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.