An Investigation into User Text Query and Text Descriptor Construction

by

Darius Mark Pfitzner,

B.Int.Bus. (Bachelor of International Business)

B.Comp.Sci (Bachelor of Computer Science)

M.Info.Tech (Masters of Information Technology)

School of Computer Science,

Engineering and Mathematics,

Faculty of Science and Engineering

A thesis presented to the

Flinders University of South Australia

in total fulfillment of the requirements for the degree of

Doctor of Philosophy

Adelaide, South Australia, 2009

Contents

List of Figures

List of Tables

Certification

I certify that this thesis does not incorporate without acknowledgment 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.

As requested under Clause 14 of Appendix D of the *Flinders University Research Higher Degree Student Information Manual* I hereby agree to waive the conditions referred to in Clause 13(b) and (c), and thus

- Flinders University may lend this thesis to other institutions or individuals for the purpose of scholarly research;
- Flinders University may reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signed Dated

Darius Mark Pfitzner

Acknowledgements

This thesis arose in part from two years of research conducted under the guidance of Professor David Powers. From the beginning of that research until today I have had the pleasure of interacting with and working alongside a great number of very personable, collegiate and experienced people, many of which influenced my thesis through either opinion and/or guidance, and some who simply made the time more interesting and enjoyable. These people range from conference organisers and attendees, academics from other universities and members of other research institutions, such as the HCI crew at DSTO and the Thinking Head crew, to fellow members of the Artificial Intelligence Laboratory here at Flinders University. To these people, and they know who they are, I give my very humble thanks and I hope that in some way I have made their life a little more enjoyable, their research a little easier or hopefully both, and that I can do so again in the future.

Of these people I would like to especially thank professor David Powers, Dr. Richard Leibbrandt, Dr. Trent Lewis, Dr. Martin Leurssen, Mr. Graham Bignell and finally, and most importantly my wife Susan Clarkson and kids Pheobe and Mellion. David Powers has been the best and supporting supervisor a person could have as well as an excellent friend. Richard Leibbrandt has been the best room mate a person could ask for, a very good friend and someone whom I have had the great pleasure of working alongside of as a research associate. Trent Lewis and Martin Leurssen have been two younger men from whom I have learned much, and that have not only demonstrated great humility in allowing me to manage them from time to time but also have been constant sources of friendship, camaraderie, opinion and research advice. Graham Bignell has been a great source of information, guidance and advice, and has been a very special friend whom I have been able to depend on for support, and with whom I have spent much time arguing/discussing all manner of things from sport to algorithms.

Most importantly, my wife Susan and kids Pheobe and Mellion have been there every step of the emotional roller coaster of my academic career, holding my hand, and for this I can say no less than I love them!

Finally, I would like to acknowledge those several thousand people whom I bugged, cajoled, convinced, badgered, and otherwise coerced into completing my many and varied surveys. These people supplied the data, that underpinned my research, and as such have also helped me add to our race's knowledge.

Published Papers

0.1 Journal Publications

[Pfit08a] Darius Pfitzner, Richard Leibbrandt and David Powers (2008), "Characterization and evaluation of similarity measures for pairs of clusterings", Knowledge and Information Systems, published online Saturday, July 05, 2008, Web version available at http://dx.doi.org/10.10 07/s10115-008-0150-6.

[Pfit08b] Darius Pfitzner, Kenneth Treharne & David M. W. Powers (in press, accepted May 2008), "User Keyword Preference: the Nwords and Rwords Experiments", International Journal of Internet Protocol Technology: Special Issue on Intelligent Internet-based Systems: Emerging Technologies and Programming Techniques.

0.2 Conference Publications

[Powe08c] David M. W. Powers, Richard Leibbrandt, Darius Pfitzner, Martin Luerssen, Trent Lewis, Arman Abrahamyan and Kate Stevens, "Language Teaching in a Mixed Reality Games Environment", The 1st International Conference on PErvasive Technologies Related to Assistive Environments (PETRA) Workshop on "Gaming Design and Experience: Design for Engaging Experience and Social Interaction", July 15-19, 2008, Athens Greece.

[Treh08] Kenneth Treharne, Darius Pfitzner, Richard Leibbrandt & David M. W. Powers, "A Lean Online Approach to Human Factors Research", The 1st International Conference on PErvasive Technologies Related to Assistive Environments (PETRA) workshop on "Pervasive Technologies in e/m-Learning and Internet based Experiments" (PTLIE), July 15-19, 2008, Athens Greece.

[Pfit07a] Darius Pfitzner, Kenneth Treharne & David M. W. Powers (2007), "Cognitive load in text search: The Nwords and Rwords surveys", Australian Society for Cognitive Science Conference, July 9-11, 2007, Abstract.

[Pfit07b] Darius Pfitzner, Kenneth Treharne & David M. W. Powers (2007), "Cognitive Load in Text Search: the Nwords and Rwords Surveys", Joint HCSNet-HxI Workshop on Human Issues in Interaction and Interactive Interfaces, 13-14 September 2007, Australian Technology Park, Sydney, Abstract.

[Treh07a] Kenneth Treharne, Darius Pfitzner & David M. W. Powers (2007), "The versatile role of motion in visualization", Australian Society for Cognitive Science Conference, July 9-11, 2007, Abstract.

[Shill07b] Anna Shillabeer and Darius Pfitzner (2007)., "Determining Pattern Element Contribution in Medical Datasets", Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007), Ballarat, Australia. CRPIT, 68. ACS.

[Treh06a] Kenneth Treharne, Darius Pfitzner and David Powers (2006)., "Information Coding in Animation", Australian Language Technology Workshop Poster, University of Sydney, November 2006 (Extended Abstract).

Abstract

Cognitive limitations such as those described in Miller's (1956) work on channel capacity and Cowen's (2001) on short-term memory are factors in determining user cognitive load and in turn task performance. Inappropriate user cognitive load can reduce user efficiency in goal realization. For instance, if the user's attentional capacity is not appropriately applied to the task, distractor processing can tend to appropriate capacity from it. Conversely, if a task drives users beyond their short-term memory envelope, information loss may be realized in its translation to long-term memory and subsequent retrieval for task base processing.

To manage user cognitive capacity in the task of text search the interface should allow users to draw on their powerful and innate pattern recognition abilities. This harmonizes with Johnson-Laird's (1983) proposal that propositional representation is tied to mental models. Combined with the theory that knowledge is highly organized when stored in memory an appropriate approach for cognitive load optimization would be to graphically present single documents, or clusters thereof, with an appropriate number and type of descriptors. These descriptors are commonly words and/or phrases.

Information theory research suggests that words have different levels of importance in document topic differentiation. Although key word identification is well researched, there is a lack of basic research into human preference regarding query formation and the heuristics users employ in search. This lack extends to features as elementary as the number of words preferred to describe and/or search for a document. Contrastive understanding these preferences will help balance processing overheads of tasks like clustering against user cognitive load to realize a more efficient document retrieval process. Common approaches such as search engine log analysis cannot provide this degree of understanding and do not allow clear identification of the intended set of target documents.

This research endeavours to improve the manner in which text search returns are presented so that user performance under real world situations is enhanced. To this end we explore both how to appropriately present search information and results graphically to facilitate optimal cognitive and perceptual load/utilization, as well as how people use textual information in describing documents or constructing queries.

Chapter 1

Introduction

The Introductory Chapter of this thesis serves two purposes:

1. to provide background information on the context and motivation of the work described in the thesis;
2. to offer a map of the remaining chapters of this thesis including short summary descriptions of the treatment in each chapter.

The work contained in this thesis works toward a comprehensive answer to the question;

"How many words do people naturally use to describe and/or query for documents?"

Relative to this question this thesis does the following:

1. Motivates this research by describing an overarching dream within a real world context.
2. Extensively reviews the cognitive aspects of human-computer interaction.
3. Critiques and reviews the current approaches used to answer similar questions.
4. Outlines several experiments and associated results that were designed to empirically answer this question.
5. Proposes a measure that can be used in the comparison of human and automatically generated document keyword lists.

1.1 Background

The dream:

Imagine searching for textual data using a system that is so attuned to the user's information need, context and general cognitive traits that for any document search, on the first attempt and within a few seconds it returns at most a very small list of documents (say one to five) that all address the information need perfectly or near enough to it for your purposes.

This thesis is a step toward this dream which seems to become more and more distant given the rapidly increasing amounts of data being stored in electronic form around the world.

At the risk of sounding a little theatrical, the dream is in stark contrast to the reality of the research-style search of today and drives at the heart of humanity's future success. For example, a text search often sees the user set out to find what is thought to be readily available textual information from a data source like the WWW (World Wide Web) only to be frustrated by the process. This normally sees several words typed into a single line search engine interface the result of which I describe as a "Data-Avalanche". This is where the search engine returns an apparently ranked list of documents far too large to manually filter (often in the millions) that in some questionable way addresses the search criteria. Unperturbed, the user surveys the list to find only a few mildly appropriate documents in the first few pages of returns if anything at all. "That's O.K." they say to themselves having experienced this situation on what seems to be an hourly basis (especially when doing a PhD) and knowing the information required is out there somewhere and is quite possibly in the "ranked" avalanche of returns. Filled with optimism they type different seemingly targeted search criteria or extend the original criteria, and search again. This time they only receive 10,000 returns a similarly un-motivating and time-consuming result when the required information is still not near the top of the list.

This scenario highlights a critical bottleneck for decision making processes relying on rapid text based information retrieval. At the core of human success is the ability to make "informed" decisions and information is the critical component in decision-making processes. From humanity's perspective, its success has been fueled by the

individual's ability to not only store and retrieve information internally as memories but also externally in hard formats like books and recently technology based soft formats.

> If information can't be retrieved in a timely and accurate manner humanity's continuing progress will falter!

Toward the realization of "the dream", which equates to the "ultimate text search system", this thesis adds to a Masters thesis and other work by Pfitzner et al. (Pfitzner, Hobbs & Powers 2003). The Masters thesis proposed techniques and tools to guide the appropriate use of visual screen artifacts/devices/cues when designing search interfaces that present multi-dimensional data, specifically textual documents. The authors were critical of the then current graphical techniques proposed for the presentation of textual search returns. The criticism stemmed from the fact that although many of the techniques were visually appealing 2D, 2.5D, 3D and gravity/repulsive multi-dimensional approaches they lacked evidence for their ability to truly allow the user to **visually** discern groups (clusters) of topically related documents apart given the underlying need to identify the documents that best realize a better task outcome. In partial response to this observation, other work by Pfitzner and Powers (Pfitzner & Powers 2004) proposed a grid-based visual-clustering technique, described as "Vedges" (**V**ector **edges**), that allows the user to make relevance judgments on clusters presented against six dimensions as opposed to the textual list approach, or 2D, 2.5D, 3D and gravity/repulsive multi-dimensional approaches.

During the development of Vedges, it was realized that any truly graphical approach can only serve as a device that visually communicates simple characteristics of visual objects. However, in the process of making decisions to fulfill an information requirement the user needs to make fine-grained contextual decisions against topic/content characteristics of individual or groups of documents.

The effective communication of information via any medium (in this case the visual medium) requires the appropriate use of a conduit language to ensure the user can identify that data critical to the completion of a task or sub-task. The devices (not including text) used by graphical search interfaces being iconographic/semiotic in nature are linguistically low in resolution and so can only communicate a limited set of simple concepts like size, magnitude and relatedness. To describe or discern the difference

between documents or groups of similar documents the conduit language needs to be able to visually represent subtle differences of a complexity only available to textual languages. In short, basic graphical objects can be used to rapidly communicate gross differences between textual objects and *words* can be used to communicate fine-grained differences between them.

The whole point of using technology to search for textual data is that it should make the process more efficient (i.e. easier, more accurate and faster). However, the manner in which documents or groups thereof are describe using *words* will affect this efficiency. For example if one word is used to visually describe a document the user is not going to have enough information to correctly classify it or even complete the task, at the other extreme if the whole document is used the user will spend far to much time reading individual documents to identify classifying features. Somewhere along this continuum, is an optimal descriptor length, but where?

The process of identifying useful classifying words is well researched (for a general review see Baeza-Yates and Ribeiro-Neto (1999)), however traditional search systems use techniques that employ fixed heuristics (not based on user research) to guide the selection of classifier words and calculate their weightings. For example, the most popular weighting scheme used to find the most the characterizing words of a document is one known as TFIDF (Text Frequency Inverse Document Frequency). This scheme is a fast calculation that weights the words of a document given their raw document frequencies correct by the reciprocal of the number of documents they occur in across the total corpus. Mathematical speak aside, this type of calculation is the most common type of calculation, variants of which are used by all the major search engines, however it does not rely on any model of cognition or recognize in any way user capacity limits or tendencies.

Despite this lack of a valid cognitive model justifying the use or applicability of TFIDF there is no research into what positive or negative effects such fixed heuristics might have given users' will have varying information requirements, cognitive tendencies/abilities/preferences and language usages. This comes from the apparent observation that users are not homogeneous, having different cognitive traits and tendencies, and will often react differently to the same situation/question/information need so will require a system that allows for their tendencies and/or variances of ability. Simply

put, TFIDF does not and can not reflect knowledge of intent or individual ability and experience.

With respect to user cognitive ability (see Section 3.1) there are clearly limitations regarding the number of *chunks* of information (words) they can optimally manage at any one time (e.g., 7 ± 2 or 4 ± 1). These limits can also be described as preferences because when a reduction in task performance is noted, for a given task, it can be unclear whether a biophysical limit has been realized (e.g. the user naturally manages 4 chunks not 7) or a personal selective preference/tendency has been realised (e.g. the user is normally a bit lazy so does not search as far down a list before reformulating the query). The implication of such user limitations is that for any system to promote the best possible task outcome it either must allow for such user characteristics/limits by applying an appropriate user model or reliably identified general user tendencies.

Thus, we come to the research contributions of this thesis.

- The first contribution is an extensive and thorough literature review of the cognitive factors that influence the interactive information retrieval process.
- Next the empirical component of this thesis investigates the number and type of words needed to best describe documents individually and in clusters.
- Lastly, a theoretical chapter discussed clustering comparison measures and their shortcomings, before introducing a novel clustering comparison measure.

Basically, this finds its origins in the earlier suggestion that the design of "the ultimate search system" will include the presentation of document clusters that allow the user to optimally reduce the return set by throwing away clusters of documents (topically related) which have been selected primarily using cluster descriptors or by drilling down and using the document descriptors within a cluster.

The main hypothesis of this thesis regards the number and type of words and is divided into the following two parts:

1. Because the popular TFIDF like weighting schemes are based on frequency statistics and not an appropriate user model or reliably identified general user tendencies they will produce ranked list of words for documents the heads of which do not match those a user might produce for the same documents. Thus the types of words users use to describe a document will be different from those produced by the commonly used automated processes.

2. Given researched cognitive limits such as those represented by the magic numbers 7 ± 2 or 4 ± 1 (see Section 3.1.1) and their associated chunks of information users will prefer document descriptions of between 1 and 9 characterizing words (chunks). Within this range the tendency is more likely to be lower given the human bias toward energy conservation in activities like search, as demonstrated by O'Brien and Keane (O'Brien & Keane 2007). In other words users' will tend to use as few words as possible to describe a document. Related to this bias is the tendency of most users to select the first member of a search returns list without any real inspection of data presented. After this initial selection they, in a similar manner, sequentially select down the list until they reach some threshold at which they alter their search technique to a more energy consuming approach. These approaches see the user surveying in more depth the associated snippets for each entry before selecting.

To test this hypothesis a series of 4 surveys, the **Nwords** surveys, were designed to gather data in a "de-contextualized" manner. By de-contextualized it is meant that the experiments are designed so that there are no underlying mechanisms, such as fixed heuristics, that might result in data that is only relevant to a certain mechanism. This concern is the result of the observation that user models are often tested in such a way that underlying mechanism are likely to introduce contextual effects making it difficult to prove any postulate beyond the specific system (see Section 5.1). An example of this can be seen in a popular technique used to produce user Web search statistics known as Transaction Log Analysis (see Section 5.5). The main problem in this situation is that the search engine directly affects the success of any text search task through the mechanisms that deliver and order a set of results. Different search engines deliver different orderings demonstrating that the result lists are directly impacted by internal heuristics such as term/phrase weighting schemes, stopping techniques and

stemming techniques. At a research level the effects of such mechanisms are impossible to predict making the search engine itself a variable that needs strict controlling or outright removal from the process.

The last part of this thesis looks at comparing *clusterings* for the purpose of identifying which clustering approaches are best used in the creation of document clusters for the user cluster filtering (throwing away) approach described earlier. Given the user filtering process the set of document clusters (clustering) used should be comprise of clusters that relate in a manner the user might reasonable assume such as by the topic content a user is likely to describe for a document or group of documents. That topic content the user might realize is important, given part 1 of my thesis suggests that automatic approaches might realize different keywords than a user. Therefore, the comparison of automatically generated document clusters should be conducted against manually generated "Gold Standard" and the results of different clustering approaches compared to see which best match the "Gold Standard".

Finally, it is hoped that this research will lead to improvements in both the manual search return filtering process and reduction in machine process overheads realized by automatic clustering approaches. A critical problem of automatic clustering approaches is that they are renowned for their processing overheads which are typically in the range of $O(n^2)$ to $O(n^3)$. Such orders of magnitude are not practical when operating on return sets of typically a million documents consisting of approximately 800-2000 words per document. Because the clustering problem is such a complex problem if it can not practically be streamlined to anything less than such processing magnitudes the logical solution is to reduce the number of dimension used to cluster against (n) as much as possible. This can be achieved by only clustering against those dimensions that a user needs to determine the topic of a document because these are the only dimensions needed relative to the user's task. In this manner processing overheads will **not** be determined by all the words in all the documents in a return set but by the top say $1 - 9$ keywords of all the documents in the return set.

1.2 Map of this thesis

Chapters 2, 3 & 4 review aspects of cognition relative to user interaction and the task of visual search.

2 Cognitive Information Processing looks at those cognitive mechanisms that impact the user's decision making.

3 Cognitive Limitations and Load discusses user cognitive limitations that give an indication as to how many words a cluster or document descriptor should contain.

4 Visual Processing extends the discussions of the previous chapter by looking at the effects the visual system has on the interactive search/filtering task.

Chapters 5, 6 & 7 constitute the empirical contribution of this thesis.

5 Modeling Users looks at user modeling in the context of the document search task and the understanding of their internal processes and preferences.

6 Nwords describes the Nwords surveys, outlines the results and discusses how the results support the two parts of my thesis.

7 Rwords & Infields discusses extra research needed to support the design of the Nwords survey and investigate a potential problem with the design of the Nwords interactive interface to ensure the validity of any claimed postulate.

Chapter 8 Comparing Pairs of Clusterings reviews the field of clustering comparison, describes the key approaches of the field, lists a number of recognized and common measures and proposes a desiderata of desirable traits a clustering comparison measure might have. Subsequently, a new measure for the comparison of pairs of clusters is proposed and evaluated against those measures presented earlier using a specific set of five tests.

Chapter 2

Cognitive Information Processing

During interactive search, the user must make decisions to guide the process to a subjectively and contextually appropriate outcome. Decisions are the outcomes of mental processes leading to the selection of a course of action from among several alternatives. Knowledge is the basic component upon which the individual applies some form of weighting or selection scheme that allows them to arrive at a preferred path/choice (decision) among all those available. The acquisition of knowledge is reliant on complex cognitive processes such as perception, learning, communication, association and reasoning, and is based on information gathered from the individual's environment or from their own internal store of information.

To develop data presentation techniques it is evident that appropriate understanding of those cognitive factors that might impact the decision making process. The following Chapter discusses cognition relative to the processing of information specifically memory, attention and cognitive styles, and any notable impacts on the task of interactive search.

2.1 Memory

Memory is simply a mental capacity or faculty of retaining and reviving facts, events, impressions and other such perceptions, or of recalling or recognizing previous experiences. However, this simplistic description does not account for the important and intricate relationship memory has with the cognitive processes that manage and process stimuli.

The key model of memory is the "Modal Model of Memory" proposed by Atkinson & Shiffrin (1968) which distinguishes between the three memory types or modes of *sensory memory*, *short-term or working memory* and *long-term memory*. The basic model presented in Figure 2.1 describes "information processing" as an integrated model of memory for the human cognitive architecture. This model assumes information flows from the environment through sensory stores, which are parts of the perceptual system, into a short-term store. This model also includes two important limiting factors. These are that short-term memory has limited capacity (nine elements of attended information at anyone time as described by Copper (1998)), and the longer an item is resident in this store the more likely it will be transferred to long-term memory.

In the context of text search, research into word learning suggests that there are three processes by which the promotion of transference occurs. These are by observing printed characteristics (shape, font, upper or lower case), through acoustical training (reading it aloud), and, the strongest of these, through the process of making judgments about meaning and the relation between the text and a pre-existing concept or experience.

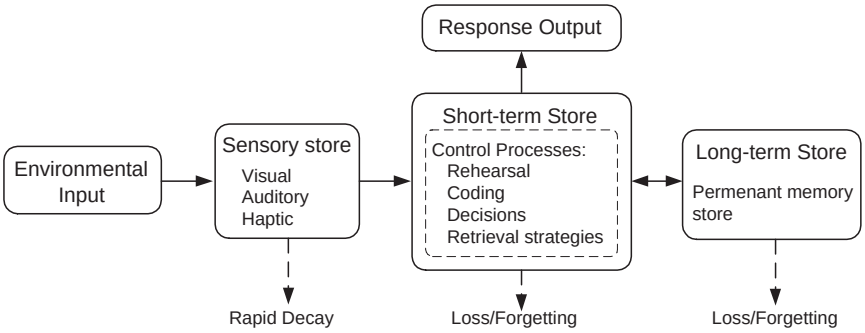


Fig. 2.1: Atkinson & Shiffrin's Modal Model of Memory

Storage and processing functions of working memory are partly distinct, because short-term storage of information does not necessarily interfere with concurrent processing (Baddeley & Hitch, 1974; Baddeley, 1986; 1990; Halford, 1993; Halford, Bain, & Maybery, 1984; Halford, Maybery, OHare, & Grant, 1994; Klapp, Marshburn, & Lester, 1983). Because of this, Baddeley (1986) postulated three systems, a visuospatial scratchpad, a phonological loop, and a central executive.

In forming these into a model Baddeley & Hitch (1974) seem to have originated the

term "working-memory" when they proposed a more complicated model for the short-term store in which they renamed it as working-memory. Of course, this model was composed of three main components: The central executive that acts as supervisory system and controls the flow of information from and to its slave systems, the phonological loop and the visuospatial sketchpad. The two subsidiary systems, the visuospatial sketchpad and the phonological loop, hold memory traces and engage in the rehearsal of information received. Repeating information input into these two subsidiary systems enable repeating executions, which after going through the central executive for a number of times, the information is better learnt and stored in the long-term store.

More recently Baddeley (2000) added the episodic buffer to his model as a third slave system. The episodic buffer comprises a "limited capacity system that provides temporary storage of information held in a multimodal code capable of binding information from the subsidiary systems, and from long-term memory, into a unitary episodic representation".

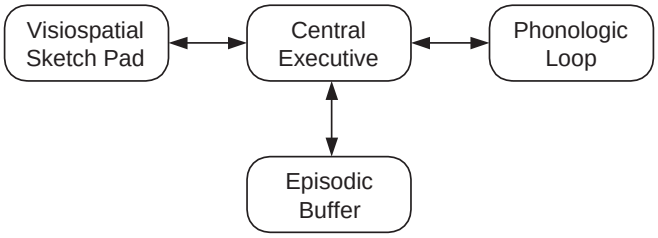


Fig. 2.2: Baddeley's model of Working Memory including Episodic Buffer

2.1.1 Sensory Memory

Sensory memory describes the process of managing information in its initial receipt. Much like the input stage of a computer task, information is input via an *input device* before being processed through to the *hard drive*. Humans have five key input devices with the physical world, which are the five primary senses of the nervous system (hearing, taste, touch, smell and sight). As a form of high speed input caching this memory is highly volatile, being lost unless attention is immediately directed toward the input source.

The stimuli sensory memory deals with sights, sounds, smells, touches, and tastes. The torrent of sensory information that is constantly demanding attention requires

rapid filtering for relevance for further processing resulting in a rapid turnover of these memories, for example, visual information is only retained for about half a second while auditory information last up to three seconds. Unless this sensory information is attended (i.e. identified, classified and meaning assigned) within a critical time span it is dumped/forgotten/lost as sensory memory is being constantly overwritten by new input. This volatility easily demonstrated with a practical experiment, simply close your eyes and you will note that the bulk of visual image vanishes quickly with a ghostly afterimage remaining. This almost instant loss demonstrates how quickly most sensory information is discarded while the afterimage demonstrates your visual short-term memory. This overwriting mechanism is, as previously stated, necessary because of the vast quantity of data involved, in an image and the continuous changes in that image. This has implications for graphical user interfaces and multimedia: if images are not displayed long enough, we will not be able to extract much information from them.

2.1.2 Working Memory

Generally, *short-term* or *working memory* is that part of an individuals information processing system that supports management of relevant information, that has passed from sensory memory, when it is no longer immediately needed to guide behaviour but may be needed in the near future for further processing. It is suggested to be an executive control function and has been shown to play a key role in goal-directed control of attention (Baddeley 1986, Desimone & Duncan 1995). The contents of working memory are discussed by Baddeley (1996) as being used in combination with stored knowledge from long-term memory which can be manipulated, interpreted and recombined to develop new knowledge, form goals, and to assist learning and interaction with the physical world.

A good definition for working memory is that by Stoltzfus (1996) who defined it in a general sense as a "mental workspace consisting of activated memory presentations that are available in a temporary buffer for manipulation during cognitive processing".

Working memory is often discussed as being a combination of both storage and processing functions, that allow for the temporary maintenance of active representations in memory and the manipulation of these representations in the service of current pro-

cessing. Tasks such as language comprehension require complex processing of current information on an ongoing basis and the preservation of continuity with previous information at all times. The result is a need for efficient operation of both the processing and storage components of working memory. Demand on working memory will vary situationally and between individuals on a continuum of expert to novice.

There seems to be general agreement that working memory plays a critical role in cognitive processing, however until recently there seems to have been uncertainty as to how best to conceptualize working memory and the role it plays in different cognitive activities. However, recent research like that of Lavie and DeFockert (2003, 2005) is starting to address this issue in their research into effects of perceptual load and those of target-stimulus degradation on distractor processing. They look at the role of working memory in distractor management. In this they suggest that distractor processing depends on the extent to which high perceptual load exhausts attention in relevant processing, and provide a dissociation between perceptual load and general task difficulty and processing speed.

An important refinement to the scope of memory was made by Cantor (1991) in research conducted using 49 undergraduates to test short-term memory span and complex working memory using short-term memory probe-recall tasks. The study assessed the relationships among short-term memory, working memory, and verbal ability. Results indicate that *short-term memory* and *working memory* are *separate* cognitive constructs, and that both short-term memory and working memory are important to verbal abilities. As is evident in this case the researchers define short-term and working memory as being separate mechanism, however it is considered that, as indicated earlier, the storage and processing is often considered to be sub mechanisms of the one process.

Once information passes through sensory memory and is deemed important by the controlling factor of the brain it moves into *short-term* or *working memory*. At this stage this information becomes part of a rich stream of subjective information and knowledge available for processing until attention is focused on another subject. This stage can be likened to computer RAM with limited capacity but designed for rapid I/O to service the current processes and like data in RAM if information is not moved to a more long term memory it can be overwritten and lost if not rehearsed. The life

time of this memory is between 15 and 30 seconds on average.

In interactive tasks with more than one stage or sub-task any information that may be required for decision making throughout the task or that becomes pertinent from one stage to make decisions in subsequent stages a recognisable artifact representing that information should remain on the screen or be represented at regular intervals or when pertinent to assure its presence in short-term memory when required.

Alan Baddeley (1986), proposed three different subsystems of short-term memory that can be considered when designing interactive task processes.

Speech system We sometimes subvocalize or whisper to ourselves to remember things like names, addresses and number sequences. This might be a consideration when audio is available in the task to maintain or reinstate pertinent information back into short-term memory to assist processes like decision or long term memory creation.

Spatiovisual sketchpad This subsystem is said to be used where we are trying to remember scenes, schema or codification, we have perceived as a whole.

Central executive The main unexplored part of short-term memory that contains short-term controls and cognitive processing.

The maintenance of information within short-term memory and realization and clarification of information in long term memory requires repetition and organisation. Repetition not in the form or immediate repetition but the revisiting of information as it starts to wane from short term memory and after it has ceased to exist in it by regular and conscious repetition. Organisation being achieved through the conscious processing of memories to link and categorise in with previous knowledge.

Miller (1956), when working at Bell Laboratories, collated experiments demonstrating that short term memory was limited to 7 ± 2 items. More recent estimates of the capacity of short-term memory are typically less than this like that of Cowan (2001) who suggests a limit of 4 items. It is also suggested that memory capacity can be increased through a process called chunking. This is discussed further in Section 3.1

The general importance of short-term memory to task completion in interactive computer tasks is that:

- Short-term memory allows one to recall something from several seconds to as long as a minute without rehearsal. If sub-tasks can be guided to completion within this period task success is more likely.
- If rehearsal is allowed, information may be remembered for even longer periods. This implies that if critical queues are presented either where it is rapidly accessible or re-presented on a regular basis it will be more likely available to the decision making process and thus increasing the likelihood of task success or quality.

2.1.3 Long-term Memory

Long-term memory is basically permanent memory, equatable to that of the information stored on a hard drive in a computer, available to us for a relatively long period of time. Retrieval of information from this mechanism is relatively fast and as Cooper (1998) describes the more frequently information is accessed such as names and phone numbers the faster the retrieval.

This form of memory is composed of several separate systems and is often described as being comprised of two major categories of memory; declarative memory and non-declarative memory. Declarative memory refers to the aspect of memory that stores facts and events where as n on-declarative (procedural) memory is the memory of skills and procedures. The relationships between these systems and structures is described by Figure 2.3 which presents a taxonomy of long-term memory proposed by Squire and Zola-Morgan (1991).

Retrieval of knowledge from long-term memory can be a complex process, for example we have all struggled to remember something, taking minutes to actually retrieve the knowledge we sought. However, between the time of initiating the retrieval and actual retrieval attention would have been devoted to other matters which points a background memory processor being invoked to affect difficult memory searches. According to the information-processing model, the retrieval process is simply a function of the cognitive processor however, as already mentioned, frequency of use plays a role as frequently or recently used items are more rapidly recalled. In these situations, both recognition and recall happen quickly and instantly. However, in the “tip of the tongue” situations, there is a noticed difference between the activation of the memory trace by cues (recognition) and the actual retrieval of the information (recall). In the light of

recognition and recall there is also evidence of a "spreading activation" as remembering of a fact often helps the recall of other related items (Sutcliffe & Slater 1995).

As demonstrated by Figure 2.3 long term memory can be categorized as Declarative, Procedural and Imaginary, however imagery/imaginary memory is often ignored in discussion.

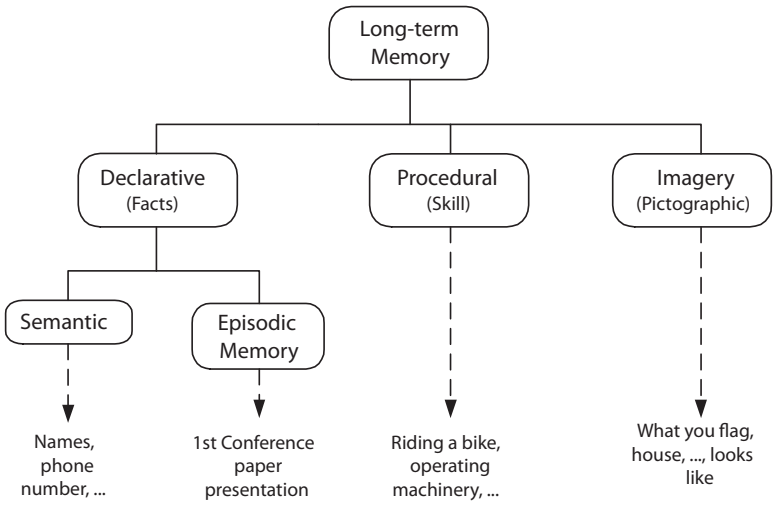


Fig. 2.3: Long-term Memory

Declarative Memory

Declarative memory is used to identify and categorize objects and events, and was subcategorised by Tulving (1983) into the categories of Episodic and Semantic memory.

Episodic memory refers to autobiographical memory for events that have a particular temporal and spatial context. Most PhD candidates can relate to this concept via memories of both those nervous feelings when delivering their first conference paper and the details about the key events of the conference. These types of memories belong to episodic memory since they are both related to a specific time and place. Episodic memory is best described by Tulving (2002),

> Episodic memory is a recently evolved, late-developing, and early-deteriorating past-oriented memory system, more vulnerable than other memory systems to neuronal dysfunction, and probably unique to humans. It makes possible mental time travel through subjective time, from the present to the past, thus allowing one to re-experience, through autonoetic awareness, one's own

> previous experiences. Its operations require, but go beyond, the semantic memory system. Retrieving information from episodic memory (remembering or conscious recollection) is contingent on the establishment of a special mental set, dubbed episodic "retrieval mode".

Semantic memories are generally factual memories about the world, including those that derives from particular events. Like names for things and place, they have very little associated context as they appear to the rememberer simply as known facts. In other words the rememberer recalls knowledge but cannot recall the context of its initial learning (Tulving 1983).

Table 2.1 presents a comparison between the characteristics of episodic and semantic memories. By comparing the episodic and semantic columns the difference between the autobiographic nature of episodic memory and the factual nature of semantic memories is clear.

Comparison of Episodic & Semantic Memory *source: http://www.dushkin.com/connectext/psy/ch07/table7.mhtml*		
Characteristic	Episodic Memory	Semantic Memory
Source	Sensation	Comprehension
Units	Events	Facts or Ideas
Organization	Temporal	Conceptual
Reference	Self	Universe
Registration	Experiential	Symbolic
Temporal	Present	Absent
Affect	More affect	Less affect
Vulnerability	More chance	Less chance
	of disruption	of disruption
Access	Deliberate	Automatic
Queries	Time? Place?	What?
Reports	Remember	Know
Development	Later in life	Early in life
Amnesia	Affected	Unaffected

Table 2.1: Episodic & Semantic Memory comparison

Procedural Memory

Procedural memory, also known as implicit memory, is the long term knowledge store of skills and procedures, or "how to" knowledge. The type of knowledge associated with this mechanism involves more than one type of sense as it is directly applied in performance of different tasks. It is a step-by-step type of knowledge that describes how to realize a certain accomplishment, like riding a bike or driving a car, which involves previous experiences to aid in the performance of a task without conscious awareness of these previous experiences.

Imagery Memory

Marschark et al. (1987) point to empirical findings from studies of memory for word and sentence lists, language comprehension and memory, and symbolic comparisons in support of the suggestion that verbal and imaginal (imagination, images, or imagery) processing systems may operate in conjunction with generic semantic memory. This type of memory is a pictorial view of the things we have seen or imagined, for instance your country's flag can be brought to your mind or that you can imagine a country's flags, being described to you, as an image.

2.1.3.1 Categories of Long Term Memory

Stage Theory

A widely accepted theory describing the manner in which we process information is "stage theory". Proposed by Atkison & Shiffrin (1968), it focuses on how information is stored in memory proposing that stimulus information is processed and stored serially, and discontinuously in three stages. These stages are presented in Figure 2.4 as *Sensory Memory* or *sensory registration*, *Short-Term Memory* and *Long Term Memory* (see Sections 2.1.1, 2.1.2 & 2.1.3). For information to be available in the medium to long term the information must be passed through all stages as information stored in long-term memory is said to be permanent. If information is not passed to long-term memory the individuals memory of it will decay relatively fast.

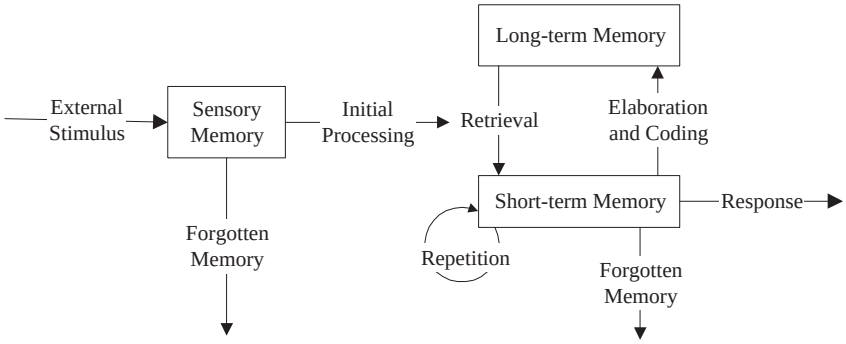


Fig. 2.4: Stage Model of Information Processing

This type of theory is often referred too as a *dual store* model (e.g. Modal Model of Memory in Section 2.1) where there are two main components of memory, that of short term and long term memory with the sensory memory considered as a volatile holding mechanism that facilitates processing. As with the stage theory these models deal with the structures of the human mind such as sensory memory, short-term and long-term memory. The structures are central to the operation of several critical processes such *attention*, *organisation*, *retrieval*, *maintenance rehearsal* and *storage*.

In short sensory memory is abstract with no associated meaning, if any meaning is realised it is created when the information reaches the central cognitive short-term memory for subsequent interpretation. The cognitive processor is responsible for object identification. The cognitive processor has an associated short-term memory used for storage of temporary working information. This information can be extracted from

the sensory processors or the long-term memory. In the modal memory model, all the short-term memories are referred to as working memory. The cognitive processor performs most of the thinking activity. The results of thinking can either be placed back in short-term memory, stored in long-term memory or passed on to the motor processor for the elicitation of behaviour.

Evidently, for interactive tasks like search, if information is pertinent to the completion of the task, especially the immediate sub-task, its screen artifact (representative metaphor) needs to remain while it is contextually relevant or until it is no longer required for the completion of the immediate task.

Levels of Processing Theory

Proposed by Craik and Lockhart (1972) the *level-of-processing* model suggests that information is processed in a number of different ways and the durability or strength of the memory trace was a direct function of the depth of processing involved. Craik and Lockhart postulated that as information is processed the individual applies different levels of elaboration. This elaboration occurs on a continuum of *perception*, *attention*, *labeling* and *meaning*. In other words, memory was the result of a successive series of analyses, each at a deeper level than the previous e.g. a shallow level might see the focus on the sound of a word while at a deeper level of processing the focus might be the words meaning.

In developing this model Craik and Lockhart make the assumptions that the deeper the level of processing, the more durable the resulting memory and that rehearsal can be relatively unimportant (e.g. a lot of rehearsal using a shallow level of processing might lead to worse memory than much less rehearsal using a deep level of processing).

Supported by a level of forensic psychology (Huitt 2003) Craik and Lockhart's key proposition was that all stimuli resulting in the activation of a sensory receptor cell are permanently stored in memory and that the access to these memories are directly affected by the level of elaboration involved. Further support for this model can be seen in Rumelhart and McClelland's (1986) Connectionistic approach (see below) that emphasizes the idea that the more connections involved in a memory the more likely it is to be remembered. Bransford (1979) later extended this model applying it to information access as well as information processing.

Parallel Distributed Processing

The theory of memory management known as Parallel Distributed Processing is an alternate view to the sequential views proposed by Atkison and Shiffrin (1968), and Craik and Lockhart (1972), that proposes that memory is managed in a concurrent manner. Rumelhart and McClelland (1986) proposed this approach suggesting that the processing of information takes place through the interaction of a large number of sectors organized into modules. Much like *neural networks* in computer science storage of memory occurs through the modification of connection weights based on the system's response to its input, that provides an opportunity for incremental storage. This model originated from the observation that systems of neural connections appeared to be distributed in a parallel array as well as several serial pathways. This implied that different types of mental processing are distributed throughout a highly complex neuronetwork (McClelland 1994).

In proposing this model McClelland also outlined eight components:

1. a set of processing units
2. a state of activation
3. an output function for each unit
4. a pattern of connectivity among units
5. a propagation rule for propagating patterns of activities through the network of connectivities
6. an activation rule for combining the inputs impinging on a unit with the current state of that unit to produce a new level of activation for the unit
7. a learning rule whereby patterns of connectivity are modified by experience
8. an environment within which the system must operate

Dictating the function of the model are three principles: the representation of information is distributed; memory and knowledge for specific things are not stored explicitly, but stored in the connections between units; learning can occur with gradual changes in connection strength with experience.

Connectionistic Approach

The connectionist model was proposed by Rumelhart and McClelland (1986) as an

extension of the parallel processing approach. It builds on the assumption that the physical structure of the brain, or its architecture, allows for the efficient processing of information. In proposing the model, Rumelhart and McClelland suggests the brain itself doesn't "know" anything rather knowledge emerges from the way in which information is processed (stored and retrieved). Given this structural view this approach emphasises that information is stored throughout the brain in multiple locations via a network of connections and that the more connections involved in a memory the more likely it is to be remembered it complements the levels-of-processing model.

In terms of the interactive task of search this model can be likened to Brin and Pages (Brin & Page 1998) approach to link weightings for improving search return results on the web. Basically they gave pages with more inward pointing links greater weights so for any given set of search terms a list of appropriate documents would be returned to the user ordered in part using the individual page weightings.

2.1.3.2 Memory Structures

Schematas are mental structures that represent some form of knowledge/understanding about the individuals world and are used to organize information and provide reference for understanding. They represent our knowledge about all concepts such as those underlying objects, situation, events, sequences of events, actions and sequences of actions. Examples of schemata include rubrics, stereotypes, social roles, scripts, world views, and archetypes. Schemata are said to be the basis for all understanding/knowledge about our world as is reflected in Piaget's (Ginsburg & Opper 1988) theory of development that proposes that children adopt a series of schemata to use in their management and realisation of understanding of their world.

Bartlett (1932) proposed the concept of schemata from studies of memory he conducted in which subjects recalled details of stories that were not actually there. From this he concluded that people must create a mental model or structure that they use as an aide for remembering. Key treatments of the concept of schemata are those by Mandler (1984), Rumelhart (1980) and Bransford & Franks (1971). Mandler expanded on a series of lectures discussing types of mental structure (such as categorical, matrix, serial, schematic, and story structures), story schemata and processing (specifically the psychological reality and psychological validity of story schema, and hierarchical

structure and the "levels effect") and the nature of scripts and scenes (including the structure of event and scene schemata, and script and scene structure and processing). Rumelhart studied the development of schemata describing them as the "building block of cognition" and proposed the concept of "tuning" as the evolutionary mechanism for schemata. Bransford & Franks demonstrated the concept of "idea acquisition and retention" experimentally in contrast to an "individual sentence memory" point of view and demonstrated that participant confidence in having heard a particular sentence is a "function of the degree to which a sentence fails to exhaust all the semantic relations characteristic of a complete idea" (p.331).

Schemata as a basis for expertise

In researching the acquisition of expertise, Chi, Glaser & Farr (1988) proposed an extension to the concept of schemata regarding learning and the differences between novice versus expert performance. The proposal was basically that experts have a set of schemas that guide perception and problem-solving which novices do not have. They suggested that increased performance of experts is evidence that new schemata are developed in long term memory through learning. Psychological studies demonstrate this principle by tracking the improvement from inefficient, slow, and frustrating to fast, and efficient. The change in performance occurs as the learner becomes increasingly familiar with the material, the cognitive characteristics associated with the material are altered so that it can be handled more efficiently by working memory.

An interesting result of work by Sweller (1988) was his Cognitive Load Theory that combined Miller's work with the schemata theory. In research into problem solving by learners, Sweller recognized that learners often use a problem solving strategy called means-ends analysis. This type of analysis requires a relatively large amount of cognitive processing capacity, which may not be devoted to schema construction. Recognizing that Miller's (1956) review suggested short term memory is limited in the number of elements it can contain simultaneously, Sweller theorized that schemata, or combinations of elements, as the cognitive structures that make up an individual's knowledge base. Simply put, schemata become chunks for expanding memory. As a result instructional designers should limit cognitive load by designing instructional materials like worked-examples, or goal-free problems.

2.1.3.3 Knowledge and Mental Models in HCI

A **mental model** is basically a description in someones mind that represents how something works in the real world. These representations are suggested to play a major role in cognition and decision-making. They are basically a psychological transformation by which an individual can acquire, code, store, recall, and decode information about phenomena. Importantly, once a model has been formed it can be used to replace consideration and analysis in order to conserve time and energy. In this section we use "mental model" to mean the same as any of cognitive map, mental map, mind map and cognitive model.

In the context of this work mental models are a key to predicting and understanding Human-Computer Interaction. However, the complexity and variability of human behaviour is difficult to describe using formal models (Suchman 1987) such as mental models. Norman (1983) nicely captures this fact when describing the properties of mental models as "contradictory, incomplete, superstitious, erroneous, and unstable, varying in time" (p.14). He also aptly expresses their importance when he states "In interacting with the environment, with others, and with the artifacts of technology, people form internal, mental models of themselves and of the things with which they are interacting. These models provide predictive and explanatory power for understanding the interaction." (p.7)

Drawings by M.C. Escher are commonly used as examples of the influence mental models have on an individual's interaction with the world. For example, Escher's 1961 Litograph of a waterfall (see Figure 2.5) clearly does not conform to visual expectation. Although, the first impressions of the lithograph are of a normal scene, upon further inspection one realizes that the water is traveling up hill to fall back to its origin to start the process again. Because the water does not perform as expected the visual imagery causes confusion because it does not fit our mental model of what water should do.

Boltzmann (1899) may have been the first to make use of a concept like a mental model is his statement "All our ideas and concepts are only internal pictures". However, Craik (1943) was the first to suggested that mental models are "small-scale models" of reality that are used in the process of reasoning about phenomenon. Basically, the mind produces a model of reality and uses it for reasoning, explanations and anticipation.

Fig. 2.5: Example of how mental models can confuse
Sourced from www.math.technion.ac.il/rl/M.C.Escher/2/

These models can be constructed from perception, imagination, or interpretation of discourse. A mental model represents explicitly what is true, but not what is false. Craik also suggested that the greater number of mental models needed to explain a phenomenon, and the greater the complexity of every model, the poorer performance is likely to be - a claim later supported by Johnson Laird (1983).

Models of the human-computer interface depend heavily on cognitive psychology. The psychological processes of attention, memory, information processing, decision making, and problem solving must be taken into account.

Models of human performance permit aspects of user interfaces to be evaluated for usability by making predictions based on task analysis and established principles of human performance (Card, Moran & Newell 1983, John & Kieras 1994, John & Kieras 1996). There have been many theories proposed to account for the low-level strategies that people use to find a known item in an unordered menu. For example Norman (1991) and Vandierendonck, Van Hoe, and De Soete (1988) suggested that people process one menu item at a time. However this was not validated empirically. There have also been conflicting theories, such as that by Card (1984) proposed that people randomly choose which item to examine next, while Lee and MacGregor (1985) provided evidence that people search systematically from top to bottom.

In modeling link evaluation and selection behaviour, Miller and Remington (2005) describe the *Threshold strategy* and the *Comparison strategy*. The threshold strategy sees the user immediately selecting and pursuing any link whose probability of success exceeds a certain threshold. On the other hand the comparison strategy sees the user first evaluating a set of links and selecting the most likely one out of the set. The threshold strategy will only be useful if the user is given enough appropriate information to select on and the list is not too large and is appropriately ordered. The comparison strategy is likely to be the most successful if there are less than thirteen link on the page as demonstrated by Lee and MacGregor (1985).

Johnson-Laird (1983) describes mental models as the basic structure of cognition. He suggested that working models are used in order to understand a phenomenon and argues that the only constraint for a mental model is that it has a similar structure to the phenomenon it represents. They are suggested to be the mapping between *propositional representation* and *mental imagery* (see Fig. 2.6). Without the mapping

one might be able to describe an object but not recognise it, having not previously seen it in context and constructed a mapping from image to proposition. During the mapping process, the mind acquires information about a new phenomena and searches previously stored models for matching semantics. If no model was found, a new model will be constructed and stored with the relevant semantics.

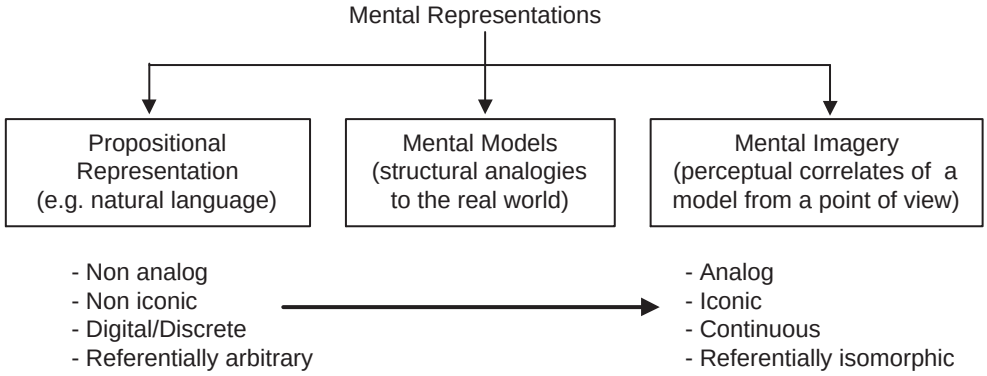


Fig. 2.6: Johnson-Laird proposition of the three types of mental representations
Sourced from www.cs.umd.edu/class/fall2002/cmsc838s/tichi/knowledge.html.

The ideal mental model explains all the aspects of a phenomenon the individual is interacting with. The theory can be summarized in terms of three principal predictions:

1. Reasoners normally build models of what is true, not what is false – a propensity that led to the discovery that people commit systematic fallacies in reasoning
2. Reasoning is easier from one model than from multiple models
3. Reasoners tend to focus on one of the possible models of multi-model problems, and are thereby led to erroneous conclusions and irrational decisions

Since human thought is involved in every day activities, mental models are applicable to almost every human interaction with nature, devices, and even interaction with other individuals. In respect to textual activities and mental models relationship to reading and comprehension Johnson-Laird (1989) suggest that a reader creates a mental model of text being read through a development of an understanding/interpretation. In this process the model is representing the phenomenon being described by the text. However, if the text does not supply enough information to clearly identify one model,

multiple competing mental models may induce reader confusion, a phenomenon less likely to happen with text that elicits one clear mental model.

In regards to text search user knowledge/experience will effect the construction of mental models and subsequent success of any set of search terms. With previous experience models may have been constructed that help guide the selection of both keywords/phrases to use in the search and in the selection of the visual representations of results that are potentially more likely to address the task at hand. However, despite mental models in this situation seeming beneficial, on a global scale if everyone is using the same search engines then they are all being influenced by similar processes such as rank ordering of list and presentation characteristics of the interface. If there are inherent flaws in the techniques used by the search engine in either the input or output mechanisms used then the mental model may become a process impeding the improvement of the system through a lake of logical analysis.

Cognitive Scientists such as Pinker (1998) often describe the mind as a computer and different aspects of mind are simply sub-routines. This has seen the development of many models of human performance by partitioning some aspects of human cognition and behavior, and logical development of a model that describes different inputs to the situation and the resultant responses.

Models of human performance are a common tool used in the development and design of computer interfaces, for example, GOMS (Goals, Operators, Methods, and Selections) (Card et al. 1983) and it's many variants (for example see Rasmussen (1983)). More recently, cognitive models have been used to simulate human capabilities in systems for developing and evaluating user interfaces (Ritter, Baxter, Jones & Young 2000). Deeper models such as ICS (Interacting Cognitive Subsystems) allow examination of the cognitive resources required to operate a particular interface (Bowman & Faconti 1999, Craik 1943, Duke & Duce 1999). These approaches are useful in identifying error-prone features in interfaces to safety-critical systems (e.g., the complex process that must be followed to enter a new flight plan into a flight management system), but they do not seem to address the most worrying kinds of problems: those associated with mode confusions and other kinds of automation surprise.

The Danish engineer Jens Rasmussen (1983) started enlisting more cognitive theory to improve the design of man-machine interface systems and thus help reduce the

potential for accidents. He began by characterising human performance in a familiar environment as goal-oriented and rule-controlled. This lead to the proposal of three qualitatively different levels of cognition in which qualitatively different types of information circulate and qualitatively different types of decision are made. These levels he referred to as "typical levels of performance".

Rasmussen's model (1983) is based on attention and classifies human performance into the three categories of *Skill-based behaviors*, *Rule-based behaviors* and *Knowledge-based behaviors*.

Skill-based behaviors (SBB) are the simplest form of behaviour, they are routine activities conducted automatically that **do not** require the conscious allocation of attention. Behaviors are skill-based when human performance is determined by stored, preprogrammed patterns of instructions. The individual is seldom able to describe how performance behaviour is controlled or what variable performance is based on. Examples of SBBs are bicycle riding or musical performance.

Rule-based behaviors (RBB) are more complex and controlled via a set of stored rules or procedures. The distinction between rule-based and skill-based is dependent on the attention applied and the individuals experience. Performance of rule-based behaviors is typically based on specific ability and the individual can often describe the rules upon which performance is based. Examples of RBBs are mathematical problem solving and system control tasks such as the discrete maneuvering of aircraft or cars.

Knowledge-based behaviors (KBB) are those in which stored rules no longer apply and a novel situation is presented for which a plan must be developed to solve a problem. In contrast to set rules, plans are often required to be changed based on the situation. Attentional resources must be allocated to the behavior and, therefore, the performance of knowledge-based behaviors is goal-controlled. An example of this is when all training fails and things have to be consciously diagnosed and responded to.

2.1.3.4 Cognitive theories of concepts

"Concepts are the glue that holds our mental world together"

(Murphy 2002)

Murphy (2002) describes concepts as mental constructs that tie our past experiences to our present interactions with the world. They embody much of our knowledge of the world, telling us what things there are and what properties they have. Concepts are tied closely to "categories" in that categorization involves the characterization of phenomenon by means of concepts. For example, an individuals concept of bird allows them to select/recognize a finch from/using the category of entities they probably call birds (Prinz 2002). In this manner concepts have been described as pattern-recognition devices that enable us to classify phenomenon and to also make inferences about them (Smith & L. 1999).

Regarding development of theory describing perception the importance of *concept* is captured by Margolis and Laurence (1999) when they state that "*concepts are the most fundamental constructs in theories of mind.*"

In Cognitive Science the study of *concepts* is generally concerned with three issues:

- how concepts are represented
- how we classify instances (exemplars) as belonging to a concept
- how we use concepts in reasoning

The idea that concept can be categorized stems from the idea of "conceptual coherence". This refers to concepts whose contents "seem to hang together, a grouping of objects that makes sense to the perceiver" (Murphy & Medin 1999). The idea of conceptual coherence comes from the notion of similarity in that phenomenon form a concept because they are similar to one another. Drawing from this idea of similarity Murphy & Medin (1999) suggest that "similarity may be the glue that makes a category learnable and useful". Categorisation via concept similarity has been suggested to be the reason behind our ability to make sense of a complex world of inter-related phenomenon in that Concepts give our world stability in that they allow us to treat nonidentical things as equivalent (Wisniewski 2002).

Similarity and cognition In the cognitive sciences *similarity* is suggested to be essential in the process of acquiring and categorizing information. Given the discussion on

the formation and use of mental models in Section 2.1.3.3 it is reasonable to suggest, as Hahn & Ramscar (2001) do, that the acquisition of conceptual knowledge involves the construction of mental representations that can facilitate interaction. The categorization of acquired knowledge sees that the comparing of the new stimuli "to previously acquired knowledge representations, and classifying it according to which pre-existing representation it most closely resembles" (Hahn & Ramscar 2001). The pivotal role of similarity is well demonstrated by Hahn, Chaterb and Richardson (2003) when they point to the fact that it is often a key concept in explanations of concepts such as memory retrieval (Hintzmann 1986), categorization (Hampton 1995, Nosofsky 1986), visual search (Duncan & Humphreys 1992), problem solving (Gick & Holyoak 1980, Holyoak & Koh 1987), learning (Gentner 1989, Ross 1984), linguistic knowledge (Bailey & Hahn 2001, Hahn & Nakisa 2000) and processing (Luce 1986), reasoning (Rips 1975), as well as social judgment (Smith & Zarate 1992).

Like any of the clustering techniques that use thresholds to determine member inclusion or exclusion Rips (1989) describes a simple method by which people might decide whether an object belongs to a category or not. The object is a member of the category if it is sufficiently similar to known category members. To decide whether an object is a category member, start with a representation of the object and a representation of the potential category, then determine the similarity of the object representation to the category representation. If this similarity value is high enough, then the object belongs to the category; otherwise, it does not.

The importance of categories can be seen in Jacob's (1991) discussion on classification and categorization in the statement "By recognizing similarities between potentially dissimilar entities, the individual is enabled to form theories, or models, of his or her environment that allow him or her to extend to new encounters the generalizations garnered from past experience". It is also suggested that categories are used to make inferences or predictions about a phenomenon. For example, if a child has seen several Indian elephants, all being fairly large with four legs, a small tail and a trunk, they would probably draw on the category of elephant they might form to and classify an African elephant as an elephant when they saw one. Spiteri (2007) suggests that in these situations similarity is the primary mechanism used in inductive thinking, "since categories whose members share similar properties have stronger inductive power than categories whose members are less similar" which is also supported

by (Heit 1997, Murphy 2002). In the communal sense a shared understanding of the nature of a category can expedite understanding and facilitate communication. If I say "I have to go home because of my dog" (Murphy 2002), because it is generally understood that dogs cannot and should not be left alone for too long, no explanation is needed.

Concepts seems to be key factors in human knowledge acquisition, understanding and communication. However, there are several theories surrounding concepts, what they are and what, and how they are used. Following is an outline of the main theories and how the relate.

2.1.3.5 Concepts and Classical theory

Concepts are defined by a set of empirically-discoverable (probably not directly observable) *necessary* and *sufficient features*. Under this theory members of a *concept* are those *exemplars* that exhibit the necessary and sufficient features that define the concept, and any exemplars that do not exhibit those features are not members of the concept. Concepts are formed through experience of enough exemplars that allow us to extract sufficient features to divide the exemplars into separate classes. Eric Margolis and Stephen Laurence (2002) suggest that the classical view is an instance of a descriptivist theory of reference, within which concepts refer to descriptions of real-world instances. Concepts in this theory account for *classification*, *category learning* and *concept representation*.

Quine (1951) seems to have been one of the early critics to question the distinction between analytic truths and synthetic truths, who described them in the first instance as truths grounded only in meanings and independent of facts, and in the second instance truths grounded in facts. An example that highlights a problem with classical theory can be seen the concept (word) "bachelor", which is clearly defined by a set of *necessary* and *sufficient features* (male and unmarried). The problem can be seen in the fact that unmarried male children are not bachelors and neither are divorcies. So the definition needs to be amended with the inclusion of the word "adult" (unmarried adult male). However, the problems continue as you need to address other individuals that break the rule like middle-aged gay men and priests. Unfortunately the list of disjunctive rules that would be needed to fix the problems of this concept would be huge and the

same problem manifests with many everyday concepts.

There has been a lot of research that demonstrated this theory as lacking. For example, Fodor, Garrett, Walker & Parkes (1980) demonstrated a lack of evidence for the existence of "definition" (a composition of concepts). Another example can be seen in work by Rosch (1973) which demonstrated that some exemplars are treated as better members (more *typical*) of a concept than others. The persistence and predictive power of *typicality* poses a serious problem for any definitional account. If concepts are represented only by their definitions, then exemplars either exhibit those features or do not, and thus are either members of a concept or not (no "gray area"/"partial membership").

Any theories that might replace the classical theory of concepts will need to address the phenomena discovered in the extensive research into its nature over the last century, phenomena such as Typicality effects, Fuzzy boundaries, Relationships between features and Contrasts between categories.

Typicality effects Some members of a category are treated as "better" members than others. This effect embodies the insight of Wittgenstein (1999) and others that concepts may have a family resemblance structure.

Fuzzy boundaries Some explanation of why category membership is not always an either-or relationship.

Relationships between features Eleanor Rosch, using her research on typicality effects and the hierarchical organization of categories, argued that our concepts must represent the structure of the world.

Contrasts between categories Concepts are not learned or represented in isolation. Instead, they are born within a web of concepts, and the relationships between concepts can influence the representation of individual concepts, as well as the classification of individual exemplars as being members of one concept or another.

2.1.3.6 Concepts and Similarity-based Theories

As the title implies, this theory is based on some comparative measurement between concepts. Similarity-based theories began with Rosch describing concepts in terms of

clusters and cluster similarity with the statement (Rosch 1999). There are two key classes of theory in this category *Prototype-theories* and *Exemplar-theories*.

Prototype Theories Rosch, Mervis, Gray, Johnson & Bayes-Braem (1976) proposed Prototype Theory to overcome problems inherent in the classical approach to accounting for the manner in which humans, and perhaps other animals, cognitively manage their perception of the world.

"The world consists of a virtually infinite number of discriminably different stimuli . One of the most basic functions of all organisms is the cutting up of the environment into classifications by which non-identical stimuli can be treated as equivalent."
(Rosch, Mervis, Gray, Johnson & Bayes-Braem 1976, 383)

The theory suggest that instances of natural concepts are defined by their resemblance to a 'prototype' that is a best or most typical example of the concept. So the instance's features are mentally represented and compared to prototype representations, which are basically those items that contain the largest number of typical features (Prinz 2002). The prototype will share the maximum number of features or attributes with other instances of that category and a minimal amount with instances of other concepts. So, prototypes consist of characteristic features rather than strict defining properties, and as such concepts have indistinct boundaries possibly represented by fuzzy sets. Generally an instance of a natural concept can be considered extremely typical, moderately typical, atypical, and borderline and thus concept member's typicality are measure by degree of similarity.

Being based on measures of similarity, prototype theories suggest that all concepts can have varying degrees of membership. For example, a sparrow is a better example of bird than is an emu, because a sparrow is associated more readily with the features that one attributes to birds; likewise, fire engine red is a better example of red than is red hair (Rosch 1999, Mervis & Rosch 1981). Classical Theory, by contrast, suggests that the meaning or definition of a concept should not change according to context (Rosch 1999).

Exemplar Theories According to the exemplar theory, (e.g., Heit & Barsalou 1996; Medin et al. 1984; Nosofsky 1988, 1992) a concept is represented by a set of particular

instances of it stored in memory (Heit & Barsalou 1996, Medin, Altom & Murphy 1984, Nosofsky 1992). In this manner defined characteristics are not needed as it is proposed that an entity is matched to a list of salient features. A new item is categorized as an instance of the concept if it is sufficiently similar to one or more previously-experienced instances via comparison of salient features.

Compared to prototype theories, exemplar theories suggest that people do not have singular definitions for concepts (say one bird definition for example) composed of multiple features found to varying degrees amongst that category of concept. Exemplar theories suggest that a concept is a member of a concept category by comparing it to the set of concepts that have been encountered and remembered previously. So when an individual sees an Emu it is compared to stored memories of other Emu's encountered. If Emu's have not been encountered before, a search of memory is conducted for entities most similar to an Emu (exemplars of) from which an individual is likely to realize similarities between other birds previously experienced (as opposed to say elephants) and conclude that this new entity is probably a bird (Smith & L. 1999).

2.1.3.7 Knowledge-based Models of Concept Formation

Theory Theory of Concepts or Theory-Theory This approach based on the our understanding of our world and the recognition feature or properties and any coocurences. In this manner concepts are learned based on our overall understanding of the world around us by noticing how often properties or features occur and co-occur. Our perception of the salience of features depends on how often we encounter them and their correlations, and on our understanding of why these properties co-occur. In this manner the formation of concepts is influenced by our theories and understanding of how features are related, for example, blackness and roundness are both frequently-occurring features of tires, yet roundness seems to be more central to tires since it is so closely linked to the function of tires (Keil 1989).

The recognition of concepts is influenced by what we already know, however concepts may also have and affect our existing knowledge (Murphy 2002, Rehder 2003*a*, Rips 1989, Slaughter, Jaakkola & Carey 1999). For example, recent experiments in the creation of self-replicating robots could cause us to question our current understanding of the biological function of reproduction. Since concepts should be consistent with

what we already know, we use our prior knowledge to decide whether a new item we encounter belongs in (as part of) an existing concept, or whether it is necessary to create a new concept (Spiteri 2007). An interesting twist to this is that research demonstrates that the essence of an entity can include features that are not readily observed, and that even if observable features change, the essence of the entity remains constant. Furthermore, although we may not be able to define what, exactly, is the essence of a raccoon, we presume that it exists (Keil 1989).

Causal-Mode Theory Like Theory-Theory, Causal-Mode Theory accounts for the effects of theoretical knowledge on our understanding of concepts, however it proposes that a greater emphasis is given to causal knowledge. This casual knowledge drives association between the features of concepts. The knowledge people have of many concepts includes not just a representation of a concepts features but also an explicit representation of the causal mechanisms that people believe link those features. People use causal models to determine a new objects category membership (Rehder 2003*a*).

As seen in Prototype theory the question of feature weighting is not new, however Causal-Mode theory differs by focusing on how feature weights are determined by peoples domain theories. Spiteri (2007) aptly demonstrates this with the following example: "straight bananas are rated as better members of the category bananas than straight boomerangs are of the category boomerangs, a result people attribute to the default feature curved occupying a more theoretically-central position in the conceptual representation of boomerang as compared with banana" (p.13). Features can be combined and certain combinations can affect realization of a coherent concept (Rehder 2003*a*).

2.2 Attention

Attention describes a phenomenon that sees certain capacities of the brain focused on a particular type of incoming information. This information is competing for those processing capacities thus the more attention that is directed toward it the more capacity is being consumed in the process. The classic cocktail party phenomenon is a good example that demonstrates attention. In the cocktail party environment, one has many competing social and environmental stimuli demanding varying levels of atten-

tion and their management. Despite this maelstrom of stimuli, the individual is still able to conduct bi-directional communications.

A well recognised model in the evaluation of attention is that by Sohlberg and Mateer (Sohlberg & Mateer 1989) which hierarchically divides attention into several clear categories. Although this model is based in the recovering of attention processes of brain damage patients after comma, its are suggested to clearly and appropriately categories different types of attention. The model describes five different kinds of attention.

Focused attention describes the ability to respond discretely to specific visual, auditory or tactile stimuli.

Sustained attention is the ability to maintain a consistent behavioral response during continuous and repetitive activity.

Selective attention refers to the capacity to maintain a behavioral or cognitive set in the face of distracting or competing stimuli. It is also suggested that it also incorporates the notion of freedom from "distractibility"

Alternating attention is the capacity for mental flexibility that allows individuals to shift their focus of attention and move between tasks having different cognitive requirements.

Divided attention suggested to be the highest level of attention which is the ability to respond simultaneously to multiple tasks or multiple task demands.

In describing the underlying mechanism of attention there are two main camps espousing theories that attempt to describe this process, that of the *Late Selection theories* and *Early Selection theories*. Early selection theories suppose that we can filter out unwanted material at an early stage of processing. Late selection theories suppose that most material is fully processed, and selection occurs only when we come to make a response.

An important trait of attention is that it is difficult to sustain as can be seen through substantial deterioration of task performance over time. This phenomenon is known as Attentional Degradation (see Section 2.2.2) that describes difficulties of attentional

maintenance and its resultant degradation that are most profound in vigilance task such as critical inspection processes. These are task such as manual security inspection, air traffic monitory and manual quality control. This type of attention is normally required for more than twenty minutes.

The importance of attention in interactive tasks is highlighted in Treisman's (1988) suggestion that attention is required for the integration of stimuli to form more complex concepts, that is, it is needed to combine separate features into one unit such as a chunk (see Section 2.2.2). This combining is critical to future performance in task achievement and realization of relevant schema, and other associated memory structures. If task relevant schema are not appropriately formed an individuals ability to bring past experience to bare on a task is hampered.

With regards to the control of selective attention there are two processes that represent the bidirectional nature of perceptual information flow and processing, that of Top-down (concept driven) processing and Bottom-up (stimulus driven) processing.

2.2.1 Early and Late Selection

The ease with which a task is completed varies and a key problem that impacts task completion is that of focusing as much cognitive power as is possible on an appropriate stimuli while not wasting it on irrelevant stimuli. In other words the ability to maintain focus on task relevant stimuli while ignoring task irrelevant stimuli is the key to any coherent cognitive function. There are two main theories addressing this issue that of Early and Late Selection. *Early selection* theories suggest that we can filter out unwanted material at an early stage of processing. The model suggests that attention shuts down processing in the unattended receptor before the mind can analyse its semantic content.

Treismans proposed *attenuation* theory to explain a set of interesting results. According to this theory, physical characteristics are attenuated early in the process sequence, but not completely filtered out, and then semantic criterion are applied. The semantic criteria are based on the individual's expectations, and are subject to change.

The *Late selection* theories suggest that most material is fully processed, and selection occurs only when we come to make a response. Content in all target receptors is

analysed semantically, but the concepts in the unattended receptor do not proceed to consciousness.

Deutsch and Deutsch (1963) suggest that filtering occurs later in the processing sequence, after the verbal content of the message has been analysed. They suggested that people can perceive many messages, but they can only respond to one. This response is performed according to a criterion, which could be either message content (semantic characteristic) or message source (physical characteristic). Thus, the bottleneck can be seen as **not** early in processing (e.g., selecting which channel to attend), but late in processing, such as at the point of response preparation.

One early attempt at identifying if filtering was an early or late process was by Treisman and Geffen (1967). They had subjects shadow a message on one ear, and tap whenever they heard a certain word in either ear. When the key word appeared in the attended ear, subjects tapped 87% of the time, but when the key word appeared in the unattended ear, subjects tapped 8% of the time. In short this suggested by them as evidence that early selection is occurring.

A Hybrid Model for Selection In more recent times Lavie and associates (Lavie & Tsal 1994, Lavie 1995, Lavie 2000, De Fockert, Rees, Frith & Lavie 2001) have made significant progress toward unifying the theories of early and late selection and have suggested a hybrid model. In this model the level of perceptual load of the relevant processing is the determinant for which process, early selection or late selection, occurs. Early selection occurs under situations of high perceptual load (e.g., when many relevant stimuli are presented) that exhaust all available capacity in relevant perception. Late selection occurs under situations of low perceptual load (e.g., just one relevant stimulus is presented), because relevant perception leaves spare capacity that spills over to the processing of irrelevant items.

Empirical support for the role of perceptual load in determining the processing of irrelevant distractors has been provided in a series of experiments that manipulated the level of perceptual load in relevant processing and measured the effects on irrelevant distractor processing. These studies used various manipulations of perceptual load and several measures of distractor processing. For example, Lavie and Cox (1995, 1997) manipulated perceptual load by varying the number of stimuli among which targets had to be found or by varying the similarity between target and non-target letters in a

task of visual search. In both studies, distractor processing was measured via response competition effects as seen through target reaction times. The results from these studies showed that response competition effects from peripheral irrelevant distractors were reduced by high perceptual load in the relevant processing.

Lavie and Fox (2000) also examined the effects of perceptual load on negative priming from irrelevant distractors. They found that the extent to which an irrelevant distractor in the prime display produced negative priming effects on subsequent target response times was critically determined by the number of task-relevant stimuli in the prime displays.

2.2.2 Attentional Degradation

Attentional degradation describes difficulties humans experience in the maintenance of attention and an affect that sees the ability to attend degrade. This affect is most profound in vigilance task such as critical inspection processes that require accuracy of performance over an extended period.

Degradation in task performance during vigilance tasks is mainly affected by the level of sustained attention, signal quality, target predictability and background event rate. Higher levels of sustained attention place higher demands on mental resources so are affected by the individual's ability to maintain this heightened load. Clarity, intensity and duration of signal play a major large role in the degradation process, for example, the lower in intensity or the shorter in duration of the target signal, or the lower the contrast between target and background signals the greater the degradation in performance. Event predictability is another key impacting factor as the greater the uncertainty regarding target location and/or frequency the greater the amount and/or rate of degradation.

Norman and Bobrow (1975) were the first to research attentional degradation examining the effect on performance of several active processes competing for limited processing resources. This led them to prescribe caution when making conclusions regarding psychological processes and describe the effect of "Graceful Degradation". This describes the situation in which the human processes become overloaded and often results in a smooth degradation in task performance rather than a calamitous failure.

Their basic precept was that resources for any system are limited, and when several

processes compete for the same resources, there will be a deterioration of performance. This was supported by the previously mentioned observation of that when human processes become overloaded, a smooth degradation on task performance, rather than a calamitous failure, occurs. This was of course the property of the human processing feature they called "The principle of graceful degradation". From this principle they go further in suggesting a basic principle of operation, that of "The principle of continually available output". This implies that processes must continually output across a wide range of resource allocation, even when the output has not been completely analysed. They used these properties to examine the interactions of processors and affects on performance.

Their caution about making conclusions regarding psychological process stems from their suggestion that processes that share a resource do not interfere with one another until forced beyond an operational resource limit. In making this caution they were targeting Posner and Boies's (1971) suggestion that the nearly perfect time sharing between preparation and encoding implied that at least one of these operations did not require central processing capacity. Norman and Bobrow (1975) made the suggestion that if either process interfered with the other it could be concluded that they shared a resource despite no indication of a reciprocal relationship. It is also suggestive of a form of executive management of resource access through the application of a prioritization protocol.

2.2.3 Top-down & Bottom-up Processing in Cognition

When considering the effects of the visual component of interactive interfaces have on task performance, an important concept is the control of selective attention. This control sees two processes represent the bidirectional nature of perceptual information flow and processing, that of Top-down and Bottom-up processing.

Top-down processing is often described as a concept driven process where bottom-up processing is said to be a stimulus driven process. Simply put, they refer to processes that result in information flowing from either the top or the bottom of the information processing hierarchy, respectively (Lindsay & Norman 1972). Understanding directional aspects of process and informational flow, and the structures and their purposes is clearly central to understanding any human cognitive processes that might influence

human computer interaction and performance in task completion.

The top of the hierarchy is often described as containing high-level, abstract, and encompassing knowledge representations such as concepts, mental models, and schemata. Inversely, the bottom of the hierarchy is suggested to contain low-level, concrete, and specific knowledge representations such as visual features, lexicons, and propositions (Bruning, Schraw & Ronning 1995, Kintsch 1998).

The bottom-up model represents the passive processing of environmental information, in which current expectations play no role. It is generally thought to occur when an individual draws from some particular examples, instances, cases, or events to form a generalization, rule, or law to capture the commonality between the examples, instances, cases or events (Brown, Collins & Duguid 1989). This highlights the importance of correct cognitive load management to allow the most effective schema (see Section 2.1.3.2) formation to assist these types of process. The more experience there is the better-formed schema are along with weights and connections, thus the more we will let prior experience, and behaviour guide our current actions. An example of bottom-up processing is Induction.

The top-down model represents active processing, which makes use of higher-level information such as heuristics to make conclusions about a particular concept.

The difference between the two types of processing can be demonstrated by comparing the difference between current-task behaviour being driven by sensory perceptions (sensory stimulus/data) of the immediate environment (bottom-up) and current-task behaviour being driven by past experiences/knowledge (concepts), particularly experience of similar situations (top-down).

In the task of visual search, the individual is driven by visual input, if the scene is managed in such as way as to address the individual's capacities the outcome of a search may be improved markedly. In this respect, an understanding of top-down processing as task-directed behaviour is important. This is because the understanding of the effects of visual inputs on the achievement of a user task one must closely look at top-down processing as task-directed behaviour require top-down control of attention to manage the allocation of attention to task-relevant stimuli as opposed to task-irrelevant distractors (Lavie & Defockert 2005).

2.2.4 Selective Attention

When considering the affect of the visual component of interactive interfaces have on task performance an important concept is the control of selective attention. This control sees two processes represent the bidirectional nature of perceptual information flow and processing, that of Top-down and Bottom-up processing.

Top-down processing is often described as a concept driven process where as bottom-up processing is said to be a stimulus driven process. Simply put, they refer to processes that result in information flowing from either the top or the bottom of the information processing hierarchy, respectively (Lindsay & Norman 1972). Understanding directional aspects of process and informational flow, and the structures and their purposes is clearly central to understanding any human cognitive processes that might influence human computer interaction and performance in task completion.

The top of the hierarchy is often described as containing high-level, abstract, and encompassing knowledge representations such as concepts, mental models, and schemata. Inversely, the bottom of the hierarchy is described as containing low-level, concrete, and specific knowledge representations such as visual features, lexicons, and propositions.

An important aspect of interactive task performance is the ability of the user to direct their attention to the areas of the screen that contain information. As a cognitive process, this is known as selective attention, a process that enables selective response to individual objects in a cluttered visual field. For many years, there has been controversy over the role of attention in perceptual processing. Early Selectionists such as Broadbent (1958) suggest attention selects items for further perceptual analysis from a pre-categorical level of representation. Driver (1989) suggests that this implies that only attended stimuli will be fully categorized by the perceptual systems. By contrast, late selectionists like Duncan (1980) argue that objects are categorised pre-attentively which suggest stimuli are selected for action rather than for identification. More recently Chun and Jiang (1998, 1999, 2001) have conducted experiments and reviewed the field looking for empirical evidence on how implicit learning guides visual attention. They also looked at how attention modulates implicit learning, however their more recent studies (Jiang & Chun 2001) focus on the selective nature of attention and four experiments that investigate how selective attention modulates implicit learning in visual search which lead them to conclude that implicit learning is robust only when

relevant, predictive information is selectively attended.

Both attention and learning are key processes in human visual processing and it is evident that selective attention allows us to pick up relevant information and in doing so allows us to ignore huge quantities of irrelevant information. The problem here is that failure to attend to critical information reduces the efficiency of visual processing to the point that in extreme cases, inattention can cause functional blindness of the observers. This point is well illustrated by research on inattentional blindness (Mack & Rock 1998), the attentional blink (see Section 2.2.4.1), change blindness (see Section 2.2.4.2), as well as classic studies of selective attention such as that of Neisser & Becklen (1975).

At this point one might be tempted to think that these limitations of attention would make it impossible to appropriately attend too more than one target as seen in research like that of Duncan (1985). In this research he demonstrated that if *several targets* occur in an attention stream at once, or close in succession, people typically only detect one target and missed the others. However, this is in regards to multiple targets in one stream and not multiple streams involving one target. Alternate to this, it was demonstrated by Eriksen & Spencer (1969) and Ostry, Moray, & Marks (1976) that people can attend to several different streams of information at once for a *particular target*.

In computer-based interactive tasks the ability to visually attend appropriately plays a key role in the interactive process. Visual attention is constrained not only by the location and spacing of stimuli, but also by how the visual system groups these stimuli together or apart. The ability to attend to multiple targets is impacted by several factors. In line with Eriksen & Spencer (1969) and Ostry, Moray, & Marks (1976), Duncan (1984) demonstrated that the two-target cost, can be eliminated if the two targets to be judged are both attributes of the same object, even if these attributes are no closer together than those of two separate objects which do produce the two-target cost. A further refinement to this was proposed by Driver and Baylis (1989) who demonstrated that distractors that group with a target, through traits such as common motion, can produce more interference than closer distractors that do not group so strongly. Egly, Driver & Rafal (1994) also supported the concept that humans perform better for targets of common origin when they adapted the spatial cuing paradigm

to showed that participants perform best for targets presented at the other end of the same object than for targets the same distance away in a different object. Most recently research by Reddy & VanRullen (2007) demonstrated that competition, such as inter-stimulus spacing, can have a significant effect on visual search performance.

We have seen that visual information is missed if it is not *attended too and presented* appropriately, however of equal importance in visual processing is prior experience. People process visual information more efficiently when experience provides schemata to organize complex scenes (Biederman 1972). This gathering of such experience, is suggested to be, at least to some extent, an automated process as suggested in work of Chun & Jian (1998, 1999) which indicates that implicit learning allow perceivers to acquire useful information about the structure of the visual world. From this it is fairly clear that a determinant for what gets attended in a given situation is reliant on past experience.

Learning has long been recognised as affecting attention as seen in discussion as far back as the mid 1960's when it was suggested by Gibson, when talking about the "education of attention", that attention is affected by perceptual learning (Gibson 1966). Chun and colleagues (Chun & Jiang 1998, Chun & Jiang 1999, Chun & Nakayama 2000) have demonstrated that implicit learning of visual context guides attention toward targets in a visual search task. For example, Chun and Nakayama (2000) demonstrated that implicit traces of past views guide attention and eye movements to provide effective access to a scene's details, hence providing context and continuity to ongoing interactions with the perceptual world.

This is not to say that attention has only a one-way relationship with memory that seeing memory the driver of successful attention task, in fact it has been shown to be a bidirectional relationship that in which attention also influencing the extent/success of implicit learning. This was supported by Nissen & Bullemer (1987) who demonstrated, in their study of the relationship between learning and awareness preserved learning in amnesia patients, that learning is partly determined by the amount of attention allocated to the task.

2.2.4.1 Attentional Blink

The "attentional blink" is a well researched phenomenon that has seen differing conclusions such as those by Chun & Potter (1995), Raymond, Shapiro & Arnell (1992) and Shapiro, Raymond & Arnell (1994). It describes interference in the correct response to a target when attending a rapid serial visual presentation. The effect results in the perception of a target further along a series being impaired if the inter-target stimulus onset asynchrony is between about 100 and 500 ms. This impairment is known as an "attentional blink".

There have been several different theories proposed that attempt to account for the attentional blink, such as Inhibition theory, Interference theory, Delay-of-processing theory, Attentional Capacity theory and Two-Stage Processing theory.

The Inhibition Theory

Raymond, Shapiro & Arnell (1992) proposed that the attentional blink is produced by perceptual confusion between the target *T1* and subsequent target *T2*. They suggest that this confusion occurs during the target identification processes. Therefore, if confusion can be eliminated, then no AB should be observed. Raymond et al. suggest that one way of eliminating confusion is to have items that cannot be named.

The Interference Theory

Proposed by Shapiro, Raymond & Arnell (1994) the interference model suggests an alternative to the inhibition model. Interference Theory suggests that the AB occurs because an inappropriate item is selected out of series due to competition (interference) among the multiple items in the series. It is suggested that this interference increases with increasing series size and alternately decreases with decreasing series size.

The Attentional Capacity Theory

Duncan, Ward & Shapiro (1994) propose that visual attention is not a high-speed switching mechanism, but a sustained state during which relevant objects become available to influence behaviour which is consistent with research on monkeys by (Chelazzi, Miller, Duncan & Desimone 1993). In discussion of this model they suggest that *T1* occupies attentional capacity to the detriment of a trailing *T2* target. As such this theory suggests that the duration for which *T1* continues to occupy attentional capacity is related directly to the *T2* processing difficulty.

The Two-Stage Processing Theory

The two-stage model extends Broadbent and Broadbent's (1987) work and relates back to concepts such as Neisser's (1967) proposal that preattentive processes guide the operation of a focal attention stage. The present two-stage model proposes that the AB deficit arises from a limited-capacity stage of processing and consolidation of the target after the target has been initially detected in the first stage. Chun & Potter (1995) propose that the rapid processing of a series of items requires two sequential stages: an initial rapid-detection stage (Stage 1) in which potential targets are detected, and a second capacity-limited stage in which items are processed serially for subsequent report. Access to Stage 2 is gained by items that have been identified as potential targets in Stage 1. And, until Stage 2 finishes processing *T1*, *T2* cannot gain access to Stage 2. If *T2* arrives in Stage 1 before Stage 2 is free, its access to Stage-2 processing is delayed. The attentional blink deficit is brought about by the decay of *T2* in Stage 1 during this delay. This theory suggests that the amount of attentional blink will depend on the discriminability of *T1*. If Stage-2 processing of *T1* is not slowed down by discriminability problems, processing of *T2* is not delayed, and the attentional blink deficit is reduced or eliminated.

2.2.4.2 Change Blindness

Change blindness is a well-recognised phenomenon that sees people viewing a visual scene failing to detect substantial changes in the scene. This often occurs typically when the change in the scene coincides with some visual, disruption such as a saccade or short obscuration of the scene. The term "Change Blindness" seems to have been coined by Rensink et al (1997).

The first key research into change blindness was conducted by George McConkie and his colleagues in the late 1970s this research saw key extensions made by John Grimes (1996) how demonstrated that people miss large changes to scenes when the changes are introduced during an eye movement. For example, many people failed to notice when two people in a scene exchanged heads.

Looking at other forms of visual disruption besides eye movements that could also induce relatively poor change detection Pashler (1988) demonstrated that "subjects' performance in detecting single changes in character displays is remarkably poor when

a 67-msec offset separates the first and second display. There is a reliable but modest improvement in performancewere rather lousy at detecting changes in arrays of familiar objects when the offset was even 100 msec long" (p.371). An interesting fact about this work is that in it Pashler noted that people could only "hang on to" 4-5 objects, which to me seems to reflect the physical or processing limits all areas of cognitive study seem to encounter at some level.

In studying the change blindness phenomenon Rensink, O'Regan & Clark (1997), using the "flicker" technique (two images of scenes alternate repeatedly with a brief blank screen (80 msec.) after each image giving the display a flickering appearance) demonstrated that surprisingly large changes to a scene could be made without the observer reliably noticing them. Other studies, such as that by Levin and Simons (1997), extended the situations for which change detection is also poor such as when the change is introduced during a cut or pan in a motion picture, despite the change of the central actor in a scene. The potential problem posed by the strength of the affect of change/attentional blindness is aptly demonstrated by Simons & Levin (1998) in the description of the situation were many people failing to notice the surreptitious swapping of an actor they are talking too.

Change blindness is a strong effect that has been observed as a result of a wide variety of visual disruptions (e.g. rapid scene changes, blinks and transient noise flashed on a display). This is of real concern to the designer of computer interfaces and presentation techniques as the effect needs to be understood and accommodated to reduce task error rates.

As discussed earlier, the bottom-up model represents the passive processing of environmental information, in which current expectations and past knowledge play no role. It is generally thought to occur when an individual draws from some particular examples, instances, cases, or events to a generalization, rule, or law to capture the commonality between the examples, instances, cases or events (Brown et al. 1989). This highlights the importance of correct cognitive load management to allow the most effective schema (see Section 2.1.3.2) formation to assist these types of process. The more experience there is the better-formed schema are along with weights and connections, thus the more we will let prior experience, and behaviour guide our current actions. An example of bottom-up processing is Induction.

The top-down model represents active processing, which makes use of higher-level information such as heuristics to make conclusions about a particular concept.

The difference between the two types of processing can be demonstrated by comparing the difference between current-task behaviour being driven by sensory perceptions (sensory stimulus/data) of the immediate environment (bottom-up) and current-task behaviour being driven by past experiences/knowledge (concepts), particularly experience of similar situations (top-down).

In the task of search, the individual is driven by visual input that if managed to conform to an individual's capacities can improve. In this respect, an understanding of top-down processing as task-directed behaviour is important. This is because the understanding of the of effects visual inputs on the achievement of a user task one must closely look at top-down processing as task-directed behaviour require top-down control of attention to manage the allocation of attention to task-relevant stimuli as opposed to task-irrelevant distractors (Lavie & Defockert 2005).

2.2.5 Cueing Attention

As we saw in Section 2.2.4 the context of an object can affect the efficiency with which it is attended and schemas are developed. This can be of advantage to the interface designer if they can manage to present a visual scene with visual artifacts that relate sufficiently such that each facilitates the comprehension of one or several of the others part, in other words the cuing of attention to one part of an object facilitates the discrimination in another part (Duncan 1984, Egly, Driver & Rafal 1994). This is simple to talk about but how is it achieved? Two general techniques are to help allow for this that of Elicitation filtering and Selection Filtering.

Elicitation Filtering

In the context of text search, this technique see the elicitation of information from the user, subsequent to the initial search terms being entered, for clarification of the relevant topic of each word. In this way more specific object information can be presented that maximizes the relative task information. Two examples of this approach can be seen in one of Alta Vista's old incarnations and more recently in Yahoos Y!Q tool. Alta Vista allowed the user to select, via radio buttons, the different contex-

tual meaning for each query term while Yahoos Y!Q tool uses both query term and information highlighted on a current Webpage to ascertain search "context".

Selection Filtering

Another way to manage this situation is via an interactive graphical presentation of a return sets that allow the user to either retain appropriate documents or discard inappropriate documents. To do this on an individual item basis for your average return set size of anywhere up too and more than one million documents would be impractical. However, if a return set is organised categorically the user can be given the opportunity to either retain or discard clusters of documents. We refer to the process of retaining or discarding of documents as a Selection-Filtering process.

The Selection-Filtering process is an interactive process that can be used to realise more precise and condensed return sets. It can be implemented by taking a condensation by elimination approach or a condensation by retention approach. Condensation by elimination sees the removal of inappropriate documents/clusters while condensation by retention sees the keeping/retention of appropriate documents/clusters. Although the result of either approach is the same the processes through which they are achieved is clearly different.

When treated recursively the Selection-Filtering approach allows the user to apply finer grades of subjectivity in the production a limited set of clusters of highly topic specific documents. For this to be a rapid process, clusters need to have relatively large populations in each cycle to allow for a rapid reduction to a concise set.

A example of a condensation by retention approach can be seen in Schneiderman's (1992) Tree Map interface. Using a hierarchical clustering the Tree Map interface allows users to drill down through levels and inspect clusters by selecting a clusters' representative or if at the bottom level singletons.

When the spatial layout of the attended set of distractor's was consistently paired with the target location, target search was facilitated, but only after a few (e.g. three) repetitions. This indicates that observers were able to extract the invariant spatial layout embedded among noise produced by the random positioning of the ignored set. Thus, contextual cuing is quite robust to perturbation of the global spatial configuration (Chun & Jiang 1998). These results suggest that contextual learning can be restricted to a subset of attended events within a visual array.

Attention to Objects and Cuing

Cognitive affects of objects in a scene/interface are often studied using an approach that varies the cue participants are presented. Using this approach attention to objects is often studied by presenting subjects with displays of multiple objects and giving them cues that indicate the target's location or some other salient property (Logan 2003). In some procedures, such as those of Eriksen & Eriksen (1974), and Theeuwes (1994), each object in the display is a potential target and the cue indicates which object to judge or report. Consequently, subjects cannot respond to the target without first responding to the cue. In other procedures such as used by (Posner 1980, Posner, Inhoff, Friedrich & Cohen 1987), the target differs from the distractor's in some way and the cue merely indicates its position. This research suggests that although subjects can respond to the target without first responding to the cue, the cue still influences performance. Simply put, *valid cues* facilitate performance, speeding reaction time and increasing accuracy and *invalid cues* that indicate a location that does not contain the target impair performance, slowing reaction time and decreasing accuracy.

There does however seem to be at least one caveat to the use of cuing in task achievement which is seen in further refinement proposed by Olson & Chun (2002) who suggested that colour differences do not effect contextual cuing at all and "that spatial features play a more important role than surface features in spatial contextual cuing" (p.273). This suggests only certain screen artifacts are applicable for the use of cuing in the positive achievement of tasks.

In the conclusion of Olson & Chun (2002) they point to a very important point that supports the use of clusters in return sets. In short their work indicated that implicit learning of spatial context is robust across noise and biased toward *spatially grouped information* which is the key to the usefulness of clustering. The visual clustering of objects of common traits improves task achievement through better target object processing.

2.2.6 Attention as a Resource

Throughout the literature, attention is shown to be a key limiting factor in the successful completion of tasks involving information derived from visual objects of a display. In this light attention can be seen as a resource to manage a concept that is recognised

by research such as that by Johnston & Dark (1986) and Kahneman (1973).

In considering attention as a resource research such as that by Baddeley (1996) and Desimone (1995) suggests working memory is the key factor in selective attention processes. Following from this research by De Fockert, Rees, Frith & Lavie (2001) demonstrated that working memory affects distractor management by influencing the priority processing of relevant and irrelevant task stimuli. In turn, this affected control of visual selective attention.

Capacity Limits and Distractor Management

Research by Lavie & Cox (1997) determined that capacity limits dictate the efficiency of selective attention processes while Jiang & Chun (2001) found that the more difficult the search task is, the more likely ignored items would be filtered out early in the process and thus would produce no benefit of repetition. Lavie & Tsal proposed structural and capacity approaches to attention and suggest that "perceptual load is a major factor in determining the locus of selection" (p.183). This is all consistent with perceptual load theory, as suggested by Lavie (1995), and Lavie & Tsal (1994), that predicts enhanced attentional selectivity with increased attentional load. Perceptual load theory embraces both the selection and the resource aspects of attention and postulates a close link between the two.

The efficiency of irrelevant distractor rejection is suggested by Lavie & Tsal (1995, 1994) to depends on the perceptual load involved in the relevant processing. Perception at this level is suggested to be an automatic, involuntary and capacity dependent process. In this case *automatic* refers to perception as being characterized not in the sense that it does not require attention, but in the sense that it is not subject to complete voluntary control. Thus, Lavie & Tsal's model combines aspects of early selection approaches (limited capacity) and late selection approaches (automatic response) in which processing proceeds from relevant to irrelevant items until capacity runs out. Simply put, under lower levels of load and during relevant information processing, spare capacity spills over to process irrelevant information, and hence may lead to distraction. As such, irrelevant processing can be prevented with higher loads in relevant processing that exhausts this excess capacity.

This dual aspect model found support in research by Spink, Zhang, Fox, Gao & Tan (2004) the results from which confirmed the extent to which higher level cognitive

resources, specifically the central executive component of working memory, are absorbed by a cognitive task and that there is a real and measurable impact upon automatic processing that occurs in response to distracting items.

Capacity vs. Task Difficulty

In the search task, Huang & Pashlers (2005) suggest that attentional capacity is not affected by the efficiency of the search technique rather it is determined primarily by the nature of the task. They also say point to the pioneering research of Treisman and Gelade (1980) as revealing robust and important functional distinctions between different kinds of visual search tasks. Added to this, the results also suggest that a different hierarchies of tasks may characterize the relative visual-attention demands of different kinds of visual search tasks which is also supported by others such as Eckstein et al. (2000) and Geisler & Chou (1995). The Eckstein research presented a model that accurately predicted human experimental data on visual search accuracy in conjunctions and disjunctions of contrast and orientation. The Geisler & Chou research proposed a signal-detection model that demonstrated how the then current psychophysical models of visual discrimination might be generalized to obtain a theory that can predict search performance for a wide range of stimulus conditions.

Inattentional & Attentional Blindness

Further support for the concept of excess capacity being consumed by irrelevant object processing can be found in research into inattentional & attentional blindness by Mack & Rock (1998) and Rees, Frith & Lavie (1999). Specifically, Rees et. al. demonstrated that conditions that do not fully engage attention result in incidental processing of linguistic properties even during non-lexical tasks. Their results suggested that, under the appropriate conditions of true inattention, words can be directly fixated but not read.

The Pop-out Effect

In research by Huang & Pashler (2005) it was noted that, in speeded visual search tasks for which an observer has an opportunity to view a display as long as they choose, and for which the scene is arranged so that the target differs from uniform distractors in only one feature dimension, search time usually does not increase with the number of distractors. A real world parallel is suggested to be that of finding a person wearing red in a large crowd of people all wearing green, they seem to “pop-out”. Treisman

& Gelade (1980) suggest that the "pop-out" effect basically reflects spatially parallel processing.

The power of the pop-out effect has been clearly demonstrated by such research as that by Carter, and Nagy & Sanchez (1982, 1990). It was demonstrated that the target/distractor difference is very subtle, however, even this type of singleton search problem used a substantial display-set-size effect is observed. Carter (1982) demonstrated that search time increased as the number of display items of the target's colour increased. It was also demonstrated that search time increased when the number of display items of different colours from the target increased but only if the colour of these items was sufficiently similar to that of the target. That is to say that, if the colour of these background items was dissimilar to that of the target, then the background items had no effect on search time, however if the more similar the background item colour is to the target the more difficult the task of differentiation is. Nagy & Sanchez (1990) used two tests, one to measure the effect of display density for both small and large colour differences and distracter chromaticity. It was primarily demonstrated that with small colour differences "response time increased with display density, indicating a serial search, but with large colour differences response time was constant, indicating a parallel search" (p.1209).

In short we see that "difficulty", described by target-distractor similarity, significantly affects the efficiency of a visual search. However, Carter & Carter (1981) capture the usefulness of the pop-out effect in their research that suggested that indices of conspicuousness, relative fixation rate and search time, were shown to be related to the colour difference between the target and background objects. In this research they conclude that "colour difference be used as a tool for design and evaluation of visual displays, for construction of colour codes to optimize search time, and as a generalization of chromatic contrast in psychophysical research" (p.723).

Spotlighting and Attention

The tendency to pick out a particular region of the visual scene for more detailed processing is often referred to as "spotlighting". The metaphor derives from the fact that we have a limited foveal region and thus need to move our eye to focus visual attention for higher acuity much like that of a spotlight. Driver (2001) points out that research like that of Grindley & Townsend (1968) and Posner (1980) supports

that observation that foveating on a target is not the primary mechanism of attention. There are however, many further influential examples of the spotlight metaphor in the literature and how its is used to focus attention (e.g. Eriksen & Eriksen (1974) and Posner (1980)).

Driver & Baylis (Driver & Baylis 1989) suggest that there is an emergent consensus that space is the medium for visual attention, which is held to operate analogously to a "spotlight". The crux of this metaphor is the idea that attention selects contiguous regions of the visual field for further processing, whether this be selection for identification or for the control of action. Two variants of the spotlight model are Eriksen & St-James's (1986) *zoom-lens model*, and Downing & Pinker's (1985) *gradient model*.

Eriksen & St. James's (1986) note two main effects, one was that under certain conditions the attentional resources are attributed evenly across the display, with parallel processing of the display items, while under other conditions serial scanning of a display seems to occur. The second is that attention can be directed to a specific location in a display using pre-cues as close together as 50msec before display onset which results in improved target detection. Under the zoom-lens model, the situation where attention is directed to two regions of space not adjacent (split attention) is not be possible, however, research by Awh & Pashler (2000) carried out a study which found that split attention is in fact possible but that there may be problems with array orientation of targets, which is not consistent with the zoom-lens model. Downing & Pinker's, (Downing & Pinker 1985) gradient model is a combination of both specific-location and general region models. It suggest that attention centres around specific locations while including a distribution of attentional resources surrounding this fixation. Gradient models generally account for increased response times through increased distances. It should be noted that these positions share the fundamental assumptions that space has a unique role for visual attention, which can only be assigned to contiguous regions of the visual field.

It has been suggested that distractor interference tends to diminish with increasing distance from the target. For example, in their study of the effect of noise in search task Eriksen & Eriksen (1974) conclude that "discrimination is more difficult and time consuming at closer spacing and inhibition is more difficult when noise letters indicate the opposite response to the target" (p.143). This supports the spotlight models in which

visual attention can only be assigned to contiguous regions of the visual field. However, as Driver and Baylis (1989) point out, it also suggests that attention is assigned to perceptual groups. In their research, they demonstrated that by grouping targets and distractors through common motion a larger effect is noted than with proximity. From this, they make the conclusion that attention is directed to perceptual groups whose components are spatially dispersed and that in dynamic environments the spotlight metaphor is probably inappropriate.

2.2.7 Stimuli Intensity

The visual input from a screen is logically going to influence the performance of an individual with regards to completing a specific task. As we have seen, the magnitude or intensity of a target is one area of interest that is being addressed to ensure user cognitive loads are at an appropriate level to complete a task effectively and efficiently. Psychophysicists have been interested in this type of question for many years and in fact one might say their raison d'tre is to generally study the relationship between the strength of stimulus and perception. The key question they seek to answer is the scaling question, which is the relationship between the magnitude of the physical stimulus and the perceived magnitude. Typically, for most types of stimuli this relationship is not one to one.

The Weber-Fechner law (1834) is suggested to describe the relationship between the physical magnitudes of stimuli and the perceived intensity of the stimuli. Ernest Heinrich Weber was one of the first people to approach the study of the human response to a physical stimulus in a quantitative fashion. The Weber-Fechner law basically states that the magnitude of a subjective sensation increases proportionally to the logarithm of the stimulus intensity.

Stevens' (1957) power law also defines the relationship between the magnitude of a physical stimulus and its perceived intensity or strength. It is widely considered to supersede the Weber-Fechner law on the basis that it describes a wider range of sensations. The power law states that $S = kI^a$ where S is the sensation magnitude, k is an arbitrary constant determined by the scaling unit, I is the stimulus intensity and a is the power exponent dependent on modality.

2.2.8 Task Complexity

Just as more items in a display usually require a greater amount of search time, when a task is more complex in general, more attention is required. However, once again there are exceptions to this general case. The exceptions are automatic behaviours.

As we have seen the more a behaviour is repeated the more performance improves, this is because fewer mistakes are made and less time is required to perform the behaviour. This improvement is reflected in the learning curve, which is a plot of performance in terms of mistakes or time required by repetitions of the behaviour. Much research has focused on plotting behaviour types however the difference between these plots is simply the axes values. Along the same line as finding a model of best fit in statistics the uniformity of performance can also be plotted using logarithmic scales on the axes to produce a linear plot. This is generally described as the "power law" of learning which states that performance of every behaviour will improve in such a way that a straight line will be produced when plotting performance over time using logarithm scales on the axes. The more a behaviour is practiced the more likely it is to become automatic which will result in behaviours requiring less attention.

It can be said that practiced behaviour is somewhat beneficial in the case of interactive tasks, there is however a level of cognitive complexity involved that can lead to current tasks and processes interacting to result in an increased cognitive load or confounded outcome. For example, the Stroop task is an example of an automatic process (reading) interfering with the current process (colour naming). Seeing the words primes people to respond with the colour that is written out. Priming occurs automatically, and refers to the activation of a response or a memory (the more practiced the stronger it is likely to be), which makes that response more likely to occur in the immediate future.

2.3 Cognitive Styles

Cognitive style, or "style of thinking", describes the general approaches individuals use in thinking of, perceive and remember information, and/or their use of this type of information to solve problems. Although primarily a concept used in the areas of education and management, as a tool that characterises users it might also be of use in

the designing of interactive interfaces and techniques. Understanding cognitive styles is important because user styles will affect the way individuals process and acquire information, make decisions, solve problems and respond to the different presentation of information. Research by Boles and Pillay (1999) demonstrated that certain tasks are more suited to certain cognitive styles and that performance improvements might be realised by matching cognitive styles to type of content and presentation. There is however a caveat to the use of cognitive styles in this manner in that there are a number of models to choose from, each of which will need to be considered for its applicability to the problem being addressed and the potential cognitive style/s.

There have been several key models representing cognitive styles proposed, however it is generally recognised that for whatever style/s an individual has they are likely to be fixed characteristics of that individual. This fixed nature is allowed for by the "Cognitive Strategies" an individuals uses as they are techniques/approaches used to cope with information that does not harmonise with the individuals cognitive style.

The key models of cognitive styles fall into the two main categories of *multi-dimensional* models and *uni-dimensional* models.

The Myers-Briggs Type Indicator (MBTI)

The Myers-Briggs Type Indicator (MBTI) (see Myers (1987)) is a well known multi-dimensional cognitive style description approach widely used throughout the world. Based on the typological theories of Carl Gustav Jung it is a questionnaire that attempts to characterise individuals according to the four basic preferences of:

1. extraversion versus introversion
2. sensing versus intuitive
3. thinking versus feeling
4. judgment versus perception

This is done by characterising the individual on all four continuums (see Table 2.2).

Psychological Continuum		
Extraversion	$E \longleftrightarrow I$	Introversion
Sensing	$S \longleftrightarrow N$	iNtuition

Thinking	$T \longleftrightarrow F$	Feeling
Judging	$J \longleftrightarrow P$	Perceiving

Table 2.2: Myer-Brings psychological continuums

The Verbal-Imagery and Wholist-Analytic continuum

The two-dimensional approach developed by Riding and Cheema (1991) measures an individual's position along the two orthogonal dimensions of Wholist-Analytic and Verbal-Imagery. They suggest that individuals along the Wholist-Analytic continuum tend to process information in wholes or parts and those along the Verbal-Imagery dimension tend to represent information verbally or in mental images. It is also suggested that these two styles are not exclusive and non-interactive, or in other words most people present as a mixture of the two dimensions and their position on one dimension does not effect their position on the other. Also, members of each of the four extremes can use the style of the opposite extreme however this may, as described by Sweller (1989), imposing extraneous cognitive load and result in reduced efficiency in the learning process.

Wholists organize information into chunks to form an overall perspective of the given information and Analytics view information in conceptual groupings focusing on one grouping at a time. Verbalisers process information as words or verbal associations while Imagers relate information better with mental images or pictures.

Pillay and Wilss (1996) used Riding and Cheema's approach in their study, involving second year nursing students at Queensland University of Technology, to test for variation across eight different groups using four different lesson sets that either matched or mismatched their style. The study results indicated an interaction between online instruction and individual's preferred cognitive style. Their conclusions were indicative of a need for further research into instruction that can be tailored to individual cognitive styles to promote learning through reduced extraneous cognitive load.

The Field Dependence-Independence Model

The field dependence-independence model, designed by Witkin (1977, 1977, 1981), identifies an individual's perceptive behaviour while distinguishing object figures from the content field in which they are set. The model stemmed from Witkins (1971) use of the

Group Embedded Figures Test (GEFT) and Embedded Figure Test (EFT) he designed to identify the preferred learning style of students. Both tests where instruments designed to distinguish field-independent from field-dependent cognitive types, a rating that is claimed to be value-neutral, using confusion fields with distracting or confusing backgrounds.

Field-dependent people are likely to perceive situations globally and have more difficulty in solving problems, tend to be extrinsically motivated social learners, and achieve better in organized and structured situations. Generally, they will tend to have better interpersonal skills and when solving problems work better in teams (Witkin, Moore, Goodenough & Cox 1977). They also tend to find it more difficult to see the parts in a complex whole. Field-independent people tend to view concepts analytically and therefore finding it easier to solve problems. They prefer their own structure and organization, are intrinsically motivated while being less skilled at building interpersonal relationships, tend to be more autonomous when it comes to the development skills; that is, those skills required during technical tasks with which the individual is not necessarily familiar (Witkin, Moore, Goodenough & Cox 1977).

Convergent and Divergent Thinkers

The concepts of Convergent and Divergent thinkers was described by Guilford (1959) when developing his "structure of intellect" model. In short, divergent thinking is the ability to find as many possible answers to a particular problem and convergent thinking is the ability to find the best single answer to a problem. In his research Hudson (1966) found that conventional measures of intelligence did not always do justice to a subject's abilities. Hudson therefore contrived the concepts of a converger-diverger continuum to measure the processing of information rather than the acquisition of information by an individual. In general those that are more convergent in thinking style tend to think rationally and logically, bringing material from a variety of sources to solve a problem. This kind of thinking is particularly appropriate in science, maths and technology. On the other hand the more divergent thinkers are more creative, rapidly realising a large number of ideas or solutions working around a problem. This kind of thinker is more suited to creative pursuits and those that require thinking outside the box.

Left-brained vs. Right-brained

The "hemispherical lateralisation concept", commonly known as the "left-brain right-

brain model" of cognitive style, was described by Doyle, Ornstein and Galin (1975) in their research in which they noted differences in the power of the alpha band in signals recorded from left and right hemispheres, depending on the tasks. In this research, they expected and demonstrated, to a degree, the language and arithmetic tasks would engage primarily the left hemisphere, and spatial and musical tasks were expected to engage primarily the right hemisphere. From this and other similar research (e.g., (Galin & Ornstein 1972, Schwartz, Davidson & Maer 1975, Davidson, Schwartz, Pugash & Bromfield 1976)) this style evolved to represent an individuals cognitive tendencies in different task on continuum between extreme left-brain to extreme right-brain types. This is dependent on which associated behaviour dominates in the individual, and by how much.

The Kirton Adaption-Innovation (KAI) Theory and Inventory

The Kirton Adaption-Innovation theory was developed by Kirton (1976, 2003) to represent an individuals preferred style of creativity and problem solving. It represent this style on a continuum between the two categories of *Adaptive* and *Innovative*. From this theory Kirton developed the KAI inventory to measure the methodology an individual uses to bring about change by indicating whether they have a preference as an *adapter* or *innovator*. KAI instrument is a form containing 32 questions each of which the individual rates on a scale .

Kirton's definition of an *innovator* is a person who is "less tolerant of structure (guidelines, rules) and less respectful of consensus". An innovator will break rules and paradigms to produce a new way of doing things. On the other hand, an adaptor will have more respect for rules and structure. They prefer solving problems in a defined environment, working to do things "better" as opposed to breaking the paradigms. While the adapter thrives on structure and has a penchant for order, predictability and repeatability, the innovator seeks newness and experimentation, fails to see structure or credits structural consistency as contributing to the problem (Kirton 1976, Kirton 2003).

2.4 Concluding Observations

This section highlights some important points relative to interactive search drawn from the body of this chapter.

Sub-tasks in interactive search are generally short and the visual information critical to decision-making is often short lived, especially that of graphics based interfaces. This poses a problem if information is required for subsequent activities because of the volatile nature of *sensory memory* means this information will be lost (see Section 2.1.2). This suggests that if information is pertinent to the completion of the task, especially the immediate sub-task, its screen artifact needs to remain while it is contextually relevant or until it is no longer required to complete any relative tasks.

The volatile nature of visual information in the task of search also means short-term memory (see Section 2.1.2) plays a critical role in task realization. It allows one to recall something from several seconds to as long as a minute without rehearsal and as such if tasks can be guided to completion within this period the information in memory is more likely to leverage the quality of decision making and task completion. For more lengthy tasks, if rehearsal of critical information is allowed through presentation of queues either where it is rapidly accessible or re-presented on a regular basis it will be more likely available to the decision making process and thus increasing the likelihood of task success and quality.

Rehearsal can be said to increase the weights on connection in the brain enhancing the chance of recall much like that describe by the Parallel Distributed Processing (see Section 2.1.3.1). So, to reduce any load realized by repetitive tasks and thus free capacity for other non-common more analysis/decision making tasks, repetitive tasks involving screen artifact interaction should see the artifacts keep constant throughout the process. That is they should look the same, do the same thing and appear in the same position. This will allow for better quality rehearsal which is important because the more a behaviour is practiced the more likely it is to become automatic which will result in behaviors requiring less attention (see Section 2.2.8) relinquishing attentional resources for other more critical tasks.

The recognition of general cognitive styles in the design of interactive interfaces is important because user styles will affect the way individuals process and acquire information, make decisions, solve problems and respond to the different presentation of information (see Section 2.3). If an interface automatically adjusts to the user's cognitive style and/or allows the user change their tactics to match the interactive style of the interface. This in turn can result in better performance by better addressing of

cognitive styles.

If critical attributes of a search task can be identified, such as those clusters that most likely address the user's query, pre-attentive processing can be used to expedite the process. The usefulness of pre-attention in this situation can be seen in the fact that pre-attentive processing can be used to rapidly draw the focus of attention to a target with a unique visual feature (i.e., little or no searching is required in the pre-attentive case) (Healey 2004). This can be achieved by making the pertinent attributes more intense, move or any one of a number of attention grabbing techniques.

Finally, interactive displays need to be simple with the singular target of getting the searcher from query to successful result as efficiently as possible. This suggestion is brought about by the fact that working memory capacity is critical under conditions in which interference leads to retrieval of response tendencies that conflict with the current task (Engle, Conway, Tuholski & Shisler 1995, Engle, Kane, Tuholski & Press. 1999, Engle 2001, Engle 2002). So, by reducing interference better application of working memory will be realized which should lead to better task outcomes.

Chapter 3

Cognitive Limitations and Load

Given the problem of Data-avalanche in document search, the "ultimate search system" might allow the rapid reduction of an unmanageably large search return sets by getting the user to evaluate and discard large inappropriate categories (clusters) of documents. Given textual language efficiently conveys fine-grained topical details about textual documents, it is appropriate to describe the topical content of clusters and individual documents using text. However, if the user's abilities are not appropriately recognised when generating the cluster descriptors, the user will realise a less than optimal task outcome. For example, if one word is used to visually describe a document, the user may not have enough information to correctly classify or evaluate its relevance to their information need. At the other extreme if the entire document is used, the user will spend far too much time reading individual documents to identify classifying features and will not have the time to process the number of documents in a Data-avalanche. Somewhere along this continuum, is an optimal descriptor length, but where?

Being based on a *physical* biological system ensures that user cognitive processes will logically have limitations that will affect the amount of information a user can process at any one time. The following sections discuss specific areas of research relative to limitations of cognition that can be used as general guidance for determining how many words should be used to describe clusters and individual documents.

3.1 Cognitive limits

3.1.1 Magic numbers and memory limits?

In designing human computer interfaces one might wish to identify cognitive limits to optimize task realisation by users. Working memory is often implicated in sub-optimal task realisation, for example in proposing a framework to identify and investigate key factors that determine a Web browser's ability to assist users in performing various information retrieval tasks Head et al. (2000) identified four human limitations on short-term memory that can lead to navigation problems, these were:

1. Arriving at a particular point, and forgetting what was to be done
2. Neglecting to return from a digression
3. Neglecting to pursue a planned digression
4. Not remembering what has been visited or altered

It is logical that before trying to design for optimal task realisation one should have a feel for the research that attempts to characterize cognitive limits in interactive tasks. The following discussion outlines some well known limits and the field in general.

3.1.2 Miller's Magic Number

In the 1950's Miller (1956) was being "persecuted by an integer" that had "assaulted" him from the pages of many publications. This prompted him to write his famous 1956 paper on the "Magic Number Seven, Plus or Minus Two" (7 ± 2) in which he compiled evidence that suggestes people can process about seven *chunks* in short-term memory tasks. This discussion was in terms of the information theory concepts *chunking* & *subitizing*, and suggests that there is "some pattern governing" the occurrence of 7 ± 2. He did **not** however, suggest that 7 ± 2 was a hard and fast limit that applied to all cognitive situations or that all the limits resulted from a single mechanism. He proposes that there is an immediate memory device that has a capacity of about 7 ± 2 chunks of information, and that this is dependent on the nature of the information. This proposition draws on Hayes' (1952) findings that people, on average, are most likely to

remember a series of five to nine mono-syllabic words, as well as a series of five to nine letters, or five to nine decimal digits.

Since a single letter contains a different amount of information than a monosyllabic word, Miller established the idea of some relative measurement for memory capacity and provided evidence that different people chunk information differently. Additionally, he points out that generalizations about capacity must take into account how individuals organize the perceived information into chunks.

3.1.3 Cowan's Magic Number

Cowan (2001) proposes that Miller's magic number 7 ± 2 is meant as a rough estimate rather than a strict capacity limit. The article points to other research subsequent to Miller's that is suggestive of competing views and a more precise capacity limit, and that this was only three to five chunks. The competing views on capacity limits are as follows:

- There do exist capacity limits but they are in line with Millers 7 ± 2, e.g. (Lisman & Idiart 1995).
- Short-term memory is limited by the amount of time that has elapsed rather than by the number of items that can be held simultaneously, e.g. (Baddeley 1986).
- There is no special short-term memory faculty at all; all memory results obey the same rules of mutual interference, distinctiveness, and so on e.g. (Crowder 1993).
- There may be no capacity limits per se but only constraints such as scheduling conflicts in performance and strategies for dealing with them, e.g. (Meyer & Kieras 1997).
- There are multiple separate capacity limits for different types of material, e.g. (Wickens 1984).
- There are separate capacity limits for storage versus processing, e.g. (Daneman & Carpenter 1980, Halford, Wilson & Phillips 1998).
- Capacity limits exist that are task-specific, with no way to extract a general estimate. Cowan (2001)

In clarifying capacity limits Cowan (2001) suggests that any such limits are only useful for the analyses of information processing if the boundary conditions for observing them can be appropriately described. He also proposes four basic conditions in which chunks can be identified and capacity limits can accordingly be observed. These four conditions are:

1. when information overload limits chunks to individual stimulus items
2. when other steps are taken specifically to block the recording of stimulus items into larger chunks
3. in performance discontinuities caused by the capacity limit
4. in various indirect effects of the capacity limit

Under these conditions, rehearsal and long-term memory cannot be used to combine stimulus items into chunks of an unknown size; nor can storage mechanisms that are not capacity-limited, such as sensory memory, allow the capacity-limited storage mechanism to be refilled during recall. Furthermore, a single, central capacity limit averaging about four chunks is implicated along with other, non-capacity-limited sources. The pure short term capacity limit expressed in chunks is distinguished from compound short term capacity limits obtained when the number of separately held chunks is unclear.

Chunking Chunking, first used by Miller (1956), describes the capacity of short term memory. Miller proposed that "the process of memorization may be simply the formation of chunks, of groups of items that go together, until there are few enough chunks so that we can recall all the items" (p.95). It was stressed that a large number of seemingly disparate findings could be reconciled if we computed memory limitations not in terms of some physical unit, such as letters, but rather in terms of a psychological unit, chunks. The reformation of items into fewer items is called *recoding*. This process takes input information of multiple chunks comprised of a small number of bits of information and sees their condensation into a form with fewer chunks that might include more bits of information per chunk as well as references to long-term memory.

Medin et al. (2004) use an alternate definition and description for chunking. They describe a chunk as any meaningful group of information. Rather than storing each piece of information in the chunk in short-term memory, you can store the idea that

the chunk occurred. When you need to retrieve the information, you can remember that the chunk occurred (because this idea is in short-term memory) and then you can bring the constituents of the chunk from memory that are more permanent.

Miller (1956) and Medin et al. (2004) relate chunking to more permanent stores of memory (in the form of experience) as does Cowan (2001) who suggests that a chunk should be "defined with respect to associations between concepts in long-term memory" then relates the concepts as "a collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use". Medin et al. (2004) refer to this relationship when they suggest that "it is important to realize that chunking is a function of our **prior knowledge**. What is meaningful i.e. 'chunkable' depends on what we know, as well as what we are currently experiencing".

Sweller's (Sweller, Van Merrienboer & Paas 1998) Cognitive Load Theory combines Miller's work with Schemata Theory. Swellers work bilds on Millers work that suggests 'short term memory is limited in the number of elements it can contain simultaneously' building a theory that treats schemata, or combinations of elements, as the cognitive structures that make up an individual's knowledge base. Simply put, schemata become chunks in memory and reuse in building memory.

In general, the use of chunks can be observed in our ability to remembering long sequences of binary numbers through the process of recoding. Recoding the sequence into decimal form ensures a more compressed representation, thus requiring less capacity/resource to process. For example, the sequence 0010 1000 1001 1100 1101 1010 could easily be remembered as 2 8 9 C D A. This also demonstrates the relationships of chunks to experience/long-term memory as this example solution only works for someone who knows how to convert binary to hexadecimal numbers (i.e., the chunks are "meaningful").

3.1.4 Subitizing

It has been suggested that we use two distinct processes when enumerating: subitizing and counting (e.g., (Kaufman, Lord, Reese & Volkmann 1949, Mandler & Shebo 1982, Trick & Pylyshyn 1994)). Subitizing as opposed to counting is characterized as a parallel process whereby the elements of a visual display are automatically translated into a numerical value (representation). Proposed by Kaufman et al. (1949) "subitizing"

refers to "the rapid, confident and accurate report of the numerosity of arrays of elements presented for short durations". The number judgments of test participants for groupings of items displayed are referred to as either *counting* or *estimating*. Whether one or the other occurs is dependent on the number of elements displayed and the exposure time (i.e., estimation occurs if insufficient time is available for observers to accurately count all the items present). However, below a certain number, and within a fairly short period of time, the observer will be correct every time, an occurrence otherwise known as subitizing.

Research shows that counting requires ocular movements to locate and mark, individual objects and groups of objects in a visual field (Atkinson, Campbell & Francis 1976, Atkinson, Francis & Campbell 1976, Simon & Vaishnavi 1996). It has also been shown that the arrangement of objects in a field effects counting but not subitizing (Atkinson, Campbell & Francis 1976, Atkinson, Francis & Campbell 1976, Mandler & Shebo 1982). Research by Ross (2003) proposes that there exists neurons specifically tuned for numbers and that it is these neurons that allow the subitizing effect of numbers.

Clearly, the display complexity and field array size will dictate how effectively an individual can count every object in a field, however even for relatively low numbers the individuals ability to count the displayed items can be limited by rapid presentation and subsequent masking of items (Mandler & Shebo 1982, Mandler 1984), or by requiring observers to respond quickly (Kaufman et al. 1949). Research has shown that these interference approaches seem to restrict the ability to count items by limiting an observers ability to shift their "zone of attention" (LaBerge, Carlson, Williams & Bunney 1997) successively to different elements within the display (for general reviews see (LaBerge 1995, Pashler 1998).

Simply put, subitizing is our ability to judge the number of a collection of randomly arranged items more or less instantly. This is demonstrated by the human's capacity to instantly recognize the number of dots on the face of a rolled die without hesitation or of randomly organised dots up to a some small count limit (e.g. 7 ± 2).

3.1.5 Why is memory capacity limited?

The fact that people seem to be able to instantly recognize some number of dots in the visual field, but have to count them above some quantity is evidence of a limit but does not expose the mechanism involved. The simplest reason for this limit could be that the visual system has physical capacity to recognize up to four things much like the main bus on a computer. This is quite a reasonable argument, however given the efficiency of the human biological system one might think that limits such as 4 and 7 are a little low. Peterson and Simon (2000) offer a resolution to this problem in their proposal that the visual system can immediately recognize a set of dots because it has seen these dots in the same array before often enough to build dedicated memory/experience for optimal recognition of that visual arrangement. The number of possible configurations of dots increases exponentially by the number of dots, so if the visual system receives enough examples of four-dot configurations enough experience will be gained to recognize any of them instantly. This does not however apply to any other number of dots except when similar experience of different configuration is gained.

In the proposal by Peterson and Simon's (2000) we see a loosening of any strict limit with a shift to a more logical explanation based on "experience". This work proposes "limits" are determined by interactions between environment and the cognitive system and not on some fixed capacity limit or range. There is undoubtedly some biophysical limitations to the amount of visual information we can process. However because visual recognition requires the interaction of several systems this *experiential* approach seem to make sense of smaller capacity limits such as 4 or 7 (what ever they may represent). This is also supported by the interaction described in suggestions that these limits are connected to long-term memory (e.g., (Miller 1956, Meyer & Kieras 1997, Cowan 2001, Medin, Ross & Markman 2004)).

3.2 Cognitive Load

Memory Load refers mainly to working memory incurring losses when a vigilance task imposes a sustained load on memory and demands a continuous supply of processing resources. This is most prevalent when event frequencies are high and thus interactive tasks need to be managed via visual techniques such as those proposed for good design

by Pfitzner et al. (2003).

Alternately, Cognitive Load refers to the load on working memory during problem solving, thinking and reasoning. The study of cognitive load generally originates from work by Miller (1956) who seems to be the first to propose that there are working memory capacity limits. Miller suggests we are only able to hold seven plus or minus two digits of information in our short term memories. Other notable research into cognitive load is that by Chase and Simon (1973*a*, 1973*b*) who also used the term "chunk" except that they used it to describe how experts use their short term memories.

Cognitive Load is commonly used to described the amount of mental effort needed to process a specific amount of information or achieve a specific task. This load increases with the amount of information required to processed, and learning is inhibited when the quantity of information exceeds a certain capacity of our mental resources. As a concept, it is commonly used in human-machine and human-system interaction research, such as that by Wiggins et al. (Wiggins & O'Hare 1995, OHare, Wiggins, Williams & Wong 1998), in the identification of the information processing requirements of the learner and the demands engendered by the task and impacting systems. From the human perspective the primary impacting factor to the level of load realised in any situation is the level of expertise or experience.

Since learning involves the process of schema construction and skill automation, devoting mental resources to activities not directly related to schema construction and automation may inhibit ones learning. The development of schemata, involves the linking of information gathered by the learner through task experiences to rules associated with the task. Schemata become refined and more automated as a result of practice, and these modifications can decrease cognitive load during task performance. Therefore, training practice relative to task demands can provide learners with the opportunity to develop problem-solving schema that might reduce working memory demands during actual operations and lead to improved performance.

Highlighted in the task interaction process is the role of memory and perception, De Groot and Gobet (1978, 1996) propose that perception and memory are more important differentiators of expertise than the ability to think ahead in the search for chess moves. Chase and Simon's (1973*a*, 1973*b*) research basically paralleled, replicated and extended de Groot and Gobet's work and demonstrates that after viewing chess positions for

only a few seconds, chess masters were able to reproduce these positions much more accurately than less skilled players. In this work, they postulate that chess knowledge is used by experts to create meaningful "chunks" consisting of several chess pieces, thus enabling them to encode structured, but not random, chess configurations more quickly and accurately. Subsequent work by Gobet and Simon (1996) demonstrate that small perceptual chunks are most likely supplemented by larger structures termed "templates". In short, as novices learn, they identify relevant patterns in the world which can be combined with other patterns. This chunking of memory components has also been described as schema construction (see Section 2.1.3.2).

Branching from the field of cognitive load research is CLT (Cognitive Load Theory) which describes how the architecture of cognition has specific implications for the design of instruction. John Sweller (1988) developed CLT while studying problem solving. He suggests that problem solving by means-ends analysis requires a relatively large amount of cognitive processing capacity, which may not be devoted to schema construction. In terms of cognitive load, Sweller states that optimum learning occurs in humans when the load on working memory is kept to a minimum which in turn facilitates the changes to long term memory.

Sweller (1999) suggests cognitive load has broad implications for instructional design and generally speaking CLT can be described as the architecture of human cognition. CLT provides a general framework of empirically based guidelines that help instructional designers manage cognitive load during learning. As an information processing based theory it emphasizes the inherent limitations of working memory and uses schemas as the relevant unit of analysis.

In describing the effect of cognitive load, CLT differentiates between three types of cognitive load:

- intrinsic cognitive load,
- germane cognitive load, and
- extraneous cognitive load.

Extraneous cognitive load is due to the design of the instructional materials for which instructional designers have some ability to control. Chandler and Sweller (1991)

described "Intrinsic cognitive load" as that load related to the complexity of the material. The "complexity of material" results in instructions with inherent difficulty associated with it. For example instructions to complete a simple addition task are less complex when compared to those required for complex matrix manipulations. However, inherent difficulty of a task is generally not altered by an instructor, although Clark et al. (2006) point out that many schemata may be broken into individual "sub-schemata" to be later brought back together and described as a combined whole.

To demonstrate the control instructional designers have, Sweller (2006) outlines two possible ways to describe a square to a student, either visual or aurally. Its fairly clear that because a square is a visual concept it will be much more effective to describe it using a picture of a square than by giving a lengthy and possible difficult verbal description. The visual medium is preferred, as it does not unduly load the learner with unnecessary information. It is this unnecessary cognitive load is described as extraneous cognitive load.

Germane load is the mental processing that allows learning to take place. It is that load resulting from the processing, construction and automation of schemata. While intrinsic load is generally thought to be immutable, instructional designers can manipulate extraneous and germane load. In their discussion of germane load Sweller et al. (1998) suggest that designer should limit extraneous load and promote germane load.

3.3 Inhibiting Irrelevant Information

Section 3.2 and 4 discuss limitations in the amount of information we can cognitively process at anyone time. Aside from early filtering mechanisms like those of the *preattentive* processes in vision there is research that points to other inhibiting mechanisms that manage the processing of irrelevant information during the attentive process by the central executive and working memory mechanisms. It has been proposed that the performance of these mechanisms affect the individuals ability to manage irrelevant information and thus the individuals ability to manage cognitive load.

Comprehension and Inhibition Mechanisms

The effectiveness with which skilled and less skilled readers could use of working mem-

ory capacity in processing discourse was suggested by Perfetti & Goldman (Perfetti & Goldman 1976), and Daneman & Carpenter (Daneman & Carpenter 1980) to demonstrate that there is an active working memory component involved in reading comprehension. Baddeley (Baddeley 1986) suggests that the functioning of the *central executive* was the critical factor in reading comprehension. Further to this, Engle and his co-authors (e.g., (Cantor, Engle & Hamilton 1991, Engle, Cantor & Carullo 1992, La Pointe & Engle 1990)) also proposed that the central executive is highly involved from which it has been suggested that low-span subjects do not have the attentional resources necessary to inhibit irrelevant information a suggestion that is supported by ((Conway & Engle 1994, Engle et al. 1995)). To describe this situation Engle (Engle et al. 1995) proposed the *Inhibition-Resource Hypothesis* that attributes the difference between low-span and high-span subjects in inhibition performance to differences in attention resources localized in the central executive component of the working memory model.

Some research (e.g., (De Beni, Palladino, Pazzaglia & Cornoldi 1998, Gernsbacher 1990, Gernsbacher 1993, Meiran 1996)) can be seen supporting the hypothesis that working memory is affected by the inhibitory mechanism. This research suggests that as a result of a poor inhibitory mechanism the working memory can get overloaded with irrelevant information. The problem with subjects having difficulty in inhibiting irrelevant information is that the level of comprehension an individual realises is affected by how much appropriate information is realised. Prior to Engle's Inhibition-Resource hypothesis Hasher et al. and Stoltzfus et al. (Hasher & Zacks 1988, Hasher, Stoltzfus, Zacks & Rypma 1991, Stoltzfus, Hasher & Zacks 1996) had suggested a resource impact resultant of a inhibitory limitation was evident in the functioning of working memory impacting comprehension. Research by Hartman & Hasher (Hartmann & Hasher 1991), and Hamm & Hasher (Hamm & Hasher 1992) indicated that comprehension deficits might be the result of poor inhibitory mechanisms which impede the abandonment of no-longer-relevant thoughts. This was subsequently supported by a range of different research (e.g.,(Engle et al. 1995, Engle et al. 1999, Engle 2001, Engle 2002, May, Hasher, Zacks & Multhaup 1999, Gernsbacher 1990, Gernsbacher 1993, Meiran 1996)). Engle and associates proposed that short-term memory is an important component of general fluid intelligence and that it is a domain-free limitation in the ability to control attention. From this research, they concluded that working memory capacity, or executive attention, becomes a critical component "under conditions in which interfer-

ence leads to retrieval of response tendencies that conflict with the current task". May et al. (1999) supports the hypothesis that the functioning of working memory influences comprehension through their research into the elderly and memory performance. Gernsbacher et al. (Gernsbacher 1990, Gernsbacher 1993) and Meiran (Meiran 1996) also supported this through their research into reading comprehension showing that poor readers were less able to suppress inappropriate meanings activated by terms with ambiguous meanings.

Other models can also be seen as accounting for a critical relationship between working-memory, inhibitory mechanisms and performance. The time-based resource-sharing model of working memory by Barrouillet (Barrouillet, Bernardin & Camos 2004) predicts that lower working-memory resources reduce the amount of attentional resources available to activate knowledge from long-term memory which implies that poor working memory resources not only impair the formation of associations in long-term memory but also the retrieval of existing associations. Cowan (Cowan 1988, Cowan 1999) proposed an "embedded processes model" of working memory which also suggest that the performance of working memory effects comprehension. Finally, recent work by Imbo & Vandierendonck (Imbo & Vandierendonck 2007), and Barrouillet & Lepine (Barrouillet & Lepine 2005), also propose new models and supports the concept of working-memory resources effecting performance in comprehension and memory formation.

3.4 Short Term Memory Volatility

In tasks, such as language comprehension, that require the maintenance and rapid retrieval of immediate task relevant information for working memory processes, the volatility of information is crucial to the successful realisation of any such task. Peterson and Peterson (1959) determine the duration, or volatility, of short term memory in research that demonstrated that three letters can be recalled correctly only about 10% of the time after 18 seconds of distracting activity. In support of this both Jacoby and Bartz (1972), and Watkins and Watkins (1974) demonstrate that subjects perform differently in memory tasks if they expect to be distracted during the retention interval than if they do not, perhaps because they form a secondary memory trace. In similar research looking at the effect of priming on the successful retrieval of working memory,

it is demonstrated by Craik and Tulving (1975), and Hyde and Jenkins (1969), that the probability of success is dependent on the manner in which the subject has been primed (e.g., semantically vs. acoustically).

Under alternate conditions to Peterson and Peterson (1959), and Muter (1980) demonstrate that a better estimate, of the volatility of short term memory, can be obtained by studying forgetting under conditions in which subjects do not expect a recall test with distracting activity during the retention interval. Under these conditions, perhaps less contaminated by secondary memory involvement, three letters could be recalled correctly only about 10% of the time after only 2 or 4 seconds of distracting activity. These observations are also supported by similar results obtained by Sebrecht, Marsh and Seamon (1989) and specific research by Marsh, Sebrechts, Hicks and Landau (1997) supports and extends this work in eliminating rehearsal time as a factor contributing to working memory performance.

It is clear through research, such as that by Marsh et al. (1997), that working memory can be very volatile (persistent for less than 2 seconds) and that when distractors are involved, subject short term memory will quickly start to decay (Peterson & Peterson 1959, Hyde & Jenkins 1969, Craik & Tulving 1975). Because of this volatility, it is not only important that users have structures to aide in remembering information, but that they are not required to remember it for an extended period of time.

3.5 Performance

Interactive tasks can often produce less than optimal performance brought about by *task complexity* and/or *interactive system complexity* and/or *human cognitive limitations* such as fatigue and cognitive capacity limits. The first two are normally addressed through the redesign of the task or system. As for the third, the human cognitive system is generally said to be biologically limited through features constrained in capacity like memory, data paths and processing, much like those problems that plague computers.

The problem of cognitive performance degradation, in this context, seems to have been first addressed in the paper "On data-limited and resource limited processes" by Norman and Bobrow (1975). They suggest that cognitive functions involve many separate and independent processes working together through the exchange of informa-

tion and can be referred to as a "program". These functions/programs require input and compete for resources including processing function, communication and memory. These resources are based on brain tissues that are limited in quantity, implying a limitation of resource. The systems managing these resources/activities are a form of high level executive described as the "supervisory system".

An important concept that influences considerations for user interaction is that the limited nature of these resources results in *graceful degradation*. This concept, also discussed in Section 2.2.2, describes the situation of a *smooth degradation* in performance rather than a catastrophic failure to finish a task when these limited human processes/resources become overloaded. This smooth degradation is widely recognised as a property of the human processing system referred to as "the principle of graceful degradation".

As an incremental affect, graceful degradation will logically result in increasing error if task and/or interface induced load, that effects these limited systems, is not managed. This can be achieved through techniques like event frequency reduction or scene complexity reduction.

For discussion of Long-term memory and structural relevance see Section 2.1.3.2.

3.6 Concluding Observations

As we have seen, short term memory is limited and capacities are used to describe this restricted nature, e.g. seven plus or minus two chunks of information. This volatility means users will often forget pertinent information especially in the presence of distractions. Problems arising from this volatility are discussed in Section 2.4. The fact that we seem to chunk in some form means that if search tasks and sub-tasks can be tailored so that visually transmitted information can be naturally realised in chunks, the user is more likely to not miss or loose information through the task and thus realize a better task outcome. This might be done through approaches such as the visual presentation of document clusters using 7 ± 2 descriptors to represent each cluster, or 7 ± 2 dimensions to present the clusters against.

The usefulness of chunks can be leveraged if there are appropriate schemata in place that the user can draw on. In Section 3.1.4 we saw that people have the ability to judge

the number of a small collection of randomly arranged items more or less instantly (subitizing). However, once outside a certain range (dependent on the individual) accuracy drops off (Kaufman et al. 1949) and counting starts. Also, it was seen that the success of counting is dependent on the manner in which the objects are displayed.

In Section 3.4 it is demonstrated through research such as Marsh's (1997) that short term memory can be very volatile and when distractors are involved users experience greater levels of difficulty (Peterson & Peterson 1959, Hyde & Jenkins 1969, Craik & Tulving 1975). Because of this volatility, it is not only important that users have structures to aide in remembering information, but that they are not required to remember them for an extended period of time.

Schemata are learned through repetition such as the presenting of groups more appropriate documents in a visual field in a similar manner to each other (e.g, similar intensity, angle, shape, colour and so on). Well developed schemata make it easier to remember items that fit within a schema. Thus, experts with well developed schemata outperform novices so an interactive interface should deliver consistent interactive devices that match the users cognitive style (see Section 2.3) to allow the novice to become an expert as quickly as possible. Also, interaction device design should draw on the experience of the expert in the delivery of information because they will have a better understanding of what information is important in a specific task.

The realisation of experience and expertise in interactive device design is further supported by developments in the theory of cognitive limits (see Section 3.1.5). Peterson and Simon (2000) proposed a shift from strict limits (e.g., 4 or 7) to a more logical explanation of "limit" to one based on "experience" as a representation of interaction between environment and the cognitive system, and not one based on some fixed capacity limit or range. As such interface design should recognise the affects of experience and expertise.

In research by Atkinson, Campbell and Francis (1976, 1976), and Mandler and Shebo (1982) it is demonstrated that visual afterimages can also effect subitizing and counting. This raises the question of "can afterimages or the mechanisms behind them be used in a manner to improve task realisation such as by giving the user a task-relevant residual image?". This may include such concepts such as spatial relationship, colour, shape and grouping.

Overall, this chapter points to the need to manage the number of things in chunks and groupings to optimize the realization of any interactive text search tasks. The thesis draws on the research presented, to target the number and type of words needed to identify the topic of a cluster of documents or document in a visualisation ("How many words do people naturally use to describe/query for documents?"). This research program has also spawned and supported related projects not discussed in this thesis that examine the impact of visual attributes (Treharne, Pfitzner, Leibbrandt & Powers 2008, Treharne, Pfitzner & Powers 2007, Treharne, Pfitzner & Powers 2006) and emotional queues (Powers, Leibbrandt, Pfitzner, Luerssen, Lewis, Abrahamyan & Stevens 2008).

Chapter 4

Visual Processing

User interaction with a visual search interface (the screen) is affected by how information is presented. When looking at the cognitive aspect of user interaction the field of cognitive psychology is surveyed to develop an understanding for how visual information is processed and managed. This involves looking at how the human stores and retrieves visual information, and the processing of raw aural input in both the perception and cognition stages.

The previous sections discuss specific areas of research relative to limitations of cognition. This chapter follows on from this with discussion looking closely at visual aspects of user information realization such as what do we see, how do we see it and what are the general effects on cognition.

4.1 What do we see?

The key to understanding "what we see" lies in the definition of *see*. If by "see" we mean any visual stimulus that is realised optically, whether being cognitively processed or not, then we could say yes we see everything in our visual field. However, a more reasonable definition of "see" should include the concept of *recognition* which implies that to "see" we perceptually and cognitively process in such a way as to realise structure, form, and/or meaning, for a given visual stimulus.

It might be easy to think that we see everything in our visual field, however our very rich visual environment (containing a relatively large amount of available stimulus)

is far more than we can cognitively process at anyone time. This lack of processing ability is recognized by an abundance of literature , for example (Miller 1956, Carlson, Sullivan & Schneider 1989, Just & Carpenter 1992, Conway & Engle 1996, Halford et al. 1998, Cowan 2001, Kosara, Miksch & Hauser 2002), that either demonstrates and/or theorises limits on how much stimulus we can process. In what ever form a limit is represented, be it the number 4 or 7, or binary or quaternary, or what ever representation may be used, there is clearly evidence for a limit and this limit is relatively small compared to the amount of information available perceptually from our visual field.

At a biophysical level this lack of processing ability is managed by differentiating regions of the retina. The retina detects photons of light with photoreceptive cells organised into regions of different densities. Located in the centre of the macula region of the retina is the *fovea* (also known as the *fovea centralis*) which is the most dense of these light receptive regions. It is responsible for our sharp high fidelity central vision necessary for tasks such as reading, watching television, driving, and any activity requiring a high level of visual detail. Surrounding the fovea region is the *parafovea belt* which is a region of moderate density photoreceptive cells. Surrounding this is the ring of cells is the *perifovea*, a region of photoreceptive cells that delivers below optimal visual acuity. Beyond these foveal regions is a larger peripheral area that delivers information of low resolution that detects gross events like general shapes, colours, size and movement. This final region comprises most of the retina compared to the small high acuity regions.

Given this progression from very small region of high acuity out to very large region of low acuity, any interactive task that needs to visually attract the attention of a user to a specific area of the screen need not use high fidelity graphical events. This is because the region of the eye that is likely to receive the stimulus from any such event will most likely be the outer low fidelity region. However, it is obvious that the level of detail required when attention is gained will need to be tailored to the requirements of the task. For example, if a task requires "reading" then higher fidelity is required compared to a "button pressing" task where targets are larger and the task is simpler.

In HCI the primary information transmission channel is the visual channel and although the staged foveal mechanism addresses some difference between available visual

information and processing power by reducing the mass of input through staged acuity it does not account for the manner in which we cognitively process the still very rich visual channel. So in terms of "what we see", on a gross visual level we seem to see everything and the perceptual experience seems full and rich, but on another level we need to pay attention to something in our line of sight to develop an intricate understanding for what we are looking at. In relation to perception and the cognitive processes involved in reconciling these opposing situations it has been hypothesized that there are two kinds of psychological processes involved *Preattentive* and *Attentive processes*.

Visual Search Paradigm Of the many experimental techniques developed to study the characteristics of pre-attentive and attentive processes the technique most relevant to the *search* topic of this thesis is called the *visual search paradigm*. In this paradigm, participants are shown visual displays containing varying numbers of objects and are asked to determine whether a pre-specified target is included in the display.

A simple example of this might be where a person is asked to look for a green circle in a display containing blue circles and green squares. The dependent measure in this paradigm is the time required to complete the search (as indicated by a selection being made). The primary independent variable is the display size or the number of items in the display. Increasing response times with increasing display times suggests that attention is needed to find the target. In contrast, if increasing the display size does not affect search time, the search is said to be based on visual properties that are processed pre-attentively.

4.2 Attentive Processing

Attentive processes see the perceiver controlling the locus of attention. In this sense, the high acuity region of the fovea is applied to one component of the visual field at a time such that all of the available visual processing capacity is focused on a very small segment of the total field. This allows enough specific information for the perceiver to do things like know who or what something is or an object's purpose/function.

4.3 Preattentive Processing

Pre-attentive processing of visual information occurs independently of the focus of attention and is performed automatically on the entire visual field detecting basic features of objects in the display. This occurs on such features as colours, closure, line ends, contrast, tilt, curvature and size that are extracted from the visual display by the pre-attentive system. Subsequently these are combined by the focused attention system into coherent objects. Pre-attentive processing is done quickly, effortlessly and in parallel without any attention being focused on the display. Pre-attentive processing occurs automatically to all of the objects in the visual field whether they are the focus of attention or not. Because of this, these processes are largely responsible for the phenomenological rich feel of visual perception (Treisman 1985, Treisman 1986). Typically, tasks that can be performed on large multi-element displays in less than 200 to 250 milliseconds are considered pre-attentive (Healey, Booth & Enns 1996).

In HCI, if low-level visual system (staged acuity) and pre-attentive processes can be harnessed during visualization, attention might be more efficiently and effectively drawn to areas of potential interest in a display. Obviously, this cannot be accomplished in an ad-hoc fashion so the visual features assigned to different data attributes must take advantage of the strengths of our visual system, must be well suited to the analysis needs of the viewer, and must not produce any visual interference effects that could mask information in a display. Table 4.1 lists some of the visual features that have been identified as pre-attentive. Experiments in psychology have used these features to perform the following pre-attentive visual tasks:

Target detection users rapidly and accurately detect the presence or absence of a "target" element with a unique visual feature within a field of distractor elements

Boundary detection users rapidly and accurately detect a texture boundary between two groups of elements, where all of the elements in each group have a common visual property

Region tracking users track one or more elements with a unique visual feature as they move in time and space

Counting and estimation users count or estimate the number of elements with a unique visual feature

Visual Features	**Associated Research**
line/shape orientation	(Julsz & Bergen 1987),(Sagi & Julsz 1985*a*), (Wolfe, Friedman-Hill, Stewart & O'Connell 1992), (Weigle, Emigh, Liu, Taylor, Enns & Healey 2000).
length, width	(Sagi & Julsz 1985*b*),(Treisman & Gormican 1988).
closure	(Julsz & Bergen 1987),(Enns 1986), (Treisman & Souther 1985).
size	(Treisman & Gelade 1980),(Healey & Enns 1998), (Healey & Enns 1999).
curvature	(Treisman & Gormican 1988).
density, contrast	(Healey & Enns 1998),(Healey & Enns 1999).
number, estimation	(Sagi & Julsz 1985*b*),(Healey, Booth & Enns 1993), (Sagi & Julsz 1985*a*),(Trick & Pylyshyn 1994).
colour (hue)	(Nagy & Sanchez 1990),(Nagy, Sanchez & Hughes 1990), (D'Zmura 1991),(Yokoi & Uchikawa 2005), (Kawai, Uchikawa & Ujike 1995),(Bauer, Jolicoeur & Cowan 1996), (Healey et al. 1996),(Bauer, Jolicoeur & Cowan 1998), (Healey & Enns 1999),(Treisman 1985).
intensity, binocular luster	(Beck, Prazdny & Rosenfeld 1983),(Treisman & Gormican 1988), (Wolfe & Franzel 1988).
intersection	(Julsz & Bergen 1987).
terminators	(Treisman 1985),(Julsz & Bergen 1987).
3D depth cues, stereoscopic depth	(Enns 1990),(Nakayama & Silverman 1986), (Julsz 1971).
flicker	(Gebb, Mowbray & Byham 1955),(Mowbray & Gebhard 1955), (Mowbray & Gebhard 1960),(Brown 1965), (Julsz 1971),(Huber & Healey 2005).
direction of motion	(Nakayama & Silverman 1986),(Driver, McLeod & Dienes 1992), (Huber & Healey 2005).
velocity of motion	(Tynan & Sekuler 1982),(Nakayama & Silverman 1986), (Driver et al. 1992),(Chey, Grossberg & Mingolla 1997), (Hohnsbein & Mateeff 1998),(Huber & Healey 2005).
lighting direction	(Enns 1990).
3D orientation	(Enns & Rensink 1990),(Enns & Rensink 1991), (Liu, Healey & Enns 2003).
artistic properties	(Healey 2001),(Healey & Enns 2002), (Healey, Enns, Tateosian & Remple 2004).

Table 4.1: Preattentive visual features and associated research [1]

[1] adapted from Healey (2004) and Hearst (2003)

4.4 Distractors and visual processing

When using the *visual search paradigm* in any form, all information other than the intended target can be described as distractor information. That is visual information that might impact search performance by impacting the amount of cognitive capacity for the filtering task. The impact might be seen in the performance different sub-activities of the search process and even in a manner opposite to what one might expect. For example McSorley and Findlay (2003) report a set of results showing that increasing the number of distracting elements in a visual-search task improved oculomotor search performance and as a consequence improved perceptual selection.

So what might be the prime mechanism behind the effect of distractors on search performance? Evidence by Shisler, Conway, Tuholski & Engle (1995) demonstrated that working memory (see Section 2.1.2) affected task performance through the amount of negative priming of distractors when presented to subjects as targets. However, from their research it was not clear if high loads on working memory affected the inhibition of distractors, or reduced their encoding into memory.

The effect of memory on task-directed behaviour has become more evident in studies using the Stroop-like paradigms, for example (Kane & Engle 2003), to observed differences in working memory span correlated to performance. Kane and Engle (2003) demonstrated that low-span subjects make a more erroneous response to a distracting incongruent word in the Stroop task than high-span subjects. This implied that the capacity of memory was affecting the individuals control of distractor response.

The evidence from these Stroop like tests has been further tested and supported by a series of experiments by Lavie and colleagues (De Fockert et al. 2001, Lavie 2000, Lavie, Hirst, De Fockert & Viding 2004). These experiments showed that working memory affects distractor management by influencing the priority processing of relevant and irrelevant task stimuli. In research by Lavie (2005) results supported the proposition that working memory affected the irrelevant stimuli rejection process during visual search tasks. Lavie suggested that this demonstrates reduced working memory availability for the selective attention task should result in a reduced ability to attend to relevant stimuli. She also proposed that if one search item is a strong competitor for selection then rejection of the competing distractor should be dependent on the availability of working memory for the goal directed control in the search task.

The point to be made about distractor effects in the search task is aptly demonstrated in research by Lavie (2000) and Jiang (2001) that showed that when processing task-relevant stimuli, under high perceptual load, distractor perception can be eliminated (early selection scenario) if the load is sufficiently high. It was also been shown that loads such as that on working memory impact the effect of irrelevant stimuli under lower perceptual loads (late selection scenario). However, as indicated by Wolfe, Friedman-Hill, Stewart & O'Connell (1992, 1994) there is a compromise to be made between the limits on parallel visual processing (pre-attentive), the demands of a complex visual world and the different processing mechanisms employed, and as such the appropriate load for a given context needs be realised.

4.4.0.1 Distractor Effects

Different user contexts in interactive tasks often results in the presentation of more information than is relevant to the user task and is often required to guide the selection and presentation of subsequent and more relative information. Any information that is not relevant is considered a distractor; this includes information emanating from the users environmental as well as that of the display. Because the environment is far too extensive and complex to consider at this stage it is ignore in the following discussion of research into distractors.

Evidence by Shisler et al. (Shisler, Conway, Tuholski & Engle 1995) demonstrates that working memory affected task performance through the amount of negative priming of distractors when presented to subjects as targets. This research clearly demonstrated that high loads on working memory affected the inhibition of distractors or reduced their encoding into memory.

The effect of memory on task-directed behaviour was further defined by stroop tests that demonstrated a correlation between differences in working memory span and performance. An example of this can be seen in work by Kane and Engle (Kane & Engle 2003) that showed that low-span subjects make a more erroneous response to a distracting incongruent word in the Stroop task than high-span subjects. This implied that the capacity of memory was affecting the individual's control of distractor response.

The evidence from Stroop like tests has been further tested and supported by

a series of experiments by Lavie, De Fockert and others (Lavie 2000, De Fockert et al. 2001, Lavie et al. 2004). These experiments showed that working memory affects distractor management by influencing the priority processing of relevant and irrelevant task stimuli. When publishing the results of further research into the area Lavie (Lavie & Defockert 2005) suggested that this demonstrates that a reduced working memory availability for the selective attention task should result in a reduced ability to attend to relevant stimuli. This further research supported the proposition when it demonstrated that working memory affected the irrelevant stimuli rejection process during visual search tasks.

Lavie (Lavie & Defockert 2005) proposes that if one search item is a strong competitor for selection then rejection of the competing distraction should be dependent on the availability of working memory for the control goal directed control of the search task. It is also suggested that loads such as that on working memory impact the affect of irrelevant stimuli under lower perceptual loads (late selection scenario). In research by Lavie and DeFockert (Lavie & Defockert 2003, Lavie & Defockert 2005) studying the effects of perceptual load and target-stimulus degradation on distractor processing, they suggest that distractor processing depends on the extent to which high perceptual load exhausts attention in relevant processing, and provide a dissociation between perceptual load and general task difficulty and processing speed.

Other research by Lavie (Lavie 2005) also suggested that high cognitive load might to be useful depending to the context. The research demonstrated that distractor processing depends on the type and level of load involved in the processing of target stimuli. It was shown that high perceptual load could eliminate distractor processing while high load on frontal cognitive processes increases distractor processing. Further to this, it was shown that when processing task-relevant stimuli involving high perceptual load distractor perception can be enhanced (early selection scenario) if the load is sufficiently high (Lavie 2000, Jiang & Chun 2001).

4.4.0.2 Reading Comprehension

De Beni's (1998)research suggests that working memory affects reader comprehension. This is due to the overburdening of working memory with irrelevant information. The resolution for this problem logically is to encourage the reader to only maintain cru-

cial information in working memory. Further to this Daneman and Carpenter (1980) theorized that an active working memory component is involved in reading comprehension and demonstrated that the correlation between reading comprehension and a short-term memory task was lower than the correlation between reading comprehension measures and listening span. Drawing from this Baddeley (1986) used it to support his theory that the central executive plays a critical role in comprehension, especially reading comprehension. Baddeley's proposition has subsequently found wide support, for example Cantor, Engle & Hamilton (1991); Engle, Cantor & Carullo (1992); La Pointe & Engle (1990).

The idea that attentional resources such as working memory are involved in the management of stimuli especially irrelevant stimuli was supported by Conway & Engle (1994), Engle, Conway, Tuholski & Shisler (1995) and De Beni, Palladino, Pazzaglia & Cornoldi (1998)). Adding to this, in experiments conducted by (Hasher et al. 1991), (Hasher & Zacks 1988) and (Stoltzfus et al. 1996) it was found that elderly people have difficulty in retrieving, quickly and efficiently, the antecedent information necessary to form an inference. This happened particularly in oral presentation, in comparison to written presentation, that is, when subjects had to maintain all the text information memorized, confirming that working memory is centrally involved in comprehension.

From all this it is hard to deny that working memory is involved in reading tasks and comprehension. This implies that when presenting users with written information working memory constraints must be taken into account when optimising for human performance.

The relationship between working memory and reading comprehension was also tested by Engle (1992) using three experiments and four hypothesies. In the first 2 experiments, a moving window procedure was used to present the operation-word and reading span tasks. High-span and low-span subjects did not differentially trade off time on the elements of the tasks and the "to be remembered word". Furthermore, the correlation between span and comprehension was undiminished when the viewing times were paired out. Experiment 3 compared a traditional experimenter-paced simple word-span and a subject-paced span in their relationship with comprehension. The experimenter-paced word-span correlated with comprehension but the subject-paced span did not. The results of all three experiments support a general capacity explana-

tion for the relationship between working memory and comprehension.

In short, working memory has processing and storage functions that compete for a limited capacity. More demanding processes consume more of the available capacity there by decreasing the amount of additional information that can be stored and processed in working memory.

This study investigated whether individual differences in working memory span are associated with different working memory management strategies during a reading task. In Experiment 1, probe questions were presented on line during reading to determine whether thematic information was maintained in working memory throughout comprehension. The data indicated that readers across the range of working memory span maintained thematic information in working memory throughout the reading of a given passage. In Experiment 2, sentence reading times and accuracy for both topic and detail questions were measured in two conditions: when topic sentences were present and when topic sentences were absent. Subjects performed similarly across the range of working memory span in the topic-present condition, but lower span subjects performed more poorly on detail questions in the topic-absent condition. In Experiment 3, the topic-present condition of the second experiment was replicated, except that subjects expected to receive questions about details only. Thematic processing and retention of topic and detail information all increased with span.

Taken together, these results suggest that, for more difficult text processing tasks, high- and low-span subjects adopt different working memory management strategies and these strategies influence what is learned from reading the text.

4.5 Multiple Dimensions in Search

The task of visual search in the real world is not as simple as detecting a target using one single dimension (such as colour, size or shape) from a simple field such as those used in the typical visual search study. Given the complex multidimensional nature of HCI, consideration should be made for complicated tasks such as detecting multiple targets with a single perceptual dimension or with multiple different dimensions (e.g. finding a shape of specific colour). In the context of guided search, Wolfe (1994) suggests that there are limitations to our ability to search multiple targets using a single dimension

because it only provides one signal to guide attention to a single target, however multiple signals of different dimensions can be combined to guide attention in a more complex search task.

This highlights context as a major determinant to target realisation in visual search. Work by Beiderman et al. (Biederman 1972, Biederman 1982) suggests that visual context affects visual processing and points to the deterioration in performance realised as scene or task complexity increases. Their work demonstrated that time and accuracy to detect targets is affected by how the targets fit into the scene (their context). Work by Treisman et al. (Treisman 1988, Treisman & Gelade 1980) suggests the different features of visual stimuli, such as colour and size, are all extracted *pre-attentively* in parallel. However, serial attention is required to the locate each item and integrate such different features in order to produce appropriate multidimensional percept's of objects given the combinations of particular dimensions such as colours and size and the general scene.

As we acquire new information from a visual scene our percept's shift in accordance with our past experience and knowledge of similar objects and scenes, and so to some extent the process of recognising complex objects (using multiple single dimensions) depends on prior knowledge/experience. Chun et al. (Chun & Jiang 1998, Chun & Jiang 1999, Chun & Nakayama 2000) examined how contextual knowledge may be acquired through learning and proposed that the implicit learning of context can efficiently guide visual attention toward target information.

4.6 Context and Implicit Learning

The effect of context in visual search was addressed in work by Chun et al. (Chun & Jiang 1998, Chun & Jiang 1999, Chun & Nakayama 2000) who examined how contextual knowledge may be acquired through implicit learning. These studies show that implicit learning of context can efficiently guide visual attention toward target information. This is highlighted in work by Chun and Jiang (1999) which demonstrated that complex motion trajectories could be implicitly learned to help localize a moving. Thus, various visual attributes can be implicitly learned to guide visual attention toward the relevant aspects of the displays.

Contextual cueing is not restricted to learning of spatial configurations. Further studies revealed that information about the item shapes could be implicitly learned to facilitate visual search of novel objects.

4.7 Visual Feature Integration

In presenting graphical representations of data the manner in which a scene is visually processed is important in the presentation of the data for rapid and accurate discrimination of the individual data concepts represented within the scene. At a physical level "Feature Integration" is a concept that suggests visual features such as colour, shape, orientation and motion are treated modularly and by separate areas of the visual cortex. In support of this recent research by Bright et al. (Bright, Moss, Stamatakis & K. 2005) suggested that "a clear picture has emerged in which the human perirhinal cortex and neighboring anteromedial temporal structures appear to provide the neural infrastructure for making fine-grained discriminations among objects".

Theories that describe the processes involved in this generally start with Triesman Feature Integration Theory, which states that all stimuli are first processed in parallel and then serially in a conjunction search in which certain features are looked for in combination. Feature search occurs pre-attentively and is the rapid search for targets defined by primitives, whereas conjunction search is a slow serial search for targets defined by the conjunction of primitive features within the scene. Features mapped in the feature search can be combined by focused attention on the object while the act of feature combining can be influenced by previous knowledge. Interestingly, in the absence of focused attending or stored knowledge, features from different objects will be combined randomly, producing "illusory conjunctions".

It should be noted that this model does not account for the effects of the similarities between distracters and the target as recognised by Treisman & Sato (1990) and Treisman & Gelade (1980). For example Treisman & Gelade found that searching for a 'T' with 'I' and 'Y' distractors is easier than searching for a 'T' with 'I' and 'Z' distractors. This is suggested to be due to the T having features in common with the 'I' and 'Z'. The number of distractors also effected performance, such that the more distractors in a scene the greater the difference in the search times.

A further development of this model was Wolfe's (1989) "Guided Search theory" which stated that serial and parallel processing occurs simultaneously in differing amounts. At first an 'activation map' is created and objects that are similar to the target are identified. Next, the similar objects are processed serially while all other objects are processed in parallel to identify the target.

Along similar lines, Broadbent's "Filter Theory" proposes that visual information is similarly processed based on the physical characteristics of the information. However, information until a limit some processor limit is reached which, much like a hole in a wall, filters or blocks out some of the information/scene. The information is filtered based on physical characteristics and is passed through a limited capacity channel that is all or none. This filter is consciously controlled and interpretation happens post filtering.

Although this is only a short treatment of the manner in which features of a scene are processed it is clear from readings in the field that there is a focusing attribute involved and much like a spotlight the region being attended gets more resource/treatment. In wanting to graphically highlight different salient features of a data set the manner in which these features are highlighted will determine weather they become distractors or standout as critical features in need of investigation. To understand this process better the processes involved in visual attention must be better understood.

4.8 Visual Attention

The visual system cannot process fully all the objects or stimuli that at any one time are projected at the retina. To manage this situation attention mechanisms are needed to select for further processing of information that is currently task-relevant, whilst ignoring irrelevant information. There are three broad classes of theories that describe this situation *Object-based*, *Discrimination-based* and *Space-based* theories. As indicated by Neisser (1967) and Kahneman & Henik (1981) Object-based theories suggest that attention is directed to objects or perceptual groups within the visual scene that has been previously segmented on the basis of gestalt principles and that the number of separate objects that can be perceived simultaneously is limited. In subsequent work Neisser (1975) rejected the whole notion that attention involve special mechanisms suggesting that selective attention is a direct consequence of an individuals skill in perceiving.

Discrimination-based theories propose a limit on the number of separate discriminations that can be made. These discriminations are made on specific features of a scene. Examples of these features can be seen in Julesz's (Julesz 1981) work that identified a class of simple features (textons) used in discrimination such as rectangles, ellipses and certain lines. Space-based theories, such as Posner's (1980) *spotlighting*, and Eriksen & Yeh's (1985) *zoom lens*, suggest that the spatial area from which information can be taken up is limited.

Neisser's (1975) work suggests that experience/skill as well as attention seem to play fundamental roles in visual processing. Previous experience or knowledge allows selectively attention by realising contextually relevant information and ignoring vast amounts of irrelevant information. The inverse is also true in that failure to attend to critical information can reduces the efficiency of visual processing as seen in extreme cases of inattention that can result in *functional blindness*. Research that illustrates this point is that of Mack & Rock's (1998) into *attentional blindness*, Chun & Potter's (1995) and Raymond, Shapiro & Arnell's (1992) into the *attentional blink* and Rensink, O'Regan, & Clark (1997) and Simons & Levin (1998) into *change blindness*. In these cases, visible information goes unnoticed when not under focused attention. The importance of previous experience in visual processing can be seen in Beiderman's (1972) work that proposes visual information is processed more efficiently when visual experience provides schemata to organize complex scenes.

4.9 Visual Spotlighting

Human modeling research, as far back as that by Averbach and Coriell (1961) and Sperling (1963), has been demonstrating that people attend more to some parts of the visual field than others. This closely relates to the fact that our eyes are structured in such as way as to require focalised attention (see the introduction to this section) and has led to the suggestion that visual attention should be thought of as a "spotlight". Attention cannot be focused on more than two spatially separate points at once and so visual attention is based on spatial selection, because items in the visual field are distributed across space. What is in the spotlight (foveal region of retina) is attended more than what is outside of the spotlight (parafoveal region of retina).

Evidence of visual spotlight can be seen in research such as that by Eriksen and

Eriksen's (1974) in which they suggest that visual spotlight provides a means to selectively manipulate the presence or absence of response competition while keeping other task demands constant. Their research demonstrated that the placing of distractors within 1 degree of visual angle from the target made ignoring the distractors impossible. Distractors outside of that (visual) area affected performance less than those within that (visual) area, which suggests that area is focused on more, as if it was lit by a spotlight.

Posner et al. (1980, 1978) looked to further characterise spotlighting and developed the spatial cueing task to measure covert shifts of visual attention. In this task, observers are required to respond to a peripherally presented target, which is preceded by a cue that serves to direct covert visual attention to a particular location. From this research, they proposed that attention can be shifted to different locations within the visual field in one of two ways: either "overtly" or "covertly". An *overt* shift of visual attention occurs when the eyes, head or body move to align the fovea with a new object of interest. While the focus of attention may in this way coincide with the area of the visual field to which the fovea is directed, the two are also potentially dissociable. A *covert* attention shift occurs when the focus of attention moves to an area of the peripheral or parafoveal visual field independent of overt movements. You may recognise this event from the saying "looking out of the corner of your eye". Grindley and Townsend also demonstrated that attention can be shifted without eye movement.

Visual Attention and the spotlight metaphor Driver J. and Baylis (1989) suggest that space is the medium for visual attention and that it is generally recognised as being analogous to a "spotlight", as suggested for example by Broadbent (1982). They also suggest that relative positioning plays an important role in visual grouping and that position may also play a unique role in perceptual integration. A common understanding of a key aspect of the spotlight metaphor is that attention selects contiguous regions in the visual field for further processing. This is also supported by different variations on the spotlight theme, such as in the development of Eriksen and St-James's (1986) *zoom-lens model* and Downing and Pinker's (1985) *gradient model*. Research by Nissen (1985) and Triesman and Gelade (1980) supports the idea that different attributes of an object, such as its colour and shape, might be combined by means of relative position (common spatial coordinates) in otherwise separate representational systems. This research supports a spatial spotlight account of visual attention

while suggesting that spatial coordinates are a means of perceptual integration and that attention can be directed to perceptual groups whose components are spatially dispersed.

Research by Eriksen and Eriksen (1974) suggests that interfering effects of distractors diminished with increasing distance from the target, however this demonstrated that grouping targets and distractors by common motion can have more influence than their proximity which is consistent with an alternate proposition to spotlighting that proposes that attention is assigned to perceptual groups. This suggests that the spotlight metaphor seems to be limited in its account of visual attention in a dynamic environment.

4.10 Proficiency in Visual Search

To test visual search proficiency Schneider & Shiffrin (1977) presented people with displays that contained varying numbers of elements (either letters or digits). Experiments demonstrated that subjects could easily pick a digit out among a display of letters, regardless of how many other letters there were. However, picking out a letter among a display of letters was quite difficult, and was more difficult the more (distraction) letters there were.

These results suggest a difference between *consistent* and *varied* mapping. Consistent mapping occurs when targets are never distractors & distractors are never targets. Varied mapping occurs when targets may be distractors and distractors may be targets. For example, if an experiment using 4 and 8 as targets and L, P, Q, Z as distractors, then it is using consistent mapping. If the experiment uses 2, 9, K, and W as targets and 3, 9, X and W as distractors, then it is using varied mapping. Shiffrin and Schneider suggest that when a task has consistent mapping, then automaticity will be achieved. However, automaticity has a price, because changing the mapping can lead to errors as a result of automaticity.

4.11 Saccade

Saccades are fast, ballistic eye movements used to rapidly change gaze position from one region in the visual field to another. It is a function of visual search in visual

environment exploration and information gathering. Saccades are used to point the high-resolution foveae at a specific object within a scene or a spatial location of interest. Controlled foveation as an information-gathering task indicates that saccades and perception are related. Interestingly though frequent saccades normally occur without conscious deliberation and are the result of a neural decision, which specifies where the saccade should land and when to initiate it.

Beutter, Eckstein & Stone (2003) investigated the relationship between the visual processing used by saccades and perception during search by comparing saccadic and perceptual decisions under conditions in which each had access to equal visual information. Their results demonstrate that the accuracy of the first saccade provides much information about the observer's perceptual state at the time of the saccadic decision. They also demonstrated that saccades and perception use similar visual processing mechanisms for contrast detection and discrimination.

In support of the proposal that object spacing and display size impact the process of visual processing Vlaskamp, Hooge & Over (2005) found that saccade increased proportionally with spacing and fixation time decreased by a small amount with increasing spacing. This can be interpreted to say that visual span roughly scales with element spacing or in other words, the number of elements processed per fixation is kept constant. An explanation for this is that crowding limits the area that is inspected per fixation.

This highlights the importance of eye movement in HCI as source of information about user perceptual processes. What a person is looking at is assumed to indicate the thought "on top of the stack" of cognitive processes (Just & Carpenter 1976). The eyes current focus can give a trace of what is being attended in a visual display. Eye fixation can reveal the amount of processing being applied to an object in a scene. Therefore, by knowing what is important in a scene relative to a specific task and making observations about saccadic motion and fixation researchers can glean information about the visibility, meaningfulness and placement of specific elements in the display. These observations can be used to study things like human attention, problem solving, reasoning, mental imagery and search strategies (Altonen, Hyrskykari & Raiha 1998, Byrne, Anderson, Douglas & Matessa 1999, Goldberg & Kotval 1999, Henderson, Pollatsek & Rayner 1989, Inhoff 1984).

4.12 Concluding Observations

This section highlights some important points relative to interactive search drawn from the body of this chapter.

Part of the approach I describe to improve the result of an interactive text search I suggest that search return documents should be graphically presented as clusters that allow the user to dispose of irrelevant clusters of documents and thus speed the filtering of large return sets to more manageable sizes. This approach is supported by research conducted by Driver and Baylis (1989) suggesting that attention is directed to perceptual groups whose components are spatially dispersed and that in dynamic environments the spotlight metaphor is probably inappropriate.

This is one of a number of graphical approaches that can be used to expedite the interactive returns filtering process. Colour is discussed in this section and is another approach that can be used effectively to make targets pop-out of a field of potential targets as supported by the research of Carter (1982). Further to this, Carter & Carter (1981) proposed that "colour difference can be used as a tool for design and evaluation of visual displays, for construction of colour codes to optimize search time, and as a generalization of chromatic contrast in psychophysical research" (p.723).

Given the progression from very small region of high acuity to very large region of low acuity (see Section 4) the process of graphical presentation process can be optimised given this fact. This is because any interactive task that needs to visually attract the attention of a user to a specific area of the screen need not use high fidelity graphical events. This is because the region of the eye that is likely to receive the stimulus from any such event will most likely be the outer low fidelity region. However, it is obvious that the level of detail required when attention is gained will need to be tailored to the requirements of the task. For example, if a task requires reading then higher fidelity is required compared to a button pressing task where targets are larger and requiring less accuracy and acuity.

Generally speaking, if the low-level visual system and pre-attentive processes can be harnessed during visualization, attention might be more efficiently and effectively drawn to areas of potential interest in a display (see Section 4.3). Obviously, this cannot be accomplished in an ad-hoc fashion so the visual features assigned to different

data attributes must take advantage of the strengths of our visual system, must be well suited to the analysis needs of the viewer, and must not produce any visual interference effects that could mask information in a display.

Three key factors that critically affect visual attention are the manner in which space is used relative to object dispersion (size of display), how objects are grouped relative to other task relevant objects and distractors via common visual traits such as motion, proximity and colour, and lastly previous experience (see Section 4)

Chapter 5

Modeling Users

The development of better interactive search techniques requires knowledge and understanding of user interactive behaviour and the cognitive processes involved in similarity/dissimilarity recognition. With respect to interactive search many empirical studies have reported general patterns of information seeking behaviour (Choo, Detlor & Turnbull 2000). World Wide Web (WWW) usability methodologists such as Spool et al. (1999), Nielsen (2000) and Brinck et al. (2001) have drawn on a mix of case studies and empirical research in suggesting good design strategies for use during development, evaluation and specifically to identifying usability problems. Examples include research into the principles regarding the ratio of content to navigation structure on WWW pages (Nielsen 2000), the use of information scent to improve WWW site navigation (Pirolli & Card 1999), the reduction of cognitive overhead (Krug 2000) and how writing style and graphic design interact (Pirolli 2000, Brinck, Gergle & Wood 2001).

In the pursuit of modeling interactive tasks, such as; information search, the characterization of the user is a central requirement. The characterization of the user in the process of interactive search falls within the larger area of Information Retrieval (IR) system evaluation. The evaluation of systems, such as Google, Alta Vista, Excite and others, classically use recall, precision, varying forms of information theoretic and even approaches such as BookMaker (Powers 2003), to measure performance. The use of measures has often been debated (Lee 1987, Saracevic 1995, Yao, Wong & Butz 1999, Powers 2003) and still sees research into metrics and measures, however selection of what measures to use can be logically guided by the adage "horses for courses".

The evaluation of an IR system typically sees each document in a known document collection classified as relevant or non-relevant based on a set of queries. The known queries are executed using an IR system on the document collection and based on the number of relevant and non-relevant documents retrieved, recall and precision is determined. This is a systems view of relevance, with recall and precision directly related to the measure (it should be noted that these are system measures of likeness not subjective user measures of likeness) used to classify the documents against the queries entered. The whole process is *relative to the system* and *not the user*.

When a real user is introduced to the situation, they apply a level of subjectivity that makes the evaluation much more complicated. Relevance in this case is now user bound, as opposed to system bound, which is recognized as being not clearly defined (Mizzaro 1997, Saracevic 1997). This is not surprising given the discussions in Sections 2,3 & 4 that point to multiple theories, architectures and models that attempt to describe the different cognitive processes and the fact that they are often in conflict with each other. The modeling of user search processes suffers similar problems as seen in the conflicting results from the analysis of several key theories (Belkin, Oddy & Brooks 1982, Saracevic 1996). Clearly, the user and the understanding of their internal processes and preferences are critical to the improvement of interactive search.

5.1 Search, Similarity, Classification and Context

All textual search engines rely on some form of similarity, or likeness, when matching query-terms to appropriate documents; most return them as either ranked lists or groups of similar documents. Because of the problem of polysemy, classification of documents into ranked list representing contextual relevance, or into groups of similar topic is very much a subjectively bound task. Although text-search is the focus of this work similarity and classification/categorization are critical to the success of most human pursuits where any form of organisation is required.

Classification can be described as "the putting together of like things or the arranging of things according to likeness and unlikeness" (Maltby 1975, Richardson 1964, p.16, p.1). Chan (1994) interestingly describes classification more intricately as the process of "deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or

objects that have the property of characteristic in common into a class" (p.259). Most definitions of classification will contain the concepts of likeness and unlikeness.

The importance of the concept of similarity is captured by Richardson (1964) when suggesting that similarity/likeness "is the universal principle of the order of things... Likeness is so much part of the essence of all human thought, that literally there is no smallest part of the human mind which cannot be analysed into just this operation of distinguishing like and unlike and either holding to or rejecting. Likeness, in particular, is the foundation of that systematic thought carried to its ultimate which we call logic"(p.6).

Broadfield's (1946) view regarding similarity is more toward that of the relational aspect between things and suggests that likeness is not a characteristic of things. He also tempers the importance of likeness by describing it as only an indicator, stating "Resemblance is only a pointer, indicating the possibility that things might be more profoundly related" (p.6). This points to the idea that other information may be required in the process of classification and that it would be contextual by nature.

An important aspect of context is prior knowledge. As discussed in Section 2.1.3.4 prior knowledge probably contributes to an individual's representations of categories. For example, people not only know that birds have wings and that they can fly and build nests in trees, but also that birds build nests in trees because they can fly, and fly because they have wings. Many people, such evolutionists/scientists, believe that morphological features of birds such as wings are ultimately caused by the kind of DNA that birds possess (Rehder 2003*b*). In comparison, however, with the development of models that account for the effects of similarity and empirical observations, there has been relatively little development of formal models to account for the effects of such prior knowledge especially in terms of query phrase formulation, occurring before and during an information/text search event.

Regarding the importance of *context* in interactive search, the problem of polysemy is a critical issue because it affects any resultant classification for which context may illuminate the desirable. This is highlighted by Spiteri (2007) when she suggests that in the task of classification, similarity "assumes a shared or common understanding of the attributes or features that give a concept its identity" (p.2). This assumed common understanding raises the question: will different people, even people from

a similar context, understand a concept in the same manner? Two recent studies demonstrated that although participants agreed that two terms were similar, they did not agree why they were similar. Some participants said that authority control is a product of cataloging, while others that cataloging is a form of authority control (Spiteri 2004, Spiteri 2005). Although subtle, the problem highlighted by these studies is that people are likely to use different words or phrases to formulate a query because they have a different conceptualization of the problem.

5.2 Modeling Human Computer Interaction

There are often situations in interactive information retrieval in which process automation is appropriate, however the question as to what should really be automated is something that user modeling is attempting to address. Schneiderman and Macs (Shneiderman & Maes 1997) suggest that the decision as to what should be automated should be under the control of the people affected by the system. That is, only the human can recognise what is contextually pertinent. For example, despite systems becoming extremely powerful they are still only "aware" of a fraction of the user relevant contextual information and can only bring a small amount of the problem-solving process to bear on the situation (Hollan 1990). As this information is mostly internal to the user an understanding of the situation or state of problem-solving of a human can not be easily or rapidly shared (Suchman 1987).

This is not to say that the user is the only part of the interactive process that can do any amount of important work; to improve task realization (e.g. less time, cost and effort), knowledge of user interactive behaviour should be recognised and incorporated during the design and implementation of a system (Feyen, Liu, Chaffin, Jimmerson & Joseph 1999). Such processes are often very expensive and time consuming (Magrab 1997). This situation can be improved through user modeling to make predictions about human performance and mental workload for a given situation (interactive task) prior to starting any expensive developmental processes (Olson & Olson 1990).

Modeling human-computer interaction behaviour offers the ability to better leverage human abilities, such as the visual perception of patterns and recognition of context in complex decision making tasks (Scott, Lesh & Klau 2002) and hence realise improved

task outcomes. In the context of this work, user models are defined as models of human-computer interaction behaviour as opposed to mental models which are internal to the user and are used in interactions with the user's world. Models and modeling systems vary greatly, for example the WEST system, a coaching system for a game called "How the West was Won" (Burton & Brown 1982), represents an early pioneering effort to generally explore issues associated with user modeling. Alternately the model by Prabu et al. (2007) is a "cyclic model of information seeking in hyperlinked environments" used to study the relationships among perceived goal difficulty, goal success, and self-efficacy.

With an overarching objective of improving the realization of an interactive task, from the user's perspective, the importance of being able to model user behaviour is reflected in the immense amount of research into, and use of, user modeling in recent decades (Fischer 1999). Some of the better recognized models, both for psychological modeling and for product design and system evaluation, include Fitts' Law, GOMS, KLM-GOMS, CPM-GOMS and SNIF-ACT. These models are outlined in the following section.

5.2.0.1 User/Cognitive Models

Models are used to explain how some aspect of cognition is accomplished by a set of primitive computational processes. A model performs a specific cognitive task or class of tasks and produces behaviour that constitutes a set of predictions that can be compared to real world human performance data. Several different task domains have received considerable human modeling attention, such as problem solving, language comprehension and memory tasks. However the domain relative to this work is that of human-device interaction, specifically within the field of HCI in document search tasks.

Belonging to the field of cognitive psychology, cognitive modeling often uses computational techniques from the field of artificial intelligence and vice versa. Cognitive modeling focuses on functionality and computational completeness, and can be used to produce theories of human behaviour for a task and a computational system that performs the task.

Cognitive models are either symbolic, connectionist, or hybrid. A cognitive model is considered a symbolic cognitive model if it has the properties of a symbolic sys-

tem as described by Newell and Simon's (1972) Physical Symbol System Hypothesis (PSSH). A distinguishing feature of these types of systems is that they should be able to compose and interpret novel structures, including structures that denote executable processes. Connectionist models describe mental or behavioral phenomena as a series of interconnected networks of simple units and their interactions. Given that they are based on connections and units it is these that generally describe the difference between the different models of this type. For example, units in the network could represent neurons and the connections could represent synapses. Another model might make each unit in the network a word, and each connection an indication of semantic similarity.

Fitts' Law Fitts' law is a model of human psychomotor behaviour (Fitts 1954) that extends Shannon's (1949) channel capacity theorem in information theory. Shannon's channel capacity describes the effective information capacity of a communication channel. As a model of human movement Fitts' law attempts to predict the time taken to rapidly move to a target area, as a function of the distance to the target D and the size of the target W as a logarithmic function of the spatial relative error D/W.

$$MT = a + b\ log_2(2D/W + c)$$

where

MT	is the movement time.
a and b	are device dependent and empirically determined by fitting a straight line to measured data. "a" represents the start/stop time of the device and "b" the inherent speed of the device.
D	is the distance (or amplitude) of movement from start to target centre.
W	is the width of the target.
c	is a constant of 0, 0.5 or 1.

Relative to interactive user interfaces the model is normally used to describe point-and-click and drag-and-drop actions or other such actions where the user needs to position a mouse cursor over a screen target, such as a button, menu or other widget. The model is relatively strict in that:

- It applies only to movement in a single dimension and not to movement in two dimensions

- It describes simple motor response (e.g. human hand) and does not account for external influences such as software acceleration often applied to a mouse cursor
- It only applies to untrained movements, not movements where the user is practiced

GOMS The model Goals Operators Methods and Selection rules (GOMS) (Card et al. 1983) is an often used and modified approach to human computer interaction observation. The user's behaviour is modeled in terms of Goals, Operators, Methods and Selection rules, which are described below in more detail. GOMS characterises user interactions with a computer by elementary actions (these actions can be physical, cognitive or perceptual), which are used as a framework to study an interface.

GOMS techniques are particularly useful for modeling sequential operations be performed by an experienced user. Using GOMS, we can walk through the sequence and assign approximate time for each step, and calculate the cumulative performance time based on each step estimate.

Goals: Goals are what the user target achievement in enacting a task.

Operators: An operator is an action performed in service of a goal. Operators can be perceptual, cognitive, or motor acts, or a composite of these. Operators can change the user's internal mental state and/or physically change the state of the external environment.

Methods: These are sequences of operators and sub-goals needed to achieve a goal (e.g. move mouse to "File" menu scroll to "Save" and click on "Save").

Selection Rules: There are often situations where more than one method may be used to achieve a goal (e.g. alternate to the above save sequence one might choose to use the key press combination of "ctrl-s"). Thus, if there is more than one method available and known for a goal, then there is a need for selection rules to represent the user's knowledge of which method should be applied. This knowledge may come from a user's personal experience with the interface or from direct training.

There are several different GOMS variations that allow for different aspects of an interface to be accurately studied and predicted. For all of the variants, the definitions of the major concepts (Goals, Operators, Methods and Selection rules) are the same.

A major limiting factor of GOMS approaches is that they do not address user unpredictability such as that of a user's behaviour being affected by physiological and environmental factors. A second limiting factor, is that of the assumption of "experience" which makes GOMS inappropriate to the novice situation. The assumption of user knowledge/experience is error-prone as it implies the user will know what to do at any given point.

KLM-GOMS The Keystroke-Level Model (KLM) (Card, Moran & Newell 1980) is a version of GOMS used to predict task execution time from a specified design and task. In this model, execution time is estimated by listing the sequence of actions for a task and then summing the times of the individual operators. An action is defined as being at keystroke level if it is at a basic level such as pressing keys, moving the mouse, pressing buttons, and so on, as opposed to more complex actions like "log onto system".

The original KLM has six classes of operators: K for pressing a key, P for pointing to a location on screen with the mouse, H for moving hands to home position on the keyboard, M for mentally preparing to perform an action, and R for system response where the user waits for the system. For each operator, there is an estimate of execution time. Additionally, there is a set of heuristic rules to account for mental preparation time.

KLM is used to study human computer interaction and improve the usability of a interactive interfaces by analysing use performance in task realization. This allows the production of estimates of how long it takes to achieve certain tasks and what might be done to improve efficiency. KLM aims to answer the following questions:

- Can users complete the tasks/goals of the interactive search tool?
- Can users complete the tasks within a reasonable amount of time and with minimal errors?
- Can novice users learn how to complete tasks within a reasonable time?

KLM is an attractive model for researchers and interface designers as it can be used to quickly assess and compare hypotheses, designs or systems. One key limiting factor

of the approach is that the model is designed to estimate the execution time for an *expert user* familiar with the specific task who is typically faster than a user unfamiliar with the task. Another characteristics of KLM to note is that it does not account for mistakes automatically, consequently the analyst must create separate models for error sequences and perform their own sensitivity analysis.

For a description of the different operations and their expected time ranges see Kieras (1993).

CPM-GOMS CPM-GOMS stands for: Cognitive, Perceptual, and Motor and the project planning technique Critical Path Method (from which it borrows some elements) (John 1988, John 1990). It is a modeling method that combines the task decomposition of a GOMS analysis with a model of human resource usage at the level of cognitive, perceptual, and motor operations. Unlike other GOMS techniques, it recognizes that as many operations as possible will happen at any given time and that these are only governed by cognitive, perceptual, and motor processes constraints. Models of user interaction are developed using PERT (Program/Project Evaluation and Review Technique) charts from which a critical path is used to determine execution time. The interleaving and visualization of the PERT chart sequences allows for the construction of arbitrarily long sequences of behaviour. CPM-GOMS models have made accurate predictions about skilled user behaviour in routine tasks, but developing such models is tedious and error-prone (Freed, John, Matessa, Remington & Vera 2002).

SNIF-ACT Scent-based Navigation and Information Foraging in the ACT architecture is an architecture that is very relevant to the subject of this work, specifically the optimisation of text search, as it simulates user performance in unfamiliar information-seeking tasks on the World Wide Web (WWW) (Pirolli & Fu 2003). Specifically, SNIF-ACT aims to characterize user hyper link choices in WWW tasks by supplying a mechanistic account of information foraging that takes into account cognitive and perceptual limitations of the user to predict what actions a user might take.

SNIF-ACT is based on the integration of Information Foraging Theory (Pirolli & Card 1999) and the theory behind ACT-R (discussed following section). A key concept in this architecture is that of *information scent* which characterizes how users

evaluate the utility of hypermedia actions (i.e. clicking on links). It also uses a user-trace methodology (Pirolli, Fu, Reeder & K. 2002, Pirolli & Fu 2003) for studying and analysing the psychology of users performing ecologically valid tasks when interacting with the WWW. A user-trace is a record of all significant states and events during interaction based on eye tracking data, application-level logs, and think-aloud protocols.

5.2.0.2 Cognitive Architectures/Models

There is a close relationship between cognitive architectures and models in that cognitive architectures are the "overall, essential structure and process of a broadly-scoped domain-generic computational *cognitive model*, used for a broad, multiple-level, multiple-domain analysis of cognition and behaviour" (Sun 2004, p.1).

Like cognitive models, cognitive architectures can be symbolic, connectionist, or hybrid and are normally based on a set of generic rules. They provide a framework for more detailed modeling of cognitive phenomena and allow the analysis of cognition at the computational level. They have a fixed set of computational mechanisms and resources that are suggested to underlie many aspects of human cognition. However, as they do not correspond to the architectures of modern computers, such as requiring a higher degree of parallelism, they must first be emulated on computers before cognitive models can be built within them for specific tasks.

Cognitive architectures not only attempt to model behaviour, but also structural properties of the modeled systems involved. They are suggested to be essential to the development of understanding of the mind (Anderson & Lebire 1998, Newell 1990, Sun 2002). Some of the better-recognised architectures are ACT-R, CHREST, CLARION, EPIC and Soar.

ACT-R ACT-R (Adaptive Control of Thought - Rational) is a computational theory of human cognition proposed by Anderson & Lebiere (1998) that incorporates both declarative and procedural knowledge. An important assumption of ACT-R is that human knowledge can be divided into two irreducible kinds of representations: declarative and procedural. These representations are used in the simulation and understanding of how people organize knowledge and produce intelligent behaviour. Declarative and

procedural knowledge form production systems in which procedural rules act on declarative chunks. These chunks are comprised of slots containing information. Production rules are executed when they match the information in these chunk slots of which there is only ever one match. One limitation of ACT-R is that it does not account for scenarios with interleaved tasks.

CHREST CHREST (CHunk Hierarchy and REtrieval STructures) is a symbolic cognitive architecture that models human perception, learning, memory, and problem solving. It is based on the concepts of limited attention, limited short-term memories, and chunking and focuses on tracking cognitive limitations such as short-term memory and processing speed. Examples of alternative approaches can be seen in systems such as Soar and ACT-R that use productions for representing knowledge (Gobet 2001).

CHREST combines low-level mechanisms of cognition, such as monitoring of short-term memory, with high-level mechanisms, such as application of strategies. It is comprised of perception facilities for interacting with the external world, short-term memory stores (in particular, visual and verbal memory stores), a long-term memory store, and associated mechanisms for problem solving. An important aspect of CHREST is that short-term memory contains references to chunks held in long-term memory, which are recognised by the discrimination network using information acquired by the perception system (Gobet 1993, Gobet & Simon 1996, Gobet 2001).

CHREST has been applied to modeling of learning using large corpora of stimuli representative of a domain, such as child-directed speech for the simulation of children's development of language.

CLARION CLARION (Connectionist Learning with Adaptive Rule Induction ONline) is a cognitive, modular architecture that consists of a number of functional subsystems that recognises explicit and implicit representations via separate components (Sun, Merrill & Perterson 1998, Sun, Merrill & Perterson 2001, Sun 2006). The subsystems interact with each other constantly working closely together in order to accomplish cognitive processing. They can be described as follows:

action-centered subsystem controls actions.

non-action-centered subsystem maintains general knowledge.

motivational subsystem provides underlying motivations for perception, action, and cognition.

meta-cognitive subsystem monitors, directs, and modifies the operations of all the other subsystems.

CLARION is appropriate for interactive user modeling as demonstrated by it being successfully used to simulate tasks in cognitive psychology and social psychology, and to implement intelligent systems in artificial intelligence applications. Other applications it has been used for include simulation of creativity and addressing the computational basis of consciousness and artificial consciousness (Sun et al. 1998, Sun et al. 2001, Sun 2006).

EPIC Executive-Process Interactive Control (EPIC) is a cognitive architecture that aims to provide a detailed account of human perceptual and motor operations. It has been especially useful for building cognitive models in the domain of Human computer interaction (Kieras & Meyer 1994, Kieras & Meyer 1995). EPIC is generally used as a system for exploring human performance limitations that determine the effects of a particular interface design, both at low levels of specific interaction techniques, and at high levels of systems that support complex task performance in multimodal time-stressed domains. Because HCI's focus is human performance, EPIC is a good architecture to use as it allows for the analysis and comparison of interface designs by modeling human performance in different tasks (Kieras & Meyer 1994, Kieras & Meyer 1995). Evidence of EPIC's usefulness in representing interactive tasks can be seen in the fact that some of its features, specifically its perceptual/motor capabilities, have been incorporated into ACT-R, CLARION, and other cognitive architectures.

Soar State, Operator And Result (Soar or SOAR) is a symbolic cognitive architecture that is primarily used as a computational model for Artificial Intelligence research (Laird, Newell & Rosenbloom 1987). Although it is not directly used in the field of HCI it is a pertinent architecture to discuss as it is used to model cognition through the development of general artificial intelligence and thus can be used to model and describe human behaviour under certain conditions via the AI analogue.

Soar uses explicit production rules to govern its behaviour (like "if... then..."). Problem solving can be described as a search within a problem space for a goal state that is implemented by searching for the states that bring the system gradually closer to its goal). Each move in the problem space consists of a decision cycle which has an *elaboration phase*, during which pieces of knowledge pertinent to the problem, are brought in to working memory, and a *decision procedure* that uses weightings found in previous phases and assigns preferences to ultimately make a decision. If the decision procedure does not result in a course of action, Soar may use different strategies, known as weak methods to solve the impasse. These methods are appropriate to situations in which knowledge is not abundant. When a solution is found by one of these methods, Soar uses a learning technique called *chunking* to transform the course of action taken into a new rule. The new rule can then be applied whenever Soar encounters the situation again.

5.3 Identifying User Originating Thresholds

HCI has seen a broad spectrum of work into the characterisation of search strategies like browsing (e.g.,Brown & Sellen (2001), Catledge & Pitkow (1995) and Ford, Miller & Moss (2002, 2003)) and characterisation of user habits when using search engines (e.g., Ford, Miller & Moss (2002, 2003), Moukdad & Large (2001), Ozmutlu, Spink & Ozmutlu (2003), Spink, Wolfram, Jansen & Saracevic (2001), SPink & Oamutlu (2002), Su (2003), White, Rose & Ruthven (White, Jose & Ruthven 2003) and Xie citeyearxie03). In the pursuit of modeling user interactive behaviour or characterising user search preferences Transaction Log Analysis (TLA) is by far the most common technique used.

Transaction log analysis uses transaction logs to discern attributes of interactive processes between two entities. Penniman and Dominick (1980) describe the different things log data can be used for:

1. a diagnostic aid
2. group or individual user evaluation, e.g., analysis of user performance
3. system protection, e.g., diagnosis of attempts at unauthorized system access

4. system evaluation

HCI sees TLA being used to address item 1 above in that logs in this case are used to collect data about interactive events and state information (such as type, content or time data) to process and draw conclusions about search processes such as a searcher's actions, the interaction between the user and the system, and the searcher's evaluation of the results. TLA is suggested to be a grounded theory approach (Glaser & Strauss 1967) in that real-world user search characteristics are assessed to identify facts and trends that characterise interactions between searchers and the system. This can be refuted with the observation that a log can not reflect the topic or meaning.

A benefit of using TLA can be seen in the extensive nature of the WWW and private databases, with respect to servers (all with logs) and search engines (all with logs), of specifically textual as well as other information formats. This means that there is an abundance of log information to mine for interactive trends. This has translated into TLA's being used to study many different aspects of interactive search as can be seen in Spink and Jansen's (2004*b*) extensive bibliography of studies that use TLA in the Web search domain.

TLA has seen its fair deal of criticism as a research methodology such as that by Blecic et al. (1998), Hancock-Beaulieu et al. (1998), Jansen & Pooch (2001), Jansen (2006) and Phippen et al. (2004). Blecic et al. suggest that TLA only views the transaction trail and does not provide an overall picture of the user or their behaviour. Jansen & Pooch (2001) when talking about Web searching studies suggest that TLA typically lacks the context and relevance judgments of the user. Jansen (2006) points to the fact that logs are primarily data collected server-side that can not capture events such as *cut or paste*, *clicking the back button* or *selecting print*. Phippen et al. (2004) point to this technique as falling short of delivering the richness of data required for effective evaluations of other approaches. The general criticism as demonstrated here is that transaction logs do not deliver critical information needed for interactive data assessment such as user experience and topic knowledge, and cannot record the user's underlying information need which is relative by nature and contextually bound. In this vein, Kurth (1993) points out that transaction logs can only deliver data about the user's actions, not their perceptions, emotions and background skills.

In defence of TLA Jansen (2006) suggests, and logically so, that many of TLA's

suggested problems/weaknesses are not just confined to TLA and that they are also issues of many empirical methodologies. He goes on further, pointing to technological and procedural advancements that have moved to address many of the issues. Some examples can be seen in Hancock-Beaulieu et al.'s (1998) transaction logging software and online questionnaire and *Tracker* a similar research package by Choo et al. (1998, 2000) designed to elicit the user's information needs and information seeking preferences relative to their usage.

Researchers have used transaction logs to analyse a variety of Web and Database search systems, including Fireball (Holscher & Strube 2000), AltaVista (Jansen, Spink & Pederson 2005, Silverstein, Henzinger, Marais & Moricz 1999), Excite (Ross & Wolfram 2000, Spink & Jansen 2004*b*), Fast (Spink & Jansen 2004*b*), OPACs (Jones, Cunningham & McNab 1998), THOMAS (Croft, Cook & Wilder 1995) and Yandex (Buzikashvili 2000). In addition, Web search engine and Web intelligence companies use TLA to identify usage and market trends, and the effects of system changes.

5.4 User Queries and Web Search Trends

Research to identify searcher habitual characteristics such as the number of words used in the average search has seen only a small number of published works of note. This work has been conducted in the last fifteen years and despite the small amount of research it has been good quality and very informative. The key statistic to come from this research is that of a value around "2", which is the average number of phrases used in an average search on the Web, based on data from search engine logs.

In their work on the Excite logs, Jansen et al. (2000) report that Web queries were generally very short and that most users in 1997 only entered around 2.8 queries per search session, with each query having around 2.21 terms. The key reported approximate distribution of terms per query where as follows:

- $< 33\%$ of queries had *only* one term.
- $< 66\%$ of queries had *only* one or two terms.
- $< 80\%$ of queries had *only* one, two or three terms.
- approximately 4% of queries had more than six terms.

In a similar study that compared national/cultural differences in search Spink et al. (2002) reported that the mean length of Excite queries increased from May 1996 to June 1999. However, this was caveated with an observable difference between the number of terms used by US and UK searchers and the number used by Europeans searchers. Overall in 1996 the mean query length for US, UK, and European users was 1.5 terms. However, in 1999 the number of terms US and UK searchers used was 2.6, an increase of more than 1 term, as opposed to that of European users of 1.9 terms, an increase of less than a half a term. This suggests that English language queries increased in length at a greater rate than European language queries over the same period.

Subsequent to this research Spink and Jansen (2004*a*) reported that the average number of terms per Excite query had increased slightly to 2.6 by 2001 and fallen back to 2.4 by 2003. From this they concluded that general Web queries had remained relatively short with searches containing 2-3 term per query and 2-3 queries per search.

Spink, et al., (2000) found that most Web searchers only used one query, or in other words they seemed to not need to reformulate their query. The average session, ignoring identical queries, was comprised of approximately 1.6 queries. The critical value reported was that approximately 66% of user's submitted *only* one query. Interestingly, and in line with the observed increase in terms per query and queries per search, in 2002 Spink et al. (2002) reported that in 2001 approximately 44% of Exite users conducted a session with more than one query reformulation while 25% percent of users reformulated more than twice.

In regards to user visual search habits in 1999 Xu (1999) observed that from 1996 to 1999 approximately 70% of the time searchers only viewed the top ten results. In his presentation he suggested that the average users viewed 2.35 pages of results (where one page equals ten results) and that over 50% of the users did not access results beyond the first page. Supporting this Spink et al. (2002) found that more than 75% of users did not view more than two pages of results. This general trend of users not visually inspecting very many documents is supported by Jansen & Spink's (2003) research that suggested that by 2003 the average user only viewed about five Web documents per query.

From a more commercial view point much of this academic research is paralleled

and supported by commercial research such as that by the major Web search engines like Google (see $http://www.google.com/press/zeitgeist.html$) and Yahoo (see $http://buzz.yahoo.com/buzzlog/?fr=fp-buzz-morebuzz$). One of the better known set of commercial results is that by OneStat.com (OneStat.com February 2, 2004) a provider of real-time intelligence web analytics. In 2004 they reported that most people use 2 word phrases in search engines. Of all the search phrases world wide, 29.22 percent of people use 2 word phrases and 24.76 percent use 1 word phrase. The OneStat research is based on a sample of 2 million visitors, made up of 20,000 visitors in 100 countries each day and concluded that the 7 most used number of word phrases in searching the web are:

1. 2 word phrases 29.22%
2. 1 word phrase 24.76%
3. 3 word phrases 24.33%
4. 4 word phrases 12.34%
5. 5 word phrases 5.43%
6. 6 word phrases 2.21%
7. 7 word phrases 0.94%

It is fairly clear that searchers use few phrases per search, few terms per phrase, do not often reformulate and don't seem to spend much time investigating the return set. Section 5.6 discusses the problems with these types of Web log analysis and the observations that might be drawn from them. This is not to say that TLAs are not useful or wrong as they are valid statistics and can be used for comparison and supporting argument in the discussion of other similar research and results.

5.5 Web TLA Flaws

Despite the broad variety of research across the field of HCI there seems to be no research trying to quantify thresholds internal to the user, such as the number of words normally used to describe a textual object in a context free manner. In the pursuit of

identifying these general user originating thresholds in interactive search, TLA presents as a potentially appropriate technique. Although it has some generally good traits as an elicitation technique, it needs to be applied in a particular manner to allow for some of its weaknesses.

TLA's most attractive feature is that transaction logs are relatively simple to collect and mine for statistics such as thresholds internal to the user and that the logs can be generated by users in their normal interactive context. This is in contrast to strict laboratory style experiments in which the interactive user would be placed in a unnatural and contrived environment that although highly controlled would result in any data gathered representing a non-realistic situation.

In the field of HCI, TLA on Web search engine logs (see Sections 5.3 & 5.4) has been extensively used and can be relied on to deliver user interactive data like the number of words used per *Web search query*. However the results of this type of research are very specific to Web search and do not practically expand to cover heuristics such as the number of words people might normally use to *describe* a textual object. This is due to the following problems:

1. Unknown user characteristics
2. Search engine specific results
3. Unknown information requirements of user
4. Un-identifiable task outcomes
5. Polysemy, Homonymy & term treatments

Unknown user characteristics

To characterize all interactive searchers under one overarching heading such as "human", would be a mistake which is essentially what occurs in Web log analysis. It goes without saying that different categories of people might have different tendencies given different situations and that any research involving humans should attempt to recognise any significant groups in a population. In Web log analysis it is normally impossible to identify such user characteristics as age, sex and search experience from the logged data because it simply does not exist.

Search-engine specific results

The main problem in this situation is that the search engine directly affects the success of any text search task through the mechanisms that deliver and order a set of results for the user to select from. The simple example in Appendix 10.1 demonstrates that different search engines deliver different sets and orderings, and that result lists are directly impacted by internal heuristics such as term/phrase weighting schemes, stopping techniques and stemming techniques. At a research level, the effects of such mechanisms are difficult to predict or cater for making the search engine itself a variable that needs strict controlling or outright removal from the process.

Unknown information requirements of user

When searching the Web, some queries may be more successful than others leading to problems in the comparison of queries. For example, if a highly publicized Web site has received a relatively high hit count then it is more likely to appear at the top of a results list and thus more likely to be found/selected early in the search process. Alternately, if the same Web site does not have the positive ranking traits of the form, because it has not been hit as much or for any other reason, it would be less likely to appear high in the ranked list and thus less likely to be found early in the search process. In the former situation the searcher may not need to do any more than click the top entry on the returns list for rapid success, in the later the searcher may need to use more terms or alter their query to better target their information requirement and realise the appropriate Web site, and thus slow realisation of information need. Given that in Web search the information sought by the user is not identifiable other than by input words and that the success of any one search is contextually bound, the comparison between searches for anything other than very broad indicative processes might be statistically un-sound. This indicates a need to generate statistics against a known information requirement or to exclude any information requirement from the research process.

Another problem related to not knowing the user's information requirements corresponds to the problem of not knowing the targeted context of a user's query. That is, when a Web search engine logs the query terms used it cannot determine and log the targeted context of the query such as whether the user is looking for textual information, downloadable media (audio, video, picture, software, ...) or services. For example if a user inputs the query "animal cruelty" (note the average two term query

length) there is no indicator as to what the user is looking for. That is, are they looking for textual information about the deliberate harming of animals, the audio track name "Animal Cruelty" by the musical group Silent Assassin, or some animal rights video. In this situation with a two word query like this the popularity of the music by Silent Assassin might inundate the top of any returns list because of a short term popularity surge resulting in millions of downloads. In short this means Web log statistics can only be strictly interpreted as word frequencies and any analysis thereof should be careful in drawing conclusions about what these frequencies mean regarding what is actually being sought.

Un-identifiable search task outcomes

Research by Mark, Gonzalez & Harris (2005) characterises information workers (people highly likely to conduct textual searches) as being very frequent context switchers. Their research is supported by specific research into context switching in software developers (Perlow 1999), context switching in information workers (Czerwinski, Horvitz & Wilhite 2004, Hudson, Christensen, Kellogg & Erickson 2002, O'Conaill & Frohlich 1995, Rouncefield, Hughes, Rodden & Viller 1994) and that of general office worker switching (Ttard 1999) all of which support the suggestion that the average time spent on one task is 3 minutes with a range of 1 second to 15 minutes. If researchers are looking to make conclusions based on search task completion times this research flags problems as it is suggests that a search task has a good probability of being interrupted before completion and that is assuming it is completed at all.

This problem with context switching and not knowing if or when a task is completed is compounded by a problem alluded to in the above section ("Unknown user information requirements"), and that is the quality of the search task outcome. Given the mass of information available on the Web and in searchable databases there is a real chance of getting a return set that presents documents that only partially address a user's targeted information need. Given user restrictions like patience, available time, importance of perfect data, the user may make a less than optimal selection. This situation is not evident in a transaction log and so allowances must be made in my research approach.

Polysemy & Homonymy

Polysemy describes the situation in which one word or phrase may potentially have

several meanings. Homonymy describes the situation where one concept can be represented by more than one term or phrase. To present a description of a textual object to a user that includes two or more terms that have the same meaning would be a waste of limited description area. Given the user is likely to attend and process a small screen area and/or set of description terms (see Section 3.1, 3.3 & 4.12) the topic of a document should be communicated with a key and set of terms each of which describe a different aspect of the text to present as many pertinent aspects of the textual content as possible.

In Web log analysis phrase characteristics often include stop words which are problematic when trying to identify heuristics like the number and type of descriptive terms the average interactive searcher might use to describe or identify a document. Stop words basically represent closed class words and normally represent the top 40%-50% of the words that occur in a corpus (e.g., 'the', 'it', 'who', 'what', ...). These are terms that occur in varying quantities and ratios (generally dependent on the writing style of the text), and that are not descriptive of a text's topic by themselves. This is compared to other words that tend to occur much less frequently in a text and corpus that are truly descriptive of the document's topic. The inclusion of stop words in descriptions and ranking calculations may affect the results of any presented ranked list and thus the user's selection.

It is practically impossible to identify the techniques used by search companies to manage these situations, and others not identified here, for commercially obvious reasons. This makes it impossible to allow for these processes and any of their effects, thus the indexing engine and search engine should be removed from any research scenario.

5.6 TLA to Nwords

This chapter has demonstrated the applicability and usability of TLA and the manner in which this might be implemented to identify characteristic user preferences/thresholds in describing visual textual objects. It discussed and highlighted some of the characteristics of general Web search and usage statistics obtained using TLA, which are used in the following chapter to compare and contrast against. Finally, several flaws with the use of search engine TLA were identified that should be systematically addressed if any experiment to identify general textual searcher characteristics is to be sound.

In the following chapter the Nwords experiments are introduced with description of how they address the identified problems of TLA.

Chapter 6

Nwords

Selected research, results and proposed techniques contained within the following two chapters have been accepted for publishing in the following peer reviewed publications:

> Darius Pfitzner, Kenneth Treharne & David M. W. Powers (in press, accepted May 2008), "User Keyword Preference: the Nwords and Rwords Experiments", International Journal of Internet Protocol Technology: Special Issue on Intelligent Internet-based Systems: Emerging Technologies and Programming Techniques.

Some of the statistics search engines use today, accurately model human search behavior but seem to have had some negative effects. These can be seen in the situation that sees search engines dictating the topology of the Web and inappropriately biasing return sets as was demonstrated by Cho and Roy (2004) and supported by user modeling experiments such as those of Klockner, Wirschum & Jameson (2004) and O'Brien, M. and Keane (2007). This is seen through the over-promoting of popular pages and the user's tendency to start return set assessment by clicking first at the top of the list and moving down the list sequentially. This behavior is negatively self reinforcing if the search engines then give that first document a higher status in future retrievals and thus a greater likelihood of being promoted to the top of the list again and so on.

In the context of document search, the value of textual language is self-evident for searching natural language documents and is a familiar and commonly used medium for information communication. It is also suggested that if a return list is of low relevance

users may switch to a more complex assessment behavior (O'Brien & Keane 2007). This complex behavior can be described as an attempt to introduce subjectivity. Searching is a context bound task and return lists from most generally used search engines do not account for word polysemy and other such linguistically confounding traits, and thus ignore an important level of context. Thus combined with other biases brought on by the linear appraisal and first-click behaviors of users we see support for the use of alternate mechanisms, such as those drawing on the powerful visual pattern spotting ability of the human.

In the quest for user models, there has been little decontextualized research into user cognitive limits and preferences relative to the number of words a user might use to describe a document. By decontextualized I mean research that is **not** reliant on output from a specific search engine (return sets, logs and other similar outputs) which have embodied predefined heuristics that may skew results, or research that is **not** conducted in the context of wanting to evaluate a specific interface (e.g. visualization and button configuration), technique (e.g. clustering, ranking or instructional approach) or tool, or simply research that is **not** conducted in a controlled environment (laboratory) but rather in an environment the participant is likely to be comfortable and familiar with.

6.1 The Two Research Problems

Given the "ultimate search system" will present information in a graphically clustered format, the two main problems are:

1. What document traits are used to form the documents in topically similar clusters, and

2. What and how many words should be used for the clusters or dimensions labels the user will use to make context judgments against.

As has already been discussed words are the primary document attributes/traits to transmit relevant contextual information to the user. The processing overheads realized by the current and appropriate clustering algorithms range between $O(n^2)$ to $O(n^3)$. Given the size of the average query returns list, the number of words in each document and the complexity of the clustering task this would result in unrealistic time/processing

overhead. However, given the nature of a document is such that only a small proportion of the words contained within it actually define the topic, the processing overhead can be reduced by several orders of magnitude by simply reducing n across the total set.

Given the frequency characteristics of words, as described by Zipf's law, the total number of words in each document to be processed can be substantially reduced by simply throwing away the most frequent words (e.g. 'the', 'of', 'and', 'to', 'a', 'in', and 'that'). These words are structural words that have low to no topical relevance to the document and so by not processing these words a smaller and more concise list of mostly nouns, verbs and adjectives is realized for easier processing.

This still leaves the problem of exactly 'how many' and 'which words' should be kept. The simple answer is those words the user thinks are most descriptive of the document given it is their opinion driving any interactive selection process.

Traditionally, the identification of topically descriptive words has seen the use of weighting schemes like that of TFIDF used to rank the words of each document given their relative document and corpus frequencies. However, because these methods relying on word frequencies the user's opinion is ignored which begs the question of how appropriate any statistical word frequency calculation can be at identifying key topical words of a document in any process the success of which is reliant on user context and intention.

Before any automated process can be designed to model user word preference, an understanding of those words the average user employs in the description of a document is needed. Relative to the task of document search and the proposed design of the "ultimate search system" there are two situations that can be used to observe user word preferences. The first is that of keyword selection tasks, such as those used for technical publications, which require the input of keywords by users to describe a document. The second is that of the search engine query task that sees a user describing a document via a short set of query words. Although different tasks, they both elicit from users condensed lists of descriptive words for individual documents.

At this point, the following questions need to be answered:

1. how many words do users employ in searching for a document,

2. how many words are used to describe a document topic/category (to optimize cluster descriptions)

3. how well does TFIDF correspond to user preference relative to word ranking (does TFIDF rank words similarly to humans?)

To answer these questions, Nwords was implemented and deployed to help identify the number and characteristics of the words people use to *describe* and *search* for documents, and how closely participant word rankings for given word lists agree with automatically generated TFIDF rankings of the same lists.

The primary objective of this research is to quantify the number of words a broad spectrum of participants use to describe different blocks of text and hence the appropriate number of words/chunks/dimensions needed to describe individual documents and clusters of documents, and to manage the impact of process intensive clustering activities. A secondary objective is to enhance understanding of choices a user makes in selecting keywords or phrases to describe or search for a document. To do this Nwords is comprised of four different experiments presented in the form of surveys using a common look and feel Web interface (for experiment/survey descriptions see Section 6.2).

This research will enhance the document search process by improving the quality of data users filter, will reduce search time and reduce machine-processing overheads. It will also provide fundamental insights into the way humans summarize and compress information.

The Nwords research is also supported by two other studies, the Rwords and InFields studies that are presented and discussed in Chapter 7. Rwords was used to identify the TFIDF formula that most approximates participant judgment by testing participant preference for the orderings produced by five commonly known variants of the TFIDF calculation. This was needed because Nwords requires the calculation of TFIDF weightings to present participants with short lists of key words of a document. InFields was used to test for any effects variations of a text input field size and the input mechanisms themselves might have on the number of words used in certain tasks. Information of this nature was required as some anomalies were noted during the analysis of Nwords data that might have been caused by different word input field sizes and mechanisms.

6.1.1 My Hypothesis

My hypothesis regarding the terms interactive searchers might use to describe or search for a document is as follows:

1. Because the popular TFIDF-like weighting schemes are based on frequency statistics and not an appropriate user model or reliably identified general user tendencies, they will produce ranked list of words for documents the heads of which do not match those a user might produce for the same documents. Thus the types of words users use to describe a document will be different than those produced by the commonly used automated processes.

2. Given researched cognitive limits such as those represented by the magic numbers 7 ± 2 or 4 ± 1 (see Section 3.1.1) and their associated chunks of information, users will have a preference for document descriptions of between 1 and 9 characterizing words (chunks). Within this range the tendency is suggested to likely be lower, given the research supporting Cowen's number 4 ± 1 (see Section 3.1.3), and given the human bias toward energy conservation in activities like search, as demonstrated by O'Brien and Keane (2007).

   The energy conservation tendency indicates the user will tend to use as few words as possible to describe a document. Related to this bias is the tendency of most users to select the first member of a search returns list without any real inspection of data presented. After this initial selection they, in a similar manner, sequentially select down the list until reaching some threshold at which they alter their search technique to a more energy consuming approach. These more energy consumptive approaches see the user surveying in more depth the associated snippets of each entry before making a selection.

   The quantity of terms used will normally fall between 1 and a number within Cowen's limit of 4 ± 1 (see Section 3.1.3). Given the tendency of energy conservation in selection activities users will prefer to use less words (e.g. less words equals less effort and thus less energy expended) than their maximal potential. As words are equatable to *chunks* the amount used will tend to be less than some cognitive limitation like Cowen's 4 ± 1 or Miller's 7 ± 2 which is likely to be Cowen's given the supporting research behind Cowen's proposal. So the user's tendency will be

to use between at least one word and an amount matching their relative cognitive limit of $4 + 1$.

Further support for a quantity of between 1 and 5 can be seen in Web search statistics Section 5.4 that indicate that the average Web search phrase was constructed of between 1 and 5 words. This is valid support despite Section 5.5 describing flaws in the Web TLA techniques and conclusions because the statistics involved can, at a very general level, be used as indicative generic usage and thus support in kind for the hypothesis.

TFIDF (Term Frequency/Inverse Document Frequency) and its variants are often used as a way of quantifying the raw frequency of a term inside a particular document. Techniques like this are used in an attempt to automatically weight words according to how important they are in characterizing a document, however to date their cognitive relevance remains unexplored. Given part 2 of my thesis suggested TFIDF and similar approaches realize ranked list of documents the heads of which will not match those produced by users the first target of Nwords is to demonstrate that these different lists do not match.

To test part 2 of my thesis Nwords needs to demonstrate that users normally use between 1 and 9 words to describe an average sized block of text with the tendency to be toward the lower of this range. If this question can be answered with an acceptable level of accuracy, documents can be clustered across a limited set of key words or concepts using the identified quantities as indicators to the number of dimensions to be acknowledged in the process. This will minimize the overheads in the clustering process, by limiting the dimensions clustered on, while at the same time creating clusters whose qualities (a limited set of descriptive words for both cluster and member documents) allow the user to make appropriate contextual filtering decisions against a cognitively optimally sized descriptor (n chunks).

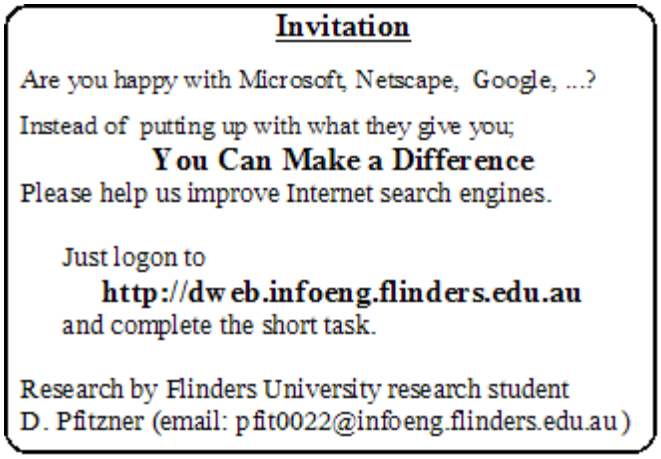
6.1.2 Participant Profile

The participant pool was drawn from several populations likely to participate in textual searches on a regular basis. They were recruited on an ad hoc basis using a small business card size invitation (see Figure 6.1) to realize a total pool size of 246 partici-

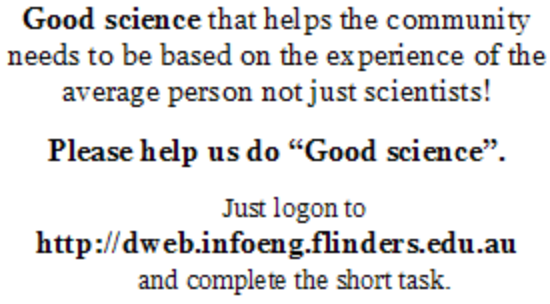
pants. The distribution of participants from the different population is outlined in the following table:

- 152 × undergraduate students (mixed humanities and science)
- 8 × academic staff (Informatics)
- 6 × administration staff (Informatics)
- 82 × HCSNet conference participants (mixed informatics/computer science, psychology, linguistics and music PhD students and academics, and some computing personnel)

However, given the above profile, with all participants living in Australia and being for the most part at English-medium tertiary education level, it is reasonable to assume that most participants would have had an adequate level of English language knowledge.



(a) Invitation Front



(b) Invitation Back

Fig. 6.1: Nwords participant invitation card

The sex and age distributions of the participants are described in Table 6.1 following.

Before Filtering								
	Survey 1		Survey 2		Survey 3		Survey 4	
Sex	Female	Male	Female	Male	Female	Male	Female	Male
-18	0	2	0	0	2	0	1	0
18-24	10	11	11	13	14	22	20	13
24-30	11	4	8	2	4	3	5	8
30-45	8	7	9	7	7	8	6	13
45+	6	0	1	2	2	1	3	2
Totals	35	24	29	24	29	34	35	36
Grand Totals	128	118						
Total Participants	246							

Table 6.1: Participant sex statistics before filtering

Due to errors noted in the results log some of the records had to be removed leading to the sex and age statistics described in Table 6.2. The reasoning and techniques used in the removal of the results in error are outline in Section 6.3.

After Filtering								
	Survey 1		Survey 2		Survey 3		Survey 4	
Sex	Female	Male	Female	Male	Female	Male	Female	Male
-18		2	0	0	2	0	1	0
18-24	10	11	11	12	14	21	20	13
24-30	11	4	8	2	2	2	5	8
30-45	8	6	5	7	7	8	5	12
45+	6	0	1	2	2	1	3	2
Totals	35	23	25	23	27	32	34	35
Grand Totals	121	113						
Total Participants	234							

Table 6.2: Participant sex statistics after filtering

6.1.3 Survey Delivery and Result Management

To emulate as closely and as reasonably possible the normal interactive environment of the user when conducting an interactive search, a Web-like interface was used to deliver the survey. In this manner the participant is likely to do the survey sitting in at least a familiar environment if not the actual environment they would normally conduct a search in (e.g. sitting in their chair in front of their computer), thus reducing any effects of an unfamiliar environment which is so often unaccounted for in *strict* laboratory experiments. For example, the undergraduate student would have sat either at a university terminal or at their computer at home and the administration staff, and academics sat at their work machines. It is uncertain where the HCSNet participants might have done the survey, however it is reasonable to assume that they would have used either their home or work machine.

The four different types of survey (see following Section 6.2) were served to the participant's machine by an Apache Web server using a mixture of Perl and Javascript to generate the dynamic content of the pages, and to log the results of the survey. The results were recorded directly to a log within the Apache directory and were also emailed to myself for redundancy purposes. Each result is marked-up using *.csv* format for ease of post processing (programmatically, spreadsheet or Matlab). The information recorded in a result depends on which survey the participant was given mainly because surveys 1, 2 & 3 generate different data compared to 4 (for survey differences see Section 6.2). Following are two examples that describe the two different types of survey result recorded.

Survey 1, 2 & 3 log result example

Sat Aug 5 11:21:44 CST 2006
, 18-24, Female, 1, 0, duc_manual_processed/2002_processed/d070_processed/fbis4-42178_processed/fbis4-42178.txt, Erich Honecker, 6, Honeckers Death, 7, Becker, 4, Court Case, 5, Santiago Chile, 3, #, 3, Honecker Death Court Case, 11:28:27-11:31:36+11:31:36-11:31:51+11:31:52-11:32:6

Survey 4 log result example

Fri Mar 31 20:50:58 CST 2006

, 45+, Female, 4, 0, duc_manual_processed/2003_processed/d31001_processed/ apw19981008.0841_processed/apw19981008.0841.txt, fouri, 5, kock, 5, evil, 1, vicencio, 1, villa, 1, black, 5, cape, 5, sake, 1, reconcili, 9, heal, 1, #, 4, , , +13:36:59-13:37:55

As mentioned, the fields of each result differ between survey types 4 and that of type 1, 2 & 3. The four surveys results are the same for the first four fields after which the sequentence and content vary to the point that some fields are left blank for processing reasons, for example:

Survey Types 1, 2 & 3 Date/Time Age Group — Sex — Survey Code — Level of Expetise — Process field — Doc Name/URL — word — level of representativeness x 10 — Timing Data

Survey Type 4 Date/Time — Age Group — Sex — Survey Code — Process field — Doc Name/URL — Level of Expetise — word — level of representativeness x 10 — Query — Timing Data

The different field contents are defined as:

- **DateTime**: Day, Date and time tag for the specific result.
- **Age Group**: Self descriptive age range selected by the participant
- **Sex**: The sex (m or f) as selected by the participant
- **Survey Code**: Indicates which survey type (1, 2, 3, or 4) this log result is for
- **Process field**:Field used in dynamic HTML processing of Web page
- **Doc Name/URL**: The address of the document (on the server) used in that survey
- #: Processing field used to indicate the change of task
- **Level of Expertise**: A number from 0 (low) to 6 (high) that represents the user's indicated level of expertise on the topic of the document used for that survey

- **word — level of representativeness**: 10 comma separated word and number sets indicating the participants' perception of the representativeness of the word for the given document
- **Query**: A search query input by the participant that they think they would use to find the given text in a normal interactive search
- **Timing Data**: A pair of start/finish times indicating the time taken to complete each task

6.1.4 Survey Documents

To lend statistical power to the experiment by reducing the variance involved, the number of documents used in the survey was limited to twenty manually classified news clippings each consisting of an average length of 514 words. Each participant had one of these clippings presented to them using an automatic selector that randomly retrieved them from the set of twenty. The document pool was in turn constructed using an automated random selector to pick from a total set of 1424 news clippings sourced from the Document Understanding Conferences (DUC) (NIST 2001) data sets of years 2001, 2002 and 2003. The Document Understanding Conference collated these documents for the study of document understanding, retrieval, and summarization, so these documents are assumed to be appropriate for this task.

6.2 The Nwords Survey

The Nwords experiments/surveys is being used to identify how many and which distinct words participants use in describing and searching for text, and to test how representative a participant thinks an automatically generated ranked lists of words (generated using a specific TFIDF function) is of a given document. To do this Nwords is actually four different experiments.

For readability purposes, the following discussion will use the numerals 1, 2, 3 and 4 to describe which of the four Surveys is being discussed and any point.

The Nwords experiment consists of four different surveys randomly selected and delivered to the participant for completion. Survey 1, 2 & 3 are designed to elicit

participant query and description term preference characteristics such as the number of terms they use and the level of importance they place on each. Survey 4 is used to measure how closely an automatically generated TFIDF ranked list of words derived from the given document agrees with the ranking a human would give to the list. The four survey paths are described in more detail following.

6.2.1 For ALL Surveys

To reduce any variance brought on by language, education or culture all survey instructions are kept simple and presented using plain English as will become evident. In this way the lesser educated non-English as first language speaker or culturally different participant should understand any instruction as much as the highly educated English first language speaking participant.

For the situation where the participant does not complete a portion of a survey correctly a Javascript error window is displayed with instructions on how to fix the error and an **OK** button.

Once the participant has completed a task they select the *Continue* button which presents them with the next task page.

All pages display a footer with standard page currency information and disclaimer. All task pages of the survey pages have a *Start Over* button in the bottom right hand corner which allows the participant to start again from the beginning for whatever reason, just like they could in a real search. However, because of security reasons we could not store personal session data which made it impossible to record the re-starts of a participant.

For maximum clarity and readability of text all task page backgrounds are light yellow and the text font used is Time New Roman.

All four Surveys end with a page thanking them for their participation and inviting them to do another survey (see Figure 6.2).

To start the experiment the participant is instructed by the invitation to navigate to the Nwords Introduction page (see Figure 6.3) at http://dweb.infoeng.flinders.edu.au which presents a simple introduction, links to further experiment and ethics information, and a *Continue* button in the middle of the screen.

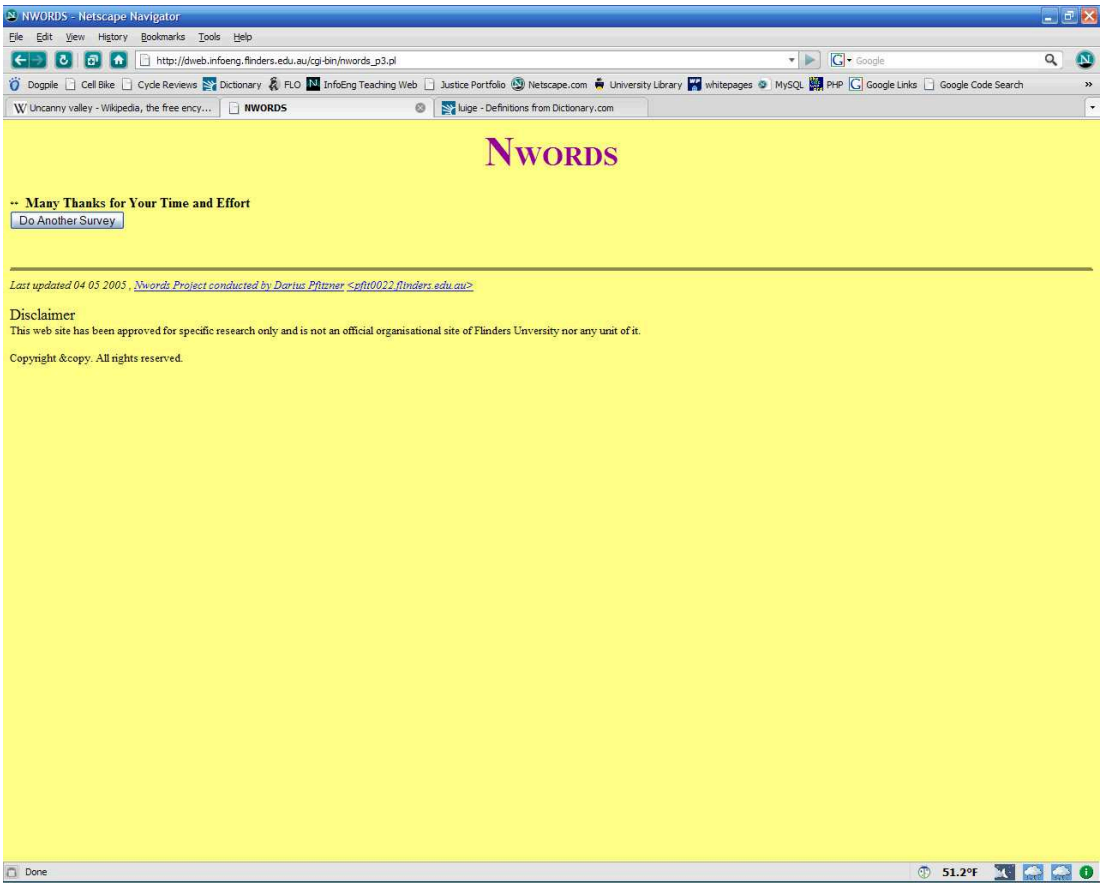


Fig. 6.2: Survey thankyou page

Upon pressing *Continue* the participant is presented with an elicitation page requiring the selection of age and sex details (see Figure 6.4) of the individual. A sex and age category must be set before the participant is allowed to progress to the next page using the *Continue* button which will display the first page of one of the four surveys.

Although at this point the different surveys begin, there is one common factor between the first pages of the four and that is that the participant is presented with a passage of text and asked to read it. As described below, the manner in which the text is displayed will differ as well as the task the participant is given.

6.2.2 Survey 1

6.2.2.1 Task 1.1

The first page of Survey 1 (see Figure 6.5) presents the participant with a passage of text and gives the three instructions:

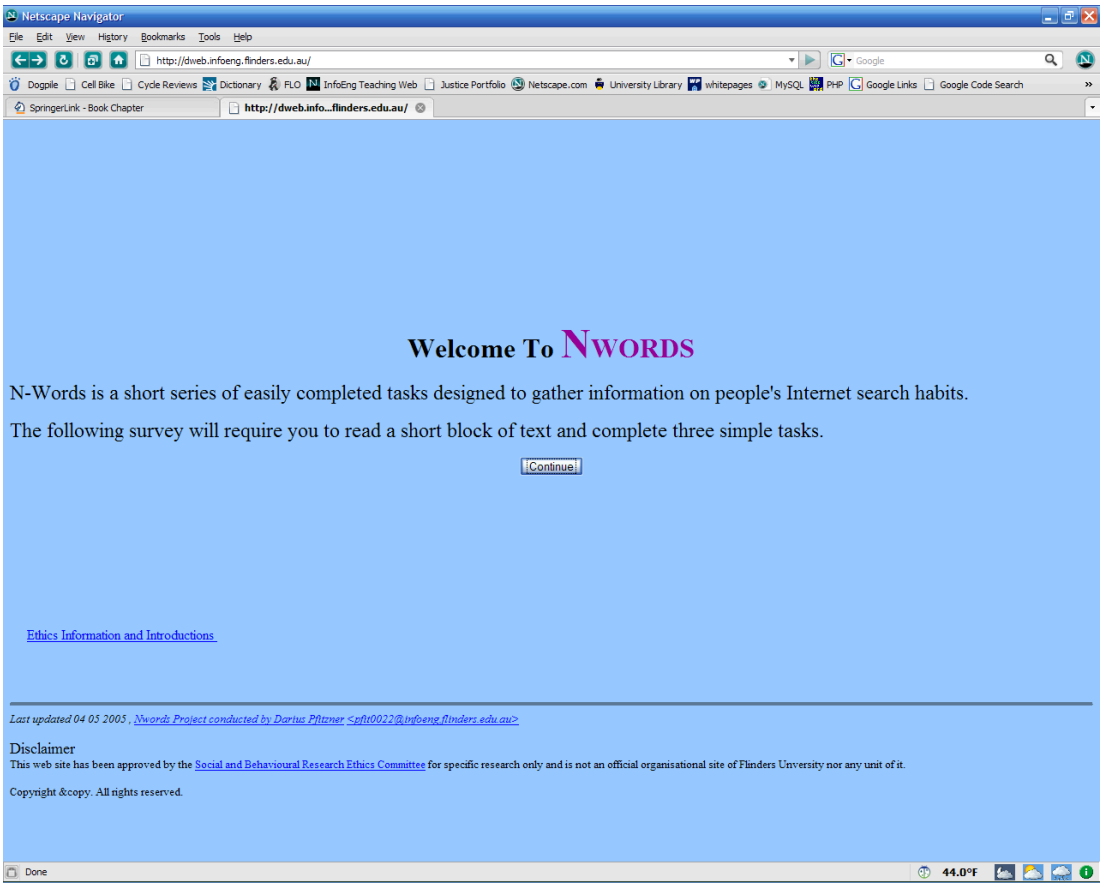


Fig. 6.3: Nwords Introduction Page

1. Read the text below.
2. In the answer field on the right, type the words &/or phrases you think best describe/represent what the text is about.
3. The words &/or phrases you choose **DO NOT** have to occur in the text.

Each words &/or phrase that is added to the list via the input field and ADD button which appends the entry to the list below the entry field which is not editable. If a word &/or phrase is input twice an error message is displayed.

6.2.2.2 Task 1.2

The second page of Survey 1 (see Figure 6.6) presents the participant with a Web query like text input field and gives the instruction:

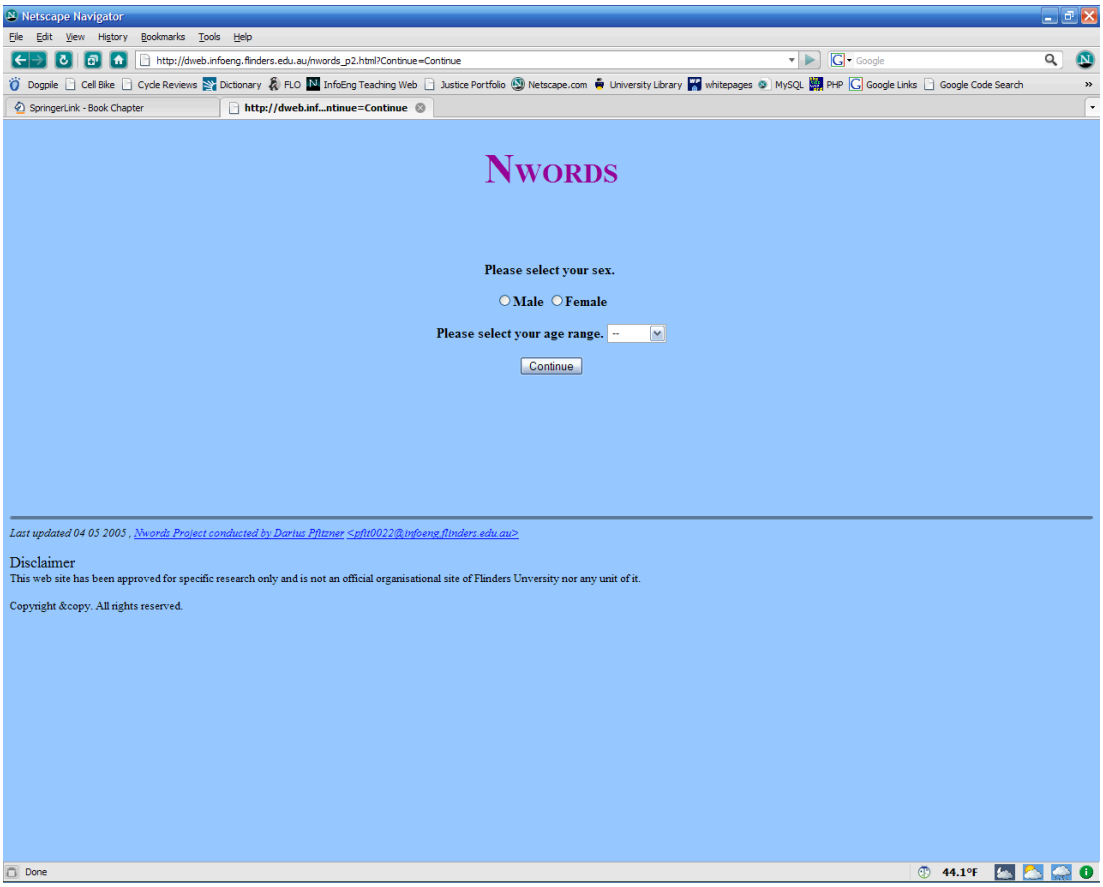


Fig. 6.4: Participant profile elicitation page

1. List the Search terms you might use to find this text using an Internet search engine.

6.2.2.3 Task 1.3

The third page of Survey 1 (see Figure 6.7) presents the participant with two table containing radio button selections.

Combined with the following instruction the first table lists the words & phrases the participant used in the previous query interface task (without any stop words).

1. For each of the words or phases below please indicate to what level it describes the original text.

Combined with the following instruction the second table is used to ask the participant how familiar they are with the topic of the text given at the beginning of the survey.

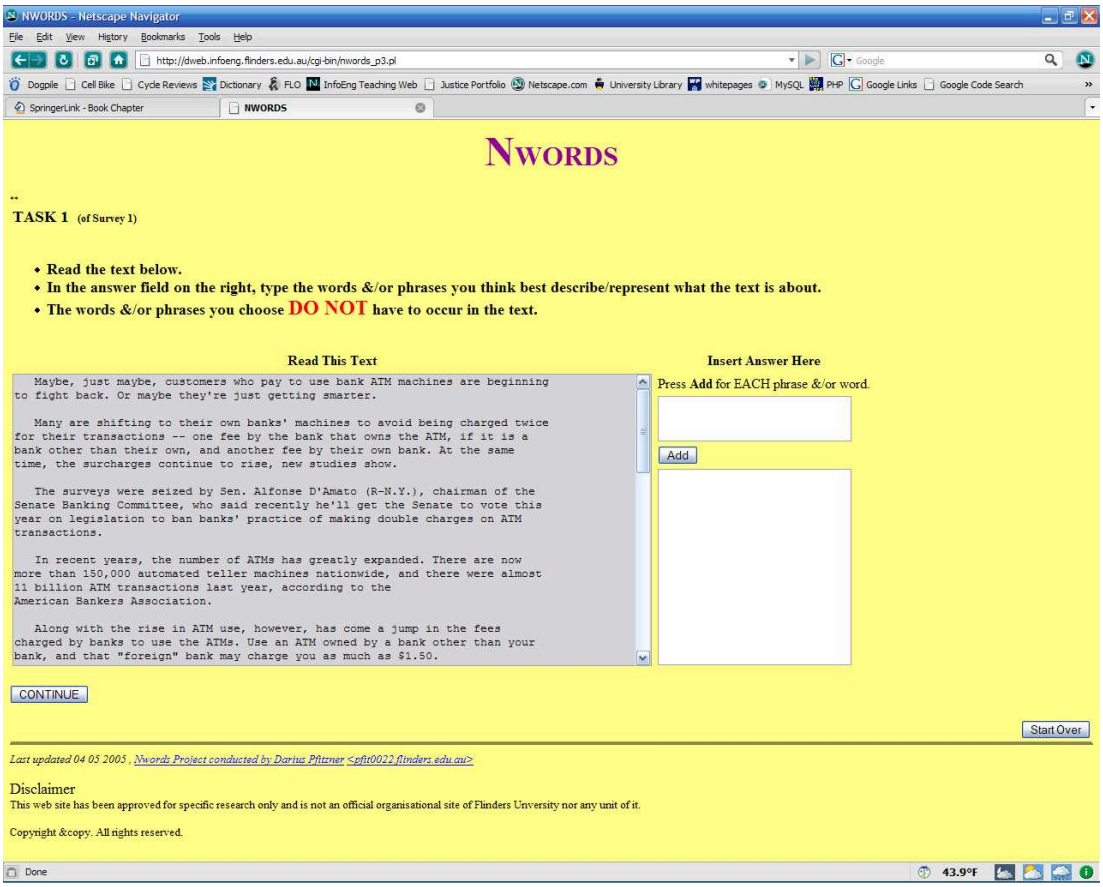


Fig. 6.5: Task 1 of Survey 1

1. Please rate your level of expertise/familiarity in the text's topic field

6.2.3 Survey 2

6.2.3.1 Task 2.1

The first page of Survey 2 (see Figure 6.8) presents the participant with a passage of text and gives the instruction:

1. Please carefully read the text below.

After the text just above the Continue button the participant is given the clear instruction:

- Before selecting Next ensure you understand what the above text is about

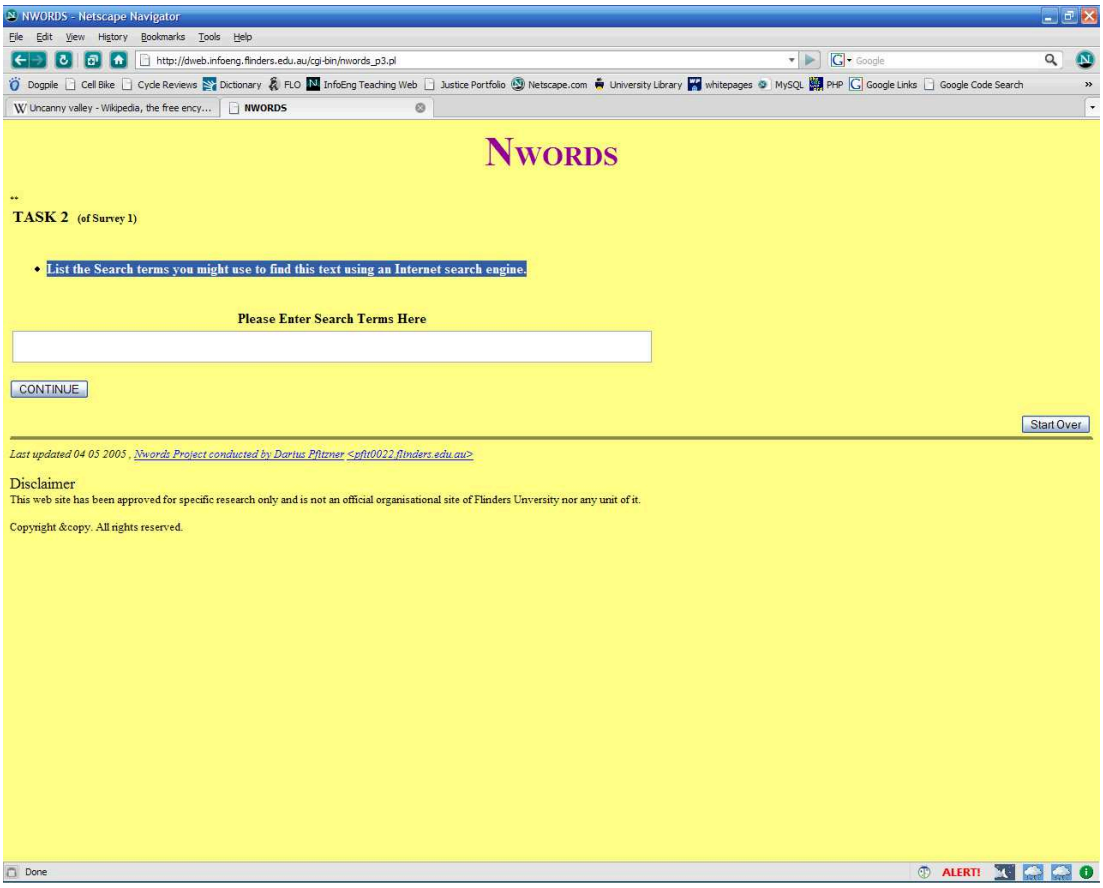


Fig. 6.6: Task 2 of Survey 1

6.2.3.2 Task 2.2

The second page of Survey 2 (see Figure 6.9) presents the participant with the same word &/or phrase input mechanism as used in Survey 1 without the accompanying text and gives the instruction:

1. **WITHOUT** Looking at the Previous Text. In the answer field, type the words &/or phrases you think best describe/represent what the text is about and press Add each time.

2. The words &/or phrases you choose **DO NOT** have to occur in the text.

6.2.3.3 Task 2.3

The third page and task of Survey 2 is identical to that found in Task 1.2 of Survey 1 (see Section 6.2.2.2).

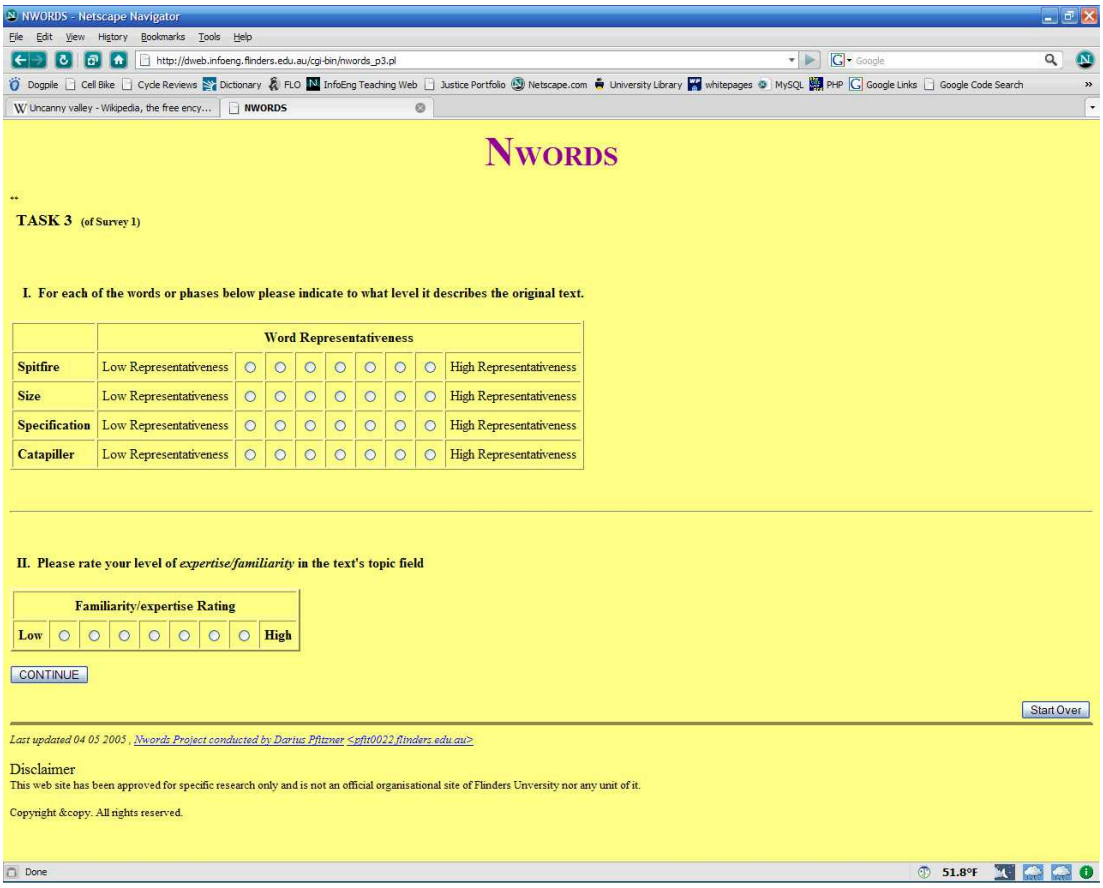


Fig. 6.7: Task 3 of Survey 1

6.2.3.4 Task 2.4

The fourth page and task of Survey 2 is identical to that found in Task 1.3 of Survey 1 (see Section 6.2.2.3).

6.2.4 Survey 3

6.2.4.1 Task 3.1

The first page of Survey 3 and instructions contained within are the same as Task 1 of Survey 1 (see Section 6.2.2) except that the words "**DO NOT** have to" in the third instruction have been replaced with the word "**MUST**", for example:

1. Read the text below.
2. In the answer field on the right, type the words &/or phrases you think best describe/represent what the text is about.

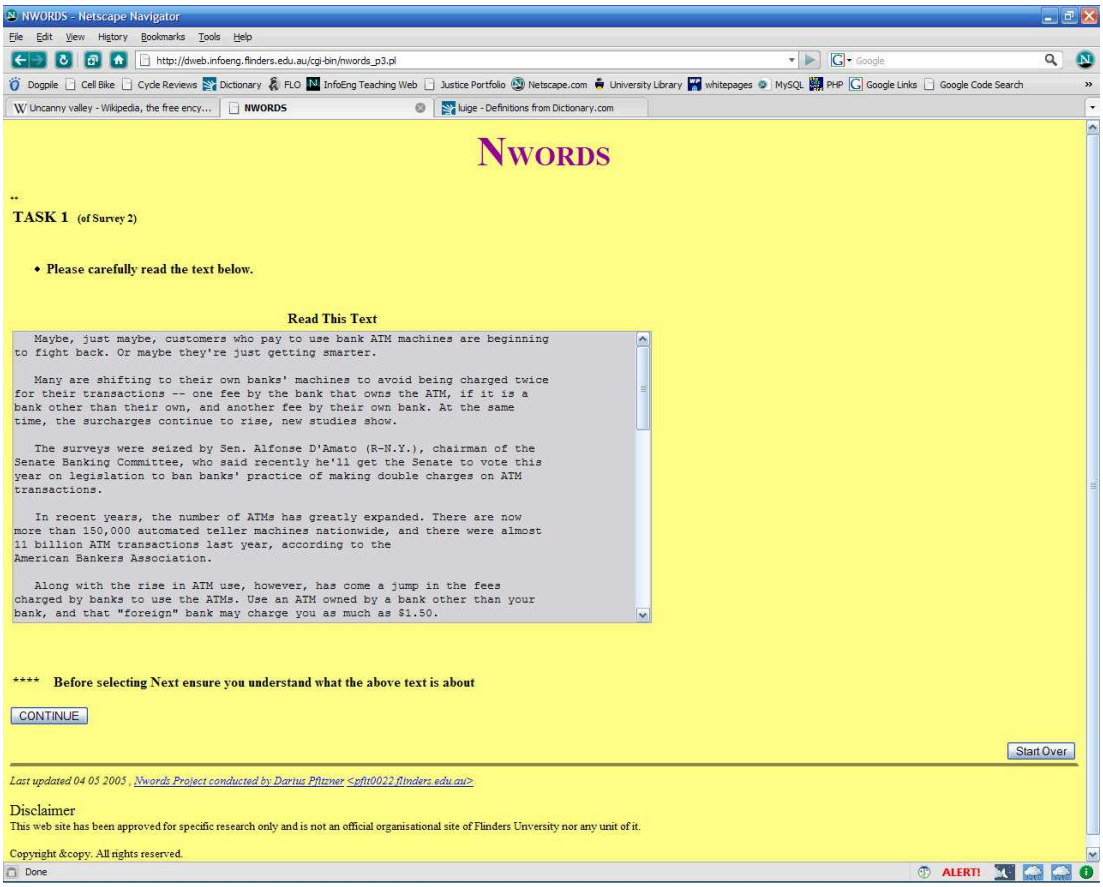


Fig. 6.8: Task 1 of Survey 2

3. The words &/or phrases you choose **MUST** occur in the text.

6.2.4.2 Task 3.2 & 3.3

The second and third pages of survey 3 are exactly the same as that of Task 1.2 and Task 1.3 of Survey 1 (see Section 6.2.2)

6.2.5 Survey 4

6.2.5.1 Task 4.1

The first page of Survey 4 and instructions within are exactly the same as Task 2.1 of Survey 2 (see Section 6.8).

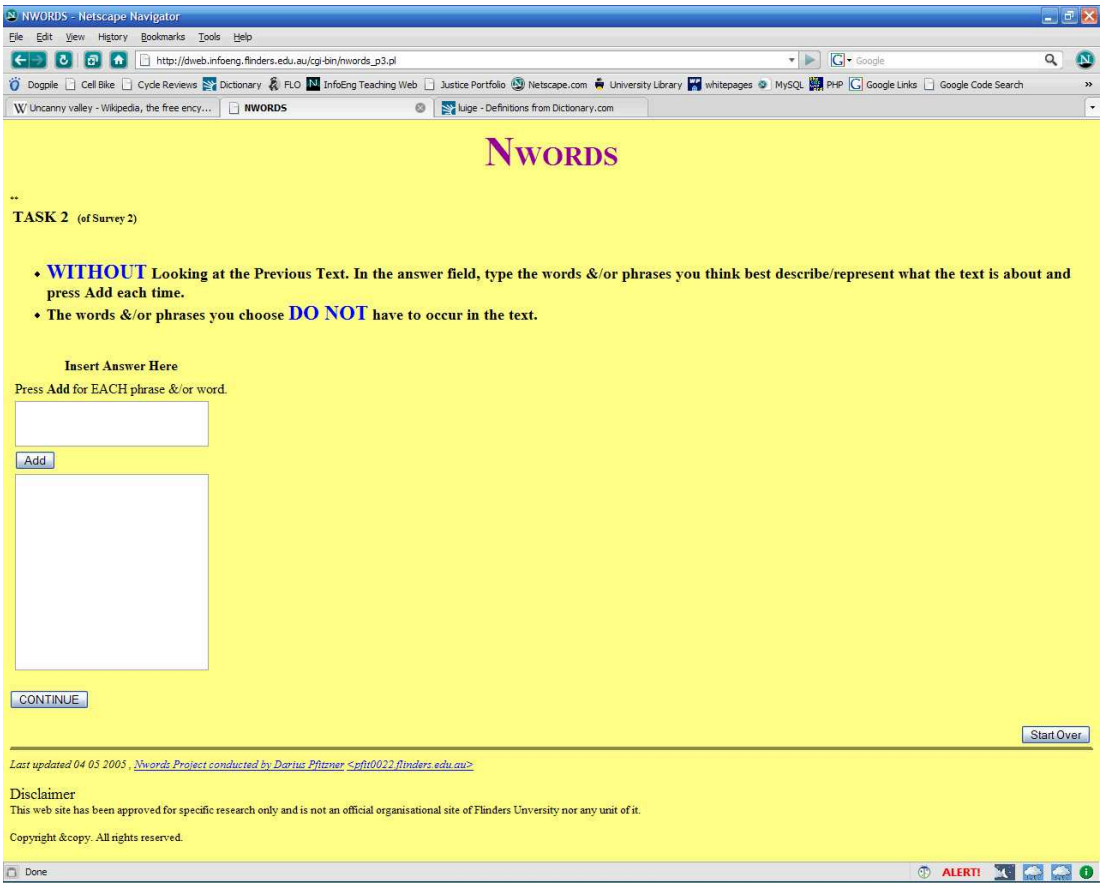


Fig. 6.9: Task 2 of Survey 2

6.2.5.2 Task 4.2

The second page of Survey 4 is similar to Task 1.3 of Survey 1 (see Section 6.2.2.3) except that the first table presents words automatically selected from the text using a specific TFIDF function as opposed to words selected by the user. It also has more radio buttons giving a greater similarity range to choose from. The instruction above the table reads as follows:

- For the words listed below, indicate how well each one describes/represents what the previous document is about.

The second table is exactly the same as that of Task 1.3 of Survey 1 (see Section 6.2.2.3).

6.2.6 Survey Reasoning

This section outlines the reasoning for each of the survey tasks. For convenience immediately following is a short outline of each survey and tasks for quick reference.

Quick Reference

Survey 1 - **WITH** access to the original text participant is asked to:

1. List those words or phrases that best *describe* the text
2. List those *search* terms they might used to find such a text on the Web
3. Rank the importance of each term input in task 1 in terms of its descriptiveness of the document
4. Rate their level of knowledge/understanding/expertise with of general topic of the text

Survey 2 - As for Survey 1 but **WITHOUT** access to the text

Survey 3 - **WITH** access to the original text participant is asked to:

1. List those words or phrases, that **only** occur in the text, that best *describe* the text
2. List those *search* terms they might used to find such a text on the Web
3. Rank the importance of each term input in task 1 in terms of its descriptiveness of the document
4. Rate their level of knowledge/understanding/expertise with the general topic of the text

Survey 4 - **WITH** access to the original text

1. The participant is presented with a selection of word groups (based on common stems) derived from the document and is asked to rate each based on their descriptiveness of the document
2. Participant is asked to rate their level of knowledge/understanding/expertise with of general topic of the text

Task Reasoning

Task 1 in Survey 1, 2 & 3 targets the testing of how many and which types of words/phrases people use to describe documents as well as building a database useful for research into other description term usage characteristics.

Task 2 in Survey 1, 2 & 3 is used to test the number and type of words people use to query documents. It also allows for relative comparisons in differentiating between the act of describing a document and searching for a document, and for future research into the identification of any patterns regarding word sequence. Word sequence patterns are of interest as search engines like Google place more importance on a word the earlier it occurs in the search term set. Because of this situation variants of this Task were created in Surveys 1, 2 & 3 to emulate different usage situations as described in the following:

Survey 1 By allowing access to the document while completing the task and allowing the use of any words including those in the document the situation is emulated in which a user might have a document at hand and wishes to locate another document like it using a search.

Survey 2 By removing access to the document while completing the task but still allowing the use of any words including those in the document the search for a known document situation is emulated. That is the searcher knows of the document with a level of detail through interactive experience and needs to find it again.

Survey 3 By allowing access to the document while completing the task but insisting all words used must occur in the document the survey mimics the description scenario of Task 1 but with a different interface and imperative. This allows us to create a comparative data set for research into the generation of automated summaries.

Task 3 in Survey 1, 2 & 3 is designed to test human rank preference for different descriptor words by asking participants to scale a set of words derived from their Task 1 terms, relative to each other and their perceived descriptiveness of the text. The set of words derived from Task 1 are generated by firstly grouping all the words together according to their parent stem. The groups of words whose stem occurred in the top

ten TFIDF list were then presented to the participant for ranking via radio button selection.

The Survey 4 is designed to test how closely the rank imposed on a list of automatically selected words by participants matches that imposed by the TFIDF formula discussed in Chapter 7.

6.2.7 Data Processing

Driven by the needed to compare the number of terms and words used by participants under the different input device and tasks the results were processed to generate counts of the number of:

1. Terms used (a term being one word-stem or a sequence of word-stems delimited by the use of the "ADD" button or by a comma in the query-word sequence). Presented under the column "**Terms**".

2. Words used (note that all words have been conflated in a stemming process). Presented under the column "**Stems**".

3. Distinct words used. Presented under the column "**Distinct Stems**".

4. Words used in more than one Term. Presented under the column "**Stems Intersections**".

5. Distinct words used in more than one Term. Presented under the column "**Distinct Stems Intersections**".

6. Distinct words used that also occurred in the list of top ten TFIDF stems. Presented under the column "**Distinct Stem / Top Ten TFIDF Intersection**".

6.3 Results Treatment

Although a cursory inspection of results prior to analysis revealed no problems, during the analysis process a possible problem/error was noted that required closer inspection. This was that for all participant responses to tasks One & Two 23% of the distinct descriptor stems and 35% of the distinct query stems did not occur in the Top 10 TFIDF listing of stems and that for task Three 11% of the distinct descriptor stems and 17% of the distinct query stems did not occur in the same listing.

This was extremely disturbing given TFIDF's extensive use, across many fields, to represent the importance of human concepts in textual situations via relevance weightings. Because these types of algorithm are used to represent concept relevance, which is subjective in nature, it is reasonable to hope they would produce ranked lists that closely relate to human lists for the same documents. This is the cause of the concern as the above observation indicates that a seemingly large proportion of the participants produced lists bear **no** relationship to the TFIDF list.

The concern over these missing terms arises from the fact that the document set used was selected via an automated random selection device to avoid human error. However, this left the study open to a possible error situation brought about by documents with complex topic structures. This situation is characterized by documents that have several clear topics that might cause participants to select one topic over the other or over represent one topic. Upon closer inspection it was realized that this was not the case as when the documents in question (target set) are compared to the total available document pool two facts are noted:

1. most available documents are present in the target set, and
2. no documents in the target set are over represented.

In other words the null set represent an expected random distribution across the population. These observations were further supported by a subjective manual analysis of all documents in the set, for which it may be concluded that although most documents had multiple sub-topic all had clear overarching topics.

As a result of this concern over the document set it was felt that the results data should be closely inspected to avoid any assumptive errors and thus ensure confidence

when making any observations or conclusions. The major concern was to avoid the error of assuming that "the results data is clean and free of errors" and thus a closer inspection of the result set was made. In short, an extremely laborious manual inspection for any possible anomalies was conducted for the total results data.

This inspection highlighted the fact that although the Nwords tasks are relatively simple and straightforward the results for some participants contained errors. Although these errors, such as misspelling or forgetting to put a space between words, were minor, due to the relatively low number of terms used on average they could have had substantial effects on final analysis results. To avoid this it was decided to hand filter the results using some fairly simple but justifiable and clearly specified rules. These were:

1. If a word is misspelt when compared to its nearest match in the relevant text it was corrected to be the same. It should be noted that there was some concern as to whether this rule was too General or not as it did not capture situations like where the intended meaning of a word might be indiscernible. However, after filtering all the results this situation was not encountered and thus did not need managing.

2. If a term clearly had several words appended together they were separated at the obvious demarcation points. No interpretative situations were encountered here as all mistakes of this type were as obvious as "HusseinClintonKing" which was converted to "Hussein", "Clinton" & "King".

3. If all the supplied terms are judged to relate in no way to the original text then the whole result is removed, for example one participant had used the terms "chicken salad" and "Omega" to describe and query for a document that discussed Russian political killings, or another participant used one irrelevant word only to describe and query for a document about schizophrenia. Although low in number the results that conformed to this rule were taken to indicate that the participant was simply mucking around.

After applying these rules 10 records were removed while 2 more were removed because of logging system errors (one simply a newline in the wrong place so was not a record and the other was missing its timing information). In total this reduced

the record set from 246 records to 234. The 10 results that conformed to rule *3* are presented in Appdenix 10.3 with an example of a randomly selected result for relative inspection.

6.3.1 Outlier Treatment

The research results presented in Section 6.4 and Section 7 uses outlier exclusion to limit the influence outliers have on the sample relative to the average distribution. Following is a short discussion regarding the background and treatment of outliers relative to this research.

Outlier management finds its roots in statistics and is a well researched field with established general and domain specific techniques (Markou & Singh 2003). Broadly speaking, there are two common approaches used; one incorporates explicit distance metrics to determine the degree to which an object is an outlier and requires relatively extensive processor and memory resource allocation; the other uses implied distance metrics, in the form of domain space quantization, to make comparisons at a high level of abstraction and avoid the extensive pair-wise comparison of members which in turn reduces memory requirement when processing large data sets (Knorr & Ng 1998, Chaudhary, Szalay & Moore 2002, Papadimitriou, Kitagawa, Gibbons & Faloutsos 2003, Chiu & Fu 2003).

Outlier exclusion is a common requirement in research brought about by observations that fall far outside a subjective judgment of what the norm might be. Outliers are often removed because they can inappropriately influence results relative to the average distribution. Exclusion is often achieved using quantitative methods such as the exclusion of any observations that fall outside a specific range like ± 2 standard deviations or even ± 1.5 standard deviations around a central value like the group mean. This is generally called "data cleaning" and is a requirement in fields like cognitive psychology where relatively small numbers of extreme outliers can completely overwhelm the final results and subsequent conclusions drawn from the analysis thereof. For example, if you are measuring reaction times in the range of say 400-700 milliseconds and several results (say resulting from erroneous "distracted reactions") occur in the range of $10 - 15$ seconds they will completely skew any results. Results with such a relatively large effect compared to the average population need to be removed before appropriate

observations can be made about the average normal population.

Relative to the results presented in Section 6.4 and 7 the upper and lower fences of the box-plot calculations, outlined in the following Section (6.3.2), are used as cut-off values for outlier removal.

6.3.2 Visual Presentation of Statistics

The research in Section 6.4 and 7 targets the identification of how many *whole* words participants use under the conditions of different experiments. Although mean statistics are used to support observation in this research because they are not restricted to whole numbers they have been used in conjunction with median and mid-quartile statistics to present *whole* number observations about word usage. A benefit of incorporating the use median statistics in statistical analysis is that they are less susceptible to adverse affects of outliers.

To present this median and mid-quartile statistics this research applies a common approach used in cognitive science, the "Box Plot", to visualize data set statistics. Box-plots are used to allow the rapid visual assessment in the recognition of central tendency, outliers, distribution characteristics and spread of data sets (Chambers, Cleveland, B. & Tukey 1983, Howell 1997).

Because there can be some variation in the manner in which box-plots are produced the statistics generated in the creation of the boxplots used are as follows:

- Median Location (ML) = $(N + 1)/2$
- Hinge Location (HL) = (ML + 1)/2
- Lower Hinge (LH) = HLth lowest score
- Upper Hinge (UH) = HLth highest score
- H-spread = UH - LH
- Lower fence = LH - 1.5(H-spread)
- Upper Fence = UH + 1.5(H-spread)
- Lower Adjacent Value = smallest value >= lower fence

- Upper Adjacent Value = largest value <= upper fence

For the readers convenience, following is a short list of key points to help with the interpretation of boxplots:

- The tops and bottoms of each "box" are the 25^{th} and 75^{th} percentiles of the samples, respectively. The distances between the tops and bottoms are the interquartile ranges.
- The line in the middle of each box is the sample median. If the median is not centered in the box, it shows sample skewness.
- The "whiskers" extending above and below each box are drawn from the ends of the interquartile ranges to the furthest observations within the whisker length (the adjacent values).
- Observations beyond the whisker length are outliers. An outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box.
- The notches in the boxes display the variability of the median between samples. The width of a notch is computed so that box plots whose notches do not overlap have different medians at the 5% significance level. The significance level is based on a normal distribution assumption, but comparisons of medians are reasonably robust for other distributions. Comparing box-plot medians is like a visual hypothesis test, analogous to the t test used for means.

6.4 Nwords Results

The results of the Nwords surveys are presented in five sections. This first Section (6.4.1) presents and discusses the results using a *between-survey* perspective to compare the effect of the different task environments. The second Section (6.4.2) presents and discusses the results using a *between-task* perspective that combines the results of surveys to compare the difference between the tasks. The third Section (6.4.3) discusses analysis conducted to identify any effects that might arise from the documents themselves. The fourth Section (6.4.4) presents and discusses the results of Survey 4 which was designed to investigate TFIDF and its relationship to human preference. The fifth Section (6.4.5) presents observations from a correlation analysis of key data.

These Sections are followed by a concluding section (8.5) that summarizes the critical observations and makes several conclusions from specific observation.

Raw results for the Nwords experiments are presented in Tables 10.2 and 10.1 (with standard errors).

6.4.1 Between Surveys Results Analysis

Following are *between survey* type analysis of the Nwords survey results for surveys type 1, 2 and 3. Comparisons are presented for the number of:

- *terms* ($\geq$ 1 word) used to describe the document in context (Section 6.4.1.1)
- distinct *description* stems used to describe the document in context (Section 6.4.1.2)
- distinct *query* stems used to search for the document in context (Section 6.4.1.3)
- distinct *description* stems that also occur in the top ten TDIDF rank stems (Section 6.4.1.4)
- distinct *query* stems that also occur in the top ten TDIDF rank stems (Section 6.4.1.5)

Each comparison presents a short discussion and two box-plots of the differences between surveys 1, 2 and 3. The first plots in each comparison depict results with

outliers and the second without outliers, these are both followed immediately by a table of key statistics. There is also a table between the first statistics table and the second figure that indicates how many outliers were removed to form the second boxplot and the number of subjects for each survey. All results have been treated as discussed in Section 6.3.

Boxplots have been used in this case for the comparison of medians for rapid visual hypothesis testing as the boxplot error representation is analogous to the t-test used for means. The key to interpretation of these plots are the notches on the mid-quartile ranges of the boxplots as they display the variability of the median between samples. The width of a notch is computed so that boxplots whose notches do not overlap have different medians at the 5% significance level. In other words if the notches of two boxplot do not overlap, you can conclude, with 95% confidence, that the true medians do differ. The significance level is based on a normal distribution assumption, but comparisons of medians are reasonably robust for other distributions.

Alternate t statistics are also presented to support any claims of similarity or dissimilarity between survey medians via analysis of population means. The t-test used is a homoscedastic test that assumes that the two data sets came from distributions with the same variances. The test is used to determine whether the two surveys represent samples that are likely to have come from distributions with equal population means.

6.4.1.1 Terms Comparison

Task 1 of surveys 1, 2 & 3 is designed to help quantify **the number of terms** ($\geq$ 1 word) people normally use to **describe documents**, and to supply data for use in identifying other characteristics of the actual terms used. In respect to quantifying the number of terms normally used, Figure 6.10 presents the results for visual analysis while Tables 6.3 & 6.4 present key statistical data.

Participants on average use four terms to describe a document under the conditions of surveys 2 & 3, and three terms under the conditions of survey 1.

Table 6.3 supports the following observations. The average number of terms used to describe a document was;

1. significantly lower for survey 1 than for survey 2 (p=0.011*)
2. significantly lower for survey 1 than for survey 3 (p=0.0003*)
3. not significantly different for surveys 2 and 3 (p=0.251)

	Surveys 1 & 2		Surveys 1 & 3		Surveys 2 & 3	
Comparison of Means[1]	needn't contain		access		no access	access
	access	no access	needn't	must	needn't	must
Mean	3.098	4.085	3.098	4.586	4.085	4.586
P(T¡=t) two-tail	**0.011***		**0.0003***		**0.251**	

Table 6.3: Term usage significance statistics

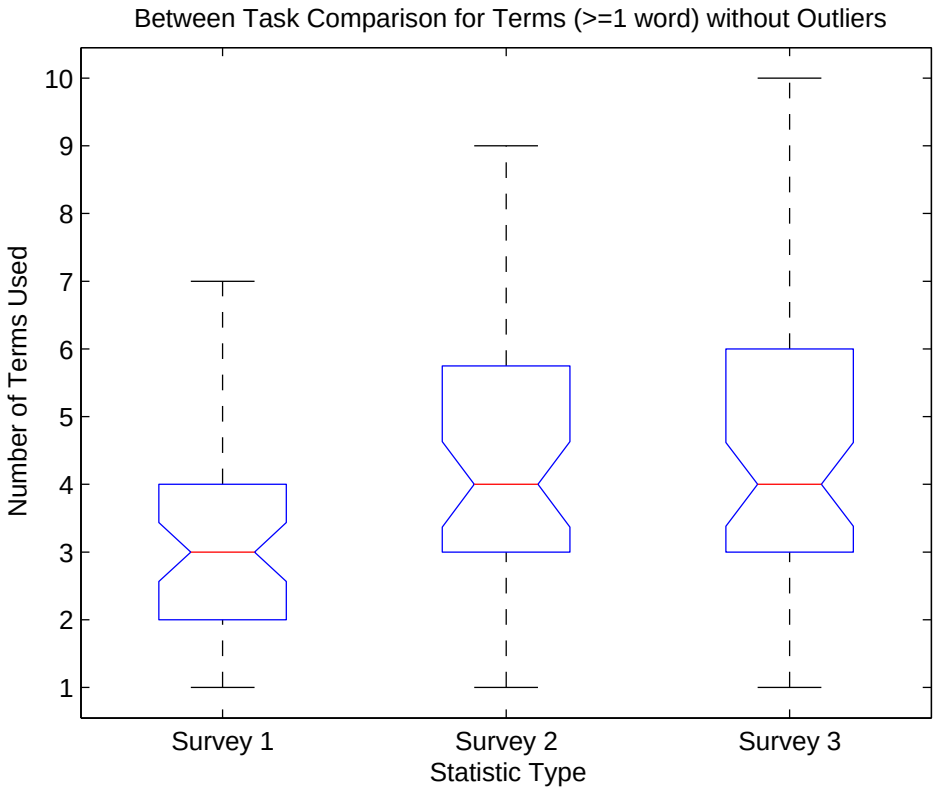


Fig. 6.10: Between surveys term ($\geq$ 1 word) usage (outliers excluded)

Critical Statistics (Outliers NOT Included)			
	Survey 1	Survey 2	Survey 3
Number of Outliers Removed	5	1	4
Number of Participants	56	48	62
Min	1	1	1
Median	3	4	4
Mean	3.098	4.085	4.586
Max	7	9	10
Std Dev	1.700	2.041	2.340
Std Err	0.240	0.301	0.310

Table 6.4: Between survey term ($\geq$ 1 word) usage (outliers excluded)

6.4.1.2 Description Stem Comparison

Task 1 of surveys 1, 2 & 3 is designed to help quantify the number of **distinct concepts** (distinct word stems) people might normally use to **describe documents**, and to supply data for use in identifying other characteristics of the actual words used. In respect to quantifying the number of concepts normally used, Figure 6.10 presents the results for visual analysis while Tables 6.5 & 6.6 present key statistical data.

Participants normally used:

five distinct stems to describe a document under the conditions of survey 1,
seven distinct stems under the conditions of survey 2 and
eight distinct stems under the conditions of survey 3.

Table 6.5 supports the following observations.

The average number of **distinct stems** used to **describe** a document was;

1. significantly lower for survey 1 than for survey 2 (p=0.008*)
2. significantly lower for survey 1 than for survey 3 (p=0.00005*)
3. significantly lower for survey 2 than for survey 3 (p=0.010*)

	Surveys 1 & 2		Surveys 1 & 3		Surveys 2 & 3	
Comparison of Means[2]	needn't contain		access		no access	access
	access	no access	needn't	must	needn't	must
Mean	5.432	7.170	5.432	9.947	7.170	9.947
P(T¡=t) two-tail	**0.008***		**0.000005***		**0.010***	

Table 6.5: Description stem usage significance statistics

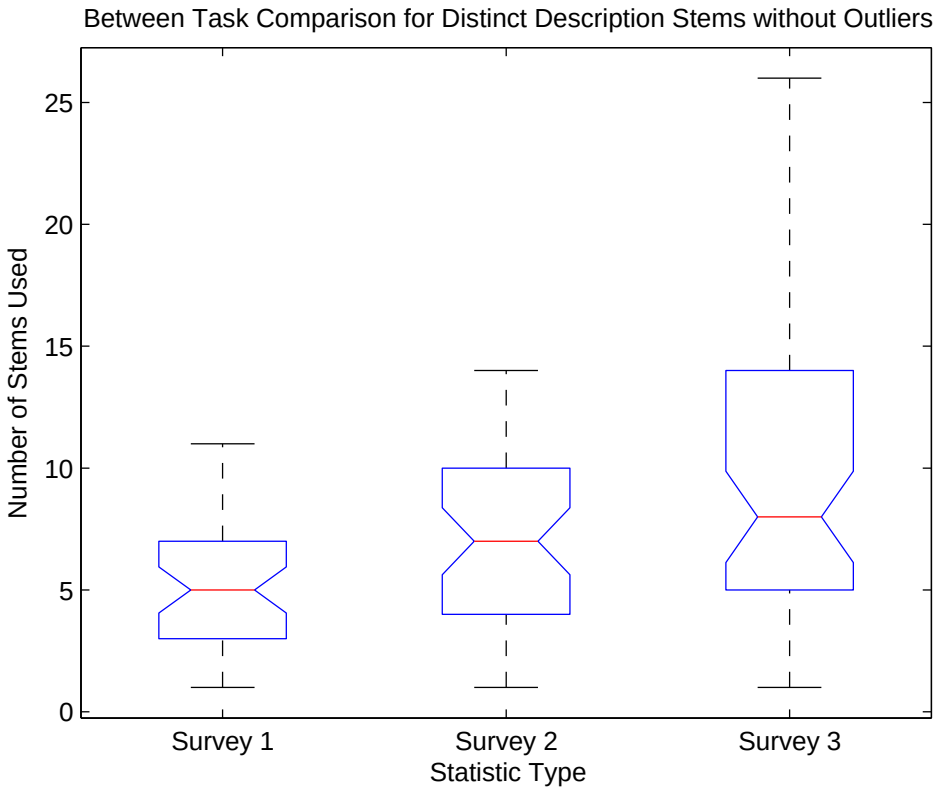


Fig. 6.11: Between survey description stem usage (outliers excluded)

Critical Statistics Outliers NOT Included			
	Survey 1	Survey 2	Survey 3
Number of Outliers Removed	6	1	5
Number of Participants	56	48	62
Min	1	1	1
Median	5	7	8
Mean	5.432	7170	9.947
Max	11	14	26
Std Dev	2.714	3.377	6.607
Std Err	0.414	0.498	0.883

Table 6.6: Between survey description **stem** usage (outliers excluded)

6.4.1.3 Query Stem Comparison

Task 2 of surveys 1, 2 & 3 is designed to help quantify the number of **distinct concepts** (distinct word stems) people might normally use to **search for** the document in context and to supply data for use in identifying other characteristics of the actual words used. In respect to quantifying the number of distinct concepts normally used, Figure 6.12 presents the results for visual analysis while Tables 6.7 & 6.8 present key statistical data.

Participants on average used **four** distinct stems to query for a document under the conditions of surveys 1, 2 & 3. The average number of **distinct stems** used to **query** for a document was;

1. not significantly different for surveys 1 and 2 (p=0.454)
2. not significantly different for surveys 1 and 3 (p=0.483)
3. not significantly different for surveys 2 and 3 (p=0.925)

	Surveys 1 & 2		Surveys 1 & 3		Surveys 2 & 3	
Comparison of Means[3]	needn't contain		access		no access	access
	access	no access	needn't	must	needn't	must
Mean	3.643	3.911	3.643	3.881	3.911	3.881
P(T¡=t) two-tail	**0.454**		**0.483**		**0.925**	

Table 6.7: Query stem usage significance statistics

Although it seems that there is no effect from the changing conditions between the surveys there exists a discrepancy between the median statistics (see boxplot Figure 6.12) and the t statistics (see Table 6.7) that should be noted. In the boxplot the median value and error region of survey 1 do not align with that of surveys 2 & 3, indicating that survey 1's conditions **did** result in participants using a different average number of distinct stems which is in disagreement with the t statistics which show strong support for **accepting** the null hypothesis. To reconcile this disagreement we look to the distribution of the population of survey 1, noting the relatively large standard deviation of **1.967**, the largest maximum value of **9** and the lowest mean of **3.643**. These are indicators of a skewed nature for the population which can also be seen in a large offset between the median and upper whisker values. The nature of the distribution indicates that the mean values are being disproportionately affected by the higher relative values of the long tail and as median statistics are more robust

to the effects of extreme and outlier values they are preferred for this analysis. This implies an alternate observation which suggests the conditions of survey 1 **did** have an effect, resulting in participants using **three** distinct stems to query for a document, not **four**.

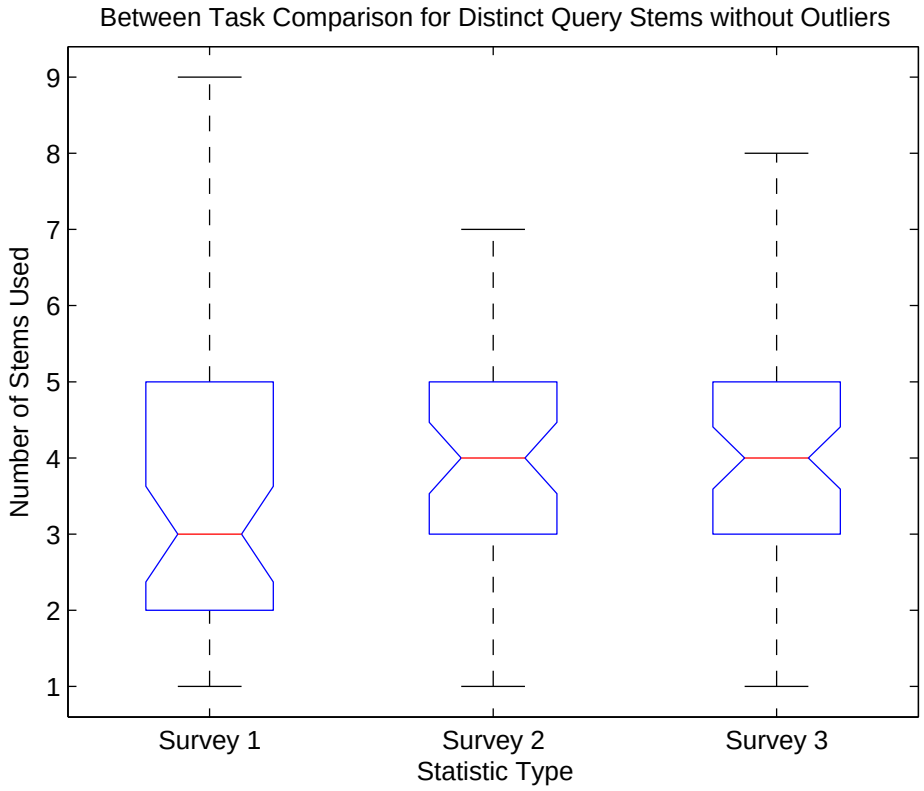


Fig. 6.12: Between survey query stems usage (outliers excluded)

Critical Statistics (Outliers NOT Included)			
	Survey 1	Survey 2	Survey 3
Number of Outliers Removed	0	2	3
Number of Participants	56	47	62
Min	1	1	1
Median	3	4	4
Mean	3.643	3.911	3.881
Max	9	7	8
Std Dev	1.967	1.520	1.662
Std Err	0.265	0.229	0.218

Table 6.8: Between survey query stems usage (outliers excluded)

6.4.1.4 Description Stem Occurrence in Top Ten TFIDF List Comparison

Task 1 of surveys 1, 2 & 3 is designed to help quantify the number of distinct concepts (word stems) people might normally use to describe documents and to supply data for use in identifying other characteristics of these stems. In respect to identifying other characteristics of these concepts Figure 6.13 and Tables 6.9 & 6.10 present key statistical data for comparative analysis of the amount of **distinct stems** participants used to **describe** the document in context that are also one of the **top ten TFIDF ranked stems** for that document.

Participants on average are likely to use **two** distinct stems, that are also one of the top ten TFIDF ranked stems, to describe a document under the conditions of surveys 1 & 2, and **three** under the conditions of survey 3.

The average number of **distinct stems** used to **query** for a document that also occur in the **top ten TFIDF stem list** was:

1. not significantly different for surveys 1 and 2 (p=0.904)
2. significantly lower for survey 1 than for survey 3 (p=0.009*)
3. significantly lower for survey 2 than for survey 3 (p=0.016*)

	Surveys 1 & 2		Surveys 1 & 3		Surveys 2 & 3	
Comparison of Means[4]	needn't contain		access		no access	access
	access	no access	needn't	must	needn't	must
Mean	1.964	2.000	1.964	2.839	2.000	2.839
P(T¡=t) two-tail	**0.904**		**0.009***		**0.016***	

Table 6.9: Description stem/TFIDF intersection significance statistics

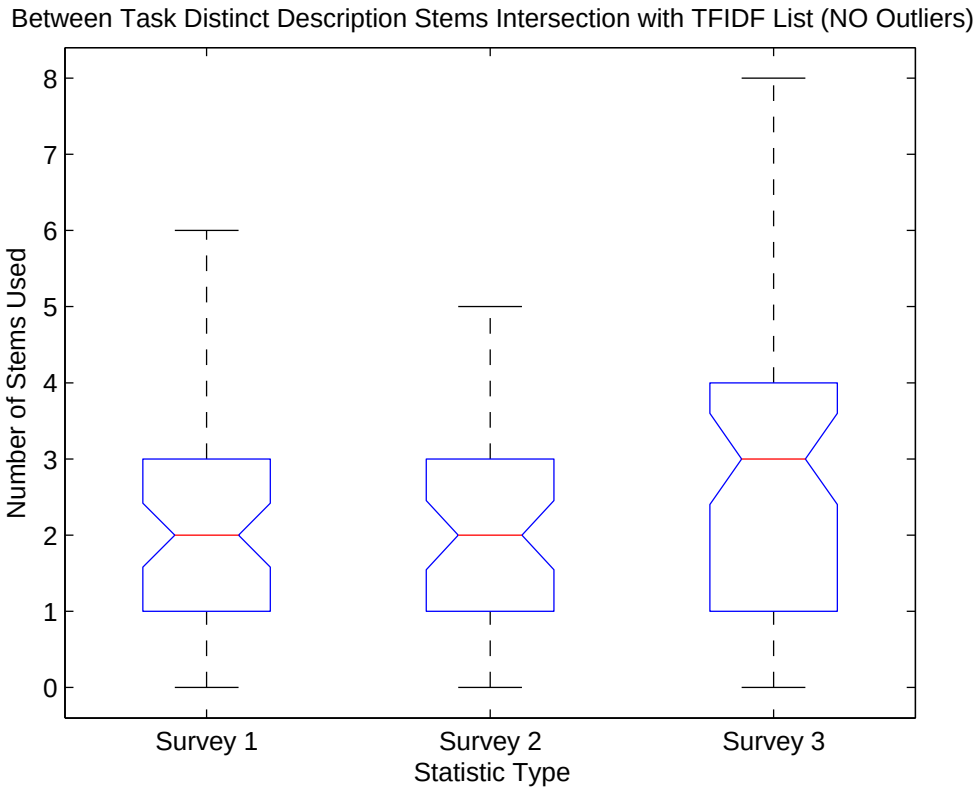


Fig. 6.13: Between survey description/TFIDF stems usage (outliers excluded)

Critical Statistics (Outliers NOT Included)			
	Survey 1	Survey 2	Survey 3
Number of Outliers Removed	1	0	0
Number of Participants	56	48	62
Min	0	0	0
Median	2	2	3
Mean	1.964	2.000	2.839
Max	6	5	8
Std Dev	1.527	1.502	1.969
Std Err	0.208	0.219	0.252

Table 6.10: Description stem/TFIDF intersection statistics

6.4.1.5 Query Stem Occurrence in Top Ten TFIDF List Comparison

Task 2 of surveys 1, 2 & 3 is designed to help quantify the number of distinct concepts (word stems) people might normally use to search for the document in context and to supply data for use in identifying other characteristics of these stems. In respect to identifying other characteristics of these concepts Figure 6.14 and Tables 6.11 & 6.12 present the results for comparative analysis of the amount of **distinct stems** participants used to **search** for a document that are also one of the top ten TFIDF ranked stems for that document.

Participants are likely to use **one** distinct stem to **describe** a document that is also one of the **top ten TFIDF ranked stems** under the conditions of all surveys.

The average number of **distinct stems** used to **describe** a document that also occur in the **top ten TFIDF stem list** was;

1. not significantly different for surveys 1 and 2 (p=0.228)
2. significantly lower for survey 1 than for survey 3 (p=0.041)
3. not significantly different for surveys 1 and 2 (p=0.559)

	Surveys 1 & 2		Surveys 1 & 3		Surveys 2 & 3	
Comparison of Means[5]	needn't contain		access		no access	access
	access	no access	needn't	must	needn't	must
Mean	1.054	1.298	1.054	1.410	1.298	1.410
P(T¡=t) two-tail	**0.228**		**0.041***		**0.559**	

Table 6.11: Query stem/TFIDF intersection significance statistics

An interesting observation to note is that survey 3 always resulted in no less than one query stem/TFIDF stem list intersection as opposed to surveys 1 & 2 that realized some queries with no intersections with the TFIDF stem list.

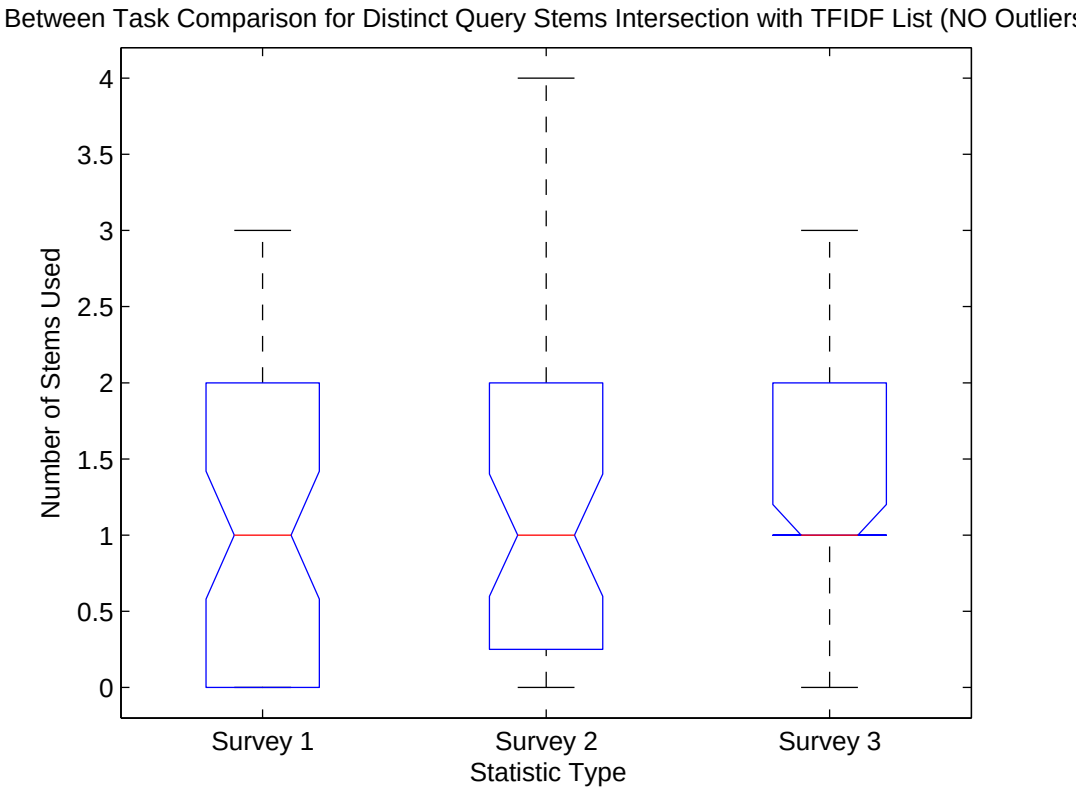


Fig. 6.14: Between survey query/TFIDF stems usage (outliers excluded)

Critical Statistics (Outliers NOT Included)			
	Survey 1	Survey 2	Survey 3
Number of Outliers Removed	0	0	1
Number of Participants	56	47	62
Min	0	0	0
Median	1	1	1
Mean	1.054	1.298	1.452
Max	3	4	3
Std Dev	0.961	1.082	0.953
Std Err	0.130	0.159	0.122

Table 6.12: Query stem/TFIDF intersection statistics

6.4.1.6 Multiple Word Instances

An issue of interest is the multiple use of words to describe and query for documents. This is of importance for the automatic generation of document summaries (descriptions) because if there is a general tendency for users to use words multiple times then a case can be made for further investigation. Investigations would look at the use of multiple word occurrences in automatically generated summaries and their treatment when encountered during text filtering and parsing processes. Be they descriptive segments of text or keyword lists, summaries should be efficient and effective in communicating the core concepts of a text. This implies that knowing if and when, and what words to repeat, is of value in the realization of succinct and appropriate summaries.

Search engines apply weightings to both the words in documents and the words of queries. Given that such weighting schemes as TFIDF and other information theoretic approaches often consider repeat occurrences as part of their calculation, knowing if repeats are important or in what situations this is true will affect the application of such schemes. For example, does multiple use of words in queries imply that the users think that those words are more important and thus should be given more status or is it simply an artifact of language syntax that should be ignored. Following from this, if users do in fact place increased importance through the use of multiple word instances then does that make them also important when automatically weighting words for tasks such as index generation, term treatment (document, query, & description terms) and during the summary generation process?

In testing participant tendency to use multiple word instances the numbers of non-distinct and distinct stems used in the **description task** (Task 1) were compared across surveys 1, 2 & 3. Tables 6.13 and 6.14, and Figure 6.15 present results for comparative analysis of the amount of **distinct stems** participants used relative to the **total stems** used to describe a document.

Participants are **NOT** likely to use multiple stem instances to describe a document under the conditions of Surveys 1, 2 & 3.

The difference between the average number of **distinct stems** compared to **total stems** used was:

1. not significantly different for Survey 1 (p=0.177)
2. not significantly different for Survey 2 (p=0.303))
3. not significantly different for Survey 3 (p=0.785))

	Survey 1 needn't contain access		Survey 2 needn't contain no access		Survey 3 must contain access	
	Total Stems	Distinct Stems	Total Stems	Distinct Stems	Total Stems	Distinct Stems
Mean	6.340	5.432	7.957	7.170	10.286	9.947
P(T≤t) two-tail	**0.177**		**0.303**		**0.785**	

Table 6.13: Multiple Description stem usage statistics

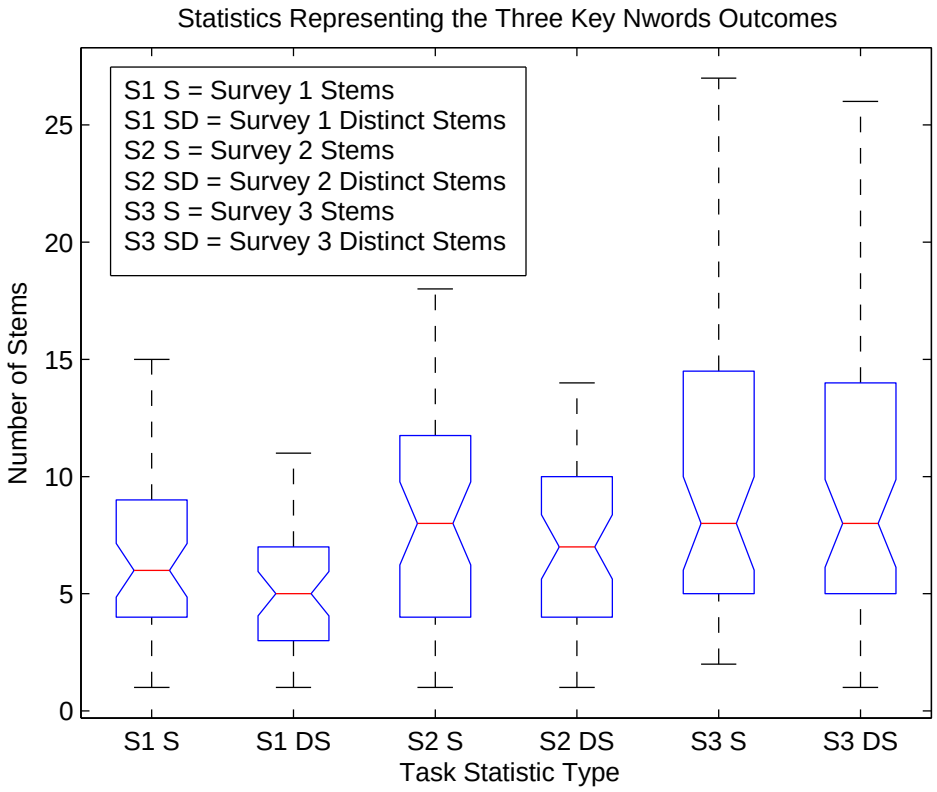


Fig. 6.15: Comparison between non-distinct and distinct description stem counts

Critical Statistics Outliers NOT Included						
	Survey 1 needn't contain access		Survey 2 needn't contain no access		Survey 3 must contain access	
	Total Stems	Distinct Stems	Total Stems	Distinct Stems	Total Stems	Distinct Stems
Number of Outliers Removed	9	12	1	1	6	5
Number of Participants	56	56	48	48	62	62
Min	1	1	1	1	2	1
Median	6	5	8	7	8	8
Mean	6.340	5.432	7.957	7.170	10.286	9.947
Max	15	11	18	14	27	26
Std Dev	3.565	2.714	3.962	3.377	6.525	6.607
Std Err	0.526	0.414	0.584	0.498	0.880	0.883

Table 6.14: Description stem usage statistics

In testing participant tendency to use multiple word instances the number of non-distinct and distinct stems used in the **query task** (Task 1) were compared across surveys 1, 2 & 3. Tables 6.15 and 6.16, and Figure 6.16 present results for comparative analysis of the amount of **distinct stems** participants used relative to the **total stems** used in a query for a document.

Participants are **NOT** likely to use multiple stem instances in a query for a document under the conditions of Surveys 1, 2 & 3.

The difference between the average number of **distinct stems** compared to **total stems** used was:

1. not significantly different for Survey 1 (p=0.827)
2. not significantly different for Survey 2 (p=0.948))
3. not significantly different for Survey 3 (p=0.0.874)

	Survey 1 needn't contain access		Survey 2 needn't contain no access		Survey 3 must contain access	
	Total Stems	Distinct Stems	Total Stems	Distinct Stems	Total Stems	Distinct Stems
Mean	3.564	3.643	3.844	3.822	3.931	3.881
P(T≤t) two-tail	**0.827**		**0.948**		**0.874**	

Table 6.15: Multiple Query stem usage statistics

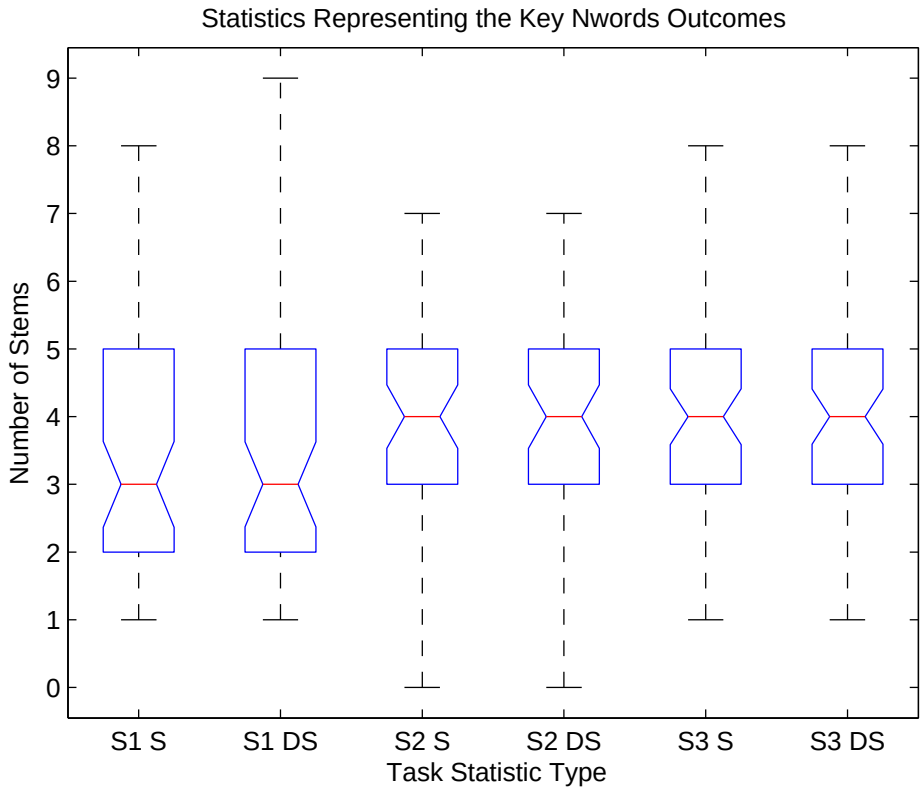


Fig. 6.16: Comparison between non-distinct and distinct query stem counts

Critical Statistics (Outliers NOT Included)						
	Survey 1 needn't contain access		Survey 2 needn't contain no access		Survey 3 must contain access	
	Total Stems	Distinct Stems	Total Stems	Distinct Stems	Total Stems	Distinct Stems
Number of Outliers Removed	1	0	2	2	4	3
Number of Participants	56	56	48	48	62	62
Min	1	1	0	0	1	1
Median	3	4	4	4	4	4
Mean	3.564	3.643	3.844	3.822	3.931	3.881
Max	8	9	7	7	8	8
Std Dev	1.844	1.967	1.623	1.628	1.715	1.662
Std Err	0.251	0.265	0.245	0.245	0.227	0.218

Table 6.16: Query stem usage statistics

6.4.2 Analysis of Combined Survey Results

In order to develop an understanding for general participant average preferences while factoring out the varying tasks environment, and to look for any unexpected correlations the following discussion looks at compound Task data. These data are formed by combining the results of each task from surveys 1, 2 & 3 with outliers removed. In quantifying the number of terms normally used, Figure 6.17 presents agglomerate task results for visual analysis while Table 6.17 presents key statistical data. From this data six key observations are made:

1. Given a median of 4 and a mean of between 3.65 and 4.324 (with 95% confidence), we can say that on average participants of Task 1 used 4 terms to describe the document in context. This should however be tempered by the observations made in Section 6.4.1.1 that suggests that participants used on average 3 terms in survey 1 as opposed to 4 terms for surveys 2 & 3. This difference is recognizable in the second quartile range being larger than the third quartile range.

2. It is difficult to make specific and high confidence observations about the average number of distinct descriptor stems used by participants for the combined results of Task 1; however, we can make some general observations. Given a standard error of 0.362 and p=0.716 we can say with confidence (0.05) that on average participants of Task 1 used 7.603 distinct stems to describe the document in context. However, as a measure of central tendency the median statistics of Figure 6.17 are more informative as they demonstrated the skewed nature of this set and indicate that participants normally use 6 (whole stems) distinct stems to describe a document.

3. Given a median of 4, mean of 3.933 and standard error of 0.148 we can say, with a confidence level of 0.291 at a probability of 95%, that on average participants across all surveys use 4 distinct stems in a query to search for the document in context.

4. The number of descriptor stems used that also occur in the top ten TFIDF weighted stems list was relatively low all falling within a tight band. Given a median of 2, mean of 2.706 and standard error of 0.136 we can say, with a confidence level of 0.268 at a probability of 95%, that on average the number of distinct stems used to describe the document in context that are also one of the top ten TFIDF weighted stems is only 2. As mentioned there is an observable tight and low tendency across all three survey contexts in this data as seen in a standard deviation of 1.622 and quartile range of between 1 and 4. This suggest

that TFIDF weightings do not match the weightings applied by participants in this context.

5. Given a median of 1, mean of 1.239 and standard error of 0.075 we can say, with a confidence level of 0.149 at a probability of 95%, that on average the number of distinct stems used in a query for a given document, that are also one of the top ten TFIDF weighted stems, is only 1. When considering this observation one should also note the standard deviation of 0.961, and the very tight mid-quartile range of 1 to 2, which implies a very tight and low normal range of average across all three survey contexts. Again, this suggest that TFIDF weightings do not match the weightings applied by participants in this context especially considering the low average number of stems and small range used in this task.

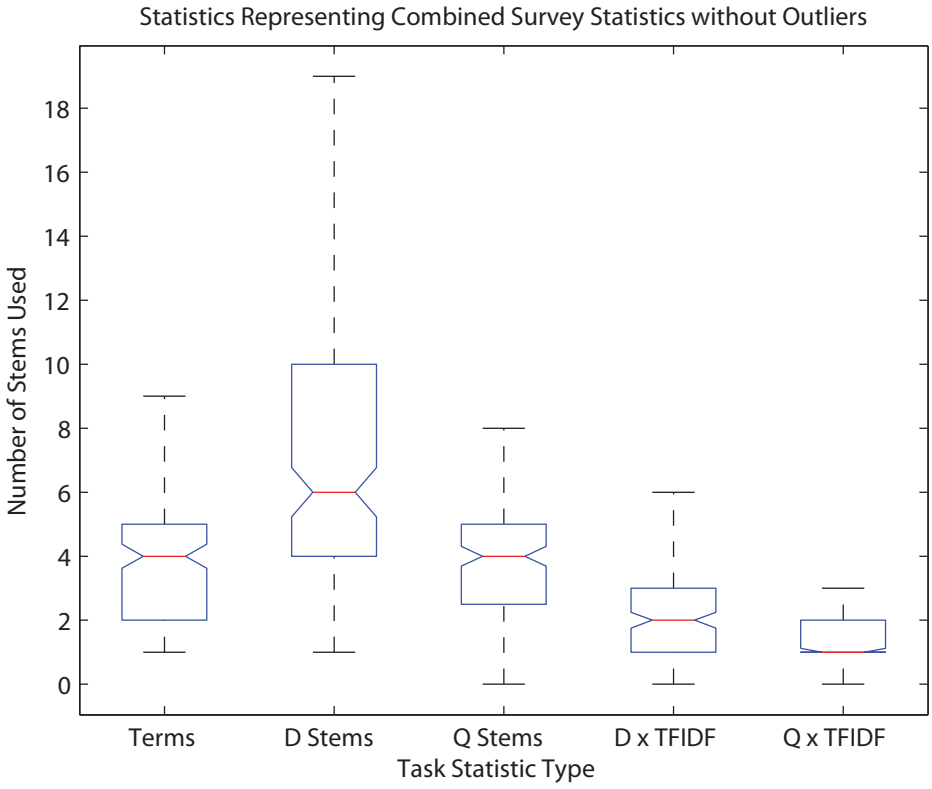


Fig. 6.17: Between task agglomerate term, description and query usage

Critical Statistics Outliers Not Included					
	Terms	Descriptor Stems	Query	Descriptor/ Top TFIDF Stem Intersections	Query/ Top TFIDF Stem Intersection

Number of Outliers Removed	8	15	6	6	2
Number of Participants	166	166	166	165	165
Min	1	1	1	1	0
Median	4	6	4	2	1
Mean	3.987	7.603	3.933	2.706	1.239
Max	9	19	8	6	3
Std Dev	2.144	4.454	1.890	1.622	0.961
Std Err	0.171	0.362	0.148	0.136	0.075
Conf. Lev. (95.0%)	0.337	0.716	0.291	0.268	0.149

Table 6.17: Agglomerate term, description and query usage statistics

Finally, as noted participants used **six** distinct stems to describe a document compared to **four** distinct stems to search for the same document. From this observation two further observations can be made:

1. Since participants use six distinct stems to describe a document and only two of them are also in the top ten TFIDF ranked stems we can make the prediction that on average only 33.33% of stems used to describe a text will also be a top ten TFIDF ranked stems.

2. Since participants use four distinct stems to query for a document and only one of them is also one of the top ten TFIDF ranked stems we can make the prediction that on average only 25% of stems used to query for a text will also be top ten TFIDF ranked stems.

These intersections further highlight the previous claim that TFIDF weightings do not match the weightings applied by participants in this context or more precisely the TFIDF weighting algorithm used in this experiment results in top ten stem lists that do not represent those lists of stems used by participants in describing and querying for documents.

6.4.3 Effects of DOCUMENT

To ensure any conclusions about the Nwords results are not in any way influenced by effects originating from variations of the documents themselves, such as differing complexities, size, number of sub-topics or any other effect, multiple ANOVA's of different types were conducted. These were used to test each of the key statistics (terms, descriptor stems & query stems) by sequentially comparing the mean for each document against the means for all other documents for the relative statistic. If there is evidence of any effects of DOCUMENT further experiments would be required to quantify their influence on the results and to identify whether corrections are possible or survey re-design is required.

Due to this multiple test approach the ANOVA's were initially conducted using a Post Hoc treatment with Bonferroni correction to reduce the probability of Type I errors (i.e. rejecting H_0 when H_0 is true). The Bonferroni correction is based on Student's t statistic and adjusts the observed significance level for the fact that multiple comparisons are made, and is applicable to finite observations. However, the results of these tests suggested there was absolutely no effect of any statistical significance, as seen in all comparisons resulting in extreme P-values of 1.000. This result was both surprising and concerning at the same time as one would reasonably expect some variance of statistical significance between all tests.

The nature of the rejection of the null hypothesis across all comparisons suggested a need to, at the very least, investigate a little further. This led to the acknowledgment that the Bonferroni approach, normally used in multiple test situations like this, is recognised (Holm 1979, Thomas, Siemiatycki, Dewar, Robins, Goldberg & Armstrong 1985, Rothman 1990, Perneger 1998, Rice 1989, Thompson 2002) as having a tendency to overly and inappropriately reduce the statistical power of rejecting an incorrect H_0 in each test. Most recently this was tested and identified, relative to the standard Bonferroni correction and the sequential Bonferroni procedure, by Jennisons and Moller (2003) as a tendency to exacerbate any existing problem of low statistical power.

Although it is common place to report only highly significant effects Nakagawa (2004) suggests that all effects should be acknowledged and that the use of Bonferroni like corrections and the practice of reviewers demanding their use should be discouraged. Alternately, Nakagawa suggests that because P-values do not indicate the degree of experimental effect present (as noted by Cohen (1990, 1994) and Yoccoz (1991)), effect sizes (confidence intervals) should be reported alongside of P-values to allow the reader to evaluate the relative importance of results and interpret non-significant results. Thus, any analysis should be rigorous such that it produces figures that give the researcher and reviewer an appropriate understanding of the data in-

volved to allow them to draw appropriate conclusions based on statistical inference and personal/professional experience.

In the spirit of rigorous treatment and given the importance of factoring out any effect of document, to make reliable conclusions a better understanding of the data was needed given the unforeseen results of the Bonferroni adjustment. As such a simple ANOVA without Post Hoc treatment was conducted for the three variables (terms, descriptor stems & query stems). This analysis again suggested that there were no significant effects, as seen in three F-values of significance greater than 0.05. However, unlike the Bonferroni results the F-values for the term count and query stem count displayed significance values much closer to the rejection point of 0.05. Subsequently, an ANOVA using a Post Hoc LSD (Least Significant Dimension) approach was conducted to see if any significant results did occurr and how they presented.

LSD (Least Significant Difference) basically uses the smallest difference between means that would be statistically significant and if the actual difference is greater than that, then results are regarded as statistically significant. It was used as it does not control the overall probability of rejecting the hypotheses that some pairs of means are different, like Bonferroni adjustment and their likes, while in fact they are equal, i.e. it doesn't matter if you are comparing 1 pair of means or a 100, no adjustment is made for the number of comparisons.

The results of these tests display two slight anomalies that might be described as non-random. In the multiple pairwise comparison of the Post Hoc analysis these were observed, when testing query stems, in the rejection of the null hypothesis 60% and 55% of the time for documents 14 and 18 respectively. This can be interpeted as indicating that when compared to the means of the other documents these means were observable different and indicating some effect of document. However, from subsequent manual inspection of document structural and general characteristics, like the number of words and paragraphs, no notable differences between these documents and the others were observable. Alongside these observations, general document statistics (see below Table 6.18) only highlighted one **expected** anomaly, in the document means query statistics, that of a skew of 1.28. This was accompanied by relatively uninteresting and normal statistics such as low standard error and deviation, and a fairly tight *logistic* style distribution (kurtosis = 1.484). Given these results it is suggested that these two document means are simply part of an expected normal distribution and do not indicate any effect of document.

	Query Stems	Terms	Descriptor Stems
Mean	3.851	4.467	9.401
Standard Error	0.188	0.274	0.556

Median	3.757	4.200	9.500
Standard Deviation	0.842	1.226	2.488
Sample Variance	0.710	1.503	6.188
Kurtosis	1.484	-1.326	1.014
Skewness	1.280	0.408	0.294
Range	3.250	3.556	11.100
Minimum	2.750	3.000	4.500
Maximum	6.000	6.556	15.600
Count	20	20	20

Table 6.18: Document Effect Query Stem Means Statistics

6.4.4 Human vs. Automated Rank Sequence

Survey Four is designed to test how closely human ranking of a set of ten top TFIDF ranked word stem derived from a given document correlates to that of the natural ordering of the TFIDF ranks themselves. The process used to identify the TFIDF weighting formula used is discussed in Section 7. The results of Survey Four are thus a number of pairs of ranked lists of the same word stems, one representing the human defined sequence and the other the sequence defined by the natural order of the TFIDF weightings.

By demonstrating a level of agreement between the lists of a list-pair it could be suggested that there is a level of implicit agreement between human judgment and the weighting scheme used to generate the list. If it can be shown that there is a significant correlation between the relative orderings (human and automated) then it can be suggested that the TFIDF weighting scheme used to generate the original list of terms closely approximates human judgment for the task of "ordering list of keywords derived from a specific document".

To compare the different sequences of each list pair Spearman's ρ was used as it is a Pearson's r (product-moment coefficient) correlation adjusted to work not on the original variables but on the variables transformed into rank-orders. It is a non-parametric measure of correlation that does not require the assumption that the relationship between the variables is linear, nor does it require variables measured on interval scales and so is appropriate for variables measured ordinally.

For each participant the Spearman's ρ correlation coefficient was calculated between the the ranked list of stems derived from the list of words supplied by the participant in Survey Four and a ranked list of top ten TFIDF stems generated using an appropriate

TFIDF calculation (see Chapter 7). The analysis of this data presents fairly clear results given the null hypothesis ("there is no association between the two ranked sequences"). From seventy-one observations only nine significant P-values ($p < 0.05$) were observed, of which only four were highly significant ($p < 0.01$). This suggests that there is on average no significant relationship between the orderings imposed by humans and those imposed by the natural ordering of the TFIDF weightings. In fact it is appropriate, given the multiple comparison nature of this test, to apply the relatively conservative Bonferroni correction to this set of comparisons. In doing so, no significant **p** values were observed at a confidence level of 95%. This goes to conclusively demonstrate that the participant did not rank words in the same manner as produced by the TFIDF calulation despite it being the most prefered ranking scheme as demonstrated in Section 7.

Spearman Rank Correlation Statistics						
N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
71	1.6819	-.8609	.8211	.1326	.4035	.163

Table 6.19: Correlations of key statistics

6.4.5 Correlations Analysis

Following are observations from a correlation analysis of all key statistics and the age, sex and experience categories. Table 6.20 presents all key values from the correlation analysis for which the following observations are made:

Sex-Experience The correlation value of 0.195 between *sex* and *experience* is significant with a P-value of 0.012 (< 0.05) where Female was encoded as zero and Male as 1, and experience took seven discrete values from one (low) to seven(high). This indicates that males are more likely to indicate they are more experienced that females.

Distinct query stems - Terms With a value of 0.349 a positive correlative relationship between the number of distinct query stem and number of terms is significant with a P-value of 4.08E-06 (< 0.05). This indicates that with higher query stems counts higher term counts will be observed.

Distinct query stems - Distinct description stems With a value of 0.355 a positive correlative relationship between the number of distinct query stems and number of terms is significant with a P-value of 2.66E-06 (< 0.05). This indicates that with higher query stem counts higher description counts will be observed.

Pearson Correlation Statistics							
Statistic Type	age category	experience level	sex category	terms	distinct description stems	distinct description stems in top ten TFIDF	distinct query stems
experience level	-0.070						
sex category	-0.095	**0.195**					
terms	-0.123	0.038	-0.012				
dist. descr. stems	-0.150	0.049	0.026	0.527			
dist. descr. stems in top ten TFIDF	-0.061	0.056	0.065	0.460	0.622		
dist. query stems	-0.124	0.005	-0.064	**0.349**	**0.355**	0.152	
query stems in top ten TFIDF	0.006	0.063	0.067	0.223	0.132	0.461	0.306
P-values							
experience level	0.369						
sex category	0.223	**0.012**					
terms	0.113	0.630	0.874				
dist. descr. stems	0.054	0.533	0.743	3.10E-13			
dist. descr. stems in top ten TFIDF	0.439	0.473	0.407	4.39E-10	3.59E-19		
dist. query stems	0.112	0.951	0.415	**4.08E-06**	**2.66E-06**	0.051	
query stems in top ten TFIDF	0.937	0.421	0.388	0.004	0.089	4.09E-10	0.00006

Table 6.20: Correlations of key statistics

6.5 Conclusion

This research has undertaken experiments the results of which cast light onto aspects of human preference in the tasks of describing documents and searching for documents. To do this the Nwords experiment has been proposed to investigate generic human preferences regarding the number of words used to describe or to search for a document. Secondly, it has been used to investigate how well a TFIDF weighting scheme commonly used to present lists of documents ordered against word ranks maps to the mental representations of humans.

The general goals of the Nwords research project as defined in Section 6.1 were:

1. to determine how many words users employ in searching for a document,
2. to determine how many words are used to describe a document topic/category (to optimize cluster descriptions)
3. to test TFIDF's cognitive validity (does TFIDF rate words similarly to the user?)

These experiments were carefully designed to manage the impact of all possible variables and despite some results being indicative of each conditions having the own effect, it would be naive not to consider that in some cases there may also be a common effect.

The rest of this Section is divided into two sections. The first, presents short discussions that state and outline important observations made. The second, presents concluding remarks relating to key observations and the goals of the Nwords research.

Important Observations

In Section 6.4.1.1 it was demonstrated that participants use a median of **four** terms to describe a document under the conditions of surveys 2 & 3, and **three** terms under the conditions of survey 1. This suggests that when the task conditions are more restrictive, as seen in the removal of document access in survey 2 and the forcing of participants to only use words from the experiment document in survey 3, participants tend to use more terms to describe the document. One possible explanation for this may be that this increased restriction forces the participant into an increased cognitive load state as they need more cognitive resources to produce descriptive terms from less information. Because of this the participant maybe placed in a state of greater uncertainty and thus may use more terms in an attempt to allow for the perceived chance of greater error. However, these remarks are purely speculative and would require further investigation.

In Section 6.4.1.2 it was demonstrated that participants normally used **five** distinct stems to describe a document under the conditions of survey 1, **seven** distinct stems under the conditions of survey 2 and **eight** distinct stems under the conditions of survey 3. Again, one possible explanation for this may be that the survey conditions result in increasing amounts of stems used to allow for the perception of a greater chance of error.

In Section 6.4.1.3 it was demonstrated that participants used, on average, **four** distinct stems to query for a document under the conditions of surveys 1, 2 & 3. However, discrepancies between the mean and median statistics were observed and an alternate observation was proposed that stated "the conditions of survey 1 **did** have an effect, resulting in participants using **three** distinct stems to query for a document not **four**". The difference between these observations has no real consequence and again, one possible explanation for this may be that the survey conditions result in increasing amounts of stems used to allow for a participant's anticipation of a greater chance of error.

In Section 6.4.1.4 it was demonstrated that participants used, on average, **two** distinct stems, that are also one of the top ten TFIDF ranked stems, to describe a document under the conditions of surveys 1 & 2, and **three** under the conditions of survey 3. This difference between survey 3 and the others can be explained by the observation that survey 3 forces the participant to only use words from the text in context, limiting their potential pool of concepts to select from, whereas surveys 1 & 2 allow the participant to use any known concepts.

In Section 6.4.1.5 it was demonstrated that participants used, on average **one** distinct stem to describe a document that is also one of the top ten TFIDF ranked stems under the conditions of all surveys.

In Section 6.4.1.6 it was demonstrated that participants are not likely to use multiple stem concepts to either describe a document or in a query for a document under the conditions of Surveys 1, 2 & 3. It is suggested however that there is in fact a tendency for participants to use multiple stem instances to **describe** the document in context. This tendency can be seen when the mean values are rounded which in both cases results in different whole number values. Rounding is conducted to realize whole numbers and thus a relative equivalence to the use of whole words. Further support for this perspective is seen in the median values that match the rounded mean values. This combination of observations suggest that in real terms participants **do** display a tendency to use multiple stem instances.

Concluding Discussion

One key aspect of Nwords was to identify how many terms or key words subjects use

to characterize or search for a document. Toward this end, it has been demonstrated that participants used 2 to 3 times the number of distinct words to describe a document than distinct words to search for the same document.

Given it has been demonstrated that participants generally use, on average, **six** distinct stems to describe a document compared to **four** distinct stems to search for the same document, two subsequent observations can be made:

1. Since participants use six distinct stems to describe a document and only **two** of them are also one of the top ten TFIDF ranked stems we can make the prediction that on average only 33.33% of stems used to describe a text will also be in the top ten TFIDF ranked stems.

2. Since participants use four distinct stems to query for a document and only **one** of them is also one of the top ten TFIDF ranked stems we can make the prediction that on average only 25% of stems used to query for a text will also be in the top ten TFIDF ranked stems.

In Section 6.1.1, I proposed that "Given researched cognitive limits such as those represented by the magic numbers 7 ± 2 or 4 ± 1 (see Section 3.1.1) and their associated chunks of information, users will have a preference for document descriptions of between 1 and 9 characterizing words (chunks)". Relative to the task of labeling clusters of documents with concise descriptors participants generally used 5 to 8 distinct stems to describe a document. This is an important observation as it implicitly supports Miller's proposed limit (see Section 3.1.1) of 7 ± 2 as being appropriate in its use as a "rule of thumb" to describe a tendency in document description formulation. Relative to the goals of this research it implies that clusters of documents should be described using 7 ± 2 different words.

It was suggested in Section 6.1.1 that Cowan's number 4 ± 1 (see Section 3.1.3) was more likely the rule applicable in the description of how many words people might use to describe a document. It has been demonstrated that this is not the case. However, the observation that participants used between 3 and 4 stems to construct a query does support the suggestions from specific Web statistics and TLA research (see Sections 5.4 & 5.5) that people tend to use between 1 and 5 single terms in a query and taken together this suggests that Cowan's number 4 ± 1 is an appropriate "rule of thumb" describing the response tendency in query formulation.

When examining the set of human query stems across all tasks it is noted that on average a minimum of one word does not occur in the description stems set. Given the small numbers of query stems normally used, it is evident that the terms used to query

for a document will be substantially different from those used to describe the same document. I propose that this is indicative of different cognitive processes being involved which in turn indicates that Miller's number and Cowan's number are heuristics that are both useful in representing human preference but in different situations.

TFIDF is generally used to describe the representativeness of textual information for a given block of text relative to an associated corpus. I propose that if TFIDF is intended to reflect human judgment in some manner then it is fair, given its ubiquity in the document retrieval field, to expect that it would exhibit a reasonable level of psychological relevance. However, given the small size of the intersections between survey participant selected terms and those generated using a TFIDF algorithm it is evident that TFIDF **does not** reflect human preference to any reasonable degree. Furthermore, is also evident that TFIDF is more representative of human preference in the task of text description as seen in participant generated description stems being substantially more likely to intersect with the TFIDF list than participant generated query stems.

During the course of this research it was recognized that there was the potential issue of participants being influenced by the different types of input fields and associated mechanisms. The InFields research (see Section 7) was conducted to ascertain if there was an effect brought about by the input fields. This research demonstrated with a high level of confidence that the input field shapes and mechanism, under the conditions described by Nwords, did not affect the number of distinct or non-distinct stems used to describe a document and likewise to search for a document.

Chapter 7

Rwords & Infields

Darius Pfitzner, Kenneth Treharne & David M. W. Powers (in press, accepted May 2008), "User Keyword Preference: the Nwords and Rwords Experiments", International Journal of Internet Protocol Technology: Special Issue on Intelligent Internet-based Systems: Emerging Technologies and Programming Techniques.

7.1 The RWords Survey

This chapter discusses the results of a paper based survey which examines how well five common variants of the TFIDF calculation match human keyword choice or preference. The survey presented the participant with four ranked lists of top ten TFIDF words generated by different TFIDF algorithms from the same text. They were asked to read the originating text and rank each list in terms of how representative of the original text the words and their ranking were.

The survey was designed to test how well five TFIDF functions rank terms compared to a human subject. There were four TFIDF variants identified. However, equation 1a (Salton & Buckley 1988) is a scaling of equation 1b (Johnson-Laird et al., 1998) and hence results are identical. Equation (2) is found in Salton (1991), equation (3) is not strictly TFIDF but TFITF (Term Frequency Inverse Term Frequency) and is found in Salton & Buckley (1988), and equation (4) is a variant of TFIDF introduced by the authors to directly scale by the relative document size. Equations (3) and (4) differ from equation (1) by addition of a constant term (add one) as well as the different scalings (which would make no difference on their own). The functions used to produce ranked lists of words are:

$$\left(\frac{f(i,j)}{max(f(*,j))}\right) \times \log\left(\frac{N}{n_i}\right) \tag{7.1}$$

$$f(i,j) \times \log\left(\frac{N}{n_i}\right) \tag{7.2}$$

$$\left(\frac{f(i,j)}{max(f(*,j))}\right) \times \log\left(\frac{N}{c_i}\right) \tag{7.3}$$

$$0.5 + \left(\frac{0.5 \times f(i,j)}{max(f(*,j))}\right) \times \log\left(\frac{N}{n_i}\right) \tag{7.4}$$

$$1 + \left(\frac{c \times f(i,j)}{c(j)}\right) \times \log\left(\frac{N}{n_i}\right) \tag{7.5}$$

$$\tag{7.6}$$

where f(i,j) is frequency of i^{th} word in document j, Ci is the frequency of word i in the corpus, n_i is the count of documents containing the i^{th} term and N is the number of documents in the corpus, $max(f(*,j))$ is the frequency of the most frequent word in document j, c is the average size of a document in the corpus and $s(j)$ is the size of the j^{th} document. In testing the agreement between the participants, several non-parametric measures were considered and Kendall's coefficient of concordance was adopted a priori as most appropriate due to the ranked data.

7.1.1 Rwords Results Statistics

It is generally desirable to structure experiments such that a parametric analysis can be performed on the results. However, in some cases like the Rwords Survey (see Section 7.1) it is not possible and non-parametric alternatives need to be investigated.

The Rwords Survey was designed to test which of five TFIDF functions is most acceptable to humans. The results of this survey required the testing of variance between k human judges assessing the results of N different objective functions the results of which require a non-parametric approach for their analysis.

In testing the agreement between the judges, two measures were considered, Friedman's two-way analysis of variance and Kendall's coefficient of concordance. These measures are similar in that they both address hypotheses concerning k ratings of N objects and they use the same χ^2 statistic for testing.

7.1.1.1 Friedman's Statistics (F_r)

Friedman's two-way test is similar to the classical balanced two-way ANOVA, however it tests only for column effects after adjusting for possible row effects so does not test

for row effects or interaction effects. Friedman's test is used when columns represent treatment objects under study, and rows represent object ratings.

The Friedman test statistic (F_r) is distributed approximately as χ^2, with $(K-1)$ degrees of freedom, where K is the number of groups, in this case TFIDF functions, in the criterion variable, from $i = 1$ to K, N being the number of objects and T_i the sum of ranks for each group (Siegel 1956). Friedman's chi-square is then computed as:

$F_r = \left[\frac{12}{NK(K+1)} \sum_{i=1}^{K} T_i^2\right] - 3N(K+1)$

which is chi-squared distributed with $K-1$ degrees of freedom. The rejection region being:

$F_r > \chi^2_{\alpha,k-1}$

The null hypothesis for this approach indicates that there are no real differences among the n objects (TFIDF functions) in which case H_0 indicates that the ranks are random for the various judges as indicated by the sums of ranks being approximately equal. This implies that the N objects are drawn from the same statistical population and thus tests the hypothesis that there is no systematic difference in the ratings. In other words if the null hypothesis is true, the judges have produced rankings that are independent of one another or that there is no agreement among the judges with respect to which is the best TFIDF function.(Siegel 1956)

Assumptions about the data:

- All data come from populations having the same continuous distribution, apart from possibly different locations due to column and row effects.
- All observations are mutually independent.

7.1.1.2 Kendall's Coefficient of Concordance (W)

Kendalls coefficient of concordance (W) is a measure of the agreement among several judges (k) for a given set of n objects (TFIDF functions). It is a normalization of the Friedman test, restricting variance from 0 to 1 and focuses on the agreement between the k judges. When the coefficient W (0 to 1) is 1 it indicates complete inter-judge agreement, while 0 indicates complete disagreement among judges. So, the null hypothesis of Kendalls test is that the ratings of the k judges are unrelated and thus they did not agree (Siegel 1956).

To calculate W, the data is first arranged into a matrix with each row representing the ranks assigned by a particular judge to the N objects (TFIDF functions). Next the sum of the ranks R_i in each column are calculated and then divided by k to find

the average rank. Each can then be expressed as a deviation from the grand mean rank with a larger deviation indicating a greater degree of association among k sets of ranks and thus the sum of the squares of these deviations is found. Once these values are calculated Kendall's W can be calculated as follows: $W = \frac{\sum_{i=1}^{N}(\bar{R}_i - \bar{R})^2}{\frac{N(N^2-1)}{12}}$ where k = number of sets of rankings, e.g. The number of judges N = number of objects (or individuals) being ranked $\bar{R}_i$ = average of the ranks assigned to the i^{th} object $\bar{R}$ = the average (or grand mean) of the ranks assigned across all objects $\frac{N(N^2-1)}{12}$ = maximum possible sum of the squared deviations i.e., the numerator which would occur if there were perfect agreement among the k rankings, and the average rankings were $1, 2, ..., N$.

The fact that the data is in rank form their values, not their ordering, are known in advance ergo the grand mean of all the rankings is known in advance. Knowing this and that because the sum of N ranks is $\frac{N(N^2-1)}{12}$ and the mean is therefore $\frac{N-1}{12}$ this can be applied to the above formula to simplify it to:

$W = \frac{\sum_{i=1}^{N}(\bar{R}_i - \bar{R})^2}{\frac{N(N^2-1)}{12}}$

To conform to problems described by Kendall (Siegel 1956) the following assumptions were made about the data:

- Data in each row have the same rank range
- All observations are mutually independent

An X^2 that is approximately distributed as χ^2 for W can be calculated as:

$\chi^2 = X^2 = K(N-1)W$

The target for the Survey was to identify which TFIDF algorithm was most accepted by humans by asking human participants to rank the different algorithms, which makes Kendall's coefficient of concordance the preferred choice over Friedman's two-way analysis of variance.

7.1.2 Rwords Results

To identify the average human preference between different TFIDF weighting schemes, 60 respondents (k) rated 4 different TFIDF word rank sequences (n) resulting in a Kendall coefficient of concordance (W) of 0.327 indicating a significance of $p < 0.001$. This allows for the observation that, with considerable confidence, the agreement among the 60 respondents is very much higher than had their ranks been random or independent.

Given this observation, the average rank sequence can be calculated and described, with a high level of certainty, as being the most likely sequence an average normal

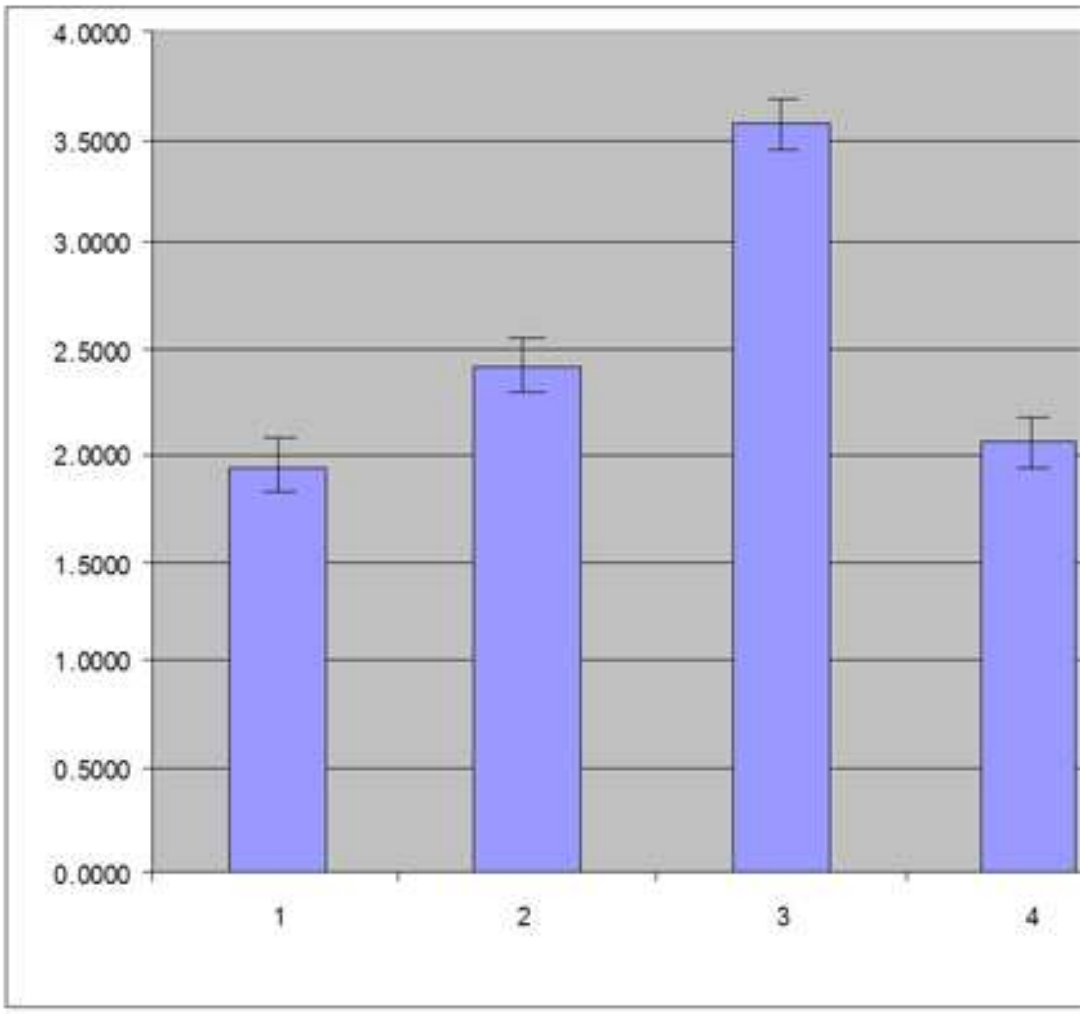


Fig. 7.1: Average ranks of four TFIDF variants (± standard error)

human would select. The results for the different TFIDF formula are graphed in Figure 7.1. Standard errors are shown concluding that there is no significant difference between equation (1) and equation (4), though there is a significant difference between the other pairs.

The results of this experiment suggest that TFIDF equation 7.1 and 7.4 performed similarly and that they performed better than equation 7.2 and far better than equation 7.3. It also informally suggested that equation 7.4 matches experienced users best and equation 7.1 matched inexperienced users.

7.2 Input Field Variants Impacting Nwords

The analysis of the Nwords survey results highlighted a possible flaw in the manner in which participants were asked to input their answers (see Section 6). It was suggested that the shape of the input fields and associated mechanisms might have influenced the number of words used to describe and query for a document. Evidently, this potential flaw might render the relevant portion of the results irrelevant to the goal of the research. To investigate this situation the "InFields" experiment was designed to describe participant word/term input characteristics under a variety of input field characteristics and task types. The primary goal of the InFields research was to determine if the different input field mechanisms used in the Nwords experiment might influence the words and terms input by participants in two common language based tasks.

Participants were asked to complete one of two possible tasks under one of two different types of input field mechanisms. The tasks were designed to replicate common keyword input and query word input activities in a highly controlled manner that held all variables constant for each participant. The different possible variables under these conditions are the document each participant is asked to read, visual characteristics of the interface, the different input field mechanisms and the task delivered to the participant. Between the participants the only dimensions varied were the question asked and the input technique used. This allowed for the identification of variance and distribution characteristics between participants of the same task and input mechanism, and the comparison between the number of words and terms used by participants of the four different survey types.

7.2.1 Survey Participants

Because the goal of this research is to determine if the different input field mechanisms used in the Nwords experiment might have adversely influenced the results, we set out to hold constant all dimensions of the original survey while only varying key element of the environment. To this end, the mixture of participants in the InFields surveys was managed to comprise approximately eighty percent undergraduate students and twenty percent equally comprised of administration staff, graduate students and teaching staff. This was the mix estimated to be approximately that of the original Nwords participant pool.

7.2.2 InFields Survey Types

The four different surveys resulted from the need to hold all interface variables constant while only varying the input field mechanism (search field or description field) and the

participant task (describe document or query for document). All four surveys required the participant to read a standard piece of text (see Appendix 10.4) and to complete one tasks. Following is a brief description of the four survey types.

Survey Type 1 (KD) Using the **Keyword input (K)** field mechanism the participant was asked the following question: "In the Description Term field below, insert as many words &/or phrases you think best **describe/represent (D)** what the text is about". This survey is presented by Figure 7.2.

Survey Type 2 (QD) Using the Query input field mechanism the participant was asked the following question: "In the Answer field below, insert as many words &/or phrases you think best **describe/represent** what the text is about". This survey is presented by Figure 7.4.

Survey Type 3 (KS) Using the **Keyword-word input (K)** field mechanism the participant was asked the following question: "In the Search Term field, insert the **Search terms (S)** you might use to find this text using an Internet or Database search engine". This survey is presented by Figure 7.3.

Survey Type 4 (QS) Using the Query-word input field mechanism the participant was asked the following question: "In the Search Term field, insert the **Search terms (S)** you might use to find this text using an Internet or Database search engine". This survey is presented by Figure 7.5.

The keyword input field mechanism emulates the conditions of common Web-based keyword input approaches, such as used by many International/National Journals and Conferences, that require users to input keywords via an "ADD" button or listing, and limited size input field. The query input field mechanism emulates a common query word input task using a wide single line input field (see Figures 7.4 & 7.5, much like that used by Google.

7.2.3 Data Treatment

This section discusses how the InFields data was treated before analysis. It outlines how raw data was cleaned, what statistics were generated and how outliers were analysis and treated.

7.2.3.1 Data Cleaning

All raw data resulting from the InFields Experiment was treated in the same manner as that of the Nwords research. Like the former research all stop-words were removed after

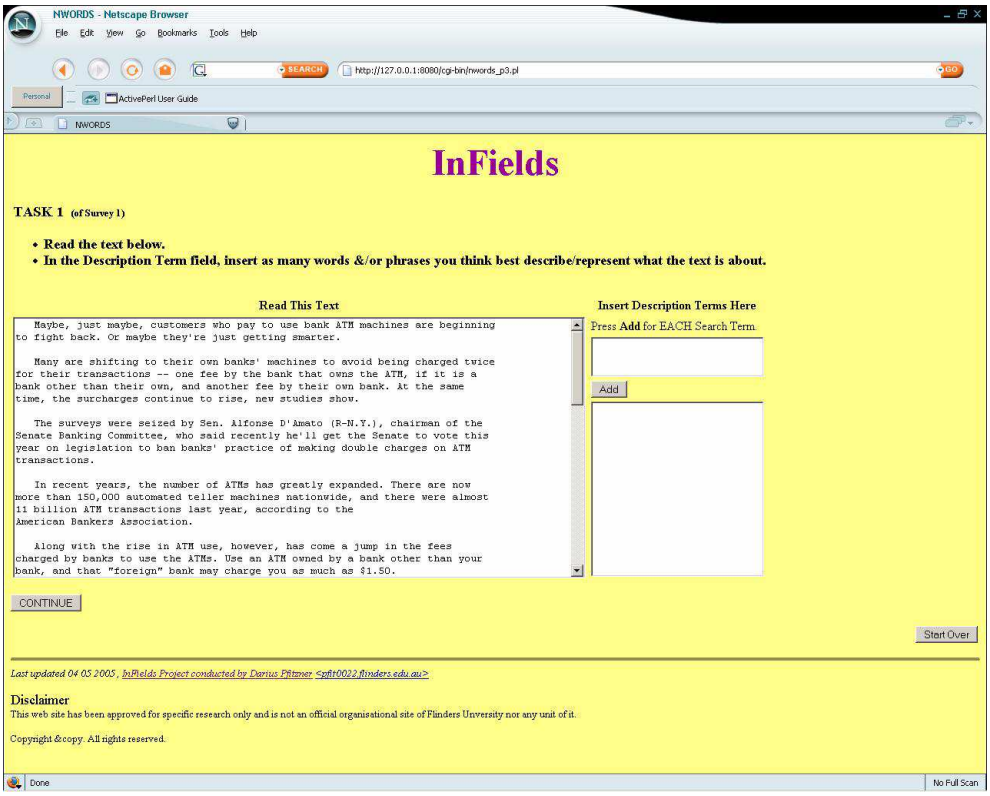


Fig. 7.2: InFields keyword input task using keyword input field

Fig. 7.3: InFields keyword input task using query-word input field

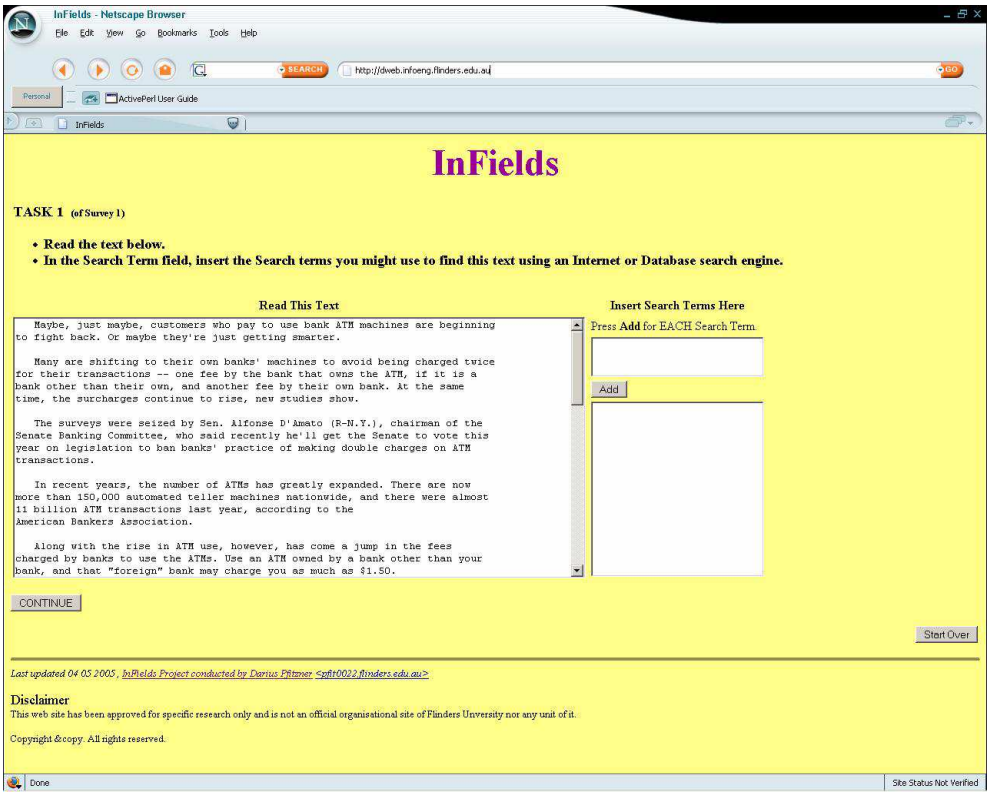


Fig. 7.4: InField query word input task using keyword input field

Fig. 7.5: InField query word input task using query-word input field

which all word stems were removed using the well known Porter stemming process.

7.2.3.2 Data Processing

To allow for appropriate comparison of results the InFields experiment data was processed into the same form as the relative Nwords data. Because we needed to compare the number of terms and words used by participants under the different input device and tasks the results were processed to generate counts of the number of:

1. Terms used (a term being one word-stem or a sequence of word-stems delimited by the use of the "ADD" button or by a comma in the query-word sequence). Presented under the column "**Terms**".
2. Words used (note that all words have been conflated in a stemming process). Presented under the column "**Stems**".
3. Distinct words used. Presented under the column "**Distinct Stems**".
4. Words used in more than one Term. Presented under the column "**Stems Intersections**".
5. Distinct words used in more than one Term. Presented under the column "**Distinct Stems Intersections**".
6. Distinct words used that also occurred in the list of top ten TFIDF stems. Presented under the column "**Distinct Stem / Top Ten TFIDF Intersection**".

7.2.4 Survey Results and Analysis

This Section presents and discusses the results of the InFields research. Each survey type subsection sequentially discusses the "statistic types" of interest from that survey type. For convenience these statistic types are associated with a number (see Section 7.2.3.2) that for cross reference purposes also occur on the x axis labels of the box-plots and table column headers presented in this section. Following is a listing of the statistic types and associated references:

Statistic 1 (T) **T**erm Count

Statistic 2 (NDS) **N**on-**D**istinct **S**tems Count

Statistic 3 (DS) **D**istinct **S**tems Count

Statistic 4 (NDSI) **N**on-**D**istinct **S**tems **I**ntersection Count

Statistic 5 (DSI) **D**istinct **S**tems **I**ntersection Count

Statistic 6 (IntTFIDF) Count of distinct stems that occur in the relative top ten TFIDF list (**Int**ersection with **TFIDF**)

7.2.5 Survey Type 1 Results Analysis

This Section discusses key characteristics of the relevant statistics of **Survey Type 1 (KD)**, as described in Table 7.1 & Figure 7.6, that combines the keyword input field with the keyword task.

Statistic 1 is characterized by a slightly skewed distribution with a median of 2, mean of 2.11 and a relatively small standard deviation. This suggests that participants normally used 2 terms (T) with little deviation from this.

Statistics 2 & 3 are combined here because of their relatively similar natures and the implications that arise from this similarity. The overlapped nature of the mid-quartile ranges and the relationships between the standard deviations and standard errors indicate that, with a high level of confidence, there is little difference between the number of distinct and non-distinct stems participants normally used. It can be said that participants normally used between 4 & 5 stems in the task and that stems are **not** normally being used in multiple terms.

Statistics 4 & 5 are combined here because of their relatively similar natures and the implications that arise from this similarity. Their skewed and small natures can be simply described as reflecting the relatively small number of terms used on average.

Statistic 6 has a skewed distribution that covers a relatively small range which suggests that with a high level of confidence it can be said that under these condition participants used on average 2 stems to describe the test document that also occurred in the top ten TFIDF stem list for that document.

	terms (1)	description stems (2)	distinct descriptions stems (3)	description stem intersections (4)	distinct description stem intersections (5)	distinct description / top 10 TFIDF intersections (6)
Mean	2.11	4.80	4.09	1.14	1.25	2.14
Std Dev	0.90	2.80	2.05	0.38	0.46	0.85
Std Err	0.21	0.66	0.48	0.09	0.11	0.20

Table 7.1: Statistics of Survey Type 1 Results

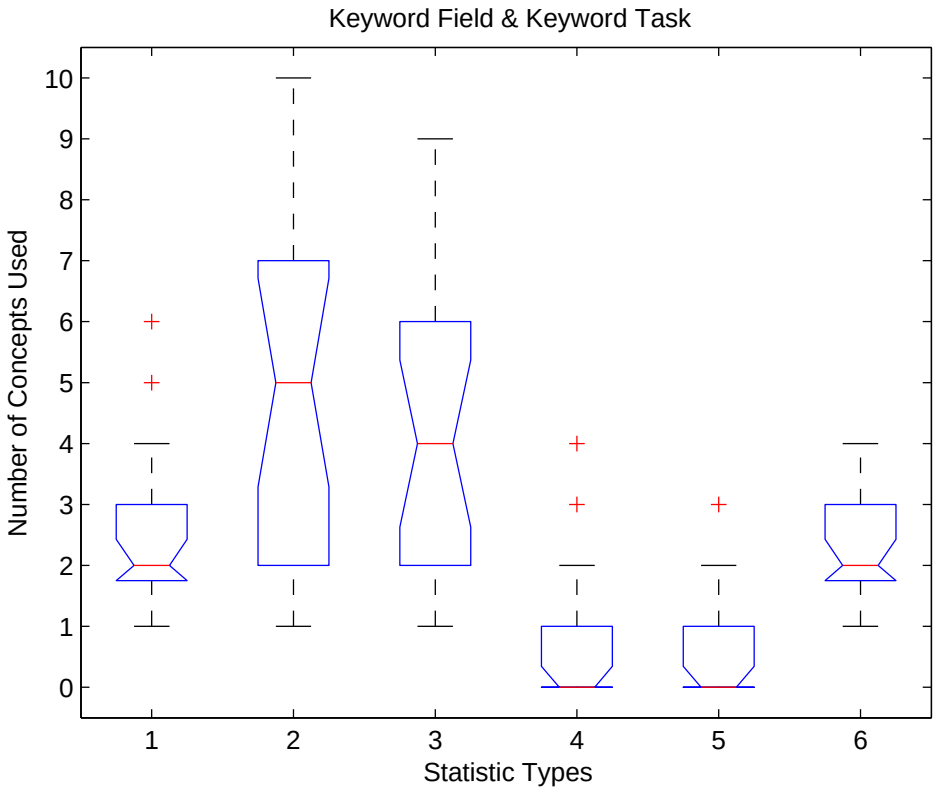


Fig. 7.6: Graphical presentation of statistics for Survey Type 1

7.2.6 Survey Type 2 Results Analysis

This Section discusses key characteristics of the relevant statistics of **Survey Type 2 (QD)**, as described in Table 7.2 & Figure 7.7, that combines the query-word input field with the keyword task.

Statistic 1 is characterized by a skewed distribution with a median of 1, a mean of 2.48 and standard deviation of 1.96. From this i can be said that participants were likely to use around 1 to 2 terms (T) in their description.

Statistics 2 & 3 are combined here because of their relatively similar natures and the implications that arise from this similarity. The overlapped nature of the mid-quartile ranges and the relationships between the standard deviations and standard errors indicate that, with a high level of confidence, there is little difference between the number of distinct and non-distinct stems participants normally used. It can be said that participants normally used between 5 & 6 stems in the task and that stems are **not** normally being used in multiple terms.

Statistics 4 & 5 are combined here because of their relatively similar natures and the implications that arise from this similarity. Their skewed and small natures can be simply described as reflecting the relatively small number of terms used on average.

Statistic 6 has a skewed distribution that covers a relatively small range. With a high level of confidence it can be said that under these condition participants used on average 3 stems to describe the test document that also occurred in the top ten TFIDF stem list for that document.

	terms (1)	description stems (2)	distinct descriptions stems (3)	description stem intersections (4)	distinct description stem intersections (5)	distinct description / top 10 TFIDF intersections (6)
Averages	2.48	6.43	5.74	3.25	2.50	2.90
Std Dev	1.97	3.71	3.21	1.75	1.27	0.79
Std Err	0.41	0.77	0.67	0.37	0.26	0.16

Table 7.2: Statistics of Survey Type 2 Results

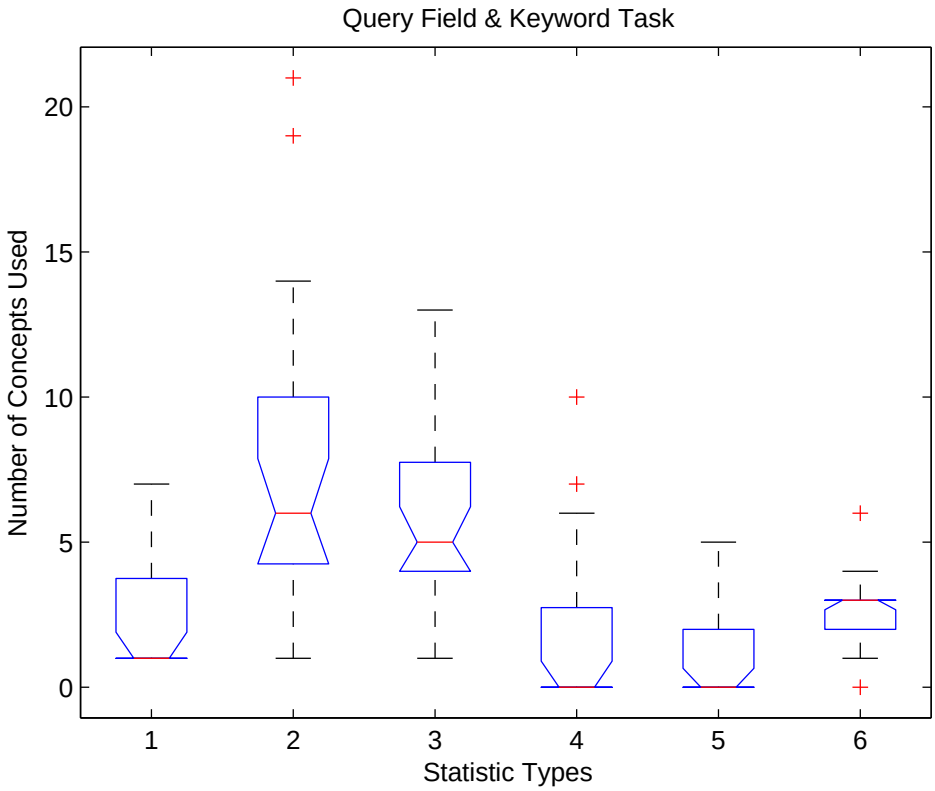


Fig. 7.7: Graphical presentation of statistics for Survey Type 2

7.2.7 Survey Type 3 Results Analysis

This Section discusses key characteristics of the relevant statistics of **Survey Type 3 (KS)**, as described in Table 7.3 & Figure 7.8, that combines the keyword input field with the query-word task.

Statistic 1 is characterized by a relatively even distributed mid-quartile with a median of 2 terms and a mean 2.48. It can be said that Participants will normally use 2 whole terms under the conditions of this test.

Statistics 2 & 3 are combined here because of their relatively similar natures and the implications that arise from this similarity. The standard deviation and long notches of the box-plots for both statistics reflects a relatively high level of variance. Given the overlapped mid-quartile ranges and closely matched notched sections it can be said with a high level of confidence that there is little difference between the number of distinct and non-distinct stems used. This indicates that stems are not being used in multiple terms.

Statistics 4 & 5 are combined here because of their relatively similar natures and the implications that arise from this similarity. The relatively small mid-quartile and variance of these results indicates that participants are likely to use on average fewer than 2 terms under these conditions.

Statistic 6 has a skewed distribution that covers a relatively small range. With a high level of confidence it can be said that under these condition participants used on average 2 stems to describe the test document that also occurred in the top ten TFIDF stem list for that document.

	terms (1)	query stems (2)	distinct query stems (3)	query stem intersections (4)	distinct query stem intersections (5)	distinct query / top 10 TFIDF intersections (6)
Averages	2.48	6.43	5.74	3.25	2.50	2.52
Std Dev	1.97	3.71	3.21	1.75	1.27	1.24
Std Err	0.41	0.77	0.67	0.37	0.26	0.26

Table 7.3: Statistics of Survey Type 3 Results

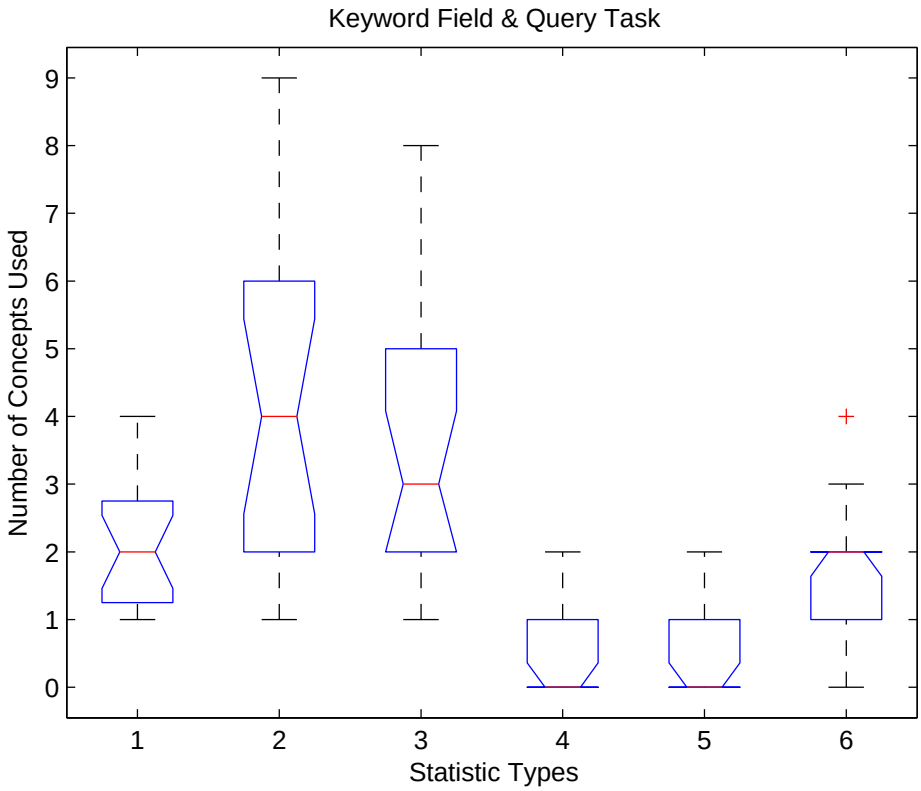


Fig. 7.8: Graphical presentation of statistics for Survey Type 3

7.2.8 Survey Type 4 Results Analysis

This Section discusses key characteristics of the relevant statistics of **Survey Type 4 (QS)**, as described in Table 7.4 & Figure 7.9, that combines the query-word input field with the query-word task.

Statistic 1 is characterized by a mid-quartile range and median of one term with a small number of outliers and a standard deviation and error of zero. This suggests that participants don't tend to use complex terms under these circumstances which is reasonable for the query task using the query interface (Google style and not a natural language style interface) given people are not being guided in any way to use anything more than a sequence of single words to search for documents. This result is in stark contrast to the other Surveys where it seems that if given the description task or the description interface under either circumstance participants seem to be encouraged to use complex concepts, as as this would suggest that the describing of a document in a single line input field, for a search, is different to using a multiple line input field or literally describing a document in either type of input mechanism.

Statistics 2 & 3 are combined here because of their relatively similar natures and the implications that arise from this similarity. In this case both statistics are basically the same especially when Statistic 2's outlier is removed for the calculation of mean, standard deviation and error. What is being presented is basically a standard normal distribution with no skew. When all the facets of this statistic are taken into account we can say that most participants are highly likely to use between three and five stems to search for this document. It can also be said that the similarity of these two statistics is an artifact of the combination of query input field and query task as was the case for Statistic 1.

Statistics 4 & 5 no observations can be made for these statistics because in both cases only one term occurred for intersections to be realized from, so these statistics are irrelevant.

Statistic 6 is skewed and covers a small range. It can be said that, with a high level of confidence participants will normally use 2 to 3 stems that also occur in the top ten TFIDF stem list for this document.

	terms (1)	query stems (2)	distinct query stems (3)	query stem intersections (4)	distinct query stem intersections (5)	distinct query / top 10 TFIDF intersections (6)
Averages	1.00	3.91	3.83	0.00	0.00	2.24
Std Dev	0.00	1.50	1.49	0.00	0.00	0.44
Std Err	0.00	0.33	0.33	0.00	0.00	0.10

Table 7.4: Statistics of Survey Type 4 Results

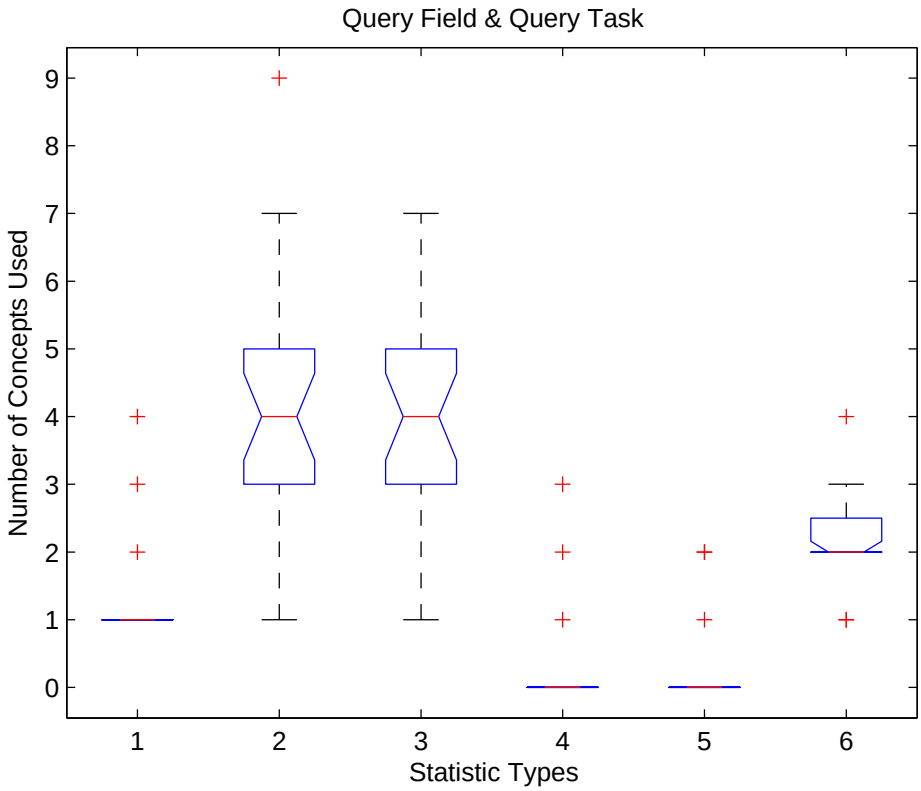


Fig. 7.9: Graphical presentation of statistics for Survey Type 4

7.2.9 Combined Results Analysis and Observations

This section discusses the key observations from the individual survey results discussion in Section 7.2.8.

Thus far general points have been discussed with a focus on the individual results for each survey. To better focus a conclusion that addresses the goal of this research

the discussion will look at the three Statistics 1, 3 & 6 (term counts, distinct stem counts and TFIDF intersections) as groups to compare and contrast the effects of the variable differences between each survey.

Following is a short analysis of results, focusing on the difference between the two different tasks. To factor out the variation of input field the results of Survey 1 & 2, and 3 & 4 have been aggregated to realize two sets of statistics differentiated by the task only.

Figure 7.10 presents Statistic 1 results of all surveys aggregated by Task type. The plot is characterized by overlapping mid-quartile ranges with clearly different medians. This suggests that, with the input field type factored out, participants are likely to use two complex terms to describe the text and only one to search for it. This can be explained as an artifact of general user tendency to use on average four distinct stems (see Section 6.4) sequentially in a single line search field to represent their query. The three outliers flag a potential difference between users characterized by English skills and/or other inherent characteristics. This opens the way for another avenue of research, that of characterizing user traits against their tendency to use more or fewer stems under different conditions.

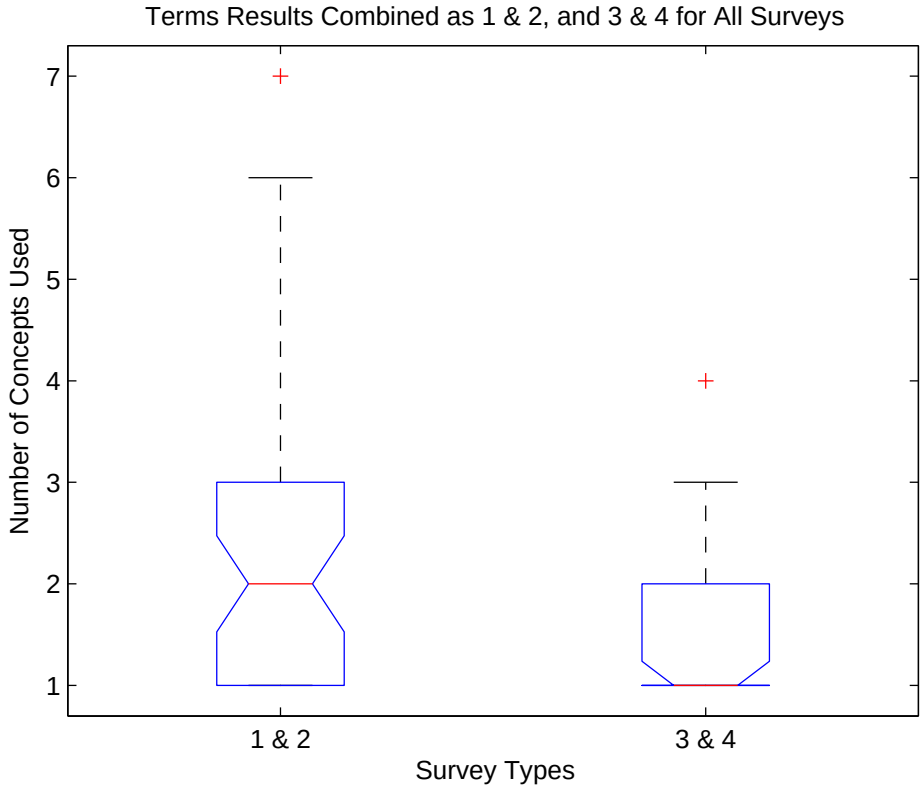


Fig. 7.10: Input Field Results Aggregated for Statistic 1

Figure 7.11 presents Statistic 3 results of all surveys aggregated by Task type. The plot describes overlapping in the mid-quartile ranges. However, the error regions for the description word task and the query word task do not overlap indicating that the medians for any two samples will be different.

This suggests that when the difference in input mechanism is factored out through aggregation the tasks result in different amounts of stems used which in turn supports observations made in the Nwords experiment.

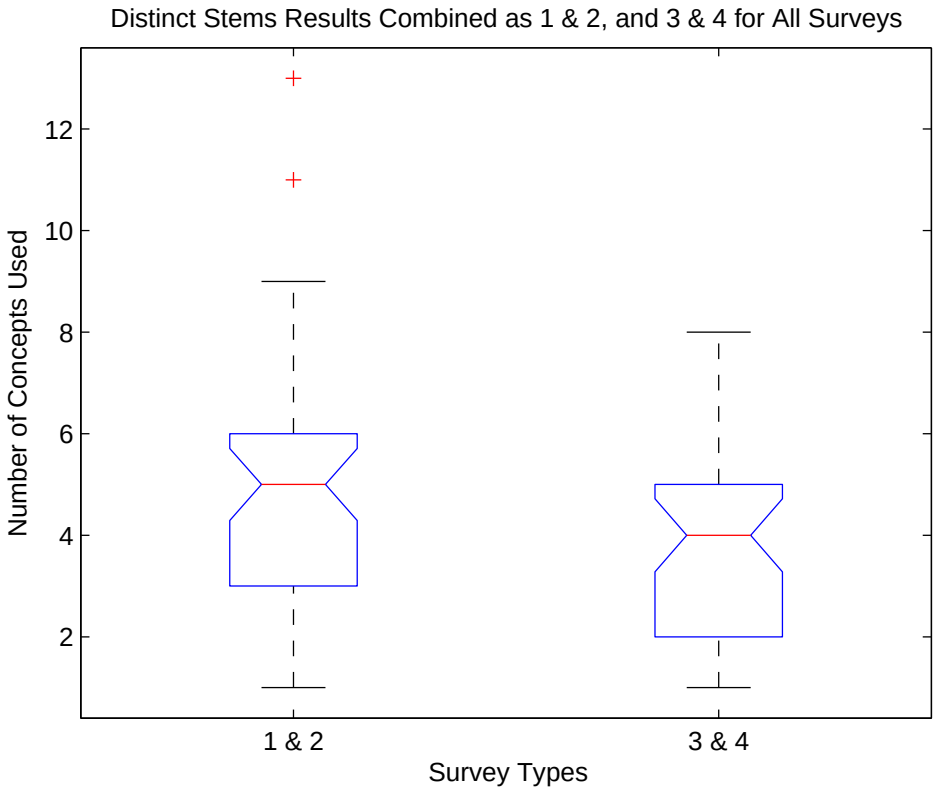


Fig. 7.11: Input Field Results Aggregated for Statistic 3

Figure 7.12 presents Statistic 6 results of all surveys aggregated by Task type. The plot is characterized by mid-quartile ranges that do not overlap and thus by different medians. From this we conclude that under these conditions participants will use a different number of stems, that intersect the top ten TFIDF list, to describe a text than they will to query for the same text.

This supports the observation and subsequent conclusions made in the Nwords experiment that, for a variety of texts, participants tended to use more stems that intersected the top ten TFIDF list, for the **description task** than they did for the **query task** (see Section 6.4).

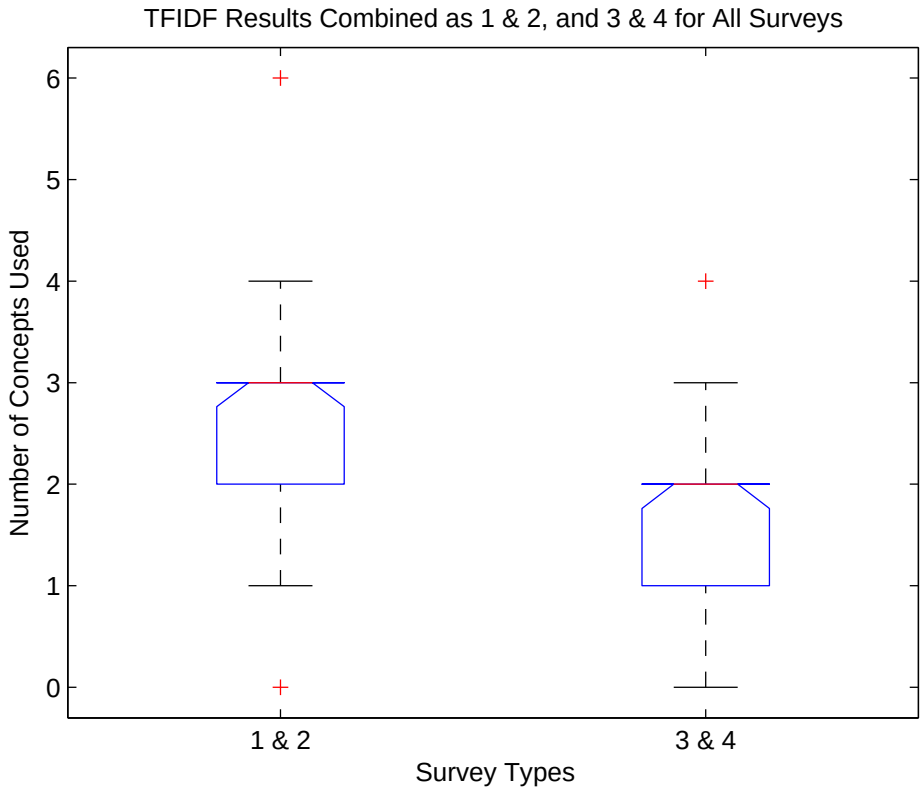


Fig. 7.12: Input Field Results Aggregated for Statistic 6

Following is a listing of the observations made in Section 7.2.8.

- Surveys 1 & 2's Statistic 1 both displayed a high level of skewing in a similar range (2-3) and although Survey 1 has outliers, Survey 2 has an upper range that would include these outliers. This suggests that there is no real difference between these statistics of the two surveys. This might be supported by an expansion of the survey to test if the outliers eventually normalize.

- Surveys 1, 2 & 3's Statistic 1 mid-quartile ranges overlap for all three Surveys indicating that with a high level of confidence it can be said that under these conditions participant responses will be the same. In other words they are likely to use between 1 and 3 (inclusive) terms.

- The tendency of participants to use several terms in Survey 1, 2 & 3 indicate that participants tend to describe a document using multiple compound concepts and not just a sequence of singular descriptors/keywords.

- Surveys 1, 2 & 3 indicate that, with a high level of confidence, there is little variance between the number of distinct and non-distinct stems used. This indicates that stems are not being used in multiple terms which is further supported by statistics 4 and 5, of all surveys, that indicate the low occurrence of non-distinct and distinct stem intersection.

- Surveys 1, 2, 3 & 4 all suggest that, with a high level of confidence, the participants will use between two and three stems to either describe or search for a document that also occur in the a list of top ten TFIDF stems. This suggests that the TFIDF weighting scheme is only moderately representative of the weighting users might use. This supports conclusions made in the Nwords experiment.

- Statistic 1 of Survey 4 suggests that participants don't tend to use complex terms under the circumstances of this survey. This result is in stark contrast to the other Surveys where if participants were given the description task or the description interface that participants seem to be encouraged to use complex concepts, as if the describing of a document in a single line input field, for a search, is different to using a multiple line input field or literally describing a document in either type of input mechanism. This is evidence enough to argue that the simplistic input field and query task combined do alter the nature of participant responses compared to the alternatives. This again raises the suggestion made in the discussion of Nwords (see Section 6.4) that:

  > Participants may interpret the multiple line input boxes as an implicit requirement to be more thorough than when replying to the single

input box. Following this it could be suggested that simply giving search engine users a bigger box may encourage them to provide a more detailed query.

This can now be extended to the suggestion that by supplying a different input field mechanism and or by coaching search engine users in their query technique (e.g. suggesting they use more complex terms or be more descriptive) they are more likely to input more complex information which can be used to better target more relevant documents to return.

Following is a short analysis of results, focusing on how the two input field mechanisms affected the outcome of the two different tasks. To factor out the variation of task the results of Survey 1 & 3, and 2 & 4 have been aggregated to realize two sets of statistics differentiated by the input mechanism only.

Figure 7.13 presents Statistic 1 **(T)** results of all surveys aggregated by input field mechanism type. The plot is characterized by similar mid-quartile ranges with clearly different medians. With a high level of confidence, we can say that the different input mechanisms were the key factor in participants normally using different amounts of terms in the description task and query task. This compounds support for the suggestion that the input field mechanisms influence the number of terms participants use.

This is evidently the first part of the answer this research was designed to elicit, that is "did the different input field mechanisms used in the Nwords experiment influence the words and **terms** input by participants in two common language based tasks". We can say **YES** the input mechanism did influence the number of **terms** used by participants. This however has no effect on any critical conclusion made in he Nwords experiment.

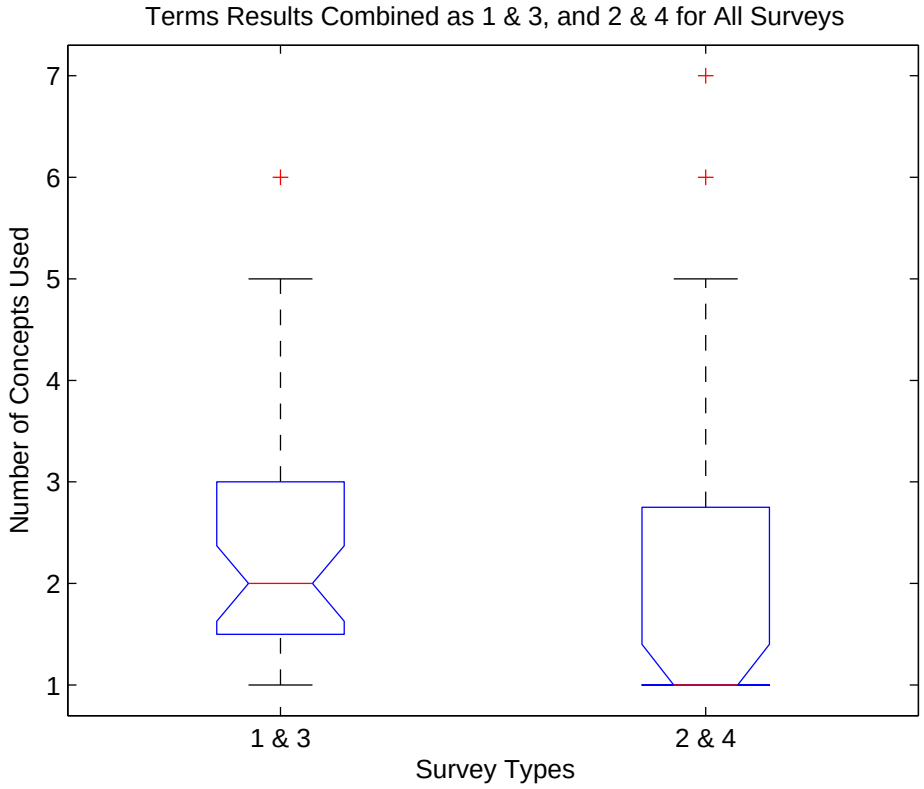


Fig. 7.13: Task results aggregated for Statistic 1

Figure 7.14 presents Statistic 3 **(DS)** results of all surveys aggregated by input field mechanism type. The plot is characterized by very similar mid-quartile ranges with the same upper and lower bounds and the same medians. With a high level of confidence we can say that the different input mechanisms had no effect on the number stems used to describe or query for the text.

We can now answer the second and key part of the answer this research was designed to elicit, that is "did the different input field mechanisms used in the Nwords experiment influence the **words** and terms input by participants in two common language based tasks". To this we can say **NO** the input mechanism did **NOT** influence the number of stems used by participants to describe or query for a text (where stems represent words of a common meaning).

This, much to the author's relief, supports the key findings of the Nwords experiment.

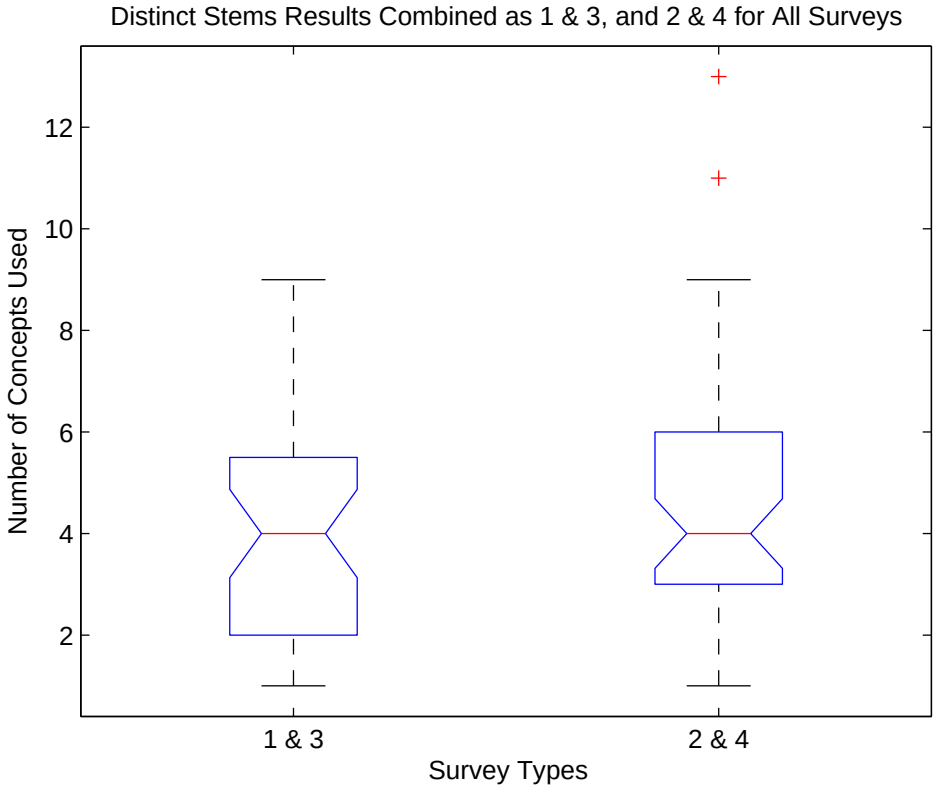


Fig. 7.14: Task Results Aggregated for Statistic 3

Figure 7.15 presents Statistic 6 **(IntTFIDF)** results of all surveys aggregated by input field mechanism type. The presentation is characterized by one plot whose median range is evenly distributed and one that is highly skewed. In addition, we note that the ranges are in fact wholly overlapped and that the two medians are the same. From this we can say that, with a high level of confidence, the different input mechanisms did not affect the number of stems that intersected the top ten TFIDF list that participants use for each Task.

This again, supports the key findings of the Nwords experiment.

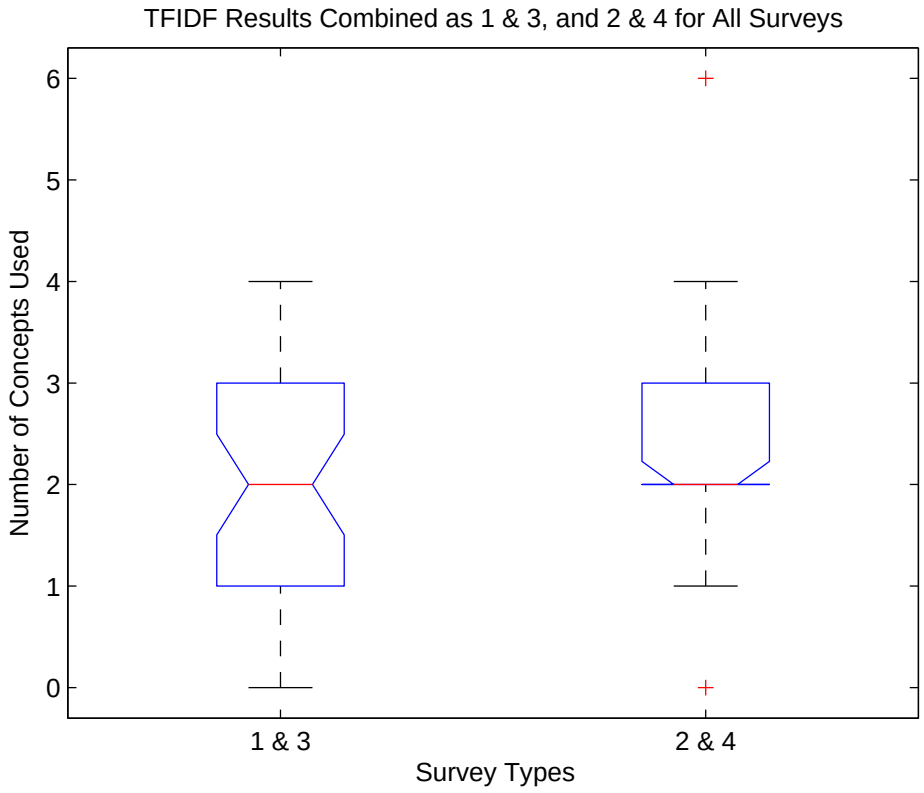


Fig. 7.15: Task Results Aggregated for Statistic 6

7.3 Rwords & Infields Research Conclusions

The Rwords experiment demonstrated that when presented with lists derived from four different TFIDF algorithms participants clearly preferred two approaches. The results indicated that TFIDF equation 7.1 and 7.4 performed similarly and that they performed better than equation 7.2 and far better than equation 7.3. It was also informally suggested that equation 7.4 matches experienced users best and equation 7.1 matched inexperienced users.

The InFields experiment resulted in three important conclusions. The first was that the input mechanism **did** influence the number of **terms** used by participants. However, although an important observation, this is of **no** consequence to any conclusions made in the Nwords experiment. Secondly, in support of the Nwords findings, the input mechanism did **not** influence the number of distinct stems used by participants to describe or query for a text. Finally, and again in support of the Nwords findings, the different input mechanisms did **not** affect the number of stems participants use that also intersected the top ten TFIDF list.

Chapter 8

Comparing Pairs of Clusterings

Research and results contained within this Chapter have been accepted for publishing in the following peer reviewed publication:

> Darius Pfitzner, Richard Leibbrandt and David Powers (in press), "Characterization and evaluation of similarity measures for pairs of clusterings", Knowledge and Information Systems, published online July 05, 2008 (http://dx.doi.org/10.10 07/s10115-008-0150-6).

8.1 Introduction

In the context of interactive document search, clustering documents based on underlying similarities is an appropriate technique facilitating the visual presentation of search results as argued in previous work by Pfitzner et al. (Pfitzner et al. 2003, Pfitzner & Powers 2004). The visual representations of these clusters would be annotated with textual labels that describe the contents of the clusters. The previous chapters have made progress toward quantifying the number of terms each textual label should be comprised of. The choice of terms would rely on the particular clustering algorithm used. In order to apply the findings of the previous Chapters an appropriate clustering algorithms needs to be chosen. This raises an important theoretical question regarding the evaluation of different algorithms and how this might be conducted.

If search return documents are to be presented for user interactive context filtering the clusters need to approximate the user's selection model as closely as possible. Until further research involving the data from the Nwords research can identify better techniques for cluster realization based on limited sets of descriptive words current clustering approaches need to be assessed for their applicability. This will allow optimal

cluster realization in the short term and supply a set of optimal standard applications that can be used for comparison purposes in future research. This Chapter begin to address this need.[1]

The work presented here characterizes a number of similarity/dissimilarity measures as applied to the context of comparing a pair of clusterings. These include measures previously proposed for this problem, and a host of other similarity/dissimilarity measures that, although they have not previously been applied to clustering comparison, are applicable. Subsequently a novel comparison measure, the *Measure of Concordance* (MoC) which addresses a number of shortcomings of existing measures, is introduced and its behaviour characterised against a number of other similarity/dissimilarity measures.

To help avoid confusion please note that the words *clusters* (data point groupings of a clustering) and *clustering* (the set of clusters that result from a clustering process) may be used in in close proximity to each other in this document.

Cluster analysis is a fundamental technique in the analysis of data across a broad spectrum of disciplines. Clustering is simply a process in which the members of a data set are divided into groups such that the members of each cluster(group) are sufficiently similar to infer they are of the same type and the members of the separate clusters are sufficiently different to infer they are of different types. The comparison of members within a data-set is normally achieved by assigning a vector of binary or numeric attributes to each member. In hard clustering, attributes are then used to compare each member to all the other members through the application of a threshold probability measure (either fixed or dynamically generated) which determines the similarity or dissimilarity between members of a cluster or between a member and the central point of a cluster.

Clustering algorithms embody the logic of forming data sets into a collection of nonempty subsets so that members in the same subset are more similar (cohesion) than members that come from different subsets (separation). The problem inherent in this process is that the maximization of cohesion and separation often causes conflict as the distance function may separate members that should be together and vice versa. When this is a likely scenario an arbitrating process such as the use of a template or gold standard may be used to help the demarcation process.

The usefulness of cluster analysis in eliciting groupings within data sets has seen extensive research and development into clustering algorithms and distance metrics. Compared to this there is relatively little research and development into measuring the

[1]The research presented in this Chapter was conducted in collaboration with Richard Leibbrandt and has been published as Pfitzner et al. (2008a)

similarity between two clusterings. In many applications the idea of clustering appears to be a useful technique used to reflect human intuitions and physical, biological or social associations or laws. A particularly challenging problem is that of introducing human understanding, interpretation and biases (context) into the problem of visualizing and interacting with data arising in an information retrieval task (such as web search). In such situations human judgment is the relevant standard for the measurement of the relevance of any clustering. This raises the question of how to compare the performance of common clustering techniques and distance measures against a human generated Gold Standard. There is a distinct lack of research and development into techniques for the comparison of clusterings (partitionings), although a small but significant amount of research and development has been done, as seen in work by Rand (1971), Fowlkes and Mallows (1983), Arabie and Boorman (1973) and Meila (2003).

More research into the comparison of clustering pairs, such as between a human-generated clustering and one automatically generated from the same set, would be timely, as techniques that optimize clustering by manipulating input parameters and/or the clustering algorithm are being employed on an increasing basis. Many of these techniques compare clusterings to other automatically generated clusterings or to a Gold Standard partitioning and often need to achieve this independently of the production algorithm or distance metric used. To do this, human-generated clusterings need to be compared to those generated automatically, a more complex task than the typical bipartite comparison of clusters. Bipartite comparison is a simple population to population correlation test whereas in the comparison of clusterings there is an extra dimension to account for in the comparative process. Cluster validation methods focus on defining cluster *cohesion* and *separation* via distance measures to represent the quality of groups of clusters; however, in comparing clusterings, the correlation between the total set of clusters as well as the individual cluster memberships needs to be considered without knowing a priori which should correspond or even having any constraint on the number of clusters matching.

8.2 Clustering Comparison Background

Association measures have been well researched and used since the late 1800's (see Section 8.2.3) to measure relative association between variables. In proposing the new measure MoC this paper looks closely at a limited but key set of association measures proposed within this period. MoC is designed to represent the difference between clustering pairs (partitions) as opposed to cluster pairs (two individual divisions of two separate partitions).

Several measures have been suggested for use in the comparison of clustering pairs. These measures can be used to compare how well different data clustering algorithms perform on a set of data. Measures are commonly summarized using a generalized 2×2 contingency (alternatively, matching or confusion) matrix to facilitate comparisons between measures. This research combined this approach with a pair counting approach, to populate the 2×2 contingency matrix (see Section 8.2.1), a convenient way to summarize the relationships between the memberships of two subclusters. Contingency tables can also be used in both asymmetric and symmetric situations as the key relationships in the contingency table can be assessed bidirectionally (see Section 8.2.1).

8.2.1 Contingency Tables & Pair Counting in Cluster Comparison

Pair counting was first applied scientifically by Thurstone (1927) through his *Law of Comparative Judgment* and is a mathematical representation of a discrimination process. These processes see comparisons made between pairs of a collection of entities with respect to the magnitudes of attributes, traits, and the like. To apply a pair counting approach to the traditional contingency matrix, firstly all the members of one clustering are incrementally paired. These pairs are then compared to all the similarly paired members of the other clustering. Using different relationships between the two member pairs of the partitions, the values in the contingency matrix are assigned as follows:

Given the partitionings P and Q of the data set D I first define data set pairs $Pairs_D$ as all the pairs realizable from the complete data set. Second, clustered pairs $Pairs_P$ and $Pairs_Q$ are those pairs of members from D that cluster together in P and Q respectively. Using $Pairs_P$, $Pairs_Q$ and $Pairs_D$ the values of the four quadrants of the contingency matrix are realised as:

$a = |Pairs_P \cap Pairs_Q|$, i.e. member pairs that occur in both partitions.

$b = |Pairs_Q \setminus Pairs_P|$, i.e. member pairs that occur in $Pairs_Q$ and **not** $Pairs_P$.

$c = |Pairs_P \setminus Pairs_Q|$, i.e. member pairs that occur in $Pairs_P$ and **not** $Pairs_Q$.

$d = |Pairs_D \setminus (Pairs_P \cap Pairs_Q)|$, i.e. non-member pairs that do **not** occur in any clusters in either partition.

We also define

$n = |Pairs_D| = a + b + c + d$

So for example given the set,

$D = \{1, 2, 3, 4, 5, 6\}$

and the partitions,

$P = \{1, 2, 3\},\ \{4, 5\},\ \{6\}$, and

$Q = \{1, 2, 4\},\ \{3, 5, 6\}$

then,

$Pairs_P = \{(1,2),\ (1,3),\ (2,3),\ (4,5)\}$,

$Pairs_Q = \{(1,2),\ (1,4),\ (2,4),\ (3,5),\ (3,6),\ (5,6)\}$, and

$Pairs_D = \{(1,2),\ (1,3),\ (1,4),\ (1,5),\ (1,6),\ (2,3),\ (2,4),$
$(2,5),\ (2,6),\ (3,4),\ (3,5),\ (3,6),\ (4,5),\ (4,6),\ (5,6)\}$

so,

$a = |Pairs_P \cap Pairs_Q| = |(1,2)| = 1$

$b = |Pairs_Q \setminus Pairs_P| = |(1,4)(2,4)(3,5)(3,6)(5,6)| = 5$

$c = |Pairs_P \setminus Pairs_Q| = |(1,3)(2,3)(4,5)| = 3$

$d = |Pairs_D \setminus (Pairs_P \cup Pairs_Q)| =$
$|(1,5)(1,6)(2,5)(2,6)(3,4)(4,6)| = 6$

$n = |Pairs_D| = 15 = a + b + c + d$

As noted earlier, contingency matrices can be used in both symmetric and asymmetric situations. In the symmetric case, there is no gold standard and so no predictive data; only the similarity of the two partitions can be measured. However, in the asymmetric case because there is a fixed target (Gold Standard) which allows for certain predictive observations to be made about the system that created the partition. In Table 8.1, each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class (Gold Standard). By comparing the actual against the predicted it is easy to see if a system is confusing two classes (i.e. commonly mislabeling one as another).

		Predicted (Cluster)		
		Pairs **in** P	Pairs **not in** P	Total
Actual (Gold Std)	Pairs **in** Q	a	b	a+b
	Pairs **not in** Q	c	d	c+d
Total		a+c	b+d	a+b+c+d=n

Table 8.1: Contingency matrix example with or without a Gold Standard (GS)

For those more comfortable with *confusion matrices* an equivalent has be supplied (see Table 8.2) to assist in translation.

	Predicted (Cluster)		Total
Actual (Gold Std)	True Positive	False Negative	a+b
	False Positive	True Negative	c+d
Total	a+c	b+d	a+b+c+d=n

Table 8.2: Alternate translation matrix

8.2.2 Clustering Comparison Criteria

Clustering algorithm selection or development aside, once a set of clusters has been realized there remains the question of quality of membership assignment relative to the initial purpose for the clustering. These techniques either treat *internal criteria*, *external features* or *relative criteria* (Halkidi, Batistikis & Vazirgiannis 2001). The relative and internal criteria approaches use Monte Carlo methods (Theodoridis & Koutroubas 1999) to evaluate whether a clustering is significantly different from chance, whereas external features are used to compare the memberships and structures of two clusterings. In this paper external criteria are used exclusively.

Internal criteria are quantities that involve the vectors of the data set themselves (e.g. proximity matrix). They are used to assess either the clustering itself or its producing algorithm by measuring characteristics like cohesion, separation, distortion and likelihood. Because these criteria are greatly affected by parameters defined *a priori*, such as number of clusters required or minimum density, internal criteria are thus sensitive to both the quality of the clustering and the *a priori* criteria used for evaluating them.

Relative criteria are used to rate a clustering by comparing it to other clusterings, produced by the same algorithm with different input parameter values. In this predefined criteria are selected to suit the algorithm and data set.

External features are used to simply measure how similar a clustering is to another clustering, gold standard or desirable-feature template and as such produce measures independent of the producing algorithm and *a priori* clustering evaluation, data set, or problem specific criteria.

In addressing the choice and comparison of clustering approaches Rand (1971) looked at clustering function characteristics and posed four questions:

1. How well does a method retrieve natural clusters?
2. How sensitive is a method to perturbations of the data?
3. How sensitive is a method to missing individuals?
4. Given two methods, do they produce different results on the same data?

Since clustering similarity/dissimilarity is not simply a comparison of two populations via some distance, membership or algorithm traits, the question of "what does it mean to compare clusterings?" must be answered.

Furthermore, when comparing clustering pair similarity without the use of a gold standard or desirable feature template, comparison measures will only be quantitative. This is to say they will not determine the degree of "goodness" regarding the clustering or its member clusters which is normally introduced through gold standards or desirable feature templates. I will thus develop a desiderata for cluster comparison methods based on external features they have in common (see Section 8.3).

8.2.3 Common Approaches in Comparing Clusterings

This section discusses two approaches commonly used in the comparison of clustering pairs. As clustering is one of the key techniques used in the exploration of data it stands to reason that one might want to compare the results of different approaches applied to the same data set for optimization, quantification or qualification purposes. The principal approaches used in clustering comparison can be described through their development of criteria, of which there are two main approaches: *pair counting* and *information theoretic*. This section briefly discusses these clustering comparison approaches. To assist in these discussions the following definitions are made:

- P represents the Left clustering
- Q represents the Right clustering
- I is the number of clusters in P where i indexes the clusters
- J is the number of clusters in Q where j indexes the clusters
- f_{ij} is the number of items in the ij^{th} fragment (the intersection of the i^{th} cluster of P & the j^{th} cluster of Q

- p_i or $f_{i.}$ is the number of items (cardinality) in the i^{th} cluster in P where $p_i = \sum_{j=1}^{J} f_{ij}$
  Note that in relation to Table 8.1 the following is true $\sum_i p_i$ = a + c.

- q_j or $f_{.j}$ is the number of items (cardinality) in the j^{th} cluster in Q where $q_j = \sum_{i=1}^{I} f_{ij}$
  Note that in relation to Table 8.1 the following is true $\sum_j q_j$ = a + b.

- n = number of items in the clustered space

8.2.3.1 Pair Counting Approaches in Clustering Comparison

As discussed previously (see Section 8.2.1) pair counting has been applied in this research to represent the relationships between the memberships of subclusters to judge how many member pairs two clusterings have in common. Following are broad discussions about the key techniques that use the pair counting approach. To assist the specific discusions of the Fowlkes and Mallows, and Rand measures these three key definitions are made:
$T_K = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij}^2 - n$

$P_K = \sum_{i=1}^{I} f_{i.}^2 - n \;=\; \sum_{p=1}^{P} p_i^2 - n$

$Q_K = \sum_{j=1}^{J} f_{.j}^2 - n \;=\; \sum_{q=1}^{Q} q_j^2 - n$

Fowlkes and Mallows Fowlkes and Mallows (1983) published the derivation of a measure of association proposed to describe the similarity between two hierarchical clusterings.

This measure was designed to represent the similarity of two trees at each level of a clustering. It ranges between 0 (maximum dissimilarity) and 1 (maximum similarity), measuring the association between two partitions of objects. Using a co-occurrence matrix to count the intersections at each level of two hierarchical trees it generates a sequence of values from which the differences are plotted. It is therefore an accumulation of the intersection counts for all relative levels of two hierarchical clusterings of the same data. Alternatively, it represents the multiple measures of similarity between the different levels of clustering and is expressed as B_k such that

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}} \tag{8.1}$$

Rand Rand (1971) proposed the measure R_K that in his words "essentially considers how each pair of data points is assigned in each clustering". R_K is described as the ratio of the sum of the number of pairs of members that occur in the same cluster in both clusterings and the number of pairs of members that don't occur in the same cluster in either clusterings compared to the total number of pairs. From this it can be said that R_K is the probability that two objects are treated alike in both clusterings.

$$R_K = \frac{\left[T_K - \frac{1}{2}P_K - \frac{1}{2}Q_K + \binom{n}{2}\right]}{\binom{n}{2}} \quad alt. \quad = \frac{a+d}{n} \tag{8.2}$$

Despite conducting fairly rigorous Monte Carlo sampling experiments to capture the characteristics of R_K and test its utility in comparing clustering methods Rand did not formally derive any properties for R_K. Fowlkes and Mallows(Fowlkes & Mallows 1983) on the other hand did derive moments of R_K for the assumptions of fixed margins, $f_{i.}$ and $f_{.j}$, and random allocation of matching counts of objects to f_{ij}.
The Rand index has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same. A problem with the Rand measure is that the expected value of two random partitions does not take a constant value (say zero).

Wallace Wallace's asymmetric criteria B^I and B^{II} (Wallace 1983) represent the probability that two data points in a cluster in one partitioning are also in the same cluster in another partitioning and are defined as;

$$B^I = \frac{a}{\sum_i^I p_i(p_i-1)/2} \qquad B^{II} = \frac{a}{\sum_j^J q_j(q_j-1)/2}$$

Precision, Recall & F Measure Another way of comparing partitionings is to use the well known *precision* and *recall* measures. For a gold standard P, then:

$$Precision \ = \ \frac{a}{\sum_i^I p_i(p_i-1)/2} \qquad Recall \ = \ \frac{a}{\sum_j^J q_j(q_j-1)/2}$$

Clearly in this case recall equals Wallace's B^I and precision equals Walace's B^{II}. A symmetric measure that combines precision and recall is the *F measure*, which is equivalent to Dice's measure (Dice 1945) and is defined as:

$$\frac{2a}{2a+b+c}$$

Other Pair Counting Measures Some researchers have proposed quantitative measures to express the degree of similarity between two clusterings. This section reviews the most influential of such work.

The use of the contingency matrix in the last three discussed measures opens up the possibility of linking cluster comparison with the larger field of association measure research. Association measures and the 2×2 contingency matrix are ubiquitous in experimental work in a range of scientific disciplines other than data clustering and data mining, e.g. psychology, biology, climatology, etc, and present an important body of research that may be drawn on. One common experimental design is to obtain categorical data from a group of entities (human subjects, animals, physical apparatus, etc.) on two different variables X and Y, where the variables can take on one of two values ($X1$ and $X2$, or $Y1$ and $Y2$). The data items are then allocated to each of the four categories $X1 - Y1$, $X1 - Y2$, $X2 - Y1$, $X2 - Y2$, and the contingency matrix displays the counts of items in each category.

In a related design, two entities or groups of entities (e.g. two biological species) are measured for the presence or absence of a set of features. The cells of the table then display the number of features common to both groups, the number of features present in the first group but not shared by the second, and vice versa, and the number of features not displayed in either group.

Contingency tables are also used in the field of animal learning, where predictable co-occurrence of two stimuli X and Y allows the animal to predict Y when presented with X, and vice versa. The table displays the number of occasions on which X is followed by Y, the number of occasions when X is absent and then Y occurs, etc.

The different scientific disciplines in which these experimental designs originated have produced a variety of quantitative measures based on the 2×2 contingency matrix, all of which essentially express the degree of similarity between the category on the columns and the category on the rows (or in the animal learning case, the degree to which X and Y are synonymous).

The case of clustering comparisons is a close fit with the second model discussed above (that of comparing two entities based on shared features). For example, if one considers that the two entities in question are two clusterings, and that the "features" that they either possess or lack correspond to the clustering together or not clustering together of two items (x, y) of an item pair, then the model fits the cluster comparison case.

This suggests that any of the similarity measures that have been developed based on 2×2 contingency matrices in other scientific contexts are valid similarity measures for

the comparison of clustering pairs. Fourty-three of these measures have been collated, several of them taken from work by Hayek (1994) on feature comparisons between amphibian species. Many of these measures may be unfamiliar in the data mining and machine learning communities, where "correctness" against a gold standard is routinely expressed in terms of precision, recall and F-measure values only. Note that, just as was seen for the Rand measure (see Section 8.2.3.1), many of these measures have the same trait of having a non-constant expected value for two random partitions for comparison, which will be demonstrated empirically in the results Sections (see Section 8.4.3).

8.2.3.2 Information Theoretic Approaches in Clustering Comparison

Information Theory is a field of mathematics that stems from the need to improve the description and quantification of data, endeavouring to reliably store and transmit this data using the least amount of information possible. The measure known as *information entropy* is used to do this and is usually expressed by the average number of bits needed to store or communicate data. *Information theoretic* approaches apply entropy in different manners to compare the difference in information between two partitions. Some different approaches used are the *Powers Measure* (Powers 2007), Meila's *Variation of Information* (Meila 2003) and NMI *Normalized Mutual Information* (Horibe 1985, Malvestuto 1986, Kvalseth 1987, Quinlan 1990, Strehl & Ghosh 2002, Fred & Jain 2003).

Entropy can be described as the information conveyed by the uncertainty that a randomly selected point belongs to a certain cluster. In the context of clustering Entropy is defined as:

$$H(C) := -\sum_{i=1}^{k} P(i)\log_2 P(i) \quad \text{where} \quad P(i) := \frac{|C_i|}{n} \tag{8.3}$$

Cat.	Reference Label	Formula	References
I	*Conditional Entropy*	$H(P\|Q)$, $H(Q\|P)$	(Lee 1987, Malvestuto 1986, Pawlak, Wong & Ziarko 1988)
	Asymmetric NMI	$\frac{I(P;Q)}{H(P)}$, $\frac{I(P;Q)}{H(Q)}$	(Kvalseth 1987, Malvestuto 1986, Quinlan 1990)
II	*Joint Entropy*	$H(P,Q)$	
	Mutual Information	$I(P;Q)$	(Knobbe & Adrianns 1996, Linfoot 1957, Quinlan 1990)
	NMI 1	$\frac{I(P;Q)}{H(P,Q)}$	(Malvestuto 1986)

Cat.	Reference Label	Formula	References
	NMI 2	$\frac{I(P;Q)}{max(H(P)+H(Q))}$	(Horibe 1985, Kvalseth 1987)
	NMI 3	$\frac{I(P;Q)}{min(H(P)+H(Q))}$	(Kvalseth 1987)
	NMI 4	$\frac{I(P;Q)}{\sqrt{(H(P)H(Q))}}$	(Strehl & Ghosh 2002)
	NMI 5	$\frac{2I(P;Q)}{H(P)+H(Q)}$	(Kvalseth 1987, Fred & Jain 2003)
III	*Lopez_Wan*	$H(P\|Q) + H(Q\|P)$	(Lopez de Mantaras 1989, Wan & Wong 1989),
	Lopez_Rajski	$\frac{H(P\|Q)+H(Q\|P)}{H(P,Q)}$	(Lopez de Mantaras 1989, Rajski 1961),
	Meila	$H(P) + H(Q) - 2I(P;Q)$	(Meila 2003)
	Powers	$\left(\frac{2H(P,Q)}{H(P)+H(Q)}\right) - 1$	(Powers 2007)

Table 8.3: Various information theoretic measures. [2]

Work by Yao, Wong and Butz (1999) critically analyses many different information-theoretic measures, obtained by combining and normalizing conditional entropy and mutual information in various ways. Some of these measures are presented in Table 8.3. Yao et al. (1999) also point out the following relationships:

1. $\frac{I(P;Q)}{H(P)} = 1 - \frac{H(P|Q)}{H(P)}$
2. $\frac{I(P;Q)}{max(H(P),H(Q))} = min\left(\frac{I(P;Q)}{H(P)}, \frac{I(P;Q)}{H(Q)}\right)$
3. $\frac{I(P;Q)}{min(H(P),H(Q))} = max\left(\frac{I(P;Q)}{H(P)}, \frac{I(P;Q)}{H(Q)}\right)$
4. $0 \leq \frac{I(P;Q)}{max(H(P),H(Q))} \leq \frac{2I(P;Q)}{H(P)+H(Q)} \leq \frac{I(P;Q)}{min(H(P),H(Q))}$
5. $H(P|Q) + H(Q|P) = H(P < Q) - I(P;Q)$
6. $\frac{2I(P;Q)}{H(P)+H(Q)} = 2\left(1 - \frac{H(P,Q)}{H(P)+H(Q)}\right)$
7. $\frac{H(P|Q)+H(Q|P)}{H(P,Q)} = 1 - \frac{I(P;Q)}{H(P,Q)}$

Conditional Entropy The conditional entropy measures how much entropy a random variable Y has remaining if you have already learned completely the value of a second random variable X. In other words it expresses how much extra information you still need to supply on average to communicate Y given that the other party knows X. The higher the conditional entropy the more an observer can predict the state of a variable, knowing the state of the other variable.

$$H(P|Q) = H(Q,P) - H(P) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \tag{8.4}$$

[2] Adapted from Yao et al., 1999. Category I = asymmetric measures; Category II = symmetric measures; Category III = distance measures.

Joint Entropy The joint entropy measures how much entropy is contained in a joint system of two random variables. In other words it is the amount of information needed on average to specify both the values and is defined as:

$$H(P,Q) = -\sum_{x \in P} \sum_{y \in Q} p(x,y) \log p(X,Y) \tag{8.5}$$

Mutual Information The Mutual Information of two random variables expresses their mutual dependence or the amount of information they have in common. In other words, it measures how much knowing one of these variables reduces the uncertainty about the other. Following is a definition for Mutual Information where $p(P)$, $p(Q)$ and $p(P,Q)$ are probabilities.

$$I(X,Y) = \sum^{Y} \sum^{X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \;=\; H(P) + H(Q) - H(P,Q) \tag{8.6}$$

Powers Whereas conditional entropy is an asymmetric measure of the information required to specify one model given the other, the Powers measure (Powers 2007) was developed to allow for the fact that it is not known which model is correct, or even better. It calculates a symmetric measure of the information required to specify the alternate model given the better model, assuming that which model is correct is unknown and the two models are equiprobable. The unnormalised measure is the average of two non-negative asymmetric measures and thus always non-negative, with 0 representing identity of the models.

$$\frac{H(P|Q) + H(Q|P)}{H(P) + H(Q)} \;=\; \left(\frac{2H(P,Q)}{H(P) + H(Q)} \right) - 1 \tag{8.7}$$

Mutual Information ($I(P,Q)$) is complementary to this model, and is also always non-negative, with 0 representing the case where $H(P)$ or $H(Q)$ is vacuous viz. has 0 entropy, indicating a trivial clustering into a single category. Whereas the Powers measure sets $[H(P)+H(Q)]/2$ as a lower bound for $H(P,Q)$, $I(P,Q)$ sets $[H(P)+H(Q)]$ as an upper bound for $H(P,Q)$. Conversely $2H(P,Q)$ is an upper bound, and $H(P,Q)$ a lower bound, for $H(P) + H(Q)$. These represent relationships between the expected entropy (the expected number of bits to represent the correct distribution given these models are equiprobably) and the joint entropy (the number of bits to represent the fragments defined by the two distributions).

Meila's Variation of Information This measure was proposed by Meila (2003) as an information theoretic to compare two clusterings of the same data. Presented as VI (Variation of Information) it measures the amount of information lost or gained in

changing from one cluster C to another C'. This measure is positive, symmetric and transitive and in Meila's words "surprisingly enough a metric". However, it should be pointed out that it is not normalized, which would improve its comparability to other measures.

$$VI(P,Q) = H(P) + H(Q) - 2I(P,Q) \tag{8.8}$$

Normalized Mutual Information There are several different approaches to the normalization of mutual information, two of these come in the form of the *coefficients of constraint* by Coombs, Dawes and Tversky (1970) and as the *uncertainty coefficient* by Press, Flannery, Teukolsky and Vetterling (1988), $C_{PQ} = \frac{I(P;Q)}{H(Q)}$ and $C_{QP} = \frac{I(P;Q)}{H(P)}$. It is clear that these two coefficients are not equal or symmetric. A symmetric alternative is that of *redundancy* $R = \frac{I(P;Q)}{H(P)+H(Q)}$. *Redundancy* obtains its minimum of zero when both variables are independent. Alternately, it reaches its maximum value of $R_{max} = \frac{min(H(X),H(Y))}{H(X)+H(Y)}$ when one of the variables is totally redundant to the other. Another symmetrical measure is that of *symmetric uncertainty* by Witten & Frank (2005) which is $U(P,Q) = 2R = 2\frac{I(P;Q)}{H(P)+H(Q)}$ which represents a weighted average of the two uncertainty coefficients.

In addition to these measures, one can also consider as cluster comparison measures Joint Entropy, Unnormalized Mutual Information and the two asymmetric versions of Conditional Entropy. Table 8.5 lists the information theoretic approaches considered in this paper along with their formulas in terms of Entropy $H(P)$ and $H(Q)$.

Formula Table - The following Tables 8.4 & 8.5 index the different formula used in testing and comparing MoC.

Pair Counting Measures			
Name	Formula	Range	Ref
Baroni Urbani & Buser 1	$\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$	$(-1, 1)$	(Baroni-Urbani & Buser 1976)
Baroni Urbani & Buser 2	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	$(0, 1)$	(Baroni-Urbani & Buser 1976)
Braun & Blanquet	$\frac{a}{a+max(b,c)}$	$(0, 1)$	(Braun-Blanquet & ; 1932)
Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$(0, 1)$	(Manning & Schutze 1999)
Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	$(-\infty, \infty)$	(Dennis, Williams & Shreeve 1998)
Dice Symmetric	$\frac{2a}{2a+b+c}$	$(0, 1)$	(Dice 1945)
Dice Asymmetric 1	$B\|A = \frac{a}{a+c}$	$(0, 1)$	(Dice 1945)
Dice Asymmetric 2	$A\|B = \frac{a}{a+b}$	$(0, 1)$	(Dice 1945)
Fager	$\frac{a}{\sqrt{[(a+c)(a+b)]^2}} - \frac{max(b,c)}{2}$	$(-\infty, 1)$	(Fager & McGowan 1963)
Faith	$\frac{2a+d}{2n}$	$(0, \infty)$	(Faith 1983)
Filkov	$b + c$	$(0, \infty)$	(Filkov & Skiena 2004)
Fowlkes Mallows	$\frac{a^2}{(a+c)(a+b)}$	$(0, 1)$	(Sorgenfrei 1958)
Forbes	$\frac{na-(a+b)(a+c)}{n \cdot min(b,c)-(a+b)(a+c)}$	$(-\infty, \infty)$	(Forbes 1925)
Forbes d	$\frac{na}{(a+b)(a+c)}$	$(0, \infty)$	(Forbes 1925)
Fossum	$\frac{n\left(a-\frac{1}{2}\right)^2}{(a+b)(a+c)}$	$(0, \infty)$	(Fossum & Haller 2004)
Gilbert Wells	$\log \frac{an}{(a+b)(a+c)}$	$(-\infty, \infty)$	(Gilbert & Wells 1966)
Goodall	$asin\sqrt{\left(\frac{a+d}{n}\right)}/(50 \times \pi)$	$(0, .57)$	(Goodall 1967)
Hamann	$\frac{(a+d)-(b+c)}{n}$	$(-1, 1)$	(Hamann 1961)
Jaccard	$\frac{a}{a+b+c}$	$(0, 1)$	(Jaccard 1901)
Johnson	$\frac{a}{a+b} + \frac{a}{a+c}$	$(0, 2)$	(Johnson 1967)
Kulczynski 1	$\frac{a}{b+c}$	$(0, \infty)$	(Kulczynski 1927)
Kulczynski 2	$\frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$	$(0, 1)$	(Kulczynski 1927)
McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$	$(-1, 1)$	(McConnaughey 1964)
Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)}$	$(-1, 1)$	(Michael 1920)
Mirkin	$2(b + c)$	$(0, \infty)$	(Mirkin 1996)
MoC	$\frac{\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{f_{ij}^2}{p_i q_j}\right)}{\left(\sqrt{IJ}\right)-1}$	$(0, 1)$	see Section 8.4.1
Mountford	$\frac{2a}{2bc+ab+ac}$	$(0, \infty)$	(Mountford 1962)
Overlap	$\frac{a}{\min((a+b),\ (a+c))}$	$(0, 1)$	(Manning & Schutze 1999)
Rand	$\frac{a+d}{n}$	$(0, 1)$	(Rand 1971)
Rogers & Tanimoto	$\frac{(a+d)}{(a+d)+2(b+c)}$	$(0, 1)$	(Rogers & Tanimoto 1960)
Russell & Rao	$\frac{a}{n}$	$(0, 1)$	(Russell & Rao 1940)
Savage	$1 - \frac{a}{a+max(b,c)}$	$(0, 1)$	(Savage 1934)
Sneath Pattern Difference	$\frac{2\sqrt{bc}}{n}$	$(0, 1)$	(Sneath 1968)
Sneath Total Difference	$\frac{b+c}{n}$	$(0, 1)$	(Sneath 1968)
Sokal & Sneath 1	$\frac{2(a+d)}{2(a+d)+(b+c)}$	$(0, 1)$	(Sokal & Sneath 1964)
Sokal & Sneath 2	$\frac{a}{a+2b+2c}$	$(0, 1)$	(Sokal & Sneath 1964)
Sokal & Sneath 3	$\frac{a+d}{b+c}$	$(0, \infty)$	(Sokal & Sneath 1964)
Sokal & Sneath 4	$\frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$	$(0, 1)$	(Sokal & Sneath 1964)
Sokal & Sneath 5	$\frac{ad}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$	$(0, 1)$	(Sokal & Sneath 1964)
Sokal & Sneath Non Metric	$\frac{b+c}{2a+b+c}$	$(0, 1)$	(Sneath & Sokal 1973)
Stiles	$log\frac{n\left(\|ad-bc\|-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$	$(-\infty, \infty)$	(Holliday, Hu & Willett 2002)
Tarwid	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	$(-1, 1)$	(Tarwid 1960)
Yules Omega	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$(-1, 1)$	(Yule 1912)

Table 8.4: Pair Counting Formula Table

Information Theoretic Measures			
Name	Formula	Range	Ref
Entropy Conditional	$H(Q,P) - H(P)$	$(0, \infty)$	(Manning & Schutze 1999)
Entropy Joint	$H(P,Q)$	$(0, \infty)$	(Manning & Schutze 1999)
NMI 1 [3]	$\frac{I(P;Q)}{H(P,Q)}$	$(0, 1)$	ref. Table 8.3
NMI 2 [3]	$\frac{I(P;Q)}{max(H(P),H(Q))}$	$(0, 1)$	ref. Table 8.3
NMI 3 [4]	$\frac{I(P;Q)}{min(H(P),H(Q))}$	$(0, 1)$	ref. Table 8.3
NMI 4 [4]	$\frac{I(P;Q)}{\sqrt{(H(P)H(Q))}}$	$(0, 1)$	ref. Table 8.3
NMI 5 [3]	$\frac{2I(P,Q)}{H(P)+H(Q)}$	$(0, 1)$	ref. Table 8.3
Asymmetric NMI [4]	$\frac{I(P;Q)}{H(P)}$, $\frac{I(P;Q)}{H(Q)}$	$(0, 1)$	ref. Table 8.3
Mutual Information	$I(P,Q) = H(P) + H(Q) - H(P,Q)$	$(0, \infty)$	(Manning & Schutze 1999)
Meila	$VI(P,Q) = H(P) + H(Q) - 2I(P,Q)$	$(0, \infty)$	(Meila 2003)
Powers [3]	$\frac{H(P\|Q)+H(Q\|P)}{H(P)+H(Q)} = \left(\frac{2H(P,Q)}{H(P)+H(Q)}\right) - 1$	$(0, 1)$	(Powers 2007)
Lopez_Wan	$H(P\|Q) + H(Q\|P)$	$(0, \infty)$	ref. Table 8.3
Lopez_Rajski [3]	$\frac{H(P\|Q)+H(Q\|P)}{H(P,Q)}$	$(0, 1)$	ref. Table 8.3

Table 8.5: Information Theoretic Formula Table

8.3 Desirable Behaviour of a Clustering Comparison Measures

The use of external criteria to compare clusterings (see Section 8.2.2) requires the comparison of two partitions via measures that reflect similarity in terms of features such as the number of clusters, cluster sizes and relative cluster memberships. This can be achieved through techniques such as *pair counting* and *information theoretic* approaches as discussed in Section 8.2.3.

Notwithstanding the innate ability of humans to spot patterns and relationships it is basically impossible to truly characterize what heuristics humans might use in the comparison of partitions. It was this observation that prompted the research presented in this paper to investigate what it is that makes partitions *different* from a human or external perspective. Interested in this question is a result of work in a number of settings, including clustering in document retrieval, human-computer-interaction modelling, and in evaluating the unsupervised induction of lexical categories from real linguistic data.

To identify perfectly matching clustering pairs is a relatively simple task, however quantifying how different a pair of partitions are is far more difficult. Different measures will have different qualities both negative and positive depending on the partitions and the desired outcome. In prelude to describing the method of comparing pairs

[3]This normalized formula is not well defined for $I = J = 1$ ($H(P) = H(Q) = 0$) but can be defined as 1 for similarity, and 0 for distance.

[4]This normalized formula is not well defined for $I = 1$ or $J = 1$ ($H(P) = 0$ or $H(Q) = 0$) but can be defined as 1 for $I = J$, and 0 for $I \neq J$.

of partitions a *desiderata* is defined to guide the selection and testing of measures process, and outline the worst cases of *Independently Codistributed Clustering Pairs*, *Complete Fragmentation* and *Conjugate Partition Pairs* used in the comparison of pairs of partitions.

Desiderata of Appropriate Measure Characteristics

1. The comparison measure $m(P, Q)$ should be independent of any concept of the 'goodness' of the individual clusterings.
2. In the absence of a Gold Standard a measure should be symmetric in regard to the two partitions i.e. $m(P, Q) = m(Q, P)$.
3. The comparison measure should range in value between 0 and 1, where 1 is a perfect match and 0 is realized by all worst case scenarios, including independence and complete fragmentation (see Sections 8.3.1 & 8.3.2).
4. There should be no dependence of the comparison measure on the number of elements - the significance of a cluster is expected to increase with size, but the evaluation of the pattern should depend only on the probabilities, and be independent of size.
5. Similarly there should be no change in the value of the measure simply as a result of changing the number of clusters.
6. The fall-off of the similarity measure should match the intuition of decreasing similarity, which may vary for different problems. For the purpose of this research, greate value is given to measurse that have the ability to indicate perfect matches between partitions, with even relatively small differences being intolerable. In the context of this research, therefore, it would be desirable for a measure to depart sharply from its best-case value as soon as the partitions begin to differ.

8.3.1 Independently Codistributed Clustering Pairs

A clear instance of a mismatch between two clusterings P and Q occurs when every cluster in P contains elements from each of the clusters in Q in the same proportions in which they are distributed among the Q clusters. In this situation the sizes of the intersections between every cluster P_i and every cluster Q_j, i.e. the values in the co-occurrence matrix between P and Q, take on their expected values according to the marginal totals.

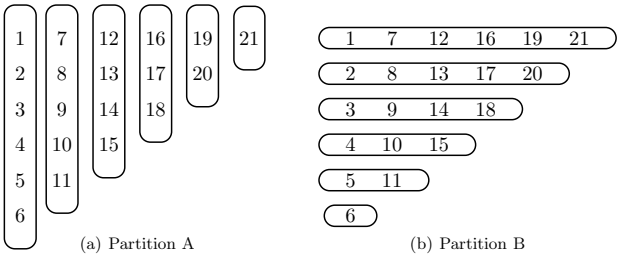


Fig. 8.1: Self Conjugate Partition example

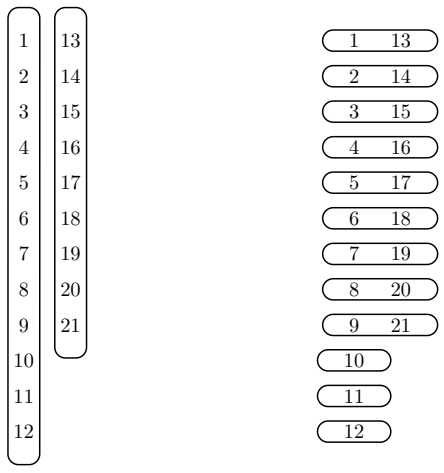
This case corresponds to a situation where no P cluster has any particular affinity for any Q cluster, so that knowing how the elements of a particular cluster P_i are distributed among the Q clusters provides no information about the value of the index i. In this case, P and Q are independently distributed.

It is desirable for a clustering comparison measure to recognise this *independently-distributed worst case*, and to take on its minimum value when the case occurs.

8.3.2 Complete Fragmentation and Conjugate Partition Pairs

Another clear "worst case" is when all intersections of classes of P and Q are singletons. the concept of "Conjugate Partition Pairs" is introduced to define a 'maximal' case which leads to *complete fragmentation*.

The conjugate of a partition is simply the 90^o rotation of the partition where the clusters change from being say the rows (see Figure 8.1(a)) of a matrix to the columns (see Figure 8.1(b)). By rotating the matrix figures 8.1(a) & 8.1(b) depict clusters with the same data points, which are the conjugate (transpose) of each other. This means that in the two partitions, every pair of clusters has only one element in common (complete fragmentation). In the pair counting approach, therefore, there are no *pairs* in common, so that cell a (representing True Positive) in the contingency matrix (see Table 8.1 and 8.2) is zero. Measures which are based on a (such as Precision and Recall, or Rand's measure) may be expected to identify this worst case relatively well. Partitions that maintain their original structure after rotation, as seen in Figures 8.1(a) and 8.1(b), are described as Self Conjugate or Symmetric Conjugate Partitions while partitions that **do not** retain their original structure (where by "retaining structure" it is meant they have the same distribution of cluster sizes), as seen in figures 8.2(a) and 8.2(b), are known as non-symmetric conjugate partitions.



(a) Partition C (b) Partition D

Fig. 8.2: Non-symmetric Conjugate Partition example

8.4 Measure of Concordance

The remainder of this Chapter introduces a novel *Measure of Concordance* (MoC) and evaluates it and a number of other similarity/dissimilarity measures as applied to the context of comparing a pair of clusterings. These include measures previously proposed for this problem, and a host of other similarity/dissimilarity measures that, although they have not previously been applied to clustering comparison, are applicable.

The evaluation of these measures is performed in the context of five test case scenarios in which certain characteristics are systematically manipulated on one member of a pair of clusterings, in order to determine whether the measures are sensitive to these manipulations. These tests are derived from a consideration of two possible worst-case matches between a pair of clusterings, which are termed Independently Codistributed Clustering Pairs and Conjugate Partition Pairs.

8.4.1 MoC Derivation & Justification

This section introduces the *Measure of Concordance* (MoC) through logical development beginning with an example. Suppose that a data set D of 36 elements has been clustered using two rival clustering algorithms, so that the first algorithm clusters the elements of D into the clustering P, and the second algorithm clusters the same elements into the clustering Q as depicted in Figure 8.3. Regarding these two clusterings, the question of interest is how to express quantitatively the extent to which they agree relative to the underlying groupings present in the dataset.

Given that there are I clusters in P, and J clusters in Q (with I and J not necessarily

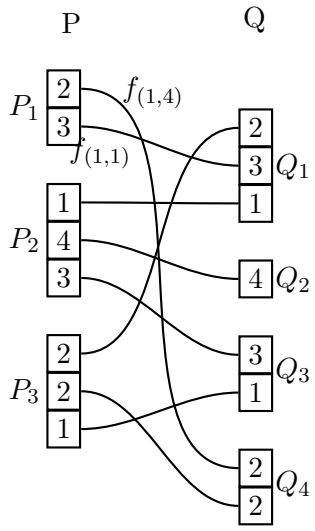


Fig. 8.3: An illustration of the division of clusters into fragments

equal), and that each individual cluster in P is referred to as P_i, and each cluster in Q as Q_j, for $i \in \{1, 2, \ldots, I\}$ and $j \in \{1, 2, \ldots, J\}$. Then any cluster P_i from P can be subdivided into smaller subclusters or *fragments*, where a fragment consists of those elements of P_i that have also been allocated to a single cluster Q_j in clustering Q, for some j. This fragment, labeled F_{ij}, is therefore the intersection between P_i and Q_j. Fragments represent instances where both clusterings agree that the elements involved "belong together", and hence represent the shared structure between the two clusterings. Clearly, if cluster P_i contains the fragment F_{ij}, then cluster Q_j also contains the same fragment F_{ij}.

The relationship between the clusterings P and Q can be expressed as a co-occurrence matrix F, with row i corresponding to cluster P_i and column j corresponding to Q_j, so that each cell of F contains the size of fragment F_{ij} as demonstrated by Table 8.6.

	Q_1	Q_2	Q_3	Q_4
P_1	3	0	0	2
P_2	1	4	3	0
P_3	2	0	1	2

Table 8.6: Fragment co-occurance matrix example

The rectangles on the left-hand side of Table 8.6 labeled P_1, P_2 and P_3 are clusters that make up a clustering P, while the rectangles labeled Q_1, Q_2, Q_3 and Q_4 on the right-hand side are clusters that make up a clustering Q. The smaller squares composing the rectangles are fragments. Lines connect a fragment in a P cluster to the corresponding fragment in the Q cluster.

To illustrate the notion of fragments, consider the situation in Figure 8.3. Here,

cluster P_1 shares a fragment of size 3 with cluster Q_1, and a fragment of size 2 with cluster Q_4. It has no fragments in common with clusters Q_2 or Q_3. All 4 elements of Cluster Q_2 are grouped together in cluster P_2, and therefore Q_2 has only one fragment of size 4. Cluster P_2, on the other hand, also contains additional fragments with clusters Q_1 and Q_3.

Recall that $|F_{ij}|$ is written as f_{ij}, $|P_i|$ as p_i and $|Q_j|$ as q_j(see Section 8.2.3). Then the proportion f_{ij}/p_i reflects the proportion of elements of P_i that are also in Q_j, so that f_{ij}/p_i is the conditional probability $P(Q_j|P_i)$ such that any element of P_i is also an element of Q_j. Likewise, f_{ij}/q_j is the conditional probability $P(P_i|Q_j)$ that any element of Q_j is also in P_i. Clearly, when $f_{ij}/p_i = 1$, the entire cluster P_i is a subset of the cluster Q_j, and conversely, if $f_{ij}/q_j = 1$, Q_j is a subset of P_i.

Next, consider the product of these two terms, ${f_{ij}}^2/p_iq_j$. This term provides a symmetric measure of mutual agreement or *mutual concordance* between the two clusters P_i and Q_j. The maximum value $f_{ij}^2/p_iq_j = 1$ is attained only when $f_{ij}/p_i = f_{ij}/q_j$, i. e. when $P_i = Q_j$.

Let S be the sum of mutual concordance over all fragments, i.e.
$S = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{f_{ij}^2}{p_iq_j}$. So S takes on its maximum value iff $f_{ij}^2/p_iq_j = 1$ for all i and j. This occurs iff $P = Q$; in this case the value of S is equal to I (which is also equal to J), the number of clusters. The minimum value of S is 1, and is attained when there is no relationship of concordance between P and Q, i.e. when every cluster P_i is broken up into fragments whose sizes reflect the overall distribution of the data set into the clusters of Q. In this case, the elements of every P_i are evenly distributed among the Q clusters (and vice versa), and fragment sizes take on their expected values given the marginal totals of the P and Q clusterings.

Figures 8.4(a), 8.4(b) and 8.4(c) illustrate the effect on S of various kinds of fragmentation. Figure 8.4(a) represents a perfect match (no fragmentation) between clusters P_i and Q_j, so that they contribute $\frac{6\times6}{6\times6} = 1$ to the sum S. In Figure 8.4(b), Q_1 has split into two clusters, Q_1 and Q_2, inducing two fragments on P_1. However, note that the contribution to S is still $\frac{4\times4}{6\times4} + \frac{2\times2}{6\times2} = 1$. In other words, merely splitting up a cluster does not detract from S (although, of course, the total number of clusters increases). S is reduced, however, by grouping the elements of the new clusters Q_1 and Q_2 together with elements that are not in P_1, as shown in Figure 8.4(c), as this causes the f_{ij}/q_j terms to decrease. In this case, the contribution to S is $\frac{4\times4}{6\times7} + \frac{2\times2}{6\times5} \approx 0.481$. The phenomenon in Figure 8.4(b) is quite general: whenever a single cluster in P (resp. Q) can be decomposed entirely into a set of clusters in Q (resp. P), then S is increased by 1. This suggests that the sum S is not entirely adequate as a measure of the amount of concordance between two clusterings. Situations analogous to Figure 8.4(b) need to

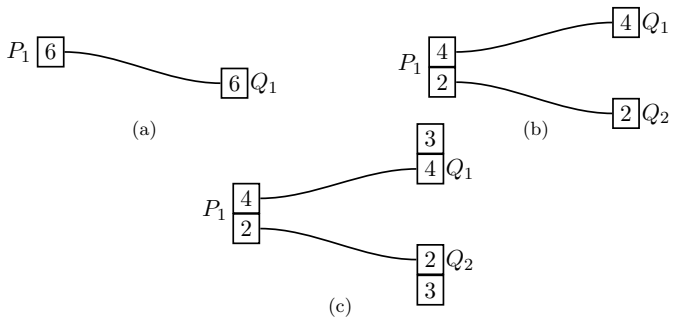


Fig. 8.4: Fragment Types

Three different situations illustrating the effect on MoC of various kinds of fragmentation. In (a), there is a complete correspondence between the two clusters. In (b), the elements of the P cluster have been distributed over two Q clusters. In (c), the elements of the P cluster have also been grouped, in the Q clusters, with elements from a different P cluster. In the MoC formula, the normalisation factor $\sqrt{IJ} - 1$ yields a lower MoC score for (b) than for (a), and the summing of mutual concordance terms ${f_{ij}}^2/p_i q_j$ yields a lower MoC score for (c) than for (b).

be penalised for "using more clusters" than situations analogous to Figure 8.4(a).

An obvious solution is to normalise S by a function of the number of clusters involved. There are a number of desirable characteristics which the normalization function and resulting normalized measure should exhibit. Firstly, as stated before, the range of values of the measure should be between 0 and 1 inclusive, with those extreme values being reserved for the worst and best cases respectively. Secondly, as the maximum attainable value for S is I when both P and Q consist of I clusters (so that $I = J$), it is appropriate in that case to normalise by I (equivalently, J). An appropriate normalization function should therefore take on the value $I = J$ in this worst case. Thirdly, even when $I \neq J$, the value of the normalization function should be of the same order of magnitude as I and J. Fourthly, the normalization function should impose a penalty in cases of relatively greater fragmentation, and should treat I and J symmetrically. Possibly the simplest normalization function that satisfies these requirements is $\sqrt{IJ}$, the geometric mean of I and J.

Finally, then, MoC is defined as as

$$MoC(P,Q) = \begin{cases} 1: & \text{if } I = J = 1 \\ \frac{1}{(\sqrt{IJ}-1)} \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{{f_{ij}}^2}{p_i q_j} - 1 \right): & \text{otherwise} \end{cases} \tag{8.9}$$

This provides a measure of the degree of concordance between two clusterings P and Q, and takes on the value 0 for independence between P and Q, and 1 when $P = Q$.

Another way to understand the MoC measure is in terms of the more familiar precision and recall measures, as follows. For every pair of clusters P_i and Q_j the $I \times J$ co-occurrence matrix F can be collapsed into a 2×2 contingency table, with the first row corresponding to P_i and the second row to all other P clusters, and likewise the first column corresponding to Q_j and the second column to all other Q clusters. Then, labeling the cells a, b, c and d as in Table 8.1 for convenience only, it is clear that $a = f_{ij}$, $b = p_i - f_{ij}$, $c = q_j - f_{ij}$, and $d = N - p_i - q_j + f_{ij}$, where N is the total number of elements in D. In this case, the mutual concordance term $f_{ij}^2/p_i q_j$ in the calculation of MoC is clearly the product of precision and recall obtained from this table. So MoC can alternatively be regarded as the (normalised) sum of the products of precision and recall over every 2×2 table induced over the cells of F.

8.4.2 Relationship to Pearson's Chi-Squared Statistic

There exists a close relationship between MoC and the familiar Pearson's chi-squared (χ^2) statistic for the independence of two variables. Chi-squared can also be used to test for independence between the two clusterings P and Q. Using the marginal totals of the co-occurrence matrix defined by clusterings P and Q, the expected value for the size of fragment f_{ij} is given by $p_i q_j / N$. The χ^2 statistic tests the goodness of fit of the obtained f_{ij} values against the expected values derived from the marginal totals. Note that the case where the obtained and expected values are the same corresponds to the case where the allocation of elements from any cluster P_i to the clusters of Q follows the same distribution as the allocation of the entire set of elements to the clusters of Q. In other words, no P_i cluster has any particular affinity for any of the Q-clusters which could distinguish it from any other P-cluster. So there is no information about how the elements will be allocated to Q clusters when given the index i of the P-cluster. Clearly, this constitutes the situation of minimal relationship (maximal independence) between P and Q, and in this case, $\chi^2 = 0$.

We can write χ^2 as

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ij} - \frac{p_i q_j}{N}\right)^2}{\frac{p_i q_j}{N}} \\
&= N \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{{f_{ij}}^2}{p_i q_j} - 2 \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} + \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{p_i q_j}{N} \\
&= N \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{{f_{ij}}^2}{p_i q_j} - 2N + N \\
&= N \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{{f_{ij}}^2}{p_i q_j} - 1 \right) \\
&= N \left(\sqrt{IJ} - 1 \right) MoC
\end{aligned}$$

Thus, MoC is a normalized form of χ^2. This means that a χ^2 value can easily be obtained given a MoC value and vice versa. From the χ^2 value, significance can be

determined; note that this gives the significance of the departure of the P and Q distributions from a purely even co-distribution based on their marginal totals, rather than the significance of the association between P and Q.

Because χ^2 can take on an arbitrarily large value, it is not directly suitable for the purpose of expressing the strength of the association between the two clusterings. It is preferable to use a measure of association derived from χ^2 which has been normalized to range from 0 to 1. This suggests that MoC is one such suitable measure.

Two other popular measures of association that are derived from χ^2, as discussed by Mirkin (2001), and that perform normalization somewhat differently, are the Cramér and Tchouproff coefficients, given by

$$\phi_c = \sqrt{\frac{\chi^2}{N(min(I-1,J-1))}}, \qquad \phi_t = \sqrt{\frac{\chi^2}{N\sqrt{(I-1)(J-1)}}}.$$

Cramer's coefficient. Tchouproff's coefficient.

Note firstly, however, that both Cramér's and Tchouproff's coefficients are undefined when either $I = 1$ or $J = 1$. This seems to be a deficiency; in the case where (say) a clustering algorithm places all data elements in one cluster, nevertheless one would want to allocate a value to its concordance with the gold standard. MoC is defined to have a value even when $I = 1$ or $J = 1$ (albeit by way of exception in the case where $I = J = 1$).

Furthermore, in some cases, Cramér's coefficient does not recognise departures from a perfect match between P and Q. Consider the situation in Figure 8.5. The sum S is 3, because P_2 can be cleanly divided into Q_2 and Q_3, so that the contribution to S from P_2, Q_2 and Q_3 is 1, as discussed above. Then $\phi_c = \sqrt{\frac{3-1}{min(3-1,4-1)}} = 1$, even though P and Q are clearly not identical. MoC does not suffer from this problem, but instead gives a value of $\frac{3-1}{\sqrt{3\times4}-1} \approx 0.812$ in this case.

Alternative normalizations of χ^2 have also been proposed, but suffer from these and other problems; for a general review see Hayek (1994).

8.4.3 Qualitative Description of Measure Behaviour

The following section describes the behaviour of the measures listed in Tables 8.4 and 8.5 under the worst-case situations discussed in Section 8.3. Rather than performing an exhaustive quantitative analysis of the behaviour of each measure, specific clustering scenarios have been devised representing each of these cases and qualitative descriptions of the behaviour of the measures under these scenarios were collated. These qualitative descriptions are presented as broad characterizations of the general behaviour of these

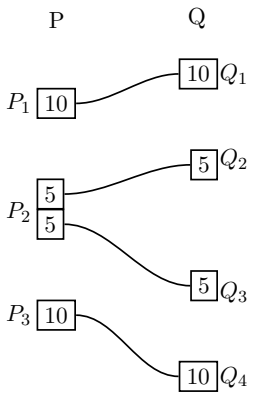


Fig. 8.5: Departure from Perfect Match Example

measures in the evenly distributed worst case, the perfectly matching best case and the conjugate worst case. These cases are characterized in terms of *range* and *shape*. Range describes the set of values attained by a function for the given test sequence, which should fall within the set of all values attainable by that function for its possible domain as described in the 'range' column of Tables 8.4 & 8.5. Shape describes the plot formed by the set of values attained by a function for the given test sequence. To avoid artifacts due to limited precision in the simulations, a measure is characterised as taking on its extreme value if it differs by no more than some fixed value ϵ from that value, and as being constant at a value over an entire scenario if it varies by less than some fixed value ϵ throughout. In Scenarios 1 to 4, ϵ was set at 0.01, and in scenarios 5, ϵ was 0.05 (the reason for the difference is explained in Section 8.4.5). For convenience the different shapes are presented in 'key' format below in Figure 8.6.

Under these conditions following observations regarding a, b, c & d (as defined in Table 8.1) are listed:

- a & d decrease proportionally to the increase in b & c as members are shifted out of common clusters resulting in a decreasing number of member pairs available as intersections (a's)
- b & c increase as members are shifted out of common clusters resulting in an increasing number of member pairs available as complements (b's & c's)

Range and Shape Key Following is a short summary of of the different graphical shapes and descriptor labels used to characterise a measure's general behaviour for each given test.

- Constant functions are indicated by the symbol C.

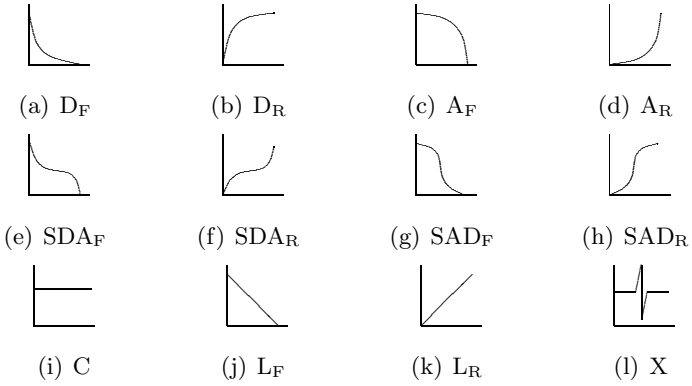


Fig. 8.6: Plot Shapes Key

	1	**2**	**3**	...	**100**
1	10^5	0	0	...	0
2	0	10^5	0	...	0
3	0	0	10^5	...	0
⋮	⋮	⋮	⋮	⋱	⋮
100	0	0	0	...	10^5

(a)

	1	**2**	**3**	...	**100**
1	-99	1	1	...	1
2	1	-99	1	...	1
3	1	1	-99	...	1
⋮	⋮	⋮	⋮	⋱	⋮
100	1	1	1	...	-99

(b)

	1	**2**	**3**	...	**100**
1	1000	1000	1000	...	1000
2	1000	1000	1000	...	1000
3	1000	1000	1000	...	1000
⋮	⋮	⋮	⋮	⋱	⋮
100	1000	1000	1000	...	1000

(c)

Fig. 8.7: Incremental eveness contingency matrices

- Linear functions are indicated by the symbol L, with a subscript of either F or R to indicate whether their value respectively falls or rises over the interval.
- Non-linear functions are described in terms of whether the absolute value of their derivative over the interval is increasing or decreasing, i.e. whether the function value is *accelerating* or *decelerating*. *Decelerating* functions are indicated by the *symbol D*, and *accelerating* functions by the *letter A*, with *subscripts F* and *R* indicating *falling* and *rising* behaviour as before.
- Sigmoid functions can be described as piecewise functions assembled from a pair of falling or a pair of rising functions, with the two members of the pair exhibiting opposite accelerating/decelerating behaviour. So for instance, the shape labelled as SAD_F in Figure 8.6(g) is constructed from an initial A_F function followed by a D_F function. The other three sigmoid functions are defined similarly.
- Functions which are undefined over the interval are indicated by a U.
- Other functions are indicated by an X, however there is only one such function (Forbes in the Incremental Independence test.)

8.4.4 Testing on Independently Distributed Clusterings

8.4.4.1 Incremental Independence

Scenario 1 demonstrates how each function reacts when incrementally increasing the amount of a partitioning that is independently distributed, while holding the number of members, fragments and clusters constant. This is achieved by comparing pairs of clusterings that range from being perfectly correlated with each other to being perfectly independent of each other, by systematically altering the original composition of clustering P according to the marginal totals of the clustering Q, in a series of 1000 increments of 0.1% at a time.

This scenario starts with the co-occurrence matrix in Figure 8.7(a) which is incrementally added to using the matrix in Figure 8.7(b). After one thousand increments, the situation of total independence depicted in Figure 8.7(c) is achieved. Under these conditions, as defined in Table 8.1, the following observations regarding a, b, c and d are listed:

- a and d decrease proportionally to the increase in b and c as members are shifted out of common clusters resulting in a decreasing number of member pairs available as intersections (a's)
- b and c increase as members are shifted out of common clusters resulting in an increasing number of member pairs available as complements (b's and c's)

Expected Results Given each function is tested against the continuum of perfect match (best) to total mismatch (worst), an appropriate result is one where the function produces a series of values that track from one extreme value to the other e.g. 1 to 0. As noted in the desiderata (see Section 8.3), a measure should depart rapidly from its best case value as soon as the partitions begin to differ. For this reason the preferred shape of a measure's curve should be either D_F, D_R, SDA_F or SDA_R.

8.4.4.2 Scaling of the Independent Case

Scenarios 2, 3 and 4 test the degree to which the different measures are insensitive to scale. In the independently co-distributed case of the previous section, there is a relationship between the number of data points in the set, the size of the fragment in each cell of the co-occurrence matrix, and the number of clusters in the two clusterings. These scenarios systematically examine the effect on the measures of manipulating an

independent co-distribution by holding either the number of data points, the fragment size or the number of clusters constant, while incrementing the other two parameters.

Under these conditions the following observations are made regarding the effect on a, b, c and d:

Scenario 2 Constant n, Cluster count increases & Fragment size decreases. This test keeps the number of elements in the data set constant, while changing the number of clusters. This has the effect of decreasing the size of each fragment for increasing cluster sizes as there are fewer available elements in each cluster to intersect. Under these conditions the following observations are made:

- a decreases to zero due to decreasing fragment size and thus reduced member pairs for intersections (a's)
- Δa decreases then increases
- b & c decrease to $\frac{n}{((I+J)/2)+1}$ as the available pairs for complements (b's and c's) decrease with the decrease in fragment size
- Δb and Δc decrease
- d increases as the available pairs for intersections (a's) or complements (b's and c's) decrease with the reduction in fragment size / pairs
- Δd decreases

Scenario 3 Constant Fragment Size, Cluster count increases and n increases. This test keeps the fragment size constant, while changing the number of clusters (this has the effect of increasing the number of elements in the data set which results in an increase in the number of fragments/pairs available as intersections (a's) or complements (b's and c's). Under these conditions the following observations are made:

- n is increasing
- a, b, c and d increase as the available fragments for intersections (a's), complements (b's and c's) or to form exclusive pairs between members in different clusters (d) increases
- $a = (p_i^2 - p_i) \times \frac{I \times J}{2}$
- b and $c = \left(\frac{n}{p_i+1}\right) + p_i$
- d increases as more exclusive pairs can be formed between members in different clusters due to the increasing number of fragments and number of clusters

- Δa, Δb, Δc and Δd are decreasing .

Scenario 4 Constant Cluster Count, Fragment Size increases and n increases. This test keeps the cluster count constant, while changing the fragment size (this has the effect of increasing the number of elements in the data set). Under these conditions the following observations are made:

- n is increasing
- a, b, c and d increase as n is increasing.

Expected Results As these scenarios are varying characteristics of an *independent co-distribution* the independence between the clusterings is maintained and so the result should be constant on the function's worst-case (perfect mismatch) value.

8.4.5 Testing Conjugate Partitions

Test 5 is used to demonstrate how each measure represents the difference between a range of different partitions and their conjugates. In addition, this test reflects a measure's ability to recognise differences between paired clustering structures. As discussed in Section 8.3.2, some partitions correspond to conjugate partitions with *similar* structure, in that they have similar distributions of cluster sizes, and some partitions correspond to conjugate partitions with *dissimilar* structure where the cluster size distributions of the two partitions are highly dissimilar. In many cases the independently co-distributed worst case and the conjugate worst case are one and the same. This happens whenever the distribution of cluster sizes is even (an equal number of elements in every cluster). For this reason, distributions of equal as well as unequal cluster sizes are considered.

To do this comparison the asymmetry of the partition was manipulated to reflect variation of structure across different membership distributions by holding n constant, and decreasing the slope of the distribution histogram by increasing the number of clusters. The value of the slope is given by $\frac{2n}{I}$, where I is the number of clusters in the partition. To produce cluster pairs (gold standard and conjugate) a fixed set of 5050 elements was used while varying the number of clusters. Specifically, the number of clusters was increased from 2, an almost asymmetric case (see Figure 8.8) to 92, the almost symmetric case (see Figure 8.9), in increments of 10. This results in a decreasing slope as the number of clusters increases where d decreases in size, while b and c increase (a is constant at zero). From each clustering generated in this way, the conjugate clustering was produced. Because of the discrete nature of the operation of conjugation, it was not always possible to generate an initial clustering of 5050

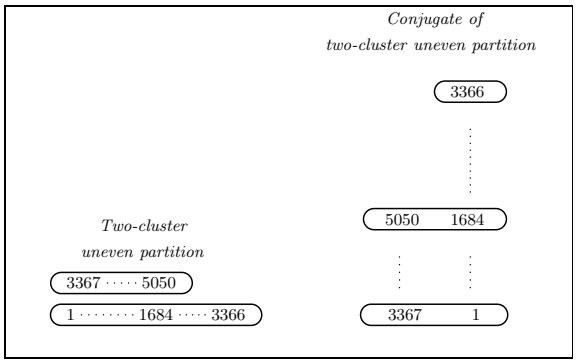


Fig. 8.8: Asymmetric uneven partition conjugate example

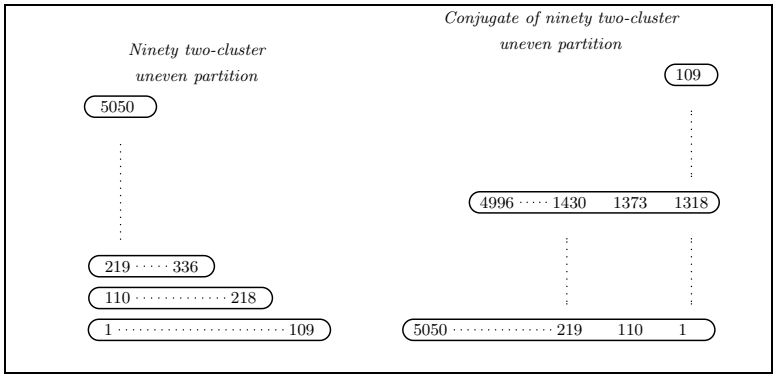


Fig. 8.9: Near symmetric uneven partition conjugate example

elements. Instead, for a given slope, the element set contained the nearest integral number of elements that would fit the slope; element set sizes ranged between 5043 and 5052. For this reason one should expect slight fluctuation in the value of a measure. In practice, the threshold ϵ value (see the introduction of this Section) of 0.05 was applied in this case.

Under these conditions the following observations are made regarding the effect on a, b, c & d:

- n is constant
- a is zero in all cases as there are no fragments with size larger than 1
- b & c increase as the slope of the first partition decreases toward the symmetric case reflected by the decreasing number of pairs available in either partition
- d decreases for the same reason b & c increase

Expected Results These scenarios represent situations of complete fragmentation. Even though the two partitions change in relative structure from being approximately

asymmetric to being symmetric, it would be expected nevertheless that this structural difference would not affect the results greatly. Measures should attain a value equal or very near to their worst case value.

8.4.6 Results

The following subsections describe the results of applying the five tests to the measures listed in Tables 8.4 and 8.5. For those pair counting measures in Table 8.4 the Contingency matrix was formed first, then the appropriate fields applied against each measure, while the information theoretic measures in Table 8.5 were applied directly to an intersection matrix derived from the pair of clusterings in question.

8.4.6.1 Incremental Independence of Clustering Pairs

This test determines whether a measure recognizes levels of *independent co-distribution* (see Section 8.3.1) across the range of total dependence to total independence (see Section 8.4.4.1). To present the results four tables have been generated that categorise measures against different common features. Table 8.10 presents measures that realise a fixed extreme in the case of total dependence between clustering pairs. Table 8.11 presents measures that realise a fixed extreme in the case of total independence. Table 8.12 presents measures that do not realise a fixed extreme in the cases of total dependence and independence. Finally, Table 8.13 presents measures that realise a fixed extreme in both the total dependence and independence cases. These tables provide the following key observations:

Type	Measure	Shape	Range
PC	Baroni Urbani & Buser 2, Braun & Blanquet, Dice, Dice Asym 1 & 2, Cosine, Fowlkes Mallows, Hamann, Jaccard, Kulczynski 2, Overlap, Sokal Sneath 1, 2, 4 & 5, Rogers & Tanimoto, Russell & Rao,	SAD_F	$1 \downarrow *$
PC	Baroni Urbani & Buser 1,	SAD_F	$1 \downarrow -*$
PC	Fager, McConnaughey	D_F	$1 \downarrow -*$
PC	Johnson	D_F	$2 \downarrow *$
PC	Filkov, Mirkin, Rand, Savage, Sneath Pattern Diff, Sneath Total Diff, Sokal & Sneath NM	D_R	$0 \uparrow *$
IT	Entropy Conditional, Lopez_Wan, Meila	D_R	$0 \uparrow *$

Table 8.10: Recognizing Fixed Extreme for Total Dependence

Type	Measure	Shape	Range
PC	Gilbert Wells,	SAD_F	$* \downarrow \sim 0$
PC	Michael	D_F	$* \downarrow \sim 0$
PC	Tarwid	SAD_F	$* \downarrow 0$
IT	Mutual Information	D_F	$* \downarrow 0$

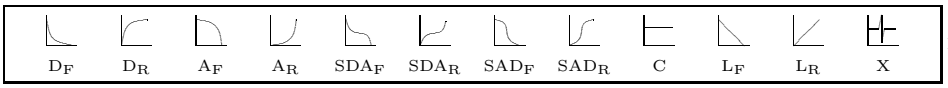
Table 8.11: Recognizing Fixed Extreme for Total Independ.

Type	Measure	Shape	Range
PC	Dennis	D_F	$* \downarrow -*$
PC	Faith, Forbes d, Fossum, Goodall, Kulczynski 1, Sokal Sneath 3	D_F	$* \downarrow *$
PC	Stiles	A_F	$* \downarrow *$
IT	Entropy Joint	D_R	$* \uparrow *$
PC	Forbes	X	$-\infty, \infty$

Table 8.12: Non-Recognition of Either Fixed Extreme

Type	Measure	Shape	Range
PC	Yules Omega	D_F	$1 \downarrow \sim 0$
PC	Mountford	D_F	$\infty \downarrow \sim 0$
IT	Lopez_Rajski	D_R	$0 \uparrow 1$
IT	Powers	D_R	$0 \uparrow 1$
IT	NMI Asy.	D_F	$1 \downarrow 0$
IT	NMI 1	D_F	$1 \downarrow 0$
IT	NMI 2	D_F	$1 \downarrow 0$
IT	NMI 3	D_F	$1 \downarrow 0$
IT	NMI 4	D_F	$1 \downarrow 0$
IT	NMI 5	D_F	$1 \downarrow 0$
*	MoD	D_F	$1 \downarrow 0$

Table 8.13: Recognizing Both Fixed Extremes



Plot Shapes Key

- If a function is a dissimilarity measure it displays an upward trend(shape $= \uparrow$), conversely if it is a similarity measure it displays a downward (shape $= \downarrow$) trend.
- The pair counting measures in Table 8.10 recognize total dependence because their numerator is wholly dependent in this situation on a or a multiple of a and d since b and c are zero.
- Because this test never realises the situation where $a = 0$ the pair counting measures in Table 8.10 do not reach their maximum for total independence as they are wholly dependent on a or a multiple a and d.
- Mutual Information would recognises total independence if it was normalised.
- There is no obvious reason why the measures in Table 8.12 do not recognize either total dependence or independence other than they are not normalised to do so.
- MoC and those other measures in Table 8.13 appropriatly attained a fixed extreme for both dependence and independence.
- With only a few exceptions most measures moved rapidly away from their best case value, addressing the requirement outlined in Section 8.3.1.

8.4.6.2 Scaling on Independent Co-distribution

This test characterises measures against three different scaling tests (see Section 8.4.4.2) to indentify those measures that realise a fixed extreme for all three scenarios. The measures are also presented in groups to enable the categorisation of measures against the following criteria:

1. Identify groups of measures that realise a fixed extreme (worst case) for all three scaling tests.
2. Identify groups of measures that realise a fixed extreme (worst case) for any one of the three scaling tests.
3. Identify groups of measures that react in the same manner under the three scenarios.
4. Identify those measures that do not fulfill categories 1, 2 or 3.

Group	Type	Measure
A	PC	Baroni Urbani & Buser 1, Baroni Urbani & Buser 2, Braun & Blanquet, Cosine, Dice, Dice Asymmetric 1, Dice Asymmetric 2, Fowlkes Mallows, Forbes, Forbes d, Fossum, Gilbert Wells, Hamann, Jaccard, Johnson, Kulczynski 1, Kulczynski 2, McConnaughey, Overlap, Savage, Sneath Pattern Diff, Sneath Total Diff, Sokal & Sneath NM, Sokal & Sneath 2, Sokal & Sneath 3, Sokal & Sneath 4, Sokal & Sneath 5, Tarwid , Yules Omega
	IT	Entropy Conditional, Entropy Joint, Lopez_Wan, Meila
B	PC	Michael, Mountford
	IT	Lopez_Rajski, Mutual Information, NMI, Powers
	*	MoC
C	PC	Faith, Filkov, Goodall, Mirkin, Rogers & Tanimoto, Russell & Rao, Sokal & Sneath 1
D	PC	Dennis, Fager, Rand, Stiles

Table 8.14: Independent Co-distribution Test Results

Table 8.14 is divided into four groupings of measures that highlight specified combinations of criterion. Group A addresses both criteria 2 and 3 by grouping measures that both realise a fixed extreme for any one of the three scaling tests, and by grouping measures that react in the same manner for the three tests. Group C addresses criterion 3 by grouping measures that react in the same manner for the three tests. Group D addresses criterion 4 by grouping measures that do not fall into any of the other categories. Group B functions produce the appropriate result described by criterion 1 by grouping measures that realise a fixed extreme for the three tests.

8.4.6.3 Incremental Conjugation of Partition Pairs

The result of conjugating a partition is a partition whose clusters have no pairs in common with its original state and is thus totally independent, one of the worst cases scenarios. Some measures however recognise structure as an measurable distributional

feature resulting in different outcomes for different shaped original partitions. This test is used to identify those measures that recognise the conjugate of a partition, either symmetric or non-symmetric, as a worst case scenario and realise a worst case fixed extreme.

Table 8.15 lists those measures that recognize structural difference and those that do not. The division in results can generally be explained by a measure either recognizing d or not, as with the inclusion of d the total space is recognized, not just some membership differential described by differences between a, b and c.

Measures that recognise cluster pair structural differences	
Type	**Measure**
PC	Dennis, Fager, Faith, Filkov, Forbes, Goodall, Hamann, Mirkin, Sneath Tot. Diff., Sokal & Sneath 1, Sokal & Sneath 3, Sokal & Sneath 4, Rand, Rogers & Tanimoto, Russell & Rao
IT	Entropy Cond., Lopez_Wan, Meila, Mutual Inf., NMI Asymmetric, NMI 2, NMI 3
Measures that do not recognise cluster pair structural differences	
Type	**Measure**
PC	Baroni Urbani & Buser 1, Baroni Urbani & Buser 2, Braun & Blanquet, Cosine, Dice, Dice Asym 1 & 2, Fowlkes Mallows, Forbes d, Fossum, Gilbert Wells, Jaccard, Johnson, Kulczynski 1, Kulczynski 2, McConnaughey, Michael, Mountford, Overlap, Savage, Sneath Pattern Diff, Sokal & Sneath NM, Sokal & Sneath 2, Sokal & Sneath 5, Stiles, Tarwid, Yules Omega
IT	Entropy Joint, Lopez_Rajski, NMI 1, NMI 4, NMI 5, Powers
*	MoC

Table 8.15: Recognition and Non-recognition of Clustering Pair Structural Differences

8.4.6.4 Comparison Based on the Combination of All Tests

There are no formal methods to categorise these measures against the manner in which they perform other than to broadly group them along lines such as being an information theoretic or pair counting based approach, or by comparing them against subjective criteria. In an attempt to categorise, and compare and contrast MoC's performance against the other measures a subjective clustering was conducted resulting in the six broad groups. The measures were grouped together if they behaved similarly across the different scenarios, "similarly" meaning that the curves proceed in the same direction (upward or downward) and between similar points. For the purpose of comparing points, consideration is made of a measure's minimal and maximal values and marked values that fall between extremes with an asterisk (as described above). So, for instance, measures in Group I are grouped together, largely because, in the case of similarity measures for example, they move from their maximal value down to a non-minimal value under the Incremental Independence scenario. The dissimilarity measures (e.g. Sokal & Sneath NM, Savage) in this group also exhibit the same pattern of behaviour but move in the opposite direction (e.g. rising from a minimal value under the same scenario). Table 8.16 is sorted into each of six groups with the behaviour of each measure under each scenario being displayed allowing the commonalities between patterns of

behaviour to be discerned.

Group I loosely consists of measures that express ratios of good to bad; that is, the number of pairs two clusterings have in common (a) divided by the number of pairs the two clusterings do not have in common ($b + c$). The majority of this group of measures are similarity measures with the exception of the Savage and Sokal & Sneath Non-Metric measures that are dissimilarity measures. These two measures still conform to being measures of good-to-bad in that the Savage measure is an inverted form of either Dice Asymmetric 1 or 2 measures depending on which is the larger, and the Sokal & Sneath Non-Metric is a comparison of the pairs two clusterings do not have in common ($b + c$) divided by the number of pairs the two clusterings have in common (a). A standout feature of this group is that it is mainly comprised of measures that for at least one test appropriately realise a fixed extreme for the worst and/or best cases. Prototypical measures in Group I include the asymmetric Dice measures ($\frac{a}{a+b}$ & $\frac{a}{a+c}$, also known as precision and recall) and Jaccard ($\frac{a}{a+b+c}$).

Group II also consists mostly of measures expressing ratios of good to bad. However, whereas Group I measures consider only contingency matrix cell a, Group II measures also take the cell d into account and express the ratios of a and b' against b and c. A prototypical measure in this group is Rand ($\frac{a+d}{a+b+c+d}$). Again, this group mainly comprises similarity measures with the exception of the Sneath Total Difference measure, Filkov's measure and Mirkin's measure, which are dissimilarity measures, and which are all based on the number of pairs the two clusterings do not have in common ($b + c$).

Group III represent measures that react similarly across the three scaling tests (see Section 8.4.6.2), realise a worst case extreme or thereabouts in the Conjugate Test (see Section 8.4.6.3) and that track sigmoidally across their ranges in the Incremental Independence test.

Group IV is a relatively small group of Information Theoretic measures grouped in this manner because they perform very similarly across all tests and all are measures of dissimilarity. The Mutual Information, Mountford and Michael measures can be seen as intermediary cases between this group and Group V.

Group V can be described as (i) for the most part, recognizing increasingly independent distributions by moving between the extreme values in their range, while (ii) being invariant in the independent case under scaling. Importantly, in the conjugate case, the measures MoC, NMI 1 and NMI 4 remain constant on thier worst-case values.

Group VI conatains all other measures that do not seem to associate significantly with any other measures.

Therefore, of all the measures only MoC, NMI 1 and NMI 4 clearly satisfied all the desired requirements (see Section 8.3). However, under close inspection of the results it was noted that for the conjugate case the Powers's measure, Lopez & Rajski's measure and NMI 5 remained constant near their worst-case value and so fundamentally also address all of the desiderata. This result clearly supports remarks by, e.g. Strehl and Ghosh (2002), Yao (1999) and Meila (2003) regarding the usefulness of Mutual Information for clustering comparison.

			Dependence to Independence		Constant Elements		Constant Fragment Size		Constant Clusters		Conjugate Pairs		Plot Shapes Key
Grp	**Type**	**Measure**	**Shape**	**Range**	**Shape**	**Range**	**Shape**	**Range**	**Shape**	**Range**	**Shape**	**Range**	
I	PC	Braun Blanquet, Cosine, Dice, Dice Asymmetric 1 & 2, Fowlkes Mallows, Jaccard, Kulczynski 2, Overlap, Sokal & Sneath 2, Sokal & Sneath 5	D_F	$1 \downarrow *$	D_F	$* \downarrow \sim 0$	D_F	$* \downarrow 0$	C	~ 0	C	0	
	PC	McConnaughey	D_F	$1 \downarrow -*$	D_F	$0 \downarrow -1$	D_F	$-* \downarrow \sim -1$	C	~ -1	C	-1	
	PC	Johnson	D_F	$2 \downarrow *$	D_F	$* \downarrow 0$	D_F	$* \downarrow \sim 0$	C	~ 0	C	0	
	PC	Forbes d	D_F	$* \downarrow *$	D_F	$* \downarrow 0$	D_F	$* \downarrow *$	D_R	$* \uparrow \sim 1$	C	0	
	PC	Kulczynski 1	D_F	$* \downarrow *$	D_F	$* \downarrow 0$	D_F	$* \downarrow \sim 0$	C	~ 0	C	0	
	PC	Sokal Sneath Non-Metric, Savage	D_R	$0 \uparrow *$	D_R	$* \uparrow 1$	D_R	$* \uparrow \sim 1$	C	~ 1	C	1	
II	PC	Rand	D_R	$0 \uparrow *$	D_F	$* \downarrow *$	D_F	$* \downarrow *$	C	~ 0	D_F	$* \downarrow \sim 0$	
	PC	Rogers & Tanimoto, Russell & Rao, Sokal & Sneath 1	D_F	$1 \downarrow *$	D_R	$* \uparrow *$	D_R	$* \uparrow *$	C	$\sim *$	D_R	$* \uparrow \sim 1$	
	PC	Faith, Goodall	D_F	$* \downarrow *$	D_R	$* \uparrow *$	D_R	$* \uparrow *$	C	$\sim *$	D_R	$* \uparrow *$	
	PC	Sneath Total Diff	D_R	$0 \uparrow *$	D_F	$* \downarrow \sim 0$	D_F	$* \downarrow *$	C	~ 0	D_F	$* \downarrow *$	
	PC	Sneath Pattern Diff	D_R	$0 \uparrow *$	D_F	$* \downarrow \sim 0$	D_F	$* \downarrow *$	C	~ 0	C	~ 0	
	PC	Filkov, Mirkin	D_R	$0 \uparrow *$	D_F	$* \downarrow *$	A_R	$* \uparrow *$	A_R	$* \uparrow *$	D_F	$* \downarrow *$	
III	PC	Baroni Urbani & Buser 1	SAD_F	$1 \downarrow -*$	SDA_F	$0 \downarrow -1$	D_F	$-* \downarrow -*$	D_R	$-* \uparrow -*$	C	-1	
	PC	Baroni Urbani & Buser 2	SAD_F	$1 \downarrow *$	SDA_F	$* \downarrow 0$	D_F	$* \downarrow *$	D_R	$* \uparrow *$	C	0	
	PC	Yules Omega	D_F	$1 \downarrow \sim 0*$	A_F	$\sim 0 \downarrow -1$	C	$-*$	D_R	$-* \uparrow \sim 0*$	C	-1	
	PC	Gilbert Wells	SAD_F	$* \downarrow \sim 0*$	A_F	$-* \downarrow -\infty$	D_F	$-* \downarrow -*$	D_R	$-* \uparrow \sim 0*$	C	$-\infty$	
	PC	Tarwid	SAD_F	$* \downarrow \sim 0*$	A_F	$\sim 0 \downarrow -1$	D_F	$-* \downarrow \sim -*$	D_R	$-* \uparrow \sim 0*$	C	-1	
IV	IT	Meila	D_R	$0 \uparrow *$	D_R	$2 \uparrow *$	D_R	$2 \uparrow *$	C	$*$	D_F	$* \downarrow *$	
	IT	Entropy Conditional	D_R	$0 \uparrow *$	D_R	$1 \uparrow *$	D_R	$1 \uparrow *$	C	$*$	D_R	$* \uparrow *$	
	IT	Entropy Joint	D_R	$* \uparrow *$	D_R	$2 \uparrow *$	D_R	$2 \uparrow *$	C	$*$	C	$*$	
	IT	Lopez_Wan	D_R	$0 \uparrow *$	D_R	$2 \uparrow *$	D_R	$2 \uparrow *$	C	$*$	D_F	$* \downarrow *$	
	IT	Mutual Information	D_F	$* \downarrow 0$	C	0	C	0	C	0	D_R	$* \uparrow *$	
	PC	Michael	D_F	$* \downarrow \sim 0*$	C	~ 0	D_R	$-* \uparrow \sim 0$	C	~ 0	C	~ 0	
V	PC	Mountford	D_F	$\infty \downarrow \sim 0$	C	~ 0	C	0	C	0	C	0	
	IT	NMI Asymmetric, NMI 3	D_F	$1 \downarrow 0$	C	0	C	0	C	0	D_F	$* \downarrow *$	
	IT	NMI 2	D_F	$1 \downarrow 0$	C	0	C	0	C	0	D_R	$\sim 0 \uparrow *$	
	IT	NMI 5	D_F	$1 \downarrow 0$	C	0	C	0	C	0	C	$\sim *$	
	IT	NMI 1, NMI 4	D_F	$1 \downarrow 0$	C	0	C	0	C	0	C	~ 0	
	*	**MoC**	D_F	$1 \downarrow 0$	C	0	C	0	C	0	C	~ 0	
	IT	Powers	D_R	$0 \uparrow 1$	C	1	C	1	C	1	C	$\sim *$	

	IT	Lopez_Rajski	D_R	$0 \uparrow 1$	C	1	C	1	C	1	C	$\sim *$
VI	PC	Fager	D_F	$1 \downarrow -*$	D_R	$-* \uparrow -*$	D_F	$-* \downarrow -*$	A_F	$-* \downarrow -*$	D_R	$-* \uparrow -*$
	PC	Dennis	D_F	$* \downarrow -*$	L_F	$-* \downarrow -*$	L_F	$-* \downarrow -*$	D_R	$-* \uparrow -*$	D_F	$-* \downarrow -*$
	PC	Forbes	X	$-\infty, \infty$	C	0	D_R	$-* \uparrow \sim 0$	C	~ 0	D_R	$-* \uparrow -*$
	PC	Hamann	D_F	$1 \downarrow *$	D_R	$\sim 0 \uparrow *$	D_R	$-* \uparrow *$	C	$\sim *$	D_R	$-* \uparrow *$
	PC	Fossum	D_F	$* \downarrow *$	D_F	$* \downarrow \sim 0$	A_R	$* \uparrow *$	A_R	$* \uparrow *$	C	~ 0
	PC	Stiles	A_F	$* \downarrow *$	D_R	$-* \uparrow *$	D_R	$-* \uparrow *$	C	$\sim *$	C	$\sim *$
	PC	Sokal & Sneath 3	D_F	$* \downarrow *$	L_R	$1 \uparrow *$	L_R	$* \uparrow *$	C	$\sim *$	D_R	$* \uparrow *$
	PC	Sokal & Sneath 4	D_F	$1 \downarrow *$	C	~ 0	D_R	$* \uparrow *$	C	$\sim *$	D_R	$* \uparrow *$

Table 8.16: Groupings of measures across. [5]

[5] Sub-groups of identical characteristics separated by lines.

8.5 Conclusion

Given the widespread use of clustering techniques in the presentation of apparently contextually pertinent data to humans and the general reliance on fixed heuristics rather than dynamic human context, there is much scope for fundamental research in this area.

Toward this end, this Chapter has focused on comparing the relative subgroup memberships of two clusterings. To study pair counting measures, 2×2 *contingency matrices* (see Figure 8.1) were used as a convenient way to summarize the relationships between the memberships of two subclusters. Although contingency tables are traditionally used to compare two populations they have been used here to compare two partitioned spaces through the application of a pair counting approach to assign values to the individual fields of the matrix. Key relationships between clustering pairs are identified by comparing relationships between the occurrence of member pairs, member non-pairs and member pairs that do not occur in common using the 2×2 contingency matrix. In addition a number of information theoretic measures were also investigated.

To compare clustering pairs using external features the two types of worst case were identified as unrelated clustering pairs and opposite partitions which are described as *Independently Codistributed Clustering Pairs* and *Conjugate Partition Pairs*. These situations were applied to the development of a measure (MoC) to appropriately recognise these cases and in the search for other measures that may react similarly or the same. They were also used to identify groups of measures with similar features to allow researchers to choose between general classes of measures exhibiting similar behavior.

MoC's logical development was supported by five tests used to demonstrate the characteristics of MoC and a selection of other measures. The individual tests produced distinct groupings as did the combined results.

The combined results (see 8.4.6.4) demonstrated that the measures in Group V conformed to many of the desiderata as stated in Section 8.3. In particular, the MoC measure, Powers's measure, Lopez & Rajski's measure and the six Normalized Mutual Information measures complied with the requirements 1, 2, 4, 5 and 6. As for requirement 3, these measures recognised the worst case of mutual independence between clusterings whereas only MoC, NMI 1 and NMI 4 strictly recognised the conjugate case (corresponding to complete fragmentation) by realising their worst case value. However, for all intents and purposes the Powers's measure, Lopez & Rajski's measure and NMI 5 appropriately recognise the conjugate case by remaining constant at a value near their worst-case value. By contrast, the unnormalized information theoretic measures in Group IV were all found to combine an absolute measure of goodness into

the comparison viz. there is an extra term that reflects the information content of the clusterings.

Chapter 9

Epilogue

This section summarises the contributions of this thesis, outlines the main conclusions, and points the way to future and ongoing work.

The ever expanding mass of data is a restriction in time-critical human-decision based processes reliant on automated text search systems. While automated systems lack the insight humans bring to the decision making process they do offer brute force iterative processing power applicable to surface processing large amounts of data quickly. Given this situation it stands to reason, as it has been suggested, that a logical marriage between these capabilities is to use machine processing to first-pass filter the data-avalanche resultant from a textual search. Human context is then introduced by allowing the human to look through broad categories derived from the data allowing them to throw away those categories not relevant to their search to result in a smaller more finely tuned return set. Finally, machine processing is used again in the application of traditional ranking techniques to form the return set into a relatively small highly accurate ranked list.

This whole thesis is based on the preposition that technology can accelerate the information acquisition process, specifically in searching for textual data and that it should make the process more efficient (i.e. easier, more accurate and faster). As described in Chapter 1 the work presented here originated in the context of previous work by Pfitzner et al. (Pfitzner et al. 2003, Pfitzner & Powers 2004) that proposed techniques and tools to guide the appropriate use of visual screen artifacts/devices/cues when designing search interfaces that present multi-dimensional data, specifically textual documents. In that work it was concluded that only textual languages provide an adequate conduit for the communication of fine grained difference between visual clusters of documents. Clearly, the manner in which documents or groups thereof are described using *words* will affect search efficiency. For example if one word is used to

visually describe a document the user is not going to have enough information to correctly classify it or even complete the task. At the other extreme if the whole document is used the user will spend far too much time reading individual documents to identify classifying features. Somewhere along this continuum, is an optimal descriptor length, but where?

The process of identifying useful classifying words is well researched, however traditional search systems use techniques that employ fixed heuristics (not based on user research) to guide the selection of classifier words and calculate their weightings. For example, the most popular weighting scheme used to find the most the characterizing words of a document is one known as TFIDF (Text Frequency Inverse Document Frequency). This scheme is a fast calculation that weights the words of a document given their raw document frequencies correct by the reciprocal of the number of documents they occur in across the total corpus. Variants of TFIDF are used by all the major search engines, however TFIDF does not rely on any model of cognition or recognize in any way user capacity limits or tendencies.

Despite this lack of a valid cognitive model justifying the use or applicability of TFIDF there was no research into what positive or negative effects such fixed heuristics might have, given that users will have varying information requirements, cognitive tendencies/abilities/preferences and language usages. This comes from the apparent observation that users are not homogeneous, having different cognitive traits and tendencies, and will often react differently to the same situation/question/information, so will require a system that allows for their tendencies and/or variances of ability. From these observations it was proposed that TFIDF does not and can not reflect knowledge of intent or individual ability and experience.

With respect to user cognitive ability (see Section 3.1) there are clearly limitations regarding the number of *chunks* of information (words) that users can optimally manage at any one time (e.g., 7 ± 2 or 4 ± 1). These limits can also be described as preferences because when a reduction in task performance is noted, for a given task, it can be unclear whether a biophysical limit has been realized (e.g. the user naturally manages 4 chunks not 7) or a personal selective preference/tendency has been realised (e.g. the user is normally a bit lazy so does not search as far down a list before reformulating the query). The implication of such user limitations is that for any system to promote the best possible task outcome it must allow for such user characteristics/limits by applying either an appropriate user model or reliably identified general user tendencies.

Thus, we come to the research of this thesis:

> "This thesis investigates the number and type of words needed to best describe documents individually and in clusters."

Basically, this finds its origins in the earlier suggestion that the design of the "ultimate search system" will include the presentation of document clusters that allow the user to rapidly reduce the return set by throwing away clusters of documents (topically related) which have been selected primarily using cluster descriptors or by drilling down and using the document descriptors within a cluster.

The main hypothesis of this thesis regards the number and types of words and is divided into the following two parts:

1. Because the popular TFIDF like weighting schemes are based on frequency statistics and not an appropriate user model or reliably identified general user tendencies they will produce ranked list of words for documents the heads of which do not match those a user might produce for the same documents. Thus the types of words users use to describe a document will be different from those produced by the commonly used automated processes.

2. Given researched cognitive limits such as those represented by the magic numbers 7 ± 2 or 4 ± 1 (see Section 3.1.1) and their associated chunks of information users will prefer document descriptions of between 1 and 9 characterizing words (chunks). The range described by Cowan is more likely to be favoured given the human bias toward energy conservation in activities like search, as demonstrated by O'Brien and Keane (O'Brien & Keane 2007). In other words users will tend to use as few words as possible to describe a document. Related to this bias is the tendency of most users to select the first member of a search returns list without any real inspection of data presented. After this initial selection they, in a similar manner, sequentially select down the list until they reach some threshold at which they alter their search technique to a more energy consuming approach. These approaches see the user surveying in more depth the associated snippets for each entry before selecting.

To build a compound understanding of the state of current knowledge and opinion Chapters 2, 3 & 4 reviewed the literature on aspects of cognition relevant to user interaction and the task of visual search.

Chapter 2 dealt specifically with those cognitive mechanisms that impact the user's decision making, focusing on cognition relative to the processing of information, specifically memory, attention and cognitive styles, and any notable impacts on the task of interactive search. The chapter made the following observations:

- Because of the volatile nature of *sensory memory*, information pertinent to the completion of the task, especially the immediate sub-tasks, should remain while

it is contextually relevant or until it is no longer required to complete any relevant tasks.

- Visual tasks requiring time on the order of several seconds to a minute may also exploit short-term memory.
- Task that require more than a minute may require the rehearsal of critical information through the presentation of cues.
- To free capacity for other more challenging tasks, repetitive tasks involving screen artifact interaction should see the artifacts kept constant throughout the process. That is they should look the same, do the same thing and appear in the same position.
- Individual differences in cognitive style should be reflected in the optimization of features of interactive interfaces.
- Screen artifacts can be made visually salient to exploit rapid preattentive processing.
- Finally, simple interactive displays minimise the risk of working memory interference.

Chapter 3 discussed user cognitive limitations that give an indication as to how many words a cluster or document descriptor should contain. The interest in any such limitation results from the premise that any cognitive processes are based on *physical* biological systems which ensures that user cognitive processes will logically have limitations that will affect the amount of information a user can process at any one time.

This type of review is important given the problem of Data-avalanche in document search and that the proposed "ultimate search system" would allow for the rapid reduction of unmanageably large search return sets by getting the user to evaluate and discard large inappropriate categories (clusters) of documents. Given textual language efficiently conveys fine-grained topical details about textual documents, it is appropriate to describe the topical content of clusters and individual documents using text. However, if the user's abilities are not appropriately recognised when generating the cluster descriptors, the user will realise a less than optimal task outcome.

The chapter made the following observations:

- Short term memory is limited in capacity and so, if search tasks and sub-tasks can be tailored so that visually transmitted information can be naturally realised in chunks, the user is more likely to realize a better task outcome.

- Small collections of randomly arranged items can be subitized. Larger collections are counted and the success of counting is dependent on the manner in which the objects are displayed.
- Because of the volatility of short term memory, it is important not only that users have structures (schemata) to aid in remembering information, but also that they are not required to remember them for an extended period of time.
- Interaction device design should draw on the experience of the expert in the delivery of information because they will have a better understanding of what information is important in a specific task.
- Interface design should recognise the effects of experience and expertise.
- Visual afterimage phenomena may be exploited in order to provide the user with a task relevant residual image.

Overall, this chapter points to the need to manage the number of things in chunks and groupings to optimize the realization of any interactive text search tasks. The thesis draws on the research presented, to target the number and type of words needed to identify the topic of a cluster of documents or document in a visualisation ("How many words do people naturally use to describe/query for documents?"). This research program has also spawned and supported related projects not discussed in this thesis that examine the impact of visual attributes (Treharne et al. 2008, Treharne et al. 2007, Treharne et al. 2006) and emotional cues (Powers et al. 2008).

Chapter 4 investigated visual processing looking at the effects the visual system has on the interactive search/filtering task. The discussions looked closely at visual aspects of user information realization such as "what do we see", "how do we see it" and "what are the general affects on cognition". The chapter made the following concluding observations:

- Search return documents should be graphically presented as clusters that allow the user to dispose of irrelevant clusters of documents and thus speed the filtering of large return sets to more manageable sizes.
- Colour can be used effectively to make targets pop-out of a field of potential targets.
- The level of visual detail used in a display should be tailored to the requirements of the task.

- Appropriate use of low-level visual system and preattentive processes may allow attention to be more efficiently drawn to areas of potential interest in the display.
- Three key factors that critically effect visual attention are the manner in which space is used relative to object dispersion (size of display), how objects are grouped relative to other task relevant objects and distractors via common visual traits such as motion, proximity and colour, and lastly previous experience (see Section 4)

The work in this chapter has motivated and fed into Post Graduate work currently being conducted by Kenneth Treharne. This work is taking a fine grained approach to investigating the effects different display attributes and artifacts such as motion, colour, proximity, size, shape and perceived 3D effects can have on user interactive task performance. In moving toward the realisation of "The Dream" the target of the research is to identify how display attributes and artifacts can be used to realise better task outcomes from both a physical and cognitive perspective.

Chapter 5 acknowledges the claims and observations, made in the previous chapters, that human cognition is a key factor in tasks such as interactive text search and that for any search system to be contextually effective it must rely on an effective model of human cognition. This acknowledgment is followed by an analysis of the field of user modeling in the context of the document search task and the understanding of user internal processes and preferences.

The chapter demonstrated the applicability and usability of TLA (Transaction Log Analysis) and the manner in which it might be implemented to identify characteristic user preferences/thresholds in describing visual textual objects. It discussed and highlighted some of the characteristics of general Web search and usage statistics obtained using TLA. Finally, several flaws with the use of search engine TLA were identified that should be systematically addressed if any experiment to identify general textual searcher characteristics is to be sound.

In short the chapter supplied reasoning for the Nwords Surveys and background for their design.

It became clear from the research in this chapter and the subsequent development of Nwords and InFields research surveys that future work is needed that investigates techniques that allow the gathering of data in naturalistic settings while delivering statistically sound results.

The main empirical results of this thesis are presented in Chapters 6 to 8.4.

Chapter 6 describes the Nwords surveys, outlined the results and discussed how the results supported the two parts of my thesis.

The chapter introduces the important observation that "in the quest for user models, there has been little decontextualized research into user cognitive limits and preferences relative to the number of words a user might use to describe a document". The chapter later describes how Nwords answers several problems inherent in conducting research in a highly controlled environments by still being controlled but being delivered in an environment the participant is likely to be comfortable and familiar with (the real world).

As the core data collection and analysis section of this thesis, Chapter 6 is the primary vehicle for testing my hypothesis. It did this via the Nwords surveys which were designed to quantify the number of words a broad spectrum of participants use to describe different blocks of text and hence the appropriate number of words/chunks/dimensions needed to describe individual documents and clusters of documents, and to manage the impact of process intensive clustering processes. A secondary objective is to enhance understanding of choices a user makes in selecting keywords or phrases to describe or search for a document. To do this Nwords is comprised of four different experiments presented in the form of surveys using a common look and feel Web interface (for experiment/survey descriptions see Section 6.2).

The chapter concluded in support of my two part thesis with the following observations, as well as making the important supporting observation that TFIDF does not relate well to human selection tendencies:

- Participants used 2 to 3 times the number of distinct words to describe a document than distinct words to search for the same document.
- Given it has been demonstrated that participants generally use, on average, **six** distinct stems to describe a document compared to **four** distinct stems to search for the same document, two subsequent observations can be made:
    1. On average only 33.33% of stems used to describe a text will also be in the top ten TFIDF ranked stems.
    2. On average only 25% of stems used to query for a text will also be in the top ten TFIDF ranked stems.

The Nwords research resulted in one a very important observation relative to two well recognised models of cognition, those of Miller's magic number 7 ± 2 and Cowan's number 4 ± 1. Relative to Miller's number it was noted that participants generally used 5 to 8 distinct stems to describe a document. This is an important observation as it implicitly supports Miller's proposed limit (see Section 3.1.1) of 7 ± 2 as being

appropriate in its use as a "rule of thumb" to describe human tendency in document description formulation.

As for Cowan's number, in Section 6.1.1 it was suggested that Cowan's number (see Section 3.1.3) was more likely the rule applicable in the description of how many words people might use to describe a document. Although it was demonstrated that this is not the case it was shown, using observations from this work and other Web statistics and TLA research (see Sections 5.4 & 5.5), that Cowan's number 4 ± 1 is an appropriate "rule of thumb" for the description of the response tendency in query formulation.

Further to this, when examining the set of human query stems across all tasks it was noted that on average a minimum of one word does not occur in the description stems set. Given the small numbers of query stems normally used, it is evident that the terms used to query for a document will be substantially different from those used to describe the same document. This is indicative of different cognitive processes being involved which in turn indicates that Miller's number and Cowan's number are heuristics that are both useful in representing human preference but in different situations.

Finally, TFIDF is generally used to describe the representativeness of textual information for a given block of text relative to an associated corpus. I propose that if TFIDF is intended to reflect human judgment in some manner then it is fair, given its ubiquity in the document retrieval field, to expect that it would exhibit a reasonable level of psychological relevance. However, given the small size of the intersections between survey participant selected terms and those generated using a TFIDF algorithm it is evident that TFIDF **does not** reflect human preference to any reasonable degree. Furthermore, it is also evident that TFIDF is more representative of human preference in the task of text description as seen in participant generated description stems being substantially more likely to intersect with the TFIDF list than participant generated query stems.

Although some of thesis suggests that TFIDF is an inappropriate measure to use where user information context is communicate via context words it is still a particularly usefule measure. Given the observations made at the end of Section 7.1 that suggest that TFIDF equation (4) matches experienced users best and equation (1) matches inexperienced users, future research is proposed to address aspects of this observation. Specifically, do different word weighting techniques better model word preferences for different user groups.

Because of TFIDF's ubiquity and the observations that it does have a level of relevance seems to motivate future work. This might see further investigation of alternatives or improvements to TFIDF that better model human preference. Part of this investigation might include research using the results/terms of the Nwords surveys

and term expansion techniques using systems like WordNet in consort with TFIDF to help TFIDF weight classifications that would represent different words the the same or similar meanings.

Chapter 7 documents two further experiments designed to support the Nwords research. These experiments were the Rwords and InFields experiments. In designing the Nwords surveys the TFIDF weighting scheme was needed to rank various word lists. Given there are several variants of TFIDF, Rwords was designed to identify which variant of TFIDF performed the best relative to human judgment.

The Rwords experiment demonstrated that when presented with lists derived from four different TFIDF algorithms participants clearly preferred two approaches. The results indicated that TFIDF equation 7.1 and 7.4 performed similarly and that they performed better than equation 7.2 and far better than equation 7.3. It was also informally suggested that equation 7.4 matches experienced users best and equation 7.1 matched inexperienced users.

The Infields research resulted from the analysis of the Nwords survey results which highlighted a possible flaw in the manner in which participants were asked to input their answers (see Section 6). It was suggested that the shape of the input fields and associated mechanisms might have influenced the number of words used to describe and query for a document. Evidently, this potential flaw might render the relevant portion of the results irrelevant to the goal of the research. To investigate this situation the "InFields" experiment was designed to describe participant word/term input characteristics under a variety of input field characteristics and task types. The primary goal of the InFields research was to determine if the different input field sizes and mechanisms used in the Nwords experiment might have been influencing the words and terms input by participants in two common language based tasks.

The InFields research resulted in three important conclusions:

1. The first was that the input mechanism **did** influence the number of **terms** used by participants. However, although an important observation, this is of **no** consequence to any conclusions made in the Nwords experiment.

2. Secondly, in support of the Nwords findings, the input mechanism did **not** influence the number of distinct stems used by participants to describe or query for a text.

3. Finally, and again in support of the Nwords findings, the different input mechanisms did **not** affect the number of stems participants use that also intersected the top ten TFIDF list.

The Chapter also looked at comparing *clusterings* for the purpose of identifying which clustering approaches are best used in the creation of document clusters for the user cluster filtering (throwing away) approach described earlier. Given the user filtering process the set of document clusters (clustering) used should be composed of clusters that relate in a manner the user might reasonably assume such as by the topic content a user is likely to describe for a document or group of documents. That topic content the user might realize is important, given part 1 of my thesis suggests that automatic approaches might realize different keywords than a user. Therefore, future work should include the comparison of automatically generated document clusters should be conducted against manually generated "Gold Standard" and the results of different clustering approaches compared to see which best match the "Gold Standard".

Chapter 8 was the first of two chapters that look at dissimilarity/similarity measures and their testing for applicability to the process of clustering documents as similarly as possible to that of a human. This stems from the proposition that if search return documents are to be presented for user interactive context filtering the clusters need to approximate the user's selection model as closely as possible. Although Nwords was designed to elicit data to support research into better techniques for cluster realization based on limited sets of descriptive words this work is future work and beyond the scope of one PhD. As such, current clustering approaches need to be assessed for optimal cluster realization in the short term and to supply a set of optimal standard applications that can be used for comparison purposes in future research that will result from this thesis.

In short, the chapter characterizes a number of similarity/dissimilarity measures as applied to the context of comparing a pair of clusterings. These include measures previously proposed for this problem, and a host of other similarity/dissimilarity measures that, although they have not previously been applied to clustering comparison, are applicable.

Chapter 8.4 follows from Chapter 8 and introduces the novel *Measure of Concordance* (MoC) evaluating it and a number of other similarity/dissimilarity measures, identified in Chapter 8.4, as applied to the context of comparing a pair of clusterings. These included measures previously proposed for this problem, and a host of other similarity/dissimilarity measures that, although they have not previously been applied to clustering comparison, are applicable. The critical conclusion from this sequence of research was that the MoC measure, Powers's measure, Lopez & Rajski's measure and the six Normalized Mutual Information measures all complied with the described requirements.

Future work that naturally follows from this research is the testing and comparing

of the compliant measures relative to each other. This will see the compliant measures applied to the comparison of the human clusterings found in the Document Understanding Corpus (DUC) and those produce from the same documents using a selection of common clustering approach.

Chapters 8 & 8.4 outline inital research regarding the identification of those clustering techniques that model user document categorization tendencies, for comparison purposes in future research. A natural continuation of this research is the investigation of whether those results presented hold for real data. Exploration is currently being conducted in a number of settings, including clustering in document retrieval, human-computer-interaction modelling, and in evaluating the unsupervised induction of lexical categories from real linguistic data.

In short this thesis presents support for my two part hypothesis and describes subsequent work targeted at taking the next step in "The Dream" by presenting:

1. a substantial review of the field of cognition relative to cognitive performance and perception in interactive search,
2. several sequences of research (Nwords, Rwords and InFields) the results of which support my hypothesis,
3. a background review of similarity/dissimilarity measures that might be used to compare clustering,
4. a new and novel *Measure of Concordance* (MoC) for use in clustering comparison and evaluates it and a number of other measures in the context of comparing clusterings.

What an Adventure!

Chapter 10

Appendices

10.1 Search Engine Returns Comparison

The following vastly different lists of search engine results, which were generated using the query "dog train security", demonstrates that different search engine internal heuristics and indexing characteristics, such as term/phrase weighting schemes, stopping techniques and stemming techniques impact the result of each individual search.

altavista - http://www.altavista.com
SHOP.com
nppsecurityservices.com/patrol_guard_sniffer_dog_training.html
excellentdogtraining.com
homes.aol.com/parkq8lz09e/wgmpjspfo/security − dog − training.html
ehow.com/how_2050732_kennel − train − stray − dog.html
bullwrinkle.com/training/heel.htm
cbs3.com/local/AMTRAK.Security.Train.2.657302.html
bullwrinkle.com/training/stand.htm
handi − dogs.org/training.htm
homes.aol.com/bowieqm0y4i0/tjl/security − dog − training.html

Google - Google.com
guarddogtraining.com.au/
k9centre.com/
dpi.vic.gov.au/dpi/.../A674673D760135E9CA256C160022044...
dpi.vic.gov.au/DPI/.../6AAE45F591CAC5F7CA256D780013EF...
abc.net.au/tv/waggingschool/training/ep2.htm
books.google.com.au/books?isbn = 0811729621...

citytrain.com.au/about_your_trip/security_safety/
security_staff/security_staff.asp
books.google.com.au/books?isbn = 0658010433...
smh.com.au/news/world/train − kills − woman − trying − to −
rescue − blind − dog/2008/09/23/1221935581896.*html*
abwa.asn.au/guidedogsinwa.htm

Alexa - http://www.alexa.com/search
dog.com/articles/howtocratetrain.asp
frontrangefrenzy.com/dogtraining/moredetails.h...
miscojobs.com/employers/resumes/L_1/C_6/rsm_23...
threecb.com/pets/DogTraining.html
topdogsecurity.co.uk/giz.htm
topdogsecurity.co.uk/paul.htm
petcaretips.net/crate_training_your_dog.html
*pets1st.ca/articles/*00047*CrateTrainYourDog.asp*
dooziedog.com/dog_training_articles/crate
alldogsadvice.com

Exalead - http://www.exalead.com
basenji − dog.com
boxer − dog.org/item/17?*memberid* = 2
dog − trainings.net/drug − dog − training
how − to − train − a − dog.com
stores.ebay.com/Velocity − Products − Inc_Personal − Security
*_W*0*QQcolZ*2*QQdirZQ*2*d*1*QQfclZ*1*QQfsubZ*9*QQftidZ*2*QQtZkm*
how − to − train − a − dog.com
dog − housetraining.com/search/housetrainingdog_2/index.php
thepeacetrain.org/modules.php?name = *News&file* = *article&sid* = 17
maddad.blogspot.com/2004_11_01_*maddad_archive.html*
junglebetti.diaryland.com/060719_7.*html*

10.2 Nwords Results

	Desc	Desc stms	dstct Desc stms	Desc stm X	dstct Desc stms X	dstct Desc stms in top 10 TFIDF list	Qry stms	dstct Qry stms	Qry stm X	dstct Qry stm X	dstct Qry by Desc stm X	dstct Qry stms in top 10 TFIDF list	Desc stms minus Qry stms in top 10 TFIDF List
Survey Type 1		**With access & words Need not occur**				**Number of Participant**		57					
Tot	223.00	562.00	493.00	69.00	54.00	117.00	208.00	204.00	4.00	3.00	125.00	59.00	
Avg	3.91	9.86	8.65	1.21	0.95	2.05	3.65	3.58	0.07	0.05	2.19	1.04	1.02
Std Dev	3.36	9.31	7.32	2.49	1.78	1.64	2.18	2.01	0.42	0.29	1.85	0.96	1.46
Std Err	0.06	0.16	0.13	0.04	0.03	0.03	0.04	0.04	0.01	0.01	0.03	0.02	0.03
0 Cnt						10						19	
0 % of tot						0.1754						0.333333333	
Survey Type 2 - No access			**words Need Not occur**							Number of Participant		48	
Tot	204.00	406.00	368.00	38.00	33.00	96.00	197.00	194.00	3.00	3.00	125.00	61.00	
Avg	4.25	8.46	7.67	0.79	0.69	2.00	4.10	4.04	0.06	0.06	2.60	1.27	0.73
													0.00
Std Dev	2.32	5.23	4.80	1.20	1.01	1.50	2.01	1.89	0.32	0.32	1.66	1.09	1.45
Std Err	0.05	0.11	0.10	0.03	0.02	0.03	0.04	0.04	0.01	0.01	0.03	0.02	0.03
0 Cnt						9						13	
0 % of tot						0.1875						0.270833333	
Survey Type 3 - With access			**words Must occur**							Number of Participant		64	
Tots	322.00	832.00	760.00	72.00	64.00	176.00	280.00	265.00	15.00	13.00	192.00	90.00	
Avg	5.19	13.42	12.26	1.16	1.03	2.84	4.52	4.27	0.24	0.21	3.10	1.45	1.39
Std Dev	3.27	12.48	10.62	2.35	1.97	1.97	2.95	2.57	0.72	0.60	1.94	0.95	1.74
Std ERR	0.05	0.20	0.17	0.04	0.03	0.03	0.05	0.04	0.01	0.01	0.03	0.02	0.03
0 Cnt						4					zeros count	8	
0 % of tot						0.064516129						0.129032258	
Aglomerate Figures for Tasks 1 & 2 & 3													
	Desc	Desc stms	dstct Desc stms	Desc stm X	dstct Desc stms X	dstct Desc stms in top 10 TFIDF list	Qry stms	dstct Qry stms	Qry stm X	dstct Qry stm X	dstct Qry by Desc stm X	dstct Qry stms in top 10 TFIDF list	Desc stms minus Qry stms in top 10 TFIDF List
Totals	749.00	1800.00	1621.00	179.00	151.00	389.00	685.00	663.00	22.00	19.00	442.00	210.00	
Average	4.45	10.58	9.52	1.05	0.89	2.30	4.09	3.96	0.12	0.11	2.63	1.25	1.04
Std Dev	2.98	9.01	7.58	2.01	1.59	1.70	2.38	2.16	0.48	0.41	1.82	1.00	1.55
Std ERR	0.05	0.16	0.13	0.04	0.03	0.03	0.04	0.04	0.01	0.01	0.03	0.02	0.03
0 Cnt						7.67						13.33	
0 % of tot						0.14						0.24	

Table 10.1: Results for nWords survey tasks 1-3

	With access & word need Not occurr	No access & word Need occur	With access & word Must occur	
Averages for	(Survey 1)	(Survey 2)	(Survey 3)	Totals
Number of participants	57	48	64	169
Descriptions per doc	3.91	4.25	5.19	4.45
Description Stems	9.86	8.46	13.42	10.58
Distinct Description Stems	8.65	7.67	12.26	9.52
Distinct description stems in top 10 TFIDF	2.05	2.00	2.84	2.30
Query stems	3.65	4.10	4.52	4.09
Distinct query stems	3.58	4.04	4.27	3.96
Distinct query stems in top 10 TFIDF	1.04	1.27	1.45	1.25
Number of NULL description stem TFIDF intersections	10	9	4	23
Number of NULL query stem TFIDF intersections	19	13	8	40
NULL description stem TFIDF intersections as a % of total number of participants	0.18	0.19	0.06	0.14
NULL query stem TFIDF intersectionsas a % of total number of participants	0.33	0.27	0.13	0.24
Descriptor stems minus query stems in top 10 TFIDF				
Average	1.02	0.73	1.39	1.04
Standard Deviation	1.46	1.45	1.74	1.55
Standard Error	0.03	0.03	0.03	0.03

Table 10.2: Statistics for nWords survey tasks 1-3

10.3 Nwords Error Removal

The following record has been pick randomly from the results log of the Nwords survey as an example of a result without error.

> Sat Aug 5 11:21:44 CST 2006,1824,Female,1,0,duc_manual_processed/2002 _processed/d070_processed/fbis442178_processed/fbis442178.txt, Erich Honecker,6, Honeckers Death,7, Becker,4, Court Case,5, Santiago Chile,3,#,3, Honecker Death Court Case,11:28:2711:31:36+11:31:3611:31:51+11:31:52-11:32:6

Following are the records removed from the results log of the Nwords survey as a result of applying the filtering rules outlined in Section 6.3:

> Tue Nov 1 14:24:27 CST 2005 , 45+, Male, 2, , duc_manual_processed/2003_processed/ d30010_processed/ nyt19981106.0494_processed/ nyt19981106.0494.txt, fish, 7, #, 7, and, 14:26:1714:26:19+14:26:1914:26:43+14:26:4414:27:26+14:27:26-14:27:51

> Tue Nov 1 14:24:28 CST 2005 , 1824, Male, 3, 0, duc_manual_processed/2004_processed/ task_1_2_processed/ d30059t_processed/ APW19981123_0274/APW19981123_0274.txt, chicken salad, 4, #, 7, omega, 14:27:1614:27:31+14:27:3114:27:40+14:27:40-14:27:52

> Tue Nov 1 14:25:00 CST 2005 , 45+, Male, 3, 0, duc_manual_processed/2003_processed/ d100_processed/ apw19990519.0113_processed/ apw19990519.0113.txt, chicken, 7, #, 7, the, 14:27:5614:28:14+14:28:1414:28:21+14:28:2114:28:25

> Sat Nov 5 00:01:13 CST 2005 , 2430, Male, 1, , duc_manual_processed/2003_processed/ d120_processed/ xie19970904.0283_processed/ xie19970904.0283.txt, water resource, 5, displacement comunities, 4, #, 2, characters and fontsCharacters and fonts, 23:59:150:2:0+0:2:10:4:3+0:4:40:4:44

> Fri Feb 10 11:25:13 CST 2006 , 1824, Female, 2, 0, duc_manual_processed/2004_processed/ task_5_processed/ d133c_processed/ APW19981105_0808/APW19981105_0808.txt, I didnt understand it, 1, #, 1, Bin Laden, 11:29:3911:29:47+11:29:48-11:30:24+11:30:2511:30:42+11:30:4211:31:5

Fri Feb 10 12:09:45 CST 2006 , 1824, Male, 1, , duc_manual_processed/2004_processed/ task_1_2_processed/ d30001t_processed/ APW19981116_0205/APW19981116_0205.txt, terrible, 2, #, 4, hun sen, 12:10:2512:14:17+12:14:1812:14:39+12:14:39-12:15:38

Wed Dec 6 23:28:29 CST 2006 , 3045, Female, 2, , duc_manual_processed/2003_processed/ d110_processed/nyt19980727.0091_processed/ nyt19980727.0091.txt, research science bank nwords, 6, #, 2, nwords, 0:10:310:10:45+0:10:53-0:11:14+0:11:140:11:26+0:11:260:11:41

Wed Dec 6 23:29:18 CST 2006 , 3045, Female, 2, 0, duc_manual_processed/2001_processed/ d34_processed/la0801890042_processed/ la0801890042.txt, your, 1, method, 1, is, 1, idiotic, 7, #, 2, bad science, 0:11:460:11:48+0:11:480:12:12+0:12:13-0:12:20+0:12:210:12:30

10.4 The Standard Document used in the InFields Survey

This section present the standard document presented to all participants in the InFields research described in Section 7.2.

Maybe, just maybe, customers who pay to use bank ATM machines are beginning to fight back. Or maybe they're just getting smarter.

Many are shifting to their own banks' machines to avoid being charged twice for their transactions – one fee by the bank that owns the ATM, if it is a bank other than their own, and another fee by their own bank. At the same time, the surcharges continue to rise, new studies show.

The surveys were seized by Sen. Alfonse D'Amato (R-N.Y.), chairman of the Senate Banking Committee, who said recently he'll get the Senate to vote this year on legislation to ban banks' practice of making double charges on ATM transactions.

In recent years, the number of ATMs has greatly expanded. There are now more than 150,000 automated teller machines nationwide, and there were almost 11 billion ATM transactions last year, according to the American Bankers Association.

Along with the rise in ATM use, however, has come a jump in the fees charged by banks to use the ATMs. Use an ATM owned by a bank other than your bank, and that "foreign" bank may charge you as much as $1.50.

In fact, the charge can come with a double whammy, a $3 charge if your bank also charges $1.50, which some do, for processing your withdrawal from the other bank's ATM.

"It's ridiculous that customers sometimes are charged twice to get access to their own money," said Robert Pregulman, an Atlanta spokesman for the U.S. Public Interest Research Group, a national consumer group.

That consumer sentiment apparently is reflected in two recent surveys that indicate some consumers have begun limiting their use of "foreign" bank ATMs.

A survey in April of 3,100 consumers by PSI Global, a market research firm, found heavy ATM users – those who conduct five or more transactions each month – had reduced their use of "foreign" ATMs by 22 percent in the last two years. They had reduced all ATM transactions by 4 percent.

"The behavior of the heavy users, the shifting away from 'foreign' ATMs and declining transactions, might be an indication that the market has reached, or nearly reached an optimal price point," said Mimi Rossetti, director of research for Tampa-based PSI Global.

Most banks don't charge their own customers for using bank-owned ATMs, so more consumers are using them to save ATM fees. Also, Rossetti said consumers who use "foreign" ATMs are apparently withdrawing larger amounts of cash to reduce the number of fee-charged withdrawals.

In addition to the "heavy users," PSI Global found that all ATM users in its survey had reduced their use of "foreign" ATMs by 18 percent during the last two years.

A March survey by Market Facts Inc., conducted for the ABA, found nearly two of three consumers said they had changed their ATM use to avoid paying fees. But the ABA used the survey to justify charging fees for the "convenience" of other banks' ATMs.

"The fees enable banks to set up many more ATMs on street corners, and in airports, grocery stores, malls, convenience stores and hospitals," said Walter A. Dodds Jr., ABA president. "Why should institutions give convenience away, for free, to people who don't bank with them?"

10.5 InFields Research Results

Results for the Description Task using Keyword Input Field						
	terms	description stems	distinct descriptions stems	description stem intersections	distinct description stem intersections	distinct description / top 10 TFIDF intersections
	3	7	6	1	1	2
21	2	4	4	0	0	3
	2	5	4	1	1	2
	4	7	6	1	1	2
	2	2	2	0	0	2
	1	9	9	0	0	3
	1	5	5	0	0	1
	2	6	5	1	1	3
	2	2	2	0	0	2
	6	10	7	3	2	4
	3	3	3	0	0	1
	3	5	4	1	1	3
	2	2	2	0	0	1
	1	2	2	0	0	1
	5	10	6	4	3	3
	2	5	4	1	1	2
	5	8	6	2	2	3
	1	1	1	0	0	1
	1	2	2	0	0	2
	3	3	3	0	0	2
	3	3	3	0	0	2

Averages	2.57	4.81	4.10	0.71	0.62	2.14
Std Dev	1.43	2.80	2.05	1.10	0.86	0.85
Std Err	0.31	0.61	0.45	0.24	0.19	0.19

Results for the Query Input Field for the Description Task

Participant count	terms	description stems	distinct descriptions stems	description stem intersections	distinct description stem intersections	distinct description / top 10 TFIDF intersections
	3	3	3	0	0	3
23	1	4	4	0	0	4
	1	1	1	0	0	0
	1	1	1	0	0	1
	1	1	1	0	0	0
	7	14	11	3	2	2
	1	5	5	0	0	3
	3	21	11	10	5	4
	1	10	8	2	2	3
	1	7	7	0	0	2
	4	10	7	3	2	4
	1	5	5	0	0	3
	6	10	4	6	4	3
	1	19	13	6	3	3
	1	9	9	0	0	6
	6	12	5	7	3	3
	1	3	3	0	0	3
	5	5	5	0	0	4
	3	6	4	2	1	3
	1	6	6	0	0	2
	3	8	6	2	1	3
	1	5	5	0	0	3
	4	10	8	2	2	2

Averages	2.48	7.61	5.74	1.87	1.09	2.78
Std Dev	1.97	5.28	3.21	2.82	1.50	1.31
Std Err	0.41	1.10	0.67	0.59	0.31	0.27

Results for the Keyword Input Field for the Query Task

Participant count	terms	query stems	distinct query stems	query stem intersections	distinct query stem intersections	distinct query / top 10 TFIDF intersections
	2	4	4	0	0	1
19	2	2	2	0	0	2
	1	2	2	0	0	2
	2	8	8	0	0	1
	2	2	2	0	0	1
	1	1	1	0	0	1
	1	1	1	0	0	1
	2	2	2	0	0	2
	1	2	2	0	0	0

Averages	2.05	4.11	3.68	0.42	0.37	1.79
Std Dev	0.85	2.45	2.16	0.69	0.60	1.08
Std Err	0.19	0.56	0.50	0.16	0.14	0.25

Results for the Keyword Input Field for the Query Task

Participant count	terms	query stems	distinct query stems	query stem intersections	distinct query stem intersections	distinct query / top 10 TFIDF intersections
	2	3	2	1	1	2
	1	5	5	0	0	2
	3	4	4	0	0	1
	3	7	5	2	1	2
	2	3	3	0	0	3
	2	9	8	1	1	4
	2	4	3	1	1	1
	3	6	4	2	2	4
	4	7	6	1	1	1
	3	6	6	0	0	3
Averages	2.05	4.11	3.68	0.42	0.37	1.79
Std Dev	0.85	2.45	2.16	0.69	0.60	1.08
Std Err	0.19	0.56	0.50	0.16	0.14	0.25

Results for the Query Task using Query Input Field

Participant count	terms	query stems	distinct query stems	query stem intersections	distinct query stem intersections	distinct query / top 10 TFIDF intersections
	1	6	6	0	0	1
24	1	9	7	2	2	2
	1	5	5	0	0	1
	3	5	4	1	1	2
	1	4	4	0	0	2
	1	3	3	0	0	2
	1	3	3	0	0	2
	1	3	3	0	0	2
	1	2	2	0	0	1
	1	1	1	0	0	1
	1	6	6	0	0	4
	1	4	4	0	0	4
	1	5	5	0	0	3
	1	3	3	0	0	2
	2	3	2	1	1	2
	1	2	2	0	0	2
	4	7	4	3	2	3
	1	5	5	0	0	3
	1	6	6	0	0	3
	1	4	4	0	0	2
	1	3	3	0	0	2
	1	3	3	0	0	2
Averages	1.25	4.13	3.83	0.29	0.25	2.13
Std Dev	0.74	1.80	1.49	0.75	0.61	0.85
Std Err	0.15	0.37	0.30	0.15	0.12	0.17

Results for the Query Task using Query Input Field						
Participant count	terms	query stems	distinct query stems	query stem intersections	distinct query stem intersections	distinct query / top 10 TFIDF intersections
	1	4	4	0	0	2
	1	3	3	0	0	1
Averages	1.25	4.13	3.83	0.29	0.25	2.13
Std Dev	0.74	1.80	1.49	0.75	0.61	0.85
Std Err	0.15	0.37	0.30	0.15	0.12	0.17

Bibliography

Altonen, A., Hyrskykari, A. & Raiha, K.-J. (1998), 101 spots, or how do users read menus?, *in* 'CHI 98 Human Factors in Computing Systems', ACM Press, pp. 132–139.

Anderson, J. R. & Lebire, C. (1998), *The atomic components of thought*, Erlbaum, Mahwah, NJ.

Arabie, P. & Boorman, S. S. (1973), 'Multidimensional scaling of measures of distance between partitions', *Mathematical Psychology* **10**, 148–203.

Atkinson, J., Campbell, F. W. & Francis, M. R. (1976), 'The magic number 4 0: A new look at visual numerosity judgements', *Perception* **5**(3), 327–334.

Atkinson, J., Francis, M. R. & Campbell, F. W. (1976), 'The dependence of the visual numerosity limit on orientation, colour, and grouping in the stimulus', *Perception* **5**(3), 335–342.

Atkinson, R. & Shiffrin, R. (1968), Human memory: A proposed system and its control processes, *in* K. W. Spence & J. T. Spence, eds, 'The Psychology of learning and motivation: Advances in research and theory', Vol. 2, Academic Press, New York, p. 249.

Averbach, E. & Coriell, A. S. (1961), 'Short-term memory in visicon', *Bell System Technical Journal* **40**, 309–328.

Awh, E. & Pashler, H. E. (2000), 'Evidence for split attentional foci', *Journal of Experimental Psychology Human Perception and Performance* **26**(2), 834–846.

Baddeley, A. D. (1986), *Working Memory*, Vol. 11 of *Oxford Psychology Series*, Oxford: Clarendon Press.

Baddeley, A. D. (1996), 'Exploring the central executive', *Quarterly Journal of Experimental Psychology* **49A**, 1–28.

Baddeley, A. D. (2000), 'The episodic buffer: a new component of working memory?', *Trends in Cognitive Science* **4**(11), 417–423.

Baddeley, A. D. & Hitch, G. (1974), Working memory, *in* G. Bower, ed., 'The psychology of learning and motivation: Advances in research and theory', Vol. 8, Academic Press, New York, pp. 47–89.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press Series/Addison Wesley, New York.

Bailey, T. M. & Hahn, U. (2001), 'Determinants of wordlikeness: lexical or phonotactic?', *Memory and Language* **44**, 568–591.

Baroni-Urbani, C. & Buser, M. W. (1976), 'Similarity of binary data', *Systematic Zoology* **25**(3), 251–259.

Barrouillet, P., Bernardin, S. & Camos, V. (2004), 'Time constraints and resource sharing in adults' working memory spans', *Experimental Psychology: General* **133**(1), 83–100.

Barrouillet, P. & Lepine, R. (2005), 'Working memory and children's use of retrieval to solve addition problems', *Journal of Experimental Child Psychology* **91**(3), 183–204.

Bartlett, F. C. (1932), *Remembering: A study in experimental and social study psychology*, Cambridge University Press, Cambridge, England.

Bauer, B., Jolicoeur, P. & Cowan, W. B. (1996), 'Visual search for colour targets that are or are not linearly-separable from distractors', *Vision Research* **36**(10), 1439–1465.

Bauer, B., Jolicoeur, P. & Cowan, W. B. (1998), 'The linearly separability effect in color visual search: Ruling out the additive color hypothesis', *Perception & Psychophysics* **60**(6), 1083–1093.

Beck, J., Prazdny, K. & Rosenfeld, A. (1983), A theory of textural segmentation, *in* J. Beck, K. Prazdny & A. Rosenfeld, eds, 'Human and Machine Vision', Academic Press, New York, pp. 1–39.

Belkin, N., Oddy, R. & Brooks, H. (1982), 'Ask for information retrieval: Part i. background and theory', *Documentation* **38**(2), 61–71.

Beutter, B. R., Eckstein, M. P. & Stone, L. S. (2003), 'Saccadic and perceptual performance in visual search tasks. i. contrast detection and discrimination', *Journal of the Optical Society of America* **20**(7), 1341–1355.

Biederman, I. (1972), 'Perceiving real-world scenes', *Science* **177**, 77–80.

Biederman, I. (1982), On the semantics of a glance at a scene, *in* M. Kubovy & J. R. Pomerantz, eds, 'Perceptual organization', Erlbaum, Hillsdale, NJ, pp. 213–254.

Blecic, D., Bangalore, N. S., Dorsch, J. L., Henderson, C. L., Koenig, M. H. & Weller, A. C. (1998), 'Using transaction log analysis to improve opac retrieval results', *College and Research Libraries* **59**(1), 39?50.

Boles, W. W. & Pillay, H. (1999), Cognitive styles, subject content and the design of computer based instruction, *in* 'Proceedings of the 5th International Symposium on Signal Processing and its Applications, ISSPA 99', Vol. 2, Brisbane, pp. 559–562.

Boltzmann, L. (1899), 'On the development of the methods of theoretical physics in recent times', pp. 77–100.

Bowman, H. & Faconti, G. (1999), 'Analysing cognitive behaviour using lotos and mexitl', *Formal Aspects of Computing* **11**(2), 132–159.

Bransford, J. D. & Franks, J. J. (1971), 'The abstraction of linguistic ideas', *Cognitive Psychology* **Online Media**, 331–350.

Brasnford, J. (1979), *Human cognition: Learning, understanding, and remembering*, Wadsworth Publishing Company, Belmont, CA.

Braun-Blanquet, J. & ;, N. Y. (1932), *Plant sociology: the study of plant communities*, McGraw-Hill Book Company, Inc., New York.

Bright, P., Moss, H. E., Stamatakis, E. A. & K., T. L. (2005), 'The anatomy of object processing: The role of anteromedial temporal cortex', *Experimental Psychology* **58B**(3/4), 361–377.

Brin, S. & Page, L. (1998), 'Dynamic data mining: Exploring large rule spaces by sampling'.

Brinck, T., Gergle, D. & Wood, S. D. (2001), *Usability for the Web: Designing Web sites that work*, Morgan Kaufman Publishers, San Francisco.

Broadbent, D. E. (1958), *Perception and Communication*, Pergamon, London.

Broadbent, D. E. (1982), 'Task combination and selective intake of information', *Acta Psychologica* **50**(3), 253–290.

Broadbent, D. E. & Broadbent, M. H. P. (1987), 'From detection to identification: Response to multiple targets in rapid serial visual presentation', *Perception & Psychophysics* **42**, 105–113.

Broadfield, A. (1946), *The philosophy of classification*, Grafton & Co, London.

Brown, B. & Sellen, A. (2001), Exploring users' experiences of the web, Technical Report HPL-2001-262, Publishing Systems and Solutions Laboratory, Hewlett-Packard Company.

Brown, J., Collins, A. & Duguid, P. (1989), 'Situated cognition and the culture of learning', *Educational Researcher* **18**, 32–42.

Brown, J. L. (1965), Flicker and intermittent stimulation, *in* C. H. Graham, ed., 'Vision and Visual Perception', John Wiley and Sons, Inc., New York, New York, pp. 251–320.

Bruning, R. H., Schraw, G. & Ronning, R. R. (1995), *Cognitive Psychology and Instruction*, NJ:Merrill, Englewood.

Burton, R. R. & Brown, J. S. (1982), An investigation of computer coaching for informal learning activities, *in* D. H. Sleeman & J. S. Brown, eds, 'Intelligent Tutoring Systems', Academic Press, London - New York, pp. 79–98.

Buzikashvili, N. (2000), 'The yandex study: First findings', *Internet-mathematics* pp. 95–120.

Byrne, M. D., Anderson, J. R., Douglas, S. & Matessa, M. (1999), Eye tracking the visual search of clickdown menus, *in* 'CHI 99', ACM Press, New York, pp. 402–409.

Cantor, J., Engle, R. & Hamilton, G. (1991), 'Short-term memory, working memory, and verbal abilities: How do they relate?', *Intelligence* **15**(2), 229–246.

Card, S. K. (1984), Visual search of computer command menus, *in* H. Bouma & D. G. Bouwhuis, eds, 'Attention and Performance X: Control of Language Processes', Lawrence Erlbaum Associates, London, pp. 97–108.

Card, S. K., Moran, T. P. & Newell, A. (1980), 'The keystroke-level model for user performance time with interactive systems', *Communications of the ACM* **23**(7), 396–410.

Card, S. K., Moran, T. P. & Newell, A. (1983), *The psychology of human-computer interaction*, Erlbaum, Hillsdale, NJ.

Carlson, R. A., Sullivan, M. A. & Schneider, W. (1989), 'Practice and working memory effects in building procedural skill', *Experimental Psychology: Learning, Memory, and Cognition* **15**, 517–526. span size on the algebraic task greater when substitutions were required.

Carter, E. C. & Carter, R. C. (1981), 'Color and conspicuousness', *Journal of Optical Society of American* **71**(6), 723–729.

Carter, R. C. (1982), 'Visual search with color', *Journal of Experimental Psychology: Human Perception and Performance* **8**, 127–136.

Catledge, L. D. & Pitkow, J. E. (1995), 'Characterizing browsing strategies in the world-wide web', *Computer Networks and ISDN Systems* **27**(6), 1065–1073.

Chambers, J., Cleveland, W., B., K. & Tukey, P. (1983), *Graphical Methods for Data Analysis*, Duxbury Press, Boston, Massachusetts.

Chan, L. M. (1994), *Cataloging and classification: An introduction*, 2nd edn, McGraw-Hill, New York.

Chandler, P. & Sweller, J. (1991), 'Cognitive load theory and the format of instruction', *Cognition and Instruction* **8**(4), 293–332.

Chase, W. G. & Simon, H. A. (1973*a*), The mind's eye in chess, *in* W. G. Chase, ed., 'Visual information processing', Academic Press, New York, pp. 215–281.

Chase, W. G. & Simon, H. A. (1973*b*), 'Perception in chess', *Cognitive Psychology* **4**, 55–81.

Chaudhary, A., Szalay, A. S. & Moore, A. W. (2002), Very fast outlier detection in large multidimensional data sets, *in* V. Ghanti, ed., 'ACM SIGMOD Workshop on Reasearch Issues on Data Mining and Knowledge Discovery (DMKD)', ACM Press, Madison, Wisconsin, pp. 45–52.

Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. (1993), 'A neural basis for visual search in inferior temporal cortex', *Nature* **363**, 345–347.

Chey, J., Grossberg, S. & Mingolla, E. (1997), 'Neural dynamics of motion grouping: from aperture ambiguity to object speed and direction', *Optical Society of America* **14**(10), 2570–2594.

Chi, M. T. H., Glaser, R. & Farr, M. J. (1988), *The Nature of Expertise*, Erlbaum, Hillsdale, NJ.

Chiu, A. & Fu, A. W.-c. (2003), Enhancements on local outlier detection, *in* '7th International Database Engineering and Application Symposium (IDEAS2003)', ACM PRess, Hong Kong, S.A.R., China, pp. 298–307.

Cho, J. & Roy, S. (2004), Impact of search engines on page popularity, *in* '3th International Conference on World Wide Web. WWW'04', ACM Press, New York, NY., pp. 20–29.

Choo, C., Detlor, B. & Turnbull, D. (2000), *Web work: Information seeking and knowledge work on the World Wide Web*, Kluwer Academic Publishers, Dordrecht.

Choo, C. W., Detlor, B. & Turnbull, D. (1998), A behavioral model of information seeking on the web – preliminary results of a study of how managers and it specialists use the web, *in* 'ASIS Annual Meeting', Vol. 35, Toronto, Canada, pp. 290–302.

Chun, M. M. & Jiang, Y. H. (1998), 'Contextual cueing: Implicit learning and memory of visual context guides spatial attention', *Cognitive Psychology* **36**, 28–71.

Chun, M. M. & Potter, M. C. (1995), 'A two-stage model for multiple target detection in rapid serial visual presentation', *Journal of Experimental Psychology: Human Perception and Performance* **21**(1), 109–127.

Chun, Marvin, M. & Jiang, Yuhong, H. (1999), 'Top-down attentional guidance based on implicit learning of visual covariation', *Psychological Science* **10**, 360–365.

Chun, Marvin, M. & Nakayama, K. (2000), 'On the functional role of implicit visual memory for the adaptive deployment of attention across views', *Visual Cognition* **7**(1-3), 65–81.

Clark, R., Nguyen, F. & Sweller, J. (2006), *Efficiency in Learning: Evidence-Based Guidelines to Manage Cognitive Load*, Pfeiffer, San Francisco.

Cohen, J. (1990), 'Things i have learned (so far)', *American Psychology* **45**(12), 1304–1312.

Cohen, J. (1994), 'The earth is round (p¡.05)', *American Psychology* **49**(12), 997–1003.

Conway, A. R. A. & Engle, R. W. (1994), 'Working memory and retrieval: A resource-dependent inhibition model', *Journal of Experimental Psychology: General* **123**(4), 3543–73.

Conway, A. R. & Engle, R. W. (1996), 'Individual differences in working memory capacity: more evidence for a general capacity theory', **4**(6), 577–90.

Coombs, C. H., Dawes, R. M. & Tversky, A. (1970), *Mathematical Psychology: An Elementary Introduction*, Prentice-Hall, Englewood Cliffs, NJ.

Cooper, G. (1998), 'An introduction to applications of cognitive load theory to instructional design. a presentation at the education 1998 conference'.

Cowan, N. (1988), 'Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system', *Psychological Bulletin* **104**(2), 163–191.

Cowan, N. (1999), An embedded-process model of working memory, *in* . P. S. A. Miyake, ed., 'Models of working memory: Mechanisms of active maintenance and executive control', Cambridge University Press, Cambridge, UK, pp. 62–101.

Cowan, N. (2001), 'The magical number 4 in short-term memory: A reconsideration of mental storage capacity', *Behavioral and Brain Sciences* **24**(1), 87–185.

Craik, F. I. M. & Lockhart, R. S. (1972), 'Levels of processing: A framework for memory research', *Journal of Verbal Learning and Verbal Behavior* **11**, 671–684.

Craik, F. I. M. & Tulving, E. (1975), 'Depth of processing and the retention of words in episodic memory', *Experimental Psychology: General* **104**(3), 268–294.

Craik, K. (1943), *The Nature of Explanation*, Cambridge University Press, Cambridge.

Croft, B. W., Cook, R. & Wilder, D. (1995), Providing government information on the internet: Experiences with thomas, *in* 'Second Annual Conference on the Theory and Practice of Digital Libraries (DL '95)', Austin, Texas, pp. 19–24.

Crowder, R. G. (1993), Auditory memory, *in* S. McAdams & E. Bigand, eds, 'Thinking in sound: The cognitive psychology of human audition', Oxford University Press, Oxford, England, pp. 113–145.

Czerwinski, M., Horvitz, E. & Wilhite, S. (2004), A diary study of task switching and interruptions, *in* 'SIGCHI conference on Human factors in computing systems', ACM, Vienna, Austria, pp. 175–182.

Daneman, M. & Carpenter, P. (1980), 'Individual differences in working memory and reading', *Journal of Verbal Learning and Verbal Behaviour* **19**(4), 450–466.

Davidson, R. J., Schwartz, G. E., Pugash, E. & Bromfield, E. (1976), 'Sex differences in patterns of eeg asymetry', *BiologicaI Psychology* **4**, 119–138.

De Beni, R., Palladino, P., Pazzaglia, F. & Cornoldi, C. (1998), 'Increases in intrusion errors and working memory deficit of poor comprehenders', *The Quarterly Journal of Experimental Psychology* **51**(2), 305–320.

De Fockert, J. W., Rees, G., Frith, C. D. & Lavie, N. (2001), 'The role of working memory in visual selective attention', *Science* **2912**, 1803–1806.

de Groot, A. D. (1978), *Thought and choice in chess*, 2nd edn, Mouton, The Hague.

de Groot, A. D. & Gobet, F. (1996), *Perception and memory in chess: in the heuristics of the professional eye*, Van Gorcum, Assen (The Netherlands). De Groot and Gobet propose that perception and memory are more important differentiators of expertise than is the ability to think ahead in the search for chess moves.

Dennis, R. L. H., Williams, W. R. & Shreeve, T. G. (1998), 'Faunal structures among european butterflies: evolutionary implications of bias for geography, endemism and taxonomic affiliation', *Ecography* **21**, 181–203.

Desimone, R. & Duncan, J. (1995), 'Neural mechanisms of selective visual attention', *Annual Review of Neuroscience* **18**, 193–222.

Deutsch, J. A. & Deutsch, D. (1963), 'Attention: some theoretical considerations', *Psychological Review* **70**, 80–90.

Dice, L. E. (1945), 'Measures of the amount of ecologic association between species', *Ecology* **26**(3), 297–302.

Downing, C. J. & Pinker, S. (1985), The spatial structure of visual attention, *in* M. I. Posner & O. S. M. Marin, eds, 'Attention and performance XI', Erlbaum, Hillsdale, NJ, pp. 171–188.

Doyle, J. C., Ornstein, R. & Galin, D. (1975), 'Lateral specialization of cognitive mode: Ii eeg frequency analysis', *Psychophysiology* **11**(5), 567–578.

Driver, J. (2001), 'A selective review of selective attention research from the past century', *British Journal of Psychology* **21**, 451–468.

Driver, J. & Baylis, G. C. (1989), 'Movement and visual attention: The spotlight metaphor breaks down', *Journal of Experimental Psychology: Human Perception and Performance* **15**(3), 448–456.

Driver, J., McLeod, P. & Dienes, Z. (1992), 'Motion coherence and conjunction search: Implications for guided search theory', *Perception & Psychophysics* **51**(1), 79–85.

Duke, D. & Duce, D. (1999), 'The formalization of a cognitive architecture and its application to reasoning about human computer interaction', *Formal Aspects of Computing* **11**(6), 665–689.

Duncan, J. (1980), 'The locus of interference in the perception of simultaneous stimuli', *Psychological Review* **87**(3), 272–300.

Duncan, J. (1984), 'Selective attention and the organization of visual information', *Journal of Experimental Psychology: General* **113**(4), 501–517.

Duncan, J. (1985), Visual search and visual attention, *in* M. I. Posner & O. S. M. Marin, eds, 'Attention and performance', Vol. 6, Erlbaum, Hillsdale, NJ, pp. 85–105.

Duncan, J. & Humphreys, G. (1992), 'Beyond the search surface - visual search and attentional engagement', *Experimental Psychology: Human Perception and Performance* **18**, 578–588.

Duncan, J., Ward, R. & Shapiro, K. L. (1994), 'Direct measurement of attentional dwell time in human vision', *Nature* **369**, 313–315.

D'Zmura, M. (1991), 'Color in visual search', *Vision Research* **31**(6), 951–966.

Eckstein, M. P., Thomas, J. P., Palmer, J. & Shimozaki, S. S. (2000), 'A signal detection model predicts the effects of set-size in visual search accuracy for feature, conjunction and disjunction displays', *Perception and Psychophysics* **62**(3), 425–451.

Egly, R., Driver, J. & Rafal, R. D. (1994), 'Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects', *Journal of Experimental Psychology: Human Perception and Performance* **123**(2), 161–177.

Engle, R., Cantor, J. & Carullo, J. J. (1992), 'Individual differences in working memory and comprehension: A test of four hypotheses', *Journal of Experimental Psychology: Learning, Memory and Cognition* **18**(5), 972–992.

Engle, R., Conway, A., Tuholski, S. & Shisler, R. (1995), 'A resource account of inhibition', *Journal of Psychological Science* **6**, 122–125.

Engle, R. W. (2001), What is working memory capacity?, *in* H. L. R. III, J. S. Nairne, I. Neath & M. Surprenant, eds, 'The nature of remembering: Essays in honor of Robert G. Crowder', American Psychology Association, Washington, DC, pp. 297–314.

Engle, R. W. (2002), 'Working memory capacity as executive attention', *Directions in Psychological Science* **11**(1), 19–23.

Engle, R. W., Kane, M. J., Tuholski, S. W. & Press. (1999), Individual differences in working memory capacity, what they tell us about controlled attention, general

fluid intelligence, and functions of the prefrontal cortex, *in* A. Miyake & P. Shah, eds, 'Models of working memory: Mechanisms of active maintenance and executive control', Cambridge University, Cambridge, UK, pp. 102–134.

Enns, J. T. (1986), 'Seeing textons in context', *Perception & Psychophysics* **39**, 143–147.

Enns, J. T. (1990), Three-dimensional features that pop out in visual search, *in* D. Brogan, ed., 'Visual Search', Taylor & Francis, New York, New York, pp. 37–45.

Enns, J. T. & Rensink, R. A. (1990), 'Sensitivity to three-dimensional orientation in visual search', *Psychological Science* **1**(5), 323–326.

Enns, J. T. & Rensink, R. A. (1991), 'Preattentive recovery of three-dimensional orientation from line drawings', *Psychological Review* **98**, 101–118.

Eriksen, B. A. & Eriksen, C. W. (1974), 'Effects of noise letters upon identification of a target letter in a nonsearch task', *Perception and Psychophysics* **16**(1), 143–149.

Eriksen, C. & Spencer, T. (1969), 'Rate of information processing in visual perception: Some results and methodological considerations', *Journal of Experimental Psychology Monograph* **79**(2/2), 1–16.

Eriksen, C. W. & St. James, J. D. (1986), 'Visual attention within and around the field of focal attention: a zoom lens model', *Percept Psychophysicsim* **40**(4), 225–40.

Eriksen, C. W. & Yeh, Y. (1985), 'Allocation of attention in the visual field', *Experimental Psychology: Human Perception and Performance* **11**(5), 583–597.

Fager, E. W. & McGowan, J. A. (1963), 'Zooplankton species groups in the north pacific:co-occurrences of species can be used to derive groups whose members react similarly to water-mass types.', *Science* **140**, 453–460. DOI: 10.1126/science.140.3566.453.

Faith, D. P. (1983), 'Asymmetric binary similarity measures', *Oecologia* **57**(3), 287–290.

Feyen, R., Liu, Y., Chaffin, D., Jimmerson, G. & Joseph, B. (1999), 'New software tools improve workplace design', *Ergonomics in Design: The Quarterly of Human Factors Applications,* **7**(2), 24–30.

Filkov, V. & Skiena, S. (2004), 'Heterogeneous data integration with the consensus clustering formalism', *Data Integration in the Life Sciences (DILS), International workshop No1.* **2994**, 110–123.

Fischer, G. (1999), User modeling: The long and winding road, *in* J. Kay, ed., 'UM99: User Modelling Conference', Springer Verlag, Wien New York, Banff, Canada, pp. 349–355.

Fitts, P. M. (1954), 'The information capacity of the human motor system in controlling the amplitude of movement', *Experimental Psychology* **47**, 381–391.

Fodor, J. A., Garrett, M., Walker, E. & Parkes, C. (1980), 'Against definitions', *Cognition and Instruction* **8**, 263–367.

Forbes, S. (1925), 'Method of determining and measuring the associative relations of species', *Science* **61**(1585), 518–524.

Ford, N., Miller, D. & Moss, N. (2002), 'Web search strategies and retrieval effectiveness: an empirical study', *Documentation* **58**(1), 30–48.

Ford, N., Miller, D. & Moss, N. (2003), 'Web search strategies and approaches to studying', *American Society for Information Science and Technology* **54**(6), 473–489.

Fossum, T. V. & Haller, S. M. (2004), 'Measuring card sort orthogonality', *Expert Systems* **22**(3), 139146.

Fowlkes, E., B. & Mallows, C., L. (1983), 'A method for comparing two hierarchical clusterings', *American Statistical Association* **78**(383), 553–569.

Fred, A. & Jain, A. (2003), 'Robust data clustering', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2003**.

Freed, M., John, B. E., Matessa, M., Remington, R. & Vera, A. (2002), 'Automating cpm-goms', *CHI Letters , 4 (1), .* **4**(1), 147 – 154.

Galin, D. & Ornstein, R. (1972), 'Lateral specialization of cognitive mode: an eeg study', *Psychophysiology* **9**, 412–418.

Gebb, J. W., Mowbray, G. H. & Byham, C. L. (1955), 'Difference lumens for photic intermittence', *Quarterly Journal of Experimental Psychology* **7**, 49–55.

Geisler, W. S. & Chou, K.-L. (1995), 'Separation of low-level and high-level factors in complex tasks: visual search', *Psychology Review* **102**(2), 356–78.

Gentner, D. (1989), The mechanisms of analogical learning, *in* S. Vosniadou & A. Ortony, eds, 'Similarity and analogical reasoning', Cambridge University Press, Cambridge, pp. 199–241.

Gernsbacher, M. A. (1993), 'Less skilled readers have less efficient suppression mechanisms', *Psychological Science* **4**(5), 294–298.

Gernsbacher, M. A. . P. o. L. C. a. S. B. l.-c. (1990), 'Precis of: "language comprehension as structure building."', *Psycoloquy* **3**(69).

Gibson, J. J. (1966), *The Senses considered as Perceptual Systems*, Houghton Mifflin,, Boston.

Gick, M. I. & Holyoak, K. J. (1980), 'Analogical problem-solving', *Cognitive Psychology* **12**, 306–355.

Gilbert, N. & Wells, T. C. E. (1966), 'Analysis of quadrat data', *Ecology* **54**(3), 675–685.

Ginsburg, H. P. & Opper, S. (1988), *Piaget's Theory ofIntellectual Development*, 3rd edn, Prentice Hall, N.J.

Glaser, B. G. & Strauss, A. L. (1967), *The discovery of grounded theory: Strategies for qualitative research*, Aldine Publishing Co., Chicago.

Gobet, F. (1993), *Les memoires d'un joueur d'echecs*, Editions Universitaires, Fribourg, Switzerland.

Gobet, F. (2001), 'Chunks hierarchies and retrieval structures: Comments on saariluoma and laine', *Scandinavian Journal of Psychology* **42**, 149–155.

Gobet, F. & Simon, H. A. (1996), 'Templates in chess memory: A mechanism for recalling several boards', *Cognitive Psychology* **31**, 1–40.

Goldberg, H. J. & Kotval, X. P. (1999), 'Computer interface evaluation using eye movements: Methods and constructs', *Industrial Ergonomics* **24**(6), 631–645.

Goodall, D. W. (1967), 'The distribution of the matching coefficient', *Biometics* **23**(4), 647–656.

Grimes, J. (1996), On the failure to detect changes in scenes across saccades, *in* K. Akins, ed., 'Perception & Psychophysics (Vancouver Studies in Cognitive Science)', Vol. 2, Oxford University Press, New York, pp. 89–110.

Grindley, G. C. & Townsend, V. (1968), 'Voluntary attention in peripheral vision and its effects on acuity and differential thresholds', *Quarterly Journal of Experimental Psychology* **20**(1), 11–19.

Guilford, J. P. (1959), 'Three faces of intellect', *American Psychologist* **14**, 469–79.

Hahn, U., Chater, N. & Richardson, L. (2003), 'Similarity as transformation', *Cognition* **87**, 1–32.

Hahn, U. & Nakisa, R. C. (2000), 'German inflection: single- or dual-route?', *Cognitive Psychology* **41**, 313–336.

Hahn, U. & Ramscar, M. C. A. (2001), *Similarity and categorization*, Oxford University Press, Oxford.

Halford, G. S., Wilson, W. H. & Phillips, W. (1998), 'Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology', *Behavioral Brain Sciences* **21**(6), 803–831.

Halkidi, M., Batistikis, Y. & Vazirgiannis, M. (2001), 'On clustering validation techniques', *Intelligent Information Systems* **17**, 107–145.

Hamann, U. (1961), 'Merkmalbestand und verwandtschaftsbeziehungen de farinosae: Ein beitrag zum system der monokotyledonen', *Wildenowia* **2**, 639–768.

Hamm, V. P. & Hasher, L. (1992), 'Age and the availability of inferences', *Psychology and Aging* **7**(1), 587–594.

Hampton, J. A. (1995), 'Testing the prototype theory of concepts', *Memory and Language* **34**, 686–708.

Hartmann, M. & Hasher, L. (1991), 'Aging and suppression: Memory for previously relevant information', *Psychology and Aging* **6**(4), 587–594.

Hasher, L., Stoltzfus, E. R., Zacks, R. T. & Rypma, B. (1991), 'Age and inhibition', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**(1), 163–169.

Hasher, L. & Zacks, R. (1988), 'Working memory, comprehension, and aging: A review and a new view', *The Psychology of Learning and Motivation* **22**, 193–225.

Hayek, L. C. (1994), Analysis of amphibian biodiversity data, *in* W. R. Heyer, M. A. Donnelly, R. W. McDiarmid, L.-A. C. Hayek & M. S. Foster, eds, 'Measuring and Monitoring Biological Diversity: Standard Methods for Amphibians', Smithsonian Institution Press.

Hayes, J. R. M. (1952), Memory span for several vocabularies as a function of vocabulary size, Technical report, Acoustics Laboratory, Massachusetts Institute of Technology.

Head, M., Archer, N. & Yuan, Y. (2000), 'World wide web navigation aid', *International Journal of Human-Computer Studies* **53**(2), 301–330.

Healey, C. G. (2001), Formalizing artistic techniques and scientific visualization for painted renditions of complex information spaces, *in* 'International Joint Conference on Artificial Intelligence (IJCAI 2001)', Seattle, Washington, pp. 371–376.

Healey, C. G. (2004), 'Perception in visualization'.

Healey, C. G., Booth, K. S. & Enns, J. T. (1993), Harnessing preattentive processes for multivariate data visualization, *in* 'Graphics Interface '93', Toronto, Canada, pp. 107–117.

Healey, C. G., Booth, K. S. & Enns, J. T. (1996), 'High-speed visual estimation using preattentive processing', *ACM Transactions on Human Computer Interaction* **3**(2), 107–135.

Healey, C. G. & Enns, J. T. (1998), Building perceptual textures to visualize multidimensional datasets, *in* D. Ebert, H. Hagen & H. Rushmeier, eds, 'Visualization '98', IEEE, Research Triangle Park, North Carolina, pp. 111–118.

Healey, C. G. & Enns, J. T. (1999), 'Large datasets at a glance: Combining textures and colors in scientific visualization', *IEEE Transactions on Visualization and Computer Graphics* **5**(2).

Healey, C. G. & Enns, J. T. (2002), 'Perception and painting: A search for effective, engaging visualizations', *IEEE Computer Graphics & Applications (CG&A), Special Issue on Information Visualization* **22**(2), 10–15.

Healey, C. G., Enns, J. T., Tateosian, L. G. & Remple, M. (2004), 'Perceptually-based brush strokes for nonphotorealistic visualization', *ACM Transactions on Graphics* **23**(1), 64–96.

Heit, E. (1997), Features of similarity and category-based induction, *in* E. Cambouropolos & H. Pain, eds, 'Interdisciplinary Workshop on Categorization and Similarity', University of Edinburgh, pp. 115–121.

Heit, E. & Barsalou, L. W. (1996), 'The instantiation principle in natural categories', *Memory* **4**, 413–451.

Henderson, J. M., Pollatsek, A. & Rayner, K. (1989), 'Covert visual attention and extrafoveal information use during object identification', *Perception & Psychophysics* **45**(3), 196–208.

Hintzmann, D. L. (1986), 'Schema abstraction in a multiple-trace memory model', *Psychological Review* **93**, 411–428.

Hohnsbein, J. & Mateeff, S. (1998), 'The time it takes to detect changes in the speed and direction of visual motion', *Vision Research* **38**(17), 2569–2573.

Hollan, J. D. (1990), User models and user interfaces: A case for domain models, task models, and tailorability, *in* 'AAAI-90, Eighth National Conference on Artificial Intelligence', AAAI Press/The MIT Press, Cambridge, MA, p. 1137.

Holliday, J. D., Hu, C.-Y. & Willett, P. (2002), 'Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings', *Combinatorial Chemistry and High Throughput Screening* **5**(2), 155–166.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandenavian Journal of Statistics* **6**, 65–70.

Holscher, C. & Strube, G. (2000), 'Web search behavior of internet experts and newbies', *Computer and Telecommunications Networking* **33**(1-6), 337–346.

Holyoak, K. J. & Koh, K. (1987), 'Surface and structural similarity in analogical transfer', *Cognition and Instruction* **15**, 332–340.

Horibe, Y. (1985), 'Entropy and correlation', *IEEE Transactions on Systems, Man and Cybernetics (SMC)* **SMC-15**(5), 641–642.

Howell, D. C. (1997), *Statistical Methods for Psychology*, 4th edn, Wadsworth Publishing Company.

Huang, L. & Pashler, H. (2005), 'Attention capacity and task difficulty in visual search', *Elsevier: Cognition* **94**(3), B101–B111.

Huber, D. E. & Healey, C. G. (2005), 'Visualizing data with motion', *Visualization, 2005. VIS 05. IEEE* pp. 527– 534.

Hudson, J. M., Christensen, J., Kellogg, W. A. & Erickson, T. (2002), "i'd be overwhelmed, but it's just one more thing to do": availability and interruption in research management, *in* 'CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, Minneapolis, Minnesota, USA, pp. 97–104.

Hudson, L. (1966), 'Contrary imaginations: A psychological study of the english schoolboy', *Comparative Education* **3**(1), 64–64. In this study Dr Liam Hudson argues that personality counts for as much as ability in the student's choice of subject.

Huitt, W. (2003), 'The information processing approach to cognition', *Seeboerger-Weichselbaum, Michael* .

Hyde, T. S. & Jenkins, J. J. (1969), 'The differential effects of incidental tasks on the organization of recall of a list of highly associated words', *Experimental Psychology: General* **82**, 472–481.

Imbo, I. & Vandierendonck, A. (2007), 'Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: differences between children and adults?', *Psychological Research* .

Inhoff, A. W. (1984), 'Two stages of word processing during eye fixations in the reading of prose', *Verbal Learning and Verbal Behavior* **23**(5), 612–624.

Jaccard, P. (1901), 'Distribution de la florine alpine dans la bassin de dranses. et dans quelques regiones voisines', *Naturelles Bulletin de la Societe Vaudoise des Sciences* pp. 241–272.

Jacob, E. (1991), Classification and categorization: Drawing the line, *in* B. H. Kwasnik & R. Fidel, eds, 'Advances in classification research, 2nd ASIS SIG/CR classification research workshop', Vol. 2, Learned Information, Inc., Medford, NJ, pp. 67–83.

Jacoby, L. L. & Bartz, W. H. (1972), 'Encoding processes and the negative recency effect', *Verbal Learning and Verbal Behavior* **11**, 561–565.

Jansen, B. J. (2006), 'Search log analysis: What it is, what's been done, how to do it', *Library & Information Science Research* **28**, 407–432.

Jansen, B. J. & Pooch, U. (2001), 'A review of web searching studies and a framework for future research'.

Jansen, B. J. & Spink, A. (2003), An analysis of web information seeking and use: Documents retrieved versus documents viewed, *in* '4th International Conference on Internet Computing', Las Vegas, Nevada, pp. 65–69.

Jansen, B. J., Spink, A. & Pederson, J. (2005), 'Trend analysis of altavistaweb searching', *American Society for Information Science and Technology* **56**(6), 559?570.

Jansen, B. J., Spink, A. & Saracevic, T. (2000), 'Real life, real users and real needs: A study and analysis of users queries on the web', *Information Processing and Management* **36**(2), 207–227.

Jennionsa, M. D. & Mllerb, A. P. (2003), 'A survey of the statistical power of research in behaviour ecology and animal behaviour', *Behavioral Ecology* **14**(3), 438–445.

Jiang, Yuhong, H. & Chun, Marvin, M. (2001), 'Selective attention modulates implicit learning', *Quarterly Journal of Experimental Psychology* **54A**(4), 1105–1124.

John, B. E. (1988), Contributions to Engineering Models of human-computer interaction, PhD thesis, Carnegie Mellon University.

John, B. E. (1990), Extensions of goms analyses to expert performance requiring perception of dynamic visual and auditory information, *in* 'CHI'90', ACM, Seattle, Washington, pp. 107–115.

John, B. E. & Kieras, D. E. (1994), The goms family of analysis techniques: Tools for design and evaluation, Technical Report CMU-CS-94-181, Carnegie Mellon University School of Computer Science.

John, B. E. & Kieras, D. E. (1996), 'The goms family of user interface analysis techniques: Comparison and contrast', *ACM Transactions on Computer-Human Interaction* **3**(4), 320–351.

Johnson-Laird, P. N. (1983), *Mental models: towards a cognitive science of language, inference and consciousness*, Cambridge University Press, Cambridge, UK.

Johnson-Laird, P. N. (1989), Mental models, *in* M. I. Posner, ed., 'Foundations of cognitive science', MIT Press, Cambridge, MA.

Johnson, S. C. (1967), 'Hierarchical clustering schemes', *Psychometrika* **2**(32), 241–254.

Johnston, W. A. & Dark, V. J. (1986), 'Selective attention', *Annual Review of Psychology* **37**, 43–75.

Jones, S., Cunningham, S. & McNab, R. (1998), 'Usage analysis of a digital library'.

Julesz, B. (1981), 'A theory of preattentive texture discrimination based on firstorder statistics of textons', *Biological Cybernetics* **41**(2), 131–138.

Julsz, B. (1971), *Foundations of Cyclopean Perception*, University of Chicago Press, Chicago, Illinois.

Julsz, B. & Bergen, J. R. (1987), Textons, the fundamental elements in preattentive vision and perception of textures, *in* M. A. Fischler & O. Firschein, eds, 'Readings in computer vision: issues, problems, principles, and paradigms', Vol. 62 of *Morgan Kaufmann Readings Series*, Morgan Kaufmann Publishers Inc., San Francisco, CA, U.S.A., pp. 243–256.

Just, M. A. & Carpenter, P. A. (1976), 'Eye fixations and cognitive processes', *Cognitive Psychology* **8**, 441–480.

Just, M. A. & Carpenter, P. A. (1992), 'A capacity theory of comprehension: Individual differences in working memory', *Psychological Review* **99**(1), 122–149.

Kahneman, D. (1973), *Attention and Effort*, Prentice-Hall, Englewood Cliffs, New Jersey.

Kahneman, D. & Henik, A. (1981), Perceptual organization and attention, *in* M. Kubovy & J. R. Pomerantz, eds, 'Perceptual organization', Erlbaum, Hillsdale, N.J.

Kane, M. J. & Engle, R. W. (2003), 'Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to stroop interference.', *Journal of Experimental Psychology: General* **132**, 47–70.

Kaufman, E. L., Lord, M. W., Reese, T. W. & Volkmann, J. (1949), 'The discrimination of visual number', *American Journal of Psychology* **62**, 498–525.

Kawai, M., Uchikawa, K. & Ujike, H. (1995), Influence of color category on visual search, *in* 'Annual Meeting of the Association for Research in Vision and Ophthalmology', Vol. 2991, Fort Lauderdale, Florida, U.S.A.

Keane, M. T., O'Brien, M. & Smyth, B. (2007), 'Power-law regularities in web search behavior: Are people biased in their use of search-engines?', *Communications of the ACM* .

Keil, F. C. (1989), *Concepts, kinds, and cognitive development*, MIT Press, Cambridge, MA.

Kieras, D. (1993), 'Using the keystroke-level model to estimate execution times'.

Kieras, D. E. & Meyer, D. (1994), The epic architecture for modeling human information-processing: A brief introduction, Technical Report 1 & 2, University of Michigan, Department of Electrical Engineering and Computer Science.

Kieras, D. & Meyer, D. E. (1995), An overview of the epic architecture for cognition and performance with application to human-computer interaction, Technical Report 5, University of Michigan, Electrical Engineering and Computer Science Department.

Kintsch, W. (1998), *Comprehension: A paradigm for cognition*, Cambridge University Press, New York.

Kirton, M. J. (1976), 'Adaptors and innovators: A description and measure', *Journal of Applied Psychology* **61**(5), October.

Kirton, M. J. (2003), *Adaption-Innovation: In the Context of Diversity and Change*, Routledge, London.

Klockner, L., Wirschum, N. & Jameson, A. (2004), 'Depth- and breadth-first processing of search result lists'.

Knobbe, A. J. & Adrianns, P. W. (1996), Analysis of binary association, *in* 'Knowledge Discovery and Data Mining (KDD-96)', Portland, Oregon, pp. 311–314.

Knorr, E. M. & Ng, A. R. T. (1998), Algorithms for mining distance-based outliers in large datasets, *in* A. Gupta, O. Shmueli & J. Widom, eds, 'Proceedings of the 24rd International Conference on Very Large Data Bases', Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 392–403.

Kosara, R., Miksch, S. & Hauser, H. (2002), 'Focus+context taken literally', *IEEE Computer Graphics & Applications (CG&A), Special Issue on Information Visualization* **22**(1), 22–29,.

Krug, S. (2000), *Don't Make Me Think: A Common Sense Approach to Web Usability*, 2 edn, New Riders & Peachpit.

Kulczynski, S. (1927), 'Zespoly roslin w pieninach - die pflanzenassoziationen der pieninen', *Bulletin international de l'acadmie polonaise des sciences et des lettres* **B**(2), 57–203.

Kurth, M. (1993), 'The limits and limitations of transaction log analysis', *Library Hi Tech* **11**(2), 98?104.

Kvalseth, T. O. (1987), 'Entropy and correlation: Some comments', *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-17**, 517–519.

La Pointe, L. & Engle, R. (1990), 'Simple and complex word spans as measures of working memory capacity', *Journal of Experimental Psychology: Learning, Memory and Cognition* **16**, 1118–1133.

LaBerge, D. (1995), *Attentional processing: The brains art of mindfulness*, Harvard University Press, Cambridge, Massachusetts.

LaBerge, D., Carlson, R. L., Williams, J. K. & Bunney, B. G. (1997), 'Shifting attention in visual space: Tests of moving-spotlight models versus an activity-distribution model', *Experimental Psychology: Human Perception and Performance* **23**, 1380–1392.

Laird, J. E., Newell, A. & Rosenbloom, P. S. (1987), 'Soar: An architecture for general intelligence', *Artificial Intelligence* **33**(1), 1–64.

Lavie, N. (1995), 'Perceptual load as a necessary condition for selective attention', *Journal of Experimental Psychology: Human Perceptual Performance* **21**, 451–468.

Lavie, N. (2000), Selective attention and cognitive control: dissociating attentional functions through different types of load, *in* S. M. Driver & J., eds, 'Attention and performance', Vol. XVIII, MIT press, Cambridge, Massachusetts.

Lavie, N. (2005), 'Distracted and confused?: Selective attention under load', *Elsevier: Trends in Cognitive Science* **9**(2), 75–82.

Lavie, N. & Cox, S. (1997), 'On efficiency of visual selective attention: Efficient visual search leads to inefficient distracted rejection', *Psychological Science* **8**(5), 395–398.

Lavie, N. & Defockert, J. W. (2003), 'Contrasting effects of sensory limits and the capacity limits in visual selective attention', *Perception and Psychophysics* **65**(2), 202–212.

Lavie, N. & Defockert, J. W. (2005), 'The role of working memory in attentional capture', *Psychonomic Bulletin and Review* **12**(4), 669–674.

Lavie, N. & Fox, E. (2000), 'The role of perceptual load in negative priming', *Journal of Experimental Psychology: Human Perception and Performance* **26**(3), 1038–1052.

Lavie, N., Hirst, A., De Fockert, J. W. & Viding, E. (2004), 'Load theory of selective attention and cognitive control', *Journal of Experimental Psychology: General* **133**, 339–354.

Lavie, N. & Tsal, Y. (1994), 'Perceptual load as a major determinant of locus of selection in visual attention', *Perception and Psychophysics* **56**(Online Media), 183–197.

Lee, E. & MacGregor, J. (1985), 'Minimizing user search time in menu retrieval systems', *Human Factors,* **27**(2), 157–162.

Lee, T. T. (1987), 'An information theoretic analysis of relational databases - part 1: data dependencies and information metric', *IEEE Transactions on Software Engengineering* **SE-13**(10), 1049–1061.

Levin, D. T. & Simons, D. J. (1997), 'Failure to detect changes to attended objects in motion pictures', *Psychonomic Bulletin and Review* **4**, 501–506.

Lindsay, P. & Norman, D. A. (1972), *Human information processing; an introduction to psychology*, New York: Academic Pres, New York.

Linfoot, E. H. (1957), 'An informational measure of correlation', *Information and Control* **1**, 85–87.

Lisman, J. E. & Idiart, M. A. P. (1995), 'Storage of 7 +/- 2 short-term memories in oscillatory subcycles', *Science* **267**(5203), 1512–1515.

Liu, G., Healey, C. G. & Enns, J. T. (2003), 'Target detection and localization in visual search: A dual systems perspective', *Perception & Psychophysics* **65**(5), 678–694.

Logan, G. D. (2003), 'Cumulative progress in formal theories of attention', *Annual Review of Psychology* **55**, 207–234.

Lopez de Mantaras, R. (1989), Id3 revisited: A distance-based criterion for attribute selection, *in* 'International Syposium on Methodologies for Intelligent Systems (ISMIS-89)', Charlotte, North California, U.S.A.

Luce, P. A. (1986), Neighborhoods of words in the mental lexicon, PhD thesis, Indiana University.

Mack, A. & Rock, I. (1998), *Inattentional blindness*, MIT Press, Cambridge, MA.

Magrab, E. B. (1997), *Integrated Product and Process Design and Development: The Product Realization Process*, Taylor & Francis CRC Press LLC, Boca Raton, FL, USA.

Maltby, A. (1975), *Sayers manual of classification for librarians*, 5th edn, Andr Deutsch, London.

Malvestuto, F. M. (1986), 'Statistical treatment of the information content of a database', *Information Systems* **11**(3), 211–223.

Mandler, G. & Shebo, B. (1982), 'Subitizing: An analysis of its component processes', *Experimental Psychology: General* **111**(1), 1–22.

Mandler, J. M. (1984), *Stories, Scripts and Scenes: Aspects of Schema Theory*, Lawrence Erlbaum Ass. Inc., Hillsdale, New Jersey, London.

Manning, C. D. & Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Margolis, E. & Laurence, S. (1999), *Concepts: Core readings*, MIT Press, Cambridge, MA.

Margolis, E. & Laurence, S. (2002), Concepts, *in* S. P. Stich & T. A. Warfield, eds, 'Blackwell Guide to Philosophy of Mind', Wiley, p. 417.

Mark, G., Gonzlez, V. M. & Harris, J. (2005), No task left behind?: examining the nature of fragmented work, *in* 'CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems, SESSION: Take a number, stand in line (interruptions & attention 1)', ACM Press, Portland, Oregon, USA, pp. 321–330.

Markou, M. & Singh, A. S. (2003), 'Novelty detection: a review - part 1: statistical approaches', *Signal Process* **82**(12), 2481–2497.

Marschark, M., Richman, C. L., Yuille, J. C. & Hunt, R. R. (1987), 'The role of imagery in memory: On shared and distinctive information', *Psychological Bulletin* **102**(1), 28–41.

Marsh, R. L., Sebrechts, M. M., Hicks, J. L. & Landau, J. D. (1997), 'Processing strategies and secondary memory in very rapid forgetting', *Memory & Cognition* **25**(2), 173–181.

May, C. P., Hasher, L., Zacks, R. T. & Multhaup, K. S. . (1999), 'Inhibition in the processing of garden path sentences', *Psychology and Aging* **14**(1), 304–313.

McClelland, J. L. (1994), 'The organization of memory a parallel distributed processing perspective', *Rev. Neural. (Paris)* **150**(8-Sep), 570–579.

McConnaughey, B. H. (1964), 'The determination and analysis of plankton communities', *Marine Research Indonesia Special (Penelitian Laut Di Indonesia)* **Spec. no.**, 30.

McSorley, E. & Findlay, J. M. (2003), 'The eyes can search large displays more effectively than small ones: an oculomotor paradox?[abstract]', *Vision* **3**(9).

Medin, D., Altom, M. & Murphy, T. (1984), 'Given versus induced category representations: Use of prototype and exemplar information in classification', *Experimental Psychology: Learning, Memory, and Cognition* **10**, 333–352.

Medin, D. L., Ross, B. H. & Markman, A. B. (2004), *Cognitive Psychology*, 4rd edn, John Wiley & Sons, Inc., San Francisco, CA.

Meila, M. (2003), 'Comparing clusterings by variation of information', *Proceedings of the Sixteenth Annual Conference of Computational Learning Theory (COLT)* .

Meiran, N. (1996), 'Reconfiguration of processing mode prior to task performance. journal of experimental psychology', *Learning, Memory and Cognition* **22**(6), 1423–1442.

Mervis, C. B. & Rosch, E. (1981), 'Categorization of natural objects', *ANual Review of Psychology* **32**, 89–115.

Meyer, D. E. & Kieras, D. E. (1997), 'A computational theory of executive cognitive processes and multiple-task performance: Part 1. basic mechanisms', *Psychological Review* **104**, 3–65.

Michael, E. L. (1920), 'Marine ecology and the coefficient of association: A plea in behalf of quantitative biology', *The Journal of Ecology* **8**(1), 54–59.

Miller, C. S. & Remington, R. W. (2005), 'Modeling information navigation: implications for information architecture', *Human-Computer Interaction* **19**, 225–271.

Miller, G. A. (1956), 'The magic number seven, plus or minus two: Some limits on our capacity for processing information', *Psychological Review* **63**, 81–97.

Mirkin, B. (1996), *Mathematical Classification and Clustering*, Kluwer Academic Press:, Boston-Dordrecht.

Mirkin, B. (2001), 'Eleven ways to look at the chi-squared coefficient for contingency tables', *The American Statistician* **55**(6), 111–120.

Mizzaro, S. (1997), 'Relevance: The whole history', *American Society of Information Science* **48**(9), 810–832.

Moukdad, H. & Large, A. (2001), 'Users' perceptions of the web as revealed by transaction log analysis', *Online Information Review* **25**(6), 349–358.

Mountford, M. D. (1962), An index of similarity and its application to classificatory problems, *in* P. W. Murphy, ed., 'Progress in Soil Zoology', Butterworth, London, England, pp. 43–50.

Mowbray, G. H. & Gebhard, J. W. (1955), 'Differential sensitivity of the eye to intermittent white light.', *Science* **121**(3136), 137–175.

Mowbray, G. H. & Gebhard, J. W. (1960), 'Differential sensitivity of peripheral retina to intermittent white light', *Science* **132**(3428), 672 – 674.

Murphy, G. L. (2002), *The big book of concepts*, MIT Press, Cambridge, MA.

Murphy, G. L. & Medin, D. L. (1999), The role of theories in conceptual coherence, *in* E. Margolis & S. Laurence, eds, 'Concepts: Core readings', MIT Press, Cambridge, MA, pp. 425–458.

Muter, P. (1980), 'Very rapid forgetting', *Memory & Cognition* **8**(2), 174–179.

Myers, I. B. (1987), *Introduction to Type: A Description of the Theory and Applications of the Myers-Briggs Type Indicator*, 4th edn, Consulting Psychologists Press.

Nagy, A. L. & Sanchez, R. R. (1990), 'Critical color differences determined with a visual search task', *Journal of the Optical Society of America A: Optics, Image Science, and Vision* **7**(7), 1209–1217.

Nagy, A. L., Sanchez, R. R. & Hughes, T. C. (1990), 'Visual search for color differences with foveal and peripheral vision', *Journal of the Optical Society of America A: Optics, Image Science, and Vision* **7**(10), 1995–2001.

Nakagawa, S. (2004), 'A farewell to bonferroni: the problems of low statistical power and publication bias', *Behavioral Ecology* **15**(6), 1044–1045.

Nakayama, K. & Silverman, G. H. (1986), 'Serial and parallel processing of visual feature conjunctions', *Nature* **320**, 264–265.

Neisser, U. (1967), *Cognitive psychology*, Prentice-Hall, Englewood Cliffs, New Jersey.

Neisser, U. & Becklen, R. (1975), 'Selective looking: attending to visually specified events', *Cognitive Psychology* **7**(4), 480–494.

Newell, A. (1990), *Unified theories of cognition*, Harvard University Press, Cambridge, MA.

Newell, A. & Simon, H. A. (1972), *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

Nielsen, J. (2000), *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing, Indianapolis, IN.

Nissen, M. J. (1985), Accessing features and objects: Is location special?, *in* M. I. Posner & D. S. Marin, eds, 'Attention and Performance XI', Erlbaum, Hillsdale, NJ, pp. 205–219.

Nissen, M. J. & Bullemer, P. (1987), 'Attentional requirements of learning: Evidence from performance measures', *Cognitive Psychology* **19**(1), 1–32.

NIST (2001), 'Document understanding conference (duc)'.

Norman, D. A. (1983), Some observations on mental models, *in* D. Gentner & A. L. Stevens, eds, 'Mental Models', Lawrence Erlbaum Associates.

Norman, D. & Bobrow, D. (1975), 'On data-limited and resource limited processes', *Cognitive Psychology* **7**(Book), 44–64.

Norman, K. L. (1991), *The Psychology of Menu Selection: Designing Cognitive Control of the Human/Computer Interface*, International Journal of Man-Machine Studies, Ablex Publishing Corporation, Norwood, N. J.

Nosofsky, R. (1986), 'Attention, similarity and the identification-categorization relationship', *Experimental Psychology: General* **115**, 39–57.

Nosofsky, R. (1992), Exemplars, prototypes, and similarity rules, *in* A. F. Healy, S. M. Kosslyn & R. M. Shiffrin, eds, 'From learning theory to connectionist theory: Essays in honor of William K. Estes', Vol. 1, Lawrence Erlbaum, Hillsdale, NJ, pp. 149–167.

O'Brien, M. & Keane, M. T. (2007), 'Modeling user behaviour using a search-engine', *Interactive User Interfaces IUI'07* pp. 357–361.

O'Conaill, B. & Frohlich, D. (1995), Timespace in the workplace: dealing with interruptions, *in* 'CHI '95: Conference companion on Human factors in computing systems', ACM Press, pp. 262–263.

OHare, D., Wiggins, M., Williams, A. & Wong, W. (1998), 'Cognitive task analyses for decision centered design and training', *Ergonomics in Design: The Quarterly of Human Factors Applications* **41**, 1698–1718.

Olson, I. R. & Chun, M. M. (2002), 'Perceptual constraints on implicit learning of spatial context', *Visual Cognition* **9**(3), 273–302.

Olson, J. R. & Olson, G. M. (1990), 'The growth of cognitive modeling in human-computer interaction since goms', *Human-Computer Interaction* **5**(2&3), 221–265.

OneStat.com (February 2, 2004), 'Most people use 2 word phrases in search engines according to onestat.com'.

Ostry, D., Moray, N. & Marks, G. (1976), 'Attention, practice, and semantic targets', *Journal of Experimental Psychology: Human Perception and Performance* **2**(3), 326–36.

Ozmutlu, S., Spink, A. & Ozmutlu, H. C. (2003), 'Multimedia web searching trends: 1997-2001', *Information Processing & Management* **39**(4), 611–621.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B. & Faloutsos, C. (2003), Loci: Fast outlier detection using the local. correlation integral, *in* K. R. U. Dayal & T. Vijayaraman, eds, '19th Internatational Conference on Data Engineering (ICDE)', Bangalore, India, pp. 315–326.

Pashler, H. E. (1988), 'Familiarity and the detection of change in visual displays', *Perception & Psychophysics* **44**, 369–378.

Pashler, H. E. (1998), *The psychology of attention*, MIT Press, Cambridge, Massachusetts.

Pawlak, Z., Wong, S. K. & Ziarko, W. I. J. M.-M. (1988), 'Rough sets: probabilistic versus deterministic approach', *International Journal of Man-Machine Studies* **29**(1), 81–95.

Penniman, W. & Dominick, W. (1980), 'Monitoring and evaluation of on-line information system usage', *Information Processing and Management* **116**, 17–35.

Perfetti, C. A. & Goldman, S. R. (1976), 'Discourse memory and reading comprehension skill', *Journal of Verbal Learning and Verbal Behavior* **15**(1), 33–42.

Perlow, L. A. (1999), 'The time famine: Toward a sociology of work time', *Administrative Science Quarterly* **44**(1), 57–81.

Perneger, Thomas, V. (1998), 'What's wrong with bonferroni adjustments', *British Medical Journal* **316**, 1236–1238.

Peterson, L. R. & Peterson, M. (1959), 'Short-term retention of individual verbal items', *Journal of Experimental Psychology* **58**, 193–198.

Peterson, S. A. & Simon, T. J. (2000), 'Computational evidence for the subitizing phenomenon as an emergent property of the human cognitive architecture', *Cognitive Science* **24**(1), 93–122.

Pfitzner, D., Hobbs, V. & Powers, D. (2003), 'A unified taxonomic framework for information visualization'.

Pfitzner, D. & Powers, D. M. W. (2004), Vedges: a grid based visual clustering approach to dsiplaying document return sets, Tech. rep., Flinders University (S.A.).

Phippen, A., Sheppard, L. & Furnell, S. (2004), 'A practical evaluation of web analytics', *Internet Research: Electronic Networking Applications and Policy* **14**, 284?293.

Pillay, H. & Wilss, L. (1996), 'Computer assisted instruction and individual cognitive style preferences in learning: does it matter?', *Australian Educational Computing* **11**(2), 28–33.

Pinker, S. (1998), *How the Mind Works*, Allen Lane, New York, NY.

Pirolli, P. (2000), 'A web site user model should at least predict something about users', *Internetworking* **3**(1).

Pirolli, P. & Card, S. K. (1999), 'Information foraging', *Psychological Review* **106**, 643–675.

Pirolli, P. & Fu, W.-T. (2003), Snif-act: a model of information foraging on the world wide web, *in* 'Proceedings of the 9th International Conference on User Modeling'.

Pirolli, P., Fu, W.-T., Reeder, R. & K., C. S. (2002), A user-tracing architecture for modeling interaction with the world wide web, *in* 'Advanced Visual Interfaces', ACM Press, Trento, Italy.

Posner, M. I. (1980), 'Orienting of attention', *Experimental Psychology* **32**(1), 3–25.

Posner, M. I. & Boies, S. J. (1971), 'Components of attention', *Psychological Review* **78**, 391–408.

Posner, M. I., Inhoff, A. W., Friedrich, F. J. & Cohen, A. (1987), 'Isolating attentional mechanisms: A cognitive-anatomical analysis', *Psychobiology* **15**, 107–112.

Posner, M. I., Nissen, M. J. & Ogden, W. C. (1978), Attended and unattended processing modes: the role of set for spatial locations, *in* H. L. Pick & B. J. Saltzman, eds, 'Modes of Perceiving and Processing Information', Erlbaum, Hillsdale, N.J., pp. 137–158.

Powers, D. M. W. (2003), Recall and precision versus the bookmakers, *in* 'Joint International conference on cognitive science', University of New South Whales, pp. 529–534.

Powers, D. M. W. (2007), Expected information in the transmission of an equality selection of distribution/clustering or of individual class labels, Tech. rep., Flinders University (S.A.).

Powers, D. M. W., Leibbrandt, R., Pfitzner, D., Luerssen, M., Lewis, T., Abrahamyan, A. & Stevens, K. (2008), 'Language teaching in a mixed reality games environment'.

Prabu, D., Mei, S., Hayes, A. & Fredin, E. S. (2007), 'A cyclic model of information seeking in hyperlinked environments: The role of goals, self-efficacy, and intrinsic motivation', *International Journal of Human-Computer Studies* **65**(2), 170–182.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Prinz, J. J. (2002), *Furnishing the mind: Concepts and their perceptual basis*, MIT Press, Cambridge, MA.

Quine, W. (1951), 'Two dogmas of empiricism', *Philosophical Review* **60**(1), 20–43.

Quinlan, J. R. (1990), Induction of decision trees, *in* J. W. Shavlik & T. G. Dietterich, eds, 'Readings in Machine Learning', Morgan Kaufmann. Originally published in Machine Learning 1:81–106, 1986.

Rajski, C. (1961), 'A metric space of discrete probability distributions', *Information and Control* **4**(4), 371–377.

Rand, W. M. (1971), 'Objective criteria for evaluation ofclustering methods', *Journal of the American Statistical Association* **66**(336), 846–850.

Rasmussen, J. (1983), 'Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance models', *IEEE Transactions on Systems, Man, and Cybernetics (SMC)* **13**(3), 257–266.

Raymond, J. E., Shapiro, K. L. & Arnell, K. M. (1992), 'Temporary suppression of visual processing in an rsvp task: An attentional blink?', *Journal of Experimental Psychology: Human Perception and Performance* **18**(3), 849–860.

Reddy, L. & VanRullen, R. (2007), 'Spacing affects some but not all visual searches: Implications for theories of attention and crowding', *Journal of Vision* **7**(2), 1–17.

Rees, G., Frith, C. D. & Lavie, N. (1999), 'Inattentional blindness versus inattentional amnesia for fixated but ignored words', *Science* **286**(5449), 2504–2507.

Rehder, B. (2003*a*), 'Categorization as causal reasoning', *Cognitive Science* **27**, 709–748.

Rehder, B. (2003*b*), 'A causal-model theory of conceptual representation and categorization', *Experimental Psychology: Learning, Memory, and Cognition* **29**(6), 1141–1159.

Rensink, Ronald, A., O'Regan, K. J. & Clark, J. J. (1997), 'To see or not to see: The need for attention to perceive changes in scenes', *Psychological Science* **8**(5), 368–373.

Rice, W. R. (1989), 'Analyizing tables of statistical tests', *Evolution, Vol.* **43**(1), 223–225.

Richardson, E. C. (1964), *Classification*, 3 edn, Shoe String Press, Hamden, CT.

Riding, R. & Cheema, I. (1991), 'Cognitive styles: An overview and integration', *Educational Psychology Review* **11**(3 & 4), 193–215.

Rips, L. J. (1975), 'Inductive judgements about natural categories', *Verbal Learning and Verbal Behavior* **14**, 665–685.

Rips, L. J. (1989), Similarity, typicality, and categorization, *in* S. V. . A. Ortony, ed., 'Similarity and analogical reasoning', Cambridge University Press, Cambridge, pp. 21–59.

Ritter, F. E., Baxter, G. D., Jones, G. & Young, R. M. (2000), 'Supporting cognitive models as users', *ACM Transactions on Computer-Human Interaction* **7**(2), 141–173.

Rogers, D. J. & Tanimoto, T. T. (1960), 'A computer program for classifying plants', *Science* **132**(3434), 1115–1118.

Rosch, E. (1973), On the internal structure of perceptual and semantic categories, *in* T. E. Moore, ed., 'Cognitive Development and the Acquisition of Language', Academic Press, New York, p. 279.

Rosch, E. (1999), 'Reclaiming concepts', *Consciousness Studies* **6**, 61–77.

Rosch, E., Mervis, C. B., Gray, W. . D., Johnson, D. M. & Bayes-Braem, P. (1976), 'Basic objects in natural categories', *Cognitive Psychology* **8**, 382–439.

Ross, B. (1984), 'Remindings and their effects in learning a cognitive skill', *Cognitive Psychology* **16**, 371–416.

Ross, J. (2003), 'Visual discrimination of number without counting', *Perception* **32**(7), 867 – 870.

Ross, N. C. M. & Wolfram, D. (2000), 'End user searching on the internet: An analysis of term pair topics submitted to the excite search engine', *American Society for Information Science and Technology* **51**(10), 949?958.

Rothman, K. J. (1990), 'No adjustments are needed for multiple comparisons', *Epidemiology* **1**(1), 43–6.

Rouncefield, M., Hughes, J. A., Rodden, T. & Viller, S. (1994), Working with constant interruption: Cscw and the small office, *in* 'ACM Conference on Computer Supported Cooperative Work (CSCW'94)', ACM, Chapel Hill, NC, pp. 275–286.

Rumelhart, D. (1980), Schemata: The building blocks of cognition, *in* R. Spiro, B.Bruce & W. Brewer, eds, 'Theoretical Issues in Reading and Comprehension', Erlbaum, Hillsdale, New Jersey, London.

Rumelhart, D. E. & McClelland, J. L. (1986), On learning the past tenses of english verbs, *in* 'Rumelhart, McClelland, and the PDP Research Group', Vol. 2, MIT Press, pp. 216–271.

Russell, P. F. & Rao, T. R. (1940), 'On habitat and association of species of anopheline larvae in southeastern, madras', *Malaria Institute of India* **3**, 153–178.

Sagi, D. & Julsz, B. (1985*a*), 'Detection versus discrimination of visual orientation', *Perception Perception* **13**(5), 619–628.

Sagi, D. & Julsz, B. (1985*b*), 'The "where" and "what" of vision', *Science* **228**(4704), 1217–1219.

Saracevic, T. (1995), 'Evaluation of evaluation in information retrieval'.

Saracevic, T. (1996), Modeling interaction in information retrieval: A review and proposal, *in* 'Annual Academy Meeting of American Society for Information Science', Vol. 33, pp. 3–9.

Saracevic, T. (1997), 'Users lost: reflections on the past, future, and limits of information science', *SIGIR Forum* **31**(2), 16–27.

Savage, R. M. (1934), 'The breeding behavior of the common frog, rana remporaria linn., and of the common toad bufo bufo bufo linn.', *Zoological Society of London* **55-70**.

Schneider, W. & Shiffrin, R. M. (1977), 'Controlled and automatic human information processing: 1. detection, search, and attention', *Psychological Review* **84**(1), 1–66.

Schwartz, G. E., Davidson, R. J. & Maer, F. (1975), 'Right hemisphere lateralization for emotion in the human brain: interactions with cognition', *Science* **190**, 286–288.

Scott, S. D., Lesh, N. & Klau, W. G. (2002), Investigating human-computer optimization, *in* 'ACM Conference on Human Factors in Computing Systems (CHI)', pp. 155–162.

Sebrechts, M. M., Marsh, R. L. & Seamon, J. G. (1989), 'Secondary memory and very rapid forgetting', *Memory & Cognition* **17**, 693–700.

Shannon, C. E. (1949), Communication in the presence of noise, *in* 'Institute of Radio Engineers', Vol. 37, pp. 10–21.

Shapiro, K. L., Raymond, J. E. & Arnell, K. M. (1994), 'Attention to visual pattern information produces the attentional blink in rapid serial visual presentation', *Journal of Experimental Psychology: Human Perception and Performance* **20**(2), 357–371.

Shisler, R. J., Conway, A. R. A., Tuholski, S. W. & Engle, R. W. (1995), 'Effects of visual and verbal workload on inhibition.', *Paper presented at the annual meeting of the Psychonomic Society* .

Shneiderman, B. (1992), 'Acm transactions on graphics', pp. 92–99.

Shneiderman, B. & Maes, P. (1997), 'Direct manipulation vs. interface agents', *ACM Interaction* **4**(6), 42–61.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioural Sciences*, McGraw-Hill Publishing Co., New York.

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999), 'Analysis of a very large web search engine query log', *SIGIR Forum* **33**(1), 6–12.

Simon, T. J. & Vaishnavi, S. (1996), 'Subitizing and counting depend on different attentional mechanisms: Evidence from visual enumeration in afterimages', *Perception & Psychophysics* **58**(6), 915–926.

Simons, D. J. & Levin, D. T. (1998), 'Failure to detect changes to people during a real-world interaction', *Psychonomic Bulletin and Review* **5**, 644–649.

Slaughter, V., Jaakkola, R. & Carey, S. (1999), Constructing a coherent theory: Childrens biological understanding of life and death, *in* M. Siegal & C. Peterson, eds, 'Childrens understanding of biology and health', Cambridge University Press, Cambridge, UK.

Smith, E. E. & L., M. D. (1999), The exemplar view, *in* E. Margolis & S. Laurence, eds, 'Concepts: Core readings', MIT Press, Cambridge, MA, pp. 207–221.

Smith, E. R. & Zarate, M. A. (1992), 'Exemplar-based model of social judgment', *Psychological Review* **99**, 3–21.

Sneath, P. H. A. (1968), 'Vigour and pattern in taxonomy', *General Microbiology* **54**(1), 1–11.

Sneath, P. H. A. & Sokal, R. R. (1973), *Numerical Taxonomy*, Freeman and Company, San Francisco.

Sohlberg, McKay, M. & Mateer, C. A. (1989), *Introduction to cognitive rehabilitation: theory and practice*, Guilford Press, New York.

Sokal, R. R. & Sneath, P. H. A. (1964), 'Principles of numerical taxonomy', *Systematic Zoology* **13**, 106–108.

Sorgenfrei, T. (1958), 'Molluscan assemblages from the marine middle miocene of south jutland and their environments'.

Sperling, G. (1963), 'A model for visual memory tasks', *Human Factors* **5**, 19–31.

Spink, A. & Jansen, B. J. (2004*a*), 'A study of web search trends', *Webology* **1**(2).

Spink, A. & Jansen, B. J. (2004*b*), *Web Search: Public Searching of the Web*, Kluwer Academic Publishing.

Spink, A. & Ozmutlu, H. C. (2002), 'Characteristics of question format web queries: an exploratory study', *Information Processing and Management* **38**(4), 453–471.

Spink, A., Ozmutlu, S., Ozmutlu, H. C. & Jansen, J. (2002), 'Us versus european web searching trends', *ACM SIGIR Forum* **36**(2).

Spink, A., Wolfram, D., Jansen, B. J. & Saracevic, T. (2001), 'Searching the web: the public and their queries', *American Society for Information Science and Technology* **52**(3), 226–234.

Spink, A. & Xu, J. L. (2000), 'Selected results from a large study of web searching: the excite study', *Information Research* **6**(1).

Spinks, J. A., Zhang, J. X., Fox, P. T., Gao, J. & Tan, L. H. (2004), 'More workload on the central executive of working memory, less attention capture by novel visual distractors: evidence from an fmri study', *NeuroImage* **23**, 517–524.

Spiteri, L. F. (2004), 'Word association testing and thesaurus construction'.

Spiteri, L. F. (2005), 'The use of word association in the construction of information retrieval thesauri: A pilot study', *Cataloging & Classification Quarterly* **40**(1), 55–78.

Spiteri, L. F. (2007), 'The role of causality and conceptual coherence in assessments of similarity', *Library and Information Science Research Electronic* **17**(2).

Spool, J. M., Scalon, T., Shroeder, W., Snyder, C. & DeAngelo, T. (1999), *Web site usability: a design guide*, Morgan Kaufman, San Francisco, CA.

Squire, L. R. & Zola-Morgan, S. (1991), 'The medial temporal lobe memory system', *Science* **253**(5026), 1380–1386.

Stevens, S. S. (1957), 'On the psychophysical law', *Psychological Review* **64**(3), 153–181.

Stoltzfus, E. R., Hasher, L. & Zacks, R. T. (1996), Working memory and aging: Current status of the inhibitory view, *in* J. T. E. Richardson, R. W. Engle & R. H. L. Hasher, eds, 'Working memory and human cognition', Oxford University Press, Oxford, UK, pp. 66–68.

Strehl, A. & Ghosh, J. (2002), 'Cluster ensembles - a knowledge reuse framework for combining partitionings', *Machine Learning Research* **3**, 583–617.

Su, L. T. (2003), 'A comprehensive and systematic model of user evaluation of web search engines: Ii. an evaluation by undergraduates', *American Society for Information Science and Technology* **54**(13), 1193–1223.

Suchman, L. A. (1987), *Plans and situated actions: the problem of human-machine communication*, Cambridge University Press, Cambridge.

Sun, R. (2002), *Duality of the mind: A Bottom-up Approach toward Cognition*, Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Sun, R. (2004), 'Desiderata for cognitive architectures', *Philosophical Psychology* **17**(3).

Sun, R. (2006), The clarion cognitive architecture: Extending cognitive modeling to social simulation, *in* R. Sun, ed., 'Cognition and Multi-Agent Interaction', Cambridge University Press, New York.

Sun, R., Merrill, E. & Perterson, T. (1998), A bottom-up model of skill learning, *in* '20th Cognitive Science Society Conference', Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1037–1042.

Sun, R., Merrill, E. & Perterson, T. (2001), 'From implicit skill to explicit knowledge: a bottom-up model of skill learning', *Cognitive Science* **25**(2), 203–244.

Sutcliffe, R. F. E. & Slater, B. E. A. (1995), 'Disambiguation by association as a practical method: Experiments and findings', *Journal of Quantitative Linguistics* **2**(1).

Sweller, J. (1988), 'Cognitive load during problem solving: Effects on learning', *Cognitive Science* **12**(2), 257–285.

Sweller, J. (1989), 'Cognitive technology: Some procedures for facilitating learning and problem solving in mathematics and science', *Educational Psychology Review* **81**(4), 457–466.

Sweller, J. (1999), *Instructional design in technical areas.*, Australian Education Review, No. 43., ACER Press, Camberwell, Victoria, Australia.

Sweller, J., Van Merrienboer, J. J. G. & Paas, F. G. W. C. (1998), 'Cognitive architecture and instructional design', *Educational Psychology Review* **10**(3), 251–296.

Tarwid, K. (1960), 'Szacowanie zbieznosci nisz ekologicznych gatunkow droga oceny prawdopodobienstwa spotykania sie ich w polowach.', *Ecologia Polska* **B**(6), 115–130.

Theeuwes, J. (1994), 'The effects of location cuing on redundant-target processing', *Psychological Research* **57**(1), 15–19.

Theodoridis, S. & Koutroubas, K. (1999), *Pattern recognition*, Academic Pres.

Thomas, D. C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M. & Armstrong, B. G. (1985), 'The problem of multiple inference in studies designed to generate hypotheses', *American Journal of Epidemiology* **122**(6), 1080–95.

Thompson, B. (2002), 'What future quantitative social science research could look like: Confidence intervals for effect sizes', *Educational Researcher* **31**(3), 25–32.

Thurstone, L. (1927), 'A law of comparative judgement', *Psychological Review* **34**, 278–286.

Treharne, K., Pfitzner, D., Leibbrandt, R. & Powers, D. M. W. (2008), 'A lean online approach to human factors research'.

Treharne, K., Pfitzner, D. & Powers, D. (2006), 'Information coding in animation (extended abstract)'.

Treharne, K., Pfitzner, D. & Powers, D. M. W. (2007), 'The versatile role of motion in visualization'.

Treisman, A. (1985), 'Preattentive processing in vision', *Computer Vision, Graphics, and Image Processing* **31**(2), 156–177.

Treisman, A. (1986), 'Features and objects in visual processing', *Scientific American* **255**(5), 114–125.

Treisman, A. (1988), 'Features and objects: The fourteenth bartlett memorial lecture', *Quarterly Journal of Experimental Psychology* **40A**(2), 201–237.

Treisman, A. & Geffen, G. (1967), 'Selective attention: Perception or response?', *Quarterly Journal of Experimental Psychology* **19**(4), 1–18.

Treisman, A. & Gelade, G. (1980), 'A feature integration theory of attention', *Cognitive Psychology* **12**, 97–136.

Treisman, A. & Gormican, S. (1988), 'Feature analysis in early vision: Evidence from search asymmetries', *Psychological Review* **95**(1), 15–48.

Treisman, A. & Sato, S. (1990), 'Conjunction search revisited', *Experimental Psychology: Human Perception and Performance* **8**, 459–478.

Treisman, A. & Souther, J. (1985), 'Search asymmetry: A diagnostic for preattentive processing of separable features', *Experimental Psychology: General* **114**(3), 285–310.

Trick, L. & Pylyshyn, Z. (1994), 'Why are small and large numbers enumerated differently? a limited capacity preattentive stage in vision', *Psychology Review* **101**(1), 80–102.

Ttard, F. (1999), On fragmentation of working time: a study of causes and effects on work interruptions, Technical Report Technical Report 9, IAMSR.

Tulving, E. (1983), *Elements of Episodic Memory*, Clarendon Press, Oxford.

Tulving, E. (2002), 'Episodic memory: from mind to brain', *Annual review of psychology* **53**, 1–25.

Tynan, P. D. & Sekuler, R. (1982), 'Motion processing in peripheral vision: Reaction time and perceived velocity', *Vision Research* **22**(1), 61–68.

Vandierendonck, A., Van Hoe, R. & Soete, D. (1988), 'Menu search as a function of menu organization, categorization and experience', *Acta Psychologica* **69**(3), 231–248.

Vlaskamp, B. N. S., Hooge, I. T. C. & B., O. E. A. (2005), 'Saccadic search performance: the effect of element spacing', *Experimental Brain Research* **167**, 246–259.

Wallace, D., L. (1983), 'A method for comparing two hierarchical clusterings: Comment', *American Statistical Association* **78**(383), 569–576.

Wan, S. J. & Wong, S. K. M. (1989), A measure for concept dissimilarity and its applications in machine learning, *in* 'International Conference on Computing and Information', Toronto North, Canada, pp. 23–27.

Watkins, M. J. & Watkins, O. C. (1974), 'Processing of recency items for free recall', *Experimental Psychology: General* **102**, 488–493.

Weber, E. (1834), De pulsu, resorptione, audita et tactu, *in* 'Annotationes anatomicae et physiologicae', Koehler, Leipzig.

Weigle, C., Emigh, W. G., Liu, G., Taylor, R. M., Enns, J. T. & Healey, C. G. (2000), Oriented texture slivers: A technique for local value estimation of multiple scalar fields, *in* 'Graphics Interface 2000', Montral, Canada, pp. 163–170.

White, R. W., Jose, J. M. & Ruthven, I. (2003), 'A task-oriented study on the influencing effects of query-biased summarisation in web searching', *Information Processing & Management* **39**(5), 707–733.

Wickens, C. D. (1984), Processing resources in attention, *in* R. Parasuraman & R. Davies, eds, 'Varieties of Attention', Academic Press, New York.

Wiggins, M. & O'Hare, D. (1995), 'Expertise in aeronautical weather-related decision making: A cross-sectional analysis of general aviation pilots', *Experimental Psychology: Applied Cognitive Psychology* **1**(4), 305320.

Wisniewski, E. J. (2002), Concepts and categorization, *in* D. L. Medin, ed., 'Stevens Handbook of Experimental Psychology', 3rd edn, Wiley, NewYork, pp. 467–532. page 467.

Witkin, H. A., Moore, C. A., Goodenough, D. R. & Cox, P. W. (1977), 'Field-dependent and field-independent cognitive styles and their educational implications', *Review of Educational Research* **47**(1), 1–64.

Witkin, H. A., Moore, C. A., Oltman, P. K., Goodenough, D. R., F., F., Owen, D. R. & Raskin, E. (1977), 'The role of the field-dependent and field-independent cognitive styles in academic evolution: a longitudinal study.', *Journal of Educational Psychology* **69**(3), 197–211.

Witkin, H. A., Oltman, P. K., Raskin, E. & Karp, S. A. (1971), *Group Embedded Figures Test Manual*, Consulting Psychologist Press, Palo Alto, CA.

Witkin, H. & Goodenough, D. (1981), *Cognitive Styles: Essence and Origins: Field Dependence and Field Independence*, International Academic Press, New York.

Witten, I. H. & Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Amsterdam.

Wittgenstein, L. (1999), *Philosophical Investigations*, Prentices Hall.

Wolfe, J. M. (1994), 'Guided search 2.0: A revised model of visual search', *Psychonomic Bulletin and Review* **1**(2), 202–238.

Wolfe, J. M., Cave, K. R. & Franzel, S. L. (1989), 'Guided search: An alternative to the feature integration model for visual search', *Experimental Psychology* **15**(3), 419–433.

Wolfe, J. M. & Franzel, S. L. (1988), 'Binocularity and visual search', *Perception & Psychophysics* **44**(1), 81–93.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I. & O'Connell, K. M. (1992), 'The role of categorization in visual search for orientation', *Experimental Psychology: Human Perception & Performance* **18**(1), 34–49.

Xu, J. L. (1999), 'Internet search engines: real world ir issues and challenges'.

Yao, Y. Y., Wong, S. K. M. & Butz, C. J. (1999), On information theoretic measures of attribute importance, *in* N. Zhong, ed., 'PAKDD'99', Beijing, China, pp. 133–137.

Yoccoz, G. N. (1991), 'Use, overuse, and misus of significance tests in evolutionary biology and ecology'.

Yokoi, K. & Uchikawa, K. (2005), 'Color category influences heterogeneous visual search for color', *Optical Society of America* **22**, 2309–2317.

Yule, G. U. (1912), 'On the methods of measuring association between two attributes', *Royal Society of London* **75**(6), 579–642.