-on-demand should be provided, and that an integrated full-text document view should be considered.

Presently, a system consisting of these facets and components is not advocated to replace contemporary search tools; instead, it is suggested that future work should focus on the strongest contenders emerging from this course of work, whilst addressing further: the negative assessment of motion, the preliminary state of the encoding naturalness hypotheses, and the uncontrolled variables identified in the analyses of the space and interface experiment.

The negative subjective impressions of the motion experiment should be improved; in particular, a deeper understanding of what constitutes an irritating motion is needed - given that earlier research has indicated that not all motions are distracting and irritating. A leading outcome of such work would involve establishing a set of characteristics for animated icons, which strike a balance between perceptual dominance and interpretative facility. One suggested idea is to draw comparisons between the icons of Chapter 3 and complex icons in which small features of an icon's composition are animated rather than animation applied to the whole icon (e.g. Horn, 2007). Research should address whether irritation and perceptual dominance peak as animation is applied to the whole icon in contrast to motion influencing smaller parts of the icon.

For instance, rotation motion could be depicted in only the border of an icon's shape, while the central region of the icon could be made to cyclically grow and shrink in size or to shuffle about. It is hypothesised that these motions would be less irritating since the bounds of the icon remains static. Implicitly, this has an additional effect of reducing the amplitude of grow and shuffle motions, which may also be contributing to irritation. While not manipulated in the experiment, using smaller motion amplitudes and reducing grow and shuffle motions to vibrations may have a positive effect on irritation. However, too smaller amplitude may be perceptually uncomfortable and introduce an additional factor for control.

The main goal for future work is to improve the subjective experience of the searcher in the hope that minimising stress on visual perception will optimise visual decoding. While it is envisioned that dynamic codes will always require additional time to interpret in order to establish a basis for change through time, a perceptually optimised motion might be possible that can be interpreted accurately and in quick succession. The results of the present motion research indicated that not all static cues are inter-

preted as quickly as others and not all are as quick to interpretation as hue. Yet we routinely make use of multiple static codes to encode data even though values encoded with graphical features - other than hue - may not be interpreted as quickly. If the data set is sufficiently broad in scope and the task context warrants visualisation of multiple variables in a visualisation, then further graphical devices will be needed to generate those visualisations.

An emerging trend indicates that the size of encoding paradigms is increasing and that search for these types of complex targets is supported by perception. Nowell (1997) experimented with four graphical features of an icon with each of three possible configurations; Ware (2004) suggested a limit of eight graphical features each with two possible configurations, while Brath (2009) proposed that the potential scope of shape encoding is huge. In addition, the visual search literature has gradually shifted upward the dimensionality of search targets, which were initially feature searches or two feature conjunction searches.

While the present research indicates that motion frequency, in the form presented to experiment participants in the motion experiment, may not be an appropriate way to increase the possibilities for encoding, future research may still yet argue for an appropriate use of motion frequency under precise circumstances. For instance, dynamic graphical features may be warranted when those features provide a natural way to encode a certain data attribute.

The second area for future work centres around the naturalness hypothesis as applied to information search tools. Despite the small sample size and limited encoding paradigm size, this experiment provides the first steps toward a more robust confirmation of the naturalness hypothesis. This experiment has utilised an obvious instance of naturalness i.e. mapping icon size to document size. However, immediate future work is to repeat this experiment thereby investigating the use of not only size and shape but perhaps also texture or density on representing a count of words per cluster. Further work should then seek to expand the scope of the encoding paradigm, to isolate further cases where natural encoding applies. Further research will require a large sample size and perhaps a laboratory-based replication to shed further light on the validity of this hypothesis.

It is easy to establish that an encoding is theoretically intuitive. However, it is altogether another thing to show empirically that a relationship exists between an encoding paradigm and the searcher's interpretations of those features within the context of their task. If the graphical feature used to encode a data feature violates a general pre-existing belief or expectation of the searcher, for example it would be a violation to represent big things by little things or to represent how big something is by the number of sides a representing shape has, then the expectation is that the searcher has to use explicit and effortful cognition to overcome pre-existing beliefs in order to use the interface. Accordingly, the better interface design will be that which permits a

user to engage with full capacity and efficiency, without a need to suppress interfering thoughts and without making errors based on erroneous beliefs.

Since an encoding rule involves a mapping of a graphical feature to a data feature that is then displayed in an encoding legend to facilitate lookup, the role of the actual labels in encoding legends should also be investigated. Since encoding legends feature both the graphical code and the numeric or text label, the text is likely to have an impact on the user's decoding. Specifically, what is the role of data labels that form superlative relationships such as big, bigger and biggest? Do superlative labels, rather than raw values, have a reinforcing effect on the user's learning of an encoding paradigm? Bertin (2011) argued that a natural ordering of graphical attribute values should be used where a natural ordering of the data attributes is present. Precise numbers may not necessarily be useful to search such as in the case of rank or file size, so is there any benefit using textual labels instead of numeric labels that represent recoded relevance ranking and file size into categories that can be labelled by adjectives which form superlative relationships.

In addition to an increase in scope of the encoding graphical features and data features, future work should seek as a priority, to investigate the role of the naturalness hypothesis under different conditions of cueing. In the experiment of Chapter 4, participants were not cued to the existence of data encoded by icon shape or by size. However, participants were informed regarding the role of colour coding. Yet, when requested to explain the use of colour coding in the apparatus, a less than perfect understanding was observed across the participant pool. Colour coding was a prominent aspect of the interface and participants received training on how to enact changes in the interface that would benefit their search and as a result, update the colour coding in the interface. This observation leads to two additional areas of future work.

The first is to improve the instruction material and to provision participants with additional practice trials at the beginning of the experiment. While this is perennial advice for any researcher, it is difficult to gauge at what point instruction material is appropriate and complete. This is particularly exacerbated when the research is conducted online potentially reaching a participant demographic beyond computer science students and furthermore, when participants are participating voluntarily and may be devoting limited resources to the absorption of instruction text.

The second area is to investigate the naturalness hypothesis under different levels of cueing to the encoded data: what happens to performance if participants are cued to the encoding before using the apparatus such as by actually mentioning in the training material that size corresponds to cluster size and shape corresponds to number of words. It may be the case that pre-task training is sufficient to foster the light bulb moment when the participant has a solid grasp on how the interface works, which would have otherwise come about by trial and error. However, under cued conditions, the naturalness hypothesis would still apply if performance under unnatural encoding resulted in

poorer performance than under natural encoding. Under these circumstances, the case for natural encoding would be further supported, since we typically expect to be able to operate everyday applications and devices without needing to devote huge cognitive resources to achieve such interactions.

The final themes of future research centre on the space and interface experiment, and can be divided into three parts: pop-up transparency, document full-text view and projection dimension controls. Starting with pop-up transparency, more work is needed to establish an appropriate level of transparency that optimises readability of pop-up windows. Furthermore, future work should seek to establish the degree of reliance a participant has on semantic landmarks and whether searchers rely on the semantic content of pop-up windows to establish an understanding of spatial region. If it can be found that participants make little use of the semantic landmarks i.e. small text labels, then further work could be devoted to improving the semantic content of pop-up windows. One idea of interest is to utilise keyword laden pop-up windows to represent semantic landmarks.

Nevertheless, a level of transparency is still necessary, in order to reveal document icons below the pop-up windows. Further work is required to investigate these search tool designs under greater levels of document icon density and larger result set sizes; in the least, this type of tool should be able to facilitate search for up to 1000 results as is provided by contemporary web-based search engines.

In relation to document full-text view, a more realistic delay to open a document is required. A shortcoming of the experiment of Chapter 6 meant that the overheads for opening a document in an integrated full-text view were unrealistically low. A simulated random delay when opening a document may provide a better idea regarding the role of an integrated view. Additionally, the full-text view could be extended to facilitate tabbing of interesting documents in order to offer the same strategies that searchers engage when utilising web-based search engines. These strategies allow participants to open a number of results in quick succession for subsequent batch relevance judgements. Following these adjustments, future work should move toward linking the full-text view with the document spatialisation such that participants can highlight sections or paragraphs in the full-text in order to enact change in the spatialisation.

Further work should also be directed to the control and manipulation of icon density. It is apparent that this has influenced the performance and behaviour of participants when the pop-up transparency was non-transparent and the document full-text view integrated and adjacent to the search results.

In relation to projection dimension control, further work is required to improve the utility of the theme list control through augmentations that provision a searcher with visual cues to make judgements regarding the best combination of projection axes. This should identify the utility of list sorting and list item mark up such as text enhancements

to facilitate dimension relevance judgements, based on a relevance score calculated by a similarity measure between the selected dimension and all other dimensions available for configuration.

The final area of future work centres on the role of online methodologies for like experiments. A web-based methodology to collect data was adopted for the motion and naturalness of encoding experiments. In contrast, the space and interface experiment utilised a closed-door laboratory paradigm for data collection.

One of the main design differences between the online and offline experiments was that the subjects were paid for participation in the offline, laboratory-based experiment. Ultimately, taking greater steps to understand the motivations of the participant pool should feature more prominently in future research. Future online experimentation should make use of a web analytics service in order to gauge the navigation patterns of participants within the experimental site and offer - with limited additional programming overheads - a more precise understanding of how long participants spend on aspects of training and questionnaires for instance.

In addition, unsolicited feedback was very common in the laboratory-based experiment in contrast, there was no unsolicited feedback regarding the online experiments. Additional measures should be taken to solicit general comments and requests for help from participants completing experiments over the Internet.

Furthermore, whilst the level of experiment drop out was as expected per the work of Reips (2002), additional steps should be taken to reduce the level of drop out at the instruction stage and drop out due to technical reasons. With regard to technical problems, contemporary Internet browsers are increasingly making it easier to upgrade browser plug-in technology with the press of a button, yet checks should be employed to notify participants prior to commencement of experiments to ensure the experiment apparatus will run having passed through the initial demographics and instruction stages. Better still, post-demographic technical drop out could be completely avoided by embedding the whole evaluation website into an embedded application. Alternatively, adopting native browser technologies such as HTML5 and JavaScript could also deal with the issue of technical drop out.

In relation to all forms of drop out that are attributable to the motivation of the participant, experiment designs that make participation fun and make use of social media frameworks should be investigated further to provide an additional gateway to participation.

7.6 *Contribution of Thesis*

As a whole, this thesis contributes a comprehensive overview of the components that connect together to form spatialisation-based search tools. Researchers and designers in

this field will benefit from the survey of systems presented in Chapter 2 and in addition, the taxonomy of components presented as Figure 7.1 on page 357 in this chapter.

The experiments presented in this thesis are more so exploratory pilot studies that introduce and develop previously unexplored ideas. Each study presents a number of lessons learned that will make future confirmatory research more robust and reliable. Future research should split out conditions first and then combine the findings of those individual studies into a realistic exercise. We must first sacrifice realism before we can bring together baseline findings into a realistic evaluation. Pilot studies are necessary to pre-evaluate both training material and exit questionnaires and piloting should cover pre-evaluation of the graphical features as well.

From a high level perspective, the motion and naturalness experiments have contributed to the development of an online methodology for conducting visualisation research; specifically, they highlight the need for excellent training material which is terse, yet descriptive and readable and engaging; furthermore, these studies suggest that researchers should plan for high dropout and that wherever possible, the use of crowd sourcing services should be carefully considered. Whilst one drawback of an online methodology is the potential for observational data, further richness might be added to the dataset if a researcher can capture metrics on a participants engagement, motivation and reason for experiment dropout.

The spatialisation focus of the latter chapters is an exploratory investigation as much as a call to arms to improve the usability of visualisation-based search systems. These chapters embody an acknowledgement that visualisation-based search is likely to maintain the attention of a fringe of researchers, at least into the immediate future, yet they do not hide the fact that these ideas must evolve further if such systems are to reach a mainstream audience. It is hoped that the set of qualitative criteria for spatialisation algorithms - presented in Chapter 5 - will shape future development of document layout algorithms that feature in visualisation-based search systems.

In relation to the experimental work presented in Chapter 3, Chapter 4 and Chapter 6, the contributions of this work are as follows:

The Role of Motion in Attribute Visualisation

Overall, increasing icon complexity results in an approximately linear increase of preparation time, answer time and errors made. However, the increase of time and error for increasing dimensionality is more severe as the number of motion frequency features increase.

The Role of a Natural Encoding Paradigm in Attribute Visualisation

When choosing graphical features to encode data, the prior expectations and beliefs of the user should be considered in order to produce natural data encodings. Where natural data encodings are used, it is expected that the user's interaction

and learning will be more efficient and thus, lead to a better outcome through tool use. The data are suggestive that interaction costs are lower when data encoding is more natural.

Pop-up Window Transparency in a Multiple Detail on Demand Interface

Pop-up windows superimposed over a document spatialisation should seek to optimise foreground text whilst not obstructing content below. A level at or above 50% transparency is not an appropriate level for this type of application. Providing a control to toggle between a multiple and single pop-up mode may be appropriate.

Document Full-text View Integration in Search User Interfaces

Document full-text view integration appears to strongly influence searcher behaviour due to the lower cost of accessing the document. Searchers may adopt particular strategies when the perceived cost of opening a document full-text is higher. Furthermore, the document full-text is likely influencing the spatialisation's display density and this should be investigated further.

Projection Dimension Control

If the document spatial model permits, a projection dimension rotation control should favour a theme list and not a theme cloud control. Despite the graphically rich presentation of information in the theme cloud, the simplicity and universality of the theme list prevailed and was more readily preferred; although objective search outcome measures were not strong enough to conclude which was the superior of the two alternatives.

7.7 Conclusion

Presenting a collection of search results by way of non-traditional, non-linear result presentation paradigms requires a multi-faceted effort and ultimately, design should closely consult the user of the tool. This thesis embodies this contention as evidenced by the collection of research presented. This multi-faceted effort can be summarised concisely as affecting the point, space and interface of the search result presentation.

Three experiments were presented; the results of which revealed interesting findings about the way search user interface designs influence human behaviour. As a consequence, a number of recommendations were made for future search user interface design and research. A survey of systems and a taxonomy of components testifies to the richness of research in the search result presentation area. Accordingly, the potential for future work is great.

APPENDIX

A. MOTION ENCODING EXPERIMENT: INSTRUCTIONS

Instructions, please read carefully.

These are the general instructions for the experiment. It is important you know what to do. If you find any of the instructions difficult to understand or find the experiment task is difficult please email the researcher at treh0003@finders.edu.au.

Figure 1. below shows a screen shot of our interface. The major features of this interface are a Visualisation (series of square shapes in the centre of screen), a Key/Legend (left side of interface) that ‘maps’ visual characteristics of square shapes to different text categories, and a Task Bar (bottom of interface) where your task statement appears. We describe each component with more detail directly below Figure 1.

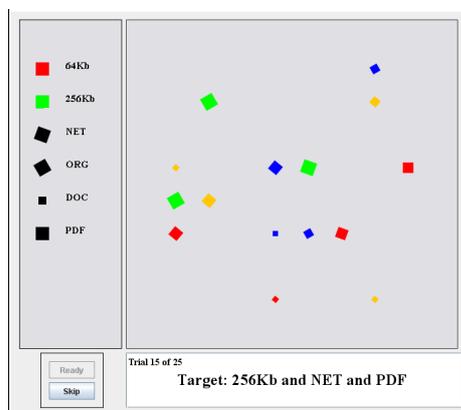


Figure 1. A screen shot of the interface

You will complete 25 search tasks. The interface will tell you how many tasks you have completed. Your job is to find the icon with attributes that match the task statement. You determine your target by matching the text categories listed [in] the task statement with the visual attributes in the key/legend and then finding the icon in the visualisation with all the visual attributes that map to the requested text attributes. Once you know the intended target, click the ‘ready’ button. Some objects will appear on screen and you must find your target object amongst distractor objects in the centre of screen. If you cannot find the target you may skip the task by pressing the ‘skip’ button.

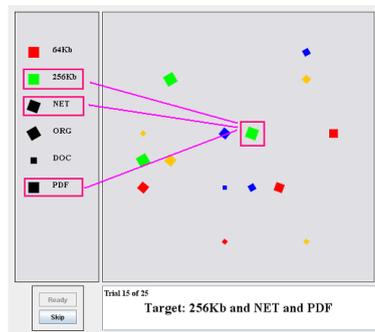
Visualisation The visualisation displays a number of different distractor objects and your (1) target object. Your target object will be a square shape with all the shape attributes that your task statement indicates. You make your answer by left mouse clicking on the target object.

Key/Legend The key (left box) ‘maps’ words to visual attributes of the square shapes e.g. 256Kb (word) to Green (visual attribute) much the same as on a road map e.g. a red circle represents a traffic light.

Task Bar The task statement (bottom box) indicates the words that you need to look up in the key and determine your answer. The task bar also displays the number of tasks you have successfully completed. If you make an incorrect answer this area will flash pink.

Example

Figure 2. below shows an annotated version of the interface to help you with understanding the task. The pink rectangles indicate which of the visual attributes will make up the target object. The target object in the centre of screen is also highlighted. Notice how the target object (highlighted in centre) has the visual properties highlighted in the key.



When you are ready to begin, click the ‘Begin’ button.

B. NATURAL ENCODING EXPERIMENT: INSTRUCTIONS I

Instructions, please read carefully.

You may only have used Internet search engines that provide you with lists of search results spread across different pages. Therefore, before you start this experiment, we would like to introduce you to our method of presenting search engine results.

Figure 1 shows a screen shot of our interface. The major features of this interface are an ‘Explorer Tree’, a Visualisation (picture) of document clusters, a Key/Legend that ‘maps’ colours to keywords and the Task Bar where your task question appears. We describe each component with more detail directly.

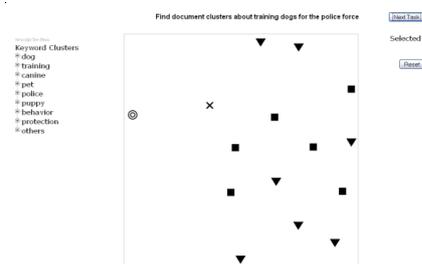


Figure 1. A screen shot of the experiment interface.

Visualisation

In preparation for this experiment, we have downloaded 120 search engine results from an online commercial search engine service. We then grouped documents into ‘document clusters’ according to their general topic using a set of keywords for each document. Finally, we arranged those document clusters so that (somewhat) similar clusters are nearby each other while document clusters that are very much unlike other clusters are further away. Each shape you see in the visualisation represents one cluster of documents.

When you mouse over each cluster (icon/shape), a popup window will appear showing you the keywords that best describe the general topic of documents inside that cluster. You should use these keywords to estimate whether or not a cluster satisfies the task statement you have been set. When you are confident that you have found the answer to the task, click on the document cluster (there could be more than one

if necessary). A red underline will appear to indicate you have successfully selected the document cluster as an answer. To continue to the next task, click the ‘next task’ button.

Explorer Tree

At the left of screen is an ‘Explorer Tree’ widget. You may have seen this widget before in the Explorer or File Manager program in your computers Operating System. You can click on each word that has a small plus ‘+’ sign next to it. When you click on a word, the tree opens up or ‘branches out’ to reveal additional words; similarly, clicking on a word with a negative ‘-’ sign will close the branch.

Rather than mousing over each cluster one by one, you can use the Explorer Tree to guide your search more efficiently. Each word in the ‘branch’ will also appear in one or more of the cluster popups in the visualisation. When you click on words inside the ‘branch’, the clusters (shapes) that contain that word will change colour. The colour assigned to the word you last clicked on will appear in the Key/Legend at the right of screen.

Key/Legend

You can select up to six (6) keywords at once. The first three (3) keywords will appear in the key/legend as coloured dots. The next three (3) keywords will appear as coloured circles. Document clusters that contain selected keywords will change colour. However, if a document cluster contains additional colours as well the resulting colour of the document cluster icon will be the combination of each relevant keyword. For instance, a cluster containing keyword A (mapped to colour red) and keyword B (mapped to colour green) will appear as yellow. Note that the colour of keywords 4,5 and 6 will effect the colour of the outline of the shape where as keywords 1,2, and 3 will effect the inside of the cluster shape.

You can reset the colour mappings by clicking the ‘reset’ button; by doing this, all document cluster shapes will return to a black colour.

Task Bar

The task bar will contain the questions that you need to answer. To submit your answers, click on the document cluster or clusters (there could be multiple answers). A red underline will appear to indicate you have successfully selected the document cluster as an answer. To continue to the next task, click the ‘next task’ button. Your answers are recorded each time you click the ‘next task’ button.

When you are ready to start, click the ‘ready’ button. If you are still unsure, click ‘help’ and we will provide you with a step by step example of how we would like you to complete the experiment.

C. NATURAL ENCODING EXPERIMENT: INSTRUCTIONS II

Training Step By Step

In our experience, it is seldom easy to explain effectively the workings of a complicated interface over the Internet! Here we offer a systematic walk through of the interface. Figure 1 shows a screen shot of the experiment interface at the beginning of a new task.

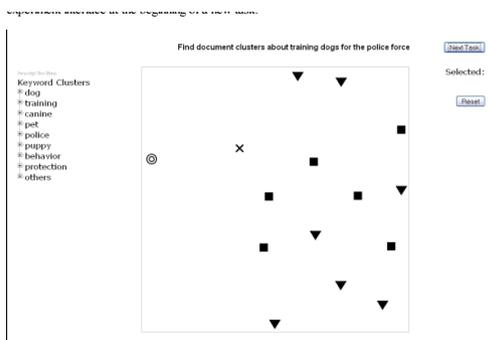


Figure 1 - screen shot of the experiment interface.

Look at the task statement at the top of Figure 1. It asks you to find documents about training dogs for the police force. Think about the types of words that describe the concept of dog training for the police force. Look at the Explorer Tree at the left of Figure 1 and try to find words that you think might describe the concept of training dogs, or that are related to the words that would describe the concept of the police force. Find and click on those words in the Explorer Tree to reveal further words for that branch. Click on other words in the open branch that you think might aid your search. Figure 2 shows a screen shot of the interface once you have clicked on a keyword group to reveal further keywords that might describe the concepts of training dogs and/or police force. In Figure 2 notice that after you click on keywords in the Explorer Tree the document clusters in the centre of screen have also changed colour. Notice also that the keywords you selected have appeared in the key/legend at the right of screen. Each keyword in the key/legend has a coloured dot or circle next to it. Figure 2 shows that the explorer tree is now open at the 'Police' branch of the Explorer Tree. Opening this branch reveals three other keywords when clicked on appear on the right hand side. The interface assigns a colour to each keyword and the mappings

between keywords and colours are displayed on the right hand side of the interface under ‘Selected:’ and above the ‘Reset’ button.

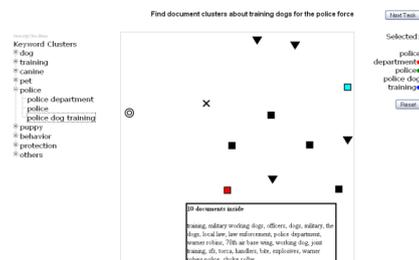


Figure 2 - screen shot of interface with keywords selected.

Now that keywords are assigned to particular colours, the document clusters in the centre of Figure 2 have changed colour according to whether or not they contain the selected keywords. The RED coloured square in the visualisation contains the keyword ‘Police Department’ and you can confirm this because when you mouse over the RED square the keywords listed in the cluster popup will contain ‘Police Department’. The CYAN coloured square is a little different. The colour CYAN is produced by mixing both BLUE and GREEN together so this reveals that the CYAN square will have both ‘Police’ and ‘Police Dog Training’ as keywords that describe documents in that cluster. You can see the content of the CYAN clusters popup in Figure 3 The contents of this pop up will contain either the number of clusters contained within the cluster and/or the total word count of all documents in side this cluster.

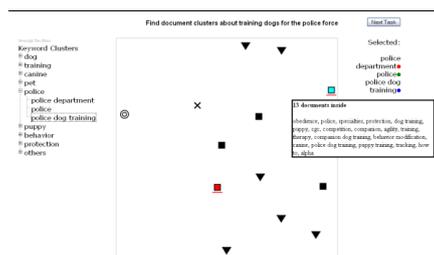


Figure 3 - document clusters saved for answer submission.

These two document clusters contain interesting keywords that are associated with ‘training dogs for the police force’. Notice that there are some unrelated keywords also document cluster is not perfect! But it does give you some indication that these two clusters could satisfy the Task statement. We do not give you the opportunity to inspect each document so provide your best answers. Left mouse click on either the RED or the CYAN or BOTH document clusters to indicate that you think they are good candidates for answering the task statement. In reality you would open these

document clusters and inspect each document to see if they completely match the task statement but that will not be necessary here. Once you have clicked on document clusters, a red underline will appear beneath the document clusters. When you are ready to proceed to the next task, click the 'Next Task' button at the top right of the interface. To start the experiment, click the 'ready' button. If you would like to go back to the previous screen and learn more about the individual components of the interface click the 'back' button.

D. SPACE AND INTERFACE EXPERIMENT: TRAINING

The following set of slides depict the initial introduction and theory component that participants in the space and interface experiment undertook. All participants regardless of experiment condition saw the following 10 slides.

Advanced Retrieval Tool Evaluation

Kenneth Treharne
AI Lab, Flinders University

Schedule of Tasks

Task	Expected Time Commitment
Demonstration and Training	15-20 minutes
Demographics Questionnaire	1-2 minutes
The Experiment	30 minutes
Exit Questionnaire	5 minutes
General Knowledge Questionnaire	5 minutes

In the following four slides below, the research assistant attempts to draw an analogy between a map of a city and a theme map, parallels are drawn between the neighbourhoods, houses and map annotations.

Training

What is a theme?
What is a theme map?
How do I use a theme map?

What is a theme?

- A theme is a topic, concept or a subject
- Often, documents will have multiple themes or topics or subject matters
- The theme might not be discussed explicitly but will contain words that are typically used when describing a theme

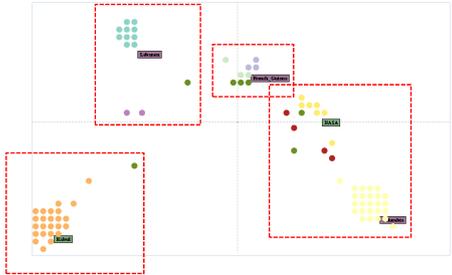
The research assistant utilises the bottom slide to draw parallels to the geographical map - i.e. each icon represents a house, a group of houses form a neighbourhood and theme map annotations correspond to the neighbourhood names.

What is a theme map?

- Think of a Street Map
- (Ignore the Roads)
- Neighbourhoods of:
 - Houses
 - Buildings
 - Areas of Interest
- Houses in an area are similar in that they all belong to a neighbourhood (labelled)



What is a theme map?

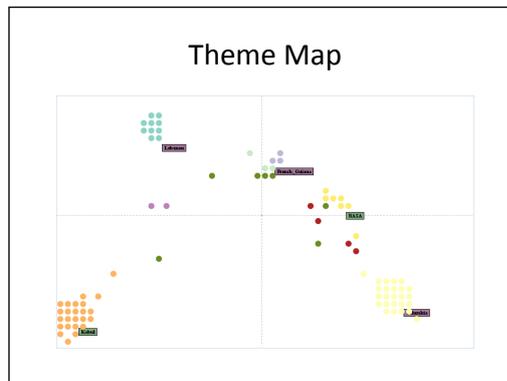


In the first slide below, an example query is discussed. The research assistant highlights how the query can be interpreted in a range of different ways depending on how the query words are entered into the search engine.

In the second slide at the bottom, the research assistant is drawing attention to how similar topics situate near by, and unconnected topics - all different interpretations by the search engine - situate at distance.

**An example – ‘Space Shuttle’ OR
Rocket OR ‘Space Craft’**

- (Vehicle) Space Shuttle
- (Vehicle) Shuttle Train
- (Vehicle) Shuttle Bus*
- (Vehicle) Shuttle Service*
- (River) Shuttle*
- (Company) Shuttle*
- (Movie) Shuttle*
- (Space) Rocket
- (War) Rocket
- (Figure of Speech) Rocketing Stock Price
- (Figure of Speech) ‘He went like a rocket’
- (Sport) The Houston Rockets



At this point, the research assistant turns to the software apparatus; the apparatus has been configured appropriately according to the session's experimental variables. Participants see a replica of the system that they will use in their experiment session. At the end of this demonstration, and throughout the entire demonstration, participants have the opportunity to ask questions and seek clarification.

Walkthrough Experiment Software

Stage 0 – Demographics
Stage 1 – List Interface
Stage 2 – Faceted Theme Map
Stage 3 – Word Cloud Theme Map
Stage 4 – Relevance Judgements
Stage 5 – Exit Questionnaire
Stage 6 – General Knowledge Questionnaire

End of Training

Any Questions?

E. SPACE AND INTERFACE EXPERIMENT: HANDOUT FOR INTEGRATED FULL-TEXT INTERFACE

Participants had the following four slides in front of them for the duration of the experiment session. These slides were printed on four A4 pages to enhance readability. There were four versions of this hand out, one for each combination of full-text integration and pop-up transparency.

How to Submit An Answer

- Open document – is it relevant?
- Close document – will still be selected
- If relevant select either Strong/Moderate/Weak relevance
- Click tag button
- If not relevant, do not rate and tag.

Here is your task, you need to find documents about the topic

Selecting search terms in text will populate this box and reshuffle the results

Tag Relevant Answers into this box

Once finished proceed with next task by clicking this button

A toggled result means you can now rate relevance

Clicking blue title will open document for preview

Go back to previous searches by clicking these buttons

Here is your task, you need to find documents about the topic

If you prefer one pop up at a time, click this button

Once finished proceed with next task by clicking this button

This is the theme map, clicking on the coloured circles will open the document in a new window

Click on words to change the configuration of the theme map. The order of factor items indicate the next best options.

Clicking on these descriptors will plot them in the theme map

Tag Relevant Answers into this box

List of results

Here is your task, you need to find documents about the topic

If you prefer one pop up at a time, click this button

Once finished, proceed with next task by clicking this button

This is the theme map, clicking on the coloured circles will open the document in a new window

Click on words to change the configuration of the theme map. The colour coding of squares indicates the next best options.

Clicking on these descriptors will plot them in the theme map

Tag Relevant Answers into this box

List of results

F. SPACE AND INTERFACE EXPERIMENT: HANDOUT FOR MODAL FULL-TEXT INTERFACE

Participants had the following four slides in front of them for the duration of the experiment session. These slides were printed on four A4 pages to enhance readability. There were four versions of this hand out, one for each combination of full-text integration and pop-up transparency.

How to Submit An Answer

- Open document – is it relevant?
- If relevant select either Strong/Moderate/Weak relevance
- Click tag button
- If not relevant, do not rate and tag.

Here is your task, you need to find documents about the topic

Go back to previous searches by clicking these buttons

A toggled result means you can now rate relevance

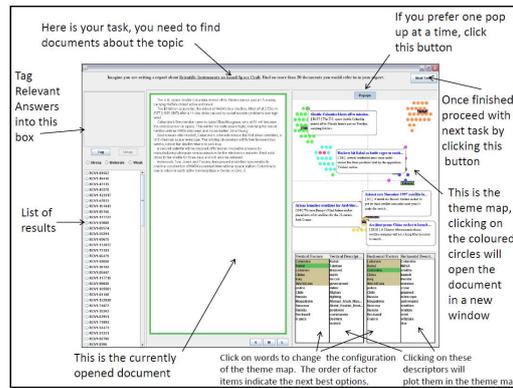
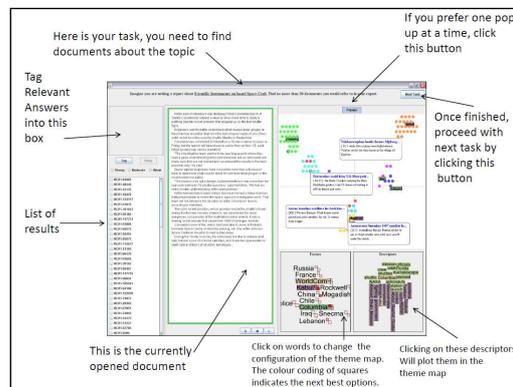
Clicking blue title will open document for preview

Once finished, proceed with next task by clicking this button

If you cannot read the text easily, you can make it bigger.

This is the currently opened document

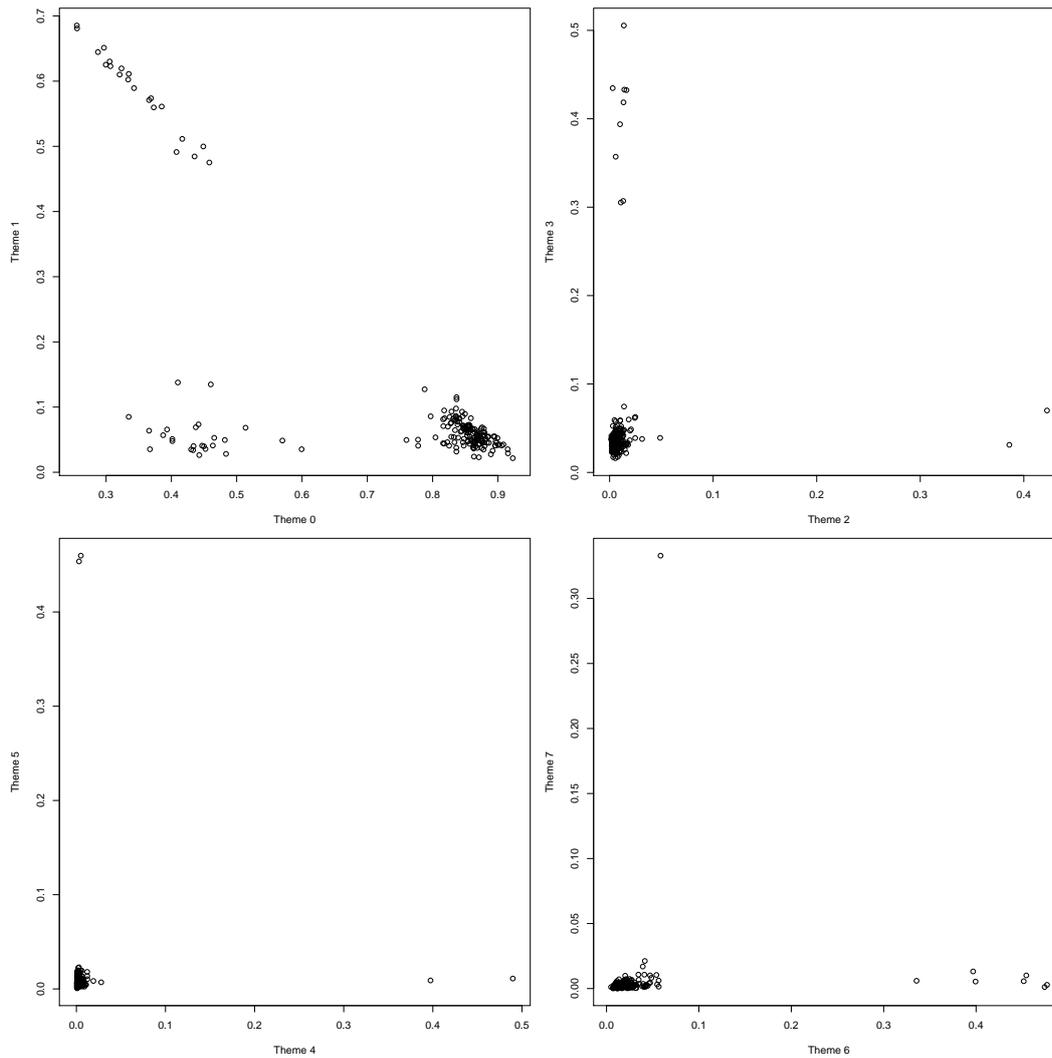
The screenshot shows a search results page with a list of search results on the left and a preview of the selected document on the right. Annotations with arrows point to various UI elements: a task instruction at the top left, navigation buttons at the top right, a search result with a blue title, a preview window, and a zoom control at the bottom right.



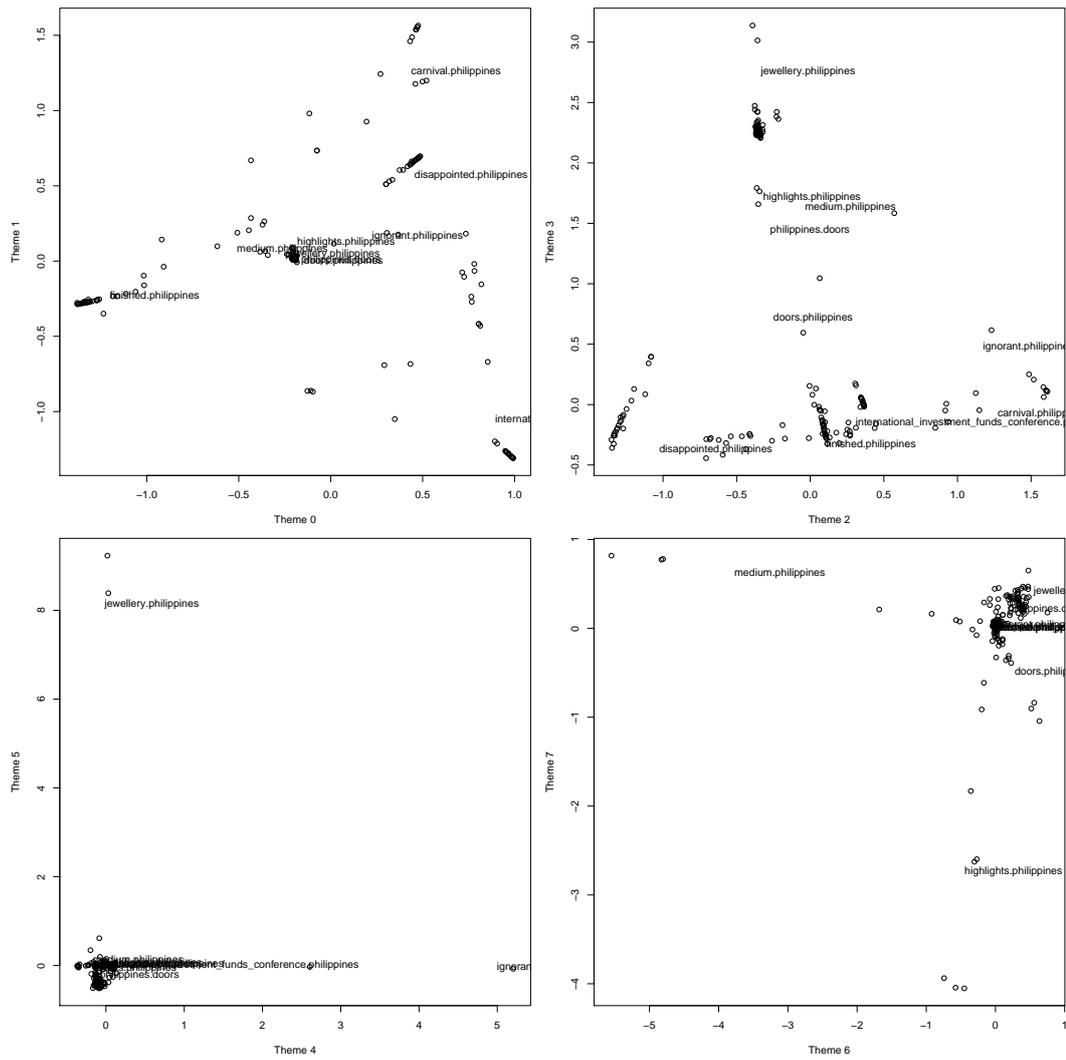
G. SPACE AND INTERFACE EXPERIMENT: SPATIALISATION QUALITATIVE EVALUATION

The following graphs were produced during an analysis of dimension compression (reduction) and projection algorithm evaluation in Chapter 5. Alongside each set of plots, the approach taken is listed and a brief discussion is made regarding the layout of each combination. The plots presented include dimension combinations (0,1), (2,3), (4,5), (6,7). For Multi-dimensional scaling, only combinations (0,1) and (2,3) are supplied.

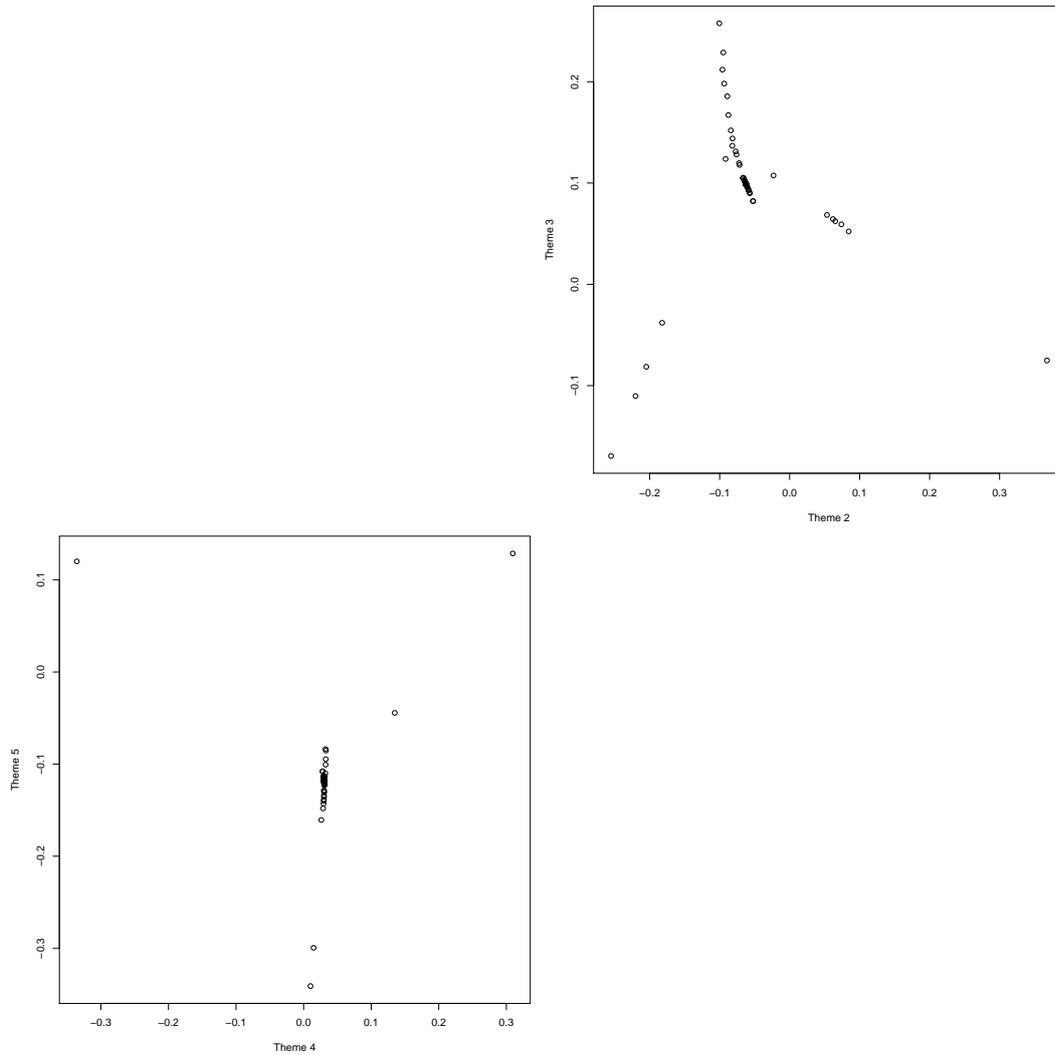
The set of plots below depict fuzzy hierarchical clustering as the compression algorithm and then projected by the raw output of the same routine. The main problem with this approach is the poor use of white space and the overly dense clusters. The (0,1) combination shows three clear clusters - though the user would spend a significant amount of time re-adjusting their view to begin interacting with a particular aspect of the data.



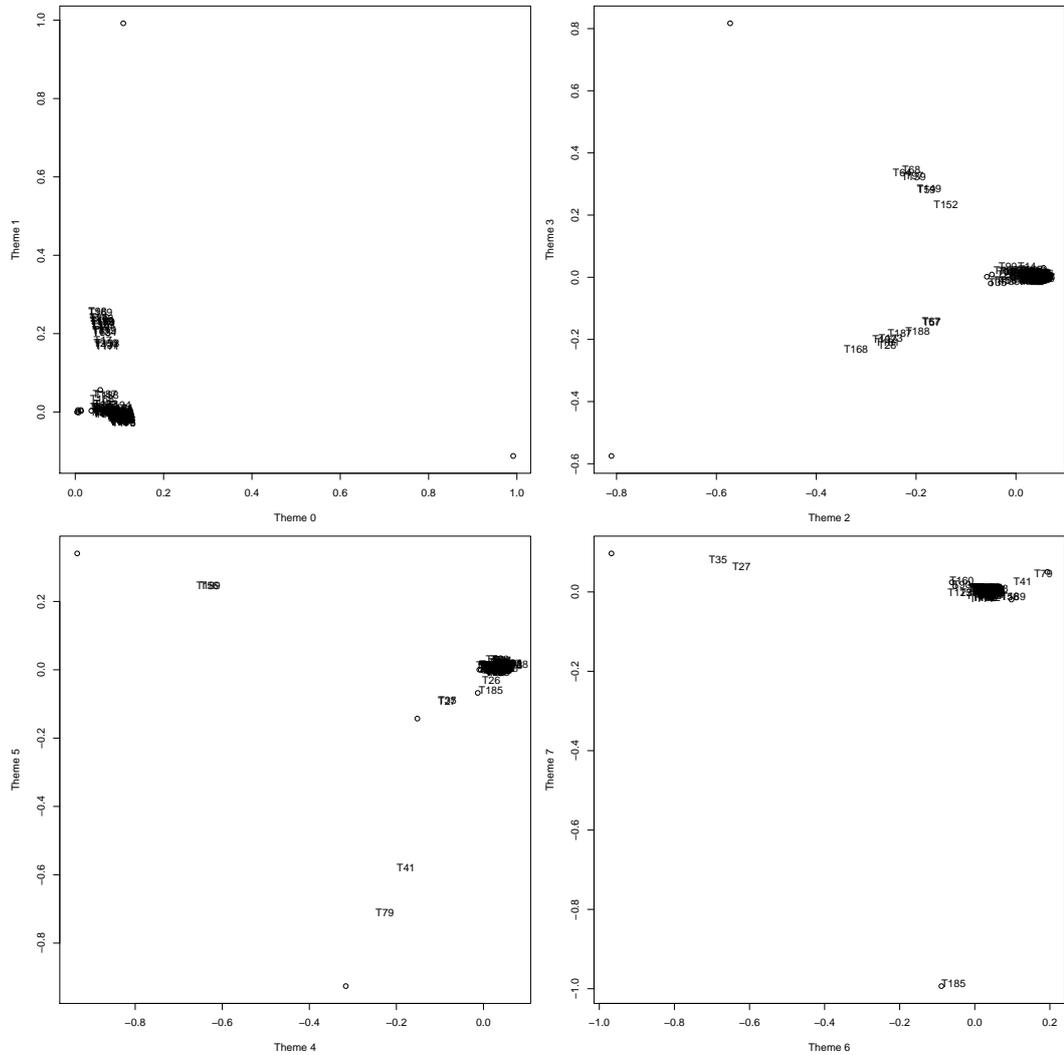
The set of plots below depicts fuzzy hierarchical clustering as the compression algorithm and correspondence analysis as the projection algorithm. The spread of points in (0,1) and possibly (2,3) are better than the example immediately prior. The textual labels indicate the position of semantic annotations in the graph and spread of annotations appears good, such that most spatial regions containing points have a semantic annotation.



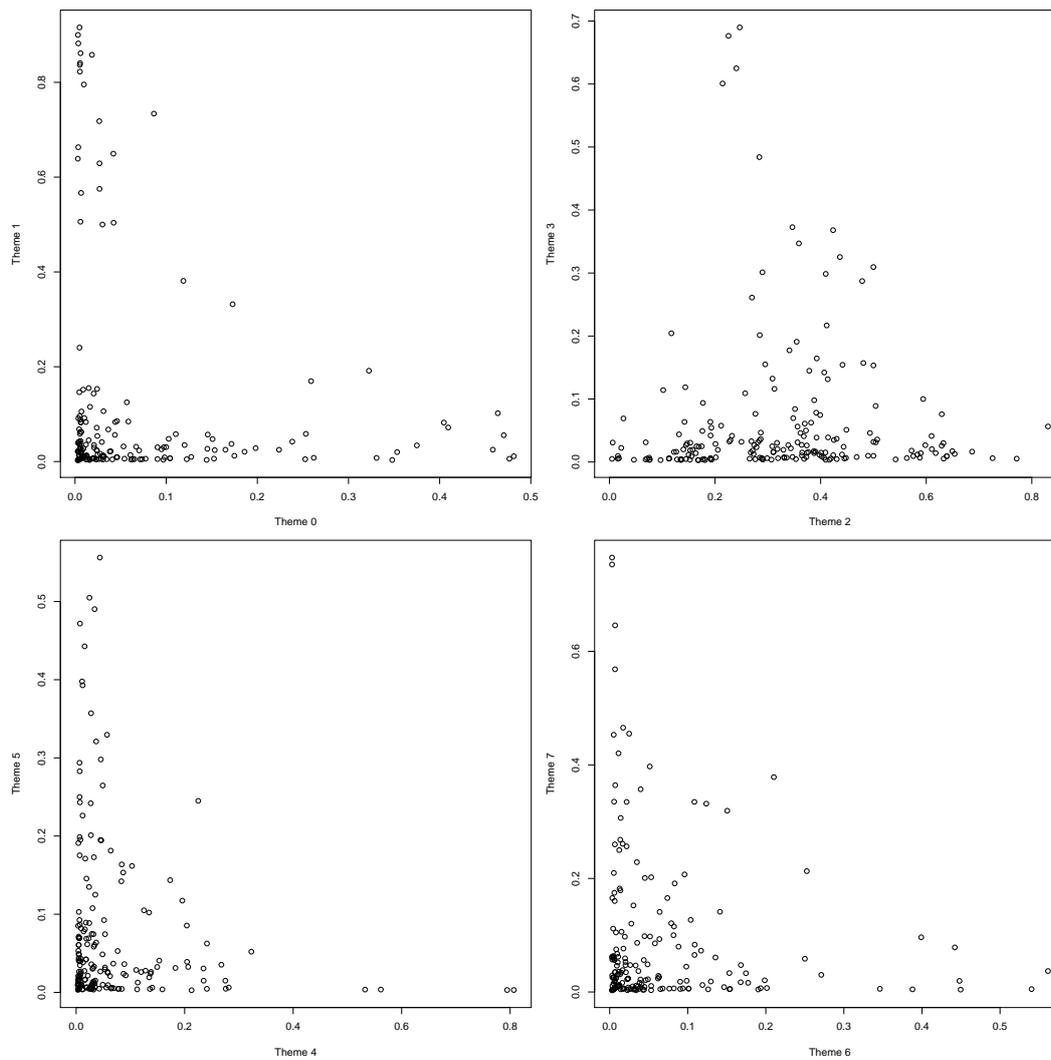
The plots below depict fuzzy hierarchical clustering as the compression algorithm and multi-dimensional scaling as the projection algorithm. There are no semantic annotations on the plots since the MDS does not provision a way to do so. The spread of points is poor, however (2,3) does show three topics in the data set.



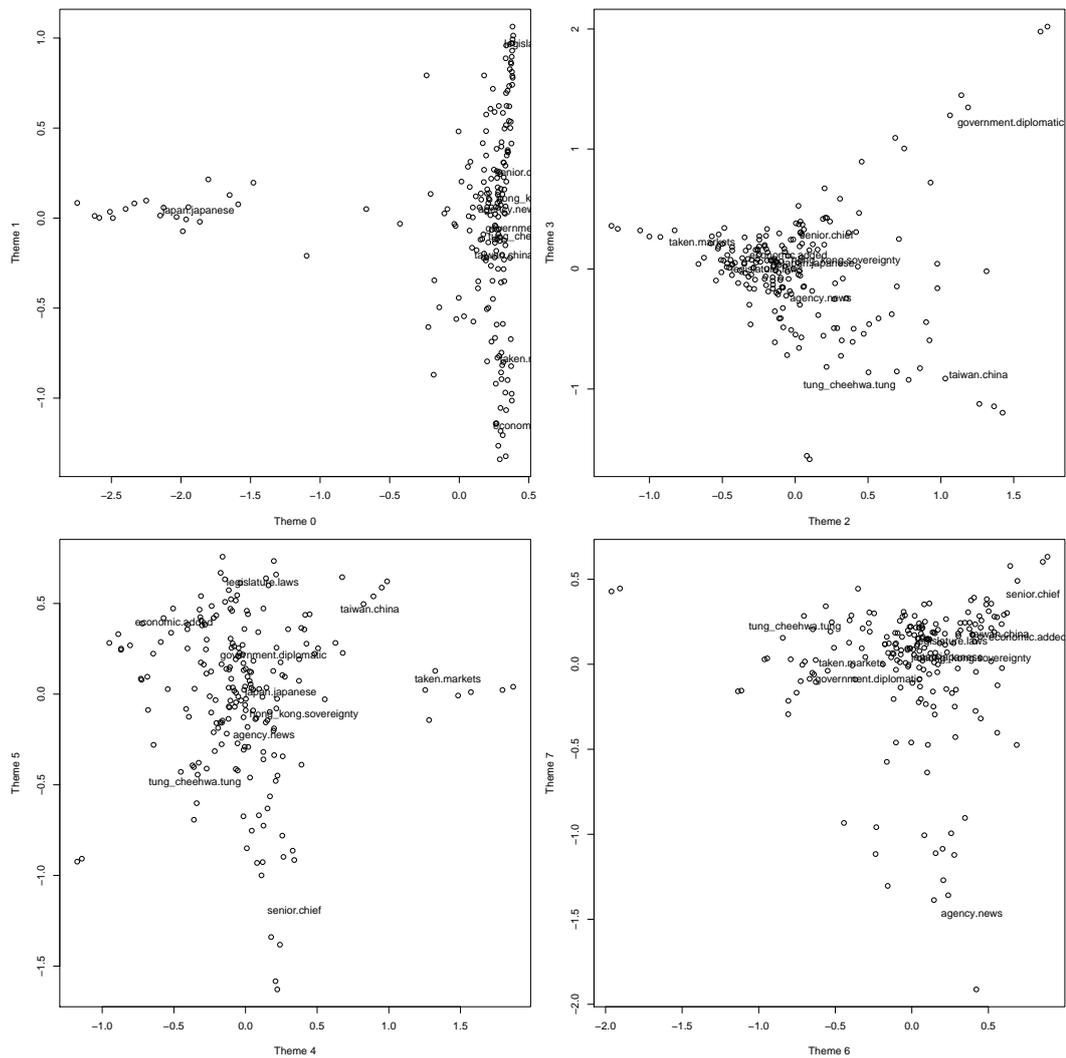
The plots below depict fuzzy hierarchical clustering as the compression algorithm and Singular Value Decomposition as the projection algorithm. The semantic annotations correspond to positioning from the terms x topics matrix. The spread of points is poor and clusters are dense; (2,3) indicates 3 topics.



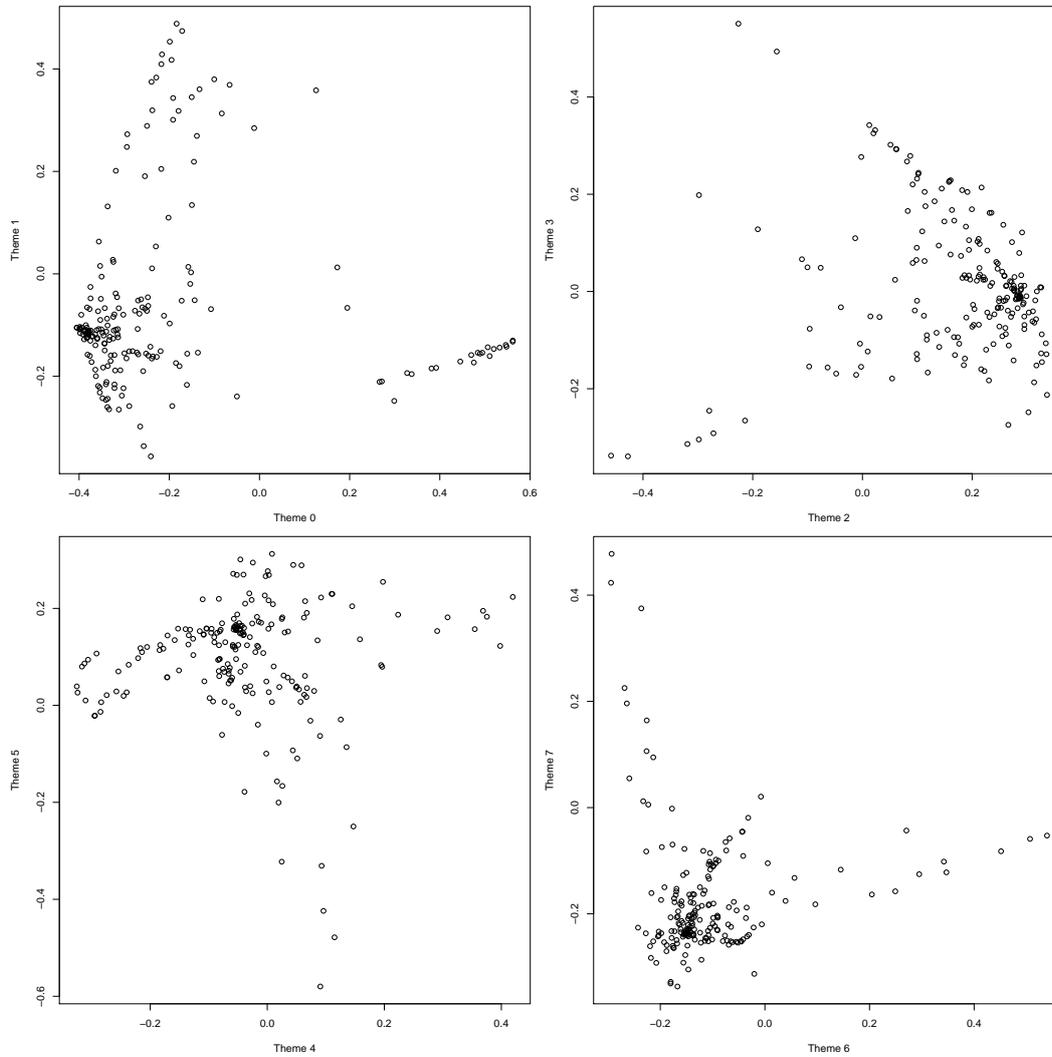
The plots below depict latent Dirichlet allocation as the compression algorithm, which is then projected by the raw output of the same routine. The spread of points is typically good across a number of dimensions combinations. There are no semantic annotations - term weights sum to one and furthermore, are not in the same space as documents. The use of white space is not yet maximally optimal.



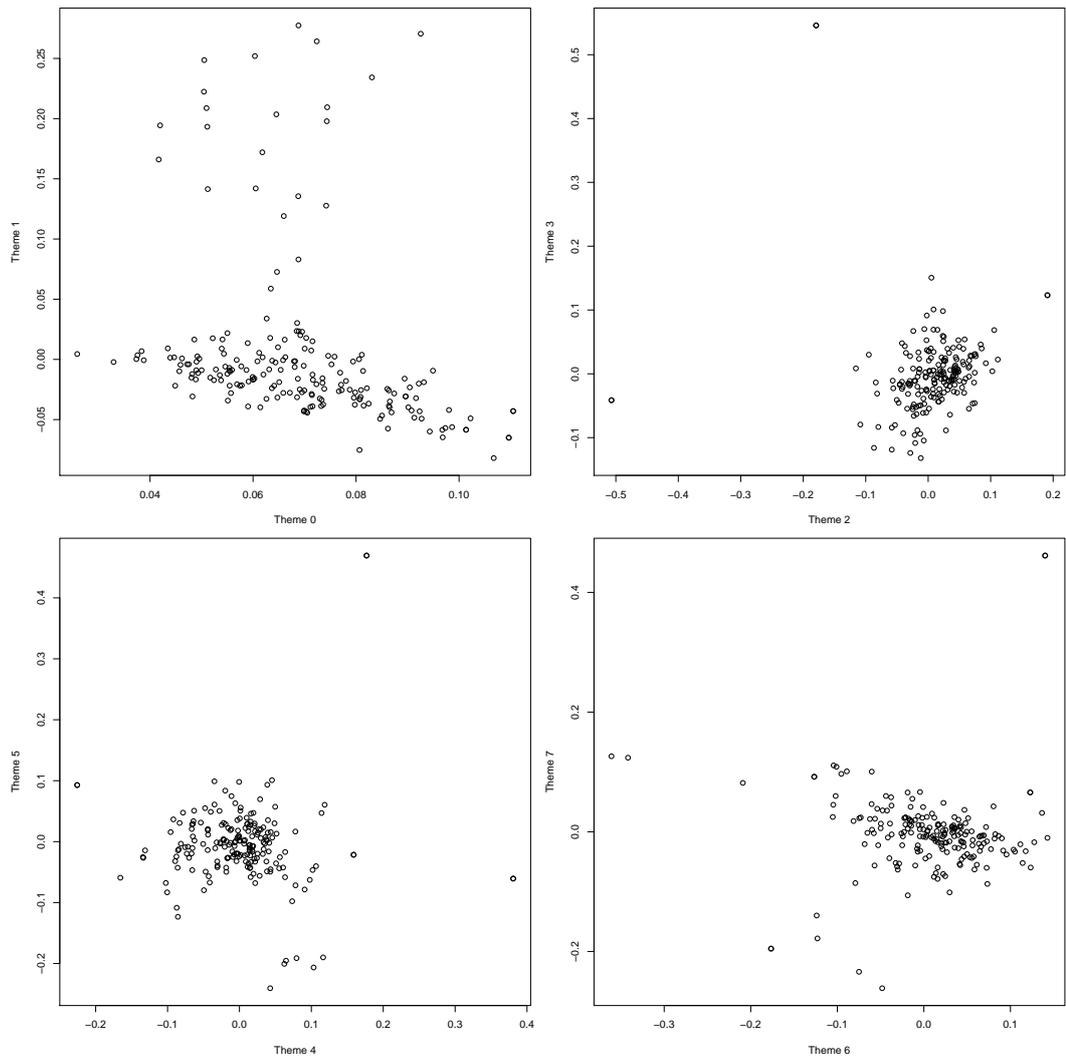
The plots below depict Latent Dirichlet Allocation as the compression algorithm and Correspondence Analysis as the projection algorithm. Arguably, this combination of algorithms may not provide a convincing demonstration of good use of white space with (0,1) containing a significantly dense clustering along the right hand edge of the plot. However, the other combinations are moderately good and the layout of semantic labels is similarly good. For another example of theme maps produced by the Latent Dirichlet Allocation and Correspondence analysis, refer to the training material in Appendix D on page 384.



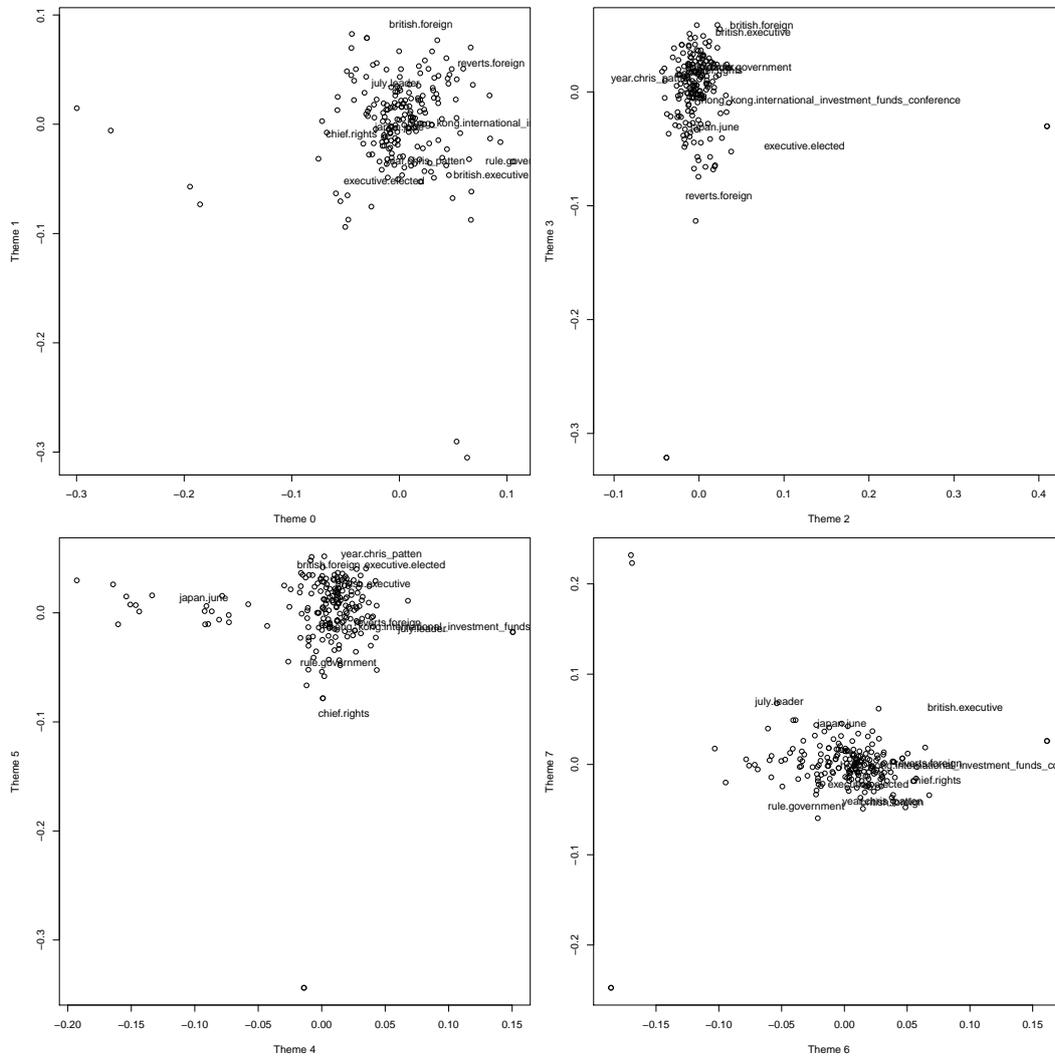
The plots below depict Latent Dirichlet Allocation as the compression algorithm and Multi-Dimensional Scaling as the projection algorithm. The spread of icons is moderately good, though there are no semantic annotations. Furthermore, there is a tendency for dense clusters about the origin.



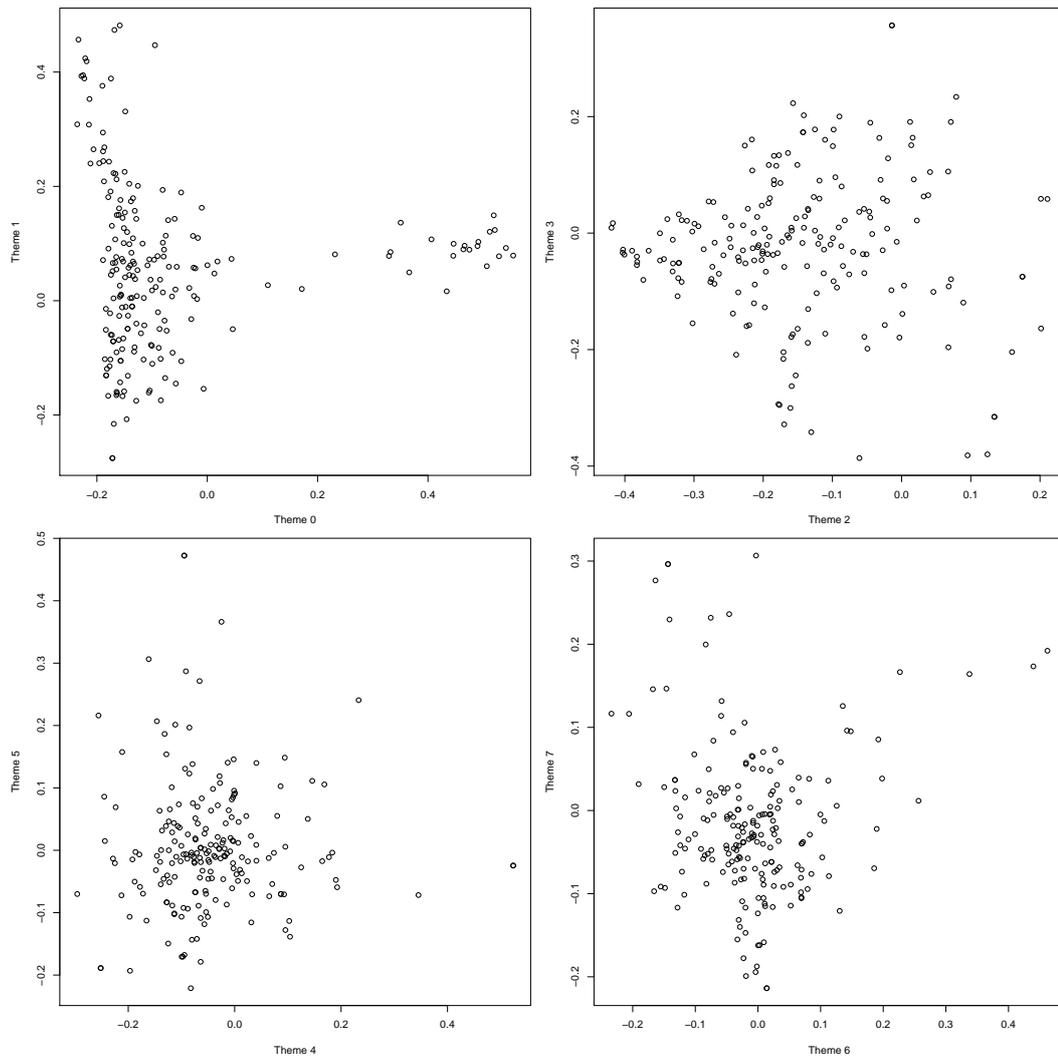
The plots below depict Singular Value Decomposition as the compression algorithm and the projection algorithm. Overall, several stray outliers cause compression of documents around the origin point. The semantic annotations are not shown on this diagram in order to focus on consideration of the layout, but these would be available. The same plot with labels is depicted above.



The plots below depict Singular Value Decomposition as the compression algorithm and Correspondence Analysis as the projection algorithm. The layout is poor with a highly dense clustering about a central point and a few stray outliers.



The plots below depict Singular Value Decomposition as the compression algorithm and Multi-Dimensional Scaling as the projection algorithm. The spread of points is typically good with (2,3) potentially the best candidate seen so far. However, without semantic annotations - therefore in violation of the qualitative criteria in Chapter 5 - the resulting theme maps are unusable.



H. STATISTICAL METHODS

This appendix outlines the objective metrics and statistical methods that are found in each of the three experimental chapters. A centralised reference for the statistical methods employed was preferable to repetitive discussion in content chapters.

H.1 Object Measures

A table of objective performance and behavioural measures is depicted below in Table H.1. This table offers a brief outline of each measure, however subsequent subsections offer a more in depth insight into the role of time and more predominantly accuracy metrics, in search tool evaluation. The remaining measures of Table H.1 are included for completeness; these were collected in the experiment of Chapter 6 specifically for the purposes for investigating search user interface components.

H.1.1 Time

As is frequently found across human-computer interaction experiments, time - recorded to Millisecond precision - features in all experimental work reported herein. Whilst information search is foremost concerned with the location of information that meets the needs of the searcher, the time it takes the searcher to locate that information is inextricably tied to an overall judgement of the success of the search system. Accordingly, the faster of two search tools that perform comparably will be perceived as superior.

H.1.2 Accuracy

Similarly, task accuracy has traditionally been a mainstay of human-computer interaction research. There are two forms of accuracy measured in both the motion and naturalness experiments and in the spatialisation-based search tool experiment. In the former, a participant's accuracy score is based on a binary correct/incorrect comparison, between the searcher's selection and the experimenter's definitive answer set; namely, there is one and only one document icon - per trial - that meets the task statement criteria exactly. In contrast, in the latter experiment, accuracy is determined by a group-wise similarity measure, between a definitive gold-standard and a set of documents deemed relevant by the searcher.

Whilst there are a number of similarity metrics available to information retrieval researchers, traditionally, two measures - namely, Recall and Precision - have contributed foremost to such considerations of search quality. Conceptually, Recall and Precision are appropriate measures to use in an evaluation of search quality; that is, *how many results that are known to be relevant have turned up in this answer set and how many results in this answer set are in fact relevant?* However, the problem with two measures is exactly that - there are two - and while a perfect search tool will maximise both recall and precision, in practise, comparative evaluation of say, two systems, one of high precision and low recall, the other of low precision and high recall, is difficult to judge.

Moreover, Precision and Recall are influenced by label Bias - the tendency of a searcher to judge a document as relevant, when in fact it is perhaps not - and prevalence - the proportion of relevant documents found across the corpus (Powers, 2007). If a searcher is generally more positive in relevance judgement i.e. biased, or the corpus contains a small number of relevant cases i.e. limited prevalence, then the rate of false positive answering will naturally inflate leading to a reduction in precision. This is an important point as earlier investigations into crowd-source judgements have shown that searchers are typically more positive than expert judges (Blanco et al., 2011; P. Bailey et al., 2008). In dealing with bias and prevalence, Bookmaker Informedness defined as $Recall - InverseRecall - 1 = \frac{(Recall - Bias)}{(1 - Prevalence)}$ reflects the underlying performance, irrespective of the innate characteristics of the corpus or the judge.

Powers (Powers, 2003) likens Informedness to a measure of a gambler's betting success on a two-horse race in which the gambler wins when picking the winner or the loser of the race and loses when the horse picked to win loses - or alternatively, when the horse picked to lose, in fact wins. In either case of losing, the cost of a loss is the same regardless of the prediction outcome - namely, losing the value of the bet tendered - and in either case of correctly picking the winner or the loser, the gain is the same - namely, receiving two times the initial wager. This contrasts with Recall and Precision, which offer no benefit to a searcher correctly identifying irrelevant documents.

Since random prediction for a session of decisions would assign each of the four possible outcomes in equal proportion, but collapse into two main outcomes - namely, win or lose for the binary case - a gambler would expect to break even. In 50% of cases they would win and 50% of cases they would lose, merely only wasting the day away. However, if some 'edge' or information were available to the gambler, that could be utilised to make it easier to predict the winner, then the expected number of wins versus losses may turn in favour of the gambler, leading to the gambler being *ahead* at the end of the session.

A parallel is drawn between the gambling scenario and the relevance assessments of documents in an information retrieval task. In the experiment of Chapter 6, the searcher and experiment participant represents the punter who is predicting the relevance of

documents - either relevant or irrelevant - and the desired outcome for the searcher, a pot of relevant documents that will satisfy his/her information need. The edge offered to the participant is the provision of information scent via the interface that permits the recognition of relevant results. Thus, interface configurations that offer superior information scent and organisation will likely dominate interface configurations that do not.

Bookmaker Informedness offers a quantifiable interpretation of the fidelity of the searcher's information made available through the search tool. A score of 0.0 indicates that the searcher's performance is that expected by chance - i.e. random guessing. In contrast, a score of 1.0 is achieved by relevance judgements made by a searcher who is presented with perfect information by the tool while a score of -1.0 is achieved by relevance judgements made with exactly imperfect information. Using imperfect information to make incorrect judgements is not guessing, rather, imperfect information is purely spurious information over which participants formulate incorrect conclusions.

Notwithstanding the theoretical benefits of Bookmaker Informedness, an information retrieval application of Informedness is complicated by the practicality of recording the evaluation of irrelevant candidates. Since an interaction with a search engine typically involves visually scanning document surrogates, there are no signals inherent in the document surrogate review process from which we may infer *all* irrelevant predictions. Accordingly, much of the false irrelevant and true relevant information is lost; this contrasts with a document full-text view event that may be captured by automated means e.g. by event listeners. As a consequence, it is unknown as to how each visually scanned document would be rated: was it rated correctly as irrelevant, thereby - potentially - increasing Informedness, or, was it incorrectly rated as irrelevant, thereby - potentially - decreasing Informedness? Clearly, technology can assist to an extent, through the aid of eye-gaze tracking, or in a primitive sense, by way of buttons on each document surrogate; and furthermore, experiment design may also introduce think-aloud components into data recording processes, in which participants call out their relevance ratings in sequence.

Tab. H.1: Objective performance and behavioural metrics in use across experimental chapters

Chapter:Experiment	Measure	Metric Description	Range
Chapter 3:Motion	Preparation Time	Time between start of trial and click of 'Ready' button	0-∞
	Answer Time	Time between click of 'Ready' button and click of correct answer	0-∞
	Attempts	A count of stimuli candidates submitted as answer	[1-3]
Chapter 4:Natural	Time	Time between start of trial and click of 'Next Task' button	0-∞
	Pop-ups Triggered	Count of pop-up windows triggered by mouse-hover event	0-∞
	Accuracy	Count of answer candidates deemed to exactly match task statement; pre-existing answer set generated	[0-12]
Chapter 6: Spatialisation Experiment	Time	Time (in Milliseconds) between start of trial and click of 'Next Task' button	0-∞
	Accuracy	Bookmaker score incorporating participant's relevance rating of selected documents and definitive relevance or gold standard	
	Documents opened	Count of documents opened - and presumably read - by participant	1-∞
	Multi Pop-up Trial Proportion	Proportion of trial time spend with multi-pop-up facility active	0-100%
	Projection Rotations	Count of rotations or updates of the spatialisation projection dimensions	0-∞
	Ranked-list re-sort Actions	Count of keyword and phrase selection events that trigger re-sorting of the ranked-list	0-∞
	Ranked-list re-sort vector length	Count of unique re-sort keywords selected	1-∞

H.2 Statistical Techniques

A table of statistical techniques is depicted below in Table H.2; a brief outline of each technique is offered. This table lists the technique's name, the data suited to the test, a set of statistical assumptions that must be met in order for the statistic to be valid, an interpretation of the outcome, and the relevant chapter and purpose for which this technique is applied. Note that the statistical assumptions column lists assumptions in addition to the data being normally distributed, having independence between observations, and the data sample being randomly selected.

H.2.1 Significance Testing

For inferential statistics utilised in the experimental work of this thesis, including the T-Test, Analysis of Variance ANOVA, and Chi Squared statistic, the production of a p-value or the probability, computed assuming a true null hypothesis, of observing a mean value at least as large as the one observed (Moore and McCabe, 1998), provides statistical evidence to reject the null hypothesis; namely, that the independent variables under manipulation, have no impact on the measured dependent variables.

The threshold at which sufficient statistical evidence is available to reject the null hypothesis, is denoted by the minimum significance level α and represents the maximum permitted rate of occurrence of a type I error or false positive. Frequently, a level of $\alpha = 0.05$ is adopted; accordingly, p-values below 0.05, generated as a result the experimental work in this thesis, will be deemed to be sufficient evidence on which to reject a null hypothesis.

H.2.2 Multiple Comparisons

A first stage of statistical testing seeks to establish an effect of the independent variable(s) on dependent variables or metrics. A subsequent stage of statistical testing seeks to establish the effect of each level of the independent variable(s) deemed statistically significant in the first round of testing.

Yet with many pair-wise comparisons taking place, each with an $\alpha=0.05$, the overall chance - the family wise error rate - of a Type 1 Error is increasingly inflated. Specifically, this inflation is given by $1 - (1 - \alpha)^{0.05}$. Accordingly, pair-wise comparisons between 8 groups, resulting in $\frac{8(8-1)}{2} = 28$ separate tests, the overall false positive rate expected by chance has blown out to $1 - (1 - 0.05)^8 = 0.34$, well above the level deemed low enough for statistical significance.

To counter the effect of inflated false positive rate under multiple comparisons, Chapter 3 adopts Tukey's Honestly Significant Difference (HSD) which calculates a Tukey statistic for two means under comparison which is then evaluated for significance

- much in the same fashion as t-tests are conducted. Tukey's HSD is considered to be a conservative post-hoc test, thereby limiting the chance of if a false positive (Dowdy, Weardon, and Chilko, 2004); yet, conservative methods also run the risk that a real effect is missed by the analysis. In contrast, the Bonferroni correction - mentioned in Chapter 6 is an altogether more conservative metric. Bonferroni ensures that the overall rate of Type 1 error is maintained at the 0.05 level for the defined number of comparisons (Moore and McCabe, 1998); mitigating against type 1 error is achieved by dividing the acceptable rate of Type 1 error e.g. $\alpha = 0.05$ by the number of tests made.

Lastly, it is broadly apparent - and noted elsewhere - that the choice of approach to post hoc testing is a matter of fashion or personal taste - (Hilton and Armstrong, 2006) and is evidenced in often conflicting recommendations (e.g. Dowdy, Weardon, and Chilko, 2004; Ruxton and Beauchamp, 2008). Moreover, it is broadly apparent that this area of statistics - namely, that concerned with the process of selecting a threshold of statistical significance for the *family wise error rate* - is undergoing theoretical evolution, with some researchers (e.g. Nakagawa, 2004) citing clear consequences for the quality of research in general. Nonetheless, it renders the selection process for an appropriate statistical approach markedly unsavoury in the least. Such a situation is acute, in the case of the production of and presentation of an ensemble of exploratory research in which it is challenging to determine the greater err - to overstate or understate the effect of a human computer interface configuration. In the case of the present research, a more conservative statistical approach has been preferred in an effort to stimulate further research into effects that are more pronounced, foremost.

H.2.3 Use of Error Bars

All histogram figures in experimental chapters denote mean values and error bars. At a glance, a histogram - with confidence intervals - of the data under examination offers much information. While not a replacement for thorough statistical analysis, such graphs depict the comparative differences between group means, an interval on which we can be 95% confident that the true value of the mean is actually situated, and a rough estimate of the potential for statistically significant differences between means.

Error bars provide an appropriate indication of the potential for statistical significance as examined by Belia et al. (2005). When depicting standard error, error bars indicate values, $\pm \frac{SD}{\sqrt{N}}$ of the mean value. Error bars not only give an indication of the potential for significance, but the width of the bars indicates the precision of the domain of the true mean value (Cumming, Fidler, and Vaux, 2007). For two likely statistically significant means at the $p=0.05$ level, the error bars must be separated by a gap of about half the width of the error bars.

Similarly, two confidence intervals - calculated by $\pm t_x * \frac{SD}{\sqrt{N}}$ with t_x denoting the

t-statistic with degrees of freedom $(n - 1)$ and $x = 1 - 0.05 = 0.95$ denoting the chosen level of confidence (Belia et al., 2005) - need only overlap by a quarter of the average length of both intervals in order to illustrate a statistically significant result at the $p=0.05$ level. Accordingly, two non-overlapping confidence intervals would indicate a markedly significant difference between means.

In either the case of standard error or confidence intervals, the use of error bars to denote significant differences is contingent on two main rules. Firstly, the size of each group must exceed 10 samples. Secondly, the larger interval under consideration must not exceed the smaller by a factor of two (Belia et al., 2005).

Finally, Confidence Intervals, like the α level in post hoc testing - discussed in the previous subsection - are similarly adjustable to counter the inflated chance of a true mean value situating outside of its Confidence Interval. In the simplest sense, this is achieved by manipulating the chosen level of confidence i.e. $x = 1 - \frac{0.05}{n}$ for n post hoc tests, in such a way as to widen each single interval and to close the gap between any two confidence intervals under comparison. Beyond the simplest sense, more sophisticated techniques are advocated (e.g. Ludbrook, 2000); these stipulate a systematic choice of denominator value, e.g. the HolmBonferroni method that selects the smallest ranking index k of ordered p-values that exceed $\frac{\alpha}{n+1-k}$.

Tab. H.2: Statistical techniques in use across experimental chapters

Technique	Data Suited	Additional Assumptions	Interpretation of Results	Relevant Chapter(s): Variables
Chi Squared	Frequencies of nominal data	Nominal measurement; mutual exclusivity of variables or groups; a minimum of five in each cell of the contingency table	Are the observed frequencies greater or less than frequencies expected by chance (unlike Fisher's Exact Test)	Chapter 6: use of ranked-list re-sort facility
Fisher's Exact Test	Frequencies of nominal data	nominal measurement;	differences in proportion of groups with each outcome (Freeman and M. Campbell, 2007)	Chapter 4: Chapter 4: answer accuracy between naturalness groups
Pearson's R	Continuous	Linearity of relationship	Strength of relationship between two variables	Chapter 6: agreement between crowd and expert relevance judgements
Spearman Rank Coefficient	Ranked continuous data	Monotonic relationship	Strength of relationship between rankings of two variables	Chapter 5: similarity in order of points about a central location in two spatial areas
Paired T-Test	Continuous	Within subjects measurement	Are the sample means significantly different to each other	Chapter 6: rotation configurations on left and right side controls
Analysis of Variance ANOVA	Continuous	Homogeneity of group variances	Are the sample means significantly different to each other	Chapter 3: Preparation Time, Answer Time, Attempts; Chapter 4: Time, Pop-ups Triggered; Chapter 6: Time, Bookmaker, Documents Opened, Rotation Configurations, Multi Pop-up Trial Proportion, Ranked-list re-sort actions; Ranked-list re-sort vector length

I. SELECT PUBLICATIONS

This appendix lists refereed conference and journal papers and a book chapter that have wholly or partially originated out of my research as a PhD candidate.

- Anderson, T.A., Chen, Z., Wen, Y., Milne, M.K., Atyabi, A., Treharne, K., Matsumoto, T., Jia, X., Luerssen, M.H., Lewis, T.W., et al. (2012). Thinking Head MulSeMedia: A Storytelling Environment for Embodied Language Learning. In *Multiple Sensorial Media Advances and Applications: New Developments in MulSeMedia*. Hershey, Pennsylvania: IGI Global, pp. 182-203.
- Pfitzner, D.M., Treharne, K., and Powers, D.M.W. (2008). User Keyword Preference: the Nwords and Rwords Experiments. *International Journal of Internet Protocol Technology*, vol. 3, no, 3, pp. 149-158.
- Powers, D.M.W, Luerssen, M.H., Lewis, T.W., Leibbrandt, R.E., Milne, M.K., Pashalis, J., and Treharne, K. (2010). MANA for the Ageing. *Proceedings of the 2010 Workshop on Companionable Dialogue Systems, ACL 2010*, pp. 7-12.
- Treharne, K., and Powers, D.M.W., 2009. Search Engine Result Visualisation: challenges and opportunities. *Proceedings of international symposium on web visualization*, pp. 633-638.
- Treharne, K., Pfitzner, D.M., Leibbrandt, R.E., and Powers, D.M.W. (2008). A lean online approach to human factors research. *Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments - PETRA'08*, pp. 57.
- Leibbrandt, R.E., Luerssen, M.H., Matsumoto, T., Treharne, K., Lewis, T.W., Santi, M.L., and Powers, D.M.W. (2008). An immersive game-like teaching environment with simulated teacher and hybrid world. *Animation, multimedia, IPTV and edutainment: proceedings of CGAT'08*, pp. 215-222.
- Atyabi, A., Anderson, T.A., Treharne, K., and Powers, D.M.W., (2011). Magician Simulator. *Eleventh International Conference on Control, Automation, Robotics and Vision -ICARCV'10*.
- Luerssen, M.H., Leibbrandt, R.E., Lewis, T.W., Pashalis, J., Treharne, K., Pfitzner, D.M. and, Powers, D.M.W. (2009). MANA - An Embodied Calendar for the Aged, *Thinking Systems Joint Symposium*, pp. 127-127.

- Treharne, K., Powers, D.M.W., and Leibbrandt, R. (2012). Optimising Visual and Textual in Search User Interfaces. Proceedings of the 24th Australian Computer-Human Interaction Conference - OzCHI'12. pp. 616-619.
- Atyabi, A., Powers, D.M.W., Anderson, T.F., Treharne, K. and Leibbrandt, R. (2013). Magician Simulator: From Simulation to Robot Teaming, Journal of Next Generation Information Technology, (in press).

BIBLIOGRAPHY

- Adelson, E. (1995). *The Checker Shadow Illusion*. URL: http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html (visited on 03/2013).
- Agrawal, R. et al. (Feb. 2009). “Diversifying search results”. In: *Proceedings 2nd ACM International Conference on Web Search and Data Mining - WSIDM'09*. Barcelona, Spain, pp. 5–14.
- Akhavi, M., M. Rahmati, and N. Amini (Aug. 2007). “3D Visualization of Hierarchical Clustered Web Search Results”. In: *Computer Graphics, Imaging and Visualisation - CGIV'07*. Bangkok, Thailand, pp. 331–446.
- Alhenshiri, A., C. Watters, and M. Shepherd (Jan. 2011). “User Behaviour during Web Search as Part of Information Gathering”. In: *Proceedings of the 44th Hawaii International Conference on System Sciences - HICSS'011*. Hawaii, USA.
- Alonso, O. and S. Mizzaro (July 2009). “Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment”. In: *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'09*. Massachusetts, USA, pp. 15–16.
- Alonso, O., D. Rose, and B. Stewart (2008). “Crowdsourcing for Relevance Evaluation”. In: 42.2, pp. 9–15.
- Andrews, K. et al. (2002). “The InfoSky visual explorer: exploiting hierarchical structure and document similarities”. In: *Information Visualization 1.3-4*, pp. 166–181.
- Aula, A. (Sept. 2004). “Enhancing the Readability of Search Result Summaries”. In: *Proceedings of the 18th British HCI Group Annual Conference - HCI'2004*. Leeds, England, pp. 1–4.
- Azar, B. (2000). “Online Experiments: Ethically fair or foul?” In: *Monitor on Psychology* 31.4. URL: <http://www.apa.org/monitor/apr00/fairorfoul.aspx> (visited on 03/2013).
- Baecker, R. and I. Small (1990). “Animation at the Interface”. In: ed. by B. Laurel. *The Art of Human-Computer Interface Design*. Addison-Wesley.
- Baecker, R., I. Small, and R. Mander (Apr. 1991). “Bringing Icons to Life”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'91*. New Orleans, USA, pp. 1–6.
- Bailey, B. and J. Konstan (2006). “On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state”. In: *Computers in Human Behaviour* 22.4.
- Bailey, P. et al. (July 2008). “Relevance Assessment: Are Judges Exchangable and Does it Matter?” In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'08*. Singapore, pp. 667–674.
- Baird, B. and J. Zollinger (2007). “Method and Systems for User Activated Automated Searching”. US7308439. HyperThink LLC.

- Baker, S. (2013). *Helping Computers Understand Language*, Google Official Blog. URL: <http://googleblog.blogspot.com.au/2010/01/helping-computers-understand-language.html> (visited on 03/2013).
- Balatsoukas, P., A. O'Brien, and A. Morris (2010). "Design factors affecting relevance judgment behaviour in the context of metadata surrogates". In: *Journal of Information Science* 36.6, pp. 780–797.
- Balatsoukas, P. and I. Ruthven (2010). "The use of relevance criteria during predictive judgment: an eye tracking approach". In: *Proceedings of Annual Meeting of the Association for Information Science and Technology - ASIS&T'10* 47.1, pp. 1–10.
- Barr, P., R. Biddle, and J. Noble (2002). *A Taxonomy of User-Interface Metaphors*. Tech. rep. CS-TR-02-11. Wellington, New Zealand: Victoria University of Wellington.
- Bartram, L. (1997). *Perceptual and Interpretative Properties of Motion for Information Visualization*. Tech. rep. CMPT-TR:1997-15. British Columbia, Canada: Simon Fraser University.
- (Nov. 1998). "Perceptual and Interpretative Properties of Motion for Information Visualization". In: *Proceedings of New Paradigms in Information Visualization - NPIV'97*. Nevada, USA, pp. 3–7.
- (2001). "Enhancing Visualization with Motion". Ph.D Thesis. British Columbia, Canada: Simon Fraser University.
- Bartram, L. and C. Ware (2002). "Brushing and Filtering with Motion". In: *Journal of Information Visualization* 1.1, pp. 66–79.
- Bartram, L., C. Ware, and T. Calvert (July 2001). "Moving icons: detection and distraction". In: *Proceedings of the 8th International Conference on Human-Computer Interaction - INTERACT'01*. Tokyo, Japan, pp. 157–165.
- (2003). "Moticons: detection, distraction and task". In: *International Journal Human-Computer Studies* 58.5, pp. 515–545.
- Bates, M. (1979). "Information Search Tactics". In: *Journal of the American Society for Information Science* 30.5, pp. 205–214.
- (1989). "The design of browsing and berry picking techniques for the online search interface". In: *Online Review* 13.5, pp. 407–431.
- Baudisch, P. and C. Gutwin (Apr. 2004). "Multiblending: displaying overlapping windows simultaneously without the drawbacks of alpha blending". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'04*. Vienna, Austria, pp. 367–374.
- Baudisch, P., D. Tan, et al. (Oct. 2006). "Phosphor: Explaining Transitions in the User Interface: by Using Afterglow Effects". In: *Proceedings of User Interface Software and Technology - UIST'06*. Montreux, Switzerland, pp. 169–178.
- Baumgärtner, S. et al. (Apr. 2007). "2D Meets 3D: A Human-Centered Interface for Visual Data Exploration". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems: Work-in-Progress - CHI'07*. San Jose, USA, pp. 2273–2278.
- Bederson, B. and C. Klein (Apr. 2005). "Benefits of Animated Scrolling". In: *Extended Abstracts on Human Factors in Computing Systems - CHI EA '05*. Portland, USA, pp. 1965–1968.
- Bekos, M.A. et al. (2007). "Boundary Labelling: Models and Efficient Algorithms for Rectangular Maps". In: *Computational Geometry* 36.3, pp. 215–236.
- Belia, S. et al. (2005). "Researchers Misunderstand Confidence Intervals and Standard Error Bars". In: *Psychological Methods* 10.4, pp. 389–396.

- Belkin, N., R. Oddy, and H. Brooks (1982). "ASK for Information Retrieval: Parts I and II". In: *Journal of Documentation* 38.2,3.
- Benford, S. et al. (1999). "Three Dimensional Visualization of the World Wide Web". In: *ACM Computer Surveys* 31.25.
- Benking, H. and A. Judge (Sept. 1994). "Design Considerations for Spatial Metaphors: reflections on the evolution of viewpoint transportation systems". In: *European Conference on Hypermedia Technology - ECHT'94*. Edinburgh, Scotland.
- Bennet, K. (1993). "Encoding Apparent Motion in Animated Mimic Displays". In: *Human Factors* 35.4, pp. 673–691.
- Benyon, D. (2001). "The new HCI? Navigation of information space". In: *Knowledge-Based Systems* 14.8, pp. 425–430.
- Benyon, D. and K. Höök (July 1997). "Navigation in Information Spaces: Supporting the Individual". In: *Proceedings of the IFIP International Conference on Human-Computer Interaction - INTERACT'97*. Sydney, Australia, pp. 39–46.
- Berendonck, C. and T. Jacobs (Feb. 2003). "Bubbleworld A New Visual Information Retrieval Technique". In: *Proceedings of Australasian Symposium on Information Visualisation*. Adelaide, Australia, pp. 47–56.
- Bertin, J. (2011). *Semiology of Graphics: Diagrams, Networks, and Maps*. ESRI Press.
- Bétrancourt, M. and S. Caro (1998). "Intégrer des Informations en Escamots dans les Textes Techniques : quels Effets sur les Processus Cognitifs / Integrating pop-pop window information in technical texts : what effects on cognitive process?" In: *Hypertextes et hypermédias*, pp. 157–173.
- Bladh, T. (2006). "A Taxonomy and Survey of User Interface Animation / Towards an Understanding of Dynamics in Information Visualization". Licentiate Thesis. Luleå, Sweden: LuleåUniversity of Technology.
- Blanco, R. et al. (July 2011). "Repeatable and Reliable Search System Evaluation using Crowdsourcing". In: *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'11*. Beijing, China, pp. 923–932.
- Blei, D., Y. Ng, and M. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022.
- Bodner, R. and I. MacKenzie (Nov. 1997). "Using animated icons to present complex tasks". In: *Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative Research - CASCON'97*. Toronto, Canada, pp. 281–291.
- Bonnel, N., A. Cotarmanac'h, and A. Morin (July 2005). "Meaning Metaphor for Visualizing Search Results". In: *Proceedings of the 9th International Conference on Information Visualisation - IV'05*. London, England, pp. 467–472.
- Bonnel, N., V. Lemaire, et al. (Feb. 2006). "Effective Organization and Visualization of Web Search Results". In: *Proceedings of 24th European International Multi-Conference on Internet and Multimedia Systems and Applications - IASTED*. Innsbruck, Austria, pp. 209–216.
- Borg, I. and P. Geonen (2005). *Modern Multidimensional Scaling: Theory and Applications*. 2nd ed. New York, USA: Springer.
- Borlund, P. (2005). "The Concept of Relevance in IR". In: *Journal of the American Society for Information Science and Technology* 53.10, pp. 913–925.
- Börner, K., C. Chen, and K. Boyack (2003). "Visualizing Knowledge Domains". In: *Annual Review of Information Science and Technology* 37.1, pp. 179–255.

- Brath, R. (Oct. 1997). "Concept Demonstration Metrics for Effective Information Visualization". In: *Proceedings IEEE Symposium on Information Visualization - InfoVis'97*. Toronto, Canada, pp. 108–111.
- (July 2009). "The Many Dimensions of Shape". In: *Proceedings of 13th International Conference on Information Visualisation - IV'09*. Keynote Lecture. Barcelona, Spain.
- Breimer, E., J. Cotler, and R. Yoder (2012). "Video vs. Text for Lab Instruction and Concept Learning". In: *Journal of Computing Sciences in Colleges* 27.6, pp. 42–48.
- Brewer, C. (1999). "Color use guidelines for data representation". In: *Proceedings of the Section on Statistical Graphics*. American Statistical Association, pp. 55–60.
- Broder, A. (2002). "A Taxonomy of Web Search". In: *SIGIR Forum* 36.2, pp. 3–10.
- Buja, A. et al. (2008). "Data Visualization with Multidimensional Scaling". In: *Journal of Computational and Graphical Statistics* 17.2, pp. 444–472.
- Butavicius, M. and M. Lee (2007). "An empirical evaluation of four data visualization techniques for displaying short news text similarities". In: *International Journal of Human-Computer Studies* 65.11, pp. 931–944.
- Byrne, M. (Apr. 1993). "Using Icons to Find Documents: Simplicity is Critical". In: *Conference on Human Factors in Computing Systems - INTERCHI'93*. Amsterdam, The Netherlands, pp. 446–453.
- (2002). "Reading vertical text: Rotated vs. Marquee". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 46.17, pp. 1633–1635.
- Card, S. and J. Mackinlay (Oct. 1997). "The Structure of the Information Visualisation Design Space". In: *Proceedings of IEEE Symposium on Information Visualisation - InfoVis'97*. Phoenix, USA, pp. 92–99.
- Carey, M., F. Kriwaczek, and S. Ruger (Nov. 2000). "A Visualization Interface for Document Searching and Browsing". In: *Proceedings of Workshop on New Paradigms in Information Visualization and Manipulation*. Washington DC, USA, pp. 24–28.
- Caro, S. (July 1997). "Pop-Up Windows and Information Retrieval". In: *Proceedings of 6th International Conference on Human-Computer Interaction - INTERACT'97*. Sydney, Australia, pp. 573–574.
- Carpineto, C., S. Osinski, et al. (2009). "A Survey of Web Clustering Engines". In: *ACM Computing Surveys* 41.3, Article 17.
- Carpineto, C. and G. Romano (2012). "A Survey of Automatic Query Expansion in Information Retrieval". In: *ACM Computing Surveys* 44.1.
- Carroll, J. (1997). "Human-Computer Interaction: Psychology as a Science of Design". In: *Annual Review of Psychology* 48.4, pp. 61–83.
- Carswell, M. and C. Wickens (1987). "Information Integration and the Object Display An Interaction of task demands and display superiority". In: *Ergonomics* 30.3, pp. 511–527.
- Chabot, C. (2009). "Demystifying Visual Analytics". In: *IEEE Computer Graphics and Applications* 29.2, pp. 84–87.
- Chaffin, R. and D. Herrmann (1984). "The similarity and diversity of semantic relations". In: *Memory and Cognition* 12.2, pp. 134–141.
- Chaparro, A. et al. (1999). "The impact of age on computer input device use: Psychophysical and physiological measures". In: *International Journal of Industrial Ergonomics* 24.5, pp. 503–513.
- Chen, C. (2005). "Top 10 Unsolved Information Visualization Problems". In: *IEEE Computer Graphics and Applications* 25.4, pp. 12–16.

- Chen, C. and K. Börner (2002). "Top Ten Problems in Visual Interfaces to Digital Libraries". In: *Visual Interfaces to Digital Libraries*. Ed. by C. Chen and K. Börner. Vol. LNCS 2539. Berlin, Germany: Springer, pp. 226–231.
- Chen, C. and Y. Yu (2000). "Empirical studies of information visualization: a meta-analysis". In: *International Journal of Human-Computer Studies* 53.5, pp. 851–866.
- Cho, E. and S. Myaeng (Apr. 2000). "Visualization of Retrieval Results using DART". In: *Proceedings of Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO*. Paris, France, pp. 1434–1439.
- Choi, H. and S. Johnson (2005). "The Effect of Context-Based Video Instruction on Learning and Motivation in Online Courses". In: *The American Journal of Distance Education* 19.4, pp. 215–227.
- Christ, R. (1975). "Review and analysis of color coding research for visual displays". In: *Human Factors* 17.6, pp. 542–570.
- Chung, W., H. Chen, and J. Nunamaker (2005). "A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration". In: *Journal of Management Information Systems* 21.4, pp. 57–84.
- Church, K. and B. Smyth (Feb. 2009). "Understanding the Intent behind Mobile Information Needs". In: *Proceedings of International Conference on Intelligent User Interfaces - IUI'09*. Sanibel Island, USA, pp. 247–256.
- Church, K., B. Smyth, et al. (Sept. 2008). "A large scale study of European mobile search behaviour". In: *Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI*. Amsterdam, The Netherlands, pp. 13–22.
- Clarkson, E., K. Desai, and J. Foley (n.d.). "ResultMaps: Visualization for Search Interfaces". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6, pp. 1057–1064.
- Cleveland, W. and R. McGill (1984). "Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods". In: *Journal of the American Statistical Association* 79.387, pp. 531–554.
- Clough, P. et al. (2013). "Examining the Limits of Crowdsourcing for Relevance Assessment". In: 17.4, pp. 32–38.
- Cockburn, A. (Jan. 2004). "Revisiting 2D vs. 3D Implications on Spatial Memory". In: *Proceedings of 5th Australasian User Interface Conference - AUIC'2004*. Dunedin, New Zealand, pp. 25–32.
- Cockburn, A. and B. McKenzie (2001). "3D or not 3D? Evaluating the Effect of the Third Dimension in a Document Management System". In: *CHI Letters* 3.1, pp. 434–441.
- Coren, S., C. Porac, and L. Ward (1979). *Motion, Sensation and Perception*. 1st ed. New York, USA: Academic Press.
- Cribbin, T. and C. Chen (July 2001). "Exploring Cognitive Issues in Visual Information Retrieval". In: *Proceedings of 8th Conference of Human-Computer Interaction - INTERACT'01*. Tokyo, Japan, pp. 166–173.
- Cumming, G., F. Fidler, and D. Vaux (2007). "Error bars in experimental biology". In: *Journal of Cell Biology* 177.1, pp. 7–11.
- deAngeli, A., A. Sutcliffe, and J. Hartmann (June 2006). "Interaction, Usability and Aesthetics: What Influences Users? Preferences?" In: *Proceedings of the 6th Conference on Designing Interactive Systems - DIS'06*. University Park, USA, pp. 271–280.

- Deerwester, S. et al. (1990). "Indexing by Latent Semantic Analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.
- Dehaene, S., S. Bossini, and P. Giraux (1993). "The mental representation of parity and number magnitude". In: *Journal of Experimental Psychology: General* 122.3, pp. 371–396.
- Der, G. and I. Deary (2006). "Age and Sex Differences in Reaction Time in Adulthood: Results from the United Kingdom Health and Lifestyle Survey". In: *Psychology and Aging* 21.1, pp. 62–73.
- diGiacomo, E. et al. (Mar. 2008). "WhatsOnWeb+: An Enhanced Visual Search and Clustering Engine". In: *IEEE Pacific Visualization Symposium - PacificViz'08*. Kyoto, Japan, pp. 167–172.
- Dong, O. Hoeber Y. (2008). "Evaluating the Effectiveness of Term Frequency Histograms for Supporting Interactive Web Search Tasks". In: *Proceedings of the 7th ACM Conference on Designing Interactive Systems - DIS'08*. Cape Town, South Africa, pp. 360–368.
- Dormal, V. and M. Pesenti (2007). "Numerosity-length interference: a Stroop experiment". In: *Experimental Psychology* 54.4, pp. 289–297.
- Doumont, J. (2002). "Magical Numbers: the seven-plus-or-minus-two myth". In: *IEEE Transactions on Professional Communication* 45.2, pp. 123–127.
- Dourish, P. and M. Chalmers (Aug. 1994). "Running Out of Space: Models of Information Navigation". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'04*. Glasgow, Scotland, pp. 23–26.
- Dowdy, S., S. Weardon, and D. Chilko (2004). In: *Statistics for Research*. Ed. by W. Shewhart and S. Wilks. 3rd ed. New Jersey, USA: Wiley-Interscience.
- Drori, O. (2000). "The Benefits of Displaying Additional Internal Document Information on Textual Database Search Result Lists". In: *Research and Advanced Technology for Digital Libraries LNCS 1923*, pp. 69–82.
- Drori, O. and N. Alon (2003). "Using Documents Classification for Displaying Search Results List". In: *Journal of Information Science* 29.2, pp. 97–106.
- Dubroy, P. and R. Balakrishnan (Apr. 2010). "A study of tabbed browsing among Mozilla Firefox users". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'10*. Atlanta, USA, pp. 673–682.
- Duncan, J. and G. Humphreys (1989). "Visual Search and Stimulus Similarity". In: *Psychological Review* 96.3, pp. 433–458.
- Dykes, J., J. Wood, and A. Slingsby (2010). "Rethinking Map Legends with Visualization". In: *IEEE Transactions on Visualisation and Computer Graphics* 16.2, pp. 890–899.
- Eckersley, P. (July 2010). "How Unique is Your Web Browser". In: *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*. Berlin, Germany, pp. 1–18.
- Elias, J., W. Westerman, and M. Haggerty (2007). "Multi-touch gesture dictionary". US20070177803. Inc. Apple Computer.
- Ellis, G. and A. Dix (2007). "A Taxonomy for Cluster Reduction". In: *Transactions on Computer Graphics and Visualization* 13.6, pp. 1216–1223.
- Elmqvist, N., P. Dragicevic, and J. Fekete (2008). "Rolling the Dice: Multidimensional Visual Exploration using Scatter plot Matrix Navigation". In: *Transactions on Visualization and Computer Graphics* 14.6, pp. 1141–1148.

- Fabrikant, S., D. Montello, et al. (2004). "The Distance-Similarity Metaphor in Network Display Spatializations". In: *Cartography and Geographic Information Science* 31.4, pp. 237–252.
- Fabrikant, S., M. Ruocco, et al. (Sept. 2002). "The First Law of Cognitive Geography: Distance and Similarity in Semantic Space". In: *Proceedings of GIS Science*. Boulder, USA, pp. 31–33.
- Farrington, J. (2011). "Seven Plus or Minus Two". In: *Performance Improvement Quarterly* 23.4, pp. 113–116.
- Feinburg, J. (2010). "Wordle". In: *Beautiful Visualization - looking at data*. Ed. by J Steele and N Illinsky. O'Reilly Media.
- Few, S. (2008). *Practical Rules for Using Color in Charts*. URL: http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf (visited on 03/2013).
- Finkel, J., T. Grenager, and C. Manning (June 2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics - ACL 2005*. Ann Arbor, USA, pp. 363–370.
- Foenix-Riou, B. (2006). "When Search Engines Play at Maps: Visualization Technologies". In: *Online* 30.2, pp. 29–32.
- Forsell, C. and J. Johansson (May 2010). "A Heuristic Set for Evaluation in Information Visualization". In: *Proceedings of the Working Conference on Advanced Visual Interfaces - AVI'10*. Rome, Italy, pp. 199–206.
- Fortuna, B., M. Grobelnik, and D. Mladenic (2005). "Visualisation of a Text Corpus". In: *Informatica* 29.4, pp. 497–502.
- Fox, S. et al. (2005). "Evaluating Implicit Measures to Improve Web Search". In: *ACM Transaction on Information Systems* 23.2, pp. 147–168.
- Freeman, J. and M. Campbell (2007). "The analysis of categorical data: Fisher's exact test". In: 16.1, pp. 11–12.
- Frick, A., M. Bächtiger, and U. Reips (1999). *Financial incentives, personal information and drop-out rate in online studies*. Current Internet Science: trends, techniques, results. Zurich, Switzerland: Online Press.
- Furnas, G. et al. (1987). "The Vocabulary Problem in Human-System Communication". In: *Communications of the ACM* 30.11, pp. 964–971.
- Gabriel-Petit, P. (2006). *Color Theory for Digital Displays: A Quick Reference: Part 1*. URL: <http://www.uxmatters.com/mt/archives/2006/01/color-theory-for-digital-displays-a-quick-reference-part-1.php> (visited on 03/2013).
- Galitz, W. (2007). *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and User Interface Design*. John Wiley & Sons Inc.
- Gamberini, L. et al. (2008). "A Game a Day Keeps the Doctor Away: A Short Review of Computer Games in Mental Healthcare". In: *Journal of Cyber Therapy and Rehabilitation* 1.2, pp. 127–145.
- Garner, W. (1974). *The Processing of Information and Structure*. New Jersey, USA: Lawrence Erlbaum Associates.
- Gerken, J. et al. (2009). "Lessons learned from the design and evaluation of visual information-seeking systems". In: *International Journal of Digital Libraries* 10.2, pp. 49–66.
- Gershon, N. (Oct. 1992). "Visualization of fuzzy data using generalized animation". In: *IEEE 3rd Conference on Visualisation - Vis'92*. Amherst, USA, pp. 268–273.

- Gevers, W. and J. Lammertyn (2005). "The hunt for SNARC". In: *Psychology Science* 47.1, pp. 10–21.
- Google (2013). *Inside Search: Basic Search Help*. URL: <http://support.google.com/websearch/bin/answer.py?hl=en&answer=134479&from=35889&rd=2#stemming> (visited on 03/2013).
- Goulding, A. (2001). "Information Poverty or Overload?" In: *Journal of Librarianship and Information Science* 33.3, pp. 109–111.
- Gowda, K. and E. Diday (1991). "Symbolic clustering using a new dissimilarity measure". In: *Pattern Recognition* 24.6, pp. 567–578.
- Granitzer, M. et al. (June 2003). "WebRat: Supporting Agile Knowledge Retrieval through Dynamic, Incremental Clustering and Automatic Labelling of Web Search Result Sets". In: *Proceedings of the 12th IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises - WETICE'03*. Linz, Austria, pp. 296–301.
- Granka, L., T. Joachims, and G. Gay (July 2004). "Eye tracking Analysis of User Behaviour in WWW Search". In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'04*. Sheffield, England, pp. 478–479.
- Green, C. (1997). *Classics in the history of psychology: Laws of Organization in Perceptual Forms*. First published as *Untersuchungen zur Lehre von der Gestalt II*, *Psychologische Forschung*, vol. 4, pp. 301–350, translation in Ellis, W. 1938. A source book of Gestalt psychology, pp. 71–88, London: Routledge & Kegan Paul. URL: <http://psychclassics.yorku.ca/Wertheimer/Forms/forms.htm> (visited on 03/2013).
- Grewal, R. et al. (Nov. 2000). "A novel interface for representing search-engine results". In: *IEE Colloquium: Lost in the Web - Navigation on the Internet*. London, England, pages.
- Griffin, A. et al. (2006). "Comparison of Animated and Static Small-Multiple Maps for Visually Identifying Space-Time Clusters". In: *Annals of the Association of American Geographers* 94.4, pp. 740–753.
- Griffiths, M. (2002). "The educational benefits of videogames". In: *Education and Health* 20.3, pp. 47–51.
- Group, Algorithmics (2009). *MDSJ: Java Library for Multidimensional Scaling (Version 0.2)*. URL: <http://www.inf.uni-konstanz.de/algo/software/mdsj/> (visited on 03/2013).
- Gurr, C. (1998). "On the Isomorphism, or Lack of it, of Representations". In: *Visual Language Theory*. Ed. by In K. Marriott and B. Meyer, pp. 293–305.
- Haas, K. et al. (July 2011). "Enhanced Results for Web Search". In: *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'11*. Beijing, China, pp. 725–734.
- Han, J. and M. Kamber (2001). *Data Mining: Concepts and Techniques*. San Diego, USA: Academic Press.
- Hardin, L. (2002). "Problem-solving concepts and theories". In: *Journal of Veterinary Medicine* 30.3, pp. 226–229.
- Harrison, B., G. Kurtenbach, and K. Vicente (Nov. 1995). "An Experimental Evaluation of Transparent User Interface Tools and Information Content". In: *Proceedings of 8th ACM Symposium on User Interface Software and Technology - UIST'95*. Pittsburgh, USA, pp. 81–90.

- Harrison, C. et al. (May 2011). “Kineticons: Using Iconographic Motion in Graphical User Interface Design”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'11*. Vancouver, Canada, pp. 1999–2008.
- Harrower, M. and C. Brewer (2003). “ColorBrewer.org: An Online tool for Selecting Colour Schemes for Maps”. In: *The Cartographic Journal* 40.1, pp. 27–37.
- Hartmann, K. et al. (2005). “Metrics for Functional and Aesthetic Label Layouts”. In: vol. LNCS 3638. Smart Graphics, pp. 115–126.
- Hartson, R. (1998). “Human-Computer Interaction: Interdisciplinary roots and trends”. In: *The Journal of Systems and Software* 43.2, pp. 103–118.
- Harward, V. et al. (2008). “The iLab Share Architecture: A Web Services Infrastructure to Build Communities of Internet Accessible Laboratories”. In: *Proceedings of the IEEE* 96.6, pp. 931–950.
- Havre, S. et al. (Oct. 2001). “Interactive Visualization of Multiple Query Results”. In: *Proceedings IEEE Symposium on Information Visualization - InfoVis'01*. San Diego, USA, pp. 105–112.
- Healey, C., R. St Amant, and M. Elhaddad (Oct. 1999). “ViA: A Perceptual Visualization Assistant”. In: *Proceedings of 28th Advanced Imagery Pattern Recognition Workshop - AIPR'99*. Ed. by W.R. Oliver. 3D Visualization for Data Exploration and Decision Making. Washington DC, USA, pp. 2–11.
- Healey, C., K. Booth, and J. Enns (May 1993). “Harnessing Pre-attentive Processes for multivariate data visualization”. In: *Proceedings of the Graphics Interface Conference*. Toronto, Canada, pp. 107–117.
- Healey, C. and J. Enns (2011). “Attention and Visual Memory in Visualization and Computer Graphics”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.7, pp. 1170–1188.
- Hearst, M. (May 1995). “TileBars: Visualization of Term Distribution Information in Full Text Information Access”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'95*. Denver, USA, pp. 59–65.
- (1999). “User Interfaces and Visualization”. In: Baeza-Yates, R. and B. Ribeiro-Neto. *Modern Information Retrieval*. New York, USA: ACM Press. Chap. 10.
- (2006). “Clustering versus faceted categories for information exploration”. In: *Communications of the ACM* 49.4, pp. 59–61.
- (2009). *Search User Interfaces*. New York, USA: Cambridge University Press.
- Hearst, M. and J. Pedersen (Aug. 1996). “Re-examining the Cluster Hypothesis Scatter/Gather on Retrieval Results”. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'96*. Zurich, Switzerland, pp. 76–84.
- Hegarty, M. (2011). “The Cognitive Science of Visual-Spatial Displays: Implications for Design”. In: *Topics in Cognitive Science* 3.3, pp. 446–474.
- Helmholtz, H. (1852). “On the theory of compound colors”. In: *Philosophical Magazine* 4.28, pp. 519–534.
- Hemmje, M., C. Kunkel, and A. Willet (July 1994). “LyberWorld- A Visualization User Interface Supporting Full text Retrieval”. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'94*. Dublin, Ireland, pp. 249–259.
- Henrich, J., S. Heine, and A. Norenzayan (2010). “The weirdest people in the world?”. In: *Behavioral and Brain Sciences* 33.2-3, pp. 61–135.
- Hering, E. (1878). *Zür lehre vom lichtsinn*. Vienna, Austria: Gerold.

- Hilton, A. and R. Armstrong (2006). “Stat Note 6 - post hoc ANOVA tests”. In: pp. 34–36.
- Hoeber, O. (2012). “Human-Centred Web Search”. In: ed. by C. Jouis et al. *Next Generation Search Engines: Advanced Models for Information Retrieval*. IGI Global, pp. 217–238.
- Hoeber, O. and H. Liu (Aug. 2010). “Comparing Tag Clouds, Term Histograms and Term Lists for Enhancing Personalized Web Search”. In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Toronto, Canada, pp. 309–313.
- Hoeber, O. and X. Yang (July 2006). “The visual exploration of web search results using HotMap”. In: *Proceedings of the 10th International Conference on Information Visualisation = IV’10*. London, England, pp. 157–165.
- (2008). “Evaluating WordBars in Exploratory Web Search Scenarios”. In: *Information Processing and Management* 44.2, pp. 485–510.
- (2009). “HotMap: Supporting visual explorations of web search results”. In: 60.1, pp. 90–110.
- Hofmann, P. (2008). “Localising and Internationalising Graphics and Visual Information: Commentary”. In: *IEEE Transactions on Professional Communication* 50.2, pp. 91–92.
- (Nov. 2009). “Stimulating the icon-omy: new directions in minimising and visualising information”. In: *Proceedings of 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group of the Human Factors and Ergonomics Society of Australia - OZCHI’09*. Keynote Lecture. Melbourne, Australia.
- Holten, D. and J. vanWijk (Apr. 2009). “A User Study on Visualizing Directed Edges in Graphs”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI’09*. Boston, USA, pp. 2299–2308.
- Horn, C. (2007). “Natural metaphors for information visualization”. Masters Thesis. Boston, USA: Massachusetts College of Art.
- Hornbæk, K., B. Bederson, and C. Plaisant (2002). “Navigation Patterns and Usability of Zoomable User Interfaces with and without an Overview”. In: *ACM Transactions on Computer-Human Interaction* 9.4, pp. 362–389.
- Hornbæk, K. and M. Hertzum (2011). “The Notion of Overview in Information Visualization”. In: *International Journal of Human-Computer Studies* 60.7/8, pp. 509–525.
- Horton, J. and L. Chilton (June 2010). “The labor economics of paid crowdsourcing”. In: *Proceedings of the 11th ACM Conference on Electronic Commerce - EC’10*. Cambridge, USA.
- Hu, P., P. Ma, and P. Chau (1999). “Evaluation of user interface designs for information retrieval systems: a computer-based experiment”. In: *Decision Support Systems* 27.1-2, pp. 125–143.
- Hu, X. et al. (July 2003). “LSA: The first dimension and dimensional weighting”. In: *Proceedings of 25th Annual Conference of the Cognitive Society*. Boston, USA, pp. 1–8.
- Huang, J., T. Lin, and R. White (Feb. 2012). “No Search Result Left Behind: Branching Behavior with Browser Tabs”. In: *Proceedings 5th ACM International Conference on Web Search and Data Mining - WSDM’12*. Seattle, USA, pp. 203–212.
- Huber, D. and C. Healey (Oct. 2005). “Visualizing data with motion”. In: *IEEE Visualisation - VIS’05*. Minneapolis, USA, pp. 527–534.

- Hubmann-Haidvogel, A., A. Scharl, and A. Weichselbraun (2009). “Multiple Coordinated Views for Searching and Navigating Web Content Repositories”. In: *Information Sciences* 179.12, pp. 1813–1821.
- Hulleman, J. and E. McWilliams (2011). “Visual Search: Elements, cues and configurations: No Motion Filtering in Visual Search amongst Moving Items”. In: *Journal of Vision* 11.11, Article 1301.
- Humphrey, D. and A. Kramer (1997). “Age Differences in Visual Search for Feature, Conjunction and Triple-Conjunction Targets”. In: *Psychology and Aging* 12.4, pp. 704–717.
- Ishak, E. and S. Feiner (Oct. 2004). “Interacting with Hidden Content Using Content-Aware Free-Space Transparency”. In: *Proceedings of 17th ACM Symposium on User Interface Software and Technology - UIST'04*. Santa Fe, USA, pp. 189–192.
- Jain, A., M. Murty, and P. Flynn (1999). “Data Clustering: A Review”. In: *ACM Computing Surveys* 31.3, pp. 264–323.
- Jansen, B., D. Booth, and A. Spink (2008). “Determining the informational, navigational and transactional intent of Web queries”. In: *Information Processing and Management* 44.3, pp. 1251–1266.
- Johnson, D. and T. Jankun-Kelly (May 2008). “A Scalability Study of Web-Native Information Visualization”. In: *Proceedings of the Graphics Interface Conference*. Windsor, Canada, pp. 163–168.
- Jubis, R. (1991). “Effects of Color-Coding, Retention-Interval and Task on Time to Recognize Target-Updates”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 35.20, pp. 1462–1465.
- Julien, C., J. Leide, and F. Bouthillier (2008). “Controlled User Evaluations of Information Visualization Interfaces for Text Retrieval: Literature Review and Meta-Analysis”. In: *Journal of the American Society for Information Science and Technology* 56.6, pp. 1012–1024.
- Käki, M. and A. Aula (2008). “Controlling the complexity in comparing search user interfaces via user studies”. In: *Information Processing and Management* 44.1, pp. 82–91.
- Kearney, P. (June 2005). “Cognitive Calisthenics: Do FPS computer games enhance the player’s cognitive abilities?” In: *Proceedings of 2nd Digital Games Research Conference - Changing Views: Worlds in Play - DiGRA'05*. Vancouver, Canada.
- Keim, D. et al. (2008). “Visual Analytics: Scope and Challenges”. In: ed. by S.J. Simoff et al. Vol. LNCS 4404. Visual Data Mining. Springer, pp. 76–90.
- Keller, H. and S. Lee (2003). “Ethical Issues Surrounding Human Participants Research Using the Internet”. In: *Ethics & Behavior* 13.3, pp. 211–219.
- Kelly, D. (May 2008). “Query through selection: modelling and facilitating information seeking behaviour in mobile environments”. In: *Joint HCSNet EII Workshop on Interactive and Ubiquitous Information Access*. Keynote Lecture. Sydney, Australia.
- Kerne, A. and S. Smith (Aug. 2004). “The Information Discovery Framework,” in: *Proceedings of Conference Designing Interactive Systems: Processes, Practices, Methods and Techniques - DIS'04*. Cambridge, USA, pp. 357–360.
- Kittur, A., E. Chi, and B. Suh (Apr. 2008). “CrowdSourcing User Studies with Mechanical Turk”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'08*. Florence, Italy, pp. 453–456.
- Klemmer, E. and F. Frick (1953). “Assimilation of information from dot and matrix patterns”. In: *Journal of Experimental Psychology* 45.1, pp. 15–19.

- Kobayashi, T. et al. (Feb. 2006). "Information Gathering Support Interface by the Overview Presentation of Web Search Results". In: *Proceedings of the Asia-Pacific Symposium on Information Visualisation*. Vol. 60. Tokyo, Japan, pp. 103–108.
- Koshman, S. (2005). "Testing User Interaction with a Prototype Visualization-Based Information Retrieval System". In: *Journal of the American Society for Information Science and Technology* 56.8, pp. 824–833.
- (2006). "Visualization-based information retrieval on the web". In: *Library and Information Science Research* 28.2, pp. 192–207.
- Koshman, S., A. Spink, and B. Jansen (2006). "Web Searching on the Vivisimo Search Engine". In: *Journal of the American Society for Information Science and Technology* 57.14, pp. 1875–1887.
- Kosslyn, S. (2007). *Clear and to the Point: 8 Psychological Principles for Compelling PowerPoint Presentations*. Oxford University Press.
- Krantz, J. (2012). *Psychological Research on the Net*. URL: <http://psych.hanover.edu/research/exponnet.html> (visited on 03/2013).
- Kriegel, H., P. Kröger, and A. Zimek (2009). "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering and Correlation Clustering". In: *ACM Transactions on Knowledge Discovery from Data* 3.1, Article 1.
- Kristjánsson, Ä., D. Wang, and K. Nakayama (2002). "The role of priming in conjunctive visual search". In: *Cognition* 85.1, pp. 37–52.
- Kules, B. and B. Bederson (Dec. 2004). "Categorized graphical overviews for web search results: An exploratory study using U. S. government agencies as a meaningful and stable structure". In: *Proceedings of the 3rd Annual Workshop on HCI Research in MIS*. Washington DC, USA, pp. 20–23.
- Kules, B., M. Wilson, and B. Shneiderman (2008). *From Keyword Search to Exploration: How Result Visualization Aids Discovery on the Web*. Tech. rep. 1516920080208. Washington DC, USA: The Catholic University of America.
- Kumar, H. and H. Kang (2008). "Another Face of Search Engine: Web Search APIs, in Nguyen, N.T. et al. (Eds.)" In: *Lecture Notes in Artificial Intelligence* LNCS 5027, pp. 311–320.
- Kunar, M. and D. Watson (2011). "Visual Search in a Multi-element asynchronous dynamic (MAD) world". In: *Journal of Experimental Psychology: Human Perception and Performance* 37.4, pp. 1017–1031.
- Kuo, B. et al. (May 2007). "Tag Clouds for Summarizing Web Search Results". In: *Proceedings of 16th International World Wide Web Conference*. Banff, Canada, pp. 1203–1204.
- Landauer, T., D. Laham, and M. Derr (2004). "From Paragraph to graph: Latent semantic analysis for information visualization". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.1, pp. 5214–5219.
- Larkin, J. and H. Simon (1987). "Why a Diagram is (Sometimes) worth Ten Thousand Words". In: *Cognitive Science* 11.1, pp. 65–99.
- Laxar, K. and S. Luria (1990). *Frequency of a flashing light as a navigational range indicator*. Technical Report NSMRL-1157. US Naval Submarine Medical Research Lab. URL: http://archive.rubicon-foundation.org/xmlui/bitstream/handle/123456789/8501/NSMRL_1157.pdf (visited on 03/2013).
- Lee, F. and A. Chan (Sept. 2005). "Spatial Stimulus-Response (S-R) compatibility for auditory display in diagonal arrangement". In: *Proceedings of the 4th International Cyberspace Conference on Ergonomics - CybErg'05*. Johannesburg, South Africa, pp. 368–380.

- Lee, J. (2008). "Fifty Years of Driving Safety Research". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50.3, pp. 521–528.
- Lee, P., A. Klippel, and H. Tappe (2003). "The Effect of Motion in Graphical User Interfaces". In: ed. by A. Butz et al. Vol. LNCS 2733. Smart Graphics: Lecture Notes in Computer Science, pp. 12–21.
- Lee, T. (2011). *The Death of Flash and the Quiet Triumph of Open Standards*. URL: <http://www.cato.org/publications/commentary/death-flash-quiet-triumph-open-standards> (visited on 03/2013).
- Leibbrandt, R. (2009). "Part-of-speech Bootstrapping using Lexically-Specific Frames". Ph.D Thesis. Bedford Park, Australia: Computer Science, Engineering and Mathematics, Flinders University of South Australia.
- Leporini, B., P. Andonico, and M. Buzzi (May 2004). "Designing Search Engine User Interfaces for the Visually Impaired". In: *Proceedings of the 2004 International Cross-Disciplinary Workshop of Web Accessibility - W4A*. New York, USA, pp. 57–66.
- Lerner, R. (2007). "At the Forge: Firebug". In: *Linux Journal* 2007.157, p. 8.
- Leuski, A. and J. Allan (Oct. 2000). "Lighthouse: Showing the Way to Relevant Information". In: *IEEE Symposium on Information Visualization - InfoVis'00*. Salt Lake City, USA, pp. 125–129.
- Levie, W. and R. Lentz (1982). "Effects of Text Illustrations: A Review of Research". In: *Educational Technology Research and Development* 30.4, pp. 195–232.
- Lewis, D. et al. (2004). "RCV1: A new benchmark collection for text categorization research". In: *Journal of Machine Learning Research* 5, pp. 361–397.
- Li, J., J. vanWijk, and J. Martens (Apr. 2009). "Evaluation of symbol contrast in Scatter plots". In: *IEEE Pacific Visualization Symposium - PacificVis'09*. Beijing, China, pp. 97–104.
- Limoges, S., C. Ware, and W. Knight (June 1989). "Displaying Correlations using Position, Motion, Point Size or Point Color". In: *Proceedings of the Graphics Interface Conference*. London, Canada, pp. 262–265.
- Lindberg, T. and R. Näsänen (2003). "The effect of icon spacing and size on the speed of icon processing in the human visual system". In: *Displays* 24.3, pp. 111–120.
- Lohmann, S., J. Ziegler, and L. Tetzlaff (2009). "Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration". In: ed. by Gross et al. Vol. LNCS 5726. INTERACT'09: Lecture Notes in Computer Science, pp. 392–404.
- Lowe, R. (2003). "Animation and Learning: Selective processing of information in dynamic graphics". In: *Learning and Instruction* 13.2, pp. 157–176.
- Lubin, J. and D. Fibush (1997). *Sarnoff JND Vision Model, Contribution to the IEEE Standards C-2.1.6 Compression and Processing Subcommittee*. URL: <http://www.videoclarity.com/PDF/Sarnoff%20jnd-1.pdf> (visited on 03/2013).
- Ludbrook, J. (2000). "Multiple inferences using confidence intervals". In: 27.3, pp. 212–215.
- Mack, A. et al. (2002). "What we see: Inattention and the capture of attention by meaning". In: *Consciousness and Cognition* 11.4, pp. 488–506.
- Mackinlay, J. (1986). "Automating the Design of Graphical Presentations of Relational Information". In: *ACM Transactions on Graphics* 5.2, pp. 110–141.
- MacLennan, A. (July 2005). "Cyberspace worlds for information retrieval". In: *Proceedings of The Human Dimension of Knowledge Organisation*. Barcelona, Spain, pp. 405–414.

- MacLeod, C. (1991). "Half a Century of Research on the Stroop Effect: An Integrative Review". In: *Psychological Bulletin* 109.2, pp. 163–203.
- Maglio, P. and C. Campbell (Apr. 2000). "Tradeoffs in Displaying Peripheral Information". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'00*. The Hague, The Netherlands, pp. 241–248.
- Maguitman, A. et al. (May 2005). "Algorithmic Detection of Semantic Similarity". In: *Proceedings of 14th International World Wide Web Conference - WWW'05*. Chiba, Japan, pp. 107–116.
- Mann, T. (Sept. 1999). "Visualization of WWW-Search Results". In: *Proceedings of 10th International Workshop on Database and Expert Systems Applications - DEXA '99*. Florence, Italy, pp. 264–268.
- Marchionini, G. and B. Shneiderman (1988). "Finding Facts vs. Browsing Knowledge in Hypertext Systems". In: *Computer* 21.1, pp. 70–80.
- Marchionini, G. and R. White (2008). "Find What You Need, Understand What You Find". In: *Journal of Human-Computer Interaction* 23.3, pp. 205–237.
- Matsuda, Y. et al. (2009). "An analysis of Eye Movements during Browsing Multiple Search Results Pages". In: ed. by J.A. Jako. Vol. LNCS 5610. Human-Computer Interaction Part I: Lecture Notes in Computer Science, pp. 121–130.
- Mayer, M. (2008). *The Future of Search*. URL: <http://googleblog.blogspot.com.au/2008/09/future-of-search.html> (visited on 03/2013).
- McCormac, A. et al. (2012). *Interfaces for Discourse Summarisation: A Human Factors Analysis*. Tech. rep. Bedford Park, Australia: Science, Engineering and Mathematics, Flinders University of South Australia.
- McCown, F. and M. Nelson (June 2007). "Agreeing to Disagree: Search Engines and Their Public Interfaces". In: *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries - JCDL'07*. Vancouver, Canada, pp. 309–318.
- McCrickard, S., R. Catrambone, and J. Stasko (July 2001). "Evaluating Animation in the Periphery as a Mechanism for Maintaining Awareness". In: *Proceedings of Conference on Human-Computer Interaction - INTERACT'01*. Tokyo, Japan, pp. 148–156.
- McDougall, S. and M. Curry (Sept. 2004). "More than just a picture: Icon interpretation in context". In: *Proceedings of the 1st International Workshop on Coping with Complexity*. Bath, England, pp. 73–81.
- McKee, S. and K. Nakayama (1984). "The Detection of Motion in the Peripheral Visual Field". In: *Vision Research* 24.1, pp. 25–32.
- Miller, G. (1956). "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". In: *The Psychological Review* 63, pp. 81–97.
- Monahan, J. and G. Lockhead (1977). "Identification of Integral Stimuli". In: *Journal of Experimental Psychology: General* 106.1, pp. 94–110.
- Moore, D. and G. McCabe (1998). *Introduction to the practice of statistics*. 3rd ed. New York, USA: W.H. Freeman and Company.
- Morris, C., D. Ebert, and P. Rheingans (Oct. 1999). "Experimental Analysis of the effectiveness of features in Chernoff faces". In: *SPIE Proceedings of Applied Imagery Pattern Recognition '99 - 3D Data Visualization for Data Exploration and Decision Making*. Washington, USA, pp. 12–17.
- Morrison, J., B. Tversky, and M. Bétrancourt (June 2000). "Animation: Does it facilitate learning?" In: *Proceedings of 2nd AAAI Spring Symposium on Smart Graphics*. Hawthorne, USA, pp. 53–59.

- Morse, E., M. Lewis, and K. Olsen (2002). "Testing Visual Information Retrieval Methodologies Case Study: Comparative Analysis of Textual, Icon, Graphical and Spring Displays". In: *Journal of the American Society for Information Science and Technology* 53.1, pp. 28–40.
- Morville, P. and J. Callender (2010). *Search Patterns*. California, USA: O'Reilly Media.
- Mote, K. (2007). "Fast Point-Feature Label Placement for Dynamic Visualizations". In: *Information Visualization* 6.4, pp. 249–260.
- Muramatsu, J. and W. Pratt (Sept. 2001). "Transparent Queries: Investigating Users' Mental Models of Search Engines". In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'01*. New Orleans, USA, pp. 217–224.
- Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. Boca Raton, USA: Chapman & Hall CRC.
- Nakagawa, S. (2004). "A farewell to Bonferroni: the problems of low statistical power and publication bias". In: 15.6.
- Narayanan, N. and R. Hübscher (1997). "Visual Language Theory: Towards a Human-Computer Interaction Perspective". In: ed. by K. Marriott and B. Meyer. *Visual Language Theory*. Berlin, Germany: Springer, pp. 85–127.
- NASA (2010). *Human Integration Design Handbook*. Technical Report NASA/SP-2010-3407. Washington DC, USA: National Aeronautical and Space Administration (NASA). URL: http://ston.jsc.nasa.gov/collections/TRS/_techrep/SP-2010-%203407.pdf (visited on 03/2013).
- Nesbitt, K. (Jan. 2005). "Using Guidelines to Assist in the Visualisation Design Process". In: *Proceedings Asia Pacific Symposium on Information Visualisation - APVIS'05*. Sydney, Australia, pp. 115–123.
- Newman, D. et al. (2009). "Visualizing search results and document collections using topic maps, Web Semantics: Science". In: *Services and Agents on the World Wide Web* 8.2-3, pp. 169–175.
- Nielsen, J. (2006). *F-Shaped Pattern for Reading Web Content*. URL: http://www.useit.com/alertbox/reading_pattern.html (visited on 03/2013).
- Niemelä, M. and P. Saariluoma (2003). "Layout Attributes and Recall". In: *Behaviour and Information Technology* 22.5, pp. 353–363.
- Norman, D. (2002). *The Design of Everyday Things*. Basic Books.
- Nowak, S. and S. Rürger (Mar. 2010). "How Reliable are Annotations via Crowdsourcing? A study about inter-annotator agreement for multi-label image annotation". In: *Proceedings of the 11th ACM International Conference on Multimedia Information Retrieval - MIR'10*. Pennsylvania, USA, pp. 557–566.
- Nowell, L. (1997). "Graphical Encoding for Information Visualization: Using Icon Color, Shape and Size to Convey Nominal and Quantitative Data". Ph.D Thesis. Blacksburg, USA: Virginia Polytechnic Institute and State University.
- Nowell, L., R. Schulman, and D. Hix (Oct. 2002). "Graphical encoding for information visualization: an empirical study". In: *IEEE Symposium on Information Visualization - InfoVis'02*. Boston, USA, pp. 43–50.
- Olston, C. and E. Chi (2003). "ScentTrails: Integrating Browsing and Searching on the Web". In: *ACM Transactions on Computer-Human Interaction* 10.3, pp. 1–21.
- Oppenheim, C. (1997). "Manager's Use and Handling of Information". In: *International Journal of Information Management* 17.4, pp. 239–248.

- Oppenheimer, D., T. Meyvis, and N. Davidenko (2009). “Instructional manipulation checks: Detecting satisficing to increase statistical power”. In: *Journal of Experimental Social Psychology* 45.4, pp. 867–872.
- Ostergren, M., S. Yu, and E. Efthimiadis (July 2010). “The Value of Visual Elements in Web Search”. In: *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR’10*. Geneva, Switzerland, pp. 867–868.
- Oulasvirta, A., J. Hukkinen, and B. Schwartz (July 2009). “When More Is Less: The Paradox of Choice in Search Engine Use”. In: *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR’09*. Boston, USA, pp. 516–523.
- Paley, B. (Nov. 2003). “Designing better transparent overlays by applying illustration techniques”. In: *Adjunct Proceedings of the 16th annual ACM Symposium on User Interface Software and Technology - UIST’03*. Vancouver, Canada.
- Paolacci, G., J. Chandler, and P. Ipeirotis (2010). “Running experiments on Mechanical Turk”. In: *Judgement and Decision Making* 5.5, pp. 411–419.
- Pappachan, P. and M. Ziefle (2008). “Cultural influences on the comprehensibility of icons in mobile-computer interaction”. In: *Behaviour and Information Technology* 27.4, pp. 331–337.
- Pérez, C. and A. deAntonio (2003). *Spatialization Algorithms for Text Document Set Visualization*. Tech. rep. CP-03-01. Madrid, Spain: Faculty of Information Technology, Polytechnic University of Madrid.
- Pfitzner, D. (2009). “An Investigation into User Text Query and User Text Descriptor Construction”. Ph.D Thesis. Bedford Park, Australia: Science, Engineering and Mathematics, Flinders University of South Australia.
- Pfitzner, D., V. Hobbs, and D. Powers (Feb. 2001). “A Unified Taxonomic Framework for Information Visualisation, Proceedings of The 3rd Australasian Symposium on Information Visualization - Invis.au’2004”. In: *Proceedings of the Asia-Pacific Symposium on Information Visualisation*. Adelaide, Australia, pp. 57–66.
- Pfitzner, D., K. Treharne, and D. Powers (2008). “User Keyword Preference: the Nwords and Rwords Experiments”. In: *International Journal of Internet Protocol Technology* 3.3, pp. 149–158.
- Phan, X. and L. Nguyen (2008). *A Java Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference*. URL: <http://jgibbllda.sourceforge.net> (visited on 03/2013).
- Phan, X., L. Nguyen, and S. Horiguchi (Apr. 2008). “Learning to classify short and sparse text & web with hidden topics from large-scale data collections”. In: *Proceedings of 17th International World Wide Web Conference - WWW’08*. Beijing, China, pp. 91–100.
- Pickett, R. and G. Grinstein (Aug. 1988). “Iconographic Displays for Visualizing Multidimensional Data”. In: *Proceedings of IEEE Conference on Systems, Man and Cybernetics*. Vol. 1. Beijing, China, pp. 514–519.
- Pike, W.A. et al. (2009). “The science of interaction”. In: *Journal of Information Visualization* 8.4, pp. 263–274.
- Pirolli, P. and S. Card (1999). “Information Foraging”. In: *Psychological Review* 106.4, pp. 643–675.
- Pirolli, P., S. Card, and M. van der Wege (May 2000). “The Effect of Information Scent on Searching Information Visualizations of Large Tree Structures”. In: *Proceedings*

- of the Working Conference on Advanced Visual Interfaces - AVI'00*. Palermo, Italy, pp. 161–172.
- Plaisant, C. and J. Fekete (May 1999). “Excentric Labelling: Dynamic Neighborhood Labelling for Data Visualization”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'99*. Pittsburgh, USA, pp. 512–519.
- Plaue, C. and J. Stasko (May 2007). “Animation in a Peripheral Display: Distraction, Appeal and Information Conveyance in Varying Display Configurations”. In: *Proceedings of the Graphics Interface Conference*. Montréal, Canada, pp. 135–142.
- Polowinski, J. (2009). “Widgets for Faceted Browsing, Human Interface: Part I”. In: *Human Interface: Part I: Lecture Notes in Computer Science*, pp. 601–610.
- Poole, B. (Nov. 2008). “Yahoo! and Next Generation Search Experience”. In: *HCSNet Next Generation Search Workshop*. Keynote Lecture. Melbourne, Australia.
- Powers, D. (July 2003). “Recall & Precision versus the Bookmaker”. In: *International Conference on Cognitive Science*. Sydney, Australia, pp. 529–534.
- (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Technical Report SIE-07-001. Adelaide, Australia: Flinders University of South Australia.
- Powers, D. and D. Pfizner (July 2003). “The Magic Science of Visualisation”. In: *Proceedings of the Joint International Conference on Cognitive Science*. Sydney, Australia, pp. 529–534.
- Proctor, R. and K. Vu (2006). “The Cognitive Revolution at Age 50: Has the Promise of the Human Information-Processing Approach Been Fulfilled”. In: *International Journal of Human-Computer Interaction* 21.3, pp. 253–284.
- Reilly, D. and K. Inkpen (Apr. 2007). “White Rooms and Morphing don’t mix: setting and the evaluation of visualization techniques”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'07*. San Jose, USA, pp. 111–120.
- Reips, U. (2001). “The Web Experimental Psychology Lab: Five years of data collection on the Internet”. In: *Behavior Research Methods, Instruments & Computers* 33.2, pp. 201–211.
- (2002). “Standards for Internet-Based Experimenting”. In: *Experimental Psychology* 49.4, pp. 243–256.
- Ren, P. et al. (2011). “Size Matters: Non-Numerical Magnitude Affects the Spatial Coding of Response”. In: *PLoS one* 6.8, pp. 241–249.
- Riche, N., B. Lee, and C. Plaisant (2010). “Understanding Interactive Legends: A Comparative Evaluation with Standard Widgets”. In: *Computer Graphics Forum* 29.3, pp. 1193–1202.
- Rivadeneira, W. and B. Bederson (2003). *A Study of Search Result Clustering Interfaces: Comparing Textual and Zoomable User Interfaces*. Technical Report HCIL-2003-36, CS-TR-4682. Maryland, USA: University of Maryland HCIL. URL: <http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2003-36> (visited on 03/2013).
- Robertson, G., M. Czerwinski, et al. (2009). “Selected Human Factors Issues in Information Visualization”. In: *Reviews of Human Factors and Ergonomics* 5.1, pp. 41–81.
- Robertson, G., R. Fernandez, et al. (2008). “Effectiveness of Animation in Trend Visualization”. In: *Transactions on Computer Graphics and Visualization* 14.6, pp. 1325–1232.

- Rodrigues, J. et al. (2007). “The Spatial-Perceptual Design Space: a new Comprehension for Data Visualization”. In: *Information Visualization* 6.4, pp. 261–279.
- Ropinski, T. and B. Preim (Feb. 2008). “Taxonomy and Usage Guidelines for Glyph-based Medical Visualization”. In: *Simulation and Visualisation - SimVis'2008*. Magdeburg, Germany, pp. 121–138.
- Rosas, R. et al. (2003). “Beyond Nintendo: design and assessment of educational video games for first and second grade students”. In: *Computers and Education* 40.1, pp. 71–94.
- Rosenholtz, R. et al. (Apr. 2005). “Feature Congestion: A measure of Display Clutter”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'05*. Portland, USA, pp. 761–770.
- Rosling, H. (2009). *Debunking third-world myths with the best stats you've ever seen*. URL: http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html (visited on 03/2013).
- Ross, J. et al. (Apr. 2010). “Who are the crowdworkers? Shifting Demographics in Mechanical Turk”. In: *Extended Abstracts on Human Factors in Computing Systems - CHI EA'10*. Atlanta, USA, pp. 2725–2734.
- Russell, D. et al. (Jan. 2006). “Being literate with large document collections: Observational studies and cost structure tradeoffs”. In: *Proceedings 39th Hawaii International Conference on System Sciences - HICSS'06*. Kauai, USA, pp. 55–62.
- Ruxton, G. and G. Beauchamp (2008). “Time for some a priori thinking about post hoc testing”. In: 19.3, pp. 690–693.
- Sabol, V. et al. (July 2009). “Visual Knowledge Discovery in Dynamic Enterprise Text Repositories”. In: *Proceedings of 13th International Conference on Information Visualisation - IV'09*. Barcelona, Spain, pp. 361–368.
- Sandvig, C. and D. Bajwa (2011). “User Perceptions of Search Enhancements in Web Search”. In: *Journal of Computer Information Systems* 52.2, pp. 22–32.
- Santiago, J. et al. (2007). “Time (also) flies from left to right”. In: *Psychonomic Bulletin & Review* 14.3, pp. 512–516.
- Saracevic, T. (Oct. 1996). “Relevance Reconsidered, In Information Science, Integration in Perspectives”. In: *Proceedings of the Second Conference on Conceptions of Library and Information Science - CoLIS 2*. Copenhagen, Denmark, pp. 201–218.
- Schick, T. and L. Vaughn (2002). *How to Think About Weird Things: Critical Thinking for a New Age*. 3rd ed. McGraw Hill.
- Scholer, F., A. Turpin, and M. Sanderson (July 2011). “Quantifying Test Collection Quality Based on Consistency of Relevance Judgements”. In: *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'11*. Beijing, China, pp. 1063–1072.
- Schwalm, N., V. Shaviv, and G. Goldschmidt (July 2000). “Can icon animation enhance human performance...or is it just another gimmick?” In: *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*. San Diego, US, pp. 323–326.
- Schwartz, B. (2005). *The Paradox of Choice: Why More Is Less*. Harper Perennial.
- Sebrechts, M. et al. (Aug. 1999). “Visualisation of Search Results: A Comparative Evaluation of Text, 2D and 3D Interfaces”. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'99*. Berkely, USA, pp. 3–10.
- Segel, E. and J. Heer (2010). “Narrative Visualization: Telling Stories with Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pp. 1139–1148.

- Seifert, C. et al. (July 2008). "On the Beauty and Usability of Tag Clouds". In: *Proceedings of 12th International Conference on Information Visualisation*. Montpellier, France, pp. 17–25.
- Setlur, V. et al. (2005). "Semanticons: Visual Metaphors as File Icons". In: *Computer Graphics Forum* 24.3, pp. 647–656.
- Shanmugasundaram, M. and P. Irani (May 2008). "The Effect of Animated Transitions in Zooming Interfaces". In: *Proceedings of the Working Conference on Advanced Visual Interfaces - AVI'08*. Napoli, Italy, pp. 396–399.
- Shiffrin, R. and R. Nosofsky (1994). "Seven Plus or Minus Two: A Commentary on Capacity Limitations". In: *Psychological Review* 101.2, pp. 357–361.
- Shneiderman, B. (Sept. 1996). "The Eyes have it: a Task by Data Type Taxonomy for Information Visualisations". In: *Proceedings of IEEE Symposium on Visual Languages*. Boulder, USA, pp. 336–343.
- (1998). *Designing the User Interface: Strategies for effective human-computer interaction*. 3rd ed. Massachusetts, USA: Addison-Wesley Longman.
- Simon, R. and S. Overmeyer (1984). "The Effect of Redundant Cues on Retrieval Time". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 26.3, pp. 315–321.
- Sinclair, J. and M. Cardew-Hall (2008). "The folksonomy tag cloud: when is it useful?" In: *Journal of Information Science* 31.1, pp. 15–29.
- Skitka, L. and E. Sargis (2006). "The Internet as Psychological Laboratory". In: *Annual Review of Psychology* 57, pp. 529–555.
- Skupin, A. and S. Fabrikant (2003). "Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization". In: *Cartography and Geographic Information Science* 30.2, pp. 95–115.
- Smilek, D., M. Dixon, and P. Merikle (2006). "Revisiting the category effect: The influence of meaning and search strategy on the efficiency of visual search". In: *Brain Research* 1080.1, pp. 73–90.
- Smith, G. et al. (2006). "FacetMap: A Scalable Search Browse Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 12.5, pp. 797–804.
- Song, R. et al. (2009). "Identification of ambiguous queries in web search". In: *Information Processing and Management* 45.2, pp. 216–229.
- Spink, A. (1997). "Study of Interactive Feedback during Mediated Information Retrieval". In: *Journal of the American Society for Information Science* 48.5, pp. 382–394.
- Spink, A. and B.J. Jansen (2006). "How are we searching the World Wide Web? A comparison of nine search engine transaction logs". In: *Information Processing and Management* 42.1, pp. 248–263.
- Spink, A., D. Wolfram, et al. (2001). "Searching the Web: The Public and their Queries". In: *Journal of the American Society for Information Science and Technology* 52.3, pp. 226–234.
- Spoerri, A. (July 2004). "Coordinated Views and Tight Coupling to Support Meta Search". In: *Proceedings 2nd International Conference of Coordinated and Multiple Views in Exploratory Visualization - CMV'04*. London, England, pp. 39–48.
- Stefaner, M. (2005). "Projection Techniques for Document Maps". Bachelor's Thesis. Osnabrück, Germany: University of Osnabrück.
- Sutcliffe, A., M. Ennis, and J. Hu (2000). "Evaluating the Effectiveness of Visual User Interfaces for Information Retrieval". In: *International Journal of Human-Computer Studies* 53.5, pp. 741–763.

- Tang, R., W. Shaw, and J. Vevea (1999). “Towards the Identification of the Optimal Number or Relevance Categories”. In: *Journal of the American Society for Information Science* 50.3, pp. 254–264.
- Tavanti, M. and M. Lind (Oct. 2001). “2D vs 3D, Implications on Spatial Memory”. In: *Proceedings IEEE Symposium on Information Visualization - InfoVis'01*. San Diego, USA, pp. 139–146.
- Thackray, R. and R. Touchstone (1990). “Effects of Monitoring under High and Low Task Load on Detection of Flashing and Colored Radar Targets”. In: *Ergonomics* 34.8, pp. 1065–1081.
- Thomas, B. and P. Calder (2001). “Applying Cartoon Animation Techniques to Graphical User Interfaces”. In: *ACM Transactions on Computer-Human Interaction* 8.3, pp. 198–222.
- Todd, P. and I. Benbasat (1994). “The Influence of Decision Aids on Choice Strategies Under Conditions of High Cognitive Load”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 24.4.
- Tolin, P. and R. Ryen (1986). “Discriminability of change in signal flash-rates”. In: *Perceptual and Motor Skills* 63.3, pp. 1259–1264.
- Tory, M. and T. Möller (2004). “Human Factors in Visualization Research”. In: *IEEE Transactions on Information Visualization and Computer Graphics* 10.1, pp. 1–13.
- Tory, M., D. Sprague, et al. (2007). “Spatialization Design: Comparing Points and Landscapes”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6, pp. 1262–1269.
- Treharne, K., D. Pfitzner, et al. (July 2008). “A Lean online approach to human factors research”. In: *Proceedings of 1st International Conference on Pervasive Technologies Related to Assistive Environments - PETRA'08*. Athens, Greece, Article 57.
- Treharne, K. and D. Powers (July 2009). “Search Engine Result Visualisation: Challenges and Opportunities, Symposium on Web Visualisation”. In: *Proceedings of 13th International Conference on Information Visualisation - IV'09*. Barcelona, Spain, pp. 633–638.
- Tudoreanu, M. and D. Hart (Apr. 2004). “Interactive Legends”. In: *Proceedings of 42nd Annual South-east Regional Conference - ACMSE'04*. Huntsville, USA, pp. 448–453.
- Tufte, E. (1990). *Envisioning Information*. Cheshire, USA: Graphics Press.
- (2001). *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, USA: Graphics Press.
- Turetken, O. and R. Sharda (2005). “Clustering-Based Visual Interfaces for Presentation of Web Search Results: An Empirical Investigation”. In: *Information Systems Frontiers* 7.3, pp. 273–297.
- Tversky, B. (2011). “Visualizing Thought”. In: *Topics in Cognitive Science* 3.3, pp. 499–535.
- Tversky, B., J. Morrison, and M. Bétrancourt (2002). “Animation: Can it Facilitate?” In: *International Journal of Human-Computer Studies* 57.4, pp. 247–262.
- Valiati, E., M. Pimenta, and C. Freitas (May 2006). “A taxonomy of Tasks for Guiding the Evaluation of Multidimensional Visualizations”. In: *Proceedings of the 2006 AVI Workshop on BEyond time and errors Novel Evaluation Methods for Information Visualization - BELIV'06*. Venice, Italy, Article1–6.
- vanOrden, K., J. Divita, and M. Shim (1993). “Redundant Use of Luminance and Flashing with Shape and Color as Highlighting Codes in Symbolic Displays”. In: *Human Factors* 35.2, pp. 195–204.

- Veerasamy, A. and R. Heikes (July 1997). “Effectiveness of a graphical display of retrieval results”. In: *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'97*. Philadelphia, USA, pp. 236–245.
- vonAhn, L. and L. Dabbish (Apr. 2004). “Labelling Images with a Computer Game”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'04*. Vienna, Austria, pp. 319–326.
- vonAhn, L., M. Kedia, and M. Blum (Apr. 2006). “Verbosity: a game for collecting common-sense facts”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'06*. Montréal, Canada, pp. 75–78.
- Ward, M. (2002). “A taxonomy of glyph placement strategies for multidimensional data visualization”. In: *Information Visualization 1.3-4*, pp. 194–201.
- (2008). “Multivariate Data Glyphs: Principles and Practice”. In: *Handbook of Data Visualization*. Ed. by C. Chen et al. Springer Handbook of Computational Statistics. Berlin, Germany: Springer, pp. 179–198.
- Ware, C. (2004). *Information Visualization: Perception for Design*. 2nd ed. California, USA: Elsevier/Morgan Kaufmann Publishers.
- Ware, C. and R. Bobrow (2004). “Motion to support rapid interactive queries on node-link diagrams”. In: *Transactions on Applied Perception 1.1*, pp. 3–18.
- (July 2006). “Motion Coding for Pattern Detection”. In: *Symposium on Applied Perception in Computer Graphics - APGV'06*. Boston, USA, pp. 107–110.
- Ware, C., J. Bonner, et al. (1992). “Moving Icons as a Human Interrupt”. In: *International Journal of Human-Computer Interaction 4.4*, pp. 341–348.
- Ware, C. and S. Limoges (1994). *Perceiving data displayed through oscillatory motion*. Technical Report TR94-089. Fredericton, Canada: University of New Brunswick.
- Web Content Accessibility Guidelines - WCAG 2.0 - Section 1.4.3 Contrast Minimum* (2008). URL: <http://www.w3.org/TR/WCAG20/> (visited on 03/2013).
- Weidenbacher, H. and M. Barnes (1997). “Target search in tactical displays with standard, single cue and redundant coding”. In: *Displays 18.1*, pp. 1–10.
- Weigle, C. et al. (May 2000). “Oriented Sliver Textures: A Technique for Local Value Estimation of Multiple Scalar Fields”. In: *Proceedings of the Graphics Interface Conference*. Montréal, Canada, pp. 163–170.
- Weigmann, D. et al. (2004). “Human Factors Aspects of Power System Flow Animation”. In: *IEEE Transactions on Power Systems 20.3*, pp. 1233–1240.
- Weiten, W. (2001). *Psychology Themes & Variations*. 5th ed. Belmont, USA: Thomson Learning Inc.
- White, R., J. Jose, and I. Ruthven (Sept. 2003). “A Granular Approach to Web Search Result Presentation”. In: *Proceedings of the 9th International Conference on Human-Computer Interaction - INTERACT'2003*. Zurich, Switzerland, pp. 213–220.
- Williams, D. and E. Reingold (2001). “Pre-attentive guidance of eye movements during triple conjunction search tasks: The effects of feature discriminability and saccadic amplitude”. In: *Psychonomic Bulletin & Review 8.3*, pp. 476–488.
- Wilson, M., P. André, and M. Schraefel (Oct. 2008). “Backward Highlighting: Enhancing Faceted Search”. In: *Proceedings of 21st ACM Symposium on User Interface Software and Technology*. Monterey, USA, pp. 235–238.
- Wise, J. (1999). “The Ecological Approach to Text Visualization”. In: *Journal of the American Society for Information Science 50.13*, pp. 1224–1233.

- Wolfe, J. (1998). "Visual Search". In: ed. by H.Pashler. *Attention*. Hove, England: Psychology Press Ltd., pp. 13–73.
- (2007). "Guided Search 4.0: Current Progress with a model of visual search". In: ed. by W. Gray. *Integrated Models of Cognitive Systems*. New York, USA: Oxford University Press, pp. 99–119.
- Wolfe, J., S. Friedman-Hill, et al. (1992). "The Role of Categorization in Visual Search for Orientation". In: *Journal of Experimental Psychology: Human Perception and Performance* 18.1, pp. 34–49.
- Wolfe, J. and T. Horowitz (2004). "What attributes guide the deployment of visual attention and how do they do it?" In: *Nature Reviews Neuroscience* 5.6, pp. 495–501.
- Woodruff, A. et al. (Mar. 2001). "Using thumbnails to search the web". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'01*. Vol. 3. 1. Seattle, USA, pp. 198–205.
- Woods, D. (1995). "The alarm problem and directed attention in dynamic fault management". In: *Ergonomics* 38.11, pp. 2371–2393.
- Woods, D. et al. (2002). "Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis". In: *Cognition, Technology and Work* 4, pp. 22–36.
- Wright, H. (2007). *Introduction to Scientific Visualization*. London, England: Springer.
- Wu, M., M. Fuller, and R. Wilkinson (2001). "Using clustering and classification approaches in interactive retrieval". In: *Information Processing and Management* 37.3, pp. 459–484.
- Yee, P. et al. (Apr. 2003). "Faceted Metadata for Image Search and Browsing". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'03*. Fort Lauderdale, USA, pp. 401–408.
- Yost, B. and C. North (Apr. 2005). "Single Complex Glyphs versus Multiple Simple Glyphs". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI'05*. Portland, USA, pp. 1889–1890.
- Zamir, O. and O. Etzioni (Aug. 1998). "Web Document Clustering: A Feasibility Demonstration". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'98*. Melbourne, Australia, pp. 46–54.
- (1999). "Grouper: A Dynamic Clustering Interface to Web Search Results". In: *Computer Networks: The International Journal of Computer and Telecommunications Networking* 31.11-16, pp. 1361–1374.
- Zhang, J. (2008). *Visualization for Information Retrieval*. The Information Retrieval Series 23. Berlin, Germany: Springer.
- Zipf, G. (1949). *Human Behavior and the principle of least effort*. Addison-Wesley.
- Zuffi, S. et al. (Sept. 2007). "Human Computer Interaction: Legibility and Contrast". In: *Proceedings of 14th International Conference on Image Analysis and Processing - ICIAP'07*. Modena, Italy, pp. 241–246.