A Comparative User Study of Web Search Interfaces: HotMap, Concept Highlighter, and Google

Orland Hoeber and Xue Dong Yang

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{hoeber, yang}@cs.uregina.ca

Abstract

Users of traditional web search engines commonly find it difficult to evaluate the results of their web searches. We suggest the use of information visualization and interactive visual manipulation as methods for improving the ability of users to evaluate the results of a web search. In this paper, we present the results of a user study that compared the search results interface provided by Google to that of two systems we have developed: HotMap and Concept Highlighter. We found that users were able to perform their searches faster with HotMap, were able to find more relevant documents with Concept Highlighter, and generally ranked these interfaces higher than Google with respect to subjective measures. When given a choice between these interfaces, participants ranked HotMap the highest, followed by Google and Concept Highlighter. These results indicate that even though the list-based representation of search results are common among search engines, visual and interactive interfaces to web search results can be more efficient, effective, and satisfying to the users.

1. Introduction

A number of studies on web search user traits have noted that users seldom view more than three pages of web search results [13, 14]. In situations where the searchers are able to craft an accurate query, it is possible that they are able to find enough relevant documents in the first few pages to satisfy their information need. However, when vague or misleading queries are used, or when the information need is inherently ambiguous, it is more common for users to either give up or re-formulate their query, rather than continue to evaluate the search results.

Part of the problem is that the list-based representation commonly used by web search engines provides little support for the users' task of deciding the relevance of the document surrogates in the search results collection. This static list promotes the evaluation of each document surrogate individually, and to some degree, in the order provided. Further, the primarily textual contents of the search results list makes it difficult to quickly evaluate the search results.

Our work has been motivated by a desire to represent features of the search results set in a visual manner, and to allow the users to interactively manipulate and explore the search results. Wise et al. noted 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" [16]. We have attempted to address this upper limit through the development of two prototype systems: HotMap [7] and Concept Highlighter [8].

The Google interface promotes the traditional model of information retrieval where a passive evaluation of the search results is supported. HotMap and Concept Highlighter extend this model through interactive search results exploration, allowing the users to take an active role in the evaluation of the search results [6]. These tools represent a step towards Yao's vision for web information retrieval support systems [17].

In this paper, we present the results of a user study that compares HotMap and Concept Highlighter against the list-based representation used by Google. Comparisons are made in terms of the time taken to complete the tasks, the perceived precision of the search results, the subjective reactions to the interfaces, and the preference ranks made by the participants.

In the following section, a review and summary of the related work is provided. In Section 3, the experimental design employed in this study is described. Section 4 presents the results of the user evaluation of the web search results interfaces, followed by a discussion about the results in Section 5. A summary of the conclusions, along with a description of future work is provided in Section 6.

2. Related Work

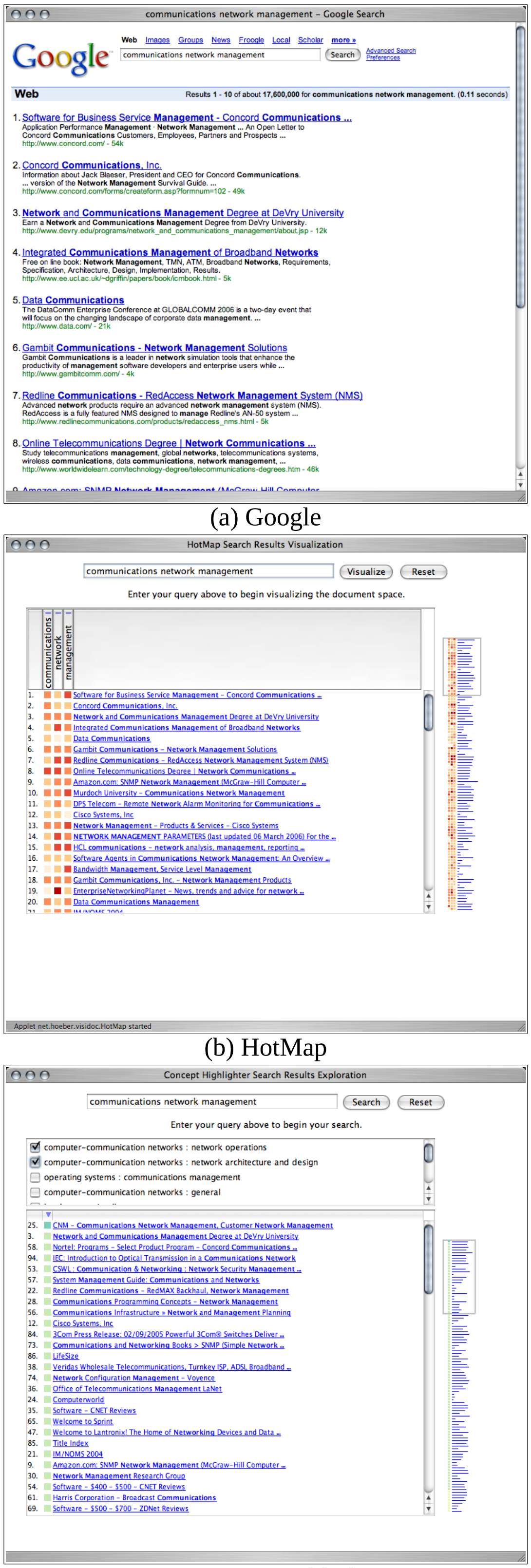
Researchers have been exploring the application of information visualization techniques to the representation of collections of documents for many years. Some noteworthy systems include Galaxy of News [11], ThemeScape [16], SeeSoft [3], and TileBars [5]. Although these techniques have been shown to be beneficial for traditional information retrieval systems, they commonly require access to the entire document contents, which is not feasible for web information retrieval.

Other systems have been developed specifically for the web, including VIEWER [2], xFind [1], WaveLens [10], and Grokker [4]. These systems all provide some degree of graphical representation of the search results, and allow the users to interactively explore the search results. Following the success of these systems, and addressing some of their shortcomings, we have developed two methods to visually represent the results of a web search which support the interactive exploration of the search results.

HotMap [7] (Figure 1b) visually depicts the frequency of each of the terms in the users' queries using a compact colour coding for each of the document surrogates in the search results. This allows "hot" documents to be easily identified with a simple glance, based on the frequent appearance of the query terms within the document surrogates. The interactive exploration of the search results is supported via dynamic re-sorting of the document surrogates based on the query term frequencies.

Concept Highlighter [8] (Figure 1c) obtains a set of relevant concepts from a concept knowledge base using the users' query terms; interactive concept-based fuzzy clustering is used to cluster the search results with respect to these concepts. As the users select the concepts that are relevant to their information need, the search results are re-sorted based on the fuzzy membership score of each document surrogate with respect to the selected concepts. Colour coding is used to visually represent the fuzzy membership scores, allowing the users to easily determine the degree to which each document surrogate belongs to the user-selected set of concepts.

These systems use a similar framework for providing coordinated views of the search results at two levels of detail. The overview map depicts the top 100 document surrogates returned by the underlying search engine in a compact and abstract representation (shown in Figures 1b and 1c on the right side). The detail window (the large list region shown in Figures 1b and 1c) shows 20 to 25 document surrogates at a time, allowing the users to determine the relevance of


communications network management – Google Search
Google
Web
Images
Groups
News
Froogle
Local
Scholar
more »
communications network management
Search
Advanced Search
Preferences

Web Results 1 - 10 of about 17,600,000 for communications network management. (0.11 seconds)

1. Software for Business Service Management - Concord Communications ...
Application Performance Management · Network Management ... An Open Letter to Concord Communications Customers, Employees, Partners and Prospects ...
http://www.concord.com/ - 54k

2. Concord Communications, Inc.
Information about Jack Blaeser, President and CEO for Concord Communications. ... version of the Network Management Survival Guide. ...
http://www.concord.com/forms/createform.asp?formnum=102 - 49k

3. Network and Communications Management Degree at DeVry University
Earn a Network and Communications Management Degree from DeVry University.
http://www.devry.edu/programs/network_and_communications_management/about.jsp - 12k

4. Integrated Communications Management of Broadband Networks
Free on line book: Network Management, TMN, ATM, Broadband Networks, Requirements, Specification, Architecture, Design, Implementation, Results.
http://www.ee.ucl.ac.uk/~dgriffin/papers/book/icmbook.html - 5k

5. Data Communications
The DataComm Enterprise Conference at GLOBALCOMM 2006 is a two-day event that will focus on the changing landscape of corporate data management. ...
http://www.data.com/ - 21k

6. Gambit Communications - Network Management Solutions
Gambit Communications is a leader in network simulation tools that enhance the productivity of management software developers and enterprise users while ...
http://www.gambitcomm.com/ - 4k

7. Redline Communications - RedAccess Network Management System (NMS)
Advanced network products require an advanced network management system (NMS). RedAccess is a fully featured NMS designed to manage Redline's AN-50 system ...
http://www.redlinecommunications.com/products/redaccess_nms.html - 5k

8. Online Telecommunications Degree | Network Communications ...
Study telecommunications management, global networks, telecommunications systems, wireless communications, data communications, network management, ...
http://www.worldwidelearn.com/technology-degree/telecommunications-degrees.htm - 46k

(a) Google

HotMap Search Results Visualization
communications network management
Visualize
Reset
Enter your query above to begin visualizing the document space.
communications
network
management

1. Software for Business Service Management – Concord Communications ...
2. Concord Communications, Inc.
3. Network and Communications Management Degree at DeVry University
4. Integrated Communications Management of Broadband Networks
5. Data Communications
6. Gambit Communications – Network Management Solutions
7. Redline Communications – RedAccess Network Management System (NMS)
8. Online Telecommunications Degree | Network Communications ...
9. Amazon.com: SNMP Network Management (McGraw-Hill Computer ...
10. Murdoch University – Communications Network Management
11. DPS Telecom – Remote Network Alarm Monitoring for Communications ...
12. Cisco Systems, Inc
13. Network Management – Products & Services – Cisco Systems
14. NETWORK MANAGEMENT PARAMETERS (last updated 06 March 2006) For the ...
15. HCL communications – network analysis, management, reporting ...
16. Software Agents in Communications Network Management: An Overview ...
17. Bandwidth Management, Service Level Management
18. Gambit Communications, Inc. – Network Management Products
19. EnterpriseNetworkingPlanet – News, trends and advice for network ...
20. Data Communications Management

Applet net.hoeber.visidoc.HotMap started

(b) HotMap

Concept Highlighter Search Results Exploration
communications network management
Search
Reset
Enter your query above to begin your search.

☑ computer–communication networks : network operations
☑ computer–communication networks : network architecture and design
☐ operating systems : communications management
☐ computer–communication networks : general

25. CNM – Communications Network Management, Customer Network Management
3. Network and Communications Management Degree at DeVry University
58. Nortel: Programs – Select Product Program – Concord Communications ...
94. IEC: Introduction to Optical Transmission in a Communications Network
53. CSWL : Communication & Networking : Network Security Management ...
57. System Management Guide: Communications and Networks
22. Redline Communications – RedMAX Backhaul, Network Management
28. Communications Programming Concepts – Network Management
56. Communications Infrastructure » Network and Management Planning
12. Cisco Systems, Inc
84. 3Com Press Release: 02/09/2005 Powerful 3Com® Switches Deliver ...
73. Communications and Networking Books > SNMP (Simple Network ...
86. LifeSize
38. Veridas Wholesale Telecommunications, Turnkey ISP, ADSL Broadband ...
74. Network Configuration Management – Voyence
36. Office of Telecommunications Management LaNet
24. Computerworld
35. Software – CNET Reviews
65. Welcome to Sprint
47. Welcome to Lantronix! The Home of Networking Devices and Data ...
85. Title Index
21. IM/NOMS 2004
9. Amazon.com: SNMP Network Management (McGraw-Hill Computer ...
30. Network Management Research Group
54. Software – $400 – $500 – CNET Reviews
61. Harris Corporation – Broadcast Communications
69. Software – $500 – $700 – ZDNet Reviews

(c) Concept Highlighter

Figure 1. Screen shots of the three interfaces evaluated in this study.

individual document surrogates to their information needs. Together, these coordinated views provide the users with the ability to interactively explore the search results both from an overview perspective and a detailed perspective. Within the same information display, the users can determine the general properties of the top 100 document surrogates, as well as the specific properties of individual document surrogates.

In preliminary evaluations, computer science graduate students were able to effectively use the exploration and evaluation features of HotMap and Concept Highlighter when conducting searches on topics related to their research. In some cases, the participants were impressed with their ability to identify documents of which they had previously been unaware. While these studies were valuable in terms of validating many of the design decisions, it was difficult to draw strong conclusions from these results since the participants each performed different searches. In this paper, we report the results of a more carefully designed user study, the details of which are described in the following section.

3. Experimental Design

3.1 Method

In order to evaluate and compare the effectiveness of the three interfaces for representing web search results, we employed a 3x2 (interface x task) within-subjects design [12]. Since all participants were already users of the Google search engine, they were exposed to this interface first. In order to reduce biasing effects, participants were assigned to one of four groups, each with a different order of exposure to HotMap and Concept Highlighter, and a different order of task assignment.

To ensure that the interfaces provided the same set of search results to each participant, the results of each web search were cached and provided to each interface. The Google interface was altered to include document numbers to facilitate data collection. Screen shots of the three interfaces are provided in Figure 1.

3.2 Procedure

Each participant completed a pre-task questionnaire, two searches with each of the three 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 a new 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 and how it was to be used in the search task.

Table 1. The relevance scores used to rate the document surrogates considered by the participants.

Score	Description
4	This document is relevant.
3	This document is probably relevant.
2	This document is probably not relevant.
1	This document is not relevant.

Each search task included a written description of the information need, along with the query to be used. These are listed below:

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"

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"

Before the first search was started with each new task, the participants were asked to rate their familiarity with the search task. 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 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,

Table 2. Features of the participant demographics.

Computer Use	10+ times per week: 95% 5-10 times per week: 0% 1-5 times per week: 5%
Computer Experience	high degree: 62% moderate degree: 38% low degree: 0%
Web Searches	10+ per week: 71% 5-10 per week: 19% 1-5 per week: 10%
Search Engine Preference	Google: 90% Yahoo: 10%
Pages Viewed	1-2 pages: 33% 3-4 pages: 29% 5-6 pages: 24% 7+ pages: 14%
Web Search Experience	high degree: 38% moderate degree: 62% low degree: 0%

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 among the interfaces to which they were exposed.

3.3 Analysis

The quantitative results (time to completion and perceived precision) 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 noted.

4. Results

4.1 Participant Demographics

21 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 2.

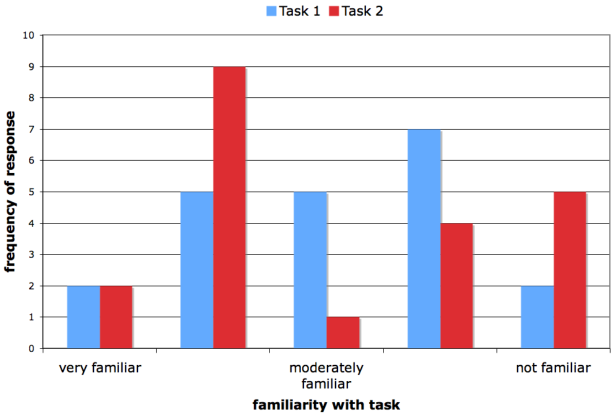


Figure 2. The pre-test familiarity with the tasks reported by the participants.

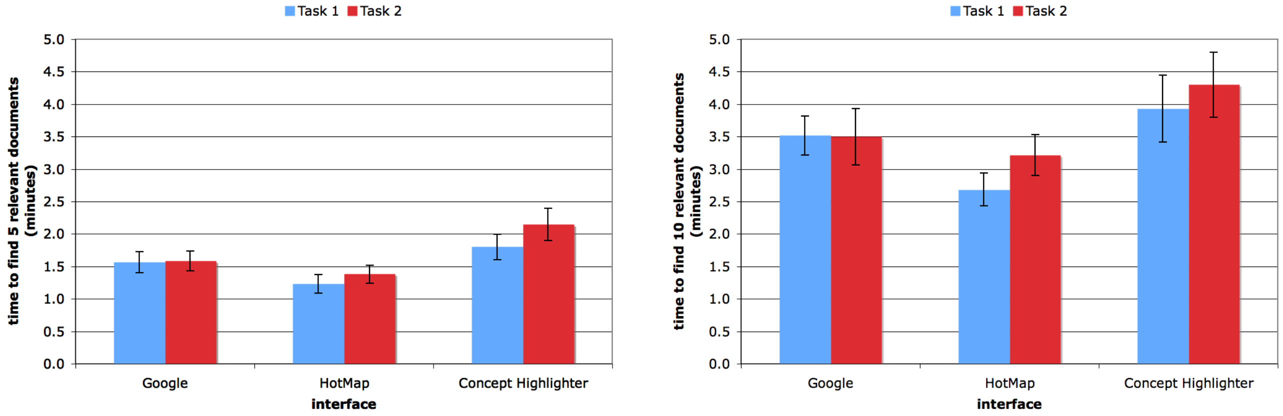
Compared to the participants in our pilot evaluations of HotMap [7], these participants would generally be considered intermediate web searchers. However, since none had been exposed to HotMap or Concept Highlighter prior to this evaluation, they would be considered novice users of these systems.

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 2. Clearly, the participants showed a wide range in familiarity with the search tasks. However, we did not find familiarity with the task to be a direct indicator of performance with any of the interfaces considered.

4.2 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, we specified two levels of fulfilment of the assigned information need: finding five relevant documents, and finding ten relevant documents. In this criteria, we considered any document assigned a relevance score of three or four as a relevant document. This represented documents the participants felt were either certainly or probably relevant (see Table 1 for the complete relevance score scale).

Figure 3 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 fastest with HotMap, followed by Google, then Concept Highlighter. However, the differences in these times did not prove to be statistically significant. The results of ANOVA tests are provided in Table 3.



(a)

(b)

Figure 3. 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 3. ANOVA tests for the time to completion data show that the differences between the task completion times are not statistically significant.

	5 Relevant Documents	10 Relevant Documents
Task 1	F(2,60) = 2.87, $p = 0.06$	F(2,60) = 2.87, $p = 0.06$
Task 2	F(2,60) = 4.69, $p = 0.01$	F(2,60) = 1.75, $p = 0.18$

These results are interesting since both HotMap and Concept Highlighter required the participants to interact with the interface features before they could evaluate the search results; by contrast, Google allowed the participants to start evaluating the search results immediately. That the times taken to find five and ten relevant documents are not significantly different among the three interface indicates that the extra work required by HotMap and Concept Highlighter can be offset by the users' ability to more easily evaluate and explore the search results.

Further, even though a large disparity existed in the experience the participants had with Google and the two new interfaces, the participants were able to complete the tasks using all the interfaces in similar times. One would expect that the task completion times for HotMap and Concept Highlighter would improve as the users become more experienced using these systems.

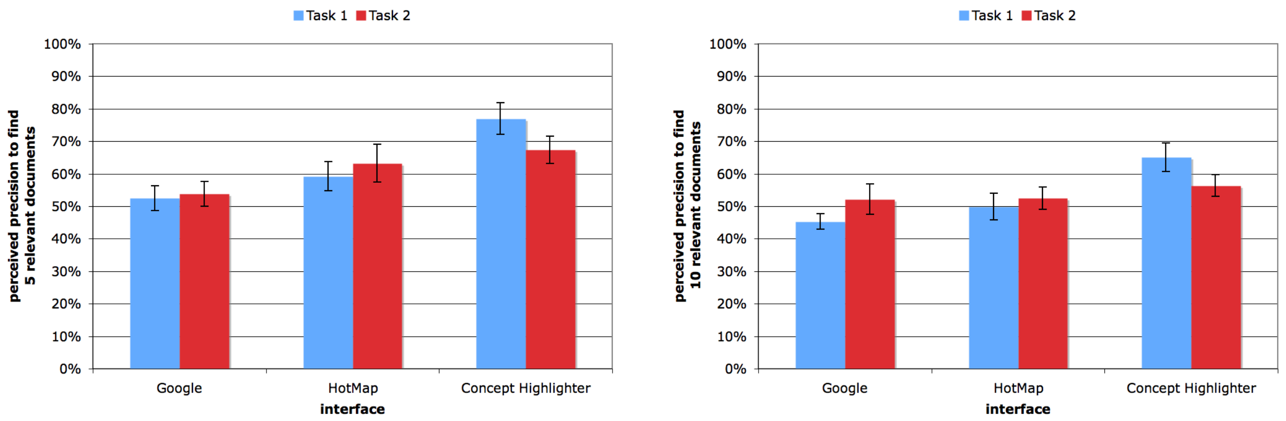
4.3 Perceived Precision

Two common measures for the performance of information retrieval systems 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 [15]. Since it is not feasible to calculate the recall metric for web information retrieval systems [9], we consider only the precision metric.

In traditional information retrieval research, test collections (such as those provided by TREC) are commonly used. These test collections usually consist of a set of documents, a set of queries, and expert relevance judgements for each document-query pair. When conducting research using live web information retrieval data, these relevance judgments are not available, making the use of the precision metric difficult.

Instead, we focus on the calculation of a metric that is inspired by the precision metric, which we call *perceived precision*. In this metric, we consider only the documents that have been viewed by the user as having been retrieved, and make use of the users' judgments of relevance. 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.

Suppose a four-point relevance scale is used, such as the one provided in Table 1. After a participant completes assigning relevance scores to a set of search results, the result is four sets of document surrogates corresponding to the four relevance scores: r_1, r_2, r_3, r_4. If we decide that documents with a relevance score of three or four are considered



(a)

(b)

Figure 4. The average perceived precision in finding five relevant documents (a) and ten relevant documents (b). The error bars represent the standard error about the mean.

Table 4. ANOVA tests for the perceived precision data show that the differences in the perceived precision for Task 1 are statistically significant, and that the differences in the perceived precision for Task 2 are not statistically significant.

	5 Relevant Documents	10 Relevant Documents
Task 1	F(2,60) = 8.22, $p < 0.001$	F(2,60) = 7.75, $p = 0.001$
Task 2	F(2,60) = 2.16, $p = 0.12$	F(2,60) = 0.37, $p = 0.69$

“relevant”, then perceived precision (pp) can be defined as:

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

Figure 4 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, followed by HotMap, then Google. For Task 1, the differences in the perceived 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 4.

The differences in statistical significance can be attributed to the 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 and Concept Highlighter to explore the search results from Task 1 than Task 2. While it is difficult to generalize the results from these two tasks, we can see that in some situations, there can be a significant improvement in the perceived precision.

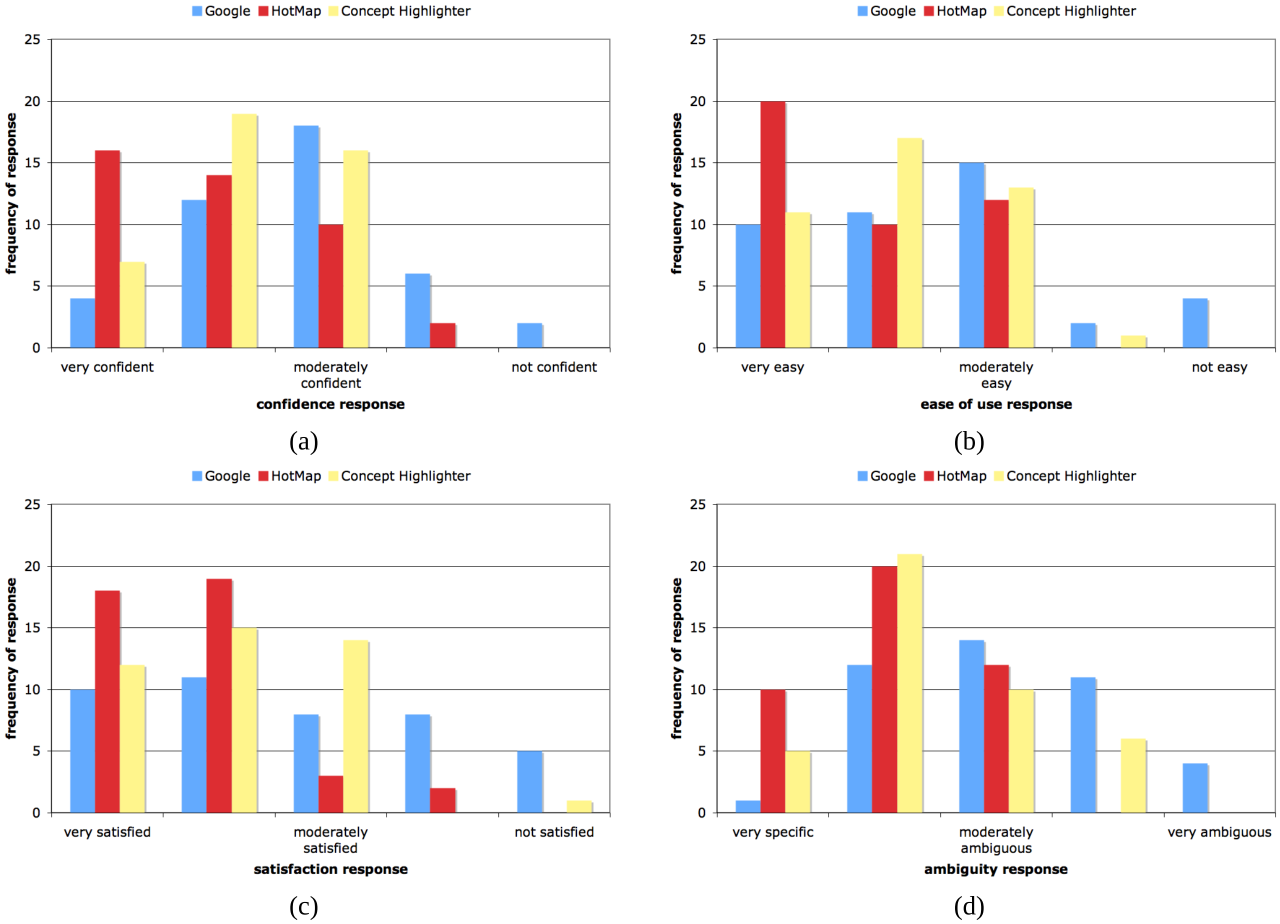
4.4 Subjective Measures

After each task, participants completed a short in-task questionnaire to measure their subjective reactions to using the assigned interface to complete 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.

For the *confidence* measure, the participants rated how confident they were in their ability to find a good set of relevant documents (Figure 5a). For all the interfaces, the responses were positively skewed; only 8% of the responses were negative. These results show that the participants were significantly more confident using HotMap than the other interfaces. Participants also showed a higher degree of confidence in Concept Highlighter than Google.

For the *ease of use* measure, the participants rated how easy they found it to use the interface to evaluate the search results (Figure 5b). For all the interfaces, the responses were positively skewed; only 6% of the responses were negative. Clearly, HotMap scored significantly higher in terms of ease of use than the other interfaces. Participants also scored Concept Highligher as marginally easier to use than Google.

For the *satisfaction* measure, the participants rated how satisfied they were in using the interface to evaluate the



(a)

(b)

(c)

(d)

Figure 5. Subjective measures of confidence in the search results (a), ease of use of the interface (b), satisfaction in the search process (c), and impressions of ambiguity in the search results (d).

search results (Figure 5c). For all the interfaces, the responses were positively skewed; 12% of the responses were negative (most of which were attributed to Google). HotMap received much higher scores in terms of confidence than the other interfaces. Concept Highlighter scored marginally higher than Google in this measure.

For the *ambiguity* measure, the participants rated how ambiguous they thought the search result set was (Figure 5d). Since the goal of the HotMap and Concept Highlighter 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. For HotMap and Concept Highlighter, the responses were positively skewed; 17% of the responses were negative (most of which were attributed to Google). The participants reported the search results to be much more specific using HotMap, followed by Concept Highlighter. The responses using Goole showed a normal distribution.

Table 5. Friedman tests for the subjective reactions show that the differences in the subjective reactions are statistically significant.

Measure	Friedman Test
confidence	$\chi^2(2) = 14.26$, $p = 0.001$
ease of use	$\chi^2(2) = 11.22$, $p = 0.004$
satisfaction	$\chi^2(2) = 14.22$, $p = 0.001$
ambiguity	$\chi^2(2) = 23.10$, $p < 0.001$

The results of Friedman tests on these responses showed them to be statistically significant. The statistics are reported in Table 5.

These positive subjective measures were also validated by the comments made by many of the participants. One participant commented that they “loved the ability to sort by keywords.” Another noted that “Concept Highlighter”

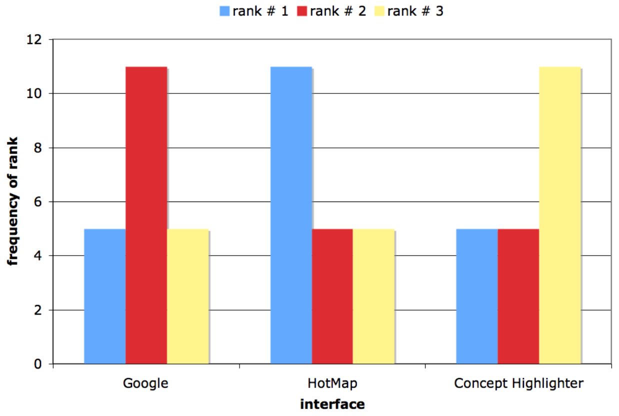


Figure 6. Interface preference ranks reported by the participants.

made the results more specific.” Although most comments were positive, some noted that the interface was too cluttered, or the font sizes to be too small.

4.5 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 Figure 6.

For Google, there was a normal distribution of rank scores, with the majority scoring the interface as their second preference. For HotMap, there was a positively skewed distribution of rank scores, with the majority scoring the interface as their first preference. For Concept Highlighter, there was a negatively skewed distribution of rank scores, with the majority scoring the interface as their third preference.

Clearly, there were a larger number of participants that indicated HotMap as their first preference. Both Concept Highlighter and Google had the same number of participants indicated these as their first preference. However, a large number of participants also indicated Concept Highlighter as their least preferable interface.

A pair-wise analysis of the results using a Wilcoxon signed ranks test showed that HotMap was preferable to both Google and Concept Highlighter, and that Google was preferable to Concept Highlighter. However, these results were not shown to be statistically significant (see Table 6). A number of participants noted that they ranked Google first because they were familiar with it, resulting in a skewing of the results in favour of Google.

Table 6. Wilcoxon signed ranks tests for the preference ranks show that the preference order is not statistically significant.

Comparison	Wilcoxon Signed Ranks Test
HotMap preferable to Google	Z = -0.974, $p = 0.33$
HotMap preferable to Concept Highlighter	Z = -1.630, $p = 0.10$
Google preferable to Concept Highlighter	Z = -0.974, $p = 0.33$

5. Discussion

The search tasks used in this evaluation were chosen to be somewhat vague in order to test situations where the search results consist of a mix of relevant and non-relevant documents. We believe this mimics many real-world situations where the users are unable to craft very specific queries to describe their information needs. In these cases, deciding which documents are relevant is a fundamental task that is not well supported by the list-based representation provided by Google and other search engines. However, providing visual representations of features of the web search results, along with tools for interactively exploring the search results, can be very beneficial to the users.

In these studies, HotMap proved to be both faster than Google, and also allowed the participants to find the relevant documents more efficiently (as indicated by the perceived precision metric). Although this performance increase over Google was not statistically significant, it is a promising result given that the participants were experienced users of Google but only novice users of HotMap.

HotMap scored significantly higher than the other interfaces in the subjective reactions, and was ranked as the top search results interface by the participants. This suggests that the simple query term visualization method implemented in this system can not only allow users to more effectively evaluate the search results, but it is also an interface that can easily be used and adopted by web searchers.

Concept Highlighter was the slowest interface tested, but at the same time resulted in the best perceived precision. Part of the reason for the poor time to completion results with this interface was the extra time required by the participants to choose the relevant concepts. This task was required before the participants could start evaluating the search results. That the perceived precision score was better with this interface indicates that this extra step is valuable in re-sorting the search results into a more meaningful (i.e., concept-oriented) order.

In terms of the subjective reactions, Concept Highlighter scored worse than HotMap, but better than Google. However, this interface was ranked as the least preferable by the participants. This may be due to the extra work required to first choose the relevant concepts. Some participants may have disliked the delay this caused in evaluating the search results, as well as the extra cognitive activity required in making these decisions. However, given the significant improvement in perceived precision in some cases, these negative aspects may be overcome with adequate training and familiarity with the interface.

It is interesting to note that the re-sorting features in both HotMap and Concept Highlighter bring attention to document surrogates that are deep within the search results. Given that users of traditional web search engines seldom venture past the third page of search results, these deep document surrogates are seldom considered for relevance. As a result, in our preliminary evaluations, a number of our colleagues were able to find documents relevant to their research areas that has previously been undetected.

6. Conclusion and Future Work

In this study, two new interfaces that promote a visual exploration of web search results were compared to the list-based representation used by Google. The participants performed the same search tasks on all three interfaces, and the search results sets presented in the interfaces were identical. Therefore, the only differences were the method by which the search results were presented to the participant, and the ability for the participant to interact with and explore the search results.

The search tasks were chosen to be somewhat vague in order to evaluate the differences in these interfaces when the search results contained a mixture of relevant and non-relevant documents. HotMap scored higher than the other two interfaces in all the measurements except for perceived precision (in which it scored the second highest). In the perceived precision measure, Concept Highlighter scored the highest, although this required additional work on the part of the participants, resulting in poorer scores in the other measurements.

The positive results of this study have validated our hypotheses that visual representations of search results features, and the visual exploration of the search results can address the inherent problems in the list-based representation used by all the major web search engines. In future work, we will integrate the features of HotMap and Concept Highlighter together into a single interface for web search results exploration, and tie this in with our previous work on interactive query refinement, resulting in a unified web information retrieval support system. Further, we plan to evaluate this system in a long-term test in order to determine how effective this system can be for experienced users.

References

[1] K. Andrews, C. Gutl, J. Moser, and V. Sabol. Search result visualization with xFind. In *Proceedings of the Second International Workshop on User Interfaces to Data Intensive Systems*, 2001.

[2] E. Berenci, C. Carpineto, V. Giannini, and S. Mizzaro. Effectiveness of keyword-based display and selection of retrieval results for interactive searches. *International Journal on Digital Libraries*, 3(3), 2000.

[3] S. Eick. Graphically displaying text. *Journal of Computational Graphics and Statistics*, 3(2), 1994.

[4] Grokker. Grokker search engine. http://www.grokker.com/.

[5] M. 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.

[6] O. Hoeber and X. D. Yang. A model for interactive web information retrieval. In *Proceedings of the International Symposium on Smart Graphics*, 2006.

[7] O. Hoeber and X. D. Yang. The visual exploration of web search results using HotMap. In *Proceedings of the International Conference on Information Visualization*, 2006.

[8] O. Hoeber and X. D. Yang. Visually exploring concept-based fuzzy clusters in web search results. In *Proceedings of the Atlantic Web Intelligence Conference*, 2006.

[9] M. Kobayashi and K. Takeda. Information retrieval on the web. *ACM Computing Surveys*, 32(2), 2000.

[10] T. Paek, S. Dumais, and R. Logan. Wavelens: A new view onto internet search results. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2004.

[11] E. 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.

[12] M. B. Rosson and J. M. Carroll. *Usability Engineering: scenario-based development of human-computer interaction*. Morgan Kaufmann, 2002.

[13] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1), 1999.

[14] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 2001.

[15] E. M. Voorhees. Overview of TREC 2004. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

[16] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of IEEE Information Visualization*, 1995.

[17] Y. Yao. Information retrieval support systems. In *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, 2002.