A STUDY ON

INTERACTIVE VISUALIZATION

FOR WEB INFORMATION RETRIEVAL

A Dissertation
Submitted to the Faculty of Graduate Studies and Research
In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in
Computer Science
University of Regina

By

Orland Hoeber

Regina, Saskatchewan

April 2007

Abstract

Although Web search has become the primary means by which people find and access information on the Web, the interfaces used by the top Web search engines have changed very little since the early days of the Web. Little support is provided for the tasks of crafting and refining queries, and subsequently exploring search results sets. The goal of this research is to explore methods by which interactive visualization can be used to support Web searchers. A novel paradigm is presented that highlights the use of subject knowledge to support interactive query refinement and interactive search results exploration. Four prototype systems are presented that represent an exploration of the design space. VisiQ provides a visual representation of a query, supporting interactive query refinement. HotMap provides a visual representation query term frequencies in the top search results, supporting interactive search results exploration. Concept Highlighter uses concept information as a means for exploring the search results. WordBars generates a term frequency histogram, supporting both interactive query refinement and interactive search results exploration. In general, the evaluations conducted on these systems produced very positive results. This research highlights the benefits of interactive visualization for Web information retrieval; the prototype systems represent instances of next-generation Web search systems that allow the user to take a more active involvement in the Web search process.

Acknowledgements

First and foremost, I would like to thank my wife Larena for all the support and encouragement she has given me throughout this process. She, along with my parents and the rest of my family have been unwavering in their support of my academic pursuits.

I would also like to thank Dr. Xue Dong Yang for his expert guidance, leadership, and support of my research. His mentorship has allowed me to grow as a researcher and academic.

Dr. Yiyu Yao and Dr. Luigi Benedicenti, as members of my supervisory committee, have provided ongoing advice and suggestions that have improved the quality of my research. To them, I am thankful.

Last but not least, I would like to thank the Department of Computer Science for their support through sessional lecturer positions, the Faculty of Graduate Studies and Research for their support through Graduate Scholarships, and the University-Industry Liaison Office for their support in the filing of the patent applications.

Post Defense Acknowledgments

I would like to thank Dr. Rahul Singh for agreeing to be the external examiner of this dissertation, and for traveling to the University of Regina to be present at the defense. I am grateful for his excellent questions and the resulting discussions.

My thanks also goes out to my supervisor, Dr. Xue Dong Yang, and my committee members, Dr. Yiyu Yao and Dr. Luigi Benedicenti. Their questions were insightful and thought-provoking.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

While it is clear that significant effort has gone into creating Web search engines that can index billions of documents and return the search results in fractions of a second [13, 23], the methods by which users craft queries, and the techniques used to present the search results to the users have remained essentially unchanged since the early days of Web search. Commonly, users simply type their query terms in a box and click the search button; the search results generated from a user's query are presented in a list-based format, with each document surrogate being represented by its title, a snippet containing the query terms in context from the source document, and the URL of the document.

Studies of Web search user behaviour have shown that a large portion of Web search queries consist of only one to three terms [48, 93]. These short queries provide an indication that users of Web search engines often have difficulties crafting queries that accurately reflect their information needs. Clearly, most Web search engines provide little support for users as they attempt to

construct and reformulate their queries; it is up to the user to decide which terms to use (both initially and during query reformulation), and manually add or remove these terms from the query. As a result of this lack of support, Web searchers seldom make subsequent modifications to their queries [89, 93].

Even if the users are able to effectively craft a query, few consider more than three pages worth of search results [89, 93]. Spink et al. noted that "the public has a low tolerance of going in depth through what is retrieved" [93]. This low tolerance may be attributed to the static representations of Web search results that are common in Web search engines. These require the users to consider each document surrogate individually, and to some degree, in the order provided. Most Web search engines provide little ability to manipulate or explore the search results.

The goal for this research is to address the fundamental issues related to these shortcomings of the current Web search engines by establishing a new paradigm for visual and interactive Web search. Through the proposed novel visual models, a set of prototype systems were developed for interactive Web search that allow users to take a more active role within the information retrieval process. The ability for users to explore and interact with their queries and the corresponding search results represents a step towards Yao's vision for Web information retrieval support systems [118].

This research acknowledges that human decision making is fundamentally important as users attempt to craft queries that capture their information needs, and as they attempt to locate documents that are relevant and useful. The development of next-generation Web search engines (of which this research is a part) will have a significant impact on how people will search for and find information in the future.

1.2 Web Search Challenges

In order to illustrate the challenges Web searchers face in crafting a query and evaluating the results of their Web searches, a hypothetical search scenario using the Google search engine is provided. Suppose a user wishes to find information related to "clustering algorithms". They may start by providing this phrase to the Google search engine.

As a result of this query, Google returns the first ten document surrogates from the nearly five million that match the query (see Figure 1.1a). The normal procedure at this point is for the user to start viewing documents that they think are relevant to their information need. Although the list-based representation provides the search results in a simple and intuitive manner for determining the relevance of individual document surrogates, it promotes a sequential evaluation of the document surrogates.

As such, it is common for a user to consider the document surrogates one-by-one, until one of the following three conditions are satisfied. The user may find enough relevant documents to successfully satisfy their information need. If few relevant documents are found, the user may decide to modify their query to generate a new set of search results; or they may give up on finding relevant documents for their current information need.

Of particular interest in this research are the two situations that are common when few relevant documents are found. Can a Web search system support the users in refining their queries? Can a Web search system assist the users in finding relevant documents that are buried deep in the search results, and thereby satisfy their information needs, rather than giving up?

In this example on "clustering algorithms", although there are a number of highly relevant documents on the general topic, the query is clearly too vague
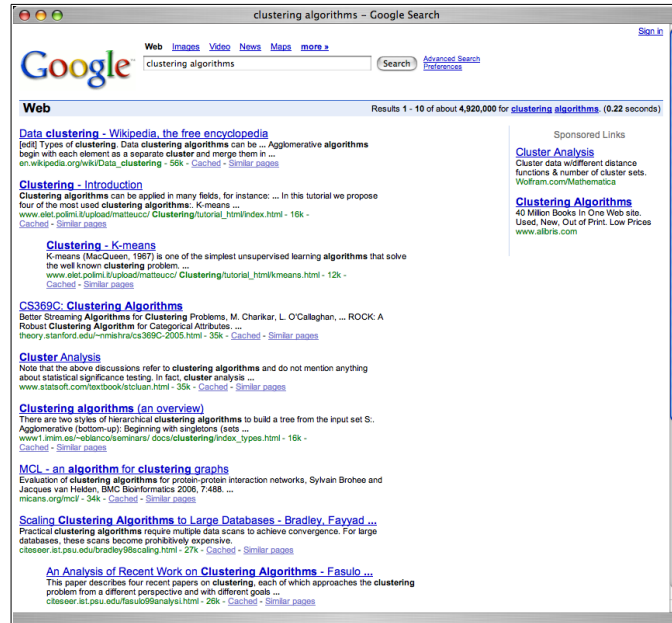
(a) Initial query
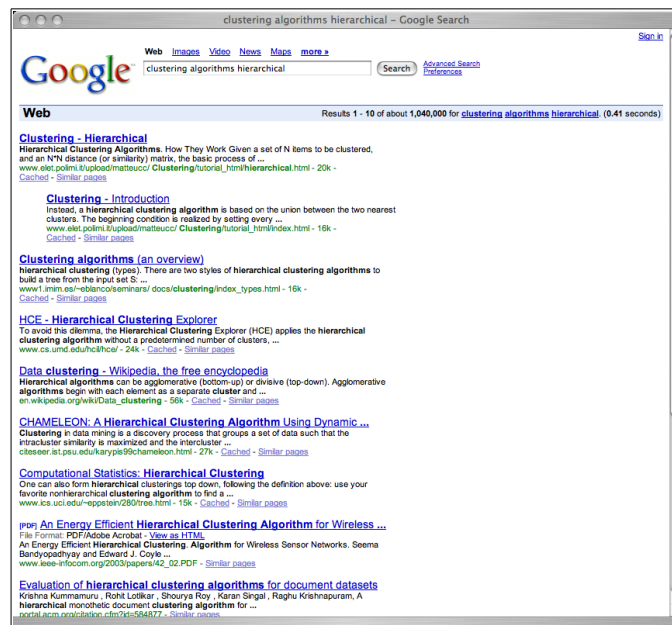
(b) Reformulated query

Figure 1.1: An illustrative example of using the Google search engine for refining a query and finding relevant documents.

to find useful and detailed information on the topic. Commonly, the next step would be to attempt to craft a more specific query. However, the only support for this process provided by Google is the availability of the query box at the top of each search results page. It is up to the user to decide which terms to add or remove from the query in order to construct a more effective description of their information needs. Further, they must manually type these new terms into the query box.

As the user reviews the list of document surrogates in the search results, as well as the source documents identified as relevant to the topic, they may identify the use of certain other terms that could be relevant to their information needs. One such term may be "hierarchical". After typing that term at the end of their query, a new set of search results is provided (see Figure 1.1b).

As with the previous set of search results, the list based representation promotes a sequential evaluation of the results from this new query. Although it may be possible for the user to satisfy their information needs based on the results from this more specific query, the users may wish to continue to reformulate their query to focus on a more specific aspect on the topic. To do so, they must again manually add or remove terms from their query, and then consider the search results in the order provided.

What is clear from this example is that the Google search engine provides little if any support for the users as they reformulate their queries to more accurately capture their information needs. Further, there is no ability to interactively explore the results of a search using Google; the search results must be considered one-by-one, and to some degree, in the order provided. The goal of this research is to explore methods by which interactive visualization can be used to support the users in their Web information retrieval tasks, and to further understand the potential benefits these visual and interactive

systems can provide to the users.

1.3   Approach

In this research, a structured approach has been employed to address the fundamental issues related to the shortcomings of the current state-of-the-art Web search interface technology. The first step in this approach has been to identify the problems and difficulties Web searchers may have with current Web search engines. These problems have been highlighted in the previous Section.

In the second stage of this research, a survey of the current methods for providing visual and interactive support for Web search tasks was conducted. In addition, models for information visualization and interaction were investigated. The results of this survey and investigation are provided in Chapter 2.

The third stage of this research was to assess the current state of technology for supporting Web information retrieval. This included analysing methods for query expansion as well as techniques for exploring and clustering Web search results. Based on this analysis, a new paradigm for Web information retrieval has been proposed that supports interactive query refinement and interactive search results exploration processes, with the goal of assisting users in fulfilling their information needs. This information is provided in Chapter 3.

Two different sources of subject knowledge were identified as being available to support the processes of interactive query refinement and interactive search results exploration: conceptual information from an external knowledge base, and supplemental information extracted from the document surrogates among the top search results of the current query. A concept knowledge base was developed in order to store concept information about computer science

Table 1.1: The design space explored in this research.

		Information Available	
		Concept Information	Top Search Results
Task	Query Refinement	VisiQ	WordBars
	Search Results Exploration	Concept Highlighter	HotMap WordBars

topics collected from the ACM Computing Classification System [1]. The top 100 search results were collected from the Google API [27] to form the basis for the search results-based methods. Although this research has focused on using only these two sources of information, others are available including Wikipedia [111] and The Open Directory Project [62] for the the concept information, and Yahoo! [116] or a traditional information retrieval system such as Apache Lucene [5] for the search results-based techniques.

A classification of the design space was identified, based on two dimensions: the tasks the users perform, and the information that is available (see Table 1.1). A design-oriented research methodology was employed to investigate the design space [20]. In this research methodology, knowledge discovery is the main contribution; this knowledge comes from the design and study of prototype systems which address problems or challenges within the research domain. This process is in contrast to research-oriented design, in which the main focus is product development. The outcomes of this design work are described in Chapters 4 (focusing on interactive query refinement) and 5 (focusing on interactive search results exploration).

Where appropriate, empirical evaluations were conducted with users in order to validate the design decisions and hypotheses made regarding the use of visualization and interaction to support Web search tasks. In addition, the methods used in HotMap and Concept Highlighter were analysed to determine

the potential effectiveness for improving user performance. The results of the experiments and evaluations conducted using the prototype systems are provided in Chapter 6.

This dissertation concludes with a summary of the research activities, a list of the contributions of this research, and an overview of future work in Chapter 7.

1.4   Contributions

One of the fundamental contributions of this research is the new paradigm for interactive Web search that extends the traditional model for Web search to support visualization and a high degree of interaction. This paradigm has formed the corner-stone for exploring methods for providing support for the users' Web search tasks through the use of visualization and interaction techniques.

In exploring the various types of subject knowledge available upon which to support the primary Web search tasks, a concept knowledge base was developed that followed conceptual semantic theories [24]. This concept knowledge base was used to support interactive query refinement and interactive search results exploration in two of the prototype systems developed in this research. This contribution breaks form the norm of using thesaurus-based knowledge bases to support information retrieval processes.

In order to explore the design space, four different prototype systems were developed. The VisiQ system (see Figure 1.2) uses the concept knowledge base to support interactive query refinement processes. A visual representation of the query space is provided, which allows the users to interactively refine the query and determine the potential outcomes of their refinement through a
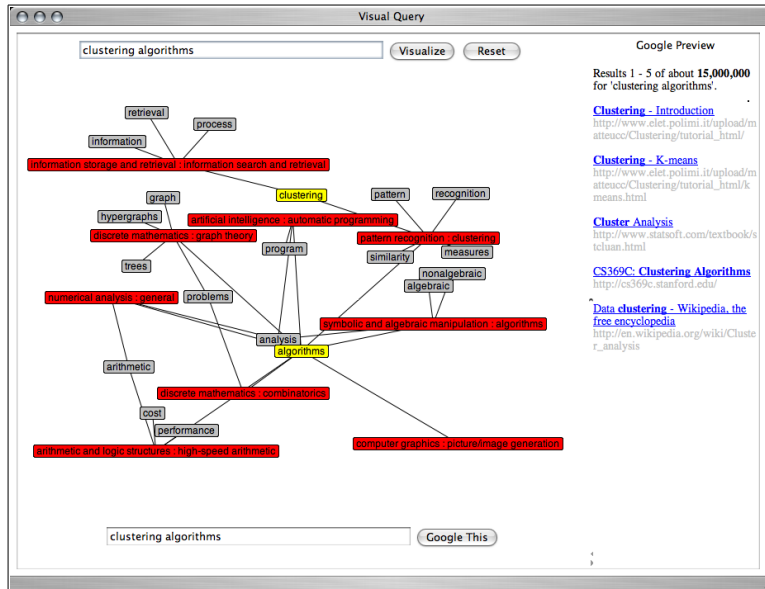
Figure 1.2: A screenshot of the VisiQ system supporting interactive query refinement.
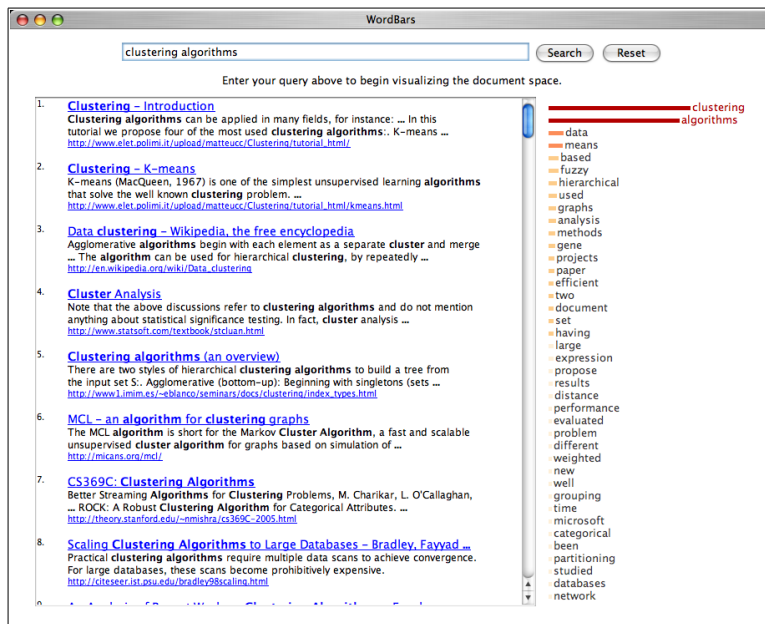
Figure 1.3: A screenshot of the WordBars system supporting interactive query refinement and interactive search results exploration.
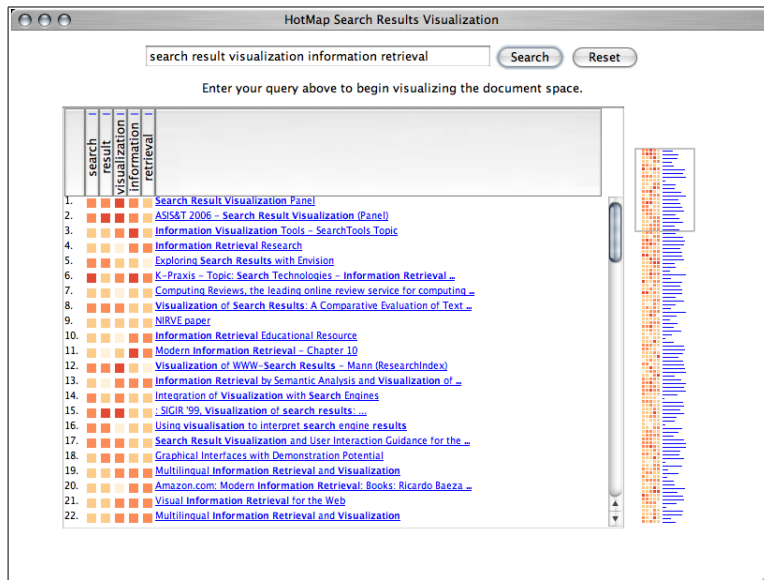
Figure 1.4: A screenshot of the HotMap system supporting interactive search results exploration.
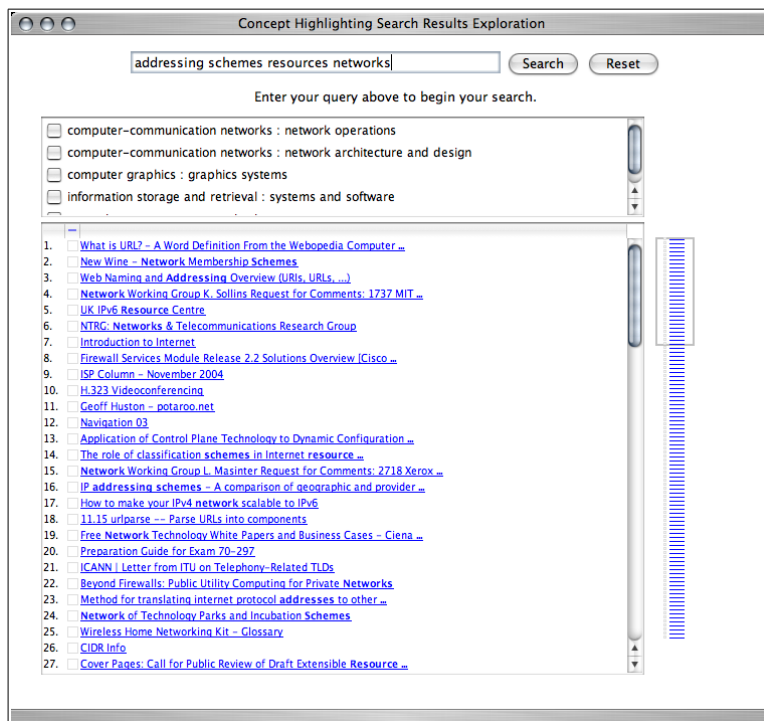
Figure 1.5: A screenshot of the Concept Highlighter system supporting interactive search results exploration.

preview of the search results.

WordBars (see Figure 1.3) uses a term frequency histogram generated from the top 100 search results for an initial query. This system supports the users in both the interactive query refinement process and the interactive search results exploration process. Users can select terms in the histogram to re-sort the search results based on their use within the search results. They can also use the histogram to select terms to add or remove from their query.

HotMap (see Figure 1.4) uses a colour-coded grid representation of the query term frequencies, which is represented at two levels of detail. The overview map supports a visual inspection of all 100 documents in the top search results; the document list window supports a detailed view of specific document surrogates. Nested sorting features are provided to support interactive search results exploration.

Concept Highlighter (see Figure 1.5) uses the concept knowledge base to provide a list of concepts that are relevant to the users' queries, along with the top 100 search results. The users are able to select concepts that are relevant, resulting in a re-sorting of the search results based on a single-pass fuzzy clustering algorithm. This supports a two-phase interactive search results exploration process.

As a result of the user-based evaluations of these systems, the fundamental hypotheses that visualization and interaction methods can be used to effectively support users as they conduct their Web search tasks is supported. Some of the techniques designed and studied in this research proved to be very effective in assisting the users to find relevant document surrogates from among the search results. This contributes greatly to the body of knowledge regarding the effective use of visualization and interaction features to support human-centric tasks.

An added contribution of this research is in the evaluation methodologies used to study the effectiveness of the prototypes developed in the course of this research. In the studies on HotMap and Concept Highlighter a number of new metrics are presented to measure the effectiveness of participants in finding relevant documents using the systems under evaluation. In addition, a categorization of the error types and their severity is used. In the study on WordBars, novel methods for making a baseline comparison, and for ranking the effectiveness of the features of the system are employed.

This dissertation contains sections that have been previously been published during the course of this research. The underlying model for interactive Web information retrieval was presented in [40]. The concept knowledge base (Section 4.2) was first published in [43]. The VisiQ system (Section 4.2) was introduced in [44]. The HotMap system (Section 5.2) was introduced in [41]. The Concept Highlighter system (Section 5.3) was introduced in [42]. The coordinated views in HotMap and Concept Highlighter were discussed in [38], and the results of the preliminary user evaluations of these systems (Sections 6.3 and 6.5) were presented in [37]. The WordBars system (Sections 4.3 and 5.4) was introduced in [39].

Chapter 2

Visualization and Interaction

2.1 Fundamentals

Information visualization is a technique for creating graphical representations of abstract data or concepts [110]. Moreover, information visualization promotes a cognitive activity in which users are able to gain understanding or insight into the data being graphically displayed by taking advantage of human visual information processing capabilities [91]. In essence, information visualization provides a link between the human vision system and computer systems [122], with the end result being an amplification of cognition [14].

At the most basic level, information visualization techniques are used when one draws graphs to visually represent data sets. However, when these data sets are large, high-dimensional, or complex, generating useful visual representations can become a challenging problem. In general, information visualization techniques allow the display of large or complex sets of abstract data in a coherent manner, allowing the viewer to compare and explore the data visually [104].

When graphically representing abstract objects in a visual representation,

there are a number of visual features that are available for representing the various dimensions or attributes of the data. These include spatial location, colour, shape, and orientation, among others. Care must be taken to select and use visual features that can be easily decoded and understood by the viewer. For example, while the spatial location and colour of an object can be perceptually separated to represent multiple dimensions in the information, this cannot be done as easily with the shape and orientation of a glyph [110]. Readers interested in the human perception system and its relevance to information visualization are directed to Ware's excellent book on the subject [110].

When designing any information visualization system, careful consideration and restraint must be employed so as to not overload the viewers with unnecessary visual complexity that is difficult to perceive and interpret. Tufte recommends a design strategy of the *smallest effective difference*: "make all visual distinctions as subtle as possible, but still clear and effective" [103].

The use of colour often poses an especially challenging dilemma. Although colour is an extremely valuable tool for encoding information within a visual display, the overuse of colour has the effect of becoming noise within the information display. As noted by Tufte, the end result of the overuse of strong colours is a "visual war with the heavily encoded information" [102]. Design principles such as those suggested by Tufte [103, 102, 104] are of value, as is an understanding of colour theory [98, 110].

A useful colour theory that has a direct implication for the use of colour in information visualization systems is the *opponent process theory* [110]. This theory proposes that there are six elementary colours which are arranged perceptually as opponent pairs along three axes: black-white (luminance), red-green, and yellow-blue (see Figure 2.1) [36]. Varying the colours monotonically
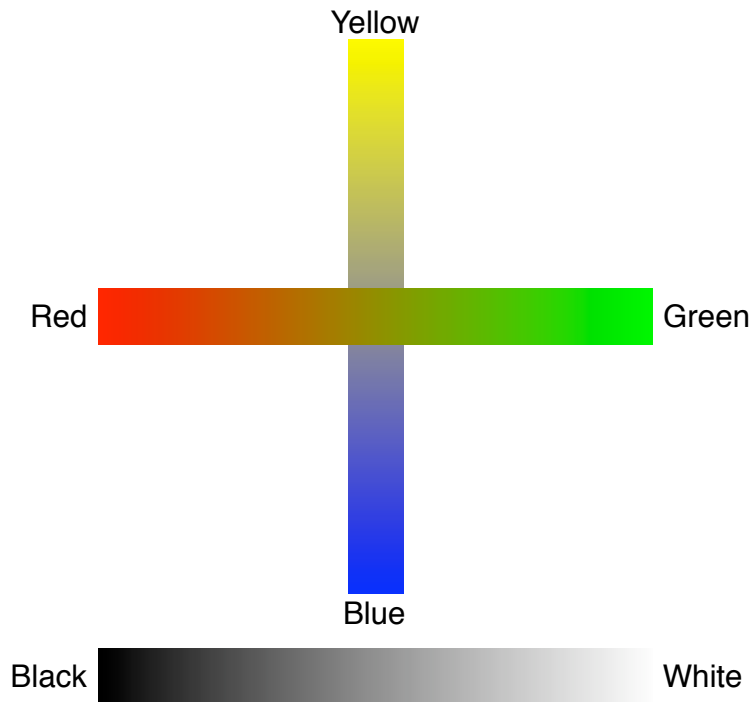
Figure 2.1: Opponent process theory of colour perception.

on one or more of the chromatic channels can produce a perceptually ordered colour scale (a feature which is not present in the rainbow colour scale, for example) [102, 110]. In addition to playing a vital role in perceiving depth, motion, and form, the luminance channel can also be valuable in making distinctions between values in an information display.

In order to further aid in the cognition supported by a good visual representation, the interactive manipulation of the visual representations is an integral part of any information visualization system [14]. Clearly, one of the great benefits of computer-generated visualizations are that they allow the users to interact with and manipulate the information being graphically represented. However, as with any interactive system, there is a challenge in ensuring that the users remain in control, and that the system provides effective feedback

that reflects the results of the user's actions [8]. As such, when discussing any visualization technique, the methods by which the user can interact with and manipulate the graphical representation are important in understanding how the information visualization system can support the users in conducting their tasks.

While the techniques for information visualization are as numerous as the types of data sets they graphically describe, this Chapter is constrained to information visualization techniques that have been applied to information retrieval systems, and in particular to those that support the visualization of queries and search results, both of which primarily contain textual data. The visualization of such unstructured or semi-structured data is difficult due to the lack of well-defined features upon which to base the visual representation [122]. Often, there is a need to identify features within the data and to construct data structures which support graphical representations.

Although many of the systems discussed in the remainder of this Chapter are designed around traditional information retrieval systems, the primary interest in this dissertation is whether and how these techniques can be applied to Web search. As such, each technique is critiqued with respect to its utility for supporting Web search tasks.

2.2   Query Visualization

The application of information visualization to queries has primarily been focused on the visualization of Boolean queries [51, 119, 4, 94]. This research in supporting Boolean queries addresses the great difficulties many users have with crafting correctly specified queries of this type [33]. In the context of visualizing free text queries, very little work has been done. For example, in

Zhu and Chen's survey of information visualization techniques for unstructured textual documents [122], the only mention of visual techniques for supporting query specification was through the use of subject hierarchies as a means for suggesting appropriate query terms. However, no specific examples were cited.

Two systems that do provide visual representations of text queries are those by Stojanovic and Joho. Stojanovic [95, 97] visually represents the current query within the context of its neighbour queries using a simple box layout (see Figure 2.2). Navigating within this query space can allow the users to express their information needs more precisely. Although the visual representation of the query space is secondary to the techniques for determining neighbour queries, a visual representation of the query space provides a meaningful organization of candidate queries. However, a requirement of this work is that the corpus being search be represented as an annotated information repository. Since this additional information is not available in the common Web search indexes, the applicability of this approach to Web search is limited.

Joho [49, 50] uses a menu structure to visually represent a hierarchy of terms to add to a query in an interactive query expansion system (see Figure 2.3). In this work, hierarchies are automatically generated based on the co-occurrence of terms within a set of retrieved documents [87]. Although this work addresses the problems of the simple list-based representation of potential query terms, it requires that a number of documents returned from the initial search be retrieved in their entirety. For Web search, this may result in the system introducing a significant delay for the users, which is contrary to the preferences of Web users for fast Web search engines [28]. Further, although the visual presentation of candidate query terms is superior to a simple list-base representation, it results in candidate query terms being hidden from view until the user navigates to the appropriate location.
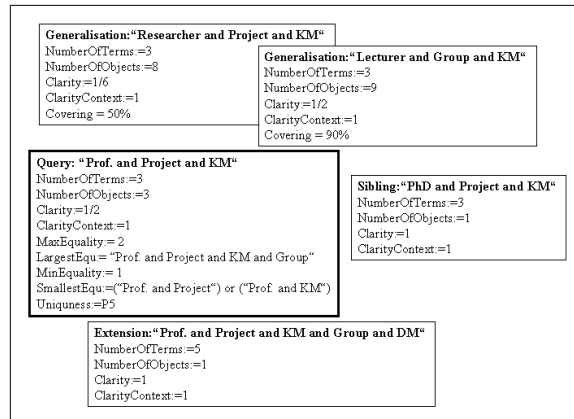
Figure 2.2: A screenshot from the work of Stojanovic [97].
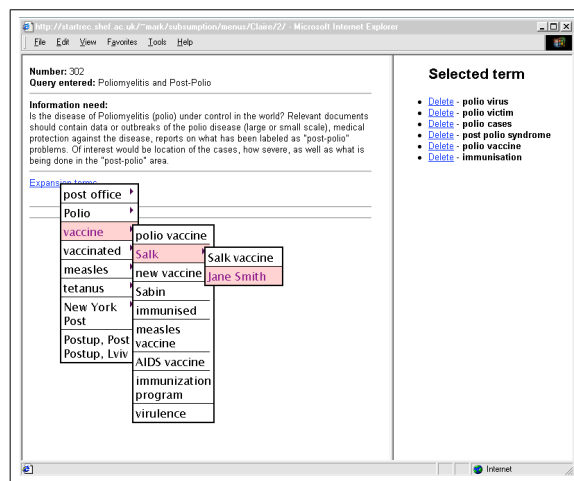
Figure 2.3: A screenshot from the work of Joho et al.[50].

Visualization of a Web search query is challenging since there is often little or no information available upon which to generate a visual representation. For example, in the absence of additional information, how can one generate a visual representation of a set of search keywords entered by a user? In Chapter 4, the VisiQ system is described which makes use of a concept knowledge base as a means of deriving additional information about a query, resulting in a visual representation of the corresponding *query space*. The WordBars system visually depicts term frequency histograms generated for all the unique terms found in the top search results from an initial query. This histogram supports the users in identifying and selecting terms to add or remove from the query.

2.3   Search Results Visualization

Many systems have been developed in recent years to represent search results in a visual manner, both for Web search and for traditional information retrieval systems. These can be categorized based on whether they provide an entirely textual representation, a visual representation based on the entire document contents, or a visual representation based on an abstraction of the document (i.e., the document surrogate). While it is not feasible to provide an exhaustive survey and analysis of all the methods for representing search results, a brief review of relevant and interesting techniques is provided in the following sections, with emphasis given to those techniques which are appropriate for Web information retrieval. A further discussion regarding the unique challenges and opportunities in Web search is provided in Chapter 3.

2.3.1 Textual Representations

While the supplemental information provided by the four primary Web search engines (Google[1], Yahoo[2], MSN[3], and AOL[4]) differs slightly, all use a textual representation of the Web search results, displaying a list of ten document surrogates per page. Within these lists, each document surrogate primarily consists of the title of the document, a snippet showing the context of the query terms, and the URL of the document.

Since the documents themselves are primarily textual, this textual representation of Web search results is a logical method for displaying this information. However, since the lists are static, they provide little ability to manipulate the search results; they only lend themselves to a sequential evaluation of ten document surrogates at a time. This textual interface style requires the users to read the information (which can be time consuming), or scan the information (which can be error prone). Further, since text can consume a large amount of space in an interface, the list based representations commonly limit the number of document surrogates displayed per page.

2.3.2 Document Visualization

Document visualization can be defined as the process of converting textual information from a document into graphical representations that can be processed visually rather than read. Since preattentive processing of certain types of graphical information is significantly faster than the non-preattentive processing required for reading [110], there is a great opportunity for taking advantage of the human visual processing capabilities when presenting textual

---

[1]http://www.google.com/
[2]http://www.yahoo.com/
[3]http://www.msn.com/
[4]http://www.aol.com/

information.

However, the representation of textual information in a visual manner is by no means a simple task. At the most fundamental level, one can think of a document as a collection of terms, represented by a high dimensional vector. Dimensional reduction techniques can be used to map a set of such document vectors into two or three dimensional space, resulting in each document occupying some point in space. The spatial proximity of two documents implies similarity, resulting in a visual clustering of documents.

These vector-based techniques have been used for the visualization of collections of documents in systems such as Galaxy of News [80] (see Figure 2.4), and ThemeScape [113, 112] (see Figure 2.5). However, accessing and viewing the information on specific documents is not well supported. Hearst noted that "although intuitively appealing, graphical overviews of large document spaces have yet to be shown to be useful and understandable for users" [33].

Other techniques retain the linear structure of the documents, and provide abstract representations of their contents. For example, in SeeSoft [19], each line of text is abstracted to a single horizontal line in the visual representation, preserving the general layout of the document (see Figure 2.6). Colour is used to highlight the lines containing specific terms within the document. While useful in some situations, the resulting visual representation does not make efficient use of display space.

Rather than retaining the layout of the document, the contents can be divided into fixed blocks, and the frequency of the query terms can be represented by colour coding in each block, as in TileBars [32] (see Figure 2.7). The result is a set of bars (one for each document) whose widths are relative to the length of the documents, and whose heights are relative to the number of
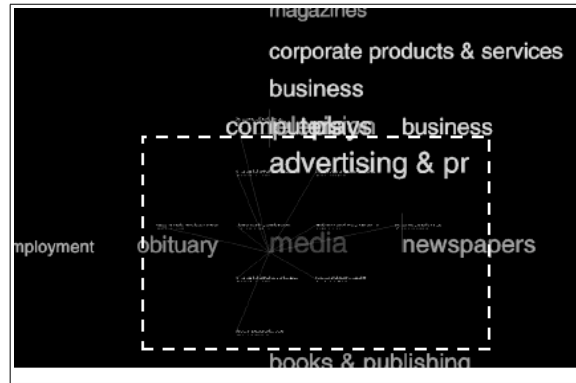
Figure 2.4: A screenshot from Galaxy of News [80].
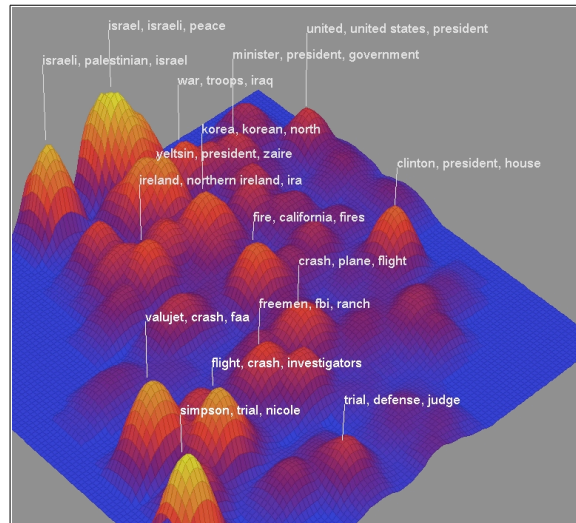
Figure 2.5: A screenshot from Themescape [113].

Figure 2.6: A screenshot from SeeSoft [19].
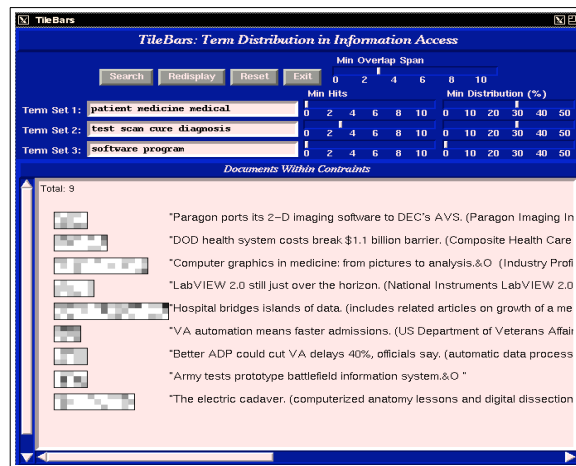
Figure 2.7: A screenshot from TileBars [32].

Figure 2.8: A screenshot from the work of Heimonen and Jhaveri [35].

query terms (or sets of query terms). Colour coding within these bars represents the frequency of use of the query terms relative to the location within the document. This results in a more compact representation than the previous example.

In the work by Heimonen and Jhaveri [35], each document is divided into four equal sized blocks (see Figure 2.8). The occurrences of all of the query terms within a 20-word window for each block are counted and depicted using a visual indicator similar to the technique used in TileBars [32]. This indicator is displayed beside each document surrogate in the list-based representation of the search results.

One common theme among these systems is that they all require access to the full textual contents of all the documents in the search results in order to generate the visual representations. Since Web search engines generally provided only document surrogates (i.e., titles, snippets, and URLs), to apply these techniques to Web search would require retrieving each document individually. The additional time required to do this supplemental document

retrieval would result in a Web search system that is unable to display the search results in real-time, contrary to the preferences of Web users [28]. Since one of the goals in this research is to provide an interactive Web search system, of interest are the techniques for the visualization of document surrogates.

2.3.3 Document Surrogate Visualization

A meta-search system is a system which makes use of the search results provided by other search engines. Since access to the textual contents of each document is not feasible for a meta-search system, the visual representation of document surrogates is a viable alternative. A document surrogate consists of summary information, attributes, and other meta-data that represent the document in the search results. Document surrogates are the primary data objects in the list-based representation used by many Web search engines, where they commonly consist of the title of the document, a snippet showing the query terms in context, the URL of the document, as well as other information. See Section 3.5 for more details on meta-search for the Web.

Envision [70] uses a highly customizable scatterplot and iconic visualization to represent the many different attributes available as part of their custom information retrieval system (see Figure 2.9). Although this visual representation was shown to be very powerful, it makes use of information that is not commonly available in the document surrogates returned by Web search engines. Further, there is an added level of complexity in this highly customizable interface that may make it too difficult for the general public to use effectively.

In VIEWER [9], all possible combinations of the query terms are generated and searched for in the document surrogates returned by the AltaVista search engine (see Figure 2.10). A histogram of these query term combinations is
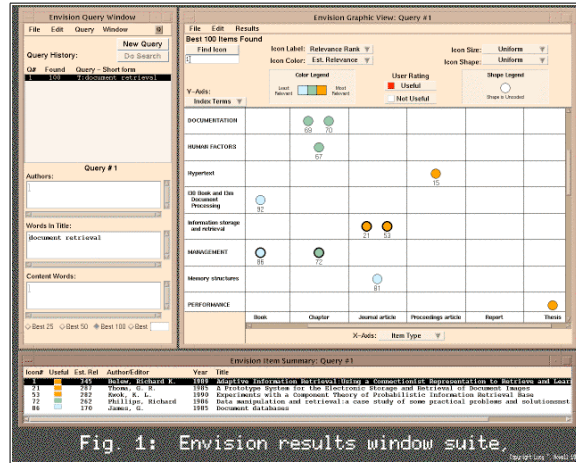
Figure 2.9: A screenshot from Envision [70].
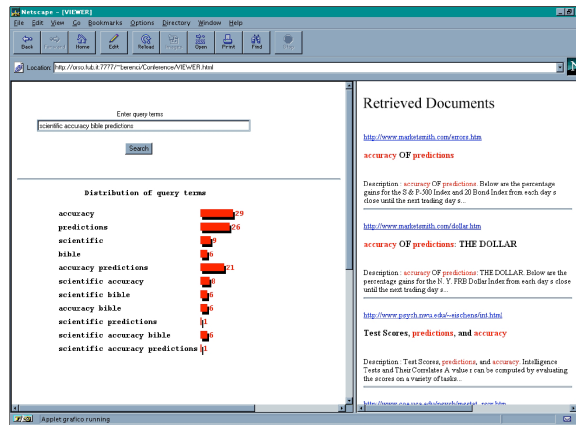
Figure 2.10: A screenshot from Viewer [9].

provided to the user, which can be used to select subsets of the search results
for further investigation. Although this system provides valuable information
to the user in terms of how the query terms are used in the search results
set, the information is provided in the context of the query terms, with little
additional information provided with respect to specific document surrogates.
While this information is of value in narrowing the search results to smaller
subsets, and perhaps reformulating queries, it provides no additional support
beyond the list-based representations when evaluating the relevance of specific
documents.

xFind [3] provides three different interfaces to a custom Web document in-
dexing system: a simple list-based representation; a scatterplot representation
similar to that in Envision [70]; and a vector-based spatial clustering repre-
sentation similar to that in ThemeScape [113] (see Figure 2.11). While these
representations of the search results take advantage of the extra information
that is available through their indexing system, this information is not available
with other search engines. Further, the spatial layout of the two visual rep-
resentations maps the document surrogates to points in the two-dimensional
display, making it difficult to view the additional information present in the
document surrogate, or to make comparisons between document surrogates.

WaveLens [73] provides a focus+context representation of the search results
allowing the users to dynamically zoom into document surrogates of interest
(see Figure 2.12). The results are provided in the traditional list-based rep-
resentation. As users move their mouse over a document surrogate, its font
size increases as the font size of the other document surrogates not in focus
decrease. This results in a fisheye lens effect. Additional text from the docu-
ment is dynamically added or removed from a document surrogate by clicking
the mouse. While this technique may make it easier for the users to read the
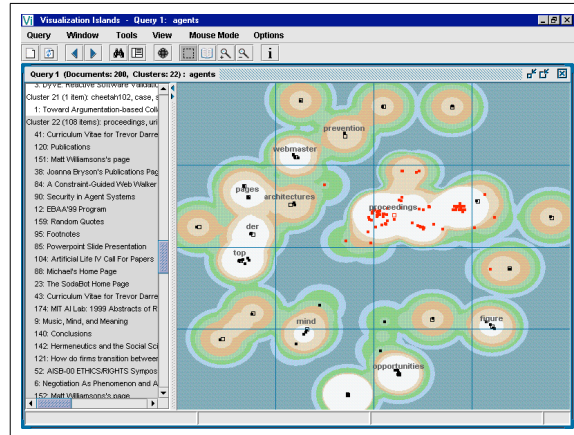
Figure 2.11: A screenshot from xFind [3].

Figure 2.12: A screenshot from WaveLens [73].

contents of the list of search results, the primary interaction method remains a sequential evaluation of the document surrogates, with little support for manipulating and exploring the search results.

Many other visualization systems exist for representing the document surrogates returned by Web search engines, including a number of publicly accessible meta-search engines such as Kartoo[5], Mooter[6], and Grokker[7]. An evaluation of the merits and problems with these systems is beyond the scope of this dissertation, although one review indicated that some of these systems do not add any support for the users in assimilating or processing the information [75].

In Chapter 5, three prototype systems are introduced that make use of visualization techniques based on the document surrogates. In HotMap, query term frequencies are visually represented at two levels of detail. In Concept Highlighter, fuzzy membership scores between each document in the search results and a set of selected relevant concepts are visually depicted. In Word-Bars, a term frequency histogram provides a visual overview of the search results set.

2.4 Models for Interaction

Two theoretical models of human-computer interaction are described in the remainder of this Chapter, with specific focus on how information visualization can support the users' tasks. These models are used to describe the benefits of the Web search interfaces developed as part of this research.

---

[5]`http://www.kartoo.com/`
[6]`http://www.mooter.com/`
[7]`http://www.grokker.com/`

2.4.1 Spence's Navigation Framework

Spence suggests a framework for the human activity of navigating within information systems [90, 91]. This navigation process is concerned with learning about the features of the information system, and interacting with the system in order to satisfy a task goal (or a set of task goals). One of the fundamental features of this framework is that it describes the navigation and interaction with a system with respect to the development and subsequent interpretation of an internal model in the mind of the user. That is, it is the user's understanding of the features and information within the system that guides their ability to make use of a system to fulfill their task goals.

Although searching and navigation are commonly thought to be two different (although complementary) activities, navigation within an information system that supports searching allows the users to formulate mental models of the information space and take appropriate steps to satisfy their information needs. Therefore, this navigation framework is a useful model for discussing the methods of interaction within Web search interfaces.

The navigation framework consists of a cycle of four cognitive activities: *browsing*, *formulating an internal model*, *interpreting the information*, and *formulating a browsing strategy*. A diagram of the framework is shown in Figure 2.13. In three of these activities, the primary information provided by the information system is the externalization of the data. While in general the externalization of the data consists of the interface of the information system, of specific interest is when this interface includes visual representations of the data.

In the *browsing* activity, the user views and interacts with the interface of the information system. Through the visual perception of the interface (i.e., the visual representation of the data) and the interaction with the features
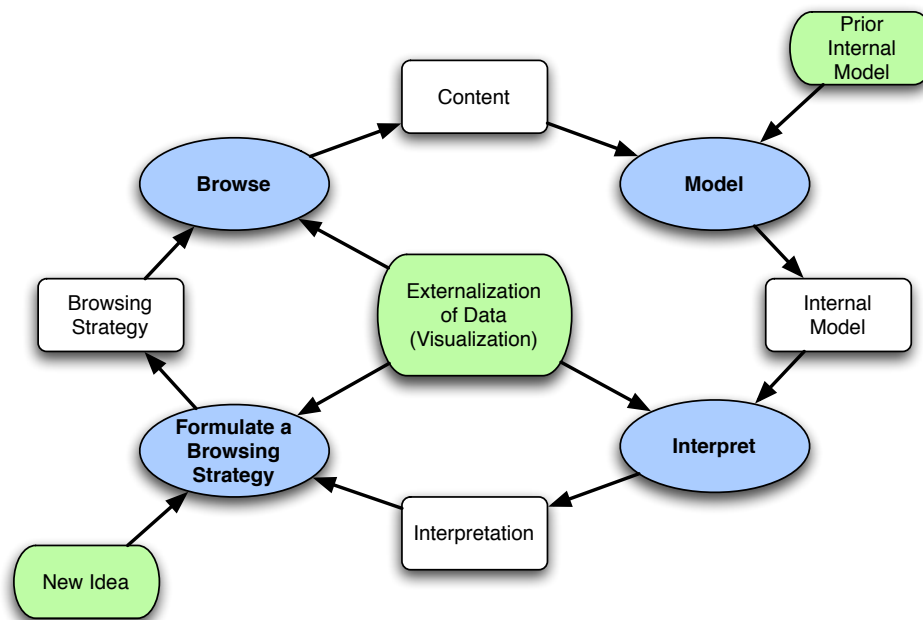
Figure 2.13: Spence's navigation framework.

of the interface, the user absorbs content that is displayed within the system. During this browsing process, the user *develops an internal model* of the content viewed. This internal model is also influence by internal models developed from prior use of the system, or from similar systems. This explains why new users of a system can easily learn the features when they are already familiar with a similar system (the internal model of the system with which they are already familiar contributes to the development of the internal model for the new system).

The ability to generate an accurate internal model of the data within an information system is greatly influenced by the manner in which the data is externalized. In particular, when information visualization techniques are used, the user's understanding of complex data can either be enhanced or hindered, depending on the appropriateness of the visualization technique. Further, the degree to which the user can easily interact with the visual representation

can also affect the user's ability to generate an accurate internal model. The manipulation of the visual representation can lead to insight into features of the information space. As such, care must be taken to choose appropriate visual representations and to allow the users to interactively manipulate the information in order to more easily understand the underlying features of the information presented to the user.

The ability of the user to understand how the process of navigation can assist them in achieving their goals depends greatly on their *interpretation* both of the internal model as well as the externalization of the data. A good visual representation of the data supports the users in both generating an accurate internal model, as well as subsequently interpreting that mental model and the visual representation with respect to their current task goals. The resulting interpretation supports the user in *formulating a subsequent browsing strategy.*

The cognitive activity of formulating a browsing strategy is rather complex and varied. A user's browsing strategy is developed based on a combination of a number of factors, including cognitive determinants (i.e., the interpretation of what the user has seen and done with the interface thus far, or a new idea the user has) or a perceptual determinant (i.e., something the user sees in the interface). The strategy may be part of a larger planned strategy with a specific goal in mind, or may be an opportunistic strategy with a new, yet possibly related, goal. The resulting browsing strategy is also influenced by the user's understanding of what they are able to do with the interface.

The output of this step is a browsing strategy that feeds back into the cognitive activity of browsing within the interface. The process continues in a cycle until the user chooses to stop using the system. In successful situations, the user interprets their internal model as having successfully satisfied their task goal. In unsuccessful situations, the user quits because they are unable

to formulate a browsing strategy that they feel will lead to satisfying their goal. These unsuccessful conditions are commonly due to problems in the externalization of the data, leading to the generation of inaccurate mental models or incorrect interpretations of the data that is available.

The starting point within this framework generally begins with the user viewing an externalization of the data (i.e., viewing the interface of the information system). From here, the user may begin with the cognitive activity of interpreting what they see, formulating a browsing strategy, or simply beginning browsing the interface. Subsequent cognitive activities and interaction follow the cycle presented in this framework.

This framework is useful for this research in that it can assist in describing the cognitive activities that a user must perform as they use an information system. This framework is used in subsequent Chapters to describe how the visualization and interaction features of the prototype Web search interfaces (VisiQ, WordBars, HotMap, and Concept Highlighter) support the users in fulfilling their Web search tasks.

2.4.2   Norman's Stages of Action

Norman's stages of action framework describes two gaps that exist between users of a system and the system itself [68, 69]. The *gulf of evaluation* is the cognitive distance between what the user sees on the display and their mental representation of their tasks and goals. The *gulf of execution* is the cognitive distance between the user's current task and the actions the system supports for fulfilling this task.

Although this framework was originally used to describe the use of any device one might interact with, this discussion focuses on its application to
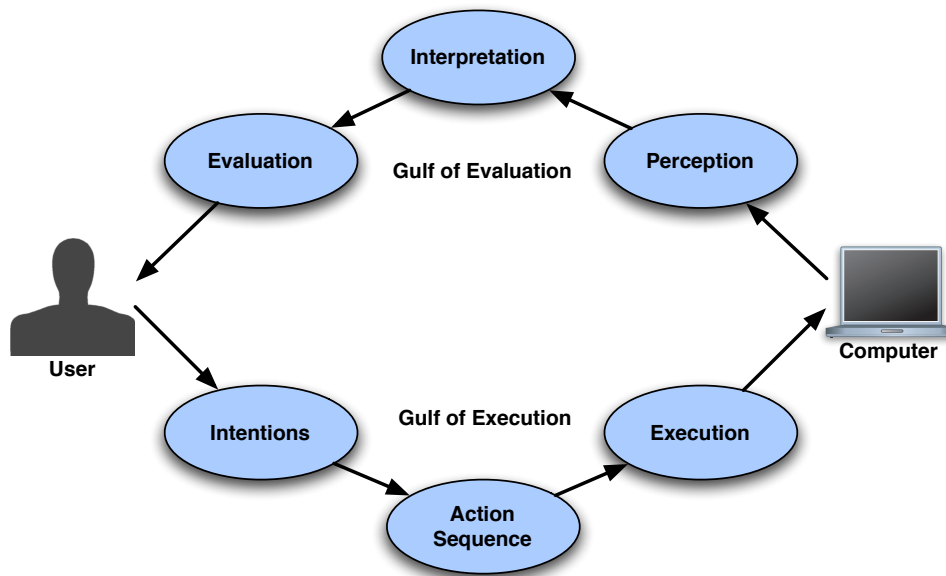
Figure 2.14: Norman's stages of action framework.

computer systems. In this context, the gulf of evaluation deals with the effectiveness of the visual display of the system (including the information visualization features); the gulf of execution deals with the ability of the users to interact with the system. As noted by Rosson and Carroll, "Norman's framework emphasizes that people's interactions with a computer are not simply a matter of looking at the display and deciding what to do" [84].

In each of the gaps between the user and the computer system, there are three stages of action that occur, as illustrated in Figure 2.14. The framework is shown as a cycle to highlight the feedback between the user performing an action with the system, and the resulting changes to the visual display. This feedback is important not only in supporting the awareness of the user in the outcome of their current action, but also in supporting the user in understanding the current state of system with respect to their overall goal.

Gulf of Evaluation

The three stages of action required by the users to cross the gulf of evaluation
are *perception*, *interpretation*, and *evaluation*. These actions are primarily
subconscious, often occurring automatically when one looks at a computer
display (or any physical device, for that matter).

In the *perception* stage, viewing the display reveals the basic structure
of the interface. The basic display elements, such as pixels, simple geometric
objects, and characters, are perceived first. These are grouped into higher-level
structures such as paragraphs, icons, and buttons. This grouping continues
into even higher-level structures based on Gestalt principles, such as proximity,
similarity, and symmetry [53, 84]. The outcome of this perception stage are
the objects and groups within the information display.

In the *interpretation* stage, the user recognizes the perceptual objects and
groups as specific types of display elements within the software system. This
interpretation is supported by the user's previous experience with the current
system, as well as their previous experience with similar systems. A direct
result of interpreting a user interface element is knowing what can be done
with it [84].

In the *evaluation* stage, the user attempts to make sense of the information
that has been perceived and interpreted with respect to their current task and
overall goal. This requires that the user integrate what they see on the display
with their internal model of the system, their current task, and their overall
goal. If the system is operating as expected, the user's internal model will be
reinforced and expanded. However, if the outcome is not what was expected,
the user may become confused and frustrated.

Gulf of Execution

The three stages of action required by the users to cross the gulf of execution are to create *intentions*, devise an *action sequence*, and then *execute* the action sequence. While expert users conducting well-understood tasks will conduct these actions subconsciously, in many situations, some of these actions may require conscious thought.

In the *intentions* stage, the user must convert their goal for using the software system into an intention to do some action within the user interface. The ability to succeed in this stage depends greatly on the user's internal model, as well as their ability to cross the gulf of evaluation in order to determine the current state of the system. The type of intention depends on the task, and is normally a high-level goal.

In the *action sequence* stage, the user converts their high-level intentions for performing some action into a specific action sequence consisting of fundamental interaction operations. For example, if the intention is to close a window, the action sequence may include moving one's hand to the mouse, moving the mouse to the control that closes the window, and clicking the mouse button. Clearly, for simple intentions, this all occurs subconsciously; for more complex intentions, a significant amount of cognitive effort may be necessary to decide how to achieve the desired outcome.

In the *execution* stage, the user must make use of the input devices of the computer system to realize the action sequence. This is the physical aspect of using the computer. As noted by Rosson and Carroll, "in some sense, execution is an inconvenience - what users really want is to accomplish their goals directly, but they must do this by carrying out a sequence of physical actions" [84].

Goals

In addition to the six stages of action described above, there is a seventh stage that is concerned with the user's goals for using the system [8]. In the *goal* stage, the user must form a goal for using the system. This goal will often begin with a real-world goal, which must be converted into a system goal. This system goal not only drives the intentions stage in the gulf of execution, but is also considered during the evaluation stage of the gulf of evaluation.

In his discussion on the stages of action framework, Norman [69] noted:

> The seven stages form an *approximate model*, not a complete psychological theory. In particular, the stages are almost certainly not discrete entities. Most behavior does not require going through all stages in sequence, and most activities will not be satisfied by single actions. There must be numerous sequences, and the whole activity may last hours or even days. There is a continual feedback loop, in which the results of one activity are used to direct further ones, in which goals lead to subgoals, intentions lead to sub-intentions. There are activities in which goals are forgotten, discarded, or reformulated.

This framework provides a valuable method for discussing information visualization systems. The gulf of evaluation can be used to illustrate how a visual representation supports the users in perceiving, interpreting, and evaluating the information being represented. The gulf of execution can be used to highlight how the visual representations support the users in forming intentions, generating action sequences, and executing those action sequences. All this

---

[8]This might be considered the zeroth stage, since the user's goal comes first when using any system.

occurs with the purposes of fulfilling a goal with the information visualization system.

Chapter 3

Web Information Retrieval

3.1 Fundamentals

3.1.1 Traditional Information Retrieval

Information retrieval is the science of automated searching within digital collections. In general, the discipline is concerned with finding a set of relevant documents from a document collection that match a user's specified information need. The goal is to automatically find all the relevant documents, while at the same time, selecting as few of the non-relevant documents as possible [105]. Traditionally, the main emphasis of information retrieval has been on document and text retrieval [109], although many of the principles also apply to the retrieval of other types of information, such as images, video, and audio [6].

In general, an information retrieval system has two inputs: the document collection to be searched and the user-supplied query; and one output: the set of documents that match the query. One of the fundamental challenges is

to specify data structures for storing documents and queries that are space-saving, yet support the efficient and accurate matching of queries to documents. Clearly, there is a great benefit to indexing the document collection once, and storing the internal representation for future querying.

Given an indexed document collection and a user-supplied query that represents their information need, the matching process seeks to find documents from the collection that are relevant to the query. This matching process is highly dependent on the methods by which the document collection is indexed, and the methods by which the user specifies their query. Common examples are boolean matching, fuzzy matching, vector-based matching, and probabilistic matching [105, 6].

The output of this matching process is a set of references to documents from the collection. It is desirable that the set of matched documents be sorted in order of relevance to the query. The goal of information retrieval is for a large portion of this set of documents to be relevant to the user's information need (commonly measured using the precision metric) and for this set of documents to contain many, if not all, of the relevant documents from the collection (commonly measured using the recall metric).

Even within early information retrieval research, feedback was cited as a method for improving information retrieval effectiveness [105]. In this early work, feedback was described in terms of the users learning better ways of specifying their queries. This has been extended to automatic and interactive methods, and is commonly referred to as query expansion [18].

3.1.2 Information Retrieval on the Web

When people today think of information retrieval, text searching, and document searching, they commonly think of Web search engines. The use of Web

search engines have become common place among Web users, and is increasingly being used in all aspects of society [56]. Nielsen reported that 88% of Web users start with a Web search engine when provided with a task to complete using the Web [67]. This is in support of earlier studies that reported nearly 85% of Web users find new Web pages using search engines [28]. Nielsen noted that instead of looking for sites to explore and use in depth, Web searchers are commonly seeking specific answers [67]. In essence, Web search engines have become answer engines.

Given the importance Web users place on search engines, information retrieval on the Web has become a very active research area, both academically and commercially. Although there are unique challenges posed by the Web for information retrieval systems, there are also unique opportunities for improvements over traditional information retrieval systems.

Although the Web can be considered a single distributed document collection, this collection has unique features that make many of the traditional approaches to information retrieval impossible or impractical to implement [117]. These features include the size of the Web (over four billion documents, and growing), the generality of the collection (documents available on virtually any topic), and the potential generality of individual documents (single documents may discuss multiple distinct topics). Yang noted that information retrieval on the Web "must deal with mostly short and unfocused queries posed against a massive collection of heterogeneous, hyper-linked documents that change dynamically" [117]. Clearly, this is a much more challenging problem than those traditionally addressed by information retrieval research.

However, there are also features of the Web that can be used advantageously to increase information retrieval performance. These features include the general link structure of the Web, the implied relevance of the linking text,

the structure of the Web documents, and the vast amounts of Web search engine usage statistics. Many Web search engines go beyond traditional term-based information retrieval, and take advantage of these additional features of the Web. A prime example of this is the PageRank algorithm used by the Google search engine [13].

Although a thorough evaluation of the techniques for conducting information retrieval on the Web is beyond the scope of this dissertation, interested readers are directed to a number of good survey papers and books on this topic [6, 52, 54, 79, 117].

One topic that appears to get little consideration in the research literature on Web information retrieval is the user interface provided by the search engine. The user interface is the gateway to the functionality provided by the system. However, there seems to be a lack of acknowledgment of the importance of the methods by which users are able to specify their queries, or the methods by which search results are provided to the users. It seems that the simple query box and the list-based representations of Web search engines have become so common place that there is little if any discussion on whether these features are providing adequate support the the users.

It is interesting to note that although user interfaces have been an active research area of traditional information retrieval for many years [6, 33], few of these results have been applied to Web information retrieval. Most Web search engines continue to use a text-based interface, even though significant advances have been made in visualization for traditional information retrieval systems (some of which were discussed in Chapter 2).

As noted by Marchionini, "Much research and development is required to produce interfaces that allow end users to be productive amid such complexity

and that protect them from information overload" [60]. Although this comment was made in regards to traditional information retrieval research, it is valid for Web information retrieval as well. This provides a clear motivation for this research on visualization and interactivity as a means of supporting users in their Web search tasks.

3.1.3   Web Search User Behaviour

A number of studies have been conducted in recent years evaluating the Web search behaviour patterns of users, as deduced from Web search transaction logs. These studies include those by Silverstein et al. [89], Jansen and Pooch [48], and Spink et al. [93].

A number of common themes regarding users' abilities to craft a query have emerged from these studies. In particular, these studies have shown that few searchers provide queries containing more than three terms; more common are the short one and two term queries [48, 93]. In addition, few searchers make use of advanced query features, and those that do have a tendency to use them incorrectly [93]. There was also a tendency for users to avoid making subsequent modifications or refinements of their queries [89, 93].

A further user trait that was identified was the lack of willingness to evaluate many Web search results. Even when users were able to effectively craft a query, few Web searchers consider more than three pages of search results [89, 93]. As noted by Spink et al., "the public has a low tolerance for going in depth through what is retrieved" [93]. Users have a strong desire to see relevant documents within the top few search results to avoid scrolling or navigating to the next page [52].

Yang summarized a collection of Web search user behaviour studies succinctly: "Web searchers in general do not want to engage in an involved

retrieval process. Instead, they expect immediate answers while expending minimum effort" [117]. What is missing from these studies, however, is an explanation for why Web searchers do not wish to engage in the information retrieval process.

An important aspect of an information system is the ability to provide information seeking functions which assist users in defining and articulating their problems, and finding solutions to these problems [60]. It is clear that the current Web search engines provide little of this assistance to the users. It is up to the user to choose a query that accurately represents their information need, with little support provided by the Web search engine beyond a box in which to type their query. When provided with the list of search results, it is up to the user to evaluate these document surrogates one-by-one, with little support for the manipulation or exploration in order to find the solutions to their information need.

This lack of support for the user's tasks highlights the need for further study on Web search interfaces and the support they provide for the searchers' fundamental tasks of crafting a query that accurately represents the information need, and evaluating the search results to find relevant documents. The goal is to shift the cognitive burden of searching the Web from the user to the system [117]. Information visualization, as noted in the previous Chapter, is an effective method for supporting and amplifying the cognition of users [14]. Kobayashi and Takeda predicted that "future systems will have better user interfaces to enable users to visually manipulate retrieved information" [52]. One of the goals of this research is to explore such interfaces for Web search.

In the remainder of this Chapter, the focus is on four aspects of Web information retrieval that are relevant to this research: query expansion, search results representations, clustering of search results, and meta-search techniques.

The Chapter concludes with the introduction of a new model for interactive Web information retrieval that focuses on interactive support for the user's tasks of query refinement and search results exploration, facilitated by subject knowledge dynamically generated based on the current query.

3.2 Query Expansion

Query expansion is the process of adding additional terms to a user's original query, with the goal of building a query that more accurately captures the user's information need. Query expansion is a well studied technique for improving information retrieval performance by improving the user's input, and represents a type of relevance feedback [105, 6]. These techniques can be classified along two dimensions: the work the user must perform in order to generate the query expansion, and the source of the information used to expand the query [18].

With respect to the work that needs to be performed by the user, there are three options: manual query expansion, automatic query expansion, and interactive query expansion. *Manual query expansion techniques* are those which require the user to do the work of evaluating, selecting, and adding new terms to their query (i.e., without any additional system support). *Automatic query expansion techniques* choose and add new terms to the user's query, without the involvement of the user other than to submit an initial query. *Interactive query expansion techniques* allow the user to interactively make choices which are then used to generate the query expansion.

With respect to the source of the information used for expanding the query, there are three options: using a subset of the initial search results, using a collection-dependent knowledge base, and using a collection-independent

Table 3.1: The classification of query expansion techniques according to the work performed by the user.

	Automatic Query Expansion	Interactive Query Expansion
Harman, 1988 [30]		$\checkmark$
Salton & Buckley, 1990 [86]		$\checkmark$
Harman, 1992 [31]	$\checkmark$	
Qiu & Frei, 1993 [77]	$\checkmark$	
Voorhees, 1994 [108]	$\checkmark$	
Chang & Hsu, 1999 [15]		$\checkmark$
Mandala et al., 1999 [58]	$\checkmark$	
Xu & Croft, 2000 [114]	$\checkmark$	
Billerbeck et al., 2003 [11]	$\checkmark$	

knowledge base. Query expansion techniques that are *based on an initial set of search results* evaluate a number of documents within the search results set from the initial query in order to obtain potential terms for adding to the query. Techniques *based on collection-dependent knowledge bases* generate the query expansion by matching the original query to a knowledge base that was created based on some or all of the collection being queried. Techniques *based on collection-independent knowledge bases* match the original query to a knowledge base that was created using a source of information that is independent of the collection being queried.

Query expansion techniques, both automatic and interactive, and based on the search results, a collection-dependent or collection-independent knowledge base, have been well studied [114, 77, 108, 86, 15, 30]. Tables 3.1 and 3.2 provide a summary of the classifications of these works.

Since the focus of the query expansion work in this dissertation is on the interactive nature of query refinement, the three interactive techniques noted in Table 3.1 warrant further discussion. The primary difference between these

Table 3.2: The classification of query expansion techniques according to the source of the information used for the query expansion.

	Initial Set of Search Results	Collection-Dependent Knowledge Base	Collection-Independent Knowledge Base
Harman, 1988 [30]	$\checkmark$	$\checkmark$	$\checkmark$
Salton & Buckley, 1990 [86]	$\checkmark$		
Harman, 1992 [31]	$\checkmark$		
Qiu & Frei, 1993 [77]		$\checkmark$	
Voorhees, 1994 [108]			$\checkmark$
Chang & Hsu, 1999 [15]	$\checkmark$		
Mandala et al., 1999 [58]		$\checkmark$	$\checkmark$
Xu & Croft, 2000 [114]	$\checkmark$		
Billerbeck et al., 2003 [11]		$\checkmark$	

techniques are the source of the query expansion information and the level of user involvement required in order to generate an expanded query.

In the work by Salton & Buckley [86], the initial query is sent to the information retrieval system, and the results are presented to the user. The users can then indicate both relevant and non-relevant documents in this result set. A new query is generated, using these relevant and non-relevant documents as the source of the query expansion, wherein the terms present in the relevant documents are given a higher importance in the new query and the terms in the non-relevant documents are given a lower importance in the new query. This technique is commonly referred to as relevance feedback, and was first proposed by Rocchio [82]. While the results of this technique have been positive, the user must spend time evaluating a number of documents in order to improve the results of their query.

Chang & Hsu [15] addressed the problem of having to tag individual documents as relevant or not relevant by clustering the search results from the initial query and allowing the user to mark clusters of documents as relevant

and non-relevant. While their technique does provide a savings in the amount of time required by the user to make relevance selections, the quality of the clusters depends primarily on the quality of results of the initial query. When the quality of the initial results is low, the clusters may contain both relevant and non-relevant documents, reducing the effectiveness of this technique.

Harman [30] used three different sources of information for new terms to add to the query, and provided lists of terms from which the user could choose. These term lists were generated via local analysis (relevance feedback), query analysis (variations of the query terms), and global analysis (term co-occurrence in the collection). While the results reported from this technique were good when users made the perfect choices from the lists of available terms, there is little evidence to show that users are able to make these perfect choices. On the contrary, studies have shown that users are often unable to make good query expansion choices when provided with a simple list of terms [85, 57].

Other techniques that use manually constructed thesauri, such as WordNet [61], have not proven to be very successful. This can be attributed to the nature of general thesauri, wherein terms are grouped by their parts of speech and relationships only exist within a given part of speech group. This restricts query expansion techniques from adding a valuable new term because it is in a different part of speech [59].

It is interesting to note that without any support for query expansion or refinement, Web searchers may refine their queries manually, based on the sets of previous search results they have seen. As described by Stojanovic [96], and supported by a number of studies of Web search behaviour [93, 89], users commonly submit short queries to a search engine in order to increase the chance of obtaining some relevant documents. They then review these documents and

make refinements to their queries based on what they have learned from their previous queries. The relevance feedback techniques described above [86, 15] provide computer support that mimics this manual process.

3.3   Search Results Representations

The most common method for representing Web search results is through a list-based representation. All of the four primary Web search engines (Google[1], Yahoo[2], MSN[3], and AOL[4]) use variations on this representation method. Although the supplemental information provided may differ slightly, these search engines provide a ranked list of ten document surrogates per page. Each document surrogate primarily consists of the title of the document, a snippet showing the context of the query terms, and the URL of the document.

Since the documents themselves are primarily textual, this textual representation of Web search results is a logical method for displaying this information. However, since these lists are static, they provide little ability to manipulate the search results; they only lend themselves to a sequential evaluation of ten document surrogates at a time. This textual interface style requires the users to read the information (which can be time consuming), or scan the information (which can be error prone). Further, since text can consume a large amount of space in an interface, the list based representations commonly limit the number of document surrogates displayed per page.

Although graphical representations of search results have been explored within the traditional information retrieval research [60, 33] (see Section 2.3),

---

[1]http://www.google.com/
[2]http://www.yahoo.com/
[3]http://www.msn.com/
[4]http://www.aol.com/

few of these techniques have been explored within the Web search setting. One of the challenges in providing a visual representation of Web search results is gaining access to enough information to generate a useful visual representation of the search results. Commonly, in traditional information retrieval systems, the visualization system has access to the entire textual contents of each document. However, for Web search, and especially for Web search interfaces that are meta-search systems, the contents of the entire documents in the search results set are not provided; the data provided is normally limited to the document surrogate (title, snippet, and URL).

If a visualization technique requires access to the entire contents of the documents in the search results set, these documents must be retrieved from the Web. This has the result of introducing significant delays in generating the visual representation, which will not be well received by Web searchers [28]. As such, many of the visualization approaches discussed in Chapter 2 are not directly applicable to Web information retrieval.

Clustering techniques have also been investigated as a means for supporting the users in their search results evaluation tasks. Commonly, these systems add additional interface elements to support the users in navigating the clusters. However, they tend to continue to employ the simple list-based representations that are common among the primary Web search engines. More details on clustering techniques are discussed later in this Chapter.

Another interesting avenue for supporting users in evaluating the results of their Web searchers is through re-sorting or re-ranking of the top Web search results provided by some underlying Web search engine. The re-sorting of search results based on Web search personalization is a rather active research field [100, 101, 78]. These systems generally provide an automated re-sorting and filtering of the search results based on the personalized profiles of the

users. Generally, these personalized profiles are created through Web mining techniques that take advantage of the vast amounts of Web search user data [54].

Other work makes use of the Web page categories provided by the Open Directory Project to re-rank Web search results [16]. The end result is similar to the personalization techniques: the search results are provided in a re-ranked order based on other knowledge present in the Web information retrieval system. However, these re-sorting systems continue to provide the search results in static list-based representation. There appears to be little research on interactive tools to allow the users to control the re-sorting methods, in personalized systems or otherwise.

3.4 Clustering Web Search Results

Clustering can be defined as the unsupervised classification of data objects into groups of similar objects (called clusters) [47]. Clustering has been explored for a number of years both for browsing text collections [17, 34] and for organizing Web search results [120, 121]. Hearst and Pedersen validated the cluster hypothesis, showing that relevant documents tended to be more similar to each other than non-relevant documents [34].

Recently, a number of publicly available Web search systems have been developed that provide clusters of Web search results to the users, and allow the users to browse the clusters to narrow down the set of search results. Many of these systems use hierarchical clustering algorithms, and primarily differ in the representation and interaction with the clusters. Two such systems are Vivisimo [107] and Grokker [29] (see Figures 3.1 and 3.2).

The hierarchical clustering algorithms commonly used by these systems
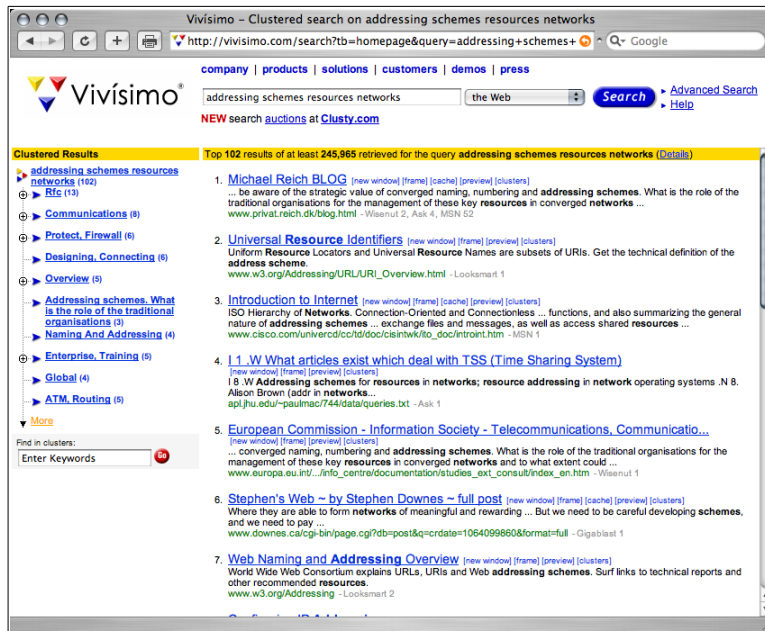
Figure 3.1: A screenshot from Vivisimo [107].
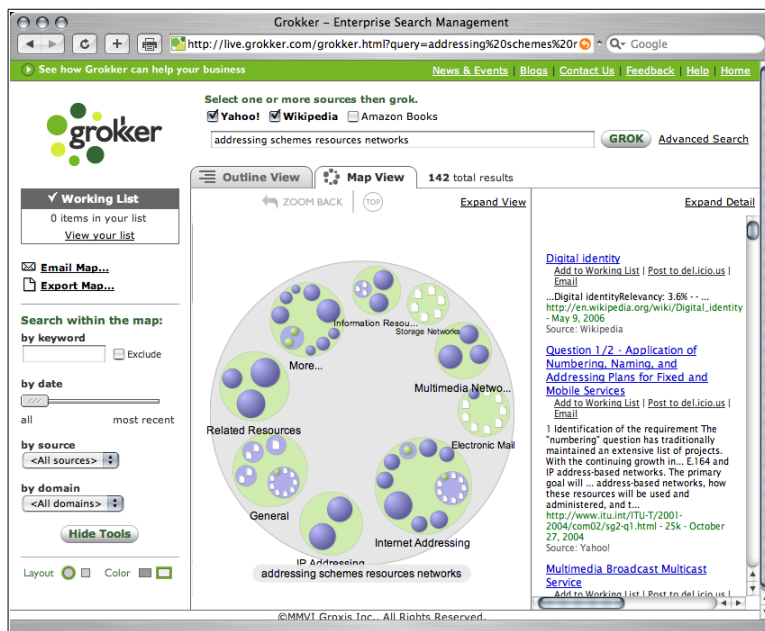
Figure 3.2: A screenshot from Grokker [29].

partition the search results at various levels of similarity [47]. This results is a tree-like structure representing the clusters, where parent clusters contain all the objects of their children clusters.

In Vivisimo, these hierarchical clusters are represented as a tree. The nodes in the tree can be expanded and collapsed in a manner similar to file directory navigation. When a tree node is selected, the document surrogates contained within that cluster are displayed in a separate frame.

In addition to providing a tree-like navigation scheme, Grokker uses a visual representation of the hierarchical cluster structure. This visual representation uses nested circles to represent the clusters and their children, and provides the ability for users to see the sizes of clusters and whether they contain additional children or document surrogates. Like Vivisimo, when a cluster is selected, the document surrogates that are contained within that cluster are displayed in a separate frame.

One of the challenges in any clustering system is to provide meaningful names for the clusters. Commonly, the names are generated by choosing the most frequent terms or phrases within the cluster (ignoring very common terms such as "is", "and", "the", etc.) [121]. The ability to choose meaningful descriptions of clusters in a Web search system has a direct impact on the ability for users to correctly navigate the clusters to find relevant document surrogates. If vague or misleading names are chosen for the clusters, this can lead to users choosing clusters that are not relevant to their information needs (resulting in the evaluation of documents that are likely not relevant), or not choosing clusters that are relevant to their information needs (resulting in missing documents that are relevant).

Commonly, the documents that are relevant to a user's information need will be distributed among multiple clusters. However, most clustering systems

do not easily support the simultaneous exploration of multiple clusters. While it is possible to view an intermediate cluster that contains all the document surrogates of its children clusters, viewing the union of an arbitrary set of clusters is generally not possible. This means that if users wish to explore multiple clusters, they must do so separately.

A further difficulty with these Web search clustering techniques is that they do not provide any additional information regarding the organization of the document surrogates within the clusters. When a cluster is selected, the documents are listed in the same order as provided by the underlying search engine. There is no indication of which documents are most similar to the cluster centroid, or the degree of membership to the cluster.

Recent research in Web search has addressed some of these shortcomings. Kules et al. categorized Web search results into meaningful and consistently named categories [55]. While this work shows promise in addressing the naming problem, the interface does not support the simultaneous selection of multiple categories, nor does it re-order the search results based on category selection.

3.5   Meta-Search Techniques

Since the focus of this research is on providing visual and interactive support for user's Web search tasks, it does not introduce new information retrieval techniques or methods. Instead, meta-search techniques are used in order to make use of the Web search services provided by other systems.

In the prototype systems developed in this research, the Google API [27] was used exclusively. However, other Web search engines also provide access to their underlying data, such as Yahoo [115] and Alexa [2]. The data provided

by these other Web search engines is similar. Although it is beyond the scope
of this research, it would be a trivial exercise to modify these prototypes to
support the data provided by other search engines.

In order to obtain a set of search results from the Google API, a well-
formed request must be sent to the Google server, which includes an access
code, the query, and the start point for the search results. The access code
is used by Google to limit the number of request that can be made by a user
per day (since the Google API is still considered a beta release). The query
is a simple text string containing a query in any format normally accepted by
Google. The start point enables the retrieval of search results that are deep in
the search results list. Other settings are also present; the reader is referred
to the documentation on the Google API for more details [27].

As the result of a request, a search results set is provided by the search
engine. This contains some data about the search results set itself, along
with ten document surrogates. Normally, this will be the first ten document
surrogates for the provided query. However, when the start point parameter is
set, a set of documents deeper in the search results set can be obtained (e.g.,
document surrogates 11 - 20).

3.5.1   Attributes of the Document Surrogate

For each document surrogate within the search results set, a number of at-
tributes are provided by the Google API. These include the title of the docu-
ment, a snippet from the document (usually showing the query terms used in
context from the original document), the URL of the document, and the Open
Directory Project [62] directory to which the document belongs (if any).

A number of attributes can be calculated based on the information within
each document surrogate. Based on the URL data, the type of the file can be

deduced from the name of the file. Using the textual data from the title and snippet, a query term vector can be calculated that represents the frequency of the query terms found within this data (this technique is used in HotMap). Higher dimensional vectors can also be created based on the occurrence of each unique term within this data (these vectors are used in both Concept Highlighter and WordBars).

3.5.2 Attributes of the Search Results Set

With each set of search results provided by the underlying search engine, there is one core attribute that is automatically provided by the search engine: the number of documents in the search results set. Although this data can be useful in determining the generality or specificity of the query, it provides little additional information to the user.

As with the individual document surrogates, other attributes of the search results set can be calculated by analyzing the contents of the top document surrogates in search results set (the top 100 document surrogates, for example). By analyzing the URL, the file type can be determined, resulting in a count for each unique file type. Alternately, the textual contents of title and snippet for all the document surrogates can be analyzed in order to determine the number of documents that use each of the query terms, as well as all possible combinations of the query terms. This data can also be used to calculate a frequency vector based on the query terms, or even a frequency vector based on all unique terms in the entire set of the top search results. Note that this last calculated attribute is a technique that is used in the WordBars system.

3.6 A New Interactive Model for Web Information Retrieval

Most Web search engine interfaces support a model of interaction based on traditional information retrieval, consisting of a cycle of query formulation, examining the search results, and either stopping or reformulating the query [33]. The problem with this model is that there is little support for the users to perform the two fundamental Web search tasks: query formulation and search results evaluation. Evidence of this inadequate support is provided in the studies of user Web search behaviour which showed that searchers commonly use very short queries and seldom view many of the search results [89, 48, 93]. Further, Web users have reported that one of the most common problems they experience when using the Web is finding information [28].

Common sense would dictate that the manual query reformulation process that is supported by current Web search engines would lead to a higher rate of success, even though the work must be done by the user. However, studies with e-commerce systems have indicated that the opposite is true: the success rate diminishes as users reformulate their queries [66]. While one can assume that during the search results evaluation process, the users were able to identify features of the documents that were not relevant to their information need, little support is provided to assist the users in employing this information in order to construct a better query. Therefore, techniques that provide support for this task may be of great benefit to Web searchers.

Even when users are able to effectively craft a query that describes their information needs, users are required to consider each document surrogate one-by-one, and to some degree, in the order provided. There is no overview of the search results, and no ability to manipulate or explore the search results.

An extension of the traditional Web information retrieval model that uses information visualization and interactive visual manipulation to support query formulation and search results exploration is proposed. Since the ability to read and assess textual information is a limiting factor for information retrieval systems [113], visual representations of subject knowledge derived from the users' queries, and the compact and effective visual representation of search results can aid the users to effectively interpreting and making sense of the information provided. A high degree of interactivity allows the users to take a more active role in the Web information retrieval process, rather than the passive role supported by the traditional model.

3.6.1 Traditional Model

The traditional model for Web information retrieval consists of two primary stages: query formulation and search results evaluation. At some point in the search results evaluation stage, the users make a decision as to whether they are satisfied with the search results or not. If they are satisfied, the Web search is considered a success. Otherwise, the users must decide whether they wish to reformulate their query. If they choose to not reformulate their query, the Web search is considered a failure.

The two primary drawbacks of the traditional model are that there is little support for the users in formulating queries, and little support for users in evaluating the search results. Applying the principles of information visualization and interactive visual manipulation in order to support the users as they perform these complex tasks results in a model that allows the users to take an active role in the Web information retrieval process. As illustrated in Figure 3.3, this model adds two new cycles in addition to the traditional query reformulation cycle: interactive query refinement and interactive search
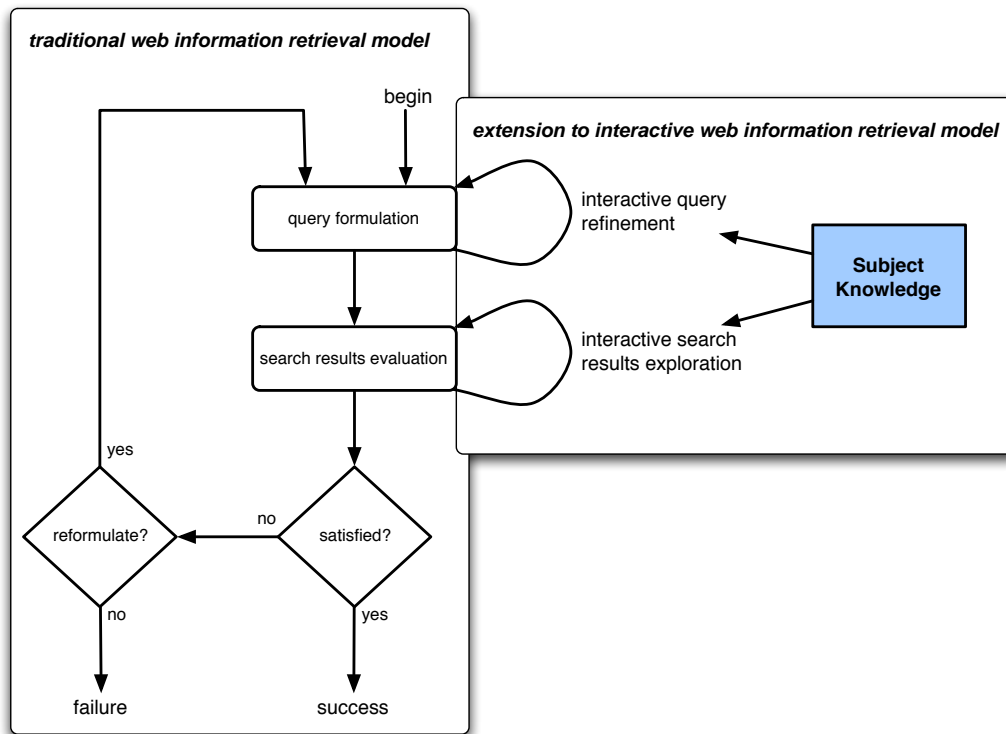
Figure 3.3: A model for interactive Web information retrieval.

results exploration. A fundamental difference from the traditional model is
the subject knowledge component which is dynamically generated based on
the users' current query. This component plays a pivotal role in both interactive cycles. A summary of the design principles in supporting this model are
provided in Table 3.3 and are elaborated upon in the following Sections.

3.6.2 Interactive Query Refinement

Within the traditional Web information retrieval model, the query reformulation cycle can not occur until the users have first formulated an initial query
and viewed the results of this search. If they make poor choices for their query
terms that result in many non-relevant documents, there is little information

Table 3.3: Design principles for interactive query refinement and interactive search results exploration.

Interactive Query Refinement
Provide subject knowledge.
Focus on query formulation.
Provide visual representations of the query.
Support a preview of the search results.
Promote exploration and experimentation.
Interactive Search Results Exploration
Support the viewing of many document surrogates.
Provide the search results in multiple levels of detail.
Provide a visual representation of the document surrogates.
Exploration decisions lead to re-ranking and re-sorting.
Indicate which documents have been viewed.

provided to allow the user improve their query. Further, if they are unwilling to view many documents (which is often the case), there is little ability to glean additional information to support the query reformulation.

The interactive query refinement loop must provide specific subject knowledge to the users that may be relevant as they craft a query. One of the challenges of query refinement is finding a source for this supplemental information that is relevant to the users' information needs. External, topic-specific knowledge bases may be necessary to resolve this challenge. There is also information present in the top search results that may be of use (assuming that an adequate initial query is provided).

Any tool that provides support for interactive query refinement should allow the user to focus on their query formulation task. It is desirable that the information provided be in the context of the query, wherein the query is the primary focus rather than the search results set. This is in contrast to the traditional model, where the query reformulation cycle promotes linear steps between query formulation, search results evaluation, and query reformulation.

Information visualization techniques can promote an easier interpretation and synthesis of the information provided regarding the users' queries. Further, the visual representations of complex data can result in much quicker comprehension than the equivalent textual representations. As such, visual representations of the query information for an interactive query refinement tool is highly recommended.

As the users formulate their queries within this interactive query refinement loop, it is important to allow them to see how they are progressing. As such, a preview of the search results should be provided. This will help to promote an understanding of whether the users' queries are an accurate reflection of their information needs.

The interactive query refinement loop should promote exploration and experimentation with the query. As the users make refinement decisions, the results of these decisions should instantly be reflected in the visual representation of the query as well as in the preview of the search results. If minimal effort is required to return the users' queries to previous states, they will be able to easily experiment with "what if" scenarios, and explore the query space.

3.6.3   Interactive Search Results Exploration

In the traditional Web information retrieval model, the search results evaluation exists as a static stage in which the users evaluate the document surrogates one-by-one, and in the order provided by the Web search engine. While this linear evaluation may be effective when there are a large portion of relevant documents in the first few pages of the search results, it is not very effective when the search results contain a mixture of relevant and non-relevant documents. It is in these cases that an interactive exploration of the search results

is beneficial.

In order to promote an effective exploration of the Web search results, many document surrogates must be retrieved and made available to the users. Providing a large number of document surrogates can allow the user to discover interesting features or patterns in the search results, and can direct the users to sets of documents that are more likely to be relevant.

Since a much larger number of document surrogates will be present in an interactive search results exploration tool, the user may wish to view both the details of specific document surrogates and the features of the entire search results collection simultaneously. Providing coordinated views at different levels of detail can allow the users to take both a micro- and a macro-view of the Web search results.

Just as information visualization techniques can allow users to more effectively make sense of their queries, generating visual representations of the document surrogates can allow the users to more effectively understand the features of the search results. Although converting the textual features of document surrogates into visual representations can be challenging, the benefits include the ability to provide a compact representation, and allow the users to "see" the information rather than "read" it.

As the users explore the Web search results, their decisions should result in a re-ranking and subsequent re-sorting of the search results. This has the effect of bringing relevant documents to a more prominent location in the search results based on the features of the search results space. Further, the exploration decisions the users make must be reflected instantly in the visual representation of the search results. This will allow the users to more easily determine the outcome of their choices, and will allow them to decide whether their decisions are bringing them closer to fulfilling their information needs.

In any system that represents Web search results, be it a static list or an exploratory system, it should be very easy for the user to access the specific documents. Further, as the users select documents to view, it should be made clear that these documents have been viewed when the user returns to the search results. This is especially important for a system that provides an abstract visual representation of the document surrogates, and allows the users to re-rank the search results.

Through the exploration of the search results, the users are directed to the documents that are more relevant, and away from documents that are less relevant. Ultimately, this is a transition from the users making decisions on a document-by-document basis, to making exploration decisions which allow individual document surrogates to be evaluated more easily. As a result, the work performed in the exploration of the search results is transferrable to the task of document surrogate evaluation.

3.6.4   Subject Knowledge

In this model, the cycles of interactive query refinement and interactive search results exploration are supported by subject knowledge. This subject knowledge is dynamically generated based on the users' current query, and provides a means for facilitating visualization and interactivity within these processes. This subject knowledge may be based on concept information derived from the current query, or it may be generated based on the top set of search results returned for the current query. Many other options for obtaining the subject knowledge exist; the primary feature is that the subject knowledge must provide supplemental information to the user to support the interactive query refinement and interactive search results exploration processes.

Chapter 4

Interactive Query Refinement

4.1 Motivation

The primary motivation for exploring visual and interactive methods to support interactive query refinement is to improve upon the current methods Web search engines provide for supporting users in crafting queries that accurately capture their information needs. In particular, current Web search engines require users to manually craft and edit their queries based on information they have gleaned from the search results list and the documents they have viewed. Interactive query refinement tools can assist users in making query refinement decisions, reducing the cognitive burden required by the users to craft queries, and increasing the quality of the queries.

The fundamental hypotheses in this research on supporting users in their interactive query refinement processes is that providing computer support through supplemental subject knowledge can allow the users to make effective choices for refining their queries. One of the challenges, however, is representing information that supports these interactive query refinement processes in a manner that allows the users to make effective choices. A further challenge

is providing interactive support that allows the users to explore the potential benefits of their decisions with respect to their primary Web search goals.

In early work on interactive query expansion, Harman [30] provided three lists of terms from which users could select to add to their queries. Although it was shown that making good selections from these lists could improve the effectiveness of the information retrieval system, it is not clear that users would be able to make such selections. Other studies have shown that users are often unable to make good query expansion choices when provided with a list of terms [85, 57]. Therefore, in this research, simple list-based representations are avoided. Instead, visual and interactive methods for presenting this information to the user are explored.

In this Chapter, two systems that support interactive query refinement are described in detail: VisiQ and WordBars. In the VisiQ system, a graphical representation of the query space is provided for user-supplied queries. Users can easily experiment and explore within this query space by selecting additional terms to add to the query, or remove terms from the query. With each change, a new query space is dynamically generated along with its corresponding graphical representation, and a preview of the top search results is provided.

In the WordBars system, all the unique terms in the top search results are counted, and a vertically-oriented term frequency histogram is provided along side the search results. Users can easily identify terms that may be relevant to their information need. They can experiment with re-sorting the search results based on these candidate terms, determining whether they are indeed relevant. These terms can readily be added to the query; terms currently in the query that are deemed to no longer be appropriate can easily be removed. The newly refined query can then be sent to the search engine. Along with

the new set of search results, a new term frequency histogram is generated to support further refinement of the query.

4.2   VisiQ

The primary goal in the development of VisiQ was to explore the use of visualization and interaction to support the users' in crafting queries that accurately reflect their information needs. However, generating a visual representation of a query is a challenging task, due to the textual nature of Web search queries. What is needed is additional information regarding the relationships between query terms and other terms that could be of potential benefit to the users as they refine their queries.

In this work, a domain-specific concept knowledge base was developed to store relationships between terms and the high-level concepts that the terms describe. A concept knowledge base was constructed using the ACM Computing Classification System [1], which provides support for expanding queries within the computer science domain.

By visually representing the query space generated by the initial query, VisiQ allows users to take advantage of their visual information processing capabilities and intelligence to determine the relationships between their query terms and the concepts they describe. Users may interactively refine their queries by choosing new terms that are also related to these concepts, or removing terms that do not adequately describe their information needs. The query space is interactively re-calculated based on these selections, and Google is accessed to provide users with a preview of the search results.

Although others have attempted to address the problem of users properly specifying their information needs by supporting natural language queries, one

of the challenges with natural language queries is properly transforming them into an internal representation that supports effective information retrieval. Some have attempted to work around the natural language problem by having the users provide their query specifications in a simple form that corresponds to the internal query representation [63]. The work on VisiQ has instead focused on building simple queries that consist of lists of terms. Although these queries may not be as effective at addressing fact-based questions, they are easier for the users to understand, and correspond to the types of queries that the users commonly provide [93].

One of the principles that has guided the design of VisiQ has been to maintain a balance between computer automation and human control [88]. There is a critical human role in the crafting of a query that accurately represents a user's information need. To this end, a prototype system has been developed that takes advantage of the computer's ability to store, retrieve, and process large amounts of data, and the human's ability for visual information processing, complex decision making, and judgement. One of the fundamental features is the method by which the human and the computer exchange information and initiate tasks. Information visualization and interaction techniques promote this exchange, allowing the users to interpret and make sense of the query space and easily refine their query to more accurately reflect their information needs.

In the remainder of this Section, the details on VisiQ are provided, along with examples illustrating its utility in supporting the users through interactive query refinement.

4.2.1   Concept Knowledge Base

The concept knowledge base used in this work draws its inspiration from the conceptual semantic theories, along with formal concept analysis. This is an extension of earlier work that used a similar technique for generating automatic query expansions based on concepts [43]. In this Section, background to these theories are provided, followed by a formal definition of the concept knowledge base, and a description of the method for automatically generating a concept knowledge base from well-defined concept information.

Semantic Theories

Semantics, the study of the meanings of words in a language, provides many theories that attempt to explain the meaning of a word or phrase. Of interest for query expansion are both the structuralist and conceptual theories of meaning [24].

The *structuralist theories of meaning* hold that in order to understand the meaning of a word, one has to understand how it functions together with, and in contrast to, other related words. The meaning of a word consists of the relationships it has with other words in the language (i.e., synonyms, homonyms, polysems, antonyms, etc.). In order for two people to communicate, they must have a similar understanding of how the words they use are related to other words in the language [24].

In the *conceptual theories of meaning*, the meaning of a word is the concept in the mind of the person using that word; communication is possible via a shared connection between words and concepts. That is, there is a mapping between the set of words in the vocabulary and the set of concepts in the minds of the users. While there will not be a single unique mapping that all users

of a language share, the individual mappings must be similar to one another over the vocabulary used in order for communication to occur [24].

The structuralist and conceptual theories of meaning are similar in the sense that they both infer the meanings of words used by an individual with respect to other objects. Therefore, the semantics of a word is not provided by a definition, a reference, or its use, as one is normally accustomed to when using a dictionary; rather the meaning of a word is provided by its relationship to other objects. The distinction between these theories is in the type of objects to which words are related (i.e., other words in the structuralist theories, and higher-level concepts in the conceptual theories). Using the conceptual semantic theory, one can infer a relationship between a pair of words if there is a concept which they have in common. The generality or specificity of this common concept provides a clue to the degree to which the terms are related; terms linked through a high-level (general) concept may have a weak relationship, whereas terms linked through a low-level (specific) concept may have a strong relationship.

Much of the previous work on query expansion, including those that use collection-dependent knowledge bases [30, 77, 58, 11] and those that use collection-independent knowledge bases [30, 108, 58], follow the structuralist theories of meaning (see Section 3.2). Queries are expanded to include additional terms that are related to the original terms in the thesauri. The success of these techniques varies depending on whether the query expansion is interactive or automatic, as well as the specificity of the thesaurus used.

The concept knowledge base used by VisiQ follows the conceptual theories of meaning. It also draws some features from formal concept analysis, which is described in the following Section.

Formal Concept Analysis

Formal concept analysis provides mathematical descriptions of concepts in terms of *extensions* and *intensions* [22]. A formal context is required in order to construct descriptions of concepts. A formal context is given by a triple (G, M, I) consisting of the set of objects G, the set of attributes M, and a binary relation I between G and M. If an object $g \in G$ has attribute $m \in M$, then $(g, m) \in I$.

A formal context can be used to derive a concept lattice. For a set of objects $A \subseteq G$, a set of attributes $A' = \{m \in M \mid \forall g \in A, (g, m) \in I\}$ can be defined. Similarly, for a set of attributes $B \subseteq M$, a set of objects $B' = \{g \in G \mid \forall m \in B, (g, m) \in I\}$ can be defined.

A formal concept in a formal context (G, M, I) is defined as a pair (A, B) where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. That is, given a set of objects A and a set of attributes B, if all the attributes that can be derived from the objects A are present in the set of attributes B, and if all the objects that can be derived from the attributes B are present in the set of objects A, then these sets A and B form a formal concept. In this case, the set of objects A represents the *extension* of the concept, and the set of attributes B represents the *intension* of the concept.

In terms of Web information retrieval, G is the set of all documents on the Web, and M is the set of all terms used in these documents. The relation I represents whether a document $g \in G$ contains term $m \in M$. Discovering a formal concept within this set of documents requires analysing subsets of documents and subsets of keywords, seeking matches between the intensions and extensions of possible concepts.

It is clear that for a set of documents as large as the Web, finding formal concepts using formal concept analysis is not feasible. Suppose, however, that

one is provided with the set of documents that are used to describe a concept, such as the set of documents that belong to a particular node in a concept hierarchy (eg. the Open Directory Project [62] or the ACM Computing Classification System [1]). While these documents may only be a subset of the extension of the concept represented by the node, they can be analysed to find a possible intension of the concept (i.e., the terms common among these documents). While this intension may not exactly match the actual intension of the concept when all possible documents on the concept are considered, if the set of documents is a good representation of the extension of the concept, the set of common terms will be a good representation of the intension of the concept.

One of the difficulties of using formal concept analysis in the domain of information retrieval is that the vocabulary or word mismatch problem [21] (in particular, the synonymy problem) may result in two terms that have a similar meaning being excluded from the intension of a formal concept. That is, because two such terms may be used interchangeably, each term may not be present in all documents that represent the extension of the concept, which precludes them from being present in the intension of the concept.

The formal concept analysis theory influenced the development of the concept knowledge base in this research. The concept knowledge base extends the formal concept analysis work by including a weight factor for each term in the intension of a concept, permitting the inclusion of intensions of a concept which are not present in all the extensions of a concept. Even though two terms are present in the same document, which is part of the extension of the concept, this does not mean that these two terms will necessarily end up being connected to this concept in the final concept knowledge base. As will be shown shortly, the weight factor is calculated as the average term frequency

in the documents that belong to the extension of the concept, normalized over the entire set of concepts. Thus, terms that are used in the extension of many of the concepts will have a much lower weight than terms that are used in the extensions of few concepts. These extremely low weight edges in the concept knowledge base can effectively be removed, since they do not represent a connection between a term and a concept that is of value.

Concept Knowledge Base

The source of information for generating the query space is the concept knowledge base, the design of which drew inspiration from the conceptual semantic theories [24] and formal concept analysis [22]. A concept knowledge base is a bipartite graph consisting of two sets of nodes: concepts and terms. In this graph, edges are assigned a weight and can only exist between concept nodes and term nodes. The result is a data structure capable of storing mappings between concepts and the intensions of the concepts (i.e., between concepts and the terms that have been used the describe these concepts).

More formally, a concept knowledge base $CKB = \{C, T, E\}$ consists of a set of concept nodes C, a set of term nodes T and a set of edges $E = \{c_i, t_j, w_{ij}\}$, where $c_i \in C$, $t_j \in T$, and c_i and t_j are related with a weight w_{ij}. The weight w_{ij} of an edge in the concept knowledge base represents the degree to which the term represents the intension of the concept.

In the initial work on the concept knowledge base [43], a method for automatically constructing such a knowledge base using a concept hierarchy such as the Open Directory Project [62] was described. For the work on VisiQ, a more specific and focused source of concept information was chosen: the ACM Computing Classification System [1]. This classification is used to categorize papers submitted to the ACM Press, and contains both a concept hierarchy

representing the various disciplines within the computer science domain, and short descriptions of these disciplines. In generating the concept knowledge base, the nodes in this concept hierarchy become the concepts (i.e., disciplines in the computer science domain), and the descriptions of these nodes are used as the source of terms.

Porter's stemming algorithm [74] is used to pre-compute the stems of all terms contained in the concept knowledge base. This algorithm automatically removes all prefixes and suffixes from words in the English language, resulting in the roots or stems of the words. This allows terms to be matched based on the stems of the words, rather than exact word matches. Therefore similar terms that share a common stem (e.g., cluster and clustering) will be treated as the same term in the concept knowledge base.

A bag-of-words approach is used to count the occurrences of term stems within each concept description. To calculate the edge weight values w_{ij}, the set of descriptions $D_i = \{d_{i1}, \ldots, d_{in}\}$ which are associated with a concept $c_i \in C$ are each evaluated. For each description d_{ik}, the set of term stems used in this description is $T_{ik} = \{t_{1,ik}, \ldots, t_{m,ik}\}$. The function $f(d_{ik}, t_j)$ is defined as the occurrence count of term stem t_j in description d_{ik}. The value for the edge weight between concept c_i and term stem t_j is given by:

$$w_{ij} = \frac{\sum_{k=1}^{n} \frac{f(d_{ik}, t_j)}{\sum_{l=1}^{m} f(d_{ik}, t_{l,ik})}}{n}$$

After all the concepts have been analysed, the edge weights from each term stem in the concept network are normalized. For a term stem t_i that is connected to r concepts whose index is given by the relation $f(x)$, $x = 1 \ldots r$,

the normalization is performed via the calculation:

$$w_{ij}^* = \frac{w_{ij}}{\sum_{k=1}^{r} w_{if(k)}}$$

In effect, the weights in the concept knowledge base represent the average term stem frequency, which is then normalized with respect to the occurrences of the term stem in other concepts. Term stems that have a high frequency of occurrence in the descriptions of a concept will result in a high edge weight between these nodes in the concept knowledge base. Term stems that have a low frequency of occurrence in the descriptions of a document will result in a low edge weight. Due to the normalization, term stems that are used frequently to describe many different concepts will also have a low edge weight to these concepts.

Using the normalized average term stem frequency rather than a simple total of the term stem occurrences reduces the impact of different description sizes and different numbers of descriptions (or documents) that represent a concept in the calculation of the weight values. In particular, without this normalization step, a single large description could provide term stem weights that overshadow the weights provided by a number of smaller descriptions; a similar situation is avoided for concepts that have a large number of descriptions in relation to the number of descriptions present for other concepts. Further, without the normalization, common term stems that are included in many of the descriptions for a concept would have a very high weight value, even though these term stems are often of little value in describing the concept.

This technique for constructing the concept knowledge base is somewhat

similar to the TF-IDF (term frequency - inverse document frequency) weighting scheme commonly used in traditional information retrieval [105, 6]. However, instead of re-weighting the term frequency by the inverse document frequency, the edge weights begin with the average term frequency, and are then normalized with respect to term usage in other concepts. Although the end result is a re-weighted term frequency, the technique used takes into account not only whether the term is used frequently, but also how many different concepts make reference to the term.

4.2.2   System Architecture and Features

The system architecture for VisiQ consists of three primary components. The *query space generation component* takes the query and the concept knowledge base as input, and produces the query space. The *query space visualization component* uses graph drawing algorithms along with techniques for colour coding and data representation to create a visual display of the query space. The *Google preview component* sends the current query to the Google search engine via the Google API [27] and displays summary information about the query results.

This system architecture is illustrated in Figure 4.1, the details of which are provided in the remainder of this Section. The user interaction steps shown in this diagram are described in Section 4.2.3.

Query Space Generation

In order to generate a visual representation of a user's query, the first step is to generate the query space. That is, the context or setting in which the query terms exist must be determined. For this purpose, the concept knowledge base
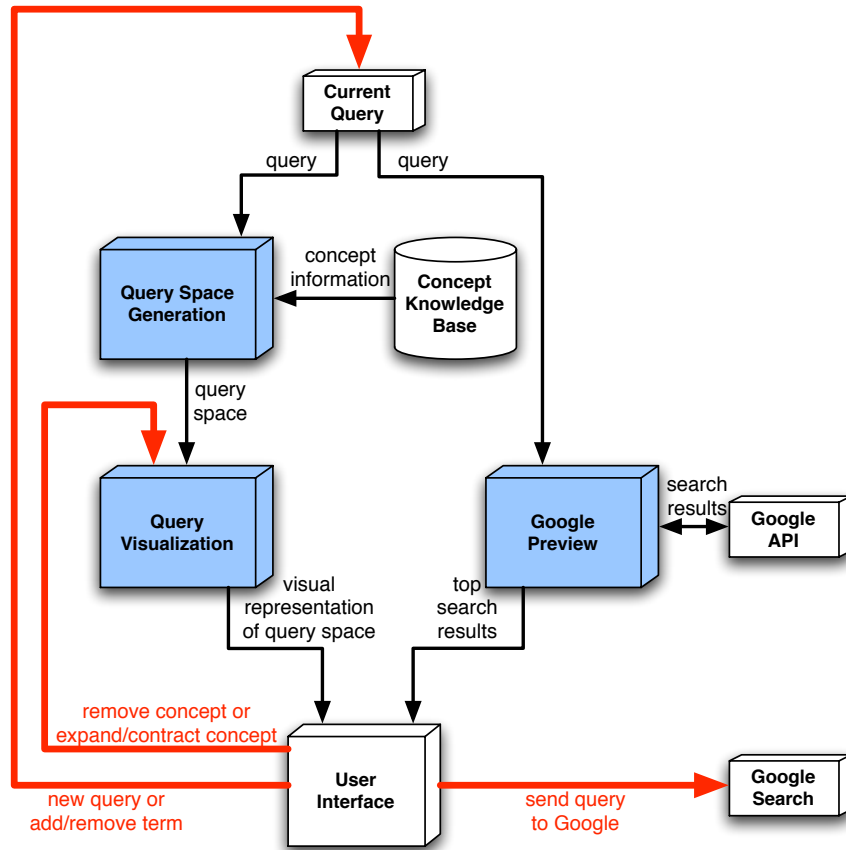
Figure 4.1: The VisiQ system architecture consists of three primary software components (coloured in blue): query space generation, query visualization, and Google preview. The interaction paths are illustrated by the thick red arrows.

which was described in the previous Section is employed.

Given a concept knowledge base $CKB = \{C, T, E\}$, and a query $Q = \{q_1, \ldots, q_n\}$ consisting of query terms q_i, the process of generating the query space $QS = \{C^*, T^*, E^*\}$ is given by the following five steps:

1. Match the query terms q_i to the term set T to obtain a $T' \subseteq T$ using the stems of the terms.

2. Obtain the set of concepts $C^* \subseteq C$ which are connected to the terms

in T'. Two parameters are used to control this operation: a weight threshold w_e, and a term ratio TR. First, all the concepts that are connected to the terms in T' with a weight greater than w_e are chosen as candidate concepts in C^*. Each of these concepts are then evaluated to determine the ratio of the terms in T' to which they are connected with a weight greater then w_e. If this ratio is less than TR, the candidate concept is dropped from C^*.

3. Obtain the set of terms $T'' \subseteq T$ which are connected to the concepts C^*. A weight threshold parameter w_d is used to control this operation. All terms that are connected to the concepts in C^* with a weight greater than w_d are chosen as the phrases in T''.

4. Perform a union of the original query terms and the new set of terms to obtain the query space terms: $T^* = Q \cup T''$.

5. Obtain the set of edges $E^* \subset E$ such that $e_{ij} \in E^*$ iff $c_i \in C^*$ and $t_j \in T^*$.

Borrowing terms from the field of morphological image processing [25], the generation of the query space can be considered as a two step process: *erosion* and *dilation*. When the process of erosion is followed by a process of dilation, the result is called *opening*. We use this terminology since the meanings in morphological image processing are analogous to the process of generating the query space.

During the erosion process (step 2 in the above algorithm), the original query is reduced to the set of concepts to which the terms are related. These concepts represent an abstraction of the original query, eliminating much of the specific detail of the query. The purpose of this step is to identify the core concepts to which the user's query is related.

In the dilation process (step 3 in the above algorithm), the core concepts identified for the query are expanded to include the terms to which they are closely related. These additional terms broaden the information present, and bridge the gaps between the concepts. The purpose of this step is to identify new terms that may be relevant to the user's query.

The opening process that results from the erosion and dilation processes being executed sequentially produces the query space. By explicitly including the original query terms as well as the edges between the concepts and the terms (steps 4 and 5 in the above algorithm), the result is a sub-graph of the concept knowledge base. Since the concept knowledge base is a bipartite graph, so is the query space.

The thresholds w_e and w_d, along with the term ratio TR, control the size of the query space generated for a given query. Through initial experiments with using the concept knowledge base generated from the ACM Computing Classification System for query expansion, these values were set at $w_e = 0.05$, $w_d = 0.10$, and $TR = 0.51$.

Adjusting these parameters will have an impact on the features of the query space. The w_e threshold parameter controls which concepts are selected from the concept knowledge base to be included in the query space. Increasing this value will cause only the concepts that are closely related to the query terms to be selected; decreasing this value will cause many concepts, even those that are loosely related to the query terms, to be included.

Similarly, the w_d threshold parameter controls which terms are selected from the concept knowledge base to be included in the query space. Large values will result in only the terms that are closely related to the selected concepts being included. Small values will allow many terms, including those that are only slightly related to the concepts, to be included in the query space.

The term ratio parameter TR becomes relevant when there are multiple terms in the query. One may assume that when a user enters multiple terms in a query, there is a relationship among these terms in regards to the user's information need. Therefore, it may not be desirable to select all the concepts that are related to any of the query terms; instead, selecting only the concepts that are related to multiple query terms may be of more value. Increasing the TR parameter will result in selecting fewer concepts that are simultaneously related to more of the query terms; decreasing the parameter will result in selecting more concepts that are simultaneously related to fewer of the query terms.

The current parameter settings represent a balance between very specific and directed searching, and exploratory searching. A more specific and focused query space can be generated by increasing all the values; a more general query space can be generated by decreasing all the values. The challenge here is that a very specific query space may result in few concepts and terms, and provide little information to the user upon which they can base their query refinement. Conversely, a very general query space may result in many concepts and terms, which would be difficult to depict visually, and difficult for the user to interpret and use to refine their query due to information overload.

Query Space Visualization

The key requirement in creating a visual representation of the bipartite subgraph that represents the query space is to ensure the users can properly perceive, interpret, and make sense of the underlying features of the query space. That is, users must be able to identify the three classes of nodes (concepts, unselected terms, and selected terms), along with the relationships between the concept nodes and the term nodes. To this end, it is necessary that the

graph drawing algorithm represent the distance between two nodes as a function of the edge weight, and ensure that the edges are always visually present to reinforce the relationship between connected concept-term pairs.

VisiQ uses the Java Universal Network/Graph Framework (JUNG) [72] as the basis for generating the visual representations of the query space. A heuristic-based graph drawing algorithm is used to satisfy two criteria: node distance must represent the edge weights, and nodes must not be placed too close to one another. The algorithm attempts to place nodes that are connected to one another a distance that is a function of their edge weight. Therefore, a concept-term pair with a high edge weight will be placed near one another, and a concept-term pair with a low edge weight will be placed away from one another. In addition, the system attempts ensure that two nodes are not placed too close to one another, reducing the overlapping and occlusion of nodes. In order to increase the speed at which a steady-state solution is found, the concept nodes remain static and only the term nodes are moved to satisfy these criteria.

Animation techniques are used to move the nodes to their locations as the graph drawing algorithm iteratively attempts to satisfy the above criteria. The purpose of this animation is to illustrate to the user that they too can interact with the visual representation. For example, users can drag the nodes to visually organize the query space, after which, the system will continue to attempt to satisfy the graph layout criteria. An alternative would have been to delay rendering the graph until a steady-state solution was found. However, this delay would have been noticeable to the users, and would not convey the dynamic nature of the visual representation.

In order to promote the perception of the different classes of nodes in the visual representation, VisiQ uses colour coding techniques [102, 98, 110].
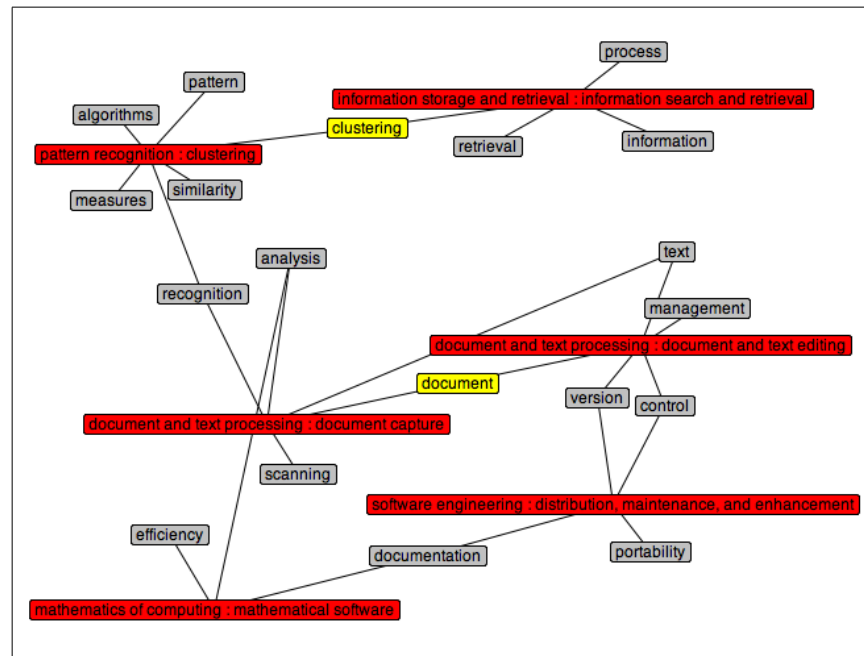
Figure 4.2: The visual representation of the query space generated by the terms "document" and "clustering".

The colours were chosen to be visually distinct from one another, representing different locations within the three-dimensional colour space of the colour opponent process model [36, 110].

Low luminance neutral colours (i.e., dark grey or black) are used to represent the underlying graph structure, including the edges between the nodes and the borders of the nodes. As shown in Figure 4.2, a medium-high luminance red is used to represent the concepts; a medium luminance yellow is used to represent the selected terms, and a medium-low luminance grey is used to represent the unselected terms. This use of colour allows the users to easily identify their original query terms (yellow), and deduce the meanings of the concepts (red) and the other terms in the query space (grey).

By default, the visual representation of the query space shows all the concepts and terms that are present. However, when the query space is large, a
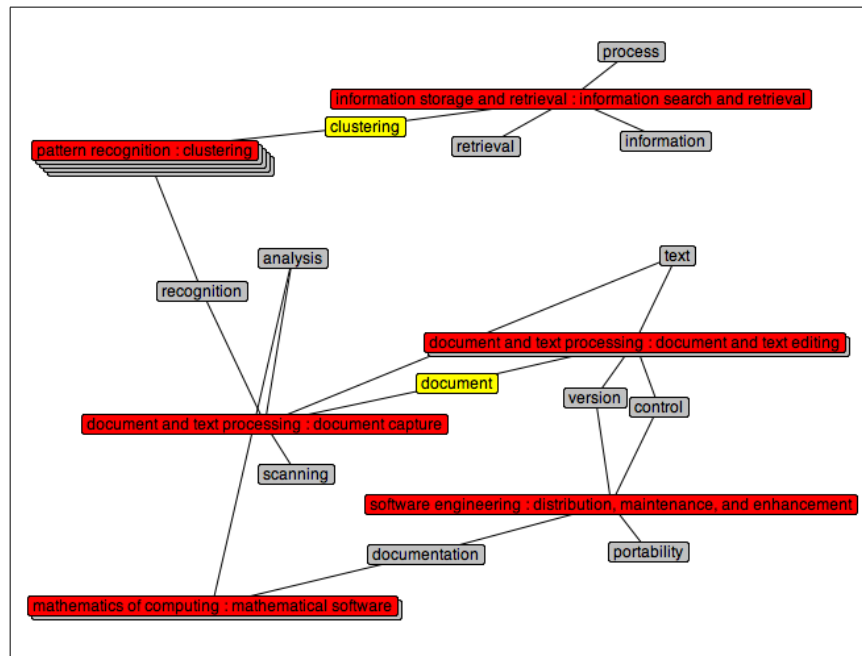
Figure 4.3: The visual representation of a query space with compacted concepts (pattern recognition: clustering and mathematics of computing: mathematical software).

more compact representation is beneficial. To this end, a technique is provided for compacting and expanding the terms that are connected to a concept in the visual representation of the query space.

When a concept is to be displayed in an expanded representation, all the terms that are connected to that concept are displayed as described above (see Figure 4.2). A compacted representation is also available, wherein the terms connected to the concept are displayed as thin layers below the concept, providing the illusion that these terms are stacked behind the concept (see Figure 4.3). However, when one of the terms connected to a compacted concept is also connected to an expanded concept, that term remains external to the compacted concepts to which it is connected until all its connected concepts are also placed in the compacted representation.

This feature allows the user to interactively adjust the level of detail of the visual representation. In effect, this as a type of localized semantic zooming [91] where information is hidden or shown, depending on the level of detail the user has chosen for the concept. The benefit of this technique is that the user can still identify the presence of terms and relationships to other terms or concepts, and can choose to view this information in more detail easily.

The user's original query terms are highlighted in the visual representation, which results in the ability to visually identify the relationships between their query and the concepts, as well as the relationships between these concepts and new terms that can be added to the query. At the most basic level, this allows the user to identify whether the terms they entered are actually appropriate for their information need. That is, the resulting concepts should have some relevance to the information the users are seeking. In addition, users can easily identify whether their terms are very general (connected to many different concepts) or very specific (connected to only a single concept).

Google Preview

In order to provide the user with an indication of the quality of their currently generated query, VisiQ provides a preview of the Google search results for the query. The Google Web API [27] is provided with the current state of the refined query; the corresponding search results consisting of the number of documents returned by the query, and the title and URL of each of the first five document surrogates, are displayed along side the visual representation of the query space. The snippet of text associated with each document surrogate is made available via a tool tip on the document surrogate title.

This preview of the results returned by Google provides the users with an additional source of information about the quality of their query. That is, as

Table 4.1: User interaction features in VisiQ.

Object	Operation	Result
any node	drag	Move the location of the node.
concept node	double-click	Toggle between expanded and compacted representations of the concept.
concept node	shift double-click	Remove the concept from the query space.
term node	double-click	Add or remove the term from the refined query.
term node	shift double-click	Remove the term from the query space.
'Google this' button	click	Open a new browser window and load the Google search engine using the refined query.

users interactively refine their query (as described in the following Section), they can determine the quality of the results of their actions. In effect, this feature provides the users with the ability to determine the outcome of their actions in an interactive manner, rather than the manual cycle of altering their query, submitting the query to Google, and viewing the results. The number of documents in the search results can provide an indication as to whether the query is becoming more general or more specific, and the individual documents in the search results can provide an indication of whether the query is yielding relevant documents.

4.2.3 User Interaction

Users of VisiQ can interactively refine their queries through the operations supported within the visual representation of the query space. These operations are visually depicted by the thick red arrows in Figure 4.1, and are described below. A summary of the operations that can be performed on the different node types within the visual representation of the query space are provided in Table 4.1.

New Query

The user can initiate a new query by simply typing in the query text box, and either pressing the Enter key, or clicking the "Visualize" button. In doing so, the system will generate the query space, produce a visual representation of this query space, and provide a preview of the Google results.

Studies of Web search behaviour have shown that users seldom make use of advanced query behaviour such as Boolean queries or term exclusion operators [89, 93]. As such, VisiQ only supports queries consisting of simple lists of terms. In future work, adding the ability for expert users to exert more control over how their query is refined may be considered. This could include phrase support, order of term control, and exclusion operators. However, these features will add another level of complexity to the methods by which the users can interact with the query space.

Add or Remove Terms From The Query

The users may interactively modify their query by double-clicking on term nodes they wish to add to the query. Newly added terms are highlighted in yellow, like the original query terms. As terms are added to the query, a new query space is automatically generated, and the visual representation is updated (i.e., concepts and terms that are no longer in the query space are removed, and those that are new are added). In addition, a new Google preview is generated, and the refined query box is updated to reflect this change.

To support the users in refining their queries (rather than just expanding them), the system also allows the user to remove terms from the query by

double-clicking a highlighted term. Doing so will return this term to the non-highlighted (grey) colour, and will regenerate the query space, update the visual representation, and reproduce the Google preview.

Note that the user may choose to remove terms that are present in their original query. This illustrates one of the key features of a query refinement system over a query expansion system: in a query expansion system, additional terms may only be added to the user's original query; in a query refinement system, the user may make modifications to the original query. This is a valuable feature in situations where the user identifies a new term that is a better description of their information need than one of the original query terms.

Remove Concepts

In reviewing the visual representation of the query space, the user may identify concepts which are not relevant to their current information need. These concepts may be present in the query space if the query terms are very general and have been used to describe multiple concepts. In order to make it easier for the user to focus on the concepts that are relevant, the system allows a user to remove a concept from the query space by holding down the control key while double clicking on the concept. In all subsequent interaction steps in the current user session, these removed concepts will not be present.

Expand or Compact Concepts

Users may choose to switch the representation of a concept between expanded mode and compacted mode by double-clicking on a concept. Switching between these different modes (as described in Section 4.2.2 and illustrated in Figures 4.2 and 4.3) allows the user to reduce the visual clutter generated by

concepts which they no longer wish to consider.

Expanding and compacting concepts also provides the users of VisiQ with a means for dealing with vague queries that generate large query spaces. Large query spaces will generally result in a visual representation that is difficult to interpret and make sense of (primarily due to edge crossings and node occlusion, as well as the inability of the graph drawing algorithm to satisfy the layout criteria). By compacting some of the concepts within the query space, these large query spaces become easier to visually process, since the number of nodes are significantly reduced. The user may then interactively expand and collapse the concepts as needed to refine their query.

Google Search Using the Refined Query

In addition to providing a preview of the Google search results as the user conducts their query refinement tasks, the system also allows the user to send the query they have created to the Google search engine [26] by clicking the "Google This" button. Doing so opens a new instance of the Web browser and submits the query to Google automatically. This allows the user to access the complete set of search results returned by Google, and to perform relevance judgement within the popular list-based search results interface.

A complete query refinement scenario using VisiQ is illustrated in Figures 4.4 - 4.9. A video showing the use of this system for visualizing and interactively refining a query is provided on the author's Web site[1].

In this scenario, the user beings with the query "document clustering". Upon viewing the visual representation of the query space, the user notices that the term "document" is not a good descriptor of the information need, and removes it from the query space. From the updated query space, the

---

[1]`http://www.cs.uregina.ca/~hoeber/VisualQuery/`
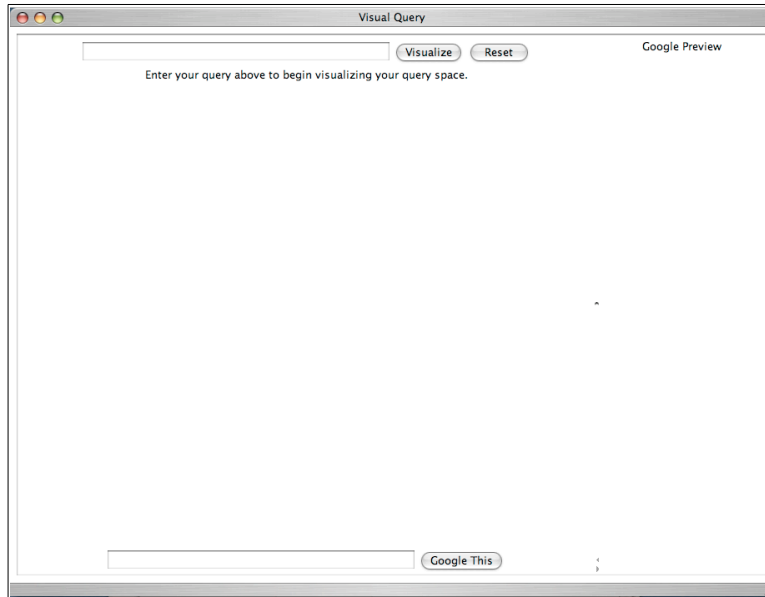
Figure 4.4: The process of visual query refinement with VisiQ begins with entering the initial query terms.
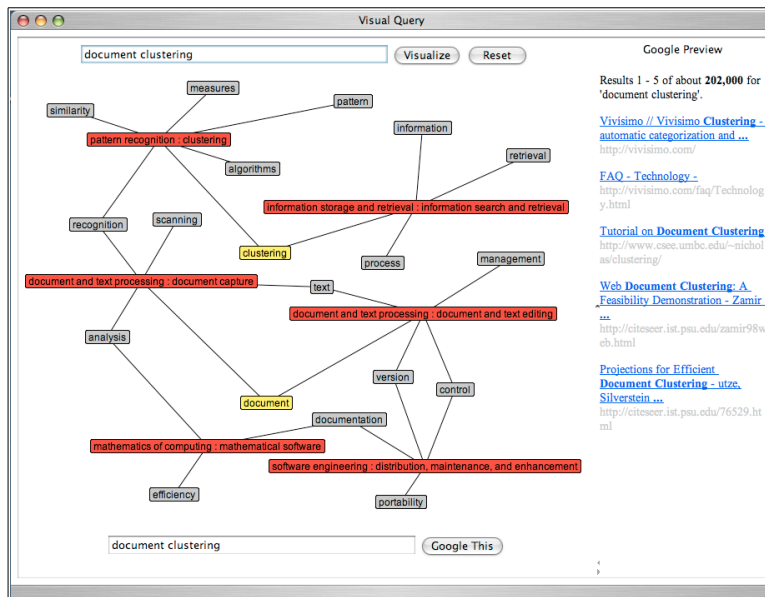
Figure 4.5: The process of visual query refinement begins with entering the initial query terms.
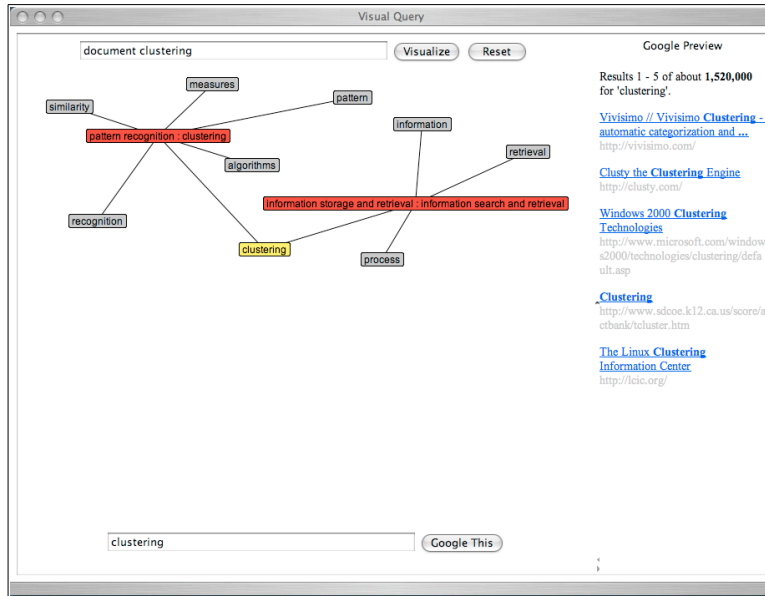
Figure 4.6: The users may choose to remove terms from the query.
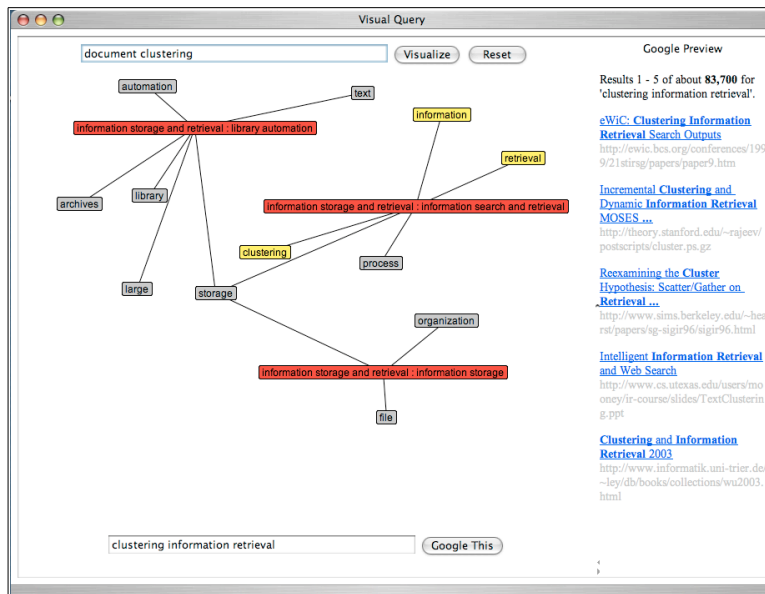
Figure 4.7: The users may choose to add terms from the query.
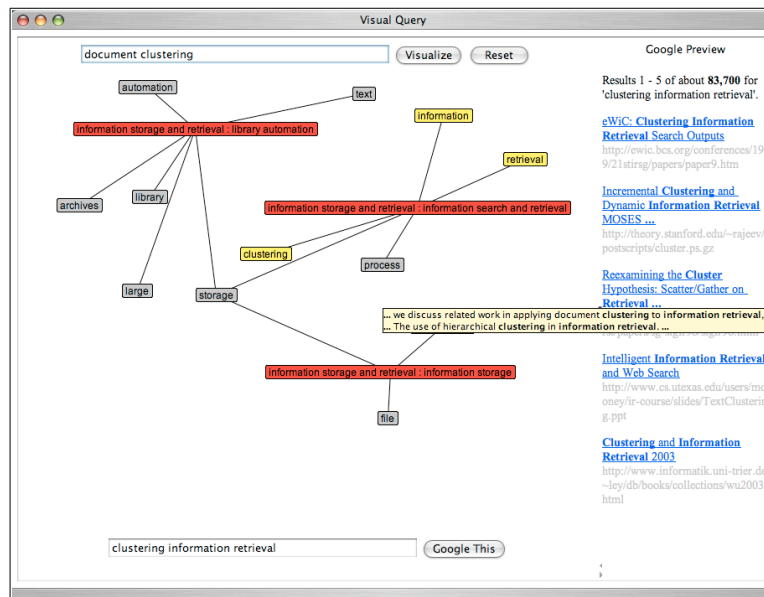
Figure 4.8: The users can determine the progress of their query refinement by consulting the Google preview.
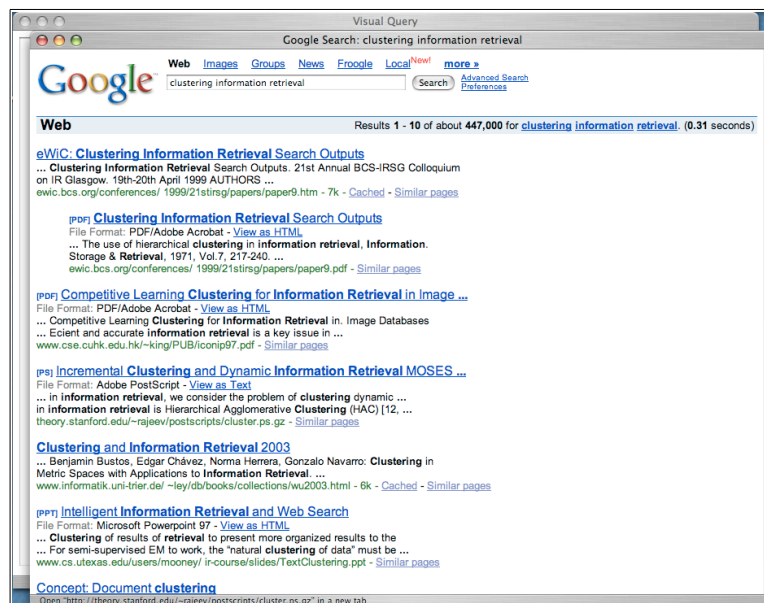
Figure 4.9: The refined query can be sent to the Google search engine with the click of a button.

user identifies "information" and "retrieval" as terms which are relevant, and adds these to the query. Using the Google preview, the user determines that the refined query is producing relevant documents, and completes the process by choosing to send the refined query to Google for further evaluation of the search results.

The benefit of using VisiQ in this example was that it illustrated to the user how the terms "document" and "documentation" are used to describe a number of concepts which are not relevant to the user's information needs. By removing the "document" term, the user is able to focus on the concepts that are relevant, and consider additional terms that might be more useful, such as "information" and "retrieval". This scenario represents one of many different situations where VisiQ can assist users in refining their queries.

Further examples are provided in the remainder of this Section to highlight four different Web search situations: starting with vague query terms, starting with specific query terms, starting with incorrect query terms, and conducting an exploratory search.

Vague Terms

Users often have difficulty in choosing specific terms to add to their query. Some terms that appear to be relevant to the user's information needs are sometimes poor choices since they are often used to describe many different topics.

For example, in the query "clustering algorithms", the term "algorithms" is rather vague. Within the concept knowledge base constructed from the ACM Computing Classification System, this term is connected with sufficient weight to eight different concepts. Within the visual representation of the query space, it is easy to identify these common terms, as well as other terms
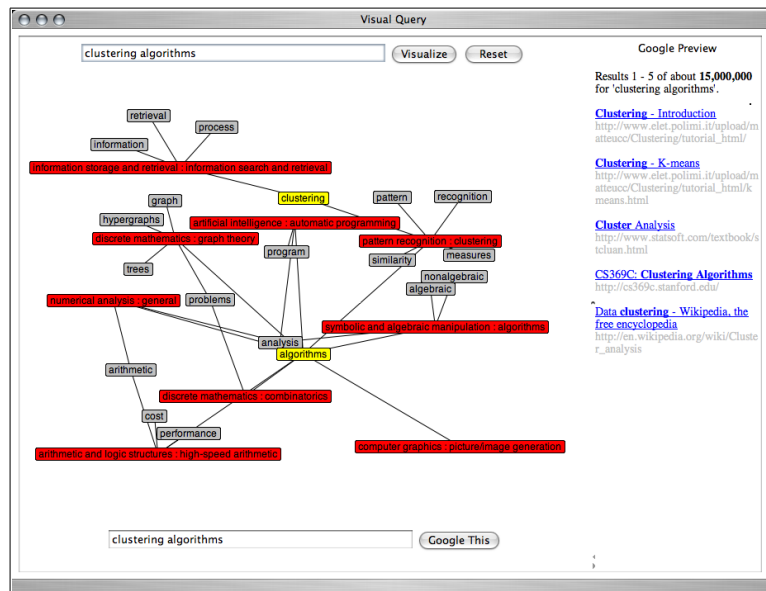
Figure 4.10: This example illustrates how VisiQ can be beneficial when the user begins with a vague query. The user is able to identify terms that are used by many concepts, and is given suggestions for other terms that might be more specific and relevant to their information needs.
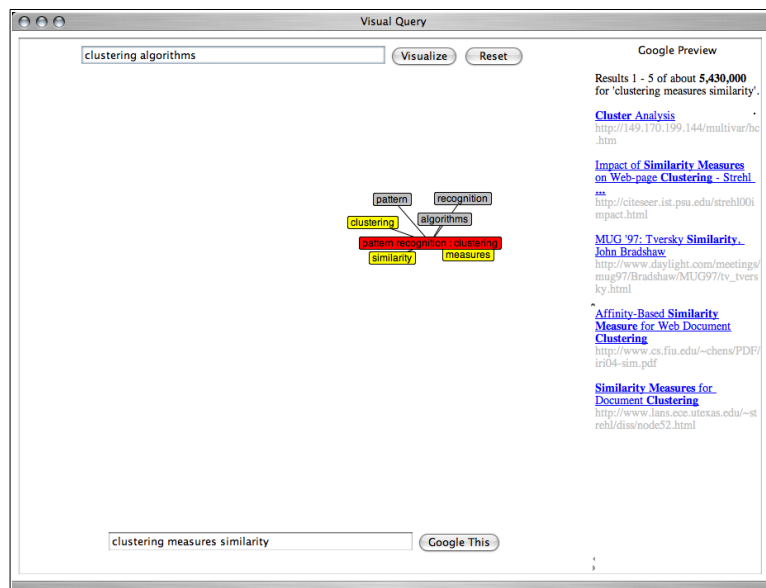
Figure 4.11: The refined query generated by removing vague terms and adding more specific terms.

that might be considered as replacements (see Figures 4.10 - 4.11).

The user may identify the concept "pattern recognition: clustering" as a relevant concept, and choose to remove the vague term "algorithm" from the query in favour of the terms "similarity" and "measures" in order to focus on a specific aspect of clustering algorithms. This example illustrates how VisiQ can be used to both identify vague terms (i.e., those that are connected to many concepts), and to suggest other terms that might be more valuable to the user's query.

Specific Terms

It is common for users who are familiar with a topic to provide a number of specific terms in their query in order to increase the likelihood of obtaining a large portion of relevant documents in the search results. VisiQ can further support this process by suggesting additional terms that are also relevant to the topic of interest to the user.

Consider the example illustrated in Figures 4.12 - 4.13. When the user submits a very specific query such as "library automation storage", VisiQ returns a query space consisting of the single concept "information storage and retrieval: library automation". Clearly, this is the concept which the user intended by this query.

A number of additional terms are also suggested as being relevant to this topic, including "archives", "text", "large", "information" and "retrieval". The user may choose to use these additional suggestions to further refine their query. For example, adding the terms "retrieval" and "archives" results in narrowing down the set of search results from 13.2 million documents to 700,000. While this is still many more documents than one would wish to consider, it does provide an indication of how VisiQ can assist the user in taking a specific
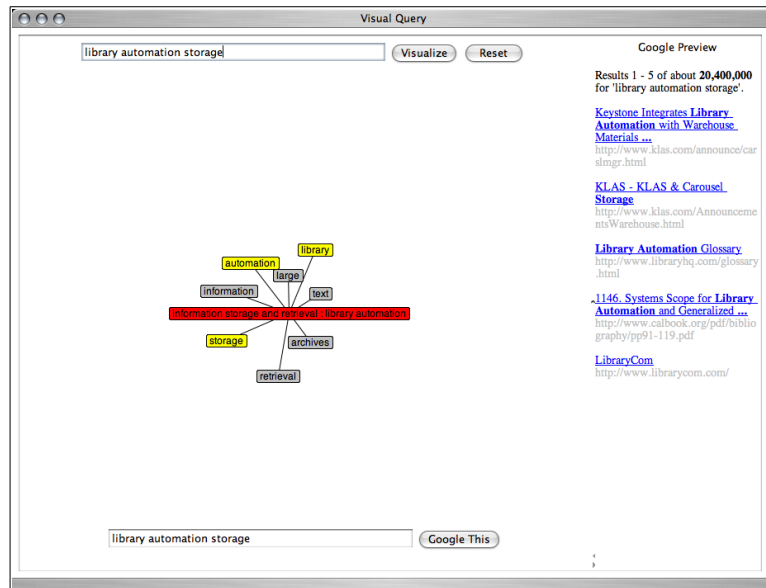
Figure 4.12: This example illustrates how VisiQ can be beneficial even when the user begins with a specific query. Not only can the user verify that all the terms are relevant for a specific topic, but they can also choose additional terms to further focus and refine their query.
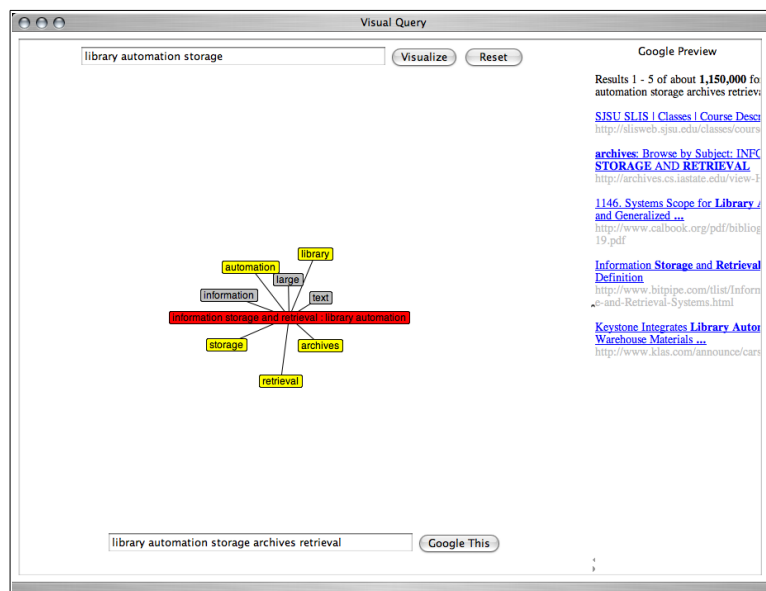
Figure 4.13: The refined query generated by adding even more specific terms.

query and making it even more specific.

Incorrect Terms

It is common for users that are unfamiliar with a topic to use incorrect terms to describe their information need. When this is the case, search engines seldom provide enough information for the user to fix this mistake.

For the query "performance complexity", the use of the term "performance" is likely not what the user wanted, and results in many documents that are not relevant to the user's intended topic of algorithmic analysis and computational complexity. Viewing the query space shown in Figure 4.14, the user can easily identify that the concepts to which the term "performance" is connected are not relevant to their information need.

By removing this incorrect term, the size of the query space is reduced, and the user is able to identify a concept that is relevant: "computation by abstract devices: complexity measures and classes". From this concept, the user can identify additional terms to add to their query, such as "classes" and "relations", arriving at a query that is a more accurate representation of their original information need (see Figure 4.15).

Exploratory Search

Often, a user will enter a somewhat vague query in order to determine what kind of information exists for this topic. This type of exploratory searching can be further supported through VisiQ.

Starting with a vague query such as "parallel operating systems" returns over 50 million documents. The query space generated by this query includes six concepts, and 27 additional terms that can be added to the query (see Figures 4.16 - 4.17). The user can easily review the concepts to gain a better
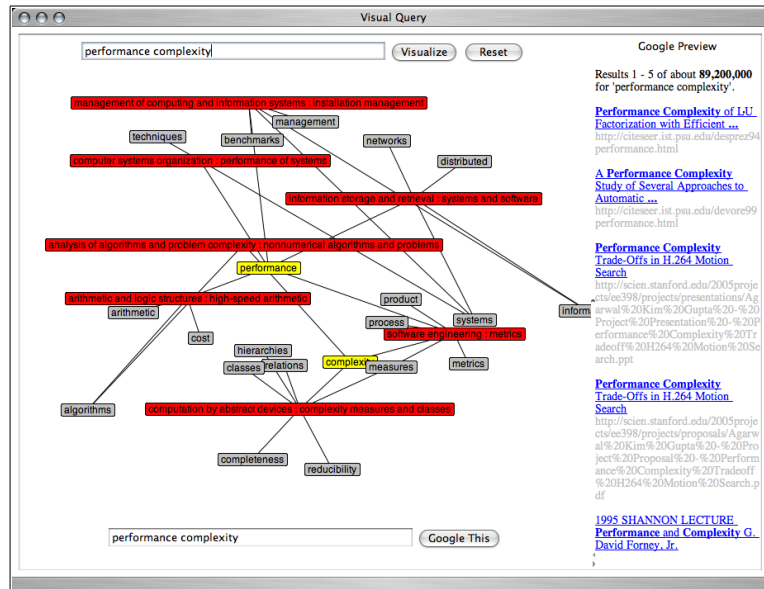
Figure 4.14: This example illustrates how VisiQ can be beneficial when the user chooses an incorrect term in their query. Incorrect terms can easily be identified since they will often be connected to concepts that are not relevant to the users information need. These can easily be removed, and more appropriate terms can be added.
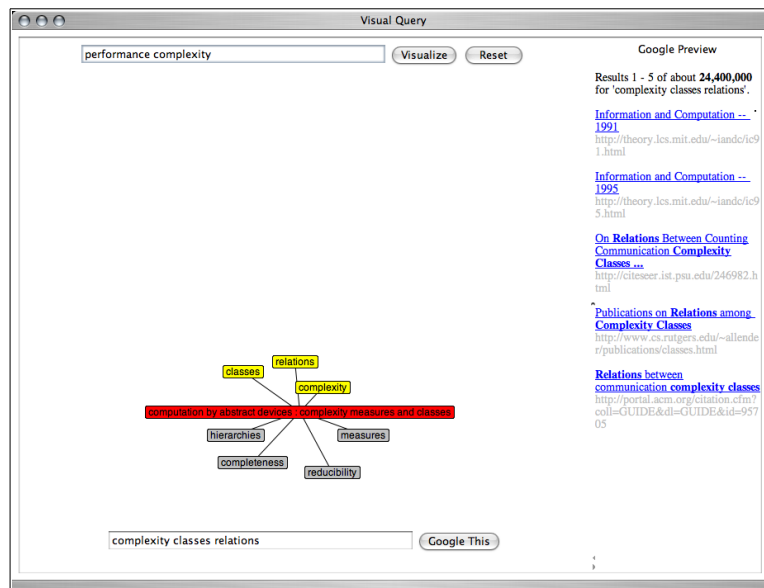
Figure 4.15: The refined query generated by replacing an incorrect term with a better alternative.
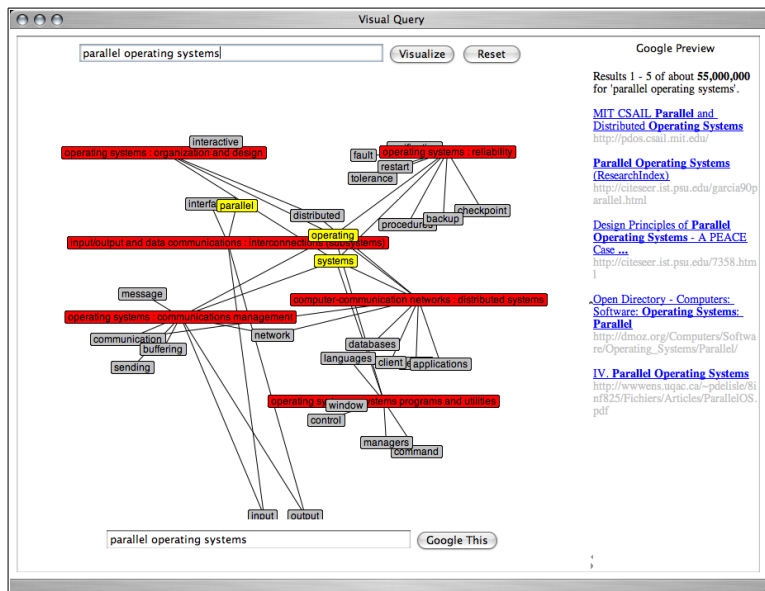
Figure 4.16: This example illustrates how VisiQ can be beneficial when the user wishes to explore a topic. The concepts related to the topic can easily be identified, along with suggestions for terms to further explore specific aspects of the topic.
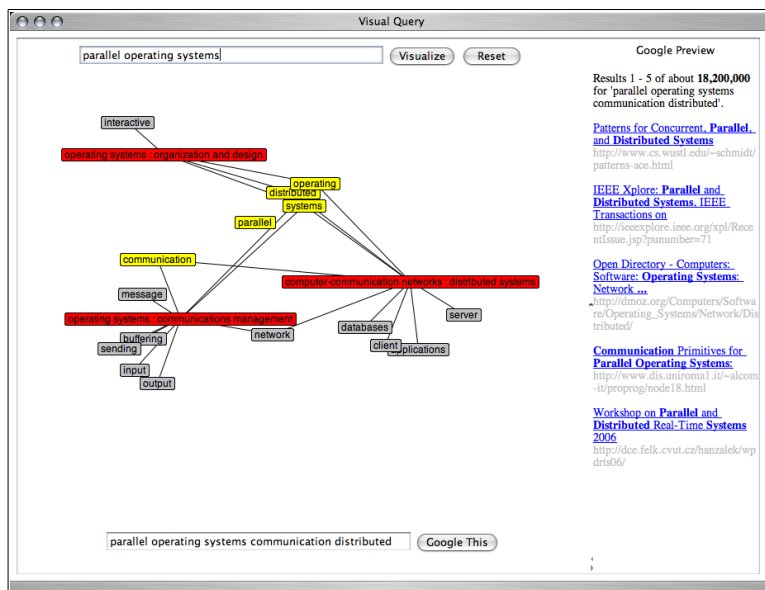
Figure 4.17: The refined query represents an aspect of the topic the user may wish to explore further.

understanding of the types of information that might be present, and can choose to focus on one or more of these concepts by adding additional terms to which they are connected.

For example, the user may decide that they would like to explore the topics of "operating systems: communications management" and "computer-communication networks: distributed systems". Adding the terms "communication" and "distributed" focuses the query on these topics.

Through the visual representation, the user is able to identify concepts of interest, as well as additional terms that might assist the user as they explore the information that is available. Further, through the interactive nature of the query refinement, the user can easily add or remove terms, and preview the search results, all in an exploratory manner.

4.2.4   Discussion

Following the classification discussed in Section 3.2, VisiQ is an interactive technique based on a collection-independent knowledge base. However, VisiQ is more accurately described as a query refinement technique, since it allows the users to additionally remove some or all of their original query terms in the crafting of a new query that accurately reflects their information needs.

One of the critiques of the work by Harman [30] is that it assumes that users can make the perfect choices when presented with a list of potential query expansion terms. Subsequent studies have shown that users are often not able to make these good choices [85, 57]. One may argue that this is a failure of the list representation used to display the potential query expansion terms. Since the list representation is not able to convey the relationships between the original query and the list of potential terms, nor the relationships between the terms themselves, users have difficulties in determining the value of adding

a term to their query.

A primary goal in the development of VisiQ was to provide the users with additional information about the potential terms that can be added to their query. This information comes in the form of the relationships between the user's query terms and the concepts they are related to, and the relationships between these concepts and the other terms that have been used to describe them. Providing this information in a visual manner allows the user to interpret and make sense of these relationships as a result of their visual information processing capabilities. That is, it is easier for the user to understand these relationships by looking at a picture than by reading a list of terms.

The Google preview assists the users in determining the results of their query refinement steps without the need to access the entire set of search results. This supports a process of query refinement exploration, in which the users can try modifying their queries, and easily determine the outcomes of these modifications.

Another important aspect of VisiQ is that the query refinement process occurs completely in the query space. The use of the concept knowledge base allows the query space to be generated completely independent of the initial search results or the collection of documents being searched. The value of this approach is best illustrated by the simple example below.

Consider the situation where a user enters a query term that is used to describe two different concepts. If one of these concepts is much more common in the corpus, or is given a higher value by the information retrieval system, then the top documents returned by the information retrieval system may all be related to this concept. In this case, the terms suggested for query expansion will be focused exclusively on this common or important concept. If the user was actually interested in the other concept, then this method of

basing the query refinement on the initial search results or the features of the document collection will have failed.

In using a concept knowledge base and therefore keeping the query refinement in the query space, this situation is avoided. That is, the user will be able to easily identify the two different concepts to which the query term is related, and can refine the query based on the concept that matches their information need (rather than the one the information retrieval system automatically selected).

It is clear that the quality of the query refinement process is directly impacted by the quality of the concept knowledge base. The concept knowledge base generated from the ACM Computing Classification System is of an adequate quality for the purposes of demonstrating the results of this research, although the query refinements are only possible within the computer science domain. Future work includes investigating the further refinement of the concept knowledge base using the top documents returned by the ACM Digital Library for each of the disciplines (concepts), as well as constructing a general purpose concept knowledge base using the Open Directory Project and Wikipedia.

Spence's Navigation Framework

As described in Section 2.4.1, Spence suggested a framework for describing the human activities involved in navigating within an information system [90, 91]. This framework provides a useful tool for discussing the features and benefits of VisiQ with respect to the cognitive activities the user conducts.

Browsing. During the browsing activity, the user views and interacts with the VisiQ system, and absorbs the content displayed within the system. The

user can easily identify their query terms, as well as view the concepts and other related terms present in the query space. Further, the Google preview presents the titles of the top five documents that match the current query, allowing the user to browse the preview of the what the search results will look like.

Formulate an Internal Model. Through the visual browsing activity, the user develops an internal model of the content viewed. Because the query space is provided in a visual representation that highlights the current query terms, the concepts to which they are related, and the additional terms that can be added to the query, VisiQ supports the user in formulating a mental model of the query space that mimics this graph-based representation. Even for queries that result in a large query space, the ability to interact with the query space through compacting and expanding concepts, and manipulating the locations of nodes, assists the user in formulating their internal model. Further, because the concepts are static (i.e., based on the information stored in the concept knowledge base), the generation of internal models for new queries are supported by the internal models developed for prior queries that resulted in common concepts.

Interpret the Information The ability of the user to understand what they are seeing is affected greatly by the internal model that has been formulated and the information they are viewing. In VisiQ, colour coding is used to provide a clear distinction between the three classes of nodes in the visual display: query terms, concepts, and additional terms. As such, little cognitive effort is needed to make the proper interpretation of the query space, both based on what they are seeing in the visual display, and based on the internal

model they have developed. A well-formulated internal model results in an easy interpretation of what the current query "looks like".

Formulate a Browsing Strategy. A browsing strategy can be formulated based on a number of different sources of information: the internal model (i.e., cognitive determinants), the visual representation being viewed (i.e., perceptual determinants), or even a new idea the user has had (also a cognitive determinant). The resulting browsing strategy may be part of planned strategy with a specific goal in mind, or an opportunistic strategy based on something the user has thought of or seen.

In VisiQ, examples of browsing strategies that are based on the interpretation of the internal model include choosing to add or remove terms from the query (planned strategy), or trying "what if" scenarios and exploratory query refinement (opportunistic strategies). Based on the visual representation of the query space, the user may notice the Google preview and choose to view some of the top ranked documents for the current query (planned strategy). Alternately, while viewing the visual representation, the user may identify some term which they had not yet considered, and may choose to add this to their query (opportunistic strategy). Further, the user may have some idea for a new search (opportunistic strategy), or for a new term to add to their current query (planned strategy).

Norman's Stages of Action Framework

Norman described two gaps that exist between the users of a system and the system itself: the gulf of evaluation and the gulf of execution [68, 69]. How the users of a system cross these gaps form Norman's stages of action framework, which is described in more detail in Section 2.4.2. As with Spence's navigation

framework, this framework provides a useful mechanism for commenting on how the features of VisiQ support the users in fulfilling their task goal of satisfying their information need.

Goals. The overall information retrieval goal is for the user to find a number of documents that are relevant to their information needs. In order to fulfill this goal, the user must first fulfill a supporting goal of crafting a query that accurately reflects their information needs. This supporting goal is translated into a system goal for using VisiQ to craft a query.

Gulf of Evaluation. While viewing the visual representation of the query space, the user will normally move through the phases of perception, interpretation, and evaluation subconsciously. In the perception phase, the user sees the pixels in the information display, and groups these into perceptual objects. The regular shapes of the nodes, as well as the visible connections between concept nodes and term nodes support the user in correctly perceiving this information. When the query space is large, there is the possibility of an occlusion problem; however, the interactive nature of the visual representation supports the user in manipulating the graph in order to separate overlapping nodes.

In the interpretation phase, the user recognizes the three different types of nodes within the graphical representation of the query space: query terms, concepts, and additional terms. The colour coding of the nodes supports the user in this interpretation. The user's previous experience with using VisiQ assists them in making a correct interpretation of the interface. Due to the animated nature of the visual display, the users are able to interpret that they too can move nodes around in the display. Further, the nodes themselves have

the appearance of buttons, providing an affordance that they can be clicked.

As a result of making a correct interpretation of what the user has seen, the user is able to evaluate the visual display with respect to their current system goal. With VisiQ, the user is able to follow the connections between their query terms and the concepts to which they are related, and follow the connections between the concepts and the other terms that may be of value in crafting the query. In effect, the user is able to make sense of the visual display in terms of their current query.

Gulf of Execution. In the intentions stage, the user converts their system goal of crafting a query that accurately reflects their information needs into an intention to make a change to the query. In VisiQ, this intention will often be in the form of adding or removing a term from the query. Other intentions are also possible, such as moving a node that is occluding another, removing a non-relevant concept from the display, or compacting a concept.

In order to fulfill the intention, the user must plan a sequence of actions within the user interface. Since many of the operations are simple click or drag operations, commonly, there is little planning required. However, there are a few more complex actions, such as using the shift key when double clicking on a node to remove it from the visual display. These actions may require conscious planning, especially when the user is a novice.

The execution stage is where the user makes the physical actions with the interface device to execute the planned action sequence. Since VisQ uses standard interface devices such as a keyboard and mouse, this stage is trivial for most computer users, requiring little cognitive effort.

After successfully executing the action sequence to add a term to the query, the query space is re-calculated and a new visual representation is provided.

All the nodes that were present in the previous query space remain in the same location; old nodes are removed and new nodes are added to the graphical representation. Further, the node the user clicked on is changed to have a yellow background, indicating that it is a query node. This new query space provides feedback to the user that they have correctly executed their action sequence.

4.2.5 Summary

The primary benefit that VisiQ provides to Web searchers is the ability to interactively refine queries based on supplemental information drawn from the concept knowledge base. For a given query, the query space is generated and displayed in an interactive graphical representation. Within this graphical representation, the users can easily identify their query terms, the concepts to which they are related, and additional terms that may be of value for the query. Within this query space, the user can choose to add or remove terms from the query, resulting in the generation of a new query space and the corresponding graphical representation. A preview of the Google search results provides an indication to the user of the progress of their query refinement tasks.

4.3 WordBars

One of the main drawbacks to the VisiQ system is that it requires that a concept knowledge base exist for the domain in which the users wish to search. If no such knowledge base exists, VisiQ is unable to support the users in the refinement of their queries. To address this drawback, methods for supporting interactive query refinement based on the top results from the initial search were explored, resulting in the development of WordBars.

The primary hypothesis in this work is that frequently used terms in the results of an initial search can provide valuable information to Web searchers, both for interactive query expansion as well as for interactive search results exploration. Information visualization techniques are employed to convey the term frequency information to the users in a compact manner that can easily be interpreted, understood, and used. Although this system supports both interactive query refinement and interactive search results exploration, the focus in this Section is on the interactive query refinement features. The interactive search results exploration features are discussed in more detail in Chapter 5.

The WordBars system retrieves the top 100 document surrogates from the Google API [27], and counts the frequencies of all the terms used within the titles and snippets. The term frequencies are sorted and depicted in a visual manner, allowing the users to easily identify the commonly used terms within the top search results. Single-clicking on any term re-sorts the search results based on the frequency of that term. Selecting multiple terms results in a re-sorting of the search results based on the combined term frequencies. Double-clicking a term either adds a new term to the query, or removes the corresponding term from the query. New search results are retrieved by clicking the search button.

As with VisiQ, a fundamental design principle in the development of Word-Bars is the balance between computer automation and human control [88]. Crafting a query that accurately represents a user's information needs is an inherently human task, as is the evaluation and exploration of the search results. While some have suggested automatically expand users' queries [114, 77, 108], this eliminates the human decisions made in the query refinement process, which are vitally important.

4.3.1 System Architecture and Features

The system architecture for WordBars consists of four primary components. The *meta-search component* obtains the set of top search results for the supplied query from the Google API. The *term frequency vector generation component* counts the occurrences of terms within the top search results, and generates vector representations of this information. The *term frequency vector visualization component* generates a visual representation of the term frequencies within the entire set of search results. The *search results re-sorting component* re-sorts the search results based on the frequency of use of user-selected terms.

This system architecture is illustrated in Figure 4.18, the details of which are provided in the remainder of this Section. The user interaction steps shown in this diagram are described in Section 4.3.2.

Meta-Search

Unlike VisiQ, which provided interactive query refinement features independent of the search results set, WordBars integrates the query refinement functionality together with the presentation of the search results. As such, WordBars uses the Google API [27] to retrieve the top 100 search results for the current query.

Since the Google API only provides ten document surrogates at a time, it is the responsibility of the meta-search component to make ten different calls to the Google API in order to collect all 100 document surrogates required by the system. In order to avoid delays introduced by the multiple requests for search results, as each set of ten document surrogates are received, they are passed to the other components requiring them. Ultimately, the user will see
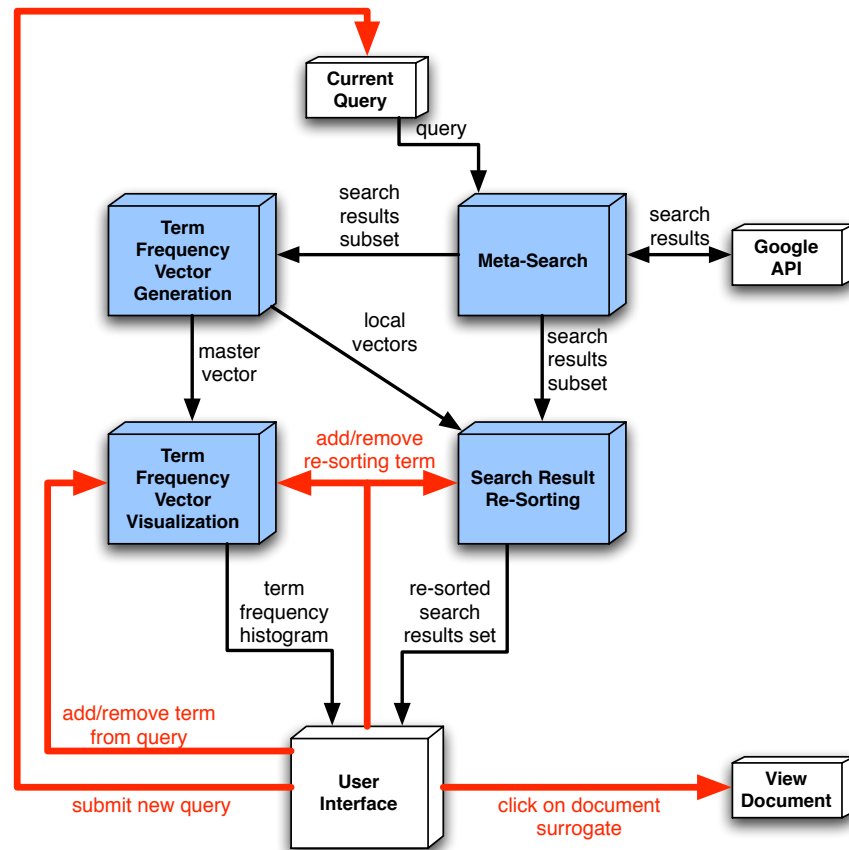
Figure 4.18: The WordBars system architecture consists of four primary software components (coloured in blue): meta-search, term frequency vector generation, term frequency vector visualization, and search results re-sorting. The interaction paths are illustrated by the thick red arrows.

the search results list and term frequency histogram grow as the document surrogates are obtained from the Google API.

Term Frequency Vector Generation

As each search results subset is provided by the meta-search component, the following steps are followed in order to generate a single master vector and a local vector for each document surrogate:

1. The title and snippet from each document surrogate in the search results subset are combined in a bag-of-words approach resulting in a document descriptor for each document surrogate.

2. Common terms and terms that contain less than three characters are eliminated from the document descriptors.

3. Porter's stemming algorithm [74] is used to reduce all terms in the document descriptors to their stem (root) forms.

4. For each document descriptor, the frequency of each unique stem is counted and a local vector is created to store this information.

5. The local vector for each document surrogate is added to the master vector.

In the master and local vectors, each dimension corresponds to a unique stem found in the search results. As new stems are found, the dimensions of the vectors are increased accordingly.

After processing a search results subset provided by the meta-search component, the master vector is sorted to ensure that the most frequent terms are always located at the top. This vector is used as the basis for visually representing the term frequencies, as explained in the following Section. A mapping between this sorted master vector and the local vectors is maintained to support adding subsequent local vectors to the master vector (as more search results subsets are processed), and re-sorting the document surrogates.

Term Frequency Vector Visualization

While some previous systems have used simple textual lists to provide recommendations for additional terms to add to the query [30], providing additional

information about the terms in a visual manner can be extremely beneficial. For example, Joho et al. showed the benefits of using a cascading menu representation of query expansion terms [49]. In WordBars, a simpler representation is employed that both allows the user to browse the available terms, as well as perceive and interpret the relative frequencies of these terms in the top search results.

The visual representation of the term frequencies consists of a vertically oriented, colour-coded histogram. Both the width of the bars in the histogram, as well as the intensities of the colours, are used to represent the frequencies of the commonly used terms in the top search results. Using multiple visual features to represent the same data attribute provides redundant coding, and results in an increase in the ease, speed, and accuracy in which the the users are able to perceive and interpret the information [84]. The colour scale was chosen to vary both on the red-green colour channel, as well as the luminance channel. Visually, this colour scale produces a heat scale, resulting in high frequency terms appearing hot, and low frequency terms appearing neutral or warm. The colour scales used in WordBars were generated using the ColorBrewer application [12].

The term labels are provided to the right of each frequency bar. The terms that are present in the query are coloured red; all others are black. This use of colour allows the users to easily identify their query terms within the histogram, as well as identify frequently used terms that are not present in the query. Further, these colour distinctions can be pre-attentively processed [110], allowing the near-instant recognition of the distinction between the query terms and the other terms.

Due to space considerations, only the 20 most frequently used terms are displayed in the term frequency histogram. While there may be relevant terms

beyond this cut-off mark, one may assume that the most beneficial terms are those that are used frequently within the top search results. Although this may not be the case in some search situations, providing a limited number of terms produces a list that can be easily browsed and is not overwhelming for the users to evaluate.

Grey boxes are used to indicate which terms the user has selected for re-sorting the search results. This provides a simple yet effective method for indicating the current state for the re-sorting of the search results. Figure 4.19a shows a screenshot of the WordBars system for a sample query; Figure 4.19b shows zoomed-in view of the visual representation of the term frequencies.

Search Results Re-Sorting

The purpose of the search results re-sorting component is to re-sort the set of search results based on the terms selected by the users from the term frequency histogram. Initially, when the search results subsets are provided by the meta-search component, the document surrogates are simply stored and passed to the user interface in the original order (as provided by the Google API). As the local vectors are calculated by the term frequency vector generation component, these are stored along with their corresponding document surrogate within this component.

When a user adds or removes a term for re-sorting purposes using the term frequency histogram, this component uses the dimensions corresponding to all the selected terms as a means for re-sorting the search results. A selection vector is generated which contains a value of one in the dimension corresponding to the selected terms, and a value of zero elsewhere. The re-sorting of the search results set is based on the calculation of the inner product between the selection vector and each of the local vectors (corresponding to each document
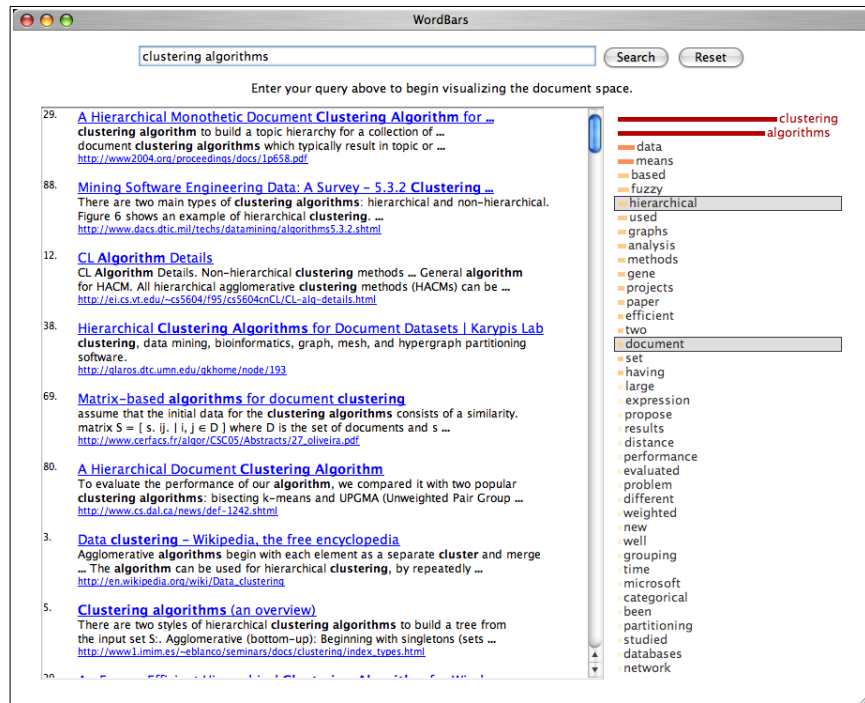
(a) WordBars
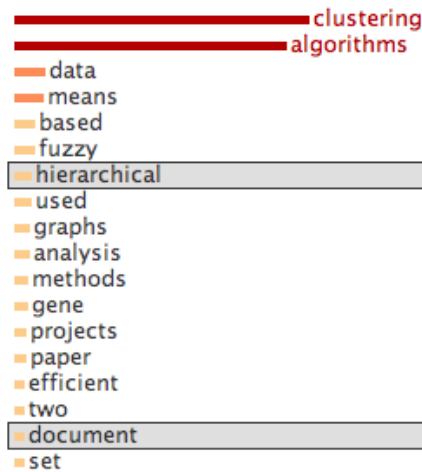
(b) term frequency histogram

Figure 4.19: A screenshot of the WordBars system (a) and a zoomed-in view of the top of the histogram (b). Note the re-sorted search results based on the selected terms in the histogram.

surrogate). The end result is that the search results set is provided to the user interface in a re-sorted order, based on the frequency of use of the selected terms.

4.3.2   User Interaction

As the search results are retrieved from the Google API, the document surrogates are automatically loaded into the document list window, and the term frequency histogram is updated as each document surrogate is processed. This has the effect of providing an animation of the growth and re-sorting of the terms used in the search results. A video showing this animation, is available on the author's Web site[2].

Once data begins to be displayed in the term frequency histogram, the user can interact with this interface without waiting for the search results to finish being retrieved. Users of WordBars can interactively refine their queries through the operations provided by the term frequency histogram and the other features of the interface. These operations are visually depicted by the thick red arrows in Figure 4.18, and are described below. A summary of the operations that can be performed using WordBars are provided in Table 4.2. The interaction methods were chosen to be as simple as possible in order to reduce the learning curve associated with using WordBars.

Investigate Term Use

Single-clicking terms within the term frequency histogram is used to initiate a re-sort of the search results displayed in the document list window based on the frequency of all the currently selected terms. Clicking a term toggles its

---

[2]`http://www.cs.uregina.ca/~hoeber/WordBars/`

Table 4.2: User interaction features in WordBars.

Object	Operation	Result
term in histogram	click	Toggle between selecting and unselecting the term. Re-sort the search results based on the selected terms.
term in histogram	double-click	Add or remove the term from the query.
'Search' button	click	Initiate a new search with the refined query.
document surrogate	click	Open the document in a new browser window.

status between selected and not selected. Selected terms are easily identified by the grey box surrounding them. This simple process allows the users to interactively investigate the use of terms within the current set of search results that may be of value for query refinement.

The users can easily add or remove terms to be used for re-sorting, resulting in an instantaneous re-sorting of the search results. This feature supports an interactive investigation of term use within the current set of search results, allowing the users to see how a term is used before committing to adding it to the query. This is an important part of the support WordBars provides for interactive search results exploration, as it supports the users in making informed query refinement choices. This feature is also of value in interactive search results exploration, and is discussed further in Section 5.4.

Add or Remove Term From The Query

Double-clicking is used to add or remove terms from the current query. All terms that are in the current query are displayed in a red font in the term frequency histogram. Double-clicking on any of these will remove that term from the query; double-clicking on any term that is currently not in the query

will add that term to the end of the query. This feature allows the users to easily refine their query based on the terms that are present in the current set of search results.

Submit the Refined Query

Adding or removing terms from the current query does not generate a new set of search results automatically. A new set of search results and the corresponding term frequency histogram are generated only once the user submits the query by clicking on the 'Search' button. Although automatically producing a new set of search results whenever the query is changed was originally proposed, this causes difficulties when the user wishes to add or remove multiple terms from the current query. As such, the query is refined using the term frequency histogram of the previous query until the user explicitly requests a new set of search results (and corresponding term frequency histogram).

4.3.3 Examples

In order to illustrate the benefits WordBars provides to Web searchers through the interactive query refinement features, four examples are provided to highlight four different Web search situations: starting with vague query terms, starting with specific query terms, starting with incorrect query terms, and conducting an exploratory search.

Vague Terms

It is common for users to provide vague queries for their Web search tasks. This may be due to incomplete knowledge on the topic of interest, a desire to explore a general topic, or choosing a query term that is inherently vague.

The search results for a vague query are often vague themselves. Sometimes these search results will all be relevant to some general topic that is clearly not specific enough to satisfy the users' information needs; other times, the search results may be relevant to two or more very different topics. In these situations, users tend to spend a lot of time considering document surrogates that are not relevant to their information need. Web search engines provide little support to help users improve their queries or to focus on more specific aspects of their information need.

With WordBars, the users can benefit from being able to easily browse the commonly used terms in the search results. Vague search results can be identified by the high frequency of the query terms, and the relatively low frequency of all the other terms in the histogram. This is due to the search results being a mixture of documents on multiple topics or sub-topics, all of which use different terms in their descriptions.

Users of WordBars benefit from the support that the system provides as they explore and browse the search results. If relevant documents are found near the top of the list after re-sorting the search results, the selected terms can be added to the query, and other less valuable terms removed. As a result, the user can first consider documents that make use of a potential new query terms, and then easily add these terms to the query. All this interaction occurs within the same interface, allowing the user to readily flip back and forth between their task of search results exploration and their task of query refinement.

Suppose the user starts with the vague initial query: "document clustering". Clearly, the query terms are used frequently; but few other terms are used consistently in many of the search results, indicating the vagueness of the initial query (Figure 4.20). The user can explore the search results by
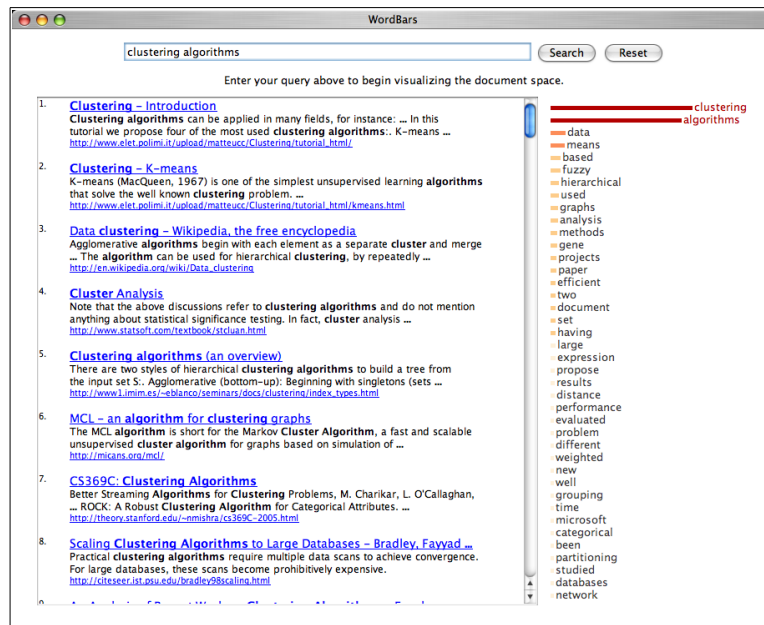
Figure 4.20: Starting with a vague query, WordBars produces a term frequency histogram with a mix of relevant and non-relevant terms.
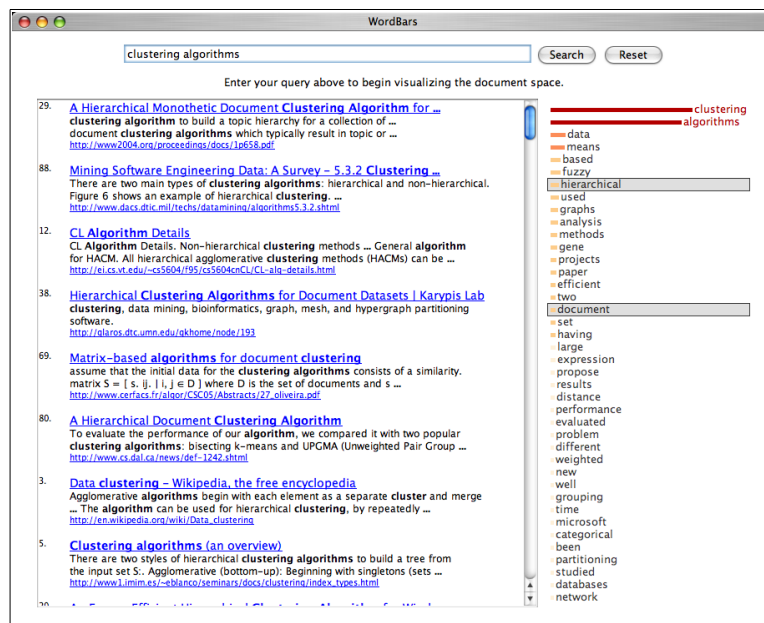
Figure 4.21: How the terms are being used in the top search results can easily be explored.
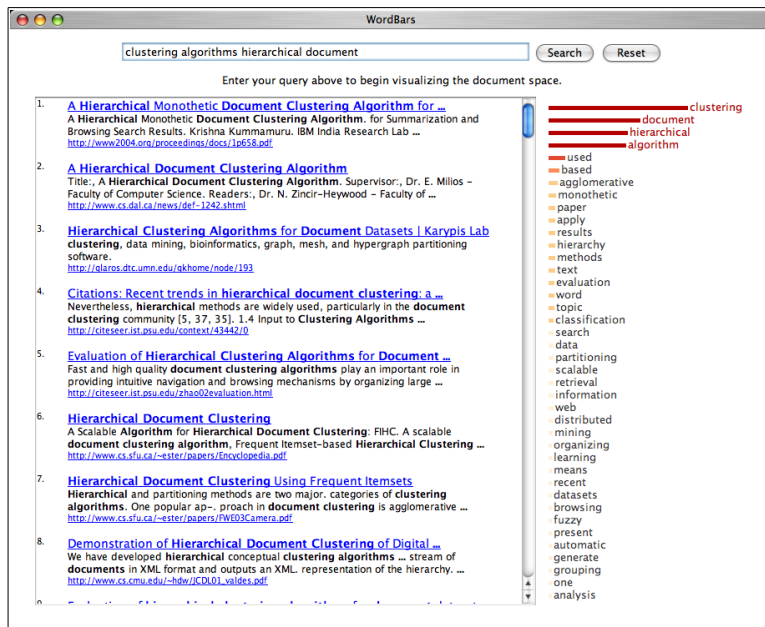
Figure 4.22: More specific terms can easily be added to the query, generating a new term frequency histogram.
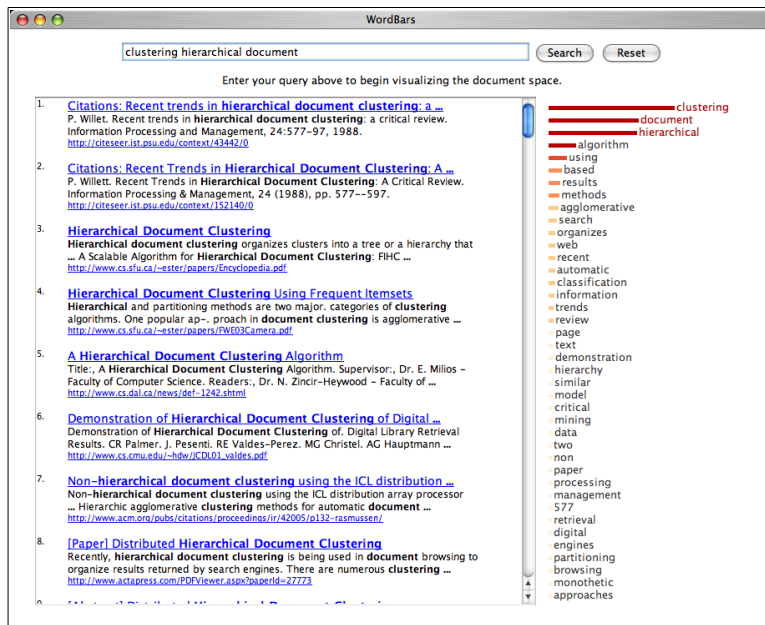
Figure 4.23: Vague terms can easily be removed from the query, generating a new term frequency histogram.

selecting terms that are better descriptors of their information need, such as "hierarchical" and "documents" (Figure 4.21). If the top documents are relevant, the user may choose to add these terms to the query by double-clicking on them (Figure 4.22). The user may decide that some of the query terms are not very descriptive, and may choose to remove these, such as "algorithms" (Figure 4.23).

From this example, it is easy to see the value of being able to re-sort and explore the search results in support of the interactive query refinement process. This is supported using the simple interaction features provided by the term frequency histogram. The terms presented in the histogram can easily be considered for relevance, and can be used to focus on a subset of the search results that are relevant to a specific sub-topic (i.e., by re-sorting the search results), or focus and refine the query with respect to this sub-topic (i.e, by adding the terms to the query).

In this scenario, the user was able to start with a vague initial query, and explore the search results using the term frequency histogram to find additional terms that were also relevant to their goal. The user recognized terms in the histogram that were relevant to their information need, and was able to evaluate how these terms were being used in the search results. This exploration is further supported by allowing the user to craft a new query using terms from the histogram, which can then be further explored.

Specific Terms

In general, when a user is able to provide a specific initial query that accurately reflects their information needs, Web search engines do a very good job of providing highly relevant documents within the first few pages of the search results. Even in these situations, there is a benefit to using WordBars.

By providing a term frequency histogram to represent the commonly used terms in the top search results, the users can easily verify that their initial query is indeed returning documents that are relevant. In these situations, many of the top terms in the histogram should be relevant to the user's information need. By providing a visual indication of the frequency of the terms, the users can easily interpret the relative frequency differences between terms. The user may use the term frequency histogram to re-sort the search results to further focus on and explore a particular aspect of the information need. The user may also add new terms to the query, resulting in a search that is even more specific than the initial query.

Suppose the user starts with a specific initial query "library automation storage". By reviewing the top terms provided in the term frequency histogram, the user can easily verify that many of the documents are relevant to their information need (Figure 4.24). The user can easily focus on a specific aspect of the search results, such as "network" and "software" by clicking on these terms in the histogram (Figure 4.25). Alternately, the user may choose to select the term "backup" to obtain a different sorting of the search results (Figure 4.26). The user may decide to add this term to their query by double-clicking on it, generating a more specific set of search results (Figure 4.27).

This scenario shows how the user is able to explore the search results using WordBars. This allows them to interactively manipulate the search results as they evaluate how various terms are being used within the document surrogates. In some cases, the users may find that exploring the search results from their initial query is sufficient to satisfy their information needs. In other cases, after exploring how some of the terms are being used in the search results, they may wish to refine their query, resulting in a new set of search results that can be further explored using the new term frequency histogram.

Figure 4.24: Starting with a specific query, WordBars produces a term frequency histogram with many relevant terms.
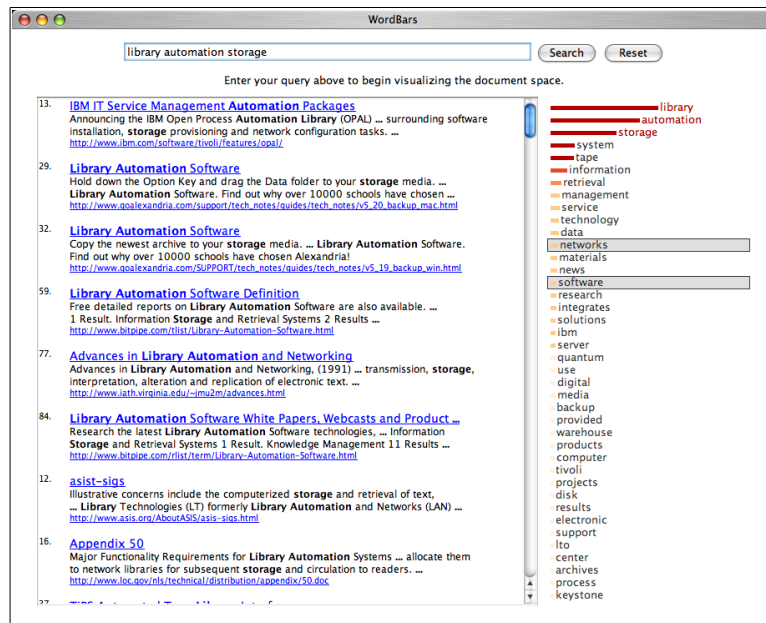
Figure 4.25: How the terms are being used in the top search results can easily be explored.
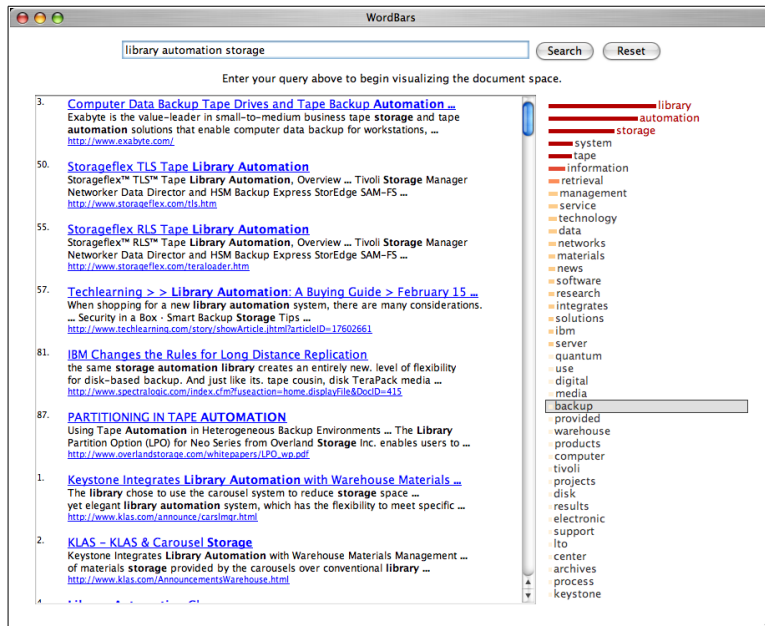
Figure 4.26: The user can easily investigate the use of many different terms.
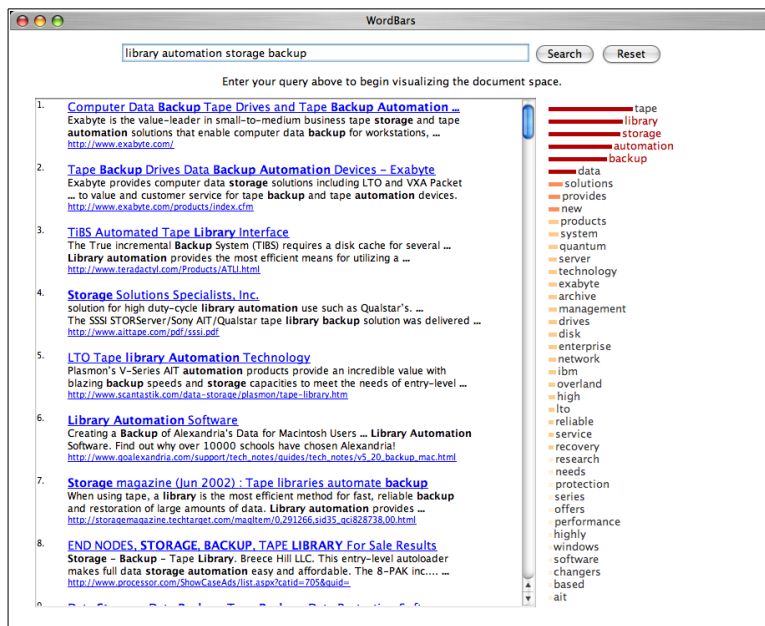
Figure 4.27: Additional terms can be added, resulting in an even more specific query.

Incorrect Terms

When users are unfamiliar with the topic upon which they wish to conduct a Web search, it is common for them to use incorrect terms to describe their information need. Although these terms may appear to the user to be relevant, they may not be commonly used to describe the desired topic. This poses a challenging problem for the users, as there is little help provided by the traditional Web search engines in refining their queries to accurately reflect their search goals. In these situations, it is probable that many of the top search results will not be relevant, making it difficult for the users to identify what the problem is with their query.

The term frequency histogram can assist the user in identifying that there is a problem with their query. If many of the terms are not relevant to their information need, this provides a clear indication of errors within the query. While browsing the histogram, the user may be able to identify some terms that are better descriptors of what they are seeking than those originally used. These terms can easily be evaluated by re-sorting the search results to bring the document surrogates that make use of them to the top of the list; the terms can then be added to the query if they are found to be of value.

Suppose a user wishes to search for documents on the topic of algorithmic analysis and computational complexity. If they are not very familiar with the topic, they may begin with a query such as "performance complexity". However, the use of the term "performance" may not be correct since it is often used to describe actual system performance rather than algorithmic analysis topics. Viewing the term frequency histogram, the user can see that there are many terms that are not directly related to the intended topic (Figure 4.28).

However, in this histogram, the user may be able to identify a few terms that may be more relevant to their information need. The search results can
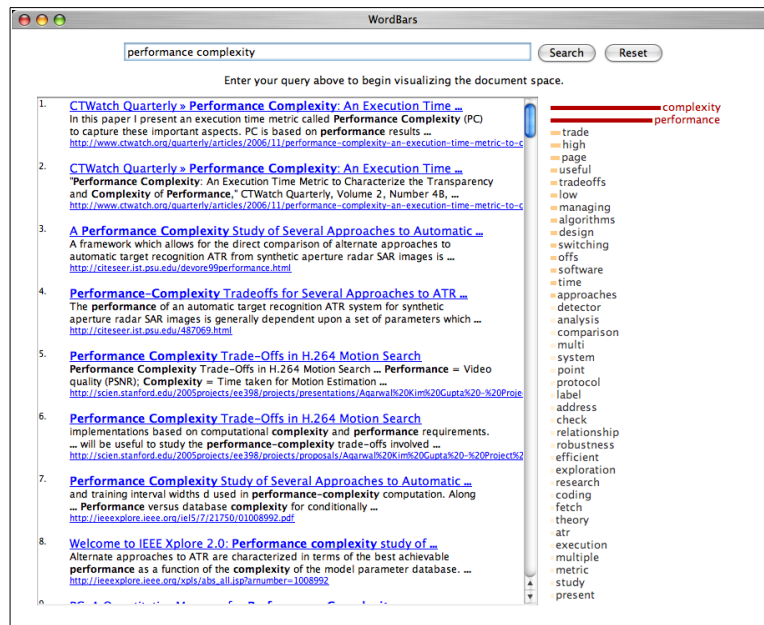
Figure 4.28: Starting with an incorrect term in the query, WordBars produces a term frequency histogram with a mix of relevant and non-relevant terms.
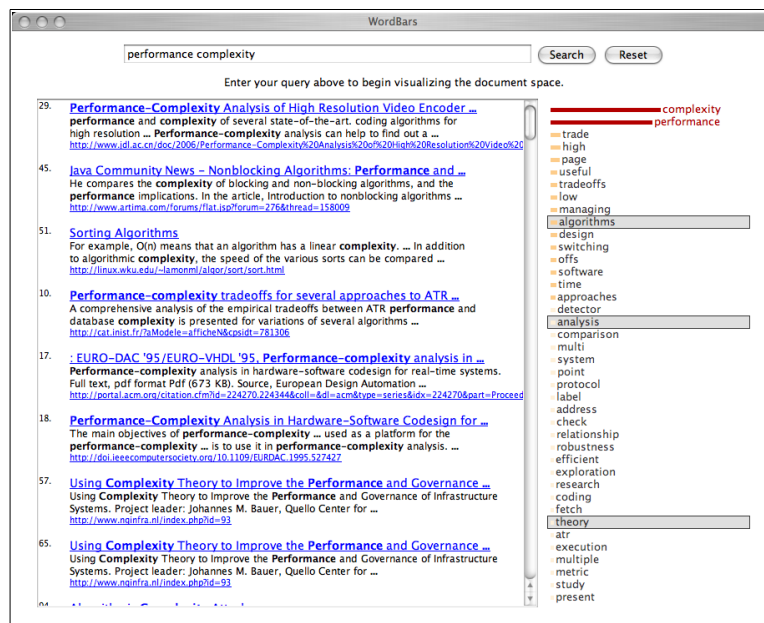
Figure 4.29: Potentially better terms can easily be explored based on how they are being used in the top search results.
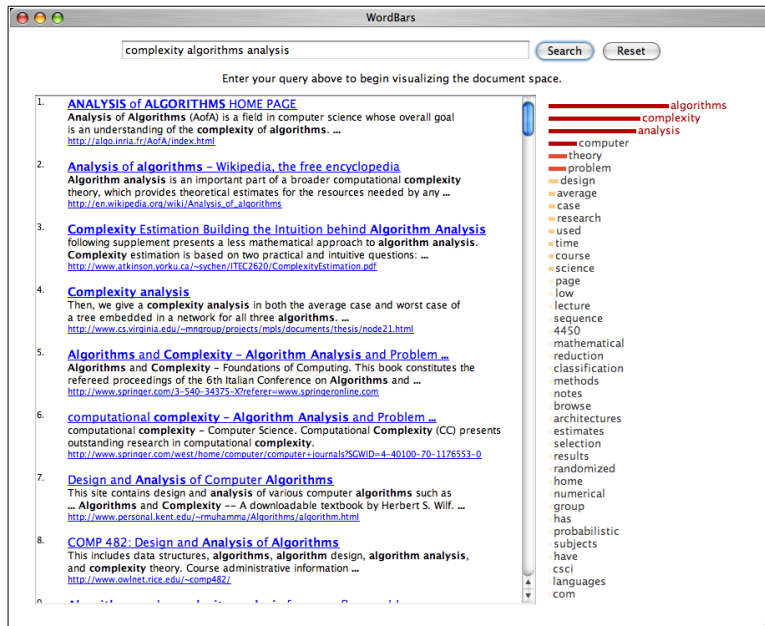
Figure 4.30: The user can replace the incorrect terms with better alternatives.
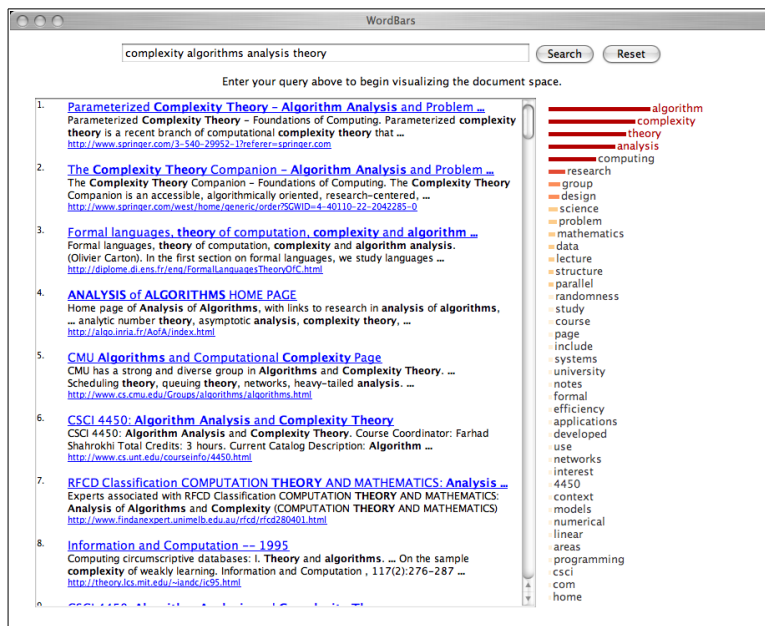
Figure 4.31: Additional terms can be added, resulting in an even more specific query.

be re-sorted based on the use of some of these terms, allowing the user to evaluate their relevance (Figure 4.29). Once the user identifies a few relevant terms, they may wish to start crafting a better query. In this example, the user may decide to add the term "algorithms", as well as replace the term "performance" with "analysis" (Figure 4.30). Viewing the subsequent term frequency histogram and the corresponding Web search results, they may wish to further refine their query by adding the term "theory" as well (Figure 4.31).

In this example, the benefits WordBars provides is clear. Even though the user began with an incorrect term, there were enough marginally relevant terms in the top search results to provide some information to the user regarding potentially better terms. While it is still up to the user to make the proper selection from the term frequency histogram, the histogram supports the recognition of relevant terms, rather than requiring the user to recall the terms.

Exploratory Search

Exploratory search is a technique used by many Web searchers, both for finding out what information exists for a topic to which they are unfamiliar, and for finding out what is new in a topic to which they are familiar. In either case, exploratory searches seldom have a specific goal at the beginning of the search, although they may develop one as they explore the topic of interest.

The features of WordBars provide support to the users in exploratory Web search processes. Starting with an initial query, the users can readily identify terms of interest within the WordBars histogram. Selecting some of these terms will move the documents that make use of them to the top of the list, making it easy for the user to decide whether this is an aspect of the search results that they would like to explore further. These terms can easily be

added to the query and a new set of search results generated. In this manner, users of WordBars are able to both explore the current search results, as well as refine their queries as they seek to explore and discover information about their topic of interest.

Suppose the user starts with a query such as "parallel operating systems", with the goal of exploring this general topic to discover what is there. Although there are many relevant documents returned as a result of this search (see Figure 4.32), the user is more interested in exploring specific aspects of this topic. As such, the user may choose to explore how specific terms such as "distributed" and "processing" are used in the top search results (see Figure 4.33).

If these re-sorted results look like a promising avenue to pursue further, the user may choose to add these terms to their query and review the subsequent search results (see Figure 4.34). From here, the user may note the term "grid" in the term frequency histogram, and choose to add this to the term in place of "distributed processing", resulting in the exploration of a related, but different, aspect of the general topic (see Figure 4.35).

In this scenario, the term frequency histogram provides the user with an overview of the contents of the search results. By browsing this information, the user can identify specific aspects of the search results set, which can easily be explored further by refining the query.

4.3.4 Discussion

Following the classification of query expansion techniques discussed in Section 3.2, WordBars is an interactive technique based on the initial set of search results. However, as with VisiQ, WordBars is more accurately described as a query refinement technique, since it allows the users to additionally remove
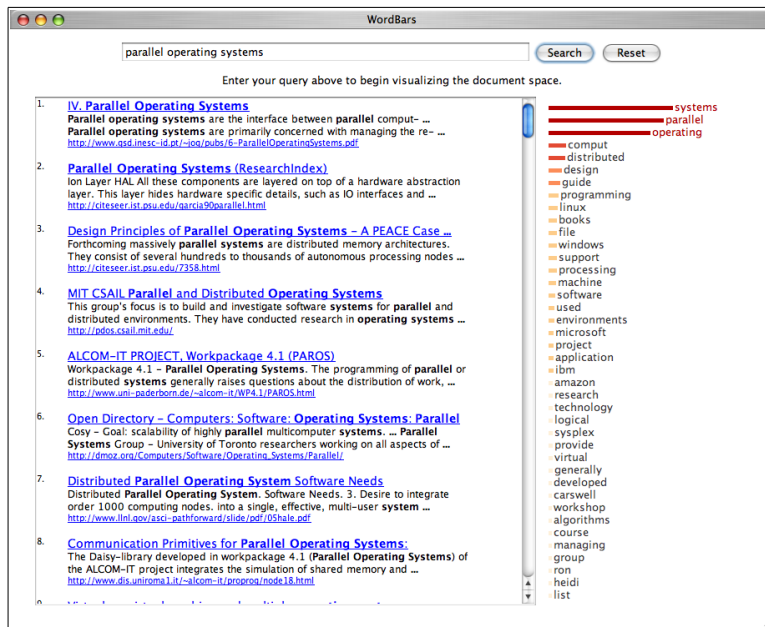
Figure 4.32: Starting with an exploratory query, many of the terms in the term frequency histogram generated by WordBars will be relevant to the general topic.
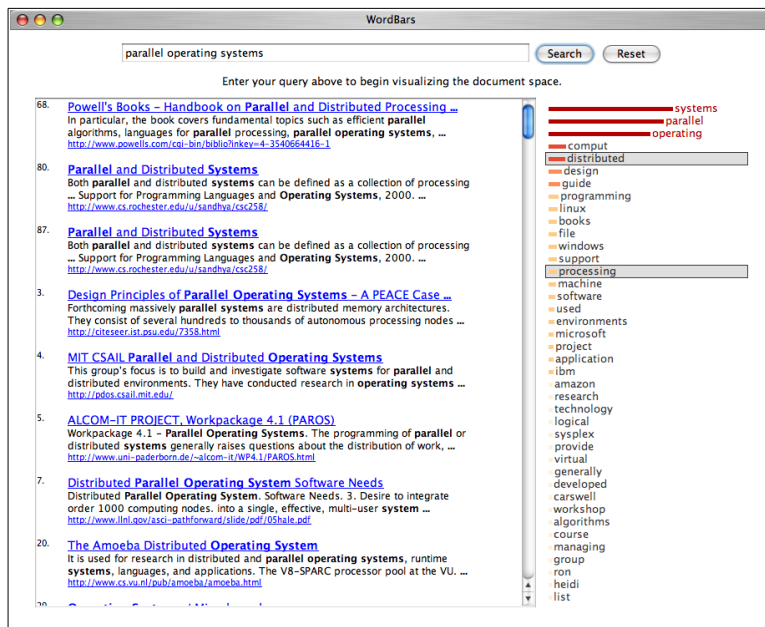
Figure 4.33: Terms of interest can easily be explored.
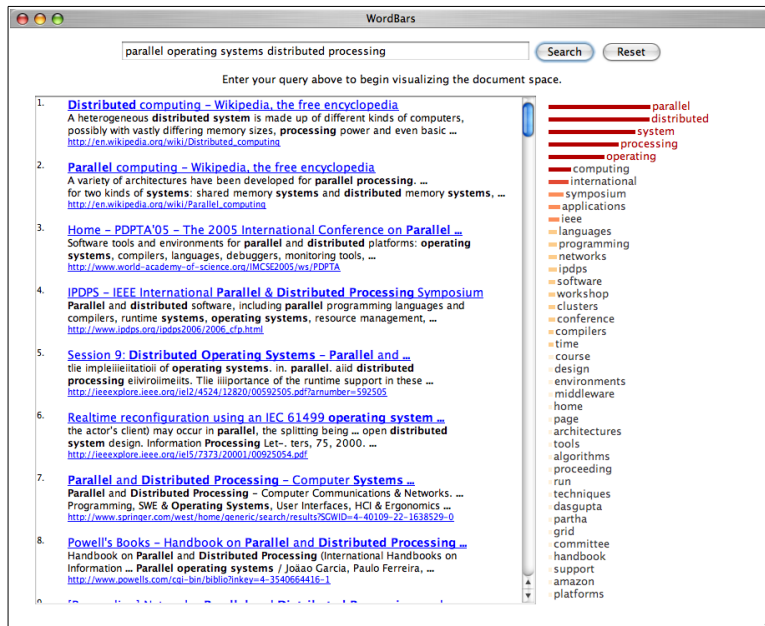
Figure 4.34: The query can be refined based on terms of interest identified in the term frequency histogram.

Figure 4.35: Alternate queries can easily be generated.

some or all of their original query terms in the process of crafting of a new query that accurately reflects their information needs.

One of the primary limitations of WordBars is that there is little ability to support the users in their exploratory Web search tasks when a very poor initial query is provided. If no relevant document surrogates are returned within the top 100 search results, then the ability to explore the use of these terms within the search results is of little value to the users. The terms that are common among these top search results will likely not be relevant to the user's information need, making it difficult for them to choose from the term frequency histogram. However, the lack of relevant terms in the term frequency histogram may indicate to the user that they need to start with a better initial query than the one provided.

Supposing that at least some of the document surrogates returned from the initial search are relevant, WordBars can be very beneficial in assisting the users in their information retrieval tasks. The term frequency histogram provides a visual indication to the users of the relative frequencies of the terms used in the top document surrogates from the search results. The users may explore the search results through re-sorting based on the terms they select as relevant to their information need. They can use these terms to further refine their query, commonly resulting in a more specific set of search results that can then be explored further.

The term frequencies in WordBars are generated from a subset of the actual document: the title and the snippet provided by the Google API. The title is often descriptive of the information within the document, and the snippet contains contextual information regarding the use of the query terms within the document. These both provide valuable information about the documents in the search results, and may even result in a better list of terms than if the

entire textual documents were considered.

Since only a simple pre-processing of the title and snippet are performed, it is possible for terms that are not meaningful to appear in the term frequency histogram. For example, the word "two" may appear somewhat frequently in the top search results for a given query, even though this word is not meaningful for search results exploration or query refinement. While it is possible to add such terms to the stop-words list, in some situations, these terms may be relevant and meaningful. As such, the stop-words list is limited to commonly used verbs, adverbs, pronouns, and prepositions, which are of little value in exploring the search results or refining the queries.

As noted previously, users may have difficulties choosing good terms to use for query expansion when presented in a list format [85, 57]. WordBars enhances the simple list-based format by providing a visual representation of the frequency of the terms, as well as an indication of which terms are present in the current query. The ability to re-sort the search results allows the users to see how potential query expansion terms are used within the top search results. This additional information allows the users to make informed decisions for query expansion that would not be possible when simply considering the terms in a list.

In the examples provided in the previous Section, the users are provided with a visual indication of the term frequencies, and are able to take advantage of their human intelligence as they use this information to interactively refine their queries. One of the key benefits of the term frequency histogram in WordBars is that it allows the users to recognize terms from the list, rather than having to recall relevant query terms for a given topic. *Recognition rather than recall* is a primary usability principle, suggesting that users should be able to see the information they need, rather than having to remember it [65]. In

WordBars, this allows the users to begin with an initial query, and use their recognition ability to explore and subsequently add or remove terms from their query, resulting in a refined query that did not require the user to remember the specific terms that are relevant to their information need.

Spence's Navigation Framework

Spence's navigation framework, as described in Section 2.4.1, describes the human activities involved in navigating within an information system [90, 91]. This framework is used to discuss the features and benefits of WordBars. The focus here is on the support of interactive query refinement through the term frequency histogram; a further discussion with regards to interactive search results exploration is provided in Section 5.4.2.

Browsing. While visually browsing the term frequency histogram, the user can see information about the commonly used terms within the top search results. The query terms can be readily identified in the histogram due to the font colour. The histogram can easily be seen to represent the relative frequencies of the terms with little or no training. By selecting terms in the histogram, the search results are re-sorted and the user can see how those terms are used within the top search results.

Formulate an Internal Model. As the user performs the browsing activities, they develop an internal model of the content being viewed. This internal model will be greatly influenced by the visual representation of the query term frequencies. Further, as the user explores the use of candidate terms within the term frequency histogram, the internal model will be supplemented by information they learn about the use of these terms within the search results.

The user's internal model will also be influenced by previous internal models generated for searches on the same or similar topics. Over time, the user will be able to learn which terms represented in the term frequency histogram are of value for their common search topics.

An important aspect in the development of an internal model for a given search is that the user is able to recognize terms within the term frequency histogram, rather than needing to recall them while constructing the query. This recognition ability provides a link between what they see in the externalization of the data (i.e., the visual representation), and their previously constructed internal models. Without the externalization of the data, the user would be required to recall their previously constructed internal models without any support from the system, requiring significant cognitive effort.

Interpret the Information Before the user can act upon the information they have gained through the use of the WordBars system, they must make a proper interpretation of their internal model, as well as what they are currently viewing in the system. The colour coding of the query terms versus the other terms found in the top search results supports the proper interpretation of which terms are currently in the query. Further, the width and colour coding of the bars within the term frequency histogram support the user in interpreting the relative frequencies of the terms. As the user explores how candidate terms are being used in the top search results, they are able to make effective interpretations of the relevance of these terms with respect to their current information need.

Formulate a Browsing Strategy. When formulating a browsing strategy, a number of different sources of information can be used: the internal model,

the visual representation being viewed, or a new idea the user has had. The resulting browsing strategy may be part of planned strategy with a specific goal in mind, or an opportunistic strategy based on something the user has thought of or seen.

Based on the interpretation of the internal model, the user may choose to pursue a browsing strategy consisting of refining the query based on what they have learned (planned strategy), or conduct a new exploratory search on a topic that is not relevant to their current goal, but is based on one or more terms they have discovered (opportunistic strategy). While viewing the WordBars user interface, the user may wish to explore a specific document in more detail (planned strategy), or may view a specific document that is not relevant to the current goal, but is of interest to the user none-the-less (opportunistic strategy). Further, as with VisiQ, the user may have some idea for a new search (opportunistic strategy), or for a new term to add to their current query (planned strategy).

Norman's Stages of Action Framework

In Norman's stages of action framework, two gaps exist between the users of a system and the system itself: the gulf of evaluation and the gulf of execution [68, 69]. The steps that users take as they cross these gaps are described in more detail in Section 2.4.2. As with Spence's navigation framework, this framework provides a useful mechanism by which the interactive query refinement features of WordBars can be discussed.

Goals   Although the overall information retrieval goal is to find a number of documents which are relevant to the user's information needs, a supporting goal of crafting a query that accurately reflects the user's information needs

must first be fulfilled. The system goal is to use the features of WordBars to craft a query that captures the user's information needs. Note that once this system goal is achieved, WordBars also supports a system goal of using the system to explore the search results and find specific relevant documents. This additional system goal is discussed with respect to WordBars support of interactive search results exploration in Section 5.4.2.

Gulf of Evaluation   During the perception stage, viewers of WordBars perceive the low level objects in the interface, and group these together into higher-level objects such as the bars and terms in the term frequency histogram. The regular shapes of the histogram bars, and the proximity of the corresponding terms support the user in making the perceptual connection of the terms to their histogram bars.

In the interpretation phase, the user identifies the histogram bars as representing some sort of value for the corresponding term. The representation of the value is provided both through the width of the bar, as well as the colour coding of the bar. This redundant coding is used to ensure a fast and accurate interpretation of the information. Although the interface does not explicitly state the meaning of the histogram values, logical interpretations include frequency, value, and importance. As users move beyond the novice user stage, prior experience will support the proper interpretations of this data.

The user will also be able to interpret that the there is a difference between the terms which use a red font and those that use a black font. Since the users likely just finished typing in their initial query terms to start the search, it requires little cognitive effort to recognize that the red terms are the query terms. This interpretation is further supported by subsequent use of the

system. Although the term frequency histogram does not provide much information to the user as to how it can be used, with a small amount of training, the user can make the proper interpretation that the histogram bars can be clicked and double-clicked.

The final stage to overcome in the gulf of evaluation is for the user to make a correct evaluation of the term frequency histogram with respect to their current system goal. The user can readily identify their query terms in the histogram, and can make sense of the meanings of the other terms as well. For the goal of crafting an accurate query, this evaluation stage will result in the identification of one or more terms that might be good candidates for the query.

Gulf of Execution   For the intentions stage, the user's system goal of crafting an accurate query is converted into an intention of adding one or more terms to the query that were identified while crossing the gulf of evaluation. Other intentions may include removing a query term, or selecting a term which to evaluate further in the search results list.

Since the action sequence required to add or remove a term is a simple double-click, little planning required. Further, since many users of the system will have had prior computer use experience, little cognitive effort is needed to double-click the mouse button.

After double-clicking a term and adding it to the query, the term is shown in the query box, and the term's font colour is changed to red to highlight that it is now in the query. This feedback provides the user with an indication that they have successfully executed the planned action sequence.

4.3.5 Summary

WordBars supports the users in interactively refining their queries by providing a graphical representation of the commonly used terms found in the top search results. This term frequency histogram allows the users to recognize terms that may be of value for their query, rather than requiring that they remember them. New terms can easily be added to the query, and poor choices can easily be removed. Refined queries can be submitted to the search engine, providing a new set of search results and a corresponding term frequency histogram. Further query refinement can then be performed based on the new query, resulting in an iterative improvement of the query.

Chapter 5

Interactive Search Results Exploration

5.1 Motivation

While it is clear that significant effort has gone into creating Web search engines that can index billions of documents and return the search results in fractions of a second [13, 23], the presentation of these search results has remained essentially unchanged since the early days of Web search. Commonly, the search results generated from a user's query are presented in a list-based format, with each document being represented by its title, a snippet containing the query terms in context, and its URL.

Even though this simple list-based representation provides the search results in a clear and effective manner for determining the relevance of individual document surrogates, it requires that each document surrogate be evaluated in turn, and to some degree, in the order provided. Further, there is little support for determining the overall properties of the search results, nor for manipulating and exploring the search results set.

It is common for Web searchers to encounter difficulties as they attempt to determine the relevance of the document surrogates that comprise Web search results lists. While some of these difficulties can be attributed to poorly crafted queries, even when the users provide a query that adequately describes their information needs, the search results are often a mixture of documents with varying degrees of relevance. This inability of Web search engines to provide highly relevant search results for users' queries can be attributed to the generality of the collection of documents being searched, the ambiguity of language, and the word mismatch problem [21].

Because most searches result in a combination of relevant and irrelevant documents to the users' information needs, the users are required to make relevance decisions on a document-by-document basis. This can be time consuming, and can result in the users giving up when a large portion of the search results considered are not relevant.

Studies evaluating the behaviour Web searchers have found that users seldom view more than three pages of search results [89, 93]. When the users are able to craft an effective query, they may be able to find enough relevant documents in the first few pages to satisfy their information needs. However, when vague or misleading queries are provided, it is more common for users to either re-formulate their query or give up, rather than continue to evaluate a set of seemingly non-relevant search results.

In evaluations of the search results from a number of vague queries, many highly relevant documents were found to be buried deep in the search results, even when there were many non-relevant documents in the first three pages of the search results. The interactive exploration of the Web search results can allow the users to access these relevant documents easily and efficiently, resulting in fewer non-relevant documents being viewed as the users seek to

fulfill their information needs.

In this Chapter, three prototype systems that have been developed to support the users' tasks of interactive search results exploration are discussed in detail. HotMap provides a visual representation of the query term frequencies in the top search results, and supports both a visual inspection and nested re-sorting of the search results to assist the users in finding relevant documents buried deep in the search results set. Concept Highlighter provides a list of concepts from which the user may choose. As a result of choosing the relevant concepts, the search results are re-sorted using a single-pass fuzzy clustering algorithm. The users can interactively select and un-select concepts, allowing them to explore the set of search results. WordBars provides a visual representation of the frequency of use of the top terms found in the search results set. Selecting terms from the term frequency histogram allows the users to re-sort the search results based on the use of the selected terms. Note that WordBars has already been described with respect to supporting interactive query refinement; the discussion in this Chapter is with respect to interactive search results exploration.

5.2 HotMap

To address the shortcomings of the list-based representation, and to promote interactive search results exploration, HotMap was developed as a means for visualizing and interactively manipulating Web search results. This meta-search system retrieves the top search results returned by the Google API [27] and presents these results in a compact visual manner that supports both visual information processing and user-directed exploration. Although the current prototype retrieves only the first 100 document surrogates from the

search results, the visual representation is compact and flexible enough to support the display of thousands of document surrogates in a single display screen.

The motivation for this work is based on a common method for evaluating the relevance of a document surrogate: identifying which of the query terms appear in the title and snippet. The list-based representation of the document surrogates often provides query term highlighting (i.e., bolding the query terms) which helps the users to identify their query terms, but still requires the users to read or at least recognize the terms. HotMap represents the query term frequencies in the form of a colour code on a heat scale. Multiple occurrences of a query term result in a dark red colour; fewer occurrences are represented by progressively lighter shades of red and orange. As shown in Figure 5.1, this colour coding allows the users to *see* the "hot" documents easily, and provides the inspiration for the name of this system: HotMap.

In HotMap, the search results are presented in a grid-based layout at two levels of detail: an overview map provides a compact representation of the top search results, and a detail window provides a focused view of approximately 20 document surrogates at a time. Coordinated scrolling between these two views allows the user to easily track the location of the detail window with respect to the overview map, as well as easily jump to a location of interest identified in the overview map. Of the document surrogate information provided by the Google API, only the title of the document is displayed in the detail window; tool tips are used to hide and dynamically show the additional details of the document surrogate, such as the snippet and the URL. To promote the manipulation and exploration of the search results, the users can perform nested sorting operations on the search results, based on the query term frequencies. Doing so automatically updates the order of the document
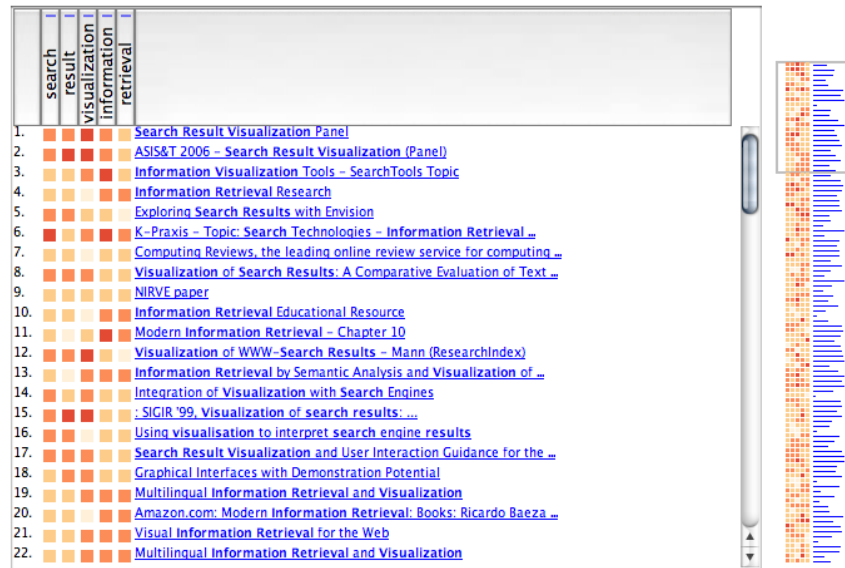
(a) HotMap screen shot

(b) magnification of overview map

Figure 5.1: The visual representation of the Web search results consists of a detail window and an overview map (a). The magnified view of the overview map (b) shows that the first document surrogate is "hot" with respect to the first two query terms and "warm" with respect to the fourth term; the second document surrogate is "hot" with respect to the third query term and "warm" with respect to the last two query terms; and the third document surrogate is "warm" with respect to the third query term and "hot" with respect to the last two query terms . These search results were returned from the query "search results visualization information retrieval".

surrogates in both the overview map and the detail window.

In the design of HotMap, the focus was on supporting the users as they evaluate and explore the search results for vague queries. For queries that are very specific, many highly relevant documents will commonly appear in the first few pages of the search results. In these situations, the list-based representations of the search results are rather effective. However, when the query is vague, the search results will commonly be a mixture of relevant and non-relevant documents, and will require the users to consider many non-relevant documents as they attempt to find relevant documents. Interactive search results exploration, as provided by HotMap, can assist the users in finding more relevant documents from vague search results.

5.2.1 System Architecture and Features

The system architecture of HotMap consists of four primary components. The *meta-search component* obtains the set of top search results for the current query from the Google API. The *document surrogate augmentation component* augments each document surrogate with information regarding the frequency of use of each of the terms in the current query. The *document surrogate visualization component* generates the visual representation of the document surrogates based on the query term frequency information. The *search results re-sorting component* re-sorts the search results based on the nested sorting selections made by the users.

This system architecture is illustrated in Figure 5.2, the details of which are provided in the remainder of this Section. The user interaction steps shown in this diagram are described in Section 5.2.2.
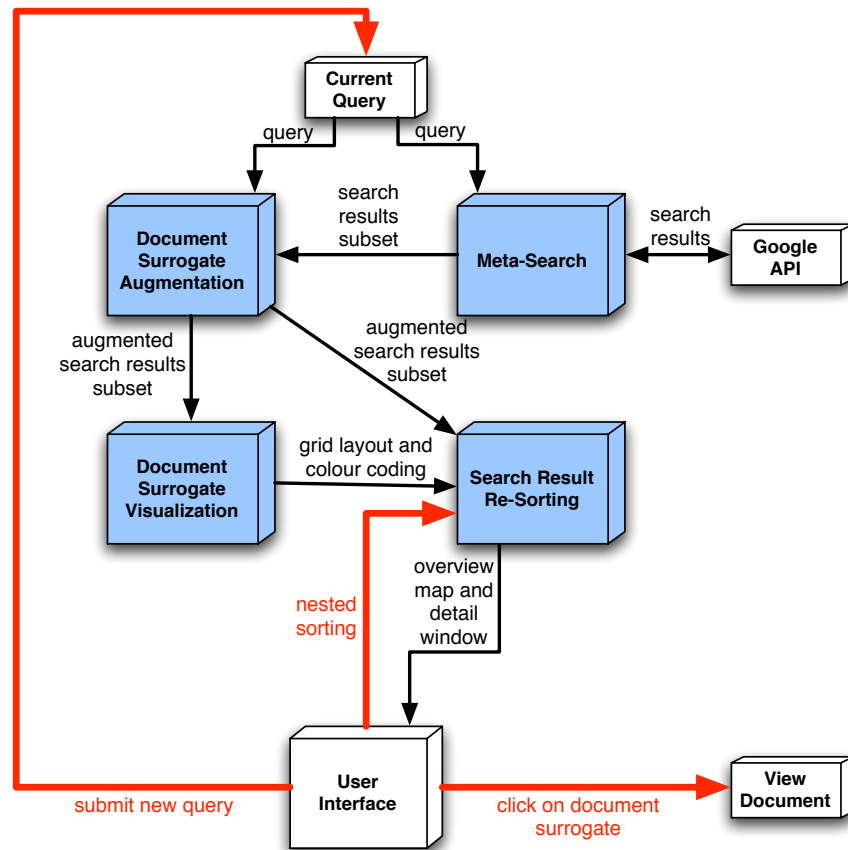
Figure 5.2: The HotMap system architecture consists of four primary software components (coloured in blue): meta-search, document surrogate augmentation, document surrogate visualization, and search results re-sorting. The interaction paths are illustrated by the thick red arrows.

Meta-Search

HotMap is a meta-search system that retrieves the top 100 document surrogates from the Google API [27] for the current query. Since the Google API only provides ten document surrogates at a time, it is the responsibility of the meta-search component to make ten different calls to the Google API in order to collect all 100 document surrogates required by the system. In order to avoid delays introduced by the multiple requests for search results, as each

set of ten document surrogates are received, these are passed to the document surrogate augmentation component for processing. Ultimately, the user will see the search results list in HotMap grow as the document surrogates are obtained from the Google API.

Document Surrogate Augmentation

HotMap augments the document surrogate information provided by the Google API by calculating an additional set of attributes for each document surrogate that represents the frequency of each of the query terms. These attributes correspond to a user's question of "how often do my query terms appear in these documents?".

As each search results subset is provided by the meta-search component, the following steps are followed for each document surrogate in order to augment the document surrogate with additional information:

1. The title and snippet for the document surrogate are combined together into a document descriptor using a bag-of-words approach.

2. Porter's stemming algorithm [74] is used to calculate the stems for each of the terms in the document descriptor.

3. The frequency of occurrence of each of the n query terms are counted in the stemmed document descriptor, resulting in frequency attributes $\{q_1, q_2, \ldots, q_n\}$.

4. The frequency attributes $\{q_1, q_2, \ldots, q_n\}$ are added to the document surrogate to produce the augmented document surrogate.

Note that matching the query terms to the terms in the document surrogate uses the stemmed versions of both the query terms and the terms in

the document surrogate. Matching based on these stems is more effective than exact word matches, since it takes into account different variations of the same root word.

Document Surrogate Visualization

The goal of any visualization technique is to support the users in *seeing* the information rather than *reading* the information. This can be a difficult goal to meet for information that is inherently textual, such as Web search results. However, the query term frequency attributes present in the augmented document surrogate can be easily represented in a visual manner.

While some argue in favour of three-dimensional layouts, not only are there problems with occlusion, but judging the relative positions of objects can be difficult [110]. Instead, HotMap uses a two-dimensional grid layout and colour coding to represent the document surrogate attributes. This results in a tile mosaic metaphor similar to that used in TileBars [32]. In the HotMap colour grid, each row represents a document surrogate in the search results, and each column represents one or more attributes or elements in the document surrogate.

Since the spatial position of an object and its colour can be perceptually separated, colour coding of the query term frequencies can be used without interfering with the spatial layout of the data [110]. Further, since colour is pre-attentively processed, this information is absorbed by the users far faster than if the users were required to read the numerical values [110].

The choice of a colour scale is not as simple as it might seem. Since it is necessary to represent an ordered sequence of values, a colour sequence that varies monotonically on at least one colour channel is required [102, 110]. A set of perceptually distinct colours on the red end of a red-green colour scale were

chosen to represent the term frequencies. This colour scale varies on both the luminance channel as well as the red-green colour channel. Visually, this colour scale appears to be a heat scale, resulting in higher frequency terms appearing hot, and lower frequency terms appearing neutral or warm. The colour scales used in HotMap were generated using the ColorBrewer application [12].

Search Results Re-Sorting

The search results re-sorting component provides the ability to re-sort the search results based on a nested sort operation requested by the user. Initially, when this component receives the augmented search results subset and the corresponding grid layout and colour coding information from the other components, no re-sorting is performed. As such, the search results are presented in the order provided by the underlying search engine.

When a user specifies a nested sorting operation, this component uses the query term frequency attributes in each augmented document surrogate as the basis for the re-sorting. A nested sorting operation will consist of an ordered list of the query terms along with the direction in which to sort for each query term in the list. Starting with the first query term in the list, the search results are sorted based on the corresponding frequency attribute, and in the direction specified (ascending or descending). For each additional query term in the list, where there are ties in the sort order, the document surrogates are sorted based on the corresponding frequency attributes.

Coordinated Views at Different Levels of Detail

As illustrated in the screen shot in Figure 5.1a, two different levels of detail of the search results are simultaneously displayed in HotMap. The detail window shows a fraction of the search results set at a time, and provides links to access

the target documents. The overview map displays the first 100 search results in a compressed format by showing the query term frequencies and an abstract representation of the document surrogate title. Coordination between these views allow the user to both investigate the document surrogates in detail, as well as gain insight into the features of the entire set of search results displayed.

Detail Window   In the detail window, a compact representation of approximately 20 document surrogates is provided. The first column contains the document surrogate number, allowing the user to easily identify the degree of importance placed on each document surrogate by the underlying search engine algorithms, as well as supporting data collection in the user evaluations. The next n columns hold the frequencies of the query terms, represented by a colour value. The final column contains the title of the document and a tool tip to access the hidden information including the snippet and the URL.

The users can click the mouse button on any title in the detail window to access the corresponding document. Doing so opens a new browser window and loads the selected document. In order to indicate to the user that this is an available option, the document titles are underlined and coloured blue. After a title has been clicked and the document has been opened, the colour of this title is turned to purple to indicate that the document has been viewed. The underlining and colour selection follow the de facto standard for new and visited links in a Web page.

Overview Map   The primary goal in the design of the overview map was to provide an indication of the features of ten pages worth of search results (i.e., 100 document surrogates) in a single compact representation. This was challenging since even for a high resolution display, there is a relatively small

number of pixels available to represent this information. Therefore, an abstract representation of the Web search results is necessary.

While many of the standard display resolutions have more horizontal space than vertical space, the overview map is displayed in a vertical orientation. Since the search results are represented in a vertical layout in the detail window, coordinating the orientation allows the users to more easily make visual connections between the two representations.

Given the small amount of vertical pixels available in many computer displays, and the large amount of search results to represent, a choice was made to use only three vertical pixels to represent each document surrogate, with a one pixel gap between adjacent objects. The result is a compact representation that consumes only 400 vertical pixels of screen space in the representation of 100 Web search results (see Figure 5.1).

Even though a small number of pixels are used to represent each document surrogate, it is still important to convey information about the features of the search results. The overview map provides a three by three pixel block to visually represent the frequency for each term in the user's query. In order to make it easier for the users to make a visual connection between the overview map and the detail window, lines were drawn in the overview map whose lengths are relative to the lengths of the document surrogate titles. These lines use the same colour as the titles in the detail window. When the link colours change in the detail window as a result of the users selecting to view specific documents, a corresponding colour change is made in the overview map for the selected document surrogates.

The end result is that the overview map provides a "zoomed out" view of the search results. The colour coding of the frequency, the length of the title line, and the link colour of the title line all provide landmarks allowing the

user to recognize the features of the search results and remain oriented within these visual representations. This consistency between the overview map and the detail window makes comparisons between the views easier for the users to perform [7].

Coordination of Views   When multiple views of the same data at different levels of detail are provided, as in HotMap, it is very important to ensure that the views remain coordinated, and that this coordination is self-evident. Synchronized scrolling provides the primary means for coordinating the detail window and the overview map. In addition, perceptual cues are provided to make the relationships between these views apparent to the user [7].

In order to aid the user in remaining aware of the coordination between the detail window and the overview map, a scroll box is placed in the overview map to indicate which document surrogates are currently being viewed in the detail window (see Figure 5.1). As the users scroll in the detail window, the location of the scroll box is automatically updated.

To support the users' task of visually exploring the search results within the overview map, clicking or dragging the mouse on any location in the overview map automatically relocates or "jumps" the scroll box to that location and scrolls the detail window to display the corresponding document surrogates. This ability to "jump" to locations of interest is valuable in HotMap, where "hot" documents can be easily identified deep within the search results simply through a visual inspection of the overview map.

Having the line length and colour in the overview map coordinated with the document title length and colour in the detail window provides visual landmarks to the user. This, together with the scroll box, allow the users to see the relationships between the document surrogates they are viewing

Table 5.1: User interaction features in HotMap.

Object	Operation	Result
overview map	click	Jumps the document list to the selected location.
overview map	drag	Scrolls the document list to the selected location.
term header	click	Re-sort the search results based on the term frequency.
term header	control-click	Re-sort the search results based on a nested term frequency.
document surrogate	click	Open the document in a new browser window.
document surrogate	hover mouse	Show the snippet and URL in a tool tip.

in the detail window and the 100 document surrogates in the overview map, and remain oriented within the visual representations as they explore the Web search results.

5.2.2 User Interaction

As the search results subsets are obtained from the Google API, and augmented with the query term frequency attributes, this information is displayed in the detail window and overview map. The user does not need to wait for the entire collection of document surrogates to be retrieved; they can begin interacting with the system immediately. The operations for interactively evaluating the search results are depicted by the thick red arrows in Figure 5.2, and are described below. A summary of the operations that can be performed using the overview map and detail window of HotMap are provided in Table 5.1.

Visual Inspection

By providing an abstract graphical representation of the top search results in a single view, the overview map supports a visual inspection of the entire search results set retrieved (i.e., 100 document surrogates). The HotMap colour coding allows the users to visually identify "hot" document surrogates - those that make frequent use of one or more of the terms in the user's query.

When such a document surrogate or group of document surrogates is identified, the user can easily scroll the detail window to the general vicinity of these document surrogates through a simple click or drag of the mouse in the overview map. A visual indication of this scroll is provided by the re-location of the scroll box over the point of interest.

The user can then direct their attention to the detail window to consider the document in more detail. Because the colour codes used in the overview map and the detail window are the same, the document surrogates identified in the overview map can readily be located in the detail window.

Nested Sorting of Search Results

The search results can be further explored through a nested sorting feature in the detail window. This feature allows the users to interactively re-sort the search results based on the frequencies of query terms they deem to be important to their information need.

This feature is initiated by clicking on the header of the column that corresponds to the query term by which the user wishes to sort Doing so will immediately re-sort the search results based on the selected query term. This

re-sorting is reflected both in the detail window and the overview map. Clicking a second time sorts the search results in the reverse order, with those document surrogates that make the least frequent use of the query terms moved to the top. Clicking a third time removes the sorting. A small arrow at the top of the column header indicates the current state of the sorting.

In addition to this single-query term sorting, a nested sorting can be initiated by holding down the control key while clicking on the column headers. The query term selected first is given the highest priority, followed by the second, third, and so on. The status of the sort is indicated by arrows of decreasing sizes for the selected query terms.

All the nested sorting features can be removed by clicking on the column header above the title of the document surrogates. This returns the search results list to the order provided by the Google API.

Deciding Relevance

An important step in using any interface that represents search results is to allow the users to decide the relevance of the document surrogates before choosing to view the corresponding document. This is supported in the detail window.

The title of each document surrogate is provided, along with the snippet and URL, which are accessed via a tool tip. This compact representation allows many document surrogates to be viewed at the same time, supporting scanning of the titles. Access to the additional information is provided as needed.

When a document surrogate is selected to be viewed (i.e., clicked), the document is opened in a new browser window. When the user returns to the HotMap system, the link colour for the selected document indicates that it

has been viewed. This feature ensures that the user is made aware of which documents have been viewed, and which have not. Being able to identify which documents have already been viewed is vitally important for any system that supports search results re-sorting, since after re-sorting the users need to be informed of which document surrogates they have already considered.

5.2.3 Example

In order to illustrate the benefits HotMap provides to the users as they seek to interactively explore their Web search results, an example is provided. When a user submits a query to HotMap, the search results are retrieved from the Google API and are added to both the detail window and the overview map as they are received. Figure 5.3 shows the status of the HotMap system after submitting the query "search results visualization information retrieval".

The user can scan the visual representations of the query term frequencies for the search results both in the detail window and the overview map. "Hot" documents can easily be identified at both levels of detail. If such a document or group of documents is found in the overview map, the user can easily jump to that location by clicking in the vicinity of the documents of interest in the overview map (see Figure 5.4).

Alternately, the user may wish to re-sort the search results based on the importance of individual query terms with respect to their information need. In this example, the user may wish to perform a nested sort first by "visualization", then "information" and "retrieval". The result of such a re-sorting of the search results is displayed in Figure 5.5.

Since only the titles of the documents are displayed in the detail window, the users must hover the mouse pointer over the titles to activate the tool tip which contains the snippet and URL of the document. This is illustrated in
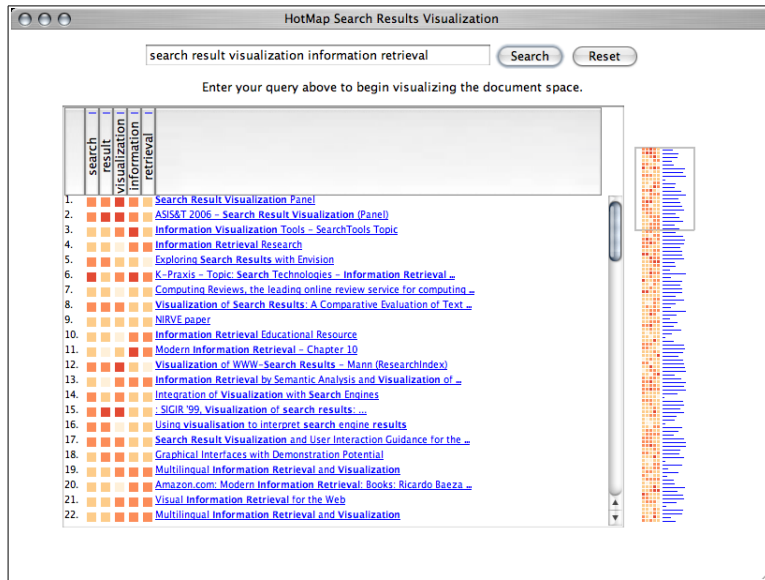
Figure 5.3: The initial search results are presented at two levels of detail in HotMap.
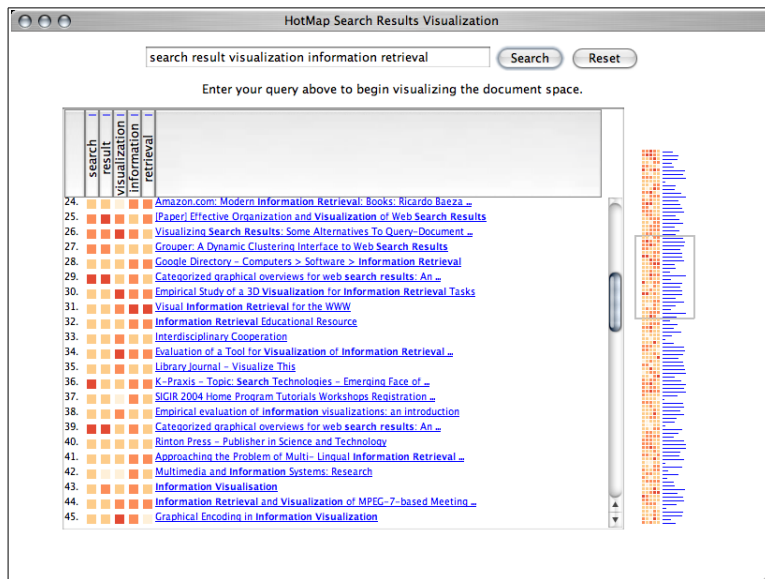
Figure 5.4: The overview map can be visually inspected to identify documents of interest.
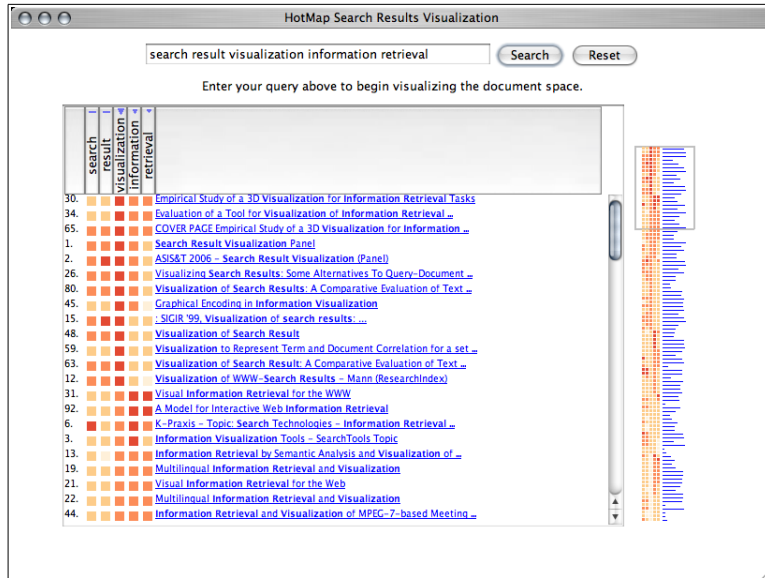
Figure 5.5: A nested sorting of the search results can bring relevant documents buried deep in the search results to a more prominent location.
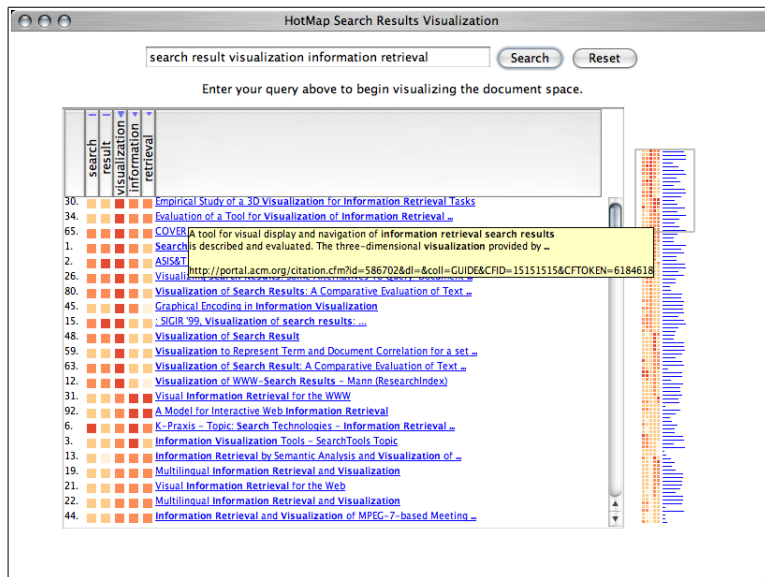
Figure 5.6: The compact representation in the detail window supports browsing; the details can be easily accessed via tool tips.

Figure 5.6. A video showing the HotMap system in use is available on the author's Web site[1].

5.2.4 Discussion

Wise et al., noting that "the need to read and assess large amounts of text that is retrieved through even the most efficient means puts a severe upper limit on the amount of text information that can be processed by any analyst for any purpose" [113], gave a very clear motivation for investigating other methods for presenting information retrieval results in non-textual manners. Providing the results of a search in a visual manner may allow this upper limit to be exceeded.

However, for a Web search interface implemented as a meta-search, there are additional constraints beyond what would be present for a traditional information retrieval system. In addition to dealing with a collection that includes billions of documents, access to the textual contents of the individual documents is not provided with the search results, and there is an expectation from the users that the results be displayed in near real-time [28]. For these reasons, it was necessary to restrict the visualizations to the information that is present in the document surrogates that are provided in the search results.

Although the process of counting the occurrences of the query terms in the document surrogates may seem simplistic, it has its basis in how some users evaluate the search results presented in the common list-based representation. That is, users will often try to find which documents make use of the terms in their query. While query term highlighting in the title and snippet do support this task, visualizing this information can be more effective since it allows the users to *see* the information rather than having to *read* the information. This

---

[1]`http://www.cs.uregina.ca/~hoeber/HotMap/`

has the added benefit of allowing the document surrogate to be presented in a far more compact manner. Since the users may still wish to see how their query terms are used in the document, access to the snippet is provided via a tool tip.

The ability to re-sort the search results based on query term frequencies support the users' tasks of manipulating and exploring the search results in order to more easily identify relevant documents. This, coupled with the ability to identify documents of interest in the overview map provides support for the interactive exploration of the search results.

Spence's Navigation Framework

Spence's navigation framework, as described in Section 2.4.1, describes the human activities involved in navigating within an information system [90, 91]. This framework is used to discuss the interactive search results exploration features of HotMap.

Browsing. While browsing the HotMap user interface, there is a large amount of information for the user to view and absorb. The visual representation of the query term frequencies for the entire set of search results can be seen in the overview map. In the detail window, specific document surrogates can be browsed (with the details being shown as the user hovers their mouse of the title text) and selected (with the document being displayed in a new window). As the user selects terms to use for re-sorting, the use can identify patterns of use of their query terms.

Formulate an Internal Model. As the user browses the top search results in HotMap, they develop an internal model for the search results set. The

visual representation of the query term frequencies supplements the user's internal model of each document surrogate with information on how frequently the query terms are appearing in the title and snippet. As the user views specific documents, this information contributes to the development of the internal model of the search results.

The internal models developed from previous searches support the development of the internal model for the current search. For example, as the user learns the meanings of the colour codes, and the mapping between the overview map and the detail window, this information becomes part of the internal model for the current search. What the user has learned about the features of the HotMap interface will support the development of future internal models while using the system.

As the user browses the interface, they may choose to view a specific document they feel may be relevant to their information need. Doing so opens the document in a new window, allowing them to read and evaluate the document in detail. The information they gain by viewing the documents further contributes to the development of their internal model, both for the specific document and for the search results set.

Interpret the Information. A fundamental step in a user being able to act upon the information they have gained through using an interface is being able to make a correct interpretation, based on what is in their internal model and what they are currently viewing. Although the colour coding used to represent the term frequencies may not be easy to decode into exact frequency values, it can be easily interpreted in terms of relative frequency of use. The users can interpret the potential importance of a document based on the query term frequency information they have gained through browsing the interface.

As the user views specific documents, their internal model will begin to include specific information that is relevant to their information need. It is in this interpretation activity that the user will decide whether they have found enough information to fulfill their information need.

Formulate a Browsing Strategy.   When formulating a browsing strategy, a number of different sources of information can be used: the internal model, the visual representation being viewed, or a new idea the user has had. The resulting browsing strategy may be part of planned strategy with a specific goal in mind, or an opportunistic strategy based on something the user has thought of or seen.

Based on the user's current interpretation of their internal model, they may choose to re-sort the search results based on one or more of their query terms, or decide that they have a chosen a good method for re-sorting the search results and wish to start evaluating the search results (planned strategies). Alternately, the user's internal model may have been influenced by the appearance of a term that is not relevant to their current task, resulting in them choosing to alter their query (opportunistic strategy). While viewing the interface, the user may choose to view a specific document that is relevant to their current task (planned strategy), or view a document that is not relevant to their current task, but is still of general interest to the user (opportunistic strategy). The user may have a new idea for something they wish to search (opportunistic strategy), or may browse the search results for a specific term or phrase they have thought of (planned strategy).

Norman's Stages of Action Framework

Norman's stages of action framework describes two gaps that exist between the users of a system and the system itself: the gulf of evaluation and the gulf of execution [68, 69]. The steps that users take as they cross these gaps are described in more detail in Section 2.4.2. As with Spence's navigation framework, this framework provides a useful means by which the interactive query refinement features of HotMap can be discussed.

Goals   The real-world goal for the user is to find a number of relevant documents that satisfy the user's information needs. In terms of a system goal, the user's task is to use the features of HotMap (visual inspection and nested re-sorting of the search results) in order to explore the search results and discover documents that are relevant.

Gulf of Evaluation   In the perception stage, the user is able to view the HotMap interface and identify the two primary interface components: the overview map and the detail window. In the overview map, the user perceives the coloured boxes and lines. The regular grid layout and variable length of lines (which represent the length of the titles of the documents) support the user in grouping the objects as rows of data. In the detail window, the user is able to perceive the table control elements as distinct from the coloured boxes and the other text, and is able to make the row-wise connection between the boxes and the text.

In the interpretation stage, the user is able to identify the objects they have perceived as specific types of display elements, such as table headers, document surrogates, term frequency indicators, overview map scroll indicators,

etc. Although a first-time user may not be able to make an accurate interpretation of the system, with a small amount of instruction and experience, the ability to accurately interpret the features of the system will be improved.

The table headers (containing the terms from the query) appear as buttons, providing an affordance that they can be clicked. The document surrogate titles appear as Web links, allowing the user to interpret them as clickable objects as well.

In the evaluation stage, the user attempts to make sense of what they are seeing with respect to their system goal of finding relevant documents. The user is able to identify the detail window as a place where document surrogates can be evaluated, the table header buttons as means for re-sorting the search results, and the overview map as a way of viewing the set of search results in a compact and abstract manner.

Gulf of Execution   Within the intentions stage, the user attempts to convert their system goal of finding relevant documents into an intention to do some action within the interface. This action may be to re-sort the search results based on the use of a specific query term; or it may be to scroll to some location of interest identified in the overview map. The user may wish to view the details of a specific document surrogate; or they may wish to view the actual document.

In order to realize the intention to use the system, the user must plan a sequence of actions, and then execute those actions. Since the user interface actions are simple clicks and drags, there is little cognitive effort involved in the action sequence stage and the execution stage. The only exception is the method for specifying a nested sorting. The user must hold down the control

key, while the column headers are selected in the order of the nested sort desired. Although this may requires some planning for novice users, experienced users have little difficulty in making such selections.

As the user executes an action sequence, the interface is instantly updated with the results of the action. When specifying a re-sorting of the search results, the user will be able to see that the search results are provided in a sorted order, both in the overview map and the detail window. This feedback indicates to the user the success of their action. Similar feedback is provided for the other types of interaction as well.

5.2.5 Summary

By providing a visual representation of the frequencies of the query terms within the top search results, HotMap supports the user in interactively exploring the search results. This exploration can occur through a visual inspection of the overview map, as the user seeks to identify documents that make frequent use of their query terms. Or it can occur through a nested re-sorting of the search results, based on the importance the users place on their query terms. The list of search results are provided in a compact format, supporting the scanning of the document titles. The snippet and URL can readily be accessed via tool tips, and the source document can readily be viewed by clicking on the title.

5.3 Concept Highlighter

A common method for assisting users in becoming more efficient as they evaluate the results of a Web search is to cluster the search results, resulting in documents that are similar to one another being grouped together [47]. In

such systems, the users can navigate the clusters in order to narrow down the search results and avoid clusters of non-relevant documents. In the best case scenario, the users will select the relevant clusters and view lists of document surrogates in which a large portion are relevant to the users' information needs.

One of the primary challenges in clustering is determining an appropriate name or description of the clusters. If this information does not correctly describe the document surrogates contained in the cluster, the users will either choose clusters that are not relevant to their information need, or will entirely miss clusters that contain relevant documents. Further problems with clustering Web search systems include an inability to select multiple clusters simultaneously, and a lack of ability to re-sort or re-order the documents within the clusters.

To address these drawbacks of clustering in Web information retrieval, Concept Highlighter was developed as a tool for visually exploring concept-based clusters in Web search results. This tool makes use of the concept knowledge base described in Section 4.2.1 in order to automatically generate a set of concepts related to the users' queries. The concepts generated from the concept knowledge base are used as the centroids for a single-pass fuzzy c-means clustering algorithm [10, 47] that is applied to the search results as they are retrieved via the Google API [27]. A list of concepts and a visual representation of the fuzzy membership scores allows the users to interactively select and change concepts of interest, and identify how this affects the clustering and re-sorting of the search results.

The primary focus in the development of Concept Highlighter was to support the users as they evaluate and explore the search results for vague queries.

For queries that are very specific, many highly relevant documents will commonly appear in the first few pages of the search results; the list-based representations of these search results are rather effective. However, when the query is vague, there is an opportunity to interactively re-order the search results to bring more of the relevant documents that are similar to the selected concepts to the top of the list.

5.3.1 System Architecture and Features

The system architecture for Concept Highlighter consists of five primary components. The *query space generation component* uses the current query and the concept knowledge base to generate a query space. The *meta-search component* obtains the set of top search results for the supplied query from the Google API. The *concept-based fuzzy clustering component* calculates fuzzy membership scores for each concept-document surrogate pair. The *clustering visualization component* generates a visual representation of the fuzzy membership scores. The *search results re-sorting component* re-sorts the search results based on the fuzzy membership scores.

This system architecture is illustrated in Figure 5.7, the details of which are provided in the remainder of this Section. The user interaction steps shown in this diagram are described in Section 5.3.2.

Query Space Generation

The first step in generating concept-based fuzzy clusters is to obtain a set of concepts associated with the users' queries. The source of the conceptual information is a *concept knowledge base*. This is the same concept knowledge base used to support users in constructing better queries using VisiQ (see
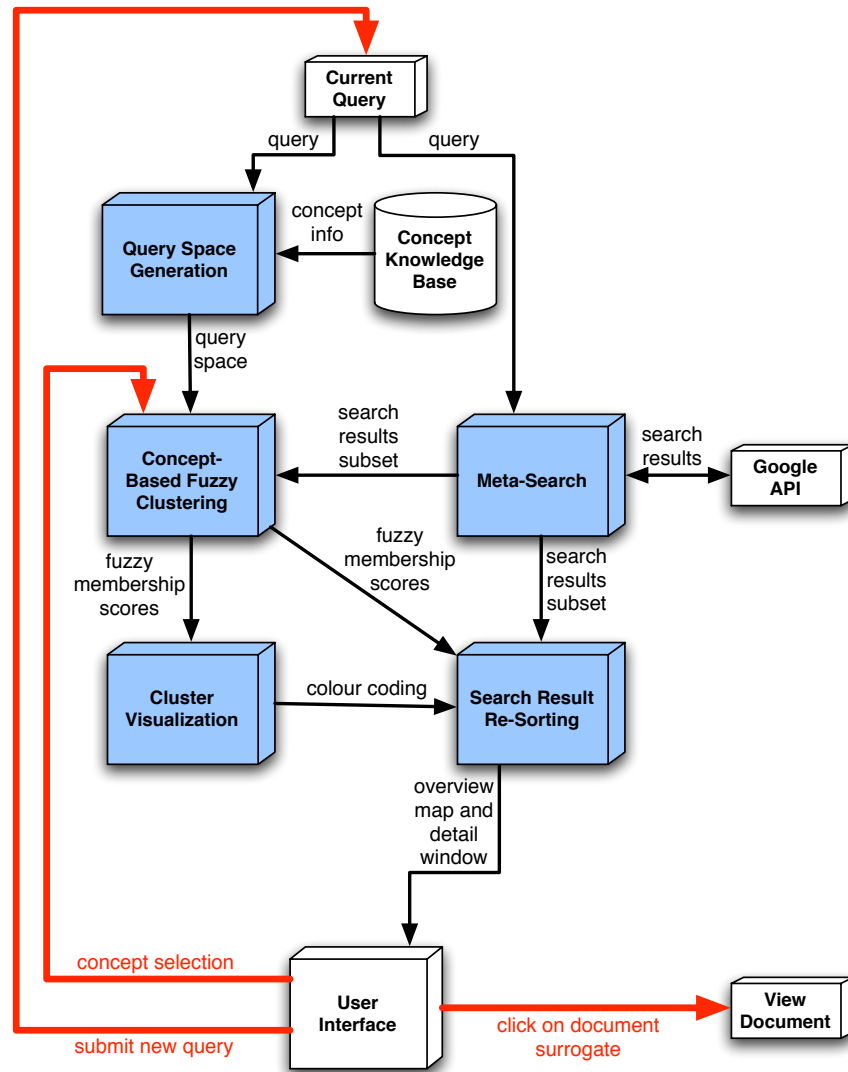
Figure 5.7: The Concept Highlighter system architecture consists of five primary software components (coloured in blue): query space generation, meta-search, concept-based fuzzy clustering, cluster visualization, and search results re-sorting. The interaction paths are illustrated by the thick red arrows.

Section 4.2).

The process for obtaining the concepts that are related to a user's query is similar to the process for generating the query space for interactive query expansion. Given a concept knowledge base $CKB = \{C, T, E\}$, and a simple query $Q = \{q_1, \ldots, q_n\}$ consisting of a set of query terms q_i, the query space $QS = \{C^*, T^*, E^*\}$ can be obtained by following the steps described in Section 4.2.2.

As with the technique used in VisiQ, the thresholds w_e and w_d, along with the term ratio TR control the size of the query space generated for a given query. Through experiments with using the concept knowledge base generated from the ACM Computing Classification System, these values are set at $w_e = 0.1$, $w_d = 0.05$, and $TR = 0.26$.

Meta-Search

As with HotMap, the meta-search component is responsible for interacting with the Google API in order to obtain the top search results for the current query. These search results subsets are provided by the Google API in groups of ten document surrogates at a time. As these subsets are obtained, they are provided to the other components within the system.

Concept-Based Fuzzy Clustering

In the work on VisiQ, the purpose for generating the query space for a user's query was to support an interactive query refinement process. In this work, the query space is used to identify potential cluster centroids that may be relevant to the users' information needs. For each concept C, a vector is created using the set of stems or terms T that were selected from the concept knowledge base. The weight of the edge E between the concept-stem pair is used to

set the magnitude of the concept vector in the dimension associated with the stem.

Therefore, as a result of this query space generation, a set of concept vectors $C^* = \{c_1, c_2, \ldots, c_m\}$ are easily obtained. If the total number of unique stems that were selected from the concept knowledge base is p, then the dimension of all vectors c_i ($i = 1 \ldots m$) is p. Further, the magnitude of the vector c_i ($i = 1 \ldots m$) on dimension j ($j = 1 \ldots p$) is given by the concept knowledge base edge weight between concept i and term j, $E_{i,j}$.

After the concepts have been obtained from the concept knowledge base, and the concept vectors have been created, the system can start processing the search results subsets being provided by the meta-search component. For each document surrogates retrieved, a single-pass fuzzy c-means clustering algorithm [10, 47] is performed. The title and snippet from the document surrogate are processed using Porter's stemming algorithm [74], and the frequency of each unique stem is calculated. These frequencies are used to generate vectors for each of the document surrogates. Although some argue against using term frequencies (TF) as the sole source of information in a text retrieval system [81], using other global information such as the inverse document frequency (IDF) is not feasible when the document surrogate vectors need to be generated as each document surrogate is retrieved (to achieve a near real-time Web information retrieval system).

Given a set of concept vectors $C^* = \{c_1, c_2, \ldots, c_m\}$ and a document surrogate vector d_i, the fuzzy membership of document surrogate d_i with respect to concept c_j is given by:

$$u_{i,j} = \frac{1}{\sum_{k=1}^{m} \left(\frac{sim(d_i, c_j)}{sim(d_i, c_k)} \right)^2}$$

In this calculation, the similarity between a document surrogate vector and a concept vector is given by the Euclidean distance metric [47]:

$$sim(x_i, x_j) = (\sum_{k=1}^{p}(x_{i,k} - x_{j,k})^2)^{1/2}$$

Normally, when evaluating the document surrogates, all unique stems would contribute to the construction of the document surrogate vector. However, since the distance calculations in this single-pass fuzzy clustering algorithm are always between concept vectors and document surrogate vectors, only the stems that are already present in the concept vectors need be considered. This reduction in the dimension of the document surrogate vectors results in an increase in the speed at which the fuzzy clusters are generated. In the prototype system, the fuzzy cluster membership scores are calculated as quickly as the underlying search engine (i.e., the Google API) can provide the document surrogates to the system.

While it is common to run the fuzzy c-means clustering algorithm in multiple passes, each time re-calculating the centroids of the clusters, in Concept Highlighter the algorithm is run in a single pass resulting in a fuzzy membership score for each concept-document surrogate pair. Although this may reduce the accuracy of the clusters, it ensures that the fuzzy clusters remain centred around the concepts.

Since the concepts represent the centroids of the clusters, the clusters can be named using the concept names. This is a valuable benefit since the names of the concepts are derived from the source knowledge upon which the concept knowledge base was constructed (in this case, the ACM Computing Classification System). Further, since the clusters always remain centred on the concepts, they are independent of the search results. Therefore, while two similar

queries will result in two different sets of search results, they will commonly result in a very similar set of concepts. This feature can assist the users learn which concepts are of interest to their general information seeking needs, and may be extended in the future to support personalized concept selection. Further, the use of stable concept names amortizes the investment the users make in learning their meanings over future searches [55].

Cluster Visualization

When visualization fuzzy clusters of document surrogates, the ultimate goal is to allow users to *see* the clusters without limiting their ability to view the entire set of document surrogates. The vector-based approach to clustering commonly leads to clusters being represented in two or three dimensional space [3, 113]. Similarly, when hierarchical clustering is used, a tree-based representation is common (e.g., Vivisimo [107] and Grokker [29]). The techniques used in Concept Highlighter depart from these methods, and instead use colour coding to represent fuzzy membership scores.

As with the colour coding in HotMap, since the spatial position of an object and its colour can be perceptually separated, colour coding of the fuzzy membership scores can be used without interfering with the spatial layout of the data [110]. In many cases, colour can be pre-attentively processed, allowing the information to be absorbed by the users faster than if they were required to read the corresponding numerical values [110]. While identifying specific values in the colour scale used in Concept Highlighter may not be pre-attentively processed, identifying a relative ordering as well as a few high values from many low values will be processed faster than reading the numerical values.

Since it is necessary to represent an ordered sequence of values, a colour sequence that varies monotonically on at least one colour channel is required

[102, 110]. A set of nine perceptually distinct colours on a yellow-green-blue colour scale were chosen to represent the fuzzy membership scores. This colour scale varies on all three colour channels: luminance, yellow-blue, and red-green. The ColorBrewer application [12] was used to select this colour scale.

Search Results Re-Sorting

The purpose of the search results re-sorting component is to re-sort the set of search results based on the concepts selected by the users. Initially, when the search results subsets are provided by the meta-search component, the document surrogates are simply stored and passed to the user interface in the original order (as provided by the Google API). As the fuzzy membership scores are calculated by the concept-based fuzzy clustering component, these are stored along with their corresponding document surrogate within this component.

When a user selects or un-selects a concept within the Concept Highlighter interface, this component uses the fuzzy membership scores corresponding to all the selected concepts as a means for re-sorting the search results. The fuzzy membership scores for all the selected concepts are added together, forming the basis by which the document surrogates are sorted. This re-sorting of the search results occurs in real-time, as the users make concept selections.

Coordinated Views at Different Levels of Detail

As illustrated in the screen-shot in Figure 5.8, two different levels of detail of the search results are simultaneously displayed in Concept Highlighter, in a manner similar to that used in HotMap. The detail window shows approximately 25 document surrogates at a time, and provides links to access the target documents. The overview map displays the first 100 search results in a
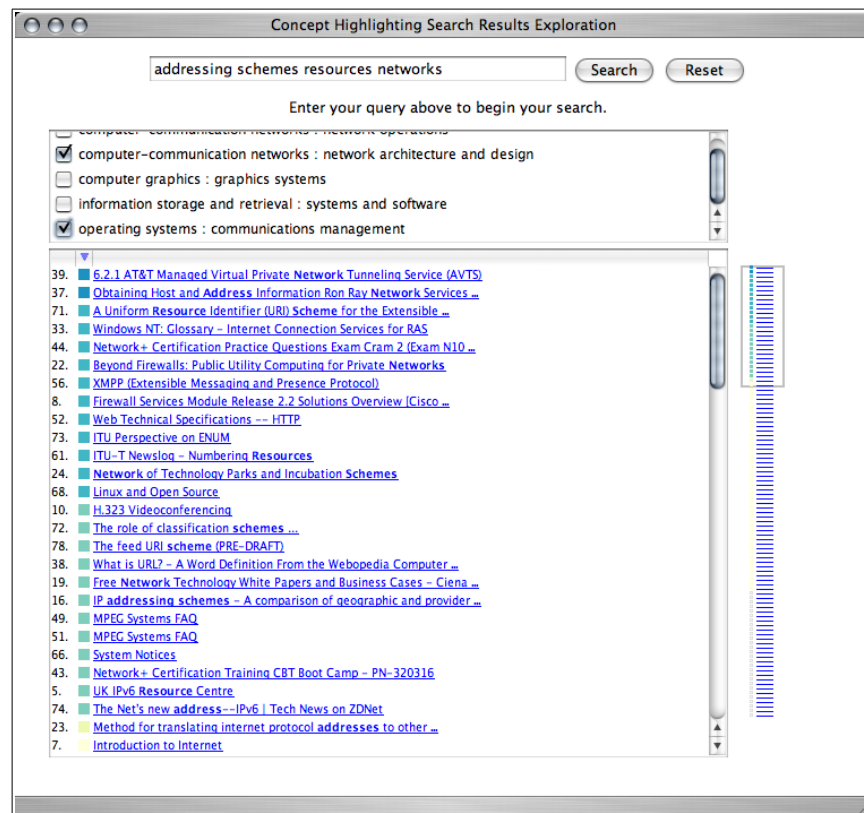
Figure 5.8: The visual representation of the Web search results consists of an overview map (right) and a detail window (left).

compressed format, showing the colour coding of the fuzzy membership scores and an abstract representation of the document surrogate title. Coordination between these views allow the user to both investigate the document surrogates in detail, as well as gain insight into the features of the entire set of search results displayed.

Detail Window and Overview Map

The detail window and overview map follow the same fundamental principles as those described for HotMap (see Sections 5.2.1 and 5.2.1). The primary difference is that instead of representing the query term frequencies visually,

Table 5.2: User interaction features in Concept Highlighter.

Object	Operation	Result
concept	check	Use concept in calculation of fuzzy membership scores. Re-sort search results.
concept	uncheck	Remove concept from calculation of fuzzy membership scores. Re-sort search results.
overview map	click	Jumps the document list to the selected location.
overview map	drag	Scrolls the document list to the selected location.
document surrogate	click	Open the document in a new browser window.
document surrogate	hover mouse	Show the snippet and URL in a tool tip.

these displays show the fuzzy membership scores for each document surrogate with respect to the selected concepts. The coordination between the detail window and the overview map in Concept Highlighter also operates in a manner similar to HotMap.

5.3.2 User Interaction

As with the other prototype systems developed, as the search results are retrieved and processed by Concept Highlighter, these are displayed in the detail window. This allows the user to start interacting with the interface to interactively explore the Web search results almost immediately. The primary operations the users can perform with the system are visually depicted by the thick red arrows in Figure 4.18, and are described below. A summary of the operations are provided in Table 5.2.

The primary method by which users of Concept Highlighter can interactively explore the search results is the selection of relevant concepts. A list of the concepts matched to the users' queries is provided at the top of the

display (see Figure 5.8). Beside each concept is a checkbox which can be used to enable or disable the corresponding fuzzy cluster.

When the user checks a cluster, the fuzzy membership scores for all the document surrogates are updated and visually represented in both the overview map and detail view. The user may check multiple concepts, the result of which generates a summation of the fuzzy membership scores corresponding to the selected concepts. Therefore, as multiple concepts are selected, the document surrogates that are nearer to both clusters are represented with a darker colour on the colour scale, indicating their higher fuzzy membership score.

As the documents that belong to the selected clusters are highlighted, the user may visually inspect both the overview map and the detail view to find relevant documents. Clicking on any location in the overview map will automatically scroll the detail view to that location. Therefore, the users can easily scan the entire 100 documents shown in the overview map, and jump to locations of interest based on the fuzzy membership score visualization. The user interaction within the detail window and the overview map is the same as that described for HotMap (see Section 5.2.2).

The interactive nature of concept cluster selection, and the re-sorting of the documents based on the total fuzzy membership score allows the users to interactively explore the search results. Using the fuzzy clusters as a means for organizing the search results in this exploration process can help to bring documents that are relevant to the users' information needs into focus, even if these documents are buried deep in the search results.

5.3.3 Example

In order to highlight the primary features of Concept Highlighter, an example is provided. When a user submits a query to the system, the concepts from

the query space are added to the concept list, and the search results retrieved from the Google API are added to both the detail window and the overview map. Figure 5.9 shows the status of Concept Highlighter after submitting the query "addressing schemes resources networks".

The first step for the user in interactively exploring the search results is to select one or more concepts that are relevant to their information need. Doing so will re-sort the search results based on the sum of the fuzzy membership scores for these selected concepts. Figure 5.10 illustrates the status of Concept Highlighter with the concepts of "computer-communications networks: network architecture and design" and "operating systems: communications management" selected.

Once the document surrogates are re-sorted based on the selected concepts, the user may evaluate the individual document surrogates for relevance. Tool tips can be used to access the snippet and URL for each of the document surrogates, as illustrated in Figure 5.11.

The user can click on any of the document surrogates in the detail window to open the corresponding document in a new window. When the user returns to Concept Highlighter, the previously selected documents will be apparent since the link colour will have changed from blue to purple (see Figure 5.12).

Note that at any time, the users may revisit their selection of relevant concepts to further explore the search results. For example, if the users find that there are many non-relevant documents near the top of the list, they can check or uncheck other concepts to achieve a new ordering of the search results. A video of Concept Highlighter in use is provided on the author's Web site[2].

---

[2]`http://www.cs.uregina.ca/~hoeber/ConceptHighlighter/`
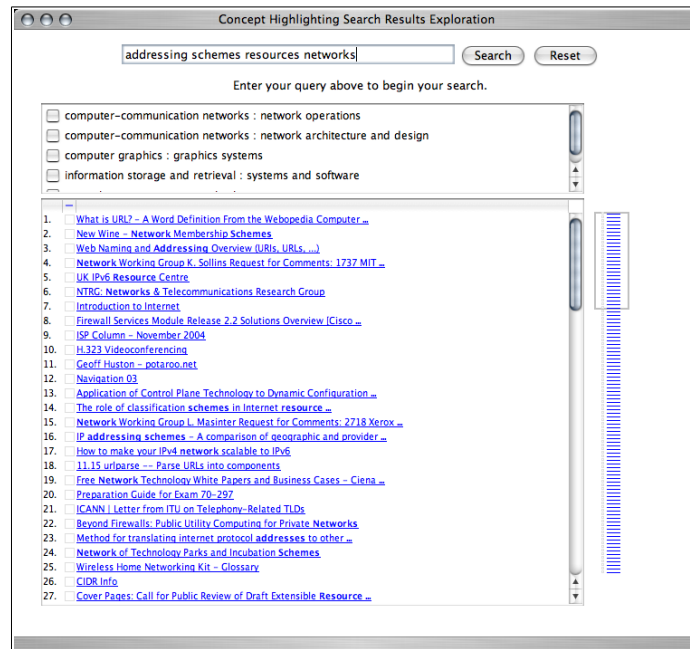
Figure 5.9: An initial query in Concept Highlighter generates a list of concepts and a set of search results.
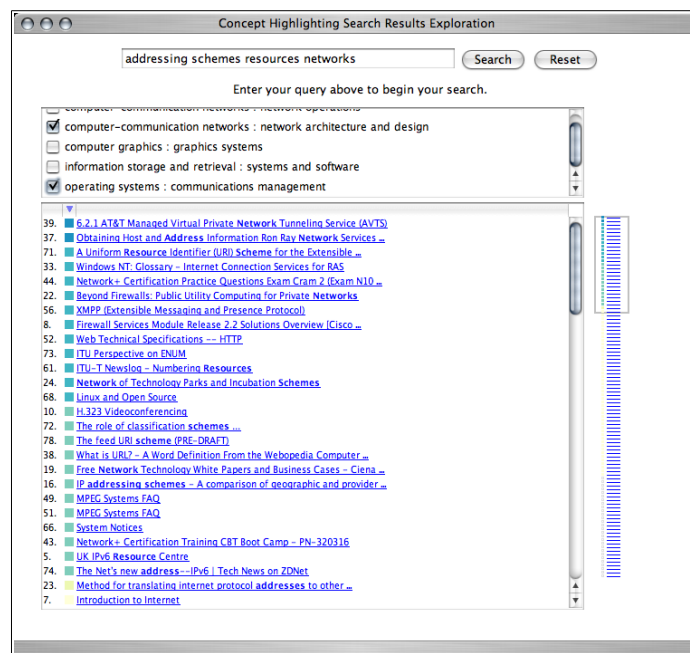
Figure 5.10: Selecting concepts results in a re-sorting of the search results based on the fuzzy membership scores.
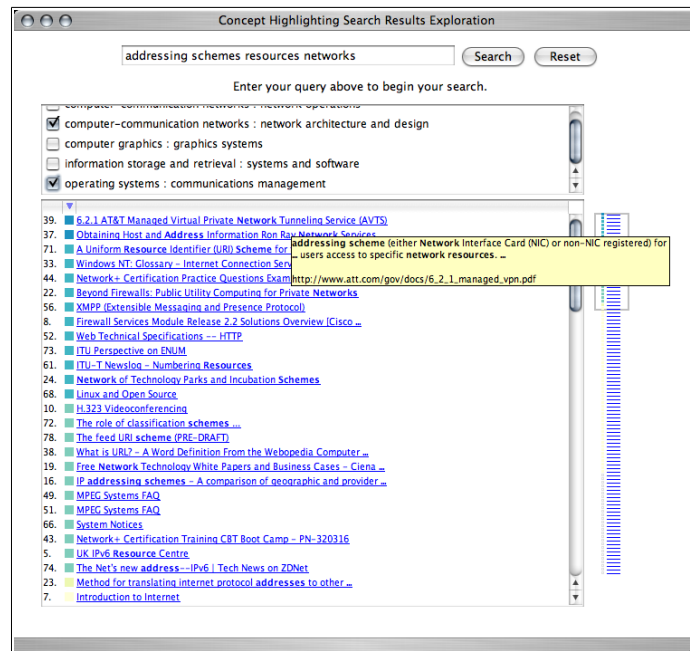
Figure 5.11: The compact representation in the detail window supports browsing; the details can be easily accessed via tool tips.
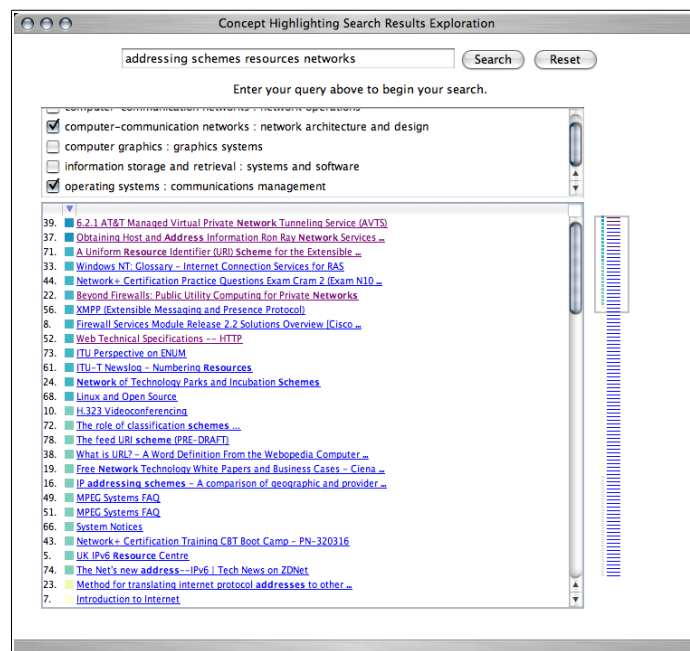
Figure 5.12: Previously visited documents can easily be identified, even when a change in the concept selection results in a new order of the search results.

5.3.4   Discussion

Even for well crafted queries, the results of Web searches often contain document surrogates of varying degrees of relevance to the users' information seeking goals. Clustering of the search results allows the users to navigate the clusters in order to narrow down the set of search results to a smaller collection containing a larger ratio of relevant documents. However, most Web search clustering systems use simple keyword-based cluster naming techniques; they do not allow the users to select multiple clusters simultaneously; and they do not organize the search results once a cluster is selected.

Concept Highlighter provides a means for generating concept-based fuzzy clusters of Web search results, and an interface for visually representing the fuzzy membership scores and interactively exploring Web search results. The visual exploration of the concept-based fuzzy clusters allows the users to interactively select the concepts they think may be relevant to their information seeking goal, and see the results of these concept selections in the highlighting of the document surrogates that belong to the corresponding fuzzy clusters.

The ability of Concept Highlighter to allow the users to find relevant document surrogates depends on the ability of the tool to match the users' queries to the concept knowledge base. If there are few concepts returned, or if all the concepts returned are relevant to the users' information needs, the ability to assist the users in narrowing down the search results is diminished. This can occur when the users' queries are very specific, and will often lead to very specific search results.

More often, the users' queries are less specific. This results in multiple concepts being selected from the concept knowledge base, and a more general collection of documents being returned from the search engine. It is in these situations Concept Highlighter can assist the users in finding relevant

documents. The interactive exploration of the Web search results using the concept-based fuzzy clusters can lead the users to groups of document surrogates that are relevant, and away from groups of document surrogates that are not relevant.

Spence's Navigation Framework

Spence's navigation framework provides a structure in which the features of Concept Highlighter can be discussed. As described in Section 2.4.1, this framework deals with the human activities involved in navigating within an information system [90, 91].

Browsing. As the user browses the Concept Highlighter interface, they will see information relating to the concepts that may be relevant to their information need, the top search results set, and a visual representation of the current fuzzy clustering status. As the user selects or un-selects concepts from the concept list, they can begin to browse and explore the relationships between the concepts and the top search results. This is represented by a colour coding, which can be viewed both in the entire set of search results (in the overview map), and for specific document surrogates (in the detail window).

Specific document surrogates can be browsed by hovering the mouse over the document title text. The user can view these documents by clicking on them, which is also considered a browsing activity.

Formulate an Internal Model. Throughout these browsing activities, the user develops an internal model for their current search. This internal model contains information related to the set of concepts, information related to the search results set and specific documents, and information related to the

relationship between the concepts and the search results set.

The internal models developed from previous searches support the development of the internal model for the current search. Since there are a finite number of concepts available in the concept knowledge base, the users will develop internal models regarding the relevance of previously seen concepts to their past Web search tasks. When these concepts appear in the current Web search task, the internal model created will be influenced by what the user previously learned about this concept.

As the user browses the interface, they may choose to view a specific document they feel may be relevant to their information need. Doing so opens the document in a new window, allowing them to read and evaluate the document in detail. The information they gain by viewing the documents further contributes to the development of their internal model for the current search task.

Interpret the Information. Before the user can act upon the information within their internal model, they must make a correct interpretation of this information, along with what they currently see in the interface. Of fundamental importance within Concept Highlighter is making a proper interpretation of the relevance of a concept with respect to their current information needs. This interpretation is supported through the user's ability to learn the meanings of the concepts as they use the system.

As relevant concepts are selected, the user must also make a correct interpretation of the colour codes that represent the fuzzy membership scores. This interpretation supports the user in identifying the potential importance of a document to the selected concepts, and therefore, to their information need.

As the user views specific documents, their internal model will begin to

include specific information that is relevant to their information need. The user will decide whether they have found enough information to fulfill their information need by interpreting the completeness of their internal model with respect to their information need.

Formulate a Browsing Strategy.   When formulating a browsing strategy, a number of different sources of information can be used: the internal model, the visual representation being viewed, or a new idea the user has had. The resulting browsing strategy may be part of planned strategy with a specific goal in mind, or an opportunistic strategy based on something the user has thought of or seen.

Based on the user's internal model, they may choose to select one or more concepts in order to re-sort the search results based on their fuzzy membership scores (planned strategy), or they may experiment with selecting marginally relevant concepts that they have no prior knowledge of (opportunistic strategy). When viewing the interface, the users may choose to view specific documents that appear to be relevant (planned strategy), or they may identify and view a document of interest that is not relevant to their current task (opportunistic strategy). The users may choose to browse the search results list looking for a term they have thought of (planned strategy), or may simply begin a new search based on something that has come to their mind (opportunistic strategy).

Norman's Stages of Action Framework

The two gaps between the user of a system and the system itself that Norman's stages of action framework describe are the the gulf of evaluation and the gulf of execution [68, 69]. The steps that users take as they cross these

gaps are described in more detail in Section 2.4.2. This framework provides a useful means by which the interactive features of Concept Highlighter can be discussed.

Goals   As with the discussion on HotMap, the real-world goal for the user is to find a number of relevant documents that satisfy the user's information needs. In terms of a system goal, the user's task is to select relevant concepts and then explore the search results in order to discover documents that fulfill the user's information need.

Gulf of Evaluation   In the perception stage, the user is able to see the fundamental interface objects of pixels, geometric objects, and characters, and group these together into higher level objects. In the list of concepts, the user is able to perceive the connection between the check boxes and the names of the concepts. In a similar manner the user is able to make a perceptual connection between the fuzzy membership colour indicators and the title of the document surrogates, both in the detail window, and the zoomed-out view of the overview map.

For the interpretation stage, the user identifies the perceived objects as specific types of user interface objects. For example, the user is able to readily interpret the check-boxes in the concept list, as well as understand that they can place a check within these boxes. They can also interpret the titles of the document surrogates as representing documents in the search results set. The Web link representation of these assists the user in interpreting that they can be clicked.

Although first-time users may have difficulty making a correct interpretation of the colour coded boxes beside each document surrogate, as they use the

system by selecting relevant concepts, they will be able to learn the meanings of these boxes. With experience, the difficulties with this interpretation will be reduced and eventually eliminated.

In the evaluation stage, the user attempts to make sense of what they are seeing with respect to their system goal of finding relevant documents. The user is able to identify the concept list as a place where concepts can be selected, the detail window as a place where the document surrogates can be evaluated, and the overview map as a compact and abstract representation of the entire set of search results.

Gulf of Execution In the intentions stage, the user attempts to convert their system goal of finding relevant documents into an intention to do some action with the interface, such as select a concept for re-sorting the search results, or view the details of a specific document surrogate, or view an actual document. The ability to choose an intention is based on the user's ability to effectively cross the gulf of evaluation.

Since the action sequence required to select a concept from the concept list requires a simple click, little planning required. Since many users of the system will have had prior computer use experience, little cognitive effort is required to click the mouse button on the desired term. The other actions are just as simple, requiring virtually no planning or thought to execute.

As the user executes an action sequence, the interface is instantly updated with the results of the action. When selecting a concept from the concept list, the user will be able to see that the search results are provided in a sorted order, both in the overview map and the detail window. This feedback indicates to the user the success of their action. Similar feedback is provided for the other types of interaction as well.

5.3.5 Summary

The goal of finding a set of relevant documents using Concept Highlighter is supported in two stages of interaction. In the first stage, the user considers the concepts provided in the concept list, and selects those that are relevant to their information need. This results in an interactive re-sorting of the search results based on the fuzzy similarity of the document surrogates to the selected concepts. In the second stage, the user considers the re-sorted document surrogates for relevance. Due to the interactive nature of the concept selection, and the speed at which the changes are reflected in the search results list, the user can easily return to the concept selection stage at any time, and then return to considering the document surrogates for relevance. The benefit in using Concept Highlighter is that macro-scale decisions in selecting relevant concepts results in re-sorted search results with those documents that are most similar to the concepts near the top of the list.

5.4 WordBars

In addition to supporting the interactive query refinement process (as described in Section 4.3) WordBars can also support interactive search results exploration. In the description of the features of WordBars, it was pointed out that the search results could be re-sorted based on the use of selected terms in order to determine whether these selected terms are appropriate for query refinement. However, once the user is satisfied with their query, they can continue to use the WordBars as a means for exploring the current search results set.

There is no operational difference in WordBars when supporting interactive search results exploration versus interactive query refinement. The difference

is in how the Web searcher makes use of the features. That is, if the user chooses to double-click on terms to add or remove from the query, then they are performing interactive query refinement steps. If instead they choose to single-click on terms to re-sort the search results, then they are performing interactive search results exploration steps.

The benefit that WordBars has over the other three systems discussed in this dissertation is that it supports both query refinement and search results exploration processes within the same interactive interface. The other three support only interactive query refinement or interactive search results exploration processes, but not both. While it is possible to combine the features of some of these systems together (e.g., VisiQ and Concept Highlighter), because the visual interfaces are fundamentally different, switching between the two tasks would require an explicit switch in the interface mode (i.e., switching from query mode to search results mode).

The visual interface of WordBars supports both tasks simultaneously, without requiring any explicit switching of modes. This is very beneficial since it does not require the users to make the conscious decision about which task they are performing. Therefore, they can switch back and forth between these tasks with ease. For example, after refining a query and then exploring the search results, if the user finds another term that they would like to add to the query, they can do so quickly and easily, and then return to their exploration tasks again. This seamless transition reduces the cognitive effort required in making such decisions.

5.4.1 Example

Although a number of examples were provided in Section 4.3.3 to illustrate the benefits of WordBars for interactive query refinement tasks, one more

example is provided here to illustrate the benefits for interactive search results exploration.

Suppose the user starts with the very specific query "search results visualization information retrieval" (perhaps after having used the WordBars system in order to craft the query). The initial set of search results will be presented in the order provided by the Google API (see Figure 5.13).

The term frequency histogram provides the user with information regarding the commonly used terms within the top search results. The user can select terms from this list that may be relevant to their information need. For example, the user may identify the term "clustering" as an avenue of interest, resulting in a re-sorting of the search results based on the use of this term (see Figure 5.14). Note how the document surrogates on clustering that were buried deep in the search results have been moved to the top of the list of search results.

The user may wish to explore other aspects of these search results. For example, they may wish to focus on those document surrogates that make use of the term "web" (see Figure 5.15). Alternately, they may identify the terms "digital" and "library" as interesting (see Figure 5.16).

With each change in the term selections made in the term frequency histogram, the search results are re-sorted based on the use of the selected terms. This re-sorting occurs instantaneously, allowing the users to easily experiment and explore the search results. As such, the users can easily make alternate selections for re-sorting, with little effort.

5.4.2 Discussion

The primary benefit WordBars provides to users is the ability to interactively manipulate the order by which the document surrogates are sorted. Most
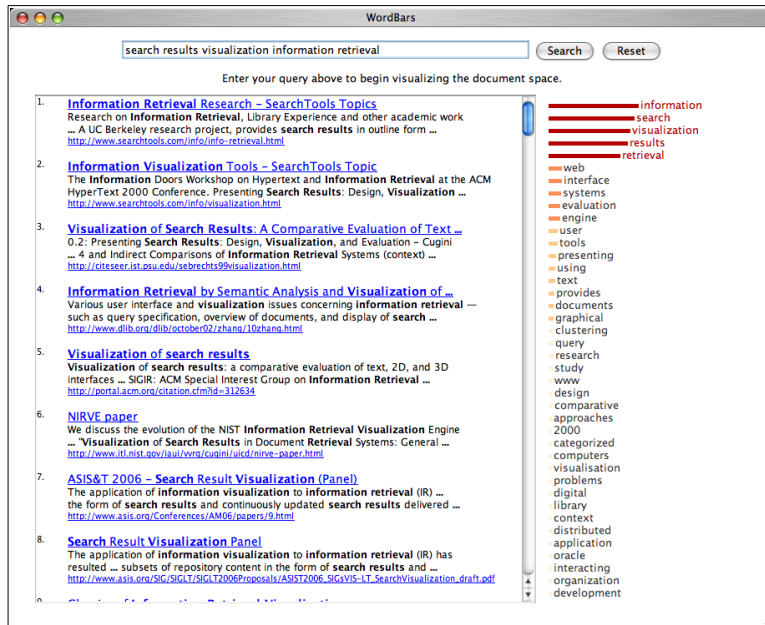
Figure 5.13: An initial query in WordBars generates a term frequency histogram and a list of the top search results.
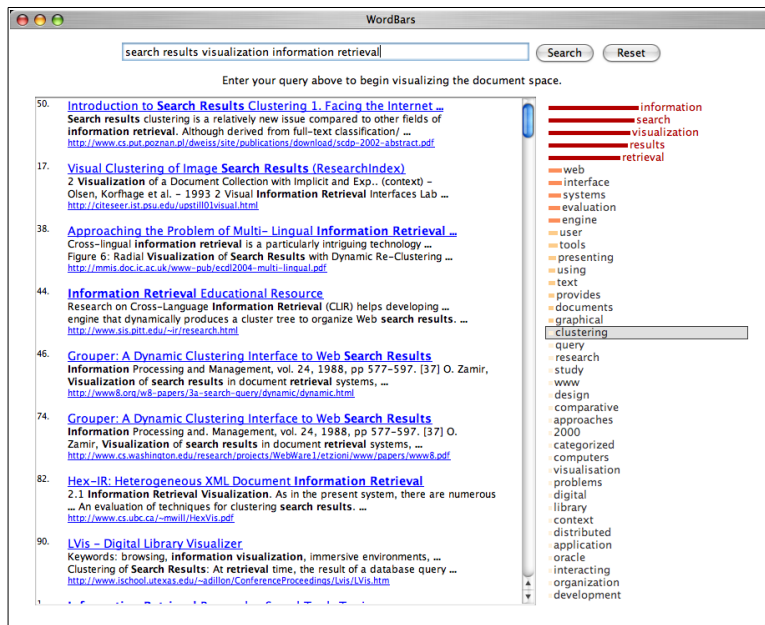
Figure 5.14: The search results can be re-sorted based on the use of the terms selected in the term frequency histogram.
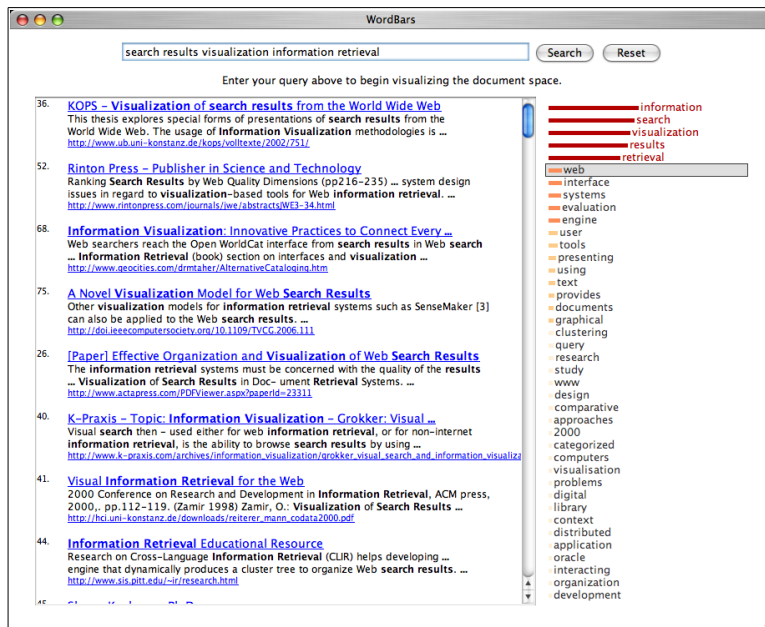
Figure 5.15: Terms can easily be selected and unselected in the term frequency histogram, resulting in an interactive re-sorting of the search results.
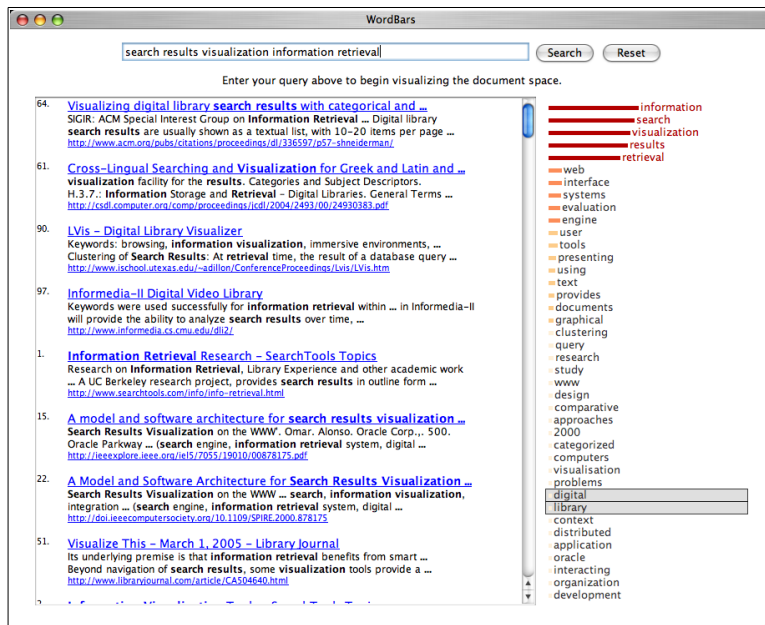
Figure 5.16: Multiple selections can be made, re-sorting the search results based on the combined frequency of all the selected terms.

Web search engines provide automated rankings of the search results based on complex and proprietary algorithms, such as PageRank [13]. However, these algorithms result in a static ordered list of search results. Interactive exploration, drawing upon the user's understanding of their information need, can allow highly relevant documents located deep in the search results to be brought to the attention of the user. This is especially valuable when the search results consist of a mixture of relevant and non-relevant documents, which is often the case.

The term frequency histogram provides to the users a clear indication of which terms are commonly used in the top search results, along with an indication of their relative frequencies of use. This visual representation can assist the users in exploring the search results, and focusing on specific aspects of their information need, without having to remember the exact terms used to describe what it is that they are seeking.

Spence's Navigation Framework

As with the previous systems, Spence's navigation framework provides a basis for discussing the features of WordBars, with respect to supporting interactive search results exploration. See Section 2.4.1 for more details on this framework. A discussion of WordBars with respect to interactive query refinement tasks is provided in Section 4.3.4.

Browsing. While visually browsing the term frequency histogram, the user can see information about the commonly used terms within the top search results. The frequency bars within the histogram can be seen to represent the relative frequencies with little or no training. By selecting terms in the histogram, the search results are re-sorted and the user can see the document

surrogates that make frequent use of those terms. The users can view the documents themselves by clicking on a document surrogate, and can easily identify which document surrogates have previously been viewed (due to a change in the link colour).

Formulate an Internal Model. As the user browses the visual interface, they develop an internal model of what they see. The user's internal model will be influenced by the visual representation of the term frequencies. As the user re-sorts the search results, their internal model is expanded with an understanding of which document surrogates make frequent use of their terms of interest. When the user chooses to view a particular document, their internal model with respect to that specific document is further re-enforced.

If the user makes frequent use of the WordBars system to search for information on related topics, the internal models from previous searches will influence the development of the internal model for the current search. For example, the user's internal model of the relevance of terms from previous searches will affect the internal model for the relevance of those terms in the current search. Over time, the user will be able to learn which terms represented in the term frequency histogram are of value for their common search topics.

An important aspect in the development of an internal model for a given search is that the user is able to recognize terms within the term frequency histogram, rather than needing to recall them while exploring the search results. This recognition ability provides a link between what they see in the externalization of the data (i.e., the visual representation), and their previously constructed internal models.

Interpret the Information. Before the user can act upon the information they have gained through the use of the WordBars system, they must make a proper interpretation of their internal model, as well as what they are currently viewing in the system. The width and colour coding of the bars within the term frequency histogram supports the user in interpreting the relative frequencies of the terms found in the top search results. As the user explores how candidate terms are being used in the top search results, they are able to make effective interpretations of how these terms are being used within the search results. Further, due to the colour coding of the document surrogate links, the users can easily interpret which documents they have already viewed, and which they have not. If the user interprets their internal model as having enough information to satisfy their information need, then they will make a decision to complete their task.

Formulate a Browsing Strategy. As noted in the previous discussions, there are three different sources of information for formulating a browsing strategy: the internal model, the visual representation, and a new idea. The resulting browsing strategy may be a planned strategy, or opportunistic.

Based on the current interpretation of the internal model, a user's browsing strategy may be to evaluate some of the top document surrogates in the list (planned strategy), or to re-sort the search results based on one or more terms they have discovered (opportunistic strategy). Based on viewing the interface, the user may choose to view a potentially relevant document by clicking on it (planned strategy), or may choose to view a document that is not relevant to the current information need, but is of interest to the user (opportunistic strategy). The user may think of a term that is relevant to their current task and actively search for that term in the term frequency histogram (planned

strategy). Of they may think of some other topic in which they are interested, and start a completely new search (opportunistic strategy).

Norman's Stages of Action Framework

Norman described two gaps that exist between the users of a system and the system itself: the gulf of evaluation and the gulf of execution [68, 69]. How the users of a system cross these gaps form Norman's stages of action framework, which is described in more detail in Section 2.4.2. As with Spence's navigation framework, this framework provides a useful mechanism for commenting on how the features of WordBars support the users in fulfilling their task goal of satisfying their information need.

Goals   In the previous discussion on WordBars (see Section 4.3.4), the focus was on fulfilling the system goal of crafting a query that describes the user's information needs. Once a query is crafted, WordBars supports the user's system goal of exploring the search results through the re-sorting features, with the purpose of finding a number of documents that are relevant to the user's information need.

Gulf of Evaluation   During the perception stage, viewers of WordBars perceive the low level objects in the interface, and group these together into higher-level objects such as the bars and terms in the term frequency histogram. The regular shapes of the histogram bars, and the proximity of the corresponding terms support the user in making the perceptual connection of the terms to their histogram bars. They also perceive the text in the detail window as distinct objects due to the regular structure (title, snippet, URL) and the gap between document surrogates.

In the interpretation phase, the user is able to easily identify the text blocks as document surrogates. This interpretation is supported by previous experience using WordBars, as well as previous Web search experience. Within the term frequency histogram, the user can identify the histogram bars as representing a value for the corresponding term. Although the interface does not explicitly state the meaning of the histogram values, logical interpretations include frequency, value, and importance. As users move beyond the novice user stage, prior experience will support the proper interpretations of this data.

Although the term frequency histogram does not provide much information to the user as to how it can be used, with a small amount of training, the user can make the proper interpretation that the histogram bars can be clicked in order to re-sort the search results. The use of the de facto standard for Web links in the list of search results supports the user in interpreting these as clickable objects leading to the source documents.

The final stage to overcome in the gulf of evaluation is for the user to make a correct evaluation of the document surrogate list and the term frequency histogram with respect to their current system goal. With a goal of finding relevant documents, the user can readily identify the document surrogates as fundamentally important objects. With experience, the user will also be able to evaluate the term frequency histogram as they seek terms that may support their system goal of finding relevant documents.

Gulf of Execution   For the intentions stage, the user's system goal of finding a number of relevant documents to their information need is converted into an intention of re-sorting the search results based on relevant terms identified in the term frequency histogram. In subsequent cycles of use, the user may form intentions to evaluate the document surrogates in the search results list,

or select to view specific documents.

Since the action sequence required to select a term from the term frequency histogram is a simple click, little planning is required. Further, since many users of the system will have had prior computer use experience, little cognitive effort is required to click the mouse button on the desired term.

After clicking on a term in the term frequency histogram, the user is provided with instant feedback in the form or re-sorting the search results. This feedback is further supported by the placement of a gray box around the selected term to indicate its selected status. These changes are easy for the user to perceive, interpret, and evaluate, and provide the user with an indication that they successfully executed their planned action sequence.

5.4.3   Summary

In addition to providing support for interactive query refinement, the term frequency histogram in WordBars supports the users in interactively exploring the search results. The visual representation of the term frequency histogram provides information to the user regarding how frequently the terms are being used, providing a summarization of the search results set. Browsing the histogram allows the users to recognize terms that are used in the top search results as being relevant to their search task. Selecting these terms results in an instant re-sorting the search results based on their use, allowing the users to easily consider these documents for relevance.

Chapter 6

Experiments and Evaluations

6.1 Purpose

In many user-oriented research fields, such as human computer interaction and information visualization, conducting user studies is not only an accepted practice, it is an expectation. User evaluations provide the researchers with a means to verify and validate design assumptions, confirm or reject hypotheses, and make comparisons between different systems and techniques. Within these domains, there are numerous accepted methods and procedures that can assist researchers as they evaluate their work. Many of these techniques are covered in textbooks in the domain of human computer interaction [64, 88, 84, 76].

A common procedure for conducting user evaluations with software is to assign the participants various tasks to perform, measure the time it takes to complete the tasks and the number of errors made, and make observations about how the software is being used. Subjective reactions by the participants can also be collected in order to measure the user's confidence and satisfaction in completing the task, and the ease of use of the software. The assigned tasks will commonly be a representative set of operations that the users would need

to perform in the real-world use of the software. Alternately, tasks may be assigned to explore specific new features of the system in order to understand whether these new features are indeed an improvement over the existing practice. It is also common to assign tasks of various degrees of difficulty in order to measure how this can affect user performance.

Applying these user evaluation methods in the study of Web search results interfaces is not as well understood as in other domains. Su noted that "the study of end-user searching on Web search engines is still in its infancy" [99] In some cases, researchers have tested their Web search interfaces using traditional information retrieval collections [9], while others have evaluated their work based solely on usage logs [121]. A number of researchers have allowed participants to conduct searches based on their own interests [121, 92, 99, 71]; others have assigned specific search task to the participants [9, 106].

Even for something as fundamental to information retrieval research as relevance judgments, there have been a wide range of methods for measuring this data. Zamir and Etzioni based the relevance on the documents viewed by the participants [121]. Berenci et al. collected binary relevance judgments from the participants [9]. Nowicki required the participants to provide the top documents in their ranked order of relevance [71], as did Vaughan [106]. Su had the participants rank the documents as well as score the documents on a three-point relevance scale [99]. Spink used a combination of a four-point relevance scale and a continuous relevance scale [92]. Even within user evaluations of traditional information retrieval systems, the measurement of relevance decisions can range from binary decisions [34] to a continuous relevance scale [46].

These studies also showed a high degree of variability in how the experiments were designed, whether a system was evaluated independently or in

comparison to other systems, how the data was collected, and by what means the data was analysed. As one might expect, there are many factors involved in choosing the methods by which to evaluate Web search systems. Since there is no single "best practices" methodology for evaluating the interfaces for Web search, in this research, methods were selected that were best suited to the evaluation of each prototype system.

As noted in Chapter 1, a design-oriented research methodology [20] was used in this work to explore ways in which visualization and interaction could be used to support users in their fundamental Web search tasks. In this methodology, the knowledge discovered comes from the design and study of the prototype systems which address the problems and challenges in the research domain. In the previous two Chapters, the design of four prototype systems has been described; in this Chapter, these systems are studied through experiments and evaluations.

The study of such prototype systems can be a challenging task. In general, there is a need to conduct controlled experiments and make comparisons to some baseline performance level. Not all aspects explored within a specific problem domain can be studied in a controlled manner. Query refinement is one such aspect of the Web search domain that proves to be exceptionally challenging for study.

There are two aspects of Web search query refinement that make its experimental evaluation difficult and error prone. The ability to effectively refine a query (manually, or with the support of a software system) is largely dependent on the user's prior knowledge on the topic and their ability to learn about the topic during the search task. This is a difficult variable to control in an experimental setting. Determining the success or failure of a query refinement task is also difficult, since for a given task, many different queries may be

crafted with varying degrees of correctness.

As such, the techniques suggested to support interactive query refinement are presented in this dissertation without study on their effectiveness. Instead, extensive examples were provided in Chapter 4 to illustrate the potential benefits of these systems in various Web search scenarios. In order to evaluate such systems, real-world evaluations in longitudinal settings are needed. These studies are beyond the scope of this dissertation.

The focus in the remainder of this Chapter is on the experimentation and evaluations of the systems which support interactive search results exploration. The goal is to be able to comment on the success of the selected visualization and interaction techniques versus the list-based representations used by Google with respect to supporting the users in being more effective, efficient, and satisfied in conducting their tasks.

In this research, two different types of studies of the prototype systems were conducted: validation of specific features, and empirical evaluations through user studies. In the validation studies, specific features of the systems were studied to verify fundamental assumptions in the system design. Two such studies were conducted in order to determine the potential effectiveness of HotMap (see Section 6.2) and Concept Highlighter (see Section 6.4) in terms of re-sorting the search results.

Two different user studies were also conducted in the course of this research. In these user studies, participants were recruited from the University community to conduct Web searches with the prototype systems in a controlled setting. In the first study, participants used the Google search engine, HotMap, Concept Highlighter, and a fourth system that proved to be ineffective (and is not discussed in this dissertation). In order to focus on the specific benefits of the prototype systems, the results of the comparison of HotMap

vs. Google and Concept Highlighter vs. Google are reported separately (see Sections 6.3 and 6.5).

In a second user study, WordBars was evaluated independently of the other two prototype systems (see Section 6.6). This second study benefited greatly from the lessons learned in the first study. The independent evaluation allowed an in-depth analysis of the various features of WordBars using two distinct classes of participants (expert and intermediate users). The initial sets of search results were presented in the order provided by Google, allowing a baseline comparison. Although the primary focus was on the evaluation of WordBars as an interactive search results exploration system, an interactive query refinement step was also included and evaluated in terms of the relevant documents found as a result.

In these user evaluations, the focus was on measuring improvements in the effectiveness, efficiency, and subjective user reactions when conducting prescribed Web search tasks. Where appropriate, Google was used as a baseline for comparison purposes, since it is the current market leader in the Web search domain. The end result is a comparison of the list-based representations of Google to the visual and interactive features of the prototype systems developed in this research.

6.2   Determining the Potential Effectiveness of HotMap

One of the fundamental assumptions in the design of HotMap is that it may be possible to move relevant documents that are buried deep in the search result to more prominent locations near the top of the list through search

results re-sorting based on query term frequencies. For vague queries, the search results will commonly consist of a mixture of relevant and non-relevant documents; often there will be relevant documents buried deep in the search results list. The re-sorting features in HotMap allow the users to specify a nested sort, resulting in document surrogates that make frequent use of the specified query terms being moved to the top of the search results list.

This re-sorting assumption is based on a manual evaluation of the search results of a number of different vague queries. For such queries, it was found that although there were many non-relevant documents in the search results, there were also many relevant documents; finding the relevant documents required the viewing of more than the common three-page limit of search results (i.e., the point at which most users stop viewing the search results [89, 93]). Many of the highly relevant documents were found to make frequent use of the query terms. Re-sorting the search results can bring some of these highly relevant documents that are buried deep in the search results near the top of the list, resulting in a higher portion of relevant documents among the top documents in the search results.

In this Section, the methods for testing this re-sorting assumption is presented, along with the experimental results.

6.2.1 Methods

To test these assumptions, four vague search tasks were composed, each consisting of descriptions of the information need and corresponding queries:

Task 1 You are a network manager for a small company. You are looking for information on tools, software, or services to assist you in managing your company's communications network (i.e., a data and/or voice network).

query: "communications network management"

Task 2 You are a software developer working on a new project that requires complex knowledge and information to be stored and maintained. You are looking for information about how to represent this information in your software system.

query: "representing knowledge information"

Task 3 You are conducting research on operating systems. You are looking for information on the communication mechanisms between processes that are running within the same computer.

query: "communication mechanism processes"

Task 4 You are looking for information relating to addressing schemes for resources within a network. Examples include IP addressing and URL addressing.

query: "addressing resources networks"

Each of these queries was submitted to the Google API, and the top 100 search results were retrieved. Each document surrogate was evaluated independently by both the author and his doctoral supervisor, and assigned relevance scores on a four-point relevance scale (see Table 6.1). Where there was disagreement in the relevance scores assigned, the document surrogate was considered again and consensus was achieved on the relevance score. Having multiple expert reviewers of the search results reduced the errors in evaluating individual document surrogates.

Table 6.1: The relevance scores used to rate document surrogates.

Score	Description
4	This document is relevant. I would definitely click on it.
3	This document is probably relevant. I would likely click on it.
2	This document is probably not relevant. I might click on it.
1	This document is not relevant. I would not click on it.

Note that in traditional information retrieval research, test collections (such as those provided by TREC, Cranfield, or CACM) are commonly used. These test collections usually consist of a set of documents, a set of queries, and expert relevance judgments for each document-query pair. When conducting research using live Web information retrieval data, the relevance judgments are not available and must be provided by the researchers. Due to the dynamic nature of Web search results, a cached copy of the search results was used to ensure that the reviewers considered exactly the same set of search results.

All unique combinations of nested sorting based on the query term frequencies were calculated, and the number of relevant documents within the top 10, top 20, and top 30 documents was counted. Note that any document assigned a score of three or four on the four-point relevance scale (see Table 6.1) was considered "relevant". This represents those documents that were marked as "relevant" and "probably relevant" by the expert reviewers.

6.2.2 Results

Table 6.2 summarizes the results of this experiment for all four tasks. Note that there are exactly 78 unique ways to specify a nested sort based on three

query terms. The results are reported separately for each query since the effectiveness of re-sorting the search results depends primarily on the specific search results returned by for each search task.

For the four search tasks evaluated in this experiment, there is much evidence to support the re-sorting assumptions. In all cases except for one (i.e., Task 4, considering 20 documents), there was at least one way to re-sort the search results based on query term frequencies that resulted in more relevant documents being present in the top documents when compared to the order provided by Google. In multiple cases, more than 35% of the ways for specifying a nested sort of the search results provided better results than the Google order. However, in other cases, there were very few ways for specifying a nested sorting in order to achieve an improvement.

For Task 4, when considering 20 documents, re-sorting the search results based on query term frequencies was unable to achieve a better order than that provided by Google, although a better order was achieved when considering 10 and 30 documents for this task. The results for Task 4 illustrate the difficulty in this type of research, where the use of the query terms does not necessarily imply the relevance of the document. In general, however, the results of this study were rather positive.

As a result of this study, it is concluded that there is sufficient support for the fundamental assumptions of the re-sorting features in HotMap. When considering the top 10 documents in the re-sorted search results, there were many different ways to specify a nested sorting of the search results that resulted in more relevant documents at the top of the list. When considering the top 20 and top 30 documents, the results were generally positive, but less conclusive.

To further illustrate the benefit of these re-sorting methods, the number of

Table 6.2: The potential effectiveness of re-sorting the search results using HotMap.

Number of sort methods...	10 Documents	20 Documents	30 Documents
Task 1: "communications network management"			
better than Google	29 (37%)	23 (29%)	31 (40%)
equivalent to Google	2 (3%)	3 (4%)	6 (8%)
worse than Google	47 (60%)	52 (67%)	41 (53%)
Task 2: "representing knowledge information"			
better than Google	28 (36%)	6 (8%)	22 (28%)
equivalent to Google	9 (12%)	7 (9%)	3 (4%)
worse than Google	41 (53%)	65 (83%)	53 (68%)
Task 3: "communication mechanism processes"			
better than Google	13 (17%)	3 (4%)	11 (14%)
equivalent to Google	16 (21%)	6 (8%)	6 (8%)
worse than Google	49 (63%)	69 (88%)	61 (78%)
Task 4: "addressing resources networks"			
better than Google	28 (36%)	0 (0%)	4 (5%)
equivalent to Google	3 (4%)	8 (10%)	0 (0%)
worse than Google	47 (60%)	70 (90%)	74 (95%)

relevant documents found using best nested sort method in HotMap was compared to the number of relevant documents found using the order of the search results provided by Google. These results are provided in Table 6.3. There was as much as a 100% improvement in the number of relevant documents in the top search results for the best re-sorting methods. The average improvement when the best re-sort method was selected, was 71% for the top 10 documents, 18% for the top 20 documents and 30% for the top 30 documents.

Note that there appears to be a trend in these results, where there is a great improvement when considering the top 10 documents, a marginal improvement when considering the top 20 documents, and a moderate improvement when considering the top 30 documents. This can be attributed to the fact that there were on average 7 relevant documents on the second page of the search results in the Google order, but only 3 relevant documents on the third page.

Table 6.3: Comparing the number of relevant documents in the Google order to the best order achieved by re-sorting in HotMap.

	10 Documents	20 Documents	30 Documents
Task 1: "communications network management"			
Google	4	12	16
Best Sort	8	16	23
percent improvement	100%	33%	44%
Task 2: "representing knowledge information"			
Google	6	13	15
Best Sort	9	15	20
percent improvement	50%	15%	33%
Task 3: "communication mechanism processes"			
Google	6	13	16
Best Sort	8	16	20
percent improvement	33%	23%	25%
Task 4: "addressing resources networks"			
Google	3	9	11
Best Sort	6	9	13
percent improvement	100%	0%	18%

Therefore, the ability to improve upon the search results order when considering the top 30 documents was greater than when considering the top 20 documents.

Although it has been shown that in many cases there exist ways to re-sort the search results in order to provide more relevant documents at the top of the list, it is not clear that a user of HotMap will be able to make the proper decisions as they explore the search results. Clearly, there are many re-sorting methods that result in an order that is worse than that provided by Google (see Table 6.2). To verify that the users are able to make good decisions when using HotMap, a user evaluation was conducted, the results of which are reported in the following section.

6.3 HotMap vs. Google

Although it has been shown that it is possible to achieve a higher number of relevant documents among the top documents in the search results for vague queries via query term re-sorting, it is not clear whether the users will be able to effectively make use of this feature to explore the search results and find relevant documents. To test the effectiveness of this nested sorting feature, as well as the other visual and interactive features of HotMap, a user study was conducted. The design and results of this study are provided in the remainder of this section.

6.3.1 Methods

To test the efficiency, effectiveness, and subjective reactions when conducting information seeking tasks, a 4x2 (interface x task) within-subjects design was employed. Although two other interfaces for representing and exploring search results were also evaluated in this study (one being Concept Highlighter, which is discussed in Section 6.5), the focus is on the comparison between HotMap and Google. Since all the participants were already frequent users of Google, they were exposed to this interface first. In order to reduce the biasing effects, the task order and remaining interface order were varied among the participant pool.

The first two tasks in Section 6.2.1 were used in this experiment. To ensure that the interfaces provided the same set of search results to each participant, the search results for each task were cached and provided to each interface. These cached results were the same as those evaluated by the expert reviewers in the potential effectiveness analysis in the previous section. The Google interface was altered to include document numbers to facilitate data collection.

6.3.2 Procedure

Each participant completed a pre-task questionnaire, two searches with each of the interfaces, an in-task questionnaire following each search, and a post-task questionnaire after all the searches were complete. The entire procedure took between 60 and 90 minutes for each participant.

Before the participants were exposed to an interface, a short training task was provided, along with a brief description of the features of the interface. This ensured that each participant had a basic understanding of the interface, how it was to be used in the search task, and how the data was to be collected.

Each search task included a written description of the information need, along with the query to be used. After submitting the search to the assigned interface, the participants were asked to use the interface to evaluate the search results. For each document surrogate considered, a relevance score on a scale from 1 to 4 was assigned (see Table 6.1). The participants were asked to speak these scores; this information was logged by the investigator, along with the elapsed time. Only the document surrogates were considered for relevance; the participants were asked to not view the actual documents. After ten document surrogates were assigned a relevance score of either three or four, the participants were instructed that the task was complete.

At the end of each task, the participants were provided with an in-task questionnaire to measure their subjective reaction and feelings regarding their experience with using the assigned interface to find documents relevant to the assigned task. These subjective measures were based on the participants' confidence in finding a good set of documents, the ease of use of the interface, satisfaction in using the interface, and impressions of ambiguity in the search results set.

After all the search tasks were completed, a post-task questionnaire was

administered to measure the participants' impressions of various features of the interfaces. In addition, the participants were asked to provide a ranking of their preference for an interface, and textual comments regarding the features of the interfaces.

6.3.3 Analysis

The quantitative results (time to task completion, perceived and corrected precision, and error rates) were analysed independently for each task using analysis of variance (ANOVA). The subjective evaluations were analysed using nonparametric Friedman tests. The preference ranks were analysed pair-wise using Wilcoxon signed ranks tests. Where relevant, the statistical significance of these tests are highlighted by a bold font.

6.3.4 Results

Participant Demographics

28 participants were recruited from undergraduate computer science courses to participate in this study. The results from the pre-task questionnaire administered to these participants are presented in Table 6.4. The participants were primarily Google users, with a moderate to high degree of Web search experience. However, since none had been exposed to HotMap prior to this study, they are considered novice users of this system.

Prior to starting the first search with each new task, the participants were asked to report their familiarity with the assigned task. These results are illustrated in Figure 6.1. Clearly, the participants showed a wide range in familiarity with the search tasks.

Table 6.4: Features of the participant demographics.

Computer Use	10+ times per week: 96%
	5-10 times per week: 0%
	1-5 times per week: 4%
Computer Experience	high degree: 64%
	moderate degree: 36%
	low degree: 0%
Web Searches	10+ per week: 75%
	5-10 per week: 18%
	1-5 per week: 8%
Search Engine Preference	Google: 93%
	Yahoo: 7%
Pages Viewed	1-2 pages: 32%
	3-4 pages: 25%
	5-6 pages: 25%
	7+ pages: 18%
Web Search Experience	high degree: 46%
	moderate degree: 54%
	low degree: 0%

Time to Task Completion

In order to measure and make comparisons among the times taken to complete the assigned information seeking tasks, it is necessary to provide clear task completion criteria. For this study, two levels of fulfillment of the assigned information need were specified: finding five relevant documents, and finding ten relevant documents. In this criteria, any document assigned a relevance score of three or four was considered a relevant document. This represented documents the participants felt were either certainly or probably relevant (see Table 6.1 for the complete relevance score scale).

Figure 6.2 illustrates the average time the participants took to find five and ten relevant documents for each of the two tasks. On average, the participants
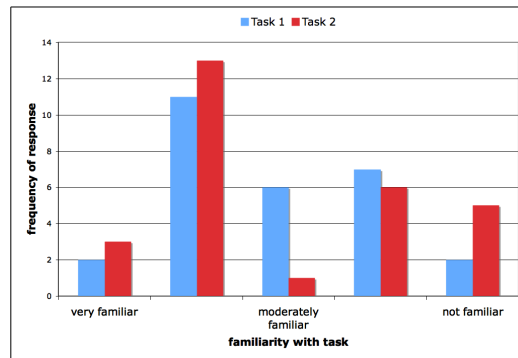
Figure 6.1: The pre-test familiarity with the tasks reported by the participants.

were able to find relevant documents faster with HotMap than Google. However, the differences in these times only proved to be statistically significant for finding ten relevant documents for Task 1. The results of ANOVA tests are provided in Table 6.5.

These results are interesting since HotMap requires the participants to interact with the interface features before they can evaluate the search results; by contrast, Google allows the participants to start evaluating the search results immediately. Even though the participants did not perform significantly faster using HotMap in all cases, neither did they perform significantly slower. This indicates that the extra work required by HotMap can be offset by the users' ability to more easily evaluate and explore the search results.

Furthermore, even though a large disparity existed in the experience the participants had with Google versus HotMap, the participants were able to complete the tasks using both interfaces in similar times. One would expect that the task completion times for HotMap would improve as the users become more experienced using these systems.
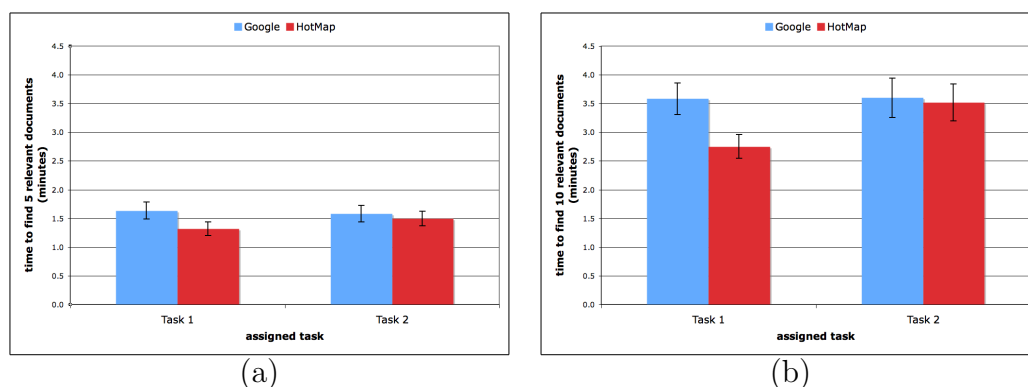
Figure 6.2: The average time to find five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 6.5: ANOVA tests for the time to completion data show that the differences between the task completion times are not statistically significant except for finding 10 documents in Task 1.

	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1, 54) = 2.94, p = 0.09$	$\mathbf{F(1, 54) = 5.74, p = 0.02}$
Task 2	$F(1, 54) = 0.25, p = 0.61$	$F(1, 54) = 0.03, p = 0.87$

Perceived Precision

Two common metrics employed in information retrieval research are precision and recall. *Precision* is the ratio of relevant documents retrieved to the total number of documents retrieved; *recall* is the ratio of relevant documents retrieved to the total number of relevant documents in the collection being searched [109]. Although Vaughan [106] indicates that it is very difficult to apply these metrics to the evaluation of Web information retrieval systems, Kobayashi and Takeda [52] suggest that only recall is unfeasible.

Instead of using the traditional precision metric which requires prior knowledge of document relevance, a new metric is defined that is inspired by the precision metric: *perceived precision*. In this metric, only the documents that

have been viewed by the users are considered as having been retrieved. Furthermore, the user's judgments of relevance are considered to be accurate and error-free. This metric measures the users' perceptions of their ability to find relevant documents in the search results, rather than the information retrieval system's ability to find relevant documents in the collections.

Perceived precision is a critical metric in the evaluation of Web search interfaces since it indicates what percent of the documents evaluated by the users were considered to be relevant to their information need. Clearly, an interface that results in a high perceived precision is of more value than one that results in a lower perceived precision. However, the definition of the traditional precision metric implies a binary distinction between relevant and non-relevant documents. When a multi-point relevance scale is used to measure relevance judgments, such as the four-point scale used in this study (see Table 6.1), this data must be partitioned into sets that represent relevant and non-relevant documents. In a technique similar to that used by [92], those documents that are given a relevance score of three or four are considered "relevant", and those given a relevance score of one or two are considered "not relevant". This results in the following formula for perceived precision:

$$pp = \frac{|r_3| + |r_4|}{|r_1| + |r_2| + |r_3| + |r_4|}$$

where $|r_i|$ represents the number of documents assigned a relevance score of i.

This partition method can be generalized to allow different levels of relevance to be used to represent "relevant" documents. For example, one might specify *strong perceived precision* using only documents that have been given a score of four as relevant, or *weak precision precision* using documents given a score of two, three, or four as relevant. However, it is not clear that participants
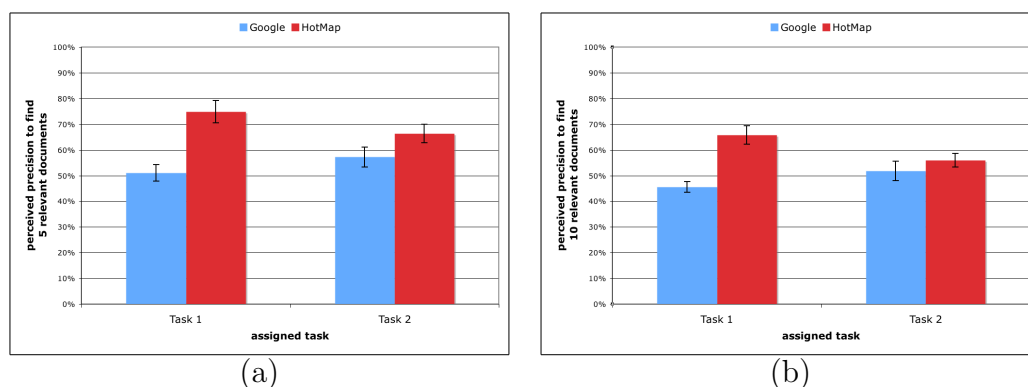
Figure 6.3: The average perceived precision (calculated assuming that there are no errors in the participants' relevance judgments) in finding five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 6.6: ANOVA tests show that the differences in the perceived precision for Task 1 are statistically significant, while that the differences in the perceived precision for Task 2 are not statistically significant.

	5 Relevant Documents	10 Relevant Documents
Task 1	$\mathbf{F(1, 54) = 19.88, p < 0.01}$	$\mathbf{F(1, 54) = 23.66, p < 0.01}$
Task 2	$F(1, 54) = 3.06, p = 0.09$	$F(1, 54) = 0.81, p = 0.37$

are able to accurately distinguish among the range of available relevance scores. Therefore, this discussion focuses on a *medium perceived precision*, which uses documents that have been given a score of three or four as relevant.

Figure 6.3 illustrates the average perceived precision the participants achieved in finding five relevant and ten relevant documents for the two assigned tasks. On average, participants were able to find a higher ratio of relevant documents using HotMap than Google. For Task 1, the differences in the perceived precision scores proved to be statistically significant; for Task 2, the differences were not be statistically significant. The results of ANOVA tests are provided in Table 6.6.

Corrected Precision

One thing that is not clear with user studies involving participant-provided relevance judgments is how to handle errors when calculating perceived precision metrics. In traditional information retrieval research, test collections such as TREC, Cranfield, and CACM are used since they include expert evaluations of the relevance of each document with respect to each test query. The use of these test collections and test queries allow the researchers to calculate the precision and recall metrics for their information retrieval algorithms in a controlled manner. However, in this research, it is not always the case that the participants will be able to make error-free relevance judgments. For example, if a participant provides a relevance score of four to a document that is not relevant, then the perceived precision metric will not be accurate.

To address this issue, a new metric called *corrected precision* is defined. Whereas the perceived precision metric is based solely on the participants' relevance scores, the corrected precision uses expert evaluation scores to correct each participant's relevance score assignments. If a participant's relevance scores differ from the expert relevance scores by zero or one (i.e., they are very similar), the participant's relevance score is used. In all other cases, where the difference is two or three (i.e., the scores are in disagreement), a relevance score of 1 is used. This corrected precision gives a more accurate indicator of actual effectiveness in the Web search tasks since only documents that are considered relevant by both the participant and the expert are considered relevant for the corrected precision metric. If the difference between these two precision metrics is large, this is an indication of significant user error (errors are discussed in more detail in the following Section).

Figure 6.4 illustrates the average corrected precision the participants achieved in finding five relevant and ten relevant documents for the two assigned tasks.
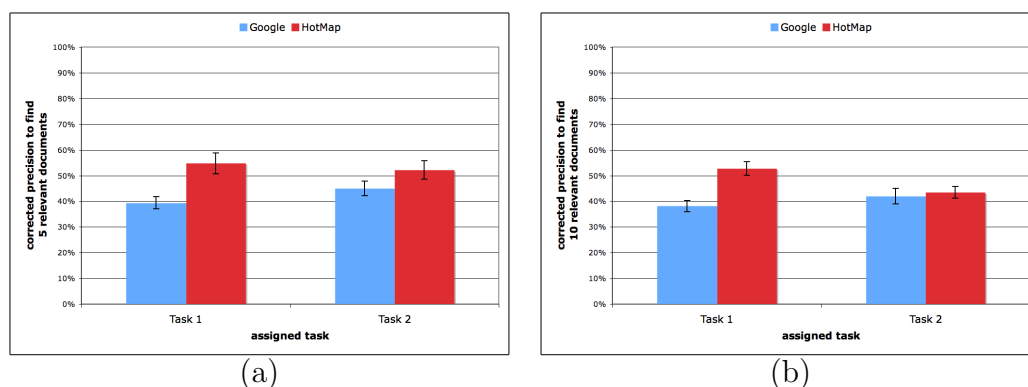
Figure 6.4: The average corrected precision (calculated using the participants' relevance judgments that are in agreement with the expert relevance judgments) in finding five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 6.7: ANOVA tests for the corrected precision data show that the differences in the precision for Task 1 are statistically significant, and that the differences in the precision for Task 2 are not statistically significant.

	5 Relevant Documents	10 Relevant Documents
Task 1	$\mathbf{F(1, 54) = 10.42, p < 0.01}$	$\mathbf{F(1, 54) = 17.31, p < 0.01}$
Task 2	$F(1, 54) = 2.42, p = 0.13$	$F(1, 54) = 0.15, p = 0.70$

On average, participants were able to find a higher ratio of relevant documents using HotMap, even when corrected for errors. For Task 1, the differences in the corrected precision scores proved to be statistically significant; for Task 2, the differences proved to not be statistically significant. The results of ANOVA tests are provided in Table 6.7.

Comparing the perceived and corrected precision in Figures 6.3 and 6.4, one can see that the general trends in these precision measures are similar. This implies that regardless of the users' abilities to evaluate documents for relevance, the visual representations and interactive exploration features of HotMap help to support the users in exploring the search results to find relevant documents.

The primary hypothesis in the development of HotMap is that participants would need to consider fewer non-relevant documents using HotMap than Google, for finding 5 and 10 relevant documents. The perceived precision and corrected precision metrics provide evidence in support of this hypotheses. High values in these metrics mean that fewer non-relevant documents were considered; low values result from many non-relevant documents being considered. On average, HotMap resulted in the participants considering fewer documents, both when seeking 5 and 10 relevant documents. This proved to be the case for both tasks, although the results were statistically significant only for Task 1.

The differences in statistical significance between Tasks 1 and 2 can be attributed to the fundamental differences between the two tasks. Although both tasks were chosen to be similar in their degree of ambiguity, the end result was that two different sets of document surrogates were evaluated by the participants. The participants may have found it easier to use HotMap to explore the search results from Task 1 than Task 2. While it is difficult to generalize the results from these two tasks to all possible Web search tasks, it is clear that in some situations, there can be a significant improvement in the perceived and corrected precision, resulting in fewer non-relevant documents being considered.

Error Rates

Calculating error rates in the relevance judgments of the participants is an important part of analysing the results of a comparative study of competing Web search interfaces. Under controlled conditions, the differences between relevance judgement error rates is a direct consequence of the interface to which the participants were exposed.

Table 6.8: The error types can be classified based on the difference between the expert relevance scores and the participant relevance scores.

	Major Error	Minor Error
Miss	-3	-2
False Hit	3	2

Using the same expert relevance scores as were produced to determine the potential effectiveness of query term re-sorting (see Section 6.2), two different classes of errors can be specified: *misses* and *false hits*. Misses occur when participants assign low relevance scores to documents that were labeled as highly relevant by experts. These errors result in the users not choosing to view relevant documents. False hits occur when participants assign high relevance scores to documents that were labeled as not relevant by experts. These errors result in the users choosing to view documents that are not relevant. The different classes of errors can have different effects on the users' perceptions of the search results evaluation experience. Too many misses can result in the users either giving up, or at least re-formulating their queries. Too many false hits are frustrating, but users may be willing to continue searching anyway.

In addition, two different levels of error severity can be specified when a relevance scale is used: *major errors* and *minor errors*. For the four-point relevance scale in Table 6.1, major errors occur when the difference between the scores for a specific document is three (i.e., a 4-document assigned a score of 1, or a 1-document assigned a score of 4). Minor errors occur when the difference between the scores for a particular document is two. Although differences of one will also be present, these errors are less significant and may be the result of personal preferences rather than actual errors in relevance judgement. A summary of this classification of errors is provided in Table 6.8.

The average number of errors for each error type is provided in Figure 6.5.

(a) major miss errors

(b) minor miss errors

(c) major false hit errors

(d) minor false hit errors
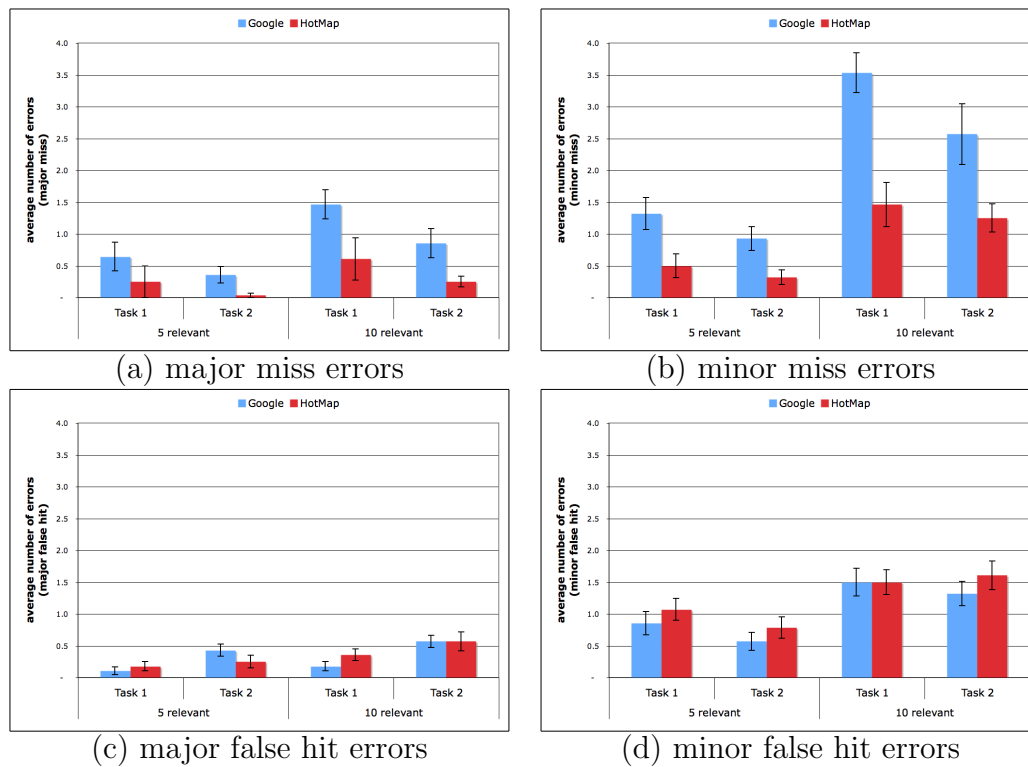
Figure 6.5: The average number of errors by classification type. The error bars represent the standard error about the mean.

As is to be expected, the number of errors occurring when finding 10 relevant documents is consistently higher than the number of errors occurring when finding 5 relevant documents. The number of miss errors, both major and minor, are significantly higher when using Google versus HotMap. Although there appear to be more false hit errors produced from HotMap in some cases, this did not prove to be statistically significant. A statistical analysis of this data is provided in Table 6.9. These analyses include the removal of one outlier data point resulting from a participant producing seven major miss errors on Task 1 using HotMap.

Given that the same information is being provided to users of both Google and HotMap, one might expect there to be little differences in the error rates.

Table 6.9: ANOVA tests for the error data show that the differences in the number of errors between HotMap and Google is statistically significant for the miss errors (when ignoring the outlier), and not statistically significant for the false hit errors.

Major Miss	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1,54) = 1.36, p = 0.25$	$\mathbf{F(1,54) = 4.49, p = 0.04}$
Task 1 (without outlier)	$\mathbf{F(1,52) = 8.21, p < 0.01}$	$\mathbf{F(1,52) = 21.68, p < 0.01}$
Task 2	$\mathbf{F(1,54) = 5.83, p = 0.02}$	$\mathbf{F(1,54) = 6.24, p = 0.02}$
Minor Miss	5 Relevant Documents	10 Relevant Documents
Task 1	$\mathbf{F(1,54) = 6.79, p = 0.01}$	$\mathbf{F(1,54) = 19.79, p < 0.01}$
Task 2	$\mathbf{F(1,54) = 7.75, p < 0.01}$	$\mathbf{F(1,54) = 6.28, p = 0.02}$
Major False Hit	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1,54) = 0.57, p = 0.45$	$F(1,54) = 2.29, p = 0.14$
Task 2	$F(1,54) = 1.71, p = 0.20$	$F(1,54) = 0.00, p = 1.00$
Minor False Hit	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1,54) = 0.73, p = 0.40$	$F(1,54) = 0.00, p = 1.00$
Task 2	$F(1,54) = 0.98, p = 0.33$	$F(1,54) = 0.92, p = 0.34$

Although the differences in false hits were negligible, the differences in misses were significant. The participants were more apt to give a relevant document a low relevance score using Google than HotMap, resulting in miss errors. While the reasons for this difference are not clear, it may stem from a general dissatisfaction with current Web search engines. This may result in an expectation of many non-relevant documents being present in the search results, and a general reluctance to view the documents that are perceived to be non-relevant (even if they are indeed relevant).

That the error rates produced using HotMap are not worse than Google (and in some case, they are better than Google) provides support in favour of the compact representation of the search results used by HotMap. Although some participants did not like that the compact representation required them to access the snippet for each document using the mouse and tool tips, this

feature did not result in a higher error rate.

Subjective Measures

After each task, participants completed a short in-task questionnaire to measure their subjective reactions to using the assigned interface as they completed the assigned task. Of interest were the participants' degree of confidence, feelings of ease of use, satisfaction in using the interface, and perceptions of ambiguity. The data from these subjective measures for both tasks is reported in Figure 6.6. The results of Friedman tests measuring statistical significance are reported in Table 6.10.

For the *confidence* measure, the participants rated how confident they were in their ability to find a good set of relevant documents (Figure 6.6a). For both tasks, the participants showed a higher degree of confidence using HotMap than Google. For Task 1, this measure is significantly more positive for HotMap. By providing the users with the ability to interactively explore the search results, they are able to take an active role in the search process, resulting in increased feelings of confidence and control.

For the *ease of use* measure, the participants rated how easy they found it to use the interface to evaluate the search results (Figure 6.6b). Although the participants reported a higher degree of ease of use with HotMap for both tasks, this is only statistically significant for Task 1. The interactive features of HotMap were designed to be simple to use, yet provide powerful abilities to explore the search results set. Clearly, many of the participants found these features simple to understand and operate.

For the *satisfaction* measure, the participants rated how satisfied they were in using the interface to evaluate the search results (Figure 6.6c). The satisfaction reported by the participants was significantly higher for HotMap than
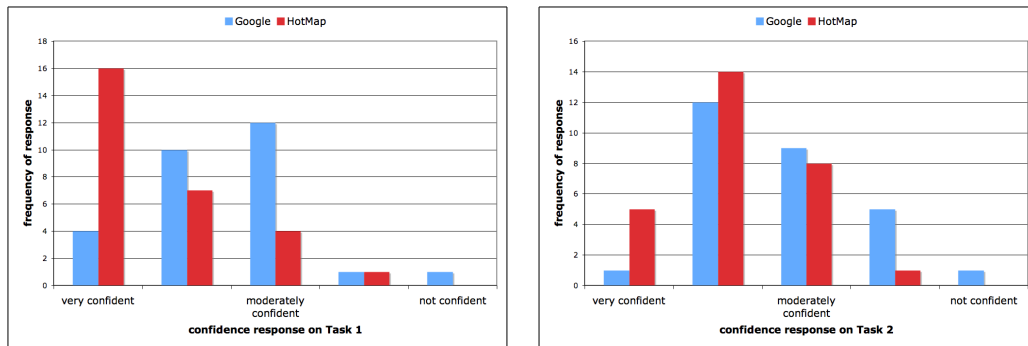
(a) confidence in ability to find a good set of relevant documents
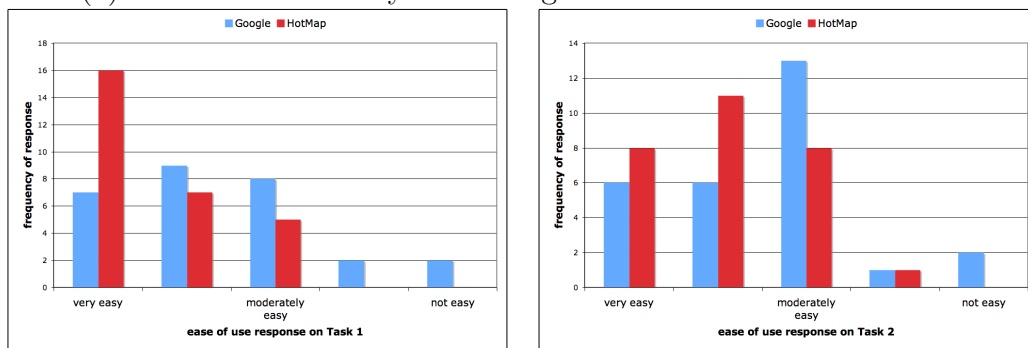
(b) ease of use for evaluating the search results
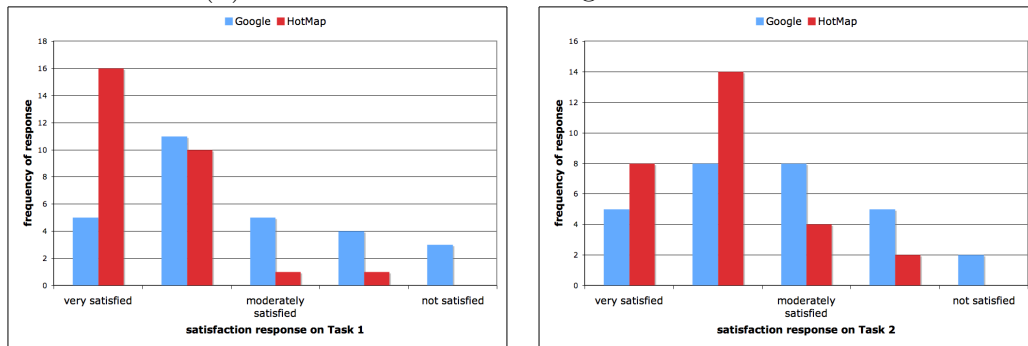
(c) satisfaction with using the interface to evaluate the search results
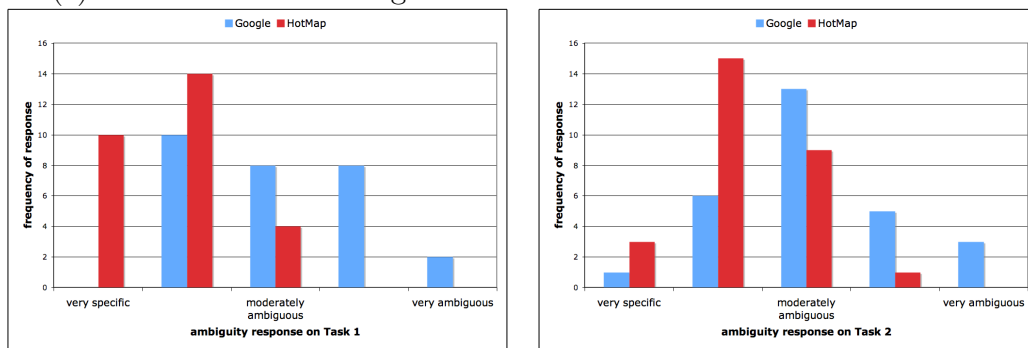
(d) ambiguity of the search results set

Figure 6.6: Subjective measures reported by the participants for Task 1 (left) and Task 2 (right). Note that the participants consistently provided more positive subjective rankings for HotMap.

Table 6.10: Friedman tests for the subjective reactions show that the differences in the subjective reactions between HotMap and Google are statistically significant for Task 1, but not so for Task 2.

Measure	Task 1	Task 2
confidence	$\chi^2(1) = \mathbf{12.25}, \mathbf{p} < \mathbf{0.01}$	$\chi^2(1) = 2.88, p = 0.09$
ease of use	$\chi^2(1) = \mathbf{8.07}, \mathbf{p} < \mathbf{0.01}$	$\chi^2(1) = 3.56, p = 0.06$
satisfaction	$\chi^2(1) = \mathbf{8.00}, \mathbf{p} < \mathbf{0.01}$	$\chi^2(1) = 2.91, p = 0.09$
ambiguity	$\chi^2(1) = \mathbf{13.76}, \mathbf{p} < \mathbf{0.01}$	$\chi^2(1) = \mathbf{8.05}, \mathbf{p} < \mathbf{0.01}$

Google on Task 1. For Task 2, there was an improvement in satisfaction, but not significantly so. As with the confidence measure, the ability to explore the search results using HotMap had a positive effect on the feelings of satisfaction from the users.

For the *ambiguity* measure, the participants rated how ambiguous they thought the search result set was (Figure 6.6d). Since the goal of HotMap was to direct the users towards more relevant documents, this measure provides an indication of the success of this goal, considering that the search results sets were identical for each interface. In both tasks, the participants reported the search results to be much more specific using HotMap than Google. This result is statistically significant for both tasks.

For most of these measures, the participants reported a significant improvement in Task 1. For Task 2, while an improvement is apparent, it is not statistically significant. As with the differences in the perceived and corrected precision metrics, this difference between the two tasks can be attributed to the participants' abilities to find relevant documents for each task. Clearly, the participants found HotMap to be very beneficial when completing Task 1. However, for Task 2, not only did the participants find it more difficult to identify relevant documents, this difficulty was also made apparent in the

Table 6.11: The majority of the participants ranked the use of HotMap above that of Google after completing the tasks. However, these results are not statistically significant

Rank	Number of Responses
HotMap ranked higher than Google	19 (68%)
HotMap ranked lower than Google	9 (32%)
Wilcoxon signed rank test	$Z = -1.89, p = 0.06$

subjective measures. Although the results for Task 2 did not prove to be statistically significant, on average, the participants were still able to perform better on this task using HotMap.

Preference Rank

After all the tasks were completed by the participants, a post-task questionnaire was administered which included a question asking the participants to rank their preference for a search results interface, assuming that the underlying search results sets are the same. These rank responses are reported in Table 6.11.

Although a large portion of the participants ranked HotMap higher than Google in terms of their interface preference, this result did not prove to be statistically significant. A number of participants noted that they ranked Google as their preferred interface because of their prior familiarity with it, resulting in a skewing of these results in favour of Google.

6.3.5 Discussion

While on average, the participants were able to complete the tasks of finding five and ten relevant documents faster using HotMap than Google, this result did not prove to be statistically significant. However, it is interesting to note

that the extra time taken to visually inspect the search results, as well as make nested sorting decisions, did not adversely affect the time taken to find relevant documents.

The ability of the participants to find relevant documents was evaluated using both perceived precision (measuring the portion of relevant documents among those that were considered by the participant) and corrected precision (taking into account the expert relevance scores to eliminate major errors in the relevance judgments). In both of these measures of precision, HotMap proved to be superior. However, the difference between Google and HotMap was only statistically significant for Task 1.

Error rates were evaluated for each participant using the expert evaluations of the search results. The participants missed significantly fewer relevant documents using HotMap than Google. While there were more false hits produced when using HotMap in some cases, this result was not shown to be significant.

As the participants completed the Web search tasks, subjective data on the topics of confidence, ease of use, satisfaction, and ambiguity were collected. For Task 1, the responses to all four measures were significantly more positive when using HotMap versus the responses provided when using Google. There was also an improvement in the responses for Task 2, although only the ambiguity response proved to be statistically significant. When asked to rank their preference for a Web search interface, 68% of the participants preferred HotMap.

From this user study, it was found that the features of HotMap can indeed be used effectively by many participants to find more relevant documents from the search results. Although the results were shown to be statistically significant for Task 1, many of the results were not significant for Task 2. The

difference in statistical significance can be attributed to the fundamental differences between the two tasks. Although both were chosen to be vague search tasks, it is possible that finding relevant documents in Task 2 was more difficult.

It should be noted that in this user evaluation, all the participants reported a moderate to high degree of Web search experience, and all except one used Google as their primary Web search engine. By contrast, none had previously been exposed to HotMap. Therefore, this study compared experienced users of Google to novice users of HotMap, resulting in a bias in favour of Google. This bias was especially apparent when the participants were asked to rank their preferred Web search interface. That the participants were able to perform better using HotMap in many cases is promising. One can reasonably expect that as users become more experienced with the features of HotMap, they will be able to more effectively make use of these features to interactively explore Web search results.

The results of these studies provide evidence in support of the fundamental assumption in the design of HotMap: that visual and interactive tools that make use of the query term frequency statistics from the top search results can provide valuable information to the users for re-sorting and exploring the search results.

6.4 Determining the Potential Effectiveness of Concept Highlighter

In Concept Highlighter, the user selects one or more concepts that are relevant to their information need, resulting in a re-sorting of the search results based

of the fuzzy similarity of each document surrogate to the selected concepts. The assumption is that document surrogates that are similar to the selected relevant concepts will also be relevant. As with the evaluation of HotMap, vague queries are used since their search results commonly consist of a mixture of relevant and non-relevant documents. It may be possible to move more relevant documents near the top of the list by re-sorting the search results based on their fuzzy membership scores to selected relevant concepts.

In this section, the methods for testing this fuzzy similarity assumption are described, along with the experimental results.

6.4.1 Methods

To test these hypotheses, the same four vague search tasks developed for the evaluation of HotMap were used, along with their search results sets and the expert evaluations for relevance to the task. The reader is referred to Section 6.2.1 for the details on these tasks and the methods for providing expert relevance scores for the search results sets.

All 2^n possible combinations of concept selections were evaluated for the test queries, where n is the number of concepts provided by the concept knowledge base for the query. There were eight concepts provided for Task 1; seven concepts for Task 2; eight concepts for Task 3; and five concepts for Task 4. For each of the concept selections, the number of relevant documents within the top 10, top 20, and top 30 documents was calculated using the expert evaluations. Any document assigned a score of three or four on the four-point relevance scale (see Table 6.1) was considered "relevant". This represents those documents that were identified as "relevant" and "probably relevant" by the experts.

Table 6.12: The potential effectiveness of re-sorting the search results using fuzzy membership scores in Concept Highlighter.

Number of sort methods...	10 Documents	20 Documents	30 Documents
Task 1: "communications network management"			
better than Google	16 (6%)	0 (0%)	5 (2%)
equivalent to Google	49 (19%)	1 (0%)	14 (6%)
worse than Google	189 (74%)	253 (100%)	235 (92%)
Task 2: "representing knowledge information"			
better than Google	19 (15%)	3 (2%)	30 (24%)
equivalent to Google	17 (13%)	3 (2%)	10 (8%)
worse than Google	90 (72%)	120 (96%)	86 (86%)
Task 3: "communication mechanism processes"			
better than Google	0 (0%)	0 (0%)	0 (0%)
equivalent to Google	1 (0%)	0 (0%)	5 (2%)
worse than Google	253 (100%)	254 (100%)	249 (98%)
Task 4: "addressing resources networks"			
better than Google	15 (50%)	0 (0%)	0 (0%)
equivalent to Google	6 (20%)	1 (3%)	1 (3%)
worse than Google	9 (30%)	29 (97%)	29 (97%)

6.4.2   Results

Table 6.12 summarizes the results of this experiment for all four tasks. The results are reported separately for each query since the effectiveness of concept-based fuzzy clustering depends primarily on the concepts chosen from the concept knowledge base for the query and the search results returned for each search task.

The results from this experiment are somewhat mixed. When considering the top 10 documents, there were multiple ways of selecting concepts that resulted in more relevant documents at the top of the list for all tasks except Task 3. But for 20 and 30 documents, the concept-based fuzzy clustering method was not effectiveness for any of the tasks.

Table 6.13: Comparing the number of relevant documents in the Google order to the best order achieved by concept-based fuzzy clustering in Concept Highlighter.

	10 Documents	20 Documents	30 Documents
Task 1: "communications network management"			
Google	4	12	16
Best Concept Selection	6	12	17
percent improvement	50%	0%	6%
Task 2: "representing knowledge information"			
Google	6	13	15
Best Concept Selection	8	15	21
percent improvement	33%	15%	40%
Task 3: "communication mechanism processes"			
Google	6	13	16
Best Concept Selection	6	11	16
percent improvement	0%	-15%	0%
Task 4: "addressing resources networks"			
Google	3	9	11
Best Concept Selection	5	9	11
percent improvement	67%	0%	0%

To further explore the potential benefit of the concept-based fuzzy clustering methods, the results from the best concept selection were compared to the search results order provided by Google. These results are provided in Table 6.13. There was as much as a 67% improvement in the number of relevant documents in the top 10 search results for the best selection of concepts. Although for one task there was no improvement, the average improvement for the best set of concepts was 38% for the top 10 documents. When considering the top 20 or top 30 documents, the improvement was good in some cases and negligible in others. In one case, the best concept selection resulted in a decrease in the number or relevant documents when compared to Google.

Since there are multiple ways of selecting concepts to achieve a superior order of the top 10 search results for most of the test queries, this can be

exploited by the users of the system as they seek relevant documents. For example, due to the ease of concept selection, the users can readily make a selection of relevant documents, and consider the top ranked search results. They can then make a change to their concept selection, and consider other documents that they have not previously seen. This exploration process may allow the users to find more relevant documents than if they make an initial selection of relevant concepts and do not re-visit this selection as they seek relevant documents from the search results list.

The ability for concept highlighter to improve the order of the search results and support the users as they explore the search results depends greatly on a number of factors. The concept knowledge base used in these evaluations was based on computer science topics provided by the ACM Computing Classification System [1]. Although the descriptions for each concept were clear and concise, the vocabulary was somewhat limited. This coupled with the limited information provided for each document surrogate (i.e., the title and snippet) may have resulted in a mismatch between the terminology in the concept knowledge base and that available in the document surrogates. As a result, the concept-based fuzzy clustering algorithm may have provided misleading fuzzy membership scores in some cases. The capabilities of Concept Highlighter can be improved by improving the quality and depth of the concept knowledge base, and by obtaining more descriptive information for each document surrogate in the search results.

Although this evaluation has shown that for most of the tasks there exist concept selections that provide more relevant documents among the top 10 search results, it is not clear that a user of this system will be able to make the proper concept selection decisions as they explore the search results. Clearly, there are many concept selections that result in an order that is worse than

that provided by Google (see Table 6.12). To explore the ability of the users to make good decisions with this interface and the concept information provided, a user evaluation has been conducted. The results of this user evaluation are reported in the following section.

6.5 Concept Highlighter vs. Google

Although it has been shown that it may be possible to achieve a higher number of relevant documents among the top documents in the search results for vague queries through concept-based fuzzy clustering, it is not clear whether the users will be able to effectively select the proper concepts in order to explore the search results and find relevant documents. To test the effectiveness of the features of Concept Highlighter, a user study was conducted. The methods, procedure, and analysis are the same as the user study for HotMap (see Section 6.3). The details are repeated in this Section for completeness.

6.5.1 Methods

To test the efficiency, effectiveness, and subjective reaction when conducting information seeking tasks, a 4x2 (interface x task) within-subjects design was employed. Although two other interfaces for representing and exploring search results were also evaluated in this study (one being HotMap, which is discussed in Section 6.3), the focus is on the comparison between Concept Highlighter and Google. Since all the participants were already frequent users of Google, they were exposed to this interface first. In order to reduce the biasing effects, the task order and remaining interface order were varied among the participant pool.

The first two tasks in Section 6.2.1 were used in this experiment. To ensure

that the interfaces provided the same set of search results to each participant, the search results for each task were cached and provided to each interface. These cached results were the same as those evaluated by the expert reviewers in the potential effectiveness analysis in the previous section. The Google interface was altered to include document numbers to facilitate data collection.

6.5.2 Procedure

Each participant completed a pre-task questionnaire, two searches with each of the interfaces, an in-task questionnaire following each search, and a post-task questionnaire after all the search tasks were complete. The entire procedure took between 60 and 90 minutes for each participant.

Before the participants were exposed to an interface, a short training task was provided, along with a brief description of the features of the interface. This ensured that each participant had a basic understanding of the interface, how it was to be used in the search task, and how the data was to be collected.

Each search task included a written description of the information need, along with the query to be used. After submitting the search to the assigned interface, the participants were asked to use the interface to evaluate the search results. For each document surrogate considered, a relevance score on a scale from 1 to 4 was assigned (see Table 6.1). The participants were asked to speak these scores; this information was logged by the investigator, along with the elapsed time. Only the document surrogates were considered for relevance; the participants were asked to not view the actual documents. After ten document surrogates were assigned a relevance score of either three or four, the participants were instructed that the task was complete.

As the participants used Concept Highlighter, they were instructed to first select the concepts they thought were relevant to the assigned task, and then

evaluate the re-ordered search results. If at any time they wished to make changes to their previous concept selections, they were encouraged to do so.

At the end of each task, the participants were provided with an in-task questionnaire to measure their subjective reactions and feelings regarding their experience with using the assigned interface to find documents relevant to the assigned task. These subjective measures were based on the participants' confidence in finding a good set of documents, the ease of use of the interface, satisfaction in using the interface, and impressions of ambiguity in the search results set.

After all the search tasks were completed, a post-task questionnaire was administered to measure the participants' impressions of various features of the interfaces. In addition, the participants were asked to provide a ranking of their preference for an interface, and textual comments regarding the features of the interfaces.

6.5.3 Analysis

As with the HotMap user evaluations, the quantitative results (time to task completion, perceived and corrected precision, and error rates) were analysed independently for each task using analysis of variance (ANOVA). The subjective evaluations were analysed using nonparametric Friedman tests. The preference ranks were analysed using pair-wise Wilcoxon signed ranks tests. Where relevant, the statistical significance of these tests are highlighted by a bold font.

6.5.4 Results

Participant Demographics

Since the evaluation of Concept Highlighter occurred simultaneously with the evaluation of HotMap, the participant groups were the same for both studies. The participant demographics, along with the participant familiarity with the tasks, are reported in Section 6.3.4.

Time to Task Completion

As with the HotMap user study, two levels of fulfillment of the assigned information need were specified: finding five relevant documents, and finding ten relevant documents. In this criteria, any document assigned a relevance score of three or four was considered as a relevant document. This represented documents the participants felt were either certainly or probably relevant (see Table 6.1 for the complete relevance score scale).

Figure 6.7 illustrates the average time the participants took to find five and ten relevant documents for each of the two tasks. On average, the participants were able to find relevant documents faster using Google than with Concept Highlighter. This is to be suspected, given that the users of Concept Highlighter needed to first make concept selections before they started evaluating the search results. These results did not prove to be statistically significant (see Table 6.14). This lack of statistical significance is promising since it implies that the time required by the participants to use the features of Concept Highlighter is not that much greater than the time required to use Google. It is possible that the extra time taken to make concept selections (and subsequent re-selections) is offset by the users' ability to more easily evaluate and explore the search results.
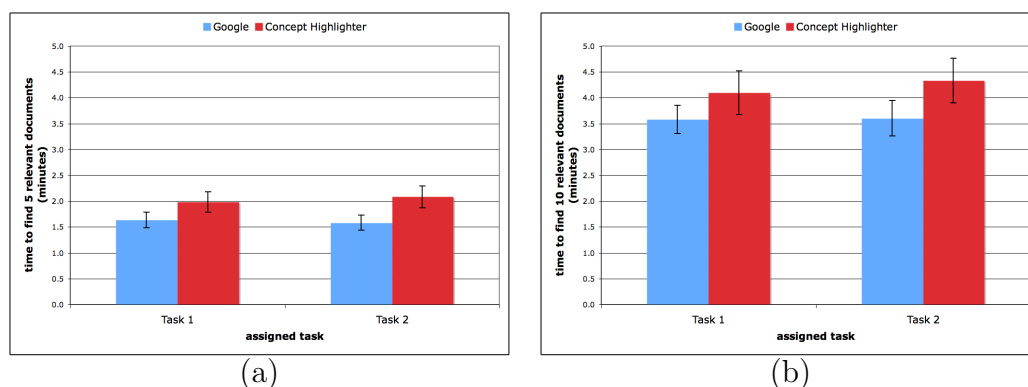
Figure 6.7: The average time to find five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 6.14: ANOVA tests for the time to completion data show that the differences between the task completion times are not statistically significant for either task.

	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1, 54) = 1.89, p = 0.18$	$F(1, 54) = 1.06, p = 0.31$
Task 2	$F(1, 54) = 3.70, p = 0.06$	$F(1, 54) = 1.74, p = 0.19$

Perceived Precision

The perceived precision metric, as described in Section 6.3.4, considers only the documents that have been viewed by the user as having been retrieved, and assumes that the users' judgments of relevance are always correct. This metric measures the users' ability to find relevant documents in the search results, rather than the information retrieval system's ability to find relevant documents in the collections. An interface that results in a high perceived precision is of more value than one that results in a lower perceived precision.

Figure 6.8 illustrates the average perceived precision the participants achieved in finding five relevant and ten relevant documents for the two assigned tasks. On average, participants were able to find a higher ratio of relevant documents using Concept Highlighter than Google, both for finding five and ten relevant
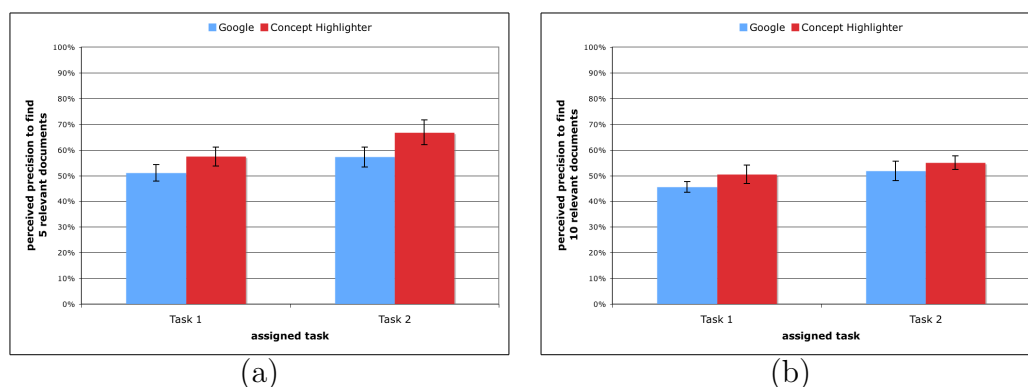
Figure 6.8: The average perceived precision (calculated assuming that there are no errors in the participants' relevance judgments) in finding five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 6.15: ANOVA tests show that the differences in the perceived precision for both tasks are not statistically significant.

	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1, 54) = 1.71, p = 0.20$	$F(1, 54) = 1.61, p = 0.21$
Task 2	$F(1, 54) = 2.40, p = 0.13$	$F(1, 54) = 0.44, p = 0.51$

documents. However, these results are not statistically significant for either task, as shown in Table 6.15.

Corrected Precision

Corrected precision was defined in Section 6.3.4. Whereas the perceived precision metric is based solely on the participants' relevance scores, the corrected precision uses expert evaluation scores to correct each participant's relevance scores. If a participant's relevance scores differ from the expert relevance scores by zero or one (i.e., they are very similar), the participant's relevance score is used. In all other cases, where the difference is two or three (i.e., the scores are in disagreement), a relevance score of 1 is used. This corrected precision gives a more accurate indicator of actual effectiveness in the Web search tasks

(a) (b)
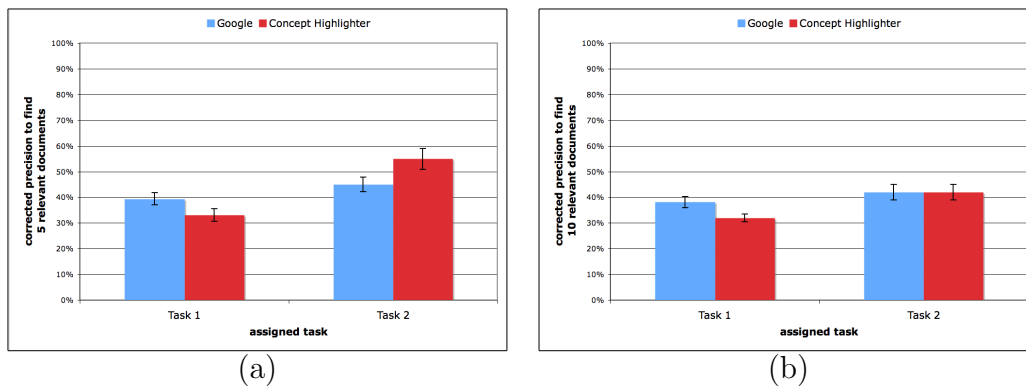
Figure 6.9: The average corrected precision (calculated using the participants' relevance judgments that are in agreement with the expert relevance judgments) in finding five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 6.16: ANOVA tests for the corrected precision data show that the results are rather varied.

	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1, 54) = 3.51, p = 0.07$	$\mathbf{F(1, 54) = 5.11, p = 0.03}$
Task 2	$\mathbf{F(1, 54) = 3.99, p = 0.05}$	$F(1, 54) = 0.00, p = 0.99$

since only documents that are considered relevant by both the participant and the expert are considered relevant for the corrected precision metric. If the difference between these two precision metrics is large, this is an indication of significant user error (errors are discussed in more detail in the following Section).

Figure 6.9 illustrates the average corrected precision the participants achieved in finding five and ten relevant documents for the two assigned tasks. Whereas the participants were able to achieve a higher average perceived precision using Concept Highlighter (as illustrated in Figure 6.8), the results are more varied for the corrected precision measure. For Task 1, the corrected precision shows that the participants performed worse using Concept Highlighter both for finding five and ten relevant documents. For Task 2, the participants still

performed better than Google for finding five relevant documents, but there is little difference for finding ten relevant documents. In terms of statistical significance, the participants performed significantly worse than Google in Task 1 for finding ten relevant documents, and significantly better than Google in Task 2 for finding five relevant documents (see Table 6.16).

The differences between the perceived and corrected precision metrics suggest that the participants made significant errors in deciding the relevance scores for some of the documents they considered. Based solely on their relevance scores, it appeared that the participants considered fewer documents using Concept Highlighter in order to find five and ten relevant documents. However, in some cases, many of these documents the participants thought were relevant while using Concept Highlighter were identified as being not relevant by the expert reviewers, resulting in a substantial reduction in the corrected precision. This casts some doubt on the effectiveness of Concept Highlighter, and merits a further analysis of the error rates.

Error Rates

In the evaluation of HotMap, two different types of errors were defined, at two different levels of severity (see Section 6.3.4). Misses occur when participants assign low relevance scores to documents that were labeled as highly relevant by experts. False hits occur when participants assign high relevance scores to documents that were labeled as not relevant by experts. For the four-point relevance scale used in these studies (as shown in Table 6.1), major errors occur when the difference between the scores for a specific document is three (i.e., a 4-document assigned a score of 1, or a 1-document assigned a score of 4). Minor errors occur when the difference between the scores for a particular document is two. A summary of this classification of errors is provided in

(a) Major Miss Errors

(b) Minor Miss Errors

(c) Major False Hit Errors

(d) Minor False Hit Errors
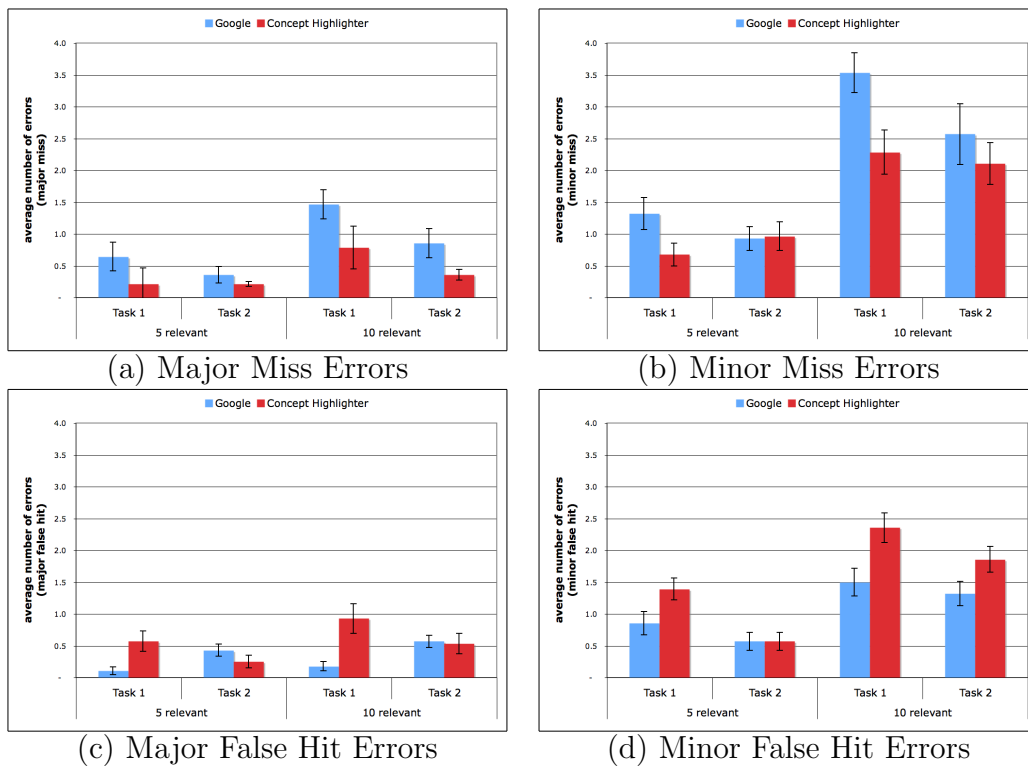
Figure 6.10: The average number of errors by classification type. The error bars represent the standard error about the mean.

Table 6.8.

The average number of errors for each error type are provided in Figure 6.10; the statistical analysis is reported in Table 6.17. As is to be expected, the number of errors occurring when finding 10 relevant documents is consistently higher than the number of errors occurring when finding 5 relevant documents. In all cases except one, the participants made more miss errors, both major and minor, using Google versus Concept Highlighter. However, this was only statistically significant for major miss errors in finding ten relevant documents for Task 1.

Table 6.17: ANOVA tests for the error data show that the differences in the number of errors between Concept Highlighter and Google are statistically significant primarily for Task 1.

Major Miss	5 Relevant Documents	10 Relevant Documents
Task 1	$F(1, 54) = 2.95, p = 0.09$	$\mathbf{F(1, 54) = 5.30, p = 0.03}$
Task 2	$F(1, 54) = 0.61, p = 0.44$	$F(1, 54) = 3.38, p = 0.07$
Minor Miss	5 Relevant Documents	10 Relevant Documents
Task 1	$\mathbf{F(1, 54) = 4.32, p = 0.04}$	$\mathbf{F(1, 54) = 7.26, p < 0.01}$
Task 2	$F(1, 54) = 0.01, p = 0.90$	$F(1, 54) = 0.64, p = 0.43$
Major False Hit	5 Relevant Documents	10 Relevant Documents
Task 1	$\mathbf{F(1, 54) = 7.57, p < 0.01}$	$\mathbf{F(1, 54) = 9.25, p < 0.01}$
Task 2	$F(1, 54) = 1.71, p = 0.20$	$F(1, 54) = 0.04, p = 0.85$
Minor False Hit	5 Relevant Documents	10 Relevant Documents
Task 1	$\mathbf{F(1, 54) = 4.51, p = 0.03}$	$\mathbf{F(1, 54) = 7.17, p < 0.01}$
Task 2	$F(1, 54) = 0.00, p = 1.00$	$F(1, 54) = 3.64, p = 0.06$

In many cases, the participants produced more false hit errors using Concept Highlighter versus Google. This increase in false hits proved to be statistically significant for Task 1, both for major and minor false hits, as well as for finding five and ten relevant documents. The high number of false hits is an indication that for Task 1, many of the participants gave high relevance scores for documents that the expert reviewers did not consider relevant. The significantly higher number of errors on Task 1 resulted in a major reduction from perceived to corrected precision for Task 1 (see Figures 6.8 and 6.9).

Although it was thought that there would be little difference in the error rates between Google and Concept Highlighter, clearly, this was not the case. In some instances, there was a significant reduction in misses when using Concept Highlighter, but also a significant increase in false hits. However, all the cases where the differences were statistically significant were from Task 1, indicating that difficulties or misunderstandings of this task may have had an effect on the error rates. Subsequently, these error rates had the effect of

significantly reducing the corrected precision metrics.

The general trend of participants missing fewer documents using Concept Highlighter provides support in favour of the compact representation of the search results used by Concept Highlighter. As with HotMap, although some participants did not like that the compact representation required them to access the snippet for each document using the mouse and tool tips, this feature did not appear to result in the participants missing relevant documents. However, this may have also contributed to the increase in false hit errors.

The general trend of participants indicating more non-relevant documents as relevant using Concept Highlighter may also stem from an increase in the expectation of documents being relevant. Because the participants first selected one or more concepts that are relevant to the assigned task, this may have given the participants the impressions that they were going to find more relevant documents. As a result, marginal documents may have appeared to be more relevant since they were moved to a more prominent location in the list.

Subjective Measures

After completing each search task, participants completed a short in-task questionnaire to measure their subjective reactions to using the assigned interface to find documents relevant to the assigned task. Of interest were each participant's degree of confidence, feelings of ease of use, satisfaction in using the interface, and perceptions of ambiguity. The data from these subjective measures for both tasks is reported in Figure 6.11. The results of Friedman tests measuring statistical significance are reported in Table 6.18.

For the *confidence* measure, the participants rated how confident they were in their ability to find a good set of relevant documents (Figure 6.11a). For

(a) confidence in ability to find a good set of relevant documents

(b) ease of use for evaluating the search results

(c) satisfaction with using the interface to evaluate the search results

(d) ambiguity of the search results set
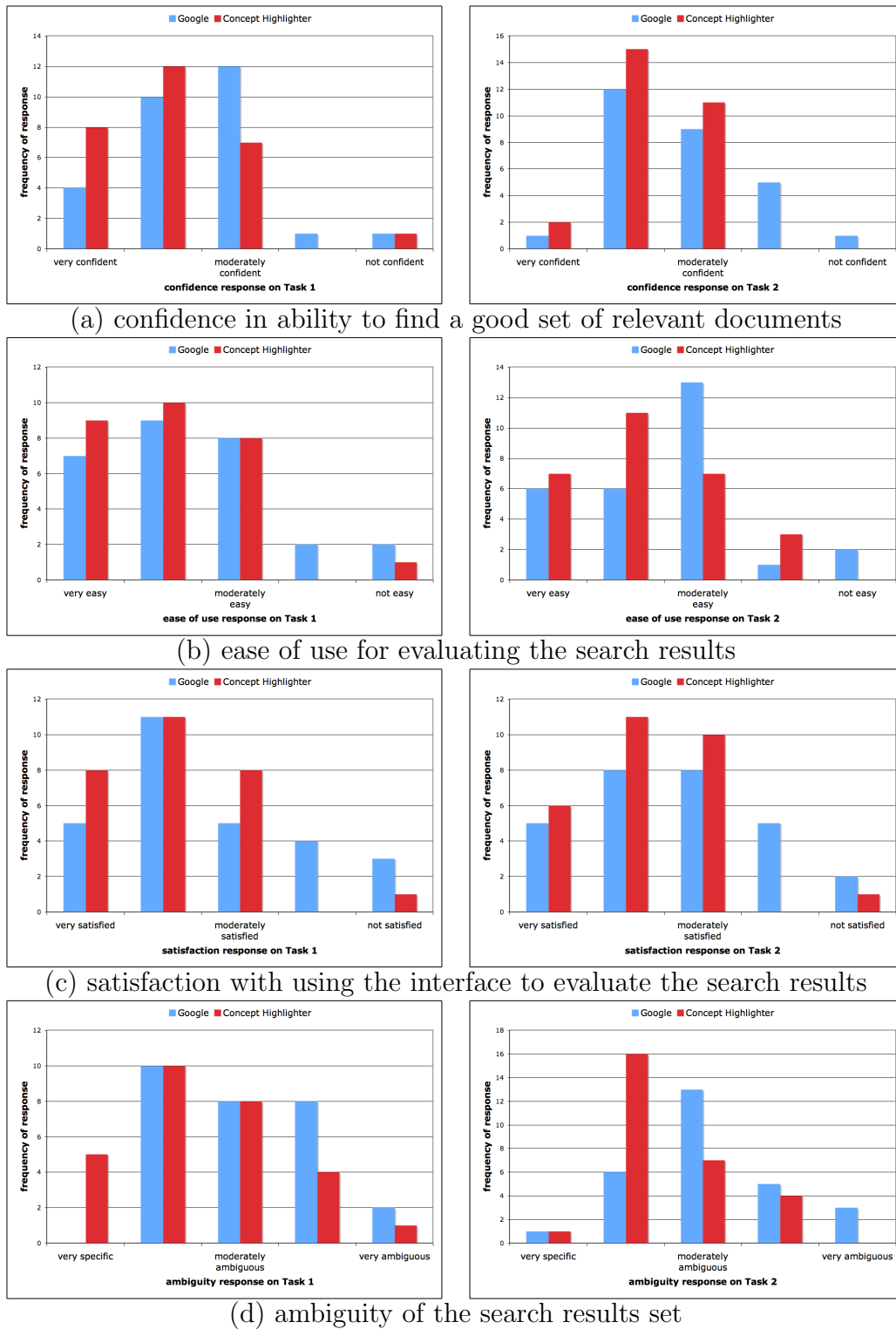
Figure 6.11: Subjective measures reported by the participants for Task 1 (left) and Task 2 (right). Note that the participants consistently provided more positive subjective responses for Concept Highlighter.

Table 6.18: Friedman tests for the subjective reactions show that the differences in the subjective reactions between Concept Highlighter and Google were only statistically significant for the satisfaction measure in Task 2, and the ambiguity measure for both tasks.

Measure	Task 1	Task 2
confidence	$\chi^2(1) = 2.88, p = 0.09$	$\chi^2(1) = 2.33, p = 0.13$
ease of use	$\chi^2(1) = 1.14, p = 0.29$	$\chi^2(1) = 2.00, p = 0.16$
satisfaction	$\chi^2(1) = 1.32, p = 0.25$	$\mathbf{\chi^2(1) = 3.77, p = 0.05}$
ambiguity	$\mathbf{\chi^2(1) = 5.56, p = 0.02}$	$\mathbf{\chi^2(1) = 5.76, p = 0.02}$

both tasks, the participants showed a higher degree of confidence using Concept Highlighter than Google. However, these results did not prove to be statistically significant. The higher confidence with Concept Highlighter can be attributed to the support this system provides to the users as they interactively explore the search results. This active role that the users take in the search process results in increased feelings of confidence and control.

For the *ease of use* measure, the participants rated how easy they found it to use the interface to evaluate the search results (Figure 6.11b). Although the participants reported a higher degree of ease of use with Concept Highlighter for both tasks, this difference did not prove to be statistically significant for either task. It is promising that even though the participants were required to perform an extra concept selection step before they could evaluate the search results, they still found the interface at least as easy to use as Google. One can expect that as they become more accustomed to the features of Concept Highlighter, this ease of use measure will increase further.

For the *satisfaction* measure, the participants rated how satisfied they were

in using the interface to evaluate the search results (Figure 6.11c). The satisfaction reported by the participants was significantly higher for Concept Highlighter than Google on Task 2. For Task 1, there was an improvement in satisfaction, but not significantly so. As with the confidence measure, the ability to explore the search results using Concept Highlighter had a positive effect on the feelings of satisfaction from the users.

For the *ambiguity* measure, the participants rated how ambiguous they thought the search result set was (Figure 6.11d). Since the goal of Concept Highlighter was to direct the users towards more relevant documents through relevant concept selection, this measure provides an indication of the success of this goal, considering that the search results sets were identical for each interface. In both tasks, the participants reported the search results to be much more specific using Concept Highlighter than Google. This result is statistically significant for both tasks.

It is interesting to note that even though many of the participants made significant errors in deciding the relevance of documents in Task 1, there was still an increase in all the subjective measures for Concept Highlighter over those reported for Google. Clearly, the participants had an impression of being more confident in the search results, that the system was easy to use, that they were satisfied with the results of the search, and that the search results were more specific. Although many of these results did not prove to be statistically significant, they do indicate a positive trend in favour of Concept Highlighter. Also, given that this evaluation compared expert Google users to novice Concept Highlighter users, the results are very promising. As the users become more experienced with the concepts available in Concept Highlighter, their subjective reactions, as well as their performance, will increase due to their ability to learn the meanings of the concepts that commonly appear for

Table 6.19: The majority of the participants ranked the use of Concept Highlighter below that of Google after completing the tasks. These results are not statistically significant

Rank	Number of Responses
Concept Highlighter ranked higher than Google	11 (39%)
Concept Highlighter ranked lower than Google	17 (61%)
Wilcoxon signed rank test	$Z = -1.13, p = 0.26$

their search tasks.

Preference Rank

After all the tasks were completed by the participants, a post-task questionnaire was administered which included a question asking the participants to rank their preference for a search results interface, assuming that the underlying search results sets are the same. These rank responses are reported in Table 6.19.

Although more than half of the participants reported a preference of Google over Concept Highlighter, this result did not prove to be statistically significant. A number of participants noted that they ranked Google as their preferred interface because of their prior familiarity with it, resulting in a skewing of these results in favour of Google. Others may have preferred the ease at which they could start considering documents for relevance using Google, as opposed to the two-step process supported by Concept Highlighter.

6.5.5 Discussion

These investigations have highlighted both the successes and challenges of Concept Highlighter. When a good set of concepts are provided by the concept knowledge base and are selected by the users, and when these concepts use a

similar vocabulary to that of the relevant search results, Concept Highlighter has a great ability to bring more relevant documents to the top of the search results. However, when the concepts selected by the users are not relevant, or when they use a different vocabulary from the relevant documents in the search results, the performance of the users as they seek relevant documents can be hampered. It is expected that the development of a more complete and detailed concept knowledge base will greatly improve the ability of concept-based fuzzy clustering to differentiate relevant and non-relevant documents based on user-controlled concept selection.

As expected, the participants in the user study took longer to find relevant documents using Concept Highlighter versus Google. While this extra time can be attributed to the added step of first choosing relevant concepts, it is promising that the extra time taken did not prove to be statistically significant. One may assume that users conducting searches on topics with which they have some familiarity will decrease the amount of time needed to choose relevant concepts. Similarly, as users become novice and expert users of Concept Highlighter, they will learn the concepts that tend to be relevant to many of their searches, resulting in a further decrease in the extra time taken.

While there was a marginal increase in the perceived precision using Concept Highlighter, the results were much more varied when considering the corrected precision. For Task 1, the participants made a significant number of errors (primarily false hits) resulting in a corrected precision that was worse than that of Google. For Task 2, there was a significant improvement in finding five relevant documents, but virtually no improvement in finding ten relevant documents. These results imply that in some cases, Concept Highlighter can be very effective, but in others, it can be a detriment (e.g., when the concepts available use a different vocabulary than the relevant documents, or when the

users make poor choices from among the concepts).

In general, the ability for users to effectively find relevant documents using this system depends greatly on their ability to first choose relevant concepts to their information need. If the task is not well understood, errors can be made in choosing the relevant concepts, resulting in an ineffective re-sorting of the search results. When good choices are made, in many cases, more relevant documents are provided among the top documents in the search results.

Although there was a reduction in user efficiency, and the results were rather varied in terms of user effectiveness, the subjective measures were positive. In all cases, the participants reported a higher degree of confidence, ease of use, and satisfaction, and that the search results were more specific when using Concept Highlighter. Although more than half of the participants reported that they preferred the use of Google, this does not come as a surprise since nearly all of them were already Google users. It is expected that these subjective responses will continue to improve as the users become more familiar with the features of Concept Highlighter.

While the results from these studies may not be conclusive, this system is illustrative of the advancements that can be made in terms of interactive search results exploration. Since the concept knowledge base used by Concept Highlighter is specific to the computer science domain, the usefulness for general Web searching is somewhat limited. However, as a research tool within specific domains, there is great value in the ability to use concepts to re-sort the search results into a more effective and relevant order. The development of a more general concept knowledge base would broaden the applicability of this tool to more general Web searching.

6.6 WordBars vs. Google

WordBars provides features that support the users in their tasks of interactive query refinement and interactive search results exploration. However, in order to improve their Web search performance, users must make appropriate selections of terms from the WordBars histogram. To test whether the users could make effective use of the WordBars features, a user study was conducted. The design and results of this study are provided in this Section.

6.6.1 Method

In order to evaluate and compare the effectiveness of the features of WordBars, a 4x3x2 (feature x task x participant group) within-subjects design was used. Unlike the previous two studies where HotMap and Concept Highlighter were compared to Google (see Sections 6.3 and 6.5), in this study, all comparisons are made among the features of WordBars.

The features tested in the evaluation are described in more detail in the following Section. In order to reduce the biasing effects, participants were provided the tasks in a pseudo-random order. Two separate participant groups were used: expert Web searchers consisting of graduate students and intermediate Web searchers consisting of undergraduate students. To ensure that the system provided the same set of search results to each participant, the results of each initial search task were cached.

6.6.2 Procedure

After completing a pre-task questionnaire, each participant was provided with a training session in which all the features of the WordBars system were explained. In addition, the procedure for the user study was described using this

training task as an example. This explanation and description of the research procedures took approximately 10 minutes.

When provided with each of the search tasks, the investigator answered any questions the participants had about the task itself. Each participant was asked to rate how well they understood the task. For each task, the participants were asked to perform the following steps:

1. Provide relevance scores for the top ten documents from the initial query in the order provided by the Google API using a four-point relevance scale (see Table 6.1).

2. Select one or more terms from the histogram in order to re-sort the search results (with the goal of moving relevant documents to the top of the list), and provide relevance scores for the top ten documents.

3. Refine the query by adding or removing terms from the query (with the goal of constructing a better query for the assigned task), and provide relevance scores for the top ten documents from the new query.

4. Select one or more terms from the histogram in order to re-sort the search results (with the goal of moving relevant documents to the top of the list), and provide relevance scores for the top ten documents.

These steps represent an example of an exploratory Web search scenario, in which a user would start with an initial query and attempt to find relevant documents in the top search results. The user would then attempt to explore certain aspects of the search results by selecting a few relevant terms from the histogram, and then evaluating the top documents in the list. Next, having learned something about the search results and the terms in the histogram, the user would attempt to refine their query and evaluate the top search results

from this new query. Having previously had success with using the histogram to re-sort the search results, the user may attempt to do so again using this new set of search results. While in real-world use, this cycle of refining the query and exploring the search results may continue until the user is satisfied that they have fulfilled their information needs, in this study the participants finish after only one cycle.

After each of these steps, an in-task questionnaire was administered regarding the participants feelings of confidence and satisfaction, and impressions of ambiguity among the search results considered. Once all three tasks were completed, a post-task questionnaire was administered. In addition, the participants were asked to provide a ranking of their preference among the features of WordBars to which they were exposed. The entire procedure took approximately 60 minutes for each participant.

Each search task included a written description of the information need, along with the initial query to be used. These were selected from the TREC 2005 HARD Track[1] test topics, and were intentionally chosen to be difficult tasks, yet understandable by a wide range of participants. The search tasks are listed below:

Task A Identify hydroelectric projects proposed or under construction by country and location. Detailed description of nature, extent, purpose, problems, and consequences is desirable.

   query: "new hydroelectric projects"

Task B Isolate instances of fraud or embezzlement in the international art trade.

---

[1] http://trec.nist.gov/data/t14_hard.html

query: "international art crime"

Task C Identify documents that discuss opposition to the introduction of the euro, the European currency.

query: "euro opposition"

As the participants evaluated the sets of search results, they were asked to speak the relevance scores using a four-point relevance scale (see Table 6.1). This information was logged by the investigator, along with the time required to evaluate the top ten documents. In addition, the time taken to choose terms for re-sorting, and the time taken to refine the queries were also logged.

In this study, the primary interest was in the Web search results interface and the ability for the participants to explore the search results and refine the queries. As such, the participants were asked to only consider the document surrogates for relevance; they were asked to not view the actual documents. The fact that a non-relevant document may appear to be relevant based on the title and snippet provided by the underlying search engine is beyond the scope of this research.

6.6.3 Analysis

Since the participants have different abilities to identify relevant documents, as well as different thresholds for viewing marginally relevant documents, performing a statistical analysis on the raw relevance data may not be meaningful. Instead, each participant's performance using the features of WordBars was compared to a baseline provided by their performance using the original order of the search results. In addition, each participant's performance in finding relevant documents was ranked, and pair-wise Wilcoxon signed rank tests were

used to identify the statistical significance of these results. The tasks were analysed both independently and in aggregate form.

The subjective evaluations, which were provided on a Likert scale, were analysed using nonparametric Friedman tests. The preference ranks were analysed using pair-wise Wilcoxon signed ranks tests. For all of these statistical tests, the expert participants were evaluated separately from the intermediate participants in order to identify the differences in their performance and subjective reactions. Where relevant, the statistical significance of these tests are noted and highlighted by a bold font.

This methodology for evaluating WordBars is a novel approach to evaluating information retrieval systems. Often, test collections are available that include test queries and expert relevance judgments for all documents in the collection. However, for Web search evaluations, it is difficult to obtain expert relevance judgments, even just for the top search results. Instead, the improvements individual participants are able to make throughout the exploratory search scenarios are evaluated.

6.6.4 Results

Participant Demographics

Twenty-four computer science students were recruited to participate in this study, and were classified into two groups. The expert group consisted of twelve graduate students; the intermediate group consisted of twelve undergraduate students (registered in first and second year computer science courses). All though there were some duplicates between the participants in this study and those in the HotMap and Concept Highlighter study, the two studies were conducted almost one year apart. The results from the pre-task questionnaire

administered to these two participant groups are presented in Table 6.20.

All the participants in this study indicated that they use a computer more than ten times per week. This should come as no surprise since all the participants were students in computer science courses or programs. As is to be expected, the expert group indicated a higher degree of computer experience, and conducted more searches per week than the intermediate group. All the participants showed a preference for the Google search engine.

In terms of pages of search results viewed, the expert group showed a tendency to view fewer pages in the search results than the intermediate users. Although it is difficult to determine the reasons for these differences, it may have to do with the ability and confidence the participants have in crafting queries. The members of the expert group may have had a high degree of confidence in constructing an effective query, resulting in fewer pages of documents being viewed before they either satisfied their information needs or refined their query. Among the intermediate group, a number of participants reported viewing many more than the three pages of search results that are common [89, 93]. This may be due to having a strong need to satisfy their information needs, but an inability to craft or refine a query; in these cases, the only alternative is to view many pages of search results.

While both groups showed similar tendencies for adding terms to their initial queries, the intermediate group expressed more likelihood of removing terms from their query than the expert group. This may be due to the expert group having more confidence in the terms selected for their initial queries than the intermediate group. It is interesting to note that both participant groups self-reported a similar degree of Web search experience.

Prior to starting each search task, the participants were asked to report their understanding of the assigned task. In asking this question, the concern

Table 6.20: Demographic features of the participant sample for the WordBars user study.

Computer Use	Expert Group	Intermediate Group
10+ times per week:	100%	100%

Computer Experience	Expert Participants	Intermediate Group
high degree:	100%	100%

Web Searches	Expert Group	Intermediate Group
10+ per week:	84%	67%
5-10 per week:	8%	25%
1-5 per week:	8%	8%

Search Engine Preference	Expert Group	Intermediate Group
Google:	100%	100%

Pages Viewed	Expert Group	Intermediate Group
1 page:	17%	8%
2 pages:	25%	8%
3 pages:	33%	17%
4 pages:	17%	17%
5 pages:	8%	25%
6 pages:	0%	0%
7 pages:	0%	17%
8 pages:	0%	8%

Likelihood of Adding Terms to a Query	Expert Group	Intermediate Group
always:	33%	42%
often:	55%	42%
sometimes:	17%	17%
seldom:	0%	0%

Likelihood of Removing Terms from a Query	Expert Group	Intermediate Group
always:	0%	25%
often:	33%	17%
sometimes:	50%	50%
seldom:	17%	8%

Web Search Experience	Expert Group	Intermediate Group
high degree:	58%	42%
moderate degree:	42%	50%
low degree:	0%	8%
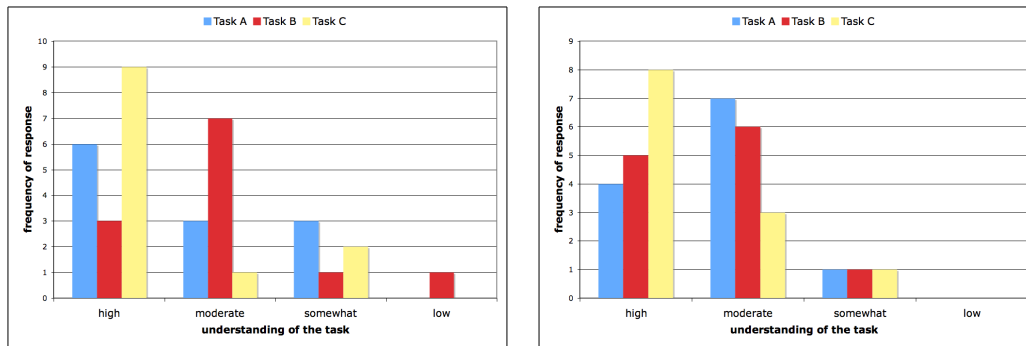
Figure 6.12: The pre-task understanding of the assigned topics, as reported by the expert group (left) and the intermediate group (right).

was not with their prior knowledge on the topic, but instead how well they understood the description of the information they were to be finding. These results are illustrated in Figure 6.12. For both participant groups, most participants reported a moderate to high degree of understanding of the topics, although there was some variability in these results between the tasks and between the groups.

Comparison to Baseline

Although it is common to use variants of the precision metric for evaluating information retrieval systems [52], these metrics generally assume the existence of expert relevance scores for the documents. In this study, obtaining such expert evaluations was difficult since for each of the three tasks, each participant was asked to craft a new query using the features of WordBars. In the worst case, with each of the 24 participants of this study generating a unique new query for each task, experts would have had to decide the relevance of the first 100 search results for 72 different queries. Clearly, this resulted in a barrier to using expert evaluations in this study.

Instead, the effectiveness of the features of WordBars were analyzed under

the assumption that the participants all made accurate relevance judgments. Although this assumption may not be valid for all participants, each participant believed they were making accurate relevance judgement choices to the best of their abilities. Therefore, evaluating the effectiveness of WordBars under this assumption provides a meaningful analysis of the performance of each participant using the features of the system.

In order to make a comparison for each participant's performance using the features of WordBars, their performance in evaluating the search results from the initial query in the original order provided by the Google API was used as the baseline. For all four sets of documents considered (original order, first re-sort, refined query, and second resort), the number of documents that were given scores of three or four (i.e., the documents that were indicated as "probably relevant" and "relevant" by the participants) were counted.

The difference between the number of relevant documents found using the baseline, and the number found for the other three sets of search results generated using the features of WordBars were counted. This allowed for the labeling of each participant as having done "better", "same" or "worse" using each feature of WordBars for each task assigned.

As is to be expected, the ability for the participants to use the features of WordBars to find more relevant documents depends greatly upon the specific search task. It is also expected that there will be differences in the abilities of the participants in the expert and intermediate groups. As such, the results of this comparison to the baseline performance data are reported separately for each task and each participant group in Figure 6.13. These results show that a large portion of the participants from both groups were able to effectively use the features of WordBars in each task to find more relevant documents than in the original order of the search results from the initial query.

(a) first re-sort (initial search results)

(b) refined query

(c) second re-sort (refined query search results)
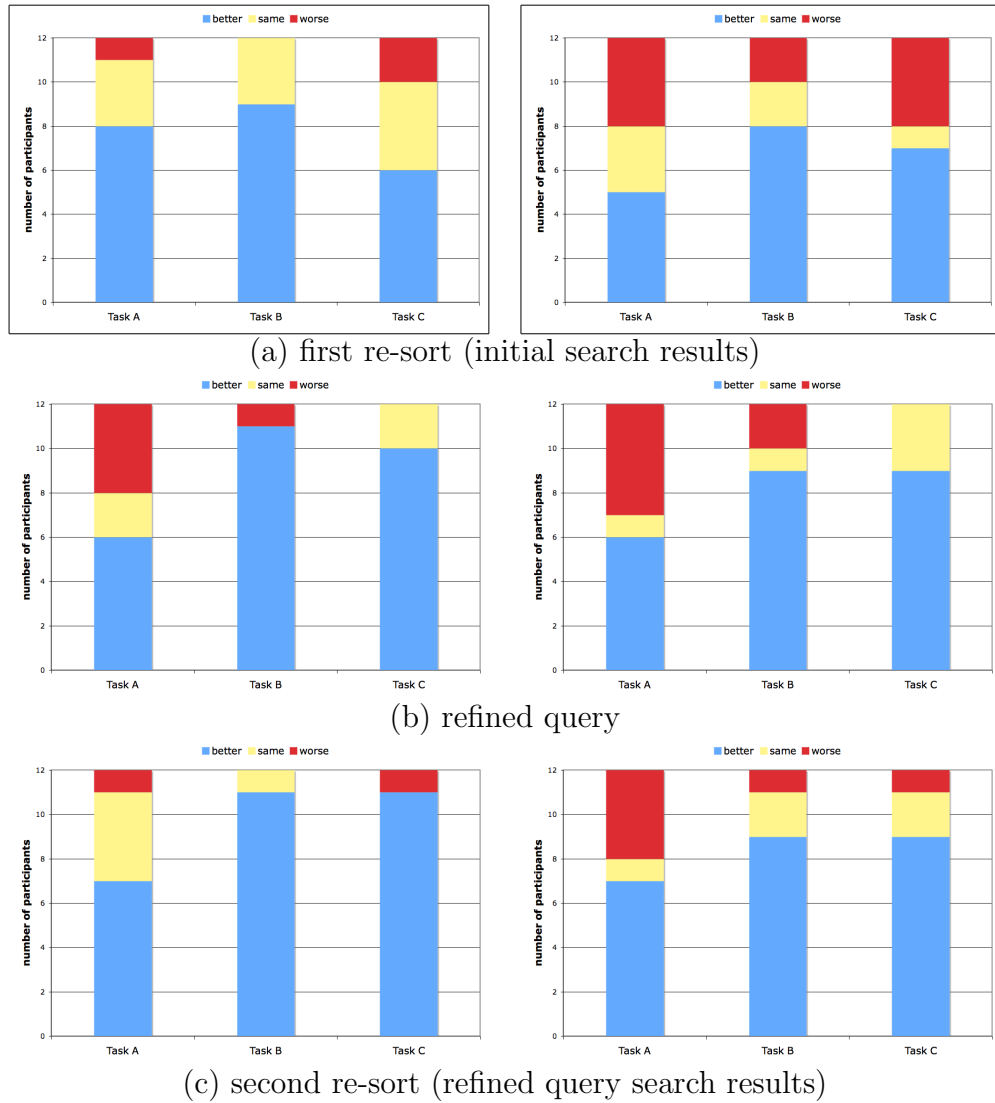
Figure 6.13: A comparison of the number of relevant documents found in the baseline (original order from the initial query) versus (a) the first re-sort, (b) the refined query, and (c) the second re-sort. The results for both the expert group (left) and the intermediate group (right) are shown.

For each of the participant groups, 108 different comparisons are reported in Figure 6.13 (i.e, 12 expert or intermediate participants, 3 WordBars features, 3 tasks). For the expert group, the participants performed worse than the original order from the initial query in only 10 instances (9%); in 19 instances (18%) the participants performed the same; in 79 instances (73%) the participants performed better. For the intermediate group, the participants performed worse than the original order from the initial query in 23 instances (21%); in 16 instances (15%) the participants performed the same; in 69 instances (64%) the participants performed better.

A visual comparison of the results between the expert and intermediate groups for each tasks reveals a similarity in the general trends of performing better, the same, or worse using the WordBars features. In particular, it appears that one to three more participants from the intermediate group achieved worse results than the baseline when compared to the expert group. This difference may be due to the different levels of experience and skill in Web searching. Clearly the expert participants were able to make more effective use of the features of WordBars.

Rank Comparison

To further analyse the performance data, a ranking of the relative performance of each participant in finding relevant documents from the original order of the search results (i.e., the baseline in the previous comparison), as well as from each of the three other sets of search results generated using the features of WordBars was produced. This provides a more detailed view of the performance of the features of WordBars, since it considers the relative performance between the various features, rather than comparing the results only to the baseline performance of the original order of the search results from

the initial query. The results of this rank comparison are provided in Figure 6.14, grouped by participant group and task (including an aggregate of all the tasks). The results of pair-wise Wilcoxon signed rank tests are provided in Table 6.21.

Task A For Task A (see Figure 6.14a and Table 6.21a), it is clear that the best results for the expert group were achieved by re-sorting the search results from the original query. Many of the expert participants had difficulties constructing a new query that accurately captured the assigned information need. However, when re-sorting the search results from the refined query, these participants were able to perform nearly as well as when they re-sorted the search results from the original order.

For the participants in the intermediate group, the best results for Task A were achieved by resorting the search results from the refined query. These participants appeared to have some difficulty in the first re-sort of the search results from the initial query, as well as in constructing an effective refined query. However, most were able to achieve good results from the second re-sorting (of the refined query).

For the expert group, the improvements over the original order of the search results proved to be statistically significant after re-sorting the results of the original query, as well as re-sorting the results from the refined query. There was also a significant improvement over the refined query after re-sorting. For the intermediate group, none of the results form Task A were statistically significant.

Task B For Task B (see Figure 6.14b and Table 6.21b), the expert group performed moderately better than the original search by re-sorting the search

(a) Task A

(b) Task B

(c) Task C

(d) All Tasks (aggregate)
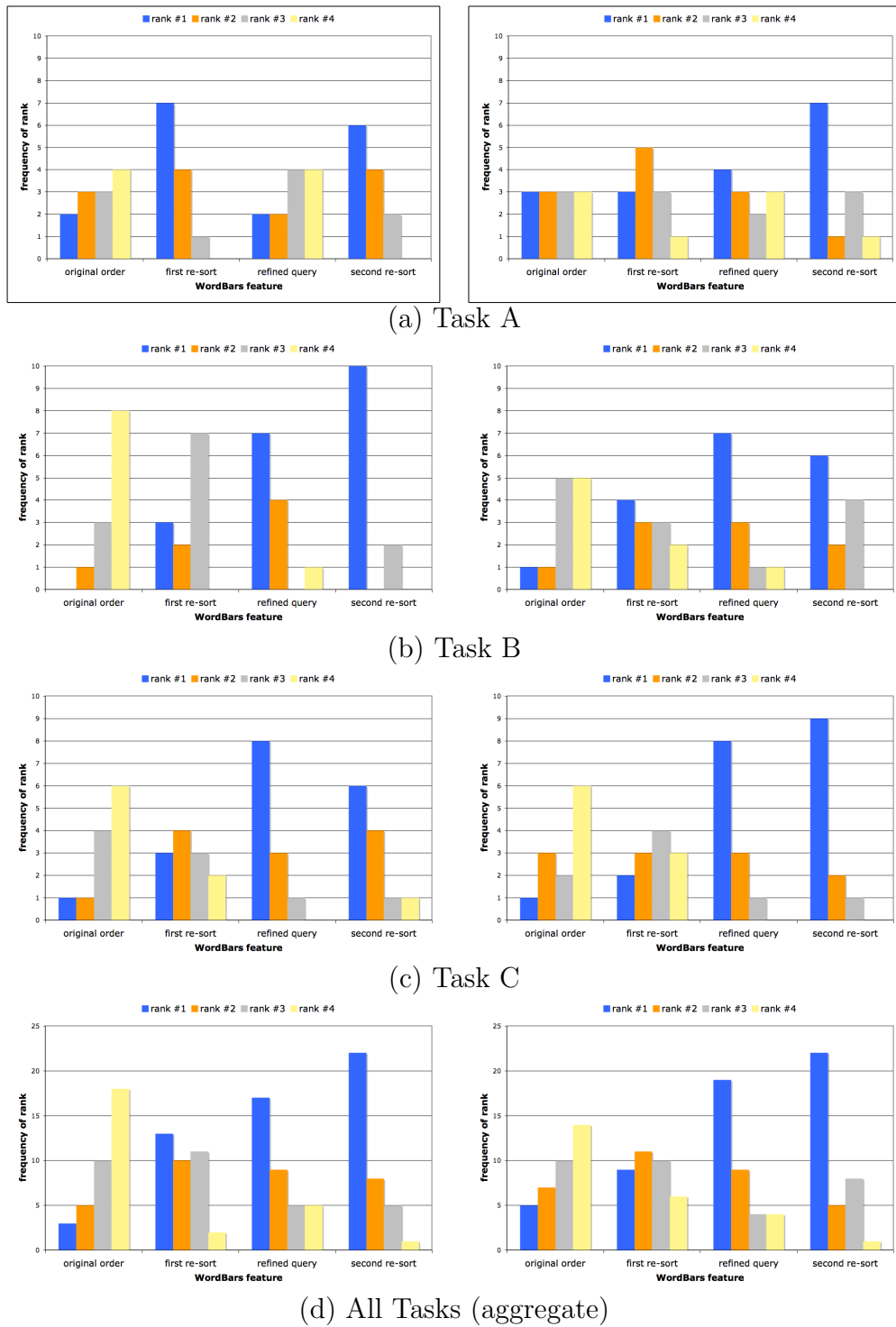
Figure 6.14: A rank comparison of the original order, the first re-sort, the refined query, and the second re-sort (of the refined query search results) for each of the three tasks as well as all the tasks aggregated together. Note that ties in the number of relevant documents result in ties in the ranking. The results for both the expert group (left) and the intermediate group (right) are shown.

Table 6.21: Pair-wise Wilcoxon signed rank tests indicate the statistical significance of the features of WordBars.

(a) Task A	Expert Group	Intermediate Group
first re-sort performed better than original order	$\mathbf{Z = -2.46, p = 0.01}$	$Z = -0.79, p = 0.43$
refined query performed better than original order	$Z = -0.58, p = 0.80$	$Z = -0.27, p = 0.79$
second re-sort performed better than original order	$\mathbf{Z = -2.27, p = 0.02}$	$Z = -1.32, p = 0.19$
second re-sort performed better than refined query	$\mathbf{Z = -2.12, p = 0.03}$	$Z = -1.03, p = 0.30$
(b) Task B	Expert Group	Intermediate Group
first re-sort performed better than original order	$\mathbf{Z = -2.76, p = 0.01}$	$Z = -1.66, p = 0.10$
refined query performed better than original order	$\mathbf{Z = -2.81, p = 0.01}$	$\mathbf{Z = -2.45, p = 0.01}$
second re-sort performed better than original order	$\mathbf{Z = -3.02, p < 0.01}$	$\mathbf{Z = -2.27, p = 0.02}$
second re-sort performed better than refined query	$Z = -0.65, p = 0.52$	$Z = -0.43, p = 0.67$ (negative ranks)
(c) Task C	Expert Group	Intermediate Group
first re-sort performed better than original order	$\mathbf{Z = -1.98, p = 0.05}$	$Z = -0.91, p = 0.36$
refined query performed better than original order	$\mathbf{Z = -2.84, p < 0.01}$	$\mathbf{Z = -2.70, p = 0.01}$
second re-sort performed better than original order	$\mathbf{Z = -2.26, p = 0.02}$	$\mathbf{Z = -2.69, p = 0.01}$
second re-sort performed better than refined query	$Z = -0.81, p = 0.42$ (negative ranks)	$Z = -0.26, p = 0.79$
(d) All Tasks (aggregate)	Expert Group	Intermediate Group
first re-sort performed better than original order	$\mathbf{Z = -4.03, p < 0.01}$	$\mathbf{Z = -1.99, p = 0.05}$
refined query performed better than original order	$\mathbf{Z = -3.20, p = 0.01}$	$\mathbf{Z = -3.18, p = 0.01}$
second re-sort performed better than original order	$\mathbf{Z = -4.34, p < 0.01}$	$\mathbf{Z = -3.79, p < 0.01}$
second re-sort performed better than refined query	$Z = -1.49, p = 0.14$	$Z = -0.54, p = 0.59$

results. In refining their queries, the participants performed much better than the original order in most cases. Re-sorting the search results from the refined query resulted in the best performance.

The participants from the intermediate group performed similarly to the expert group in that they showed a moderate improvement by re-sorting the original search results, and then a large improvement when refining the query. However, when re-sorting the refined query results, their performance decreased.

The improvements over the original order of the search results for the expert group proved to be statistically significant for all of the features of WordBars. Although the participants were able to improve upon the refined query by re-sorting these search results, this improvement was not significant. For the intermediate group, the improvement over the original order for the first re-sort was not statistically significant, although the results from the refined query and the second re-sort were. The decrease in the performance after re-sorting the refined query results did not prove to be statistically significant.

Task C   For the expert group on Task C (see Figure 6.14c and Table 6.21c), re-sorting the search results resulted in moderately better performance than the original order. Refining the query for this task produced the best results. Re-sorting the search results of this refined query resulted in a decrease in the performance compared to the refined query, but still a substantial increase over the original order.

The participants from the intermediate group showed a continued improvement over the original search results through the steps of re-sorting the original search results, refining the query, and then re-sorting the results of the refined query. The refined query and the re-sorted results from the refined query were

substantially better than the original order and the re-sorted results from the original order.

For the expert group, the improvements over the original order for all the features of WordBars proved to be statistically significant. The decrease in the performance after re-sorting the refined query results did not prove to be statistically significant. The moderate improvement by the intermediate group in re-sorting the original search results did not prove to be statistically significant. However, the refined query and the second re-sort were significant. The improvement of the second re-sort over the refined query was not significant.

Negative Performance In Task B for the intermediate group, and Task C for the expert group, there was a reduction in the performance of the participants when re-sorting the results of the refined query. This negative performance did not prove to be statistically significant in either case. It can be attributed to a number of participants finding all 10 documents to be relevant in the refined query search results, leaving no room for improvement from the re-sorting features. In other cases, poor choices of terms from the histogram resulted in a re-sorting of these search results that was not effective.

All Tasks (Aggregate) When considering the data from all tasks as an aggregate (see Figure 6.14d and Table 6.21d), the improvement in finding relevant documents using the features of WordBars is clear. The participants showed an increase in their ability to find relevant documents after re-sorting the original search results, a further increase after refining their query, and a continued increase after re-sorting the results from their refined query. These improvements over the original search results were shown to be statistically significant. The results from this analysis are very positive, and illustrate the

potential effectiveness of interactively exploring the search results and interactively refining a query through the features of the WordBars term frequency histogram.

However, when considering each task and participant group separately, it is not clear which feature of WordBars is the most effective. In some cases, the participants performed best when re-sorting the results of the original query. In other cases, the participants were able to effectively refine their queries, resulting in a significant improvement. Some participants were able to further improve their performance through re-sorting the search results from their refined queries. Although the aggregate analysis shows that with each step an improvement can be made, one of the great benefits of the exploratory features of WordBars is that each user can choose the method that works best for them for a given search task. One user may only take advantage of the search results exploration features of WordBars, whereas another may find the query refinement features most beneficial. As these users become more familiar with the features of WordBars and the support it provides for their exploratory Web search tasks, they will be able to more readily take advantage of the support WordBars provides as they seek to fulfill their information needs.

Time to Make Selections

In order to comment on the time performance of WordBars, the amount of time it took the participants to evaluate the first ten documents in the search results lists, as well as the amount of time it took them to make selections from the term frequency histogram in order to re-sort the search results and refine the queries was measured. Although the times each participant took to evaluate ten documents was consistent within a given task, the amount of time taken to re-sort the search results and refine the queries was highly variable.
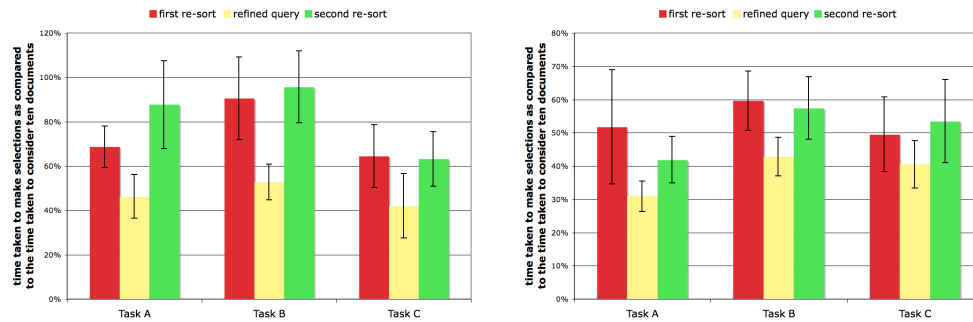
Figure 6.15: The amount of time required to make the re-sorting and query refinement selections, as compared to each participant's average time taken to consider ten documents. The error bars represent the standard error about the mean, and illustrate the high degree of variability in this data. The expert group results are shown on the left; the intermediate group results are shown on the right.

Rather than reporting absolute time values, the ratio of the time taken to make the selections versus the average time taken to evaluate ten documents is reported. This provides a more meaningful representation of this data, since it reports the amount of time the participants took to make selections from the histogram in relation to the speed at which they evaluated the search results.

This data is reported in Figure 6.15 for all three tasks and both participant groups. Clearly, both groups of participants spent more time considering the terms for re-sorting than selecting terms to add or remove from their query. Although the relative times are different for each task and participant group, the patterns are similar. That the query refinement tasks were performed quickly should not be a surprise, since each participant would have already considered the term frequency histogram while performing the first re-sort of the original search results.

The error bars in Figure 6.15 illustrate the high variability in this data. This variability can be attributed to the different techniques participants used while considering the terms in the histograms. Some participants quickly

evaluated each term, and selected all that they considered to be relevant. Others carefully considered each term, and experimented with the results of selecting and un-selecting terms. Because of the variability of this data, it is difficult to make any strong statements regarding the amount of time users might take to make useful selections using WordBars. However, one can expect that as users become more familiar with the WordBars interface, they will spend less time experimenting with the interface itself, and more time focusing on fulfilling their information needs.

Subjective Measures

After each task, participants completed a short in-task questionnaire to measure their subjective reactions to using the features of WordBars to find relevant documents for the assigned task. Of interest were the participants' feelings of confidence in completing the task, satisfaction with the search results considered, and perceptions of ambiguity among the search results set. In general, these subjective measures were very positively in favour of the features of WordBars, and showed little differences between the expert and intermediate groups.

For the *confidence* measure, the participants rated how confident they were in their ability to find a good set of relevant documents (Figure 6.16a). Whereas the confidence in the search results from the original order followed a normal distribution, the confidence from re-sorting of the original search results, refining the query, and re-sorting the refined query search results were all positively skewed. In particular, the highest degree of confidence from the expert group was reported for re-sorting the refined query results, followed closely by the refined query. The highest degree of confidence from the intermediate group was reported for refining the query, followed closely by re-sorting the

(a) confidence in ability to find relevant documents

(b) satisfaction in the search results

(c) impressions of ambiguity in the search results
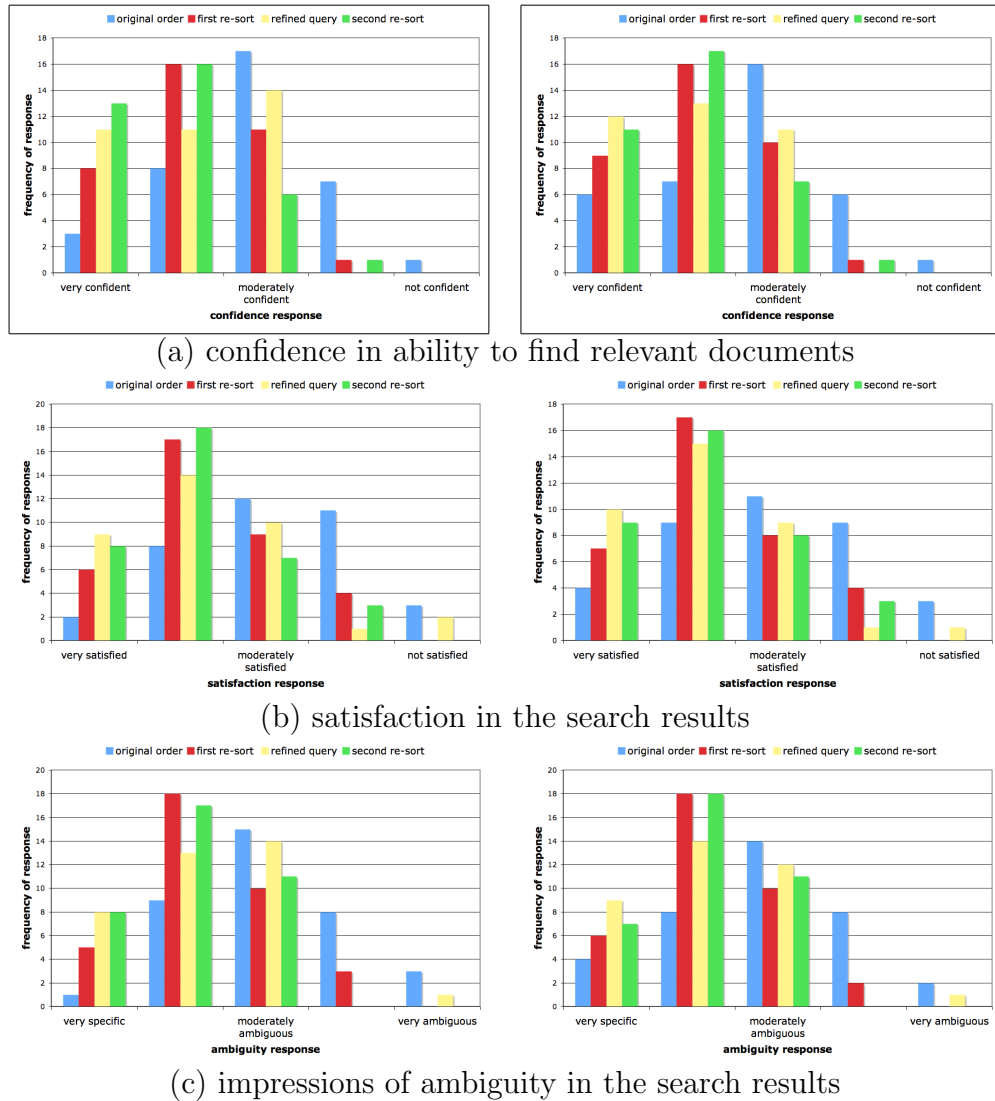
Figure 6.16: Subjective measures reported by the participants after completing each stage of the study. Note that the features of WordBars were consistently scored higher than the original order of the search results from the initial query, both for the expert group (left) and the intermediate group (right).

Table 6.22: Friedman tests for the subjective reactions show that the differences in this data are statistically significant.

Measure	Expert Group	Intermediate Group
(a) confidence	$\chi^2(3) = 36.12, p < 0.001$	$\chi^2(3) = 22.92, p < 0.001$
(b) satisfaction	$\chi^2(3) = 29.25, p < 0.001$	$\chi^2(3) = 21.80, p < 0.001$
(c) ambiguity	$\chi^2(3) = 36.17, p < 0.001$	$\chi^2(3) = 24.03, p < 0.001$

refined query results.

For the *satisfaction* measure, the participants rated how satisfied they were with the documents considered in the search results set (Figure 6.16b). Like the confidence measure, the satisfaction measure for the original order followed a normal distribution. The satisfaction measure for the three steps of using WordBars for conducting the search were all positively skewed. For both the intermediate and expert groups, the highest confidence was reported for the refined query, followed closely by the re-sorted results from the refined query.

For the *ambiguity* measure, the participants rated how ambiguous they thought the search result set was (Figure 6.16c). Again, the responses for the original order of the search results followed a normal distribution, and the responses for the features of WordBars showed positive skew. For the expert group, the responses were slightly more positive for the re-sorted search results from the refined query. For the intermediate group, the responses were in favour of the refined query.

The results of Friedman tests on these responses showed them to be statistically significant. These statistics are reported in Table 6.22.

Perference Rank

After all the tasks were completed by the participants, a post-task questionnaire was administered which included a question asking the participants to

rank their preference for the search results considered. The options were (a) search results in the original order, (b) the search results after re-sorting, (c) the search results after refining the query, and (d) the search results after re-sorting the refined query. These rank responses are reported in Figure 6.17.

The original order of the search results was almost unanimously ranked last by participants in both groups (although one participant from the intermediate group ranked it first). This result provides a clear indication that almost all the participants found value in the interactive query refinement and interactive search results exploration features of WordBars.

Among the participants in the expert group, the rank depended greatly upon their ability to refine the query. Those who were able to effectively choose relevant terms to add to their query (as well as replace or remove ambiguous terms) tended to select (c) the refined query as their top preference, followed by (d) the second re-sorting of the refined query as their second choice, and (b) the first re-sort of the original search results as their third choice. The participants who had difficulty choosing terms with which to refine their query tended to indicate (b) the first re-sorting of the original search results to be most preferable. They also had a tendency to indicate that (d) the second re-sorting of the refined query was preferable to (c) the refined query itself.

It is interesting to note that for the expert group, their preference ranks did not follow their performance when considering all tasks in aggregate (see Table 6.14d). This may be due to the difficulties some had with refining their queries, as well as their generally high level of confidence in initial queries due to their expert Web search abilities. Further, many of these participants reported viewing few pages of search results, which indicates a low tolerance for spending time searching. For some, this low tolerance may have resulted in lower rankings for the tasks that would have taken more time (i.e., refining
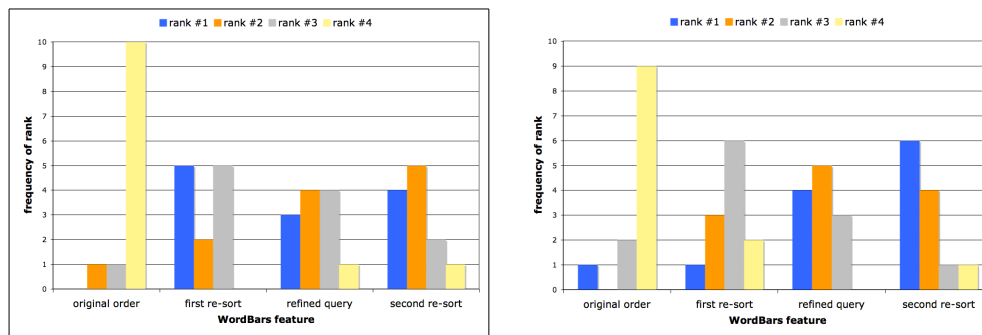
Figure 6.17: Preferences ranks reported by the participants for the features of WordBars. Note that the re-sorting and query refinement features of WordBars were almost unanimously ranked higher than the original order of the search results, both for the expert group (left) and the intermediate group (right).

and further exploring the search results).

For the intermediate group, there was a clear preference for the ability to (c) refine their queries as well as (d) subsequently re-sort the results from the refined query. This followed closely their performance when considering all tasks in aggregate (see Table 6.14d). The low rank responses for re-sorting the results of the original query may be due to the lower experience and Web search skill levels of the participants in this group. This may have resulted low confidence in the initial search results, and therefore, in their ability to re-sort and explore these search results.

A pair-wise analysis of the rank responses using Wilcoxon signed ranks tests are reported in Table 6.23. For the expert group, all the features of WordBars, including (b) the first re-sort, (c) the refined query, and (d) the second re-sort, were shown to be preferable to (a) the original order of the search results, with statistical significance. However, there was no clear preference between the features of WordBars themselves. For the intermediate group, (c) the refined query and (d) the re-sort of the refined query results were shown to be preferable to (a) the original order of the search results, with statistical

Table 6.23: Pair-wise Wilcoxon signed ranks tests for all possible comparisons of the rank data.

Comparison	Expert Group	Intermediate Group
first re-sort preferable to original order	$\mathbf{Z = -3.11, p < 0.01}$	$Z = -1.73, p = 0.08$
refined query preferable to original order	$\mathbf{Z = -2.50, p = 0.01}$	$\mathbf{Z = -2.86, p < 0.01}$
second re-sort preferable to original order	$\mathbf{Z = -2.78, p = 0.01}$	$\mathbf{Z = -2.36, p = 0.02}$
refined query preferable to first re-sort	$Z = -0.56, p = 0.58$ (negative ranks)	$Z = -1.73, p = 0.08$
second re-sort preferable to refined query	$Z = -0.58, p = 0.56$	$Z = -0.37, p = 0.71$
second re-sort preferable to first re-sort	$Z = -0.12, p = 0.90$	$\mathbf{Z = -2.17, p = 0.03}$

significance. Further, (d) the second re-sort was shown to be significantly preferable to (b) the first re-sort.

6.6.5 Discussion

There are four features of the user evaluations that are worth discussing in further detail: the tasks, the participants, the relevance judgments, and the search procedures. The three search tasks used in this study were intentionally chosen to be easy to understand, yet somewhat ambiguous. By choosing easy to understand tasks, prior knowledge or experience in the task domain was not necessary. By choosing tasks that were also somewhat ambiguous, the top search results returned by the Google API included a mixture of relevant and non-relevant documents. With very specific tasks, Google and other Web search engines perform very well, providing many highly relevant documents in the top search results, leaving little room for improvement. However, for ambiguous tasks, there is a great opportunity to improve the performance of

the users through interactive query refinement and interactive search results exploration, as this study has shown.

Conducting the user evaluations with two distinct participant groups, not only allowed the potential effectiveness of WordBars to be evaluated for these groups, but also provided opportunities for comparisons between these groups to be made in terms of the participants' abilities to use this exploratory Web search system to find relevant documents. While the aggregate performance over all three tasks was very similar for both groups, the expert group performed much better than the intermediate group when re-sorting the search results from the initial query. The time taken to make selections from the term frequency histogram were very similar for both groups, as were the subjective reactions. As is to be expected from the aggregate performance results, the intermediate participants ranked the refinement and subsequent re-sorting features as most preferable. By contrast, the expert participants also indicated the first re-sort as a preferable feature, even though their performance results for this feature were not as good as re-sorting the search results from the refined query.

Since expert evaluations of the documents were not available, the performance of the participants in terms of finding relevant documents was based solely on the relevance scores provided by the participants. This assumption that the participants were able to give accurate relevance judgments ignores the situations where the title, snippet, and URL for a document are misleading, or where the participant incorrectly interprets the information provided. These errors can be ignored since it was possible for the participants to make them during all stages of the study (including the baseline evaluation of the search results in the original order). Therefore, from the perspective of the user, the assumption of accurate relevance judgments does not invalidate the

results reported in this study. Evaluating whether users are able to accurately make relevance judgments given the limited information provided by Web search engines is worth further study, not just with WordBars, but with Web search results in general.

Even though the results of this study were very positive, and indicate that in most cases the features of WordBars can help the users to find more relevant documents, real-world use may result in even better results. This study required the participants to use WordBars in a structured and measured manner, by first evaluating the top ten documents, then re-sorting the search results and evaluating the top ten documents, followed by refining the query and evaluating the top ten documents, and then re-sorting the search results again and evaluating the top ten documents. The participants had only one chance to select terms for re-sorting, as well as for refining their query. In real-world use, users can easily experiment with "what-if" scenarios, selecting terms for re-sorting, considering a few documents, making further selections or changes, considering the results of these changes, etc. Similarly, for refining the query, if the users feel that they have made a mistake in modifying the query, they could easily return to the previous query, or attempt to make a further refinement of the query.

Chapter 7

Conclusions and Future Work

7.1 Summary of Research

The goal of this research has been to explore methods by which visualization and interaction can be used to support Web searchers in their fundamental information seeking tasks, and to further understand the potential benefits these visual and interactive systems can provide to users. A new model for interactive Web search was proposed, and a design-oriented research methodology was employed, in which prototype systems that address the fundamental problems with Web search were developed and studied. As a result, VisiQ and WordBars were developed to support interactive query refinement processes; and HotMap, Concept Highlighter, and WordBars were developed to support interactive search results exploration.

Two different user studies were conducted to evaluate the effectiveness of the interactive search results exploration systems. In the first study (n=28), the visual and interactive features of HotMap and Concept Highlighter were evaluated against the list-based representation of Web search results provided by Google in terms of efficiency, effectiveness, and subjective reactions. In this

study, HotMap proved to be an effective method for supporting the users in finding more relevant documents from among the top search results. Although Concept Highlighter showed promise in terms of allowing the users to interactively cluster and re-sort the search results, it required an extra evaluation step and introduced some errors in the relevance judgments.

In the second study (n=24), two classes of participants (expert and intermediate) used WordBars to find relevant documents for assigned Web search tasks. The original order of the search results was provided by Google, allowing a baseline comparison of the list-based representation of the Web search results to the visual and interactive features of WordBars. The results from this study were very positive, illustrating the benefits of allowing users to interactively manipulate and explore the Web search results.

In this Chapter, the primary contributions of this research are highlighted and an outline of future research work is provided.

7.2 Contributions of Research

Interactive Model for Web Information Retrieval. One of the fundamental contributions of this research has been the development of a new paradigm for interactive Web information retrieval. In the traditional model of Web search, little support is provided to users in their tasks of query formulation and search results evaluation. In this new paradigm, cycles of interactive query refinement and interactive search results exploration support the users in crafting a query and finding relevant documents from among the search results set.

This interactive model has provided a framework for exploring interactive visualization support for Web search tasks in this research. It also provides a

number of design principles that may help to guide future research on next-generation Web search systems.

Design Space. A two-dimensional design space was proposed within the domain of Web search. In one dimension are the primary tasks the users must perform, drawn from the interactive model for Web information retrieval: interactive query refinement and interactive search results exploration. In the second dimension are the types of subject knowledge that are available to support the users in their tasks: concept information and information extracted from the top search results. This design space has guided the development of the prototype systems in this research, and provides a framework for future research in next-generation Web search systems.

Concept Knowledge Base. One of the fundamental difficulties in providing a visual representation of queries in support of query refinement is the lack of information upon which to base the visualization. Queries are primarily textual and seldom provide information regarding the relationships between the query terms. The queries themselves provide little information in terms of how better queries can be constructed.

The concept knowledge base was devised to address these problems. It draws inspiration from conceptual semantic theories and formal concept analysis. Methods were developed to automatically generate a concept knowledge base from a well-formated classification of concepts (such as the ACM Computing Classification System). This concept knowledge base is used in VisiQ to generate the query space, which is then presented in a graphical representation to support interactive query refinement. In addition, the concept knowledge base is used in Concept Highlighter to form the basis for the fuzzy cluster

centroids in this interactive search results exploration system.

Prototype Development.   In the course of this research, four different prototype systems were developed to explore the design space. The modular architecture of the prototype systems supports the inclusion of other sources of information in the future, such as new knowledge bases, other search engines, etc. The architecture and implementation technologies developed in the course of creating these prototypes provides an example for future researchers in this area. Further, these prototype systems provide a basis for future development of publicly available Web search tools.

Benefits of VisiQ.   The primary benefits of VisiQ are that it provides a visual representation of the query space, allows the users to interactively refine their queries resulting in the instantaneous generation of new query spaces, and provides a preview of the Web search results. These interactive visualization features provide significant improvements over the list-based representation of potential query terms used in traditional interactive query refinement techniques. Extensive examples have been provided to illustrate the support VisiQ provides to Web searchers when they provide vague queries, specific queries, incorrect terms, and in exploratory search scenarios.

Benefits of HotMap.   HotMap provides a visual representation of the frequencies of the query terms within the top 100 search results, displaying this information at two levels of detail. The overview map allows the users to visually inspect the entire set of search results to identify potentially relevant documents. The users may also interactively re-sort the search results through a nested sorting feature.

The user study reported in this dissertation illustrated the benefits of

HotMap over Google in terms of considering fewer non-relevant documents in order to find a set of relevant documents, reduced error rates, and positive subjective measures. Although not all these results were statistically significant, the potential benefits of HotMap for interactive search results exploration are clear.

Benefits of Concept Highlighter. Concept Highlighter provides a visual and interactive interface to a single-pass fuzzy clustering of Web search results. The users can interactively select the concepts that are relevant to their information need, resulting in a re-sorting of the search results and a visual indication of the fuzzy membership score for each document in the search results. The consistent naming of the clusters based on the concept knowledge base supports the users in learning the meanings of the concepts relevant to their common information seeking goals.

The results of the Concept Highlighter user study were inconclusive, however. In some cases, there was an improvement over Google, whereas in others, Google performed better. Although the subjective reactions when using Concept Highlighter were marginally better than Google, the participants ranked Google as a more preferable system to use. While these results were not as positive or significant as those from HotMap, they do illustrate the potential benefits of the interactivity and visualization provided by Concept Highlighter in some situations.

Benefits of WordBars. The primary benefits of WordBars is that it provides a graphical representation of the frequencies of term use within the top search results, supporting both interactive query refinement and interactive search results exploration. The term frequency histogram provides an overview

of the search results set, as well as a means for selecting terms for re-sorting the search results, and adding or removing terms from the query. A key benefit is the ability to seamlessly transition between the query refinement and search results exploration tasks within the same visual interface. This reduces the cognitive overhead associated with consciously deciding to switch tasks.

The user study illustrated the benefits of the WordBars system, both for improving the queries as well as improving the performance of the participants in finding relevant documents. The participants reported a high degree of confidence and satisfaction while using the features of WordBars, and a reduction in the ambiguity of the search results sets. The features of WordBars were almost unanimously ranked higher than the original order of the search results set. These strongly positive results illustrate the benefits that WordBars provides to Web searchers.

Evaluation Methodologies An added contribution of this research has been in the evaluation methodologies used to study the effectiveness of the prototypes systems. In traditional information retrieval research, collections of locally stored documents and expert relevance scores for corresponding test queries are available for conducting evaluations of proposed techniques. However, when conducting research using live Web search results, collecting the source documents for conducting the evaluations is often not practical. Furthermore, expert relevance scores are not available for such Web documents.

In this research, a number of new evaluation methodologies were employed. In the evaluation of HotMap and Concept Highlighter, expert evaluations were manually determined for pre-cached Web search results of test queries. Two new metrics were devised based on the user and expert relevance scores: perceived precision and corrected precision. A categorization of error rates was

developed (misses and false hits) along with a severity measure (major and minor). In the evaluation of WordBars, a novel method for comparison to a baseline performance was employed. In addition, the user performance was ranked in order to determine which features were most effective. These contributions to Web search evaluation methodologies may help to guide the evaluations of future Web search systems.

7.3   Future Work

Further Exploration of the Design Space.   Opportunities exist to extend the two-dimensional design space for interactive Web search in a number of different ways. Along the *task* dimension, other tasks that support users finding information on the Web may be explored, including managing information needs over time, and storing potentially relevant documents for further evaluation either later in the current session, or in a future session. Along the *information* dimension, other sources of information may be of value in supporting the users' Web search tasks, including personalized information preferences, browsing histories, and other domain-specific knowledge bases.

A third dimension within this design space may also be explored: the *type of Web search procedure* used by the searcher. Along this dimension, users' Web search procedures may include *information seeking*, *exploratory search*, and *question answering*. This extension of the design space supports research into discovering why some systems are effective in certain situations, and less effective in others.

Further Research on Subject Knowledge.   The concept knowledge base was developed in this research to store conceptual knowledge extracted from

well-formated categorizations of domain-specific knowledge. The ACM Computing Classification System was used as the source of the information. There is value in moving towards generating similar knowledge bases for other domains, perhaps based on patent or medical databases, for example. This work may also be extended to support more general purpose knowledge bases using concept information derived from Wikipedia, the Open Directory Project, or other sources.

This concept knowledge base represents an example of subject knowledge that can be used in interactive Web search systems. Other intelligent computing techniques may also be explored in order to provide useful knowledge to the users as they conduct their interactive query refinement and interactive search results exploration tasks.

Personalization   Since subject knowledge provides the basis for the support of interactive query refinement and interactive search results exploration in all four prototype systems developed, there is an opportunity to explore the personalization of this subject knowledge. This personalization may take the form of providing custom subject knowledge based on other users searching for similar topics, or only based on the prior searching behaviour of individual users. This topic warrants further exploration, and may result in significant improvements in the capabilities of the prototype systems described in this research.

Further Evaluation of Interactive Query Refinement.   Due to the complexities involved in designing controlled experiments to evaluate the interactive query refinement tools, these systems were presented without a thorough

verification of their effectiveness. The studies with WordBars, however, illustrated a potential avenue for evaluating interactive query refinement tools indirectly through their abilities to provide relevant documents to the users. Further evaluations of VisiQ and WordBars along these lines is needed. Also, comparing the abilities of these interactive tools to some of the well-known automatic query expansion systems will illustrate the benefits of taking into account the user's intelligence and understanding of their search task. In addition, there is a need for longitudinal, real-world evaluations of these systems to validate the finding of the controlled experiments.

Further Evaluation of Interactive Search Results Exploration. Although extensive evaluations of HotMap, Concept Highlighter, and WordBars were conducted in this research, there is a need for additional evaluations of these systems. Although the results for HotMap were rather positive, there remain some features that warrant further exploration, such as the specific utility of visual inspections of the overview map, or the usability of the nested sorting features. The result of the user study with Concept Highlighter were rather mixed; further study on the use of the concept knowledge base for clustering and categorization of Web search results is needed. While the user evaluations of WordBars were very positive, it may be of value to explore other methods for generating the subject knowledge, including using TF-IDF measures and other information retrieval-inspired measures of term importance. For all three systems, there is a need for evaluations in real-world setting to validate the findings of the controlled experiments.

Combining Features into a Unified Web Search System. With the exception of WordBars, each of the prototype systems developed in this research

addressed a specific aspect of Web search: either query refinement or search results exploration. Combining the positive features of multiple prototype systems will result in the development of unified interactive and visual Web search systems that support the entire Web search process. Steps have already been taken along these lines, through the development of HotMap+WordBars. Similarly, the features of VisiQ and Concept Highlighter complement each other, and may be combined together.

VisiQ 2.0. Opportunities for further development of VisiQ include expanding the system to support a wider range of domains, improving the graph-drawing methods for representing the query space, supporting the refinement of more complex queries, and providing a tighter integration with the Web search results. In addition, VisiQ may be extended to support a personalized concept knowledge base using machine learning methods.

HotMap 2.0. The features of HotMap may be extended and improved in a number of different ways. It may be desirable to allow the users to choose between a compact representation of the Web search results that shows just the title, and an expanded representation that shows the title, snippet, and URL. Support for query phrases may also be beneficial, as would improved methods for specifying and indicating nested sorting features.

Concept Highlighter 2.0. Given the inconclusive results from the Concept Highlighter user study, there are many opportunities to improve upon this work. These include exploring different clustering and classification algorithms, providing better methods for selecting relevant concepts, and providing more informative visual representation of the fuzzy membership scores (or other scores provided by the clustering or classification algorithms used). One

of the findings of the evaluation of the fuzzy clustering techniques was that there was an increase in relevant documents among the top 10 documents, but little increase when considering the top 20 or 30. This may be exploited by providing access only to the top 10 documents for a given concept selection. Other improvements include using a colour scale that is normalized with respect to the range of values being represented, rather than representing static values as in the current implementation. As with VisiQ, there is an opportunity to employ a personalized concept knowledge base using machine learning techniques.

WordBars 2.0. Further research on the use of WordBars for both interactive query refinement and interactive search results exploration is already underway. In terms of supporting query refinement, features to support the construction of complex queries will be of value, including the ability to indicate an exclusion term, specify phrases, and to manipulate the order the terms in the query. There are also benefits in providing a visual distinction in the term frequency histogram between original query terms, new terms that have been added to the query, and other terms that are available to be added.

In terms of supporting search results exploration, users may benefit greatly from the ability to visually specify a utility function using the term frequency histogram. By indicating weights (both positive and negative) to use when re-sorting the search results set, the users will be able to exert more control over the re-sorting process, at the expense of increasing the complexity in using the system. These weights may also be used to provide a colour coding of the corresponding terms within the search results set, resulting in an easier identification of how the candidate terms are being used within the document surrogates in the search results set.

Other future work include exploring other methods from information retrieval for deducing the potential importance of terms found within the top search results. In the exploration of these techniques, it is vitally important to continue to keep the users in control of the Web search process.

Commercialization Venture. One thing that has become clear throughout the development of the prototype systems is the potential for commercial value in this research. As such, with the support of the University-Industry Liaison Office, four patent applications have been filed to protect the intellectual property associated with this research (see Appendix A). Plans are underway to develop commercial-grade Web search and domain-specific search tools based on this research.

Appendix A

Patent Applications

With the support of the University-Industry Liaison Office, four patent applications were prepared and filed based on this research. The applications are listed below:

1. O. Hoeber, X. D. Yang, Y. Yao, Method and Apparatus for Construction and Use of Concept Knowledge Base. Application filed in June 2006 in both Canada and the United States.

2. O. Hoeber, X. D. Yang, Method and Apparatus of Visual Representations of Search Results. Application filed in August 2006 in both Canada and the United States.

3. O. Hoeber, X. D. Yang, Y. Yao, Method and Apparatus for Concept-Based Visual Presentation of Search Results. Application filed in September 2006 in both Canada and the United States.

4. O. Hoeber, X. D. Yang, Interactive Web Information Retrieval Using Graphical Word Indicators. Application filed in December 2006 in both Canada and the United States.

Appendix B

User Study 1

In the first user study conducted in this research, participants used a number of different Web search systems with the goal of finding ten relevant documents. The systems used were HotMap, Concept Highlighter, Google, and a fourth system which was not very effective, and is not discussed in this dissertation.

This appendix includes the formal approval received from the Research Ethics Board for this study, as well as supplemental approval memo for changes that were made as the results of conducting a pilot study. All the evaluation documents are provided as well.

UNIVERSITY OF
REGINA

OFFICE OF RESEARCH SERVICES

M E M O R A N D U M

DATE: December 14, 2005

TO: O. Hoeber
         Computer Science

FROM: K. Arbuthnott
         Chair, Research Ethics Board

Re: **Investigating the Effectiveness of Visualization for Search Results Evaluation
         (34S0506)**

Please be advised that the University of Regina Research Ethics Board has reviewed your
proposal and found it to be:

☑ 1.   ACCEPTABLE AS SUBMITTED. Only applicants with this designation have
           ethical approval to proceed with their research as described in their applications.
           The *Tri-Council Policy Statement on Ethical Conduct for Research Involving
           Humans* requires the researcher to send the Chair of the REB annual reports and
           notice of project conclusion for research lasting more than one year (Section 1F).
           **ETHICAL CLEARANCE MUST BE RENEWED BY SUBMITTING A BRIEF
           STATUS REPORT EVERY TWELVE MONTHS.** Clearance will be revoked
           unless a satisfactory status report is received.

☐ 2.   ACCEPTABLE SUBJECT TO CHANGES AND PRECAUTIONS (SEE
           ATTACHED). Changes must be submitted to the REB and subsequently
           approved prior to beginning research. Please address the concerns raised by
           the reviewer(s) by means of a supplementary memo to the Chair of the REB. Do
           not submit a new application. Please provide the supplementary
           memorandum**, or contact the REB concerning the progress of the project,
           before **February 14, 2006** in order to keep your file active. Once changes are
           deemed acceptable, approval will be granted.

☐ 3.   UNACCEPTABLE AS SUBMITTED. Please contact the Chair of the REB for
           advice on how the project proposal might be revised.

Dr. Katherine Arbuthnott

c.   X.D. Yang, Computer Science, supervisor

KA/rr/ethics2.dot

** supplementary memorandum should be forwarded to the Chair of the Research Ethics Board at the
Office of Research Services (AH 505) or by e-mail to research.ethics@uregina.ca

UNIVERSITY OF
REGINA

OFFICE OF RESEARCH SERVICES

M E M O R A N D U M

DATE: January 17, 2006

TO: O. Hoeber
Computer Science

FROM: Katherine Arbuthnott, Chair
Research Ethics Board

RE: **Investigating the Effectiveness of Visualization for Search Results Evaluation (34S0506)**

With reference to your email of January 13, 2006, please be advised the changes have been approved as outlined for your project *Investigating the Effectiveness of Visualization for Search Results Evaluation* **(34S0506)**.

Please contact us if you have any further questions.

Sincerely,

Katherine Arbuthnott

KA/rr

P.S. Please send copies of your additional questions and task description (#6) for your file.

UNIVERSITY OF REGINA

DEPARTMENT OF COMPUTER SCIENCE

Regina, Saskatchewan
Canada S4S 0A2
phone: (306) 585.4663
fax: (306) 585.4745
www.cs.uregina.ca

INFORMED CONSENT BY SUBJECTS TO PARTICIPATE IN

Investigating the Effectiveness of Visualization for Search Results Evaluation

I understand that this form and the information it contains are given to me for my protection and full understanding of the procedures**.** My signature on this form signifies that I have received a copy of this consent form, that I understand the procedures to be used in this study and the personal risks and benefits to me in taking part, that I voluntarily agree to participate in the project, that I understand that I may withdraw my participation in this study at any time, and that my decision to participate in this study, and my subsequent involvement in it, will have absolutely no bearing on any other dealings I have with Mr. Hoeber or Dr. Yang.

Knowledge of my identity is not required. I will not be required to write my name or any other identifying information on the research questionnaires. Comments relevant to the web search tasks, or to the use of the software will be transcribed. All research materials will be held confidential by the Principal Investigators and kept in a secure location.

I agree to participate by completing a pre-task questionnaire; a set of control search tasks using a Google-like interface; a training task using each of the three visual interfaces; a set of a search tasks using each of the three visual interfaces; and a post-task questionnaire. I understand that these activities will require approximately one and a half hours, and will be conducted on campus.

I understand that this project was approved by the Research Ethics Board, University of Regina. If I have any questions or concerns about my rights or treatment as a subject, I may contact the Chair of the Research Ethics Board at 585-4775 or by e-mail: research.ethics@uregina.ca.

I may obtain copies of the results of this study, upon its completion, by contacting Dr. Yang, in care of the Computer Science Department, University of Regina.

NAME (please type or print legibly): _____

SIGNATURE: _____

WITNESS: _____

DATE: _____

Investigators:

Orland Hoeber
Department of Computer Science
University of Regina
hoeber@cs.uregina.ca
306-337-2345

Xue-Dong Yang
Department of Computer Science
University of Regina
yang@cs.uregina.ca
306-585-4692

PRE-TASK QUESTIONNAIRE

The following questions relate to your background and experience using computers as well as web search engines. Please answer each question by checking the most appropriate response immediately following each question. Your answers to these questions will allow for more accurate analysis of the data collected during the study.

1. What is your gender?

☐ **Male**   ☐ **Female**

2. To which age category do you belong?

☐ **18 - 25**   ☐ **26 - 35**   ☐ **36 - 49**   ☐ **50+**

3. How many years education have you completed past high school?

☐ **0**   ☐ **1 - 4**   ☐ **5 - 6**   ☐ **more than 6**

4. Which of the following general categories best describes your main area of interest?

☐ **Fine Arts**   ☐ **Business**   ☐ **Arts/Humanities**   ☐ **Science/Engineering**

5. How many times per week do you use a computer?

☐ **Less than 1**   ☐ **1 - 5**   ☐ **5 - 10**   ☐ **more than 10**

6. How would you rate your level of experience with using a computer?

☐ **Low**   ☐ **Medium**   ☐ **High**

7. How many web searches do you perform per week?

☐ **Less than 1**   ☐ **1 - 5**   ☐ **5 - 10**   ☐ **more than 10**

8. What is your preferred web search engine?

☐ **Google**   ☐ **Yahoo**   ☐ **MSN**   ☐ **Other (please specify)**
                                                                    _____

9. How many pages of search results do you normally review before stopping?

☐ **1 - 2**   ☐ **3 - 4**   ☐ **5 - 6**   ☐ **7 - 8**   ☐ **9 - 10**   ☐ **more than 10**

10. How would you rate your level of experience with searching the web?

☐ **Low**   ☐ **Medium**   ☐ **High**

TASK QUESTIONNAIRE (HotMap)

1. Please familiarize yourself with the following description of your search task.

"communications network management"

You are a network manager for a small company. You are looking for information on tools, software, or services to assist you in managing your company's communications network (i.e., a data and/or voice network).

2. Use HotMap to search for relevant documents. Rank the documents you consider using the following scale until you have found 10 documents that are relevant (given a rank of 4):

     1 – This document is not relevant.
     2 – This document might be relevant.
     3 – This document is probably relevant.
     4 – This document is relevant.

D	R	T	D	R	T	D	R	T	D	R	T	D	R	T

3. Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
   (not confident)   (moderately confident)   (very confident)

4. Please rate how easy you found it to use this interface to evaluate the search results:

   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
   (not easy)   (moderately easy)   (very easy)

5. Please rate how satisfied you are using this interface to evaluate the search results:

   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
   (not satisfied)   (moderately satisfied)   (very satisfied)

6. Please rate how ambiguous you thought the search results set was:

   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
   (very ambiguous)   (moderately ambiguous)   (very specific)

TASK QUESTIONNAIRE (HotMap)

1. Please familiarize yourself with the following description of your search task.

"representing knowledge information"

You are a software developer working on a new project that requires complex knowledge and information to be stored and maintained. You are looking for information about how to represent this information in your software system.

2. Use HotMap to search for relevant documents. Rank the documents you consider using the following scale until you have found 10 documents that are relevant (given a rank of 4):
      1 – This document is not relevant.
      2 – This document might be relevant.
      3 – This document is probably relevant.
      4 – This document is relevant.

D	R	T	D	R	T	D	R	T	D	R	T	D	R	T

3. Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (not confident)   (moderately confident)   (very confident)

4. Please rate how easy you found it to use this interface to evaluate the search results:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (not easy)   (moderately easy)   (very easy)

5. Please rate how satisfied you are using this interface to evaluate the search results:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (not satisfied)   (moderately satisfied)   (very satisfied)

6. Please rate how ambiguous you thought the search results set was:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
  (very ambiguous)   (moderately ambiguous)   (very specific)

TASK QUESTIONNAIRE (Concept Highlighter)

1. Please familiarize yourself with the following description of your search task.

"communications network management"

You are a network manager for a small company. You are looking for information on tools, software, or services to assist you in managing your company's communications network (i.e., a data and/or voice network).

2. Use Concept Highlighter to search for relevant documents. Rank the documents you consider using the following scale until you have found 10 documents that are relevant (given a rank of 4):
       1 – This document is not relevant.
       2 – This document might be relevant.
       3 – This document is probably relevant.
       4 – This document is relevant.

D	R	T	D	R	T	D	R	T	D	R	T	D	R	T

3. Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:
   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
  (not confident)   (moderately confident)   (very confident)

4. Please rate how easy you found it to use this interface to evaluate the search results:
   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
  (not easy)   (moderately easy)   (very easy)

5. Please rate how satisfied you are using this interface to evaluate the search results:
   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
  (not satisfied)   (moderately satisfied)   (very satisfied)

6. Please rate how ambiguous you thought the search results set was:
   ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
  (very ambiguous)   (moderately ambiguous)   (very specific)

TASK QUESTIONNAIRE (Concept Highlighter)

1. Please familiarize yourself with the following description of your search task.

"representing knowledge information"

You are a software developer working on a new project that requires complex knowledge and information to be stored and maintained. You are looking for information about how to represent this information in your software system.

2. Use Concept Highlighter to search for relevant documents. Rank the documents you consider using the following scale until you have found 10 documents that are relevant (given a rank of 4):
> **1 – This document is not relevant.**
> **2 – This document might be relevant.**
> **3 – This document is probably relevant.**
> **4 – This document is relevant.**

D	R	T	D	R	T	D	R	T	D	R	T	D	R	T

3. Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:
> ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
> (not confident)   (moderately confident)   (very confident)

4. Please rate how easy you found it to use this interface to evaluate the search results:
> ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
> (not easy)   (moderately easy)   (very easy)

5. Please rate how satisfied you are using this interface to evaluate the search results:
> ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
> (not satisfied)   (moderately satisfied)   (very satisfied)

6. Please rate how ambiguous you thought the search results set was:
> ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
> (very ambiguous)   (moderately ambiguous)   (very specific)

TASK QUESTIONNAIRE (Google)

1. Please familiarize yourself with the following description of your search task.

"communications network management"

You are a network manager for a small company. You are looking for information on tools, software, or services to assist you in managing your company's communications network (i.e., a data and/or voice network).

2. Use Google to search for relevant documents. Rank the documents you consider using the following scale until you have found 10 documents that are relevant (given a rank of 4):
    1 – This document is not relevant.
    2 – This document might be relevant.
    3 – This document is probably relevant.
    4 – This document is relevant.

D	R	T	D	R	T	D	R	T	D	R	T	D	R	T

3. Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (not confident)   (moderately confident)   (very confident)

4. Please rate how easy you found it to use this interface to evaluate the search results:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (not easy)   (moderately easy)   (very easy)

5. Please rate how satisfied you are using this interface to evaluate the search results:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (not satisfied)   (moderately satisfied)   (very satisfied)

6. Please rate how ambiguous you thought the search results set was:
    ☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
    (very ambiguous)   (moderately ambiguous)   (very specific)

TASK QUESTIONNAIRE (Google)

1. Please familiarize yourself with the following description of your search task.

"representing knowledge information"

You are a software developer working on a new project that requires complex knowledge and information to be stored and maintained. You are looking for information about how to represent this information in your software system.

2. Use Google to search for relevant documents. Rank the documents you consider using the following scale until you have found 10 documents that are relevant (given a rank of 4):
    1 – This document is not relevant.
    2 – This document might be relevant.
    3 – This document is probably relevant.
    4 – This document is relevant.

D	R	T	D	R	T	D	R	T	D	R	T	D	R	T

3. Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
(not confident)		(moderately confident)		(very confident)

4. Please rate how easy you found it to use this interface to evaluate the search results:
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
(not easy)		(moderately easy)		(very easy)

5. Please rate how satisfied you are using this interface to evaluate the search results:
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
(not satisfied)		(moderately satisfied)		(very satisfied)

6. Please rate how ambiguous you thought the search results set was:
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
(very ambiguous)		(moderately ambiguous)		(very specific)

POST-TASK QUESTIONNAIRE

The following questions relate to your experience using the visual representation of web search results to identify relevant documents. Your answer to these questions will allow for a more accurate analysis of the data collected during the study.

INSTRUCTIONS: Please rate how strongly you agree or disagree with each of the following statements by circling the appropriate number.

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	I found the *list* of search results in the Google interface easy to use.
0	1	2	3	4	5	I found the *title* of the document in the Google interface to be useful.
0	1	2	3	4	5	I found the *snippet* of the document in the Google interface to be useful.
0	1	2	3	4	5	I found the *URL* of the document in the Google interface to be useful.

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	I found the *document list window* in HotMap easy to use.
0	1	2	3	4	5	I found the *overview map* in HotMap easy to use.
0	1	2	3	4	5	I found the *tool tips* that contained the document snippet and URL in HotMap easy to use.
0	1	2	3	4	5	I found the *single-term sorting* feature in HotMap easy to use.
0	1	2	3	4	5	I found the *multi-term sorting* feature in HotMap easy to use.
0	1	2	3	4	5	I found the *colour-coding* of the query term frequency in HotMap to be useful.
0	1	2	3	4	5	I found the *coordination* between the document list window and the overview map in HotMap to be useful.
0	1	2	3	4	5	I found the ability to *jump* to documents of interest in HotMap to be useful.

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	I found the *document list window* in Concept Highlighter easy to use.
0	1	2	3	4	5	I found the *overview map* in Concept Highlighter easy to use.
0	1	2	3	4	5	I found the *tool tips* that contained the document snippet and URL in Concept Highlighter easy to use.
0	1	2	3	4	5	I found the *concept selection* in Concept Highlighter easy to use.
0	1	2	3	4	5	I found the *concept-based sorting* feature in Concept Highlighter easy to use.
0	1	2	3	4	5	I found the *colour-coding* of the concept scores in Concept Highlighter to be useful.
0	1	2	3	4	5	I found the *coordination* between the document list window and the overview map in Concept Highlighter to be useful.
0	1	2	3	4	5	I found the ability to *jump* to documents of interest in Concept Highlighter to be useful.

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	**I found the *document list window* in ConceptMap easy to use.**
0	1	2	3	4	5	**I found the *overview map* in ConceptMap easy to use.**
0	1	2	3	4	5	**I found the *tool tips* that contained the document snippet and URL in ConceptMap easy to use.**
0	1	2	3	4	5	**I found the *concept-based sorting* feature in ConceptMap easy to use.**
0	1	2	3	4	5	**I found the *colour-coding* of the concept scores in ConceptMap to be useful.**
0	1	2	3	4	5	**I found the *coordination* between the document list window and the overview map in ConceptMap to be useful.**
0	1	2	3	4	5	**I found the ability to *jump* to documents of interest in ConceptMap to be useful.**

Assuming the search results provided are the same, please rank your search interface preference (order from 1 to 4):

_____ Google
_____ HotMap
_____ Concept Highlighter
_____ ConceptMap

1. Is there any other information not provided by the software that you would find useful when deciding if a document is relevant to your search task?

2. Are there any other features that you would like to see in a visual web search tool?

3. Please make any other comments about the software and/or the search tasks.

Appendix C

User Study 2

In the second user study conducted in this research, participants used the various features of WordBars with the goal of finding as many relevant documents as possible among the first ten considered. The only system used in this study was WordBars, although the initial order of the search results was provided by Google.

This appendix includes the formal approval received from the Research Ethics Board for this study. All the evaluation documents are provided as well.

UNIVERSITY OF
REGINA

OFFICE OF RESEARCH SERVICES

M E M O R A N D U M

DATE: September 22, 2006

TO: Orland Hoeber
         Computer Science

FROM: K. Arbuthnott
         Chair, Research Ethics Board

Re: Investigating the Use of Term Frequency Histograms for Search Result Re-Ordering and Query Refinement (07S0607)

Please be advised that the University of Regina Research Ethics Board has reviewed your proposal and found it to be:

1. APPROVED AS SUBMITTED. Only applicants with this designation have ethical approval to proceed with their research as described in their applications. For research lasting more than one year (Section 1F). **ETHICAL APPROVAL MUST BE RENEWED BY SUBMITTING A BRIEF STATUS REPORT EVERY TWELVE MONTHS.** Approval will be revoked unless a satisfactory status report is received. Any substantive changes in methodology or instrumentation must also be approved prior to their implemenation.

Please correct 18 participant on email recruitment letter

2. ACCEPTABLE SUBJECT TO MINOR CHANGES AND PRECAUTIONS (SEE ATTACHED). Changes must be submitted to the REB and approved prior to beginning research. Please submit a supplementary memo addressing the concerns to the Chair of the REB.** Do not submit a new application. Once changes are deemed acceptable, ethical approval will be granted.

3. ACCEPTABLE SUBJECT TO MAJOR CHANGES AND PRECAUTIONS (SEE ATTACHED). Changes must be submitted to the REB and approved prior to beginning research. Please submit a supplementary memo addressing the concerns to the Chair of the REB.** Do not submit a new application. Once changes are deemed acceptable, ethical approval will be granted.

4. UNACCEPTABLE AS SUBMITTED. The proposal requires substantial additions or redesign. Please contact the Chair of the REB for advice on how the project proposal might be revised.

Dr. Katherine Arbuthnott

c. supervisor

**supplementary memo should be forwarded to the Chair of the Research Ethics Board at the Office of Research Services (AH 505) or by email to research.ethics@uregina.ca

UNIVERSITY OF REGINA

DEPARTMENT OF COMPUTER SCIENCE

Regina, Saskatchewan
Canada S4S 0A2
phone: (306) 585.4663
fax: (306) 585.4745
www.cs.uregina.ca

INFORMED CONSENT BY SUBJECTS TO PARTICIPATE IN

Investigating the Use of Term Frequency Histograms for Search Result Re-Ordering and Query Refinement

I understand that this form and the information it contains are given to me for my protection and full understanding of the procedures. My signature on this form signifies that I have received a copy of this consent form, that I understand the procedures to be used in this study and the personal risks and benefits to me in taking part, that I voluntarily agree to participate in the project, that I understand that I may withdraw my participation in this study at any time without any penalty, and that my decision to participate in this study, and my subsequent involvement in it, will have absolutely no bearing on any other dealings I have with Mr. Hoeber or Dr. Yang. I understand that the researchers are willing to respond to any questions about the procedures and/or goals of the study, as well as questions about the potential benefits and risks associated with the study.

Knowledge of my identity is not required. I will not be required to write my name or any other identifying information on the research questionnaires. Comments relevant to the web search tasks, or to the use of the software will be transcribed. All research materials will be held confidential by the Principal Investigators and kept in a secure location.

I agree to participate by completing a pre-task questionnaire; a training task using the web search tool; three web search tasks based on information provided; and a post-task questionnaire. I understand that these activities will require approximately one hour, and will be conducted on campus.

I understand that this project was approved by the Research Ethics Board, University of Regina. If I have any questions or concerns about my rights or treatment as a subject, I may contact the Chair of the Research Ethics Board at 585-4775 or by e-mail: research.ethics@uregina.ca.

I may obtain copies of the results of this study, upon its completion, by contacting Dr. Yang, in care of the Computer Science Department, University of Regina.

NAME (please type or print legibly): _____

SIGNATURE: _____

WITNESS: _____

DATE: _____

Investigators:
Orland Hoeber
Department of Computer Science
University of Regina
hoeber@cs.uregina.ca
306-337-2345

Xue-Dong Yang
Department of Computer Science
University of Regina
yang@cs.uregina.ca
306-585-4692

PRE-TASK QUESTIONNAIRE

The following questions relate to your background and experience using computers as well as web search engines. Please answer each question by checking the most appropriate response immediately following each question. Your answers to these questions will allow for more accurate analysis of the data collected during the study.

1. What is your gender?

☐ **Male** ☐ **Female**

2. To which age category do you belong?

☐ **18 - 25** ☐ **26 - 35** ☐ **36 - 49** ☐ **50+**

3. How many years education have you completed past high school?

☐ **0** ☐ **1 - 2** ☐ **3 - 4** ☐ **more than 4**

4. Which of the following general categories best describes your main area of interest?

☐ **Fine Arts** ☐ **Business** ☐ **Arts/Humanities** ☐ **Science/Engineering**

5. How many times per week do you use a computer?

☐ **Less than 1** ☐ **1 - 5** ☐ **5 - 10** ☐ **more than 10**

6. How would you rate your level of experience with using a computer?

☐ **Low** ☐ **Medium** ☐ **High**

7. How many web searches do you perform per week?

☐ **Less than 1** ☐ **1 - 5** ☐ **5 - 10** ☐ **more than 10**

8. What is your preferred web search engine?

☐ **Google** ☐ **Yahoo** ☐ **MSN** ☐ **Other (please specify)**
_____

9. How many pages of search results do you normally review before stopping?

☐ **1** ☐ **3** ☐ **5** ☐ **7** ☐ **9** ☐ **more than 10**
☐ **2** ☐ **4** ☐ **6** ☐ **8** ☐ **10**

10. If a query does not lead to the information you are seeking, how likely are you to add new terms to the query?

☐ **Seldom** ☐ **Sometimes** ☐ **Often** ☐ **Always**

11. If a query does not lead to the information you are seeking, how likely are you to remove terms from the query?

☐ **Seldom** ☐ **Sometimes** ☐ **Often** ☐ **Always**

12. How would you rate your level of experience with searching the web?

☐ **Low** ☐ **Medium** ☐ **High**

TASK QUESTIONNAIRE Topic A

Please familiarize yourself with the following description of your search task.

"new hydroelectric projects"

Identify hydroelectric projects proposed or under construction by country and location. Detailed description of nature, extent, purpose, problems, and consequences is desirable.

Using the original order of documents, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank
1		6	
2		7	
3		8	
4		9	
5		10	

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to re-sort the search results. Once you are satisfied, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank

Terms Selected:

Time to Select: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to add or remove from your query. When you are done, click the search button, and rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank
1		6	
2		7	
3		8	
4		9	
5		10	

New Query:

Time to Select: _____

Time to Get Results: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to re-sort the search results. Once you are satisfied, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank

Terms Selected:

Time to Select: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(very ambiguous)   (moderately ambiguous)   (very specific)

Please rate how easy you found it to use this interface to evaluate the search results:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(not easy)   (moderately easy)   (very easy)

Please rate how easy you found it to use this interface to select terms to add or remove from your query:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**

(not easy)   (moderately easy)   (very easy)

TASK QUESTIONNAIRE   **Topic B**

Please familiarize yourself with the following description of your search task.

"international art crime"

Isolate instances of fraud or embezzlement in the international art trade.

Using the original order of documents, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank
1		6	
2		7	
3		8	
4		9	
5		10	

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to re-sort the search results. Once you are satisfied, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank

Terms Selected:

Time to Select: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to add or remove from your query. When you are done, click the search button, and rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank
1		6	
2		7	
3		8	
4		9	
5		10	

New Query:

Time to Select: _____

Time to Get Results: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to re-sort the search results. Once you are satisfied, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank

Terms Selected:

Time to Select: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(very ambiguous)   (moderately ambiguous)   (very specific)

Please rate how easy you found it to use this interface to evaluate the search results:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not easy)   (moderately easy)   (very easy)

Please rate how easy you found it to use this interface to select terms to add or remove from your query:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not easy)   (moderately easy)   (very easy)

TASK QUESTIONNAIRE   **Topic C**

Please familiarize yourself with the following description of your search task.

"euro opposition"

Identify documents that discuss opposition to the introduction of the euro, the European currency.

Using the original order of documents, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank
1		6	
2		7	
3		8	
4		9	
5		10	

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to re-sort the search results. Once you are satisfied, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank

Terms Selected:

Time to Select: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to add or remove from your query. When you are done, click the search button, and rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank
1		6	
2		7	
3		8	
4		9	
5		10	

New Query:

Time to Select: _____

Time to Get Results: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(very ambiguous)   (moderately ambiguous)   (very specific)

Select one or more terms from the histogram to re-sort the search results. Once you are satisfied, rank the first 10 documents you consider on a 4-point scale.

Document	Rank	Document	Rank

Terms Selected:

Time to Select: _____

Time to Completion: _____

Please rate your confidence in your ability to find a good set of documents that are relevant to the search task:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not confident)   (moderately confident)   (very confident)

Please rate how satisfied you are with the search results set:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not satisfied)   (moderately satisfied)   (very satisfied)

Please rate how ambiguous you thought the search results set was:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(very ambiguous)   (moderately ambiguous)   (very specific)

Please rate how easy you found it to use this interface to evaluate the search results:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not easy)   (moderately easy)   (very easy)

Please rate how easy you found it to use this interface to select terms to add or remove from your query:

☐ **1**   ☐ **2**   ☐ **3**   ☐ **4**   ☐ **5**
(not easy)   (moderately easy)   (very easy)

POST-TASK QUESTIONNAIRE

The following questions relate to your experience using the visual representation of web search results to identify relevant documents. Your answer to these questions will allow for a more accurate analysis of the data collected during the study.

INSTRUCTIONS: Please rate how strongly you agree or disagree with each of the following statements by circling the appropriate number.

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	**I found it easy to evaluate the list of search results in the original order provided.**
0	1	2	3	4	5	**I found the *title* of the document to be helpful in determining the relevance.**
0	1	2	3	4	5	**I found the *snippet* of the document to be helpful in determining the relevance.**
0	1	2	3	4	5	**I found the *URL* of the document to be helpful in determining the relevance.**

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	**I found the histogram of term frequencies to be helpful as an overview of the search results.**
0	1	2	3	4	5	**I found it easy to select terms from the histogram to re-order the search results.**
0	1	2	3	4	5	**I found it easy to add new terms to the query using the histogram.**
0	1	2	3	4	5	**I found the colour coding of the histogram easy to understand.**
0	1	2	3	4	5	**I found the length of the bars in the histogram easy to understand.**
0	1	2	3	4	5	**I found the combination of length and colour coding in the histogram to be beneficial.**
0	1	2	3	4	5	**I found good terms in the histogram that I wouldn't have thought of to use in my queries.**
0	1	2	3	4	5	**I found bad terms in the histogram that clearly were not relevant to my queries.**

n/a	strongly disagree	disagree	neutral	agree	strongly agree	
0	1	2	3	4	5	**I found the search results to be more relevant after re-ordering the results based on my selected terms.**
0	1	2	3	4	5	**I found the search results to be more relevant after refining the query.**
0	1	2	3	4	5	**I found the search results to be more relevant after re-ordering the search results of the refined query.**
0	1	2	3	4	5	**I found the original search results in the original order to be the best.**

Please rank your preference of the search results considered from 1 (best) to 4 (worst):

_____ search results in original order
_____ search results after re-ordering
_____ search results after refining the query
_____ search results after refining the query and re-ordering

If you have any other comments, you may write them on the back of this page.

Bibliography

[1] ACM. ACM computing classification system. `http://www.acm.org/class/`, 2004.

[2] Alexa Internet, Inc. Alexa developer's corner. `http://www.alexa.com/site/devcorner`, 2006.

[3] Keith Andrews, Christian Gutl, Josef Moser, and Vedran Sabol. Search result visualization with xFind. In *Proceedings of the Second International Workshop on User Interfaces to Data Intensive Systems*, 2001.

[4] P. Anick, J. Brennan, R. Flynn, D. Hanssen, B. Alvey, and J. Robbins. A direct manipulation interface for boolean information retrieval via natural language query. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1990.

[5] Apache. Apache Lucene open-source search software. `http://lucene.apache.org/`, 2006.

[6] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[7] Michelle Q. Wang Baldonado, Alliston Woodruff, and Allan Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the ACM Advanced Visual Interfaces*, 2000.

[8] Benjamin B. Benderson and Ben Shneiderman. *The Craft of Information Visualization.* Morgan Kaufmann, 2003.

[9] Ezio Berenci, Claudio Carpineto, Vittorio Giannini, and Stefano Mizzaro. Effectiveness of keyword-based display and selection of retrieval results for interactive searches. *International Journal on Digital Libraries*, 3(3):249–260, 2000.

[10] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.

[11] Bodo Billerbeck, Falk Scholer, Hugh Williams, and Justin Zobel. Query expansion using associated queries. In *Proceedings of the ACM Conference on Information and Knowledge Management*, 2003.

[12] Cynthia A. Brewer. Color brewer. `www.colorbrewer.org`, 2005.

[13] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[14] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization.* Morgan Kaufmann, 1999.

[15] Chia-Hui Chang and Ching-Chi Hsu. Enabling concept-based relevance feedback for information retrieval on the WWW. *IEEE Transactions on Knowledge and Data Engineering*, 11(4), 1999.

[16] Paul-Alexandru Chirita, Wolfgang Nejdl, Raluca Paiu, and Christian Kohlschutter. Using ODP metadata to personalize search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

[17] Douglass Cutting, David Karger, Jan Pedersend, and John Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.

[18] Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, 31:121–187, 1996.

[19] Stephen Eick. Graphically displaying text. *Journal of Computational Graphics and Statistics*, 3(2):127–142, 1994.

[20] Daniel Fallman. Design-oriented human-comptuer interaction. *CHI Letters*, 5(1):225–232, 2003.

[21] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[22] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: mathematical foundations*. Springer-Verlag, 1999.

[23] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the ACM Symposium on Operating System Principles*, 2003.

[24] Cliff Goddard. *Semantic Analysis: A Practical Introduction*. Oxford University Press, 1998.

[25] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing, Second Edition.* Prentice Hall, 2002.

[26] Google. Google search engine. `http://www.google.com/`, 2005.

[27] Google. Google web API. `http://www.google.com/apis/`, 2005.

[28] Graphics, Visualization, & Usability Center. GVU's 10th WWW user survey. `http://www.gvu.gatech.edu/user_surveys/survey-1998-10/`, 1998.

[29] Grokker. Grokker search engine. `http://www.grokker.com/`, 2005.

[30] Donna Harman. Towards interactive query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988.

[31] Donna Harman. Relevance feedback revisited. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.

[32] Marti Hearst. TileBars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1995.

[33] Marti Hearst. User interfaces and visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Modern Information Retrieval.* Addison-Wesley, 1999.

[34] Marti Hearst and Jan Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[35] Tomi Heimonen and Natalie Jhaveri. Visualizing query occurrence in search result lists. In *Proceedings of the International Conference on Information Visualization*, 2005.

[36] Ewald Hering. *Outlines of a Theory of Light Sense (Grundzge der Lehr von Lichtsinn, 1920)*. Harvard University Press, 1964.

[37] Orland Hoeber and Xue Dong Yang. A comparative user study of Web search interfaces: HotMap, Concept Highlighter, and Google. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[38] Orland Hoeber and Xue Dong Yang. Exploring Web search results using coordinated views. In *Proceedings of the International Conference on Coordinated and Multiple Views*, 2006.

[39] Orland Hoeber and Xue Dong Yang. Interactive Web information retrieval using WordBars. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[40] Orland Hoeber and Xue Dong Yang. A model for interactive Web information retrieval. In *Proceedings of the International Symposium on Smart Graphics*, 2006.

[41] Orland Hoeber and Xue Dong Yang. The visual exploration of Web search results using HotMap. In *Proceedings of the International Conference on Information Visualization*, 2006.

[42] Orland Hoeber and Xue Dong Yang. Visually exploring concept-based fuzzy clusters in Web search results. In *Proceedings of the Atlantic Web Intelligence Conference*, 2006.

[43] Orland Hoeber, Xue-Dong Yang, and Yiyu Yao. Conceptual query expansion. In *Proceedings of the Atlantic Web Intelligence Conference*, 2005.

[44] Orland Hoeber, Xue-Dong Yang, and Yiyu Yao. Visualization support for interactive query refinement. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.

[45] Orland Hoeber, Xue Dong Yang, and Yiyu Yao. VisiQ: Supporting visual and interactive query refinement. *Web Intelligence and Agent Systems: An International Journal*, 2007 (to appear).

[46] Paul Jen-Hwa Hu, Pai-Chun Ma, and Patrick Y.K. Chau. Evaluation of user interface designs for information retrieval systems: a computer-based experiment. *Decision Support Systems*, 27, 1999.

[47] A.K. Jain, M.N.Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[48] Bernard J. Jansen and Udo Pooch. A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3):235–246, 2001.

[49] Hideo Joho, Claire Coverson, Mark Sanderson, and Micheline Beaulieu. Hierarchical presentation of expansion terms. In *Proceedings of the ACM Symposium on Applied Computing*, 2002.

[50] Hideo Joho, Mark Sanderson, and Micheline Beaulieu. A study of user interaction with a concept-based interactive query expansion support tool. In *Proceedings of the European Conference on IR Research*, 2004.

[51] Steve Jones. Graphical query specification and dynamic results previews for a digital library. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, 1998.

[52] Mei Kobayashi and Koichi Takeda. Information retrieval on the Web. *ACM Computing Surveys*, 32(2):114–173, 2000.

[53] Kurt Koffka. *Principles of Gestalt Psychology*. Harcourt, Brace and Company, New York, 1935.

[54] Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2(1):1–15, 2000.

[55] Bill Kules, Jack Kustanowitz, and Ben Shneiderman. Categorizing Web search results into meaningful and stable categories using fast-feature techniques. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, 2006.

[56] Steve Lawrence and C. Lee Giles. Accessibility of information on the Web. *Nature*, 400:107–109, 1999.

[57] Mark Magennis and Cornelis J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997.

[58] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[59] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. The use of WordNet in information retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 2000.

[60] Gary Marchionini. Interfaces for end-user information seeking. *Journal of the American Society for Information Science*, 43(2):156–163, 1992.

[61] George A. Miller. Wordnet: An online lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[62] Netscape. Open directory project. `http://www.dmoz.org/`, 2004.

[63] Günter Neumann and Feiyu Xu. Mining natural language answers from the web. *Web Intelligence and Agent Systems: An International Journal*, 2(2):123–135, 2004.

[64] Jakob Nielsen. *Usability Engineering*. Academic Press, 1993.

[65] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1994.

[66] Jakob Nielsen. Search: Visible and simple. Alertbox (`http://www.useit.com/alertbox/20010513.html`), May 2001.

[67] Jakob Nielsen. When search engines become answer engines. Alertbox (`http://www.useit.com/alertbox/20040816.html`), August 2004.

[68] Donald Norman. Cognitive engineering. In D. A. Norman and S. D. Draper, editors, *User Centred Systems Design*. Lawrence Erlbaum Associates, 1986.

[69] Donald Norman. *The Design of Everyday Things.* Doubleday, 1988.

[70] Lucy Nowell, Robert France, Deborah Hix, Lenwood Heath, and Edward Fox. Visualizing search results: Some alternatives to query-document similarity. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[71] Stacy Nowicki. Student vs. search engine: Undergraduates rank results for relevance. *Libraries and the Academy*, 3(3), 2003.

[72] Joshua O'Madadhain, Danyel Fisher, Scott White, and Yan-Biao Boey. The JUNG (java universal network/graph) framework. Technical Report UCI-ICS 03-17, School of Information and Computer Science, University of California, Irvine, 2003.

[73] Tim Paek, Susan Dumais, and Ron Logan. Wavelens: A new view onto internet search results. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2004.

[74] Martin Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[75] David Powers and Darius Pfitzner. The magic science of visualization. In *Proceedings of the International Conference on Cognitive Science*, 2003.

[76] Jennifer Preece, Yvonne Rogers, and Helen Sharp. *Interaction Design: beyond human-computer interaction.* Johyn Wiley & Sons, 2002.

[77] Yonggang Qiu and H. P. Frei. Concept based query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

[78] Filip Radlinski and Susan Dumais. Improving personalized Web search using result diversification. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[79] Edie M. Rasmussen. Indexing and retrieval for the Web. *Annual Review of Information Science and Technology*, 37(1):91–124, 2003.

[80] Earl Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, 1994.

[81] S. E. Robertson and K. Sparck Jones. Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory, 1997.

[82] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART System - Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[83] Yvonne Rogers. New theoretical approaches for human-computer interaction. *Annual Review of Information Science and Technology*, 38(1):87 – 143, 2004.

[84] Mary Beth Rosson and John M. Carroll. *Usability Engineering: scenario-based development of human-computer interaction*. Morgan Kaufmann, 2002.

[85] Ian Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[86] Gerald Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[87] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[88] Ben Shneiderman. *Designing the User Interface*. Addison-Wesley, 1998.

[89] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[90] Robert Spence. A framework for navigation. *International Journal of Human-Computer Studies*, 51(5):919–945, 1999.

[91] Robert Spence. *Information Visualization*. ACM Press, 2001.

[92] Amanda Spink. A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information Processing and Management*, 38(3):401–426, 2002.

[93] Amanda Spink, Dietmar Wolfram, B. J. Jansen, and Tefko Saracevic. Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[94] A. Spoerri. Infocrystal: A visualization tool for information retrieval. In *Proceedings of IEEE Visualization*, 1993.

[95] Nenad Stojanovic. Information-need driven query refinement. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, 2003.

[96] Nenad Stojanovic. On the role of query refinement in searching for information: The librarian agent query refinement process. In *Proceedings of the IEEE International Conference on Web Information Systems Engineering*, 2003.

[97] Nenad Stojanovic. Information-need driven query refinement. *Web Intelligence and Agent Systems: An International Journal*, 3(3):155–169, 2005.

[98] Maureen C. Stone. *A Field Guide to Digital Color*. A. K. Peters, 2003.

[99] Louise T. Su. A comprehensive and systematic model of user evaluation of web search enginges: Ii. an evaluation by undergraduates. *Journal of the American Society for Information Science and Technology*, 54(13), 2003.

[100] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. Adaptive Web search based on user profile construction without any effort from users. In *Proceedings of the 2004 World Wide Web Conference (WWW2004)*, 2004.

[101] Jamie Teevan, Susan Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

[102] Edward Tufte. *Envisioning Information*. Graphics Press, 1990.

[103] Edward Tufte. *Visual Explanations*. Graphics Press, 1997.

[104] Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[105] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

[106] Liwen Vaughan. New measures for search engine evaluation proposed and tested. *Information Processing and Management*, 40(4):677–691, 2004.

[107] Vivisimo. Vivisimo search engine. `http://www.vivisimo.com/`, 2005.

[108] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[109] Ellen M. Voorhees. Overview of TREC 2004. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

[110] Colin Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann, 2004.

[111] Wikimedia Foundation. Wikipedia - the free encyclopedia. `http://en.wikipedia.org/`, 2006.

[112] James A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1223–1233, 1999.

[113] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of IEEE Information Visualization*, 1995.

[114] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[115] Yahoo! Yahoo! search Web services. `http://developer.yahoo.com/ search/`, 2005.

[116] Yahoo! Yahoo! Web search engine. `http://www.yahoo.com/`, 2005.

[117] Kiduk Yang. Information retrieval on the Web. *Annual Review of Information Science and Technology*, 39(1):33 – 80, 2005.

[118] Yiyu Yao. Information retrieval support systems. In *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, 2002.

[119] Degi Young and Ben Shneiderman. A graphical filter/flow model for boolean queries: An implementation and experiment. *Journal of the American Society for Information Science*, 44(6):327–339, 1993.

[120] Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[121] Oren Zamir and Oren Etzioni. Grouper: A dynamic clustering interface to Web search results. In *Proceedings of the Eighth International World Wide Web Conference*, 1999.

[122] Bin Zhu and Hsinchun Chen. Information visualization. *Annual Review of Information Science and Technology*, 39(1):139 – 177, 2005.