Evaluating the Effectiveness of Term Frequency Histograms for Supporting Interactive Web Search Tasks

Orland Hoeber
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada
hoeber@cs.mun.ca

Xue Dong Yang
Department of Computer Science
University of Regina
Regina, SK, Canada
yang@cs.uregina.ca

ABSTRACT

Throughout many of the different types of Web searches people perform, the primary tasks are to first craft a query that effectively captures their information needs, and then evaluate the search results seeking relevant documents. However, the top Web search engines generally provide little support for users in these tasks. WordBars is a next-generation Web search interface that provides an interactive histogram representation of the most frequently appearing terms within the titles and snippets of the top 100 search results. In this paper, the results of a user study are presented in which the ability of the participants to find relevant documents using the features of WordBars is measured. Most participants were able to find more relevant documents using WordBars when compared to the original order of the search results. Subjective reactions were very positive, with all the participants rating the interactive features of WordBars highly.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Query formulation, Information filtering, Human factors, Evaluation

Keywords

web search, query refinement, search results exploration, visualization, interaction

1. INTRODUCTION

Studies on Web search user behaviour have reported that queries commonly consist of only one to three terms [9, 21]. These short queries indicate that users of Web search engines often have difficulties crafting queries that accurately reflect their information needs. Little support is provided for this task; it is up to the user to decide which terms to enter into the query, and to manually add or remove terms from their query as needed.

Even when Web searchers are able to effectively craft a query, few consider more than three pages of search results [20, 21]. Although users may be able to find the information they are seeking within the first few pages of the search results, difficulties arise when they are unable to do so. Studies have found that there is a tendency for searchers to avoid conducting an in-depth evaluation of the search results [21]. The static list-based representations of the search results require users to consider each document individually, and to some degree in the order provided. Most Web search engines provide little ability to manipulate or explore the search results further.

We have previously proposed that the traditional model for Web search be extended to include cycles of interactive query refinement and interactive search results exploration [6]. WordBars [5] represents a realization of this interactive Web search model. A term frequency histogram generated from the titles and snippets of the top search results supports users in both building a better query and exploring the search results. In this paper, the results of user studies with WordBars are reported, showing the effectiveness of using visualization and interaction to support Web search tasks.

A fundamental design principle in the development of WordBars was to strike a balance between computer automation and human control of the task [19]. Crafting a query that accurately represents a user's information needs is an inherently human task, as is the evaluation of the search results. While some have suggested methods for automatic query expansion [26, 15, 24], we believe human decision-making in query refinement is vitally important. Similarly, most Web search engines provide automated ranking of the search results based on complex and proprietary algorithms, such as PageRank [1] or HITS [11]. However, these algorithms result in static ordered lists 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.

The remainder of this paper is dedicated to a discussion of related work, an overview of the WordBars system, an outline of the experimental design, and the results from the user evaluations. The paper concludes with a discussion on the features and outcomes of the user evaluation and an overview of future work.

DIS 2008 Cape Town, South Africa

2. RELATED WORK

2.1 Interactive Query Refinement

Providing lists of terms from which a user can select for query expansion is not a new idea. Harman [3] generated three different lists from which the user could choose additional terms to add to their query. The results reported from this work were good when users made perfect choices from the lists of available terms. However, others have shown that when presented with a list of potentially relevant terms, users may have difficulties choosing good terms to add to their queries [18, 13].

Others have investigated alternate methods for presenting terms to users for query refinement purposes. For example, Joho et al. [10] generated a hierarchy of query expansion terms from the set of retrieved documents, and presented these to the user via cascading menus. Although there is value in deducing and representing the relationships among terms within the search results set, their process requires access to the contents of the entire documents within the search results, which is not feasible for interactive Web search systems.

In our work on WordBars, the potential query refinement terms are selected from the top search results returned by the underlying Web search engine. However, rather than collecting the actual document contents, the frequency statistics are based only on the title and snippet provided by the underlying search engine. 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.

Further, WordBars provides 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 users to see how potential query expansion terms are used in 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 a list of terms.

2.2 Search Results Re-Sorting

The re-sorting of search results based on Web search personalization is a rather active research field [22, 23, 16]. These systems generally provide an automated re-sorting and filtering of the search results based on the personalized profiles of the users. However, there appears to be little research on interactive tools to allow the users to control the re-sorting methods, in personalized systems or otherwise.

In our work on HotMap [7], users were able to re-sort the search results based on the frequencies of the query terms within the search results. In Concept Highlighter [8], the search results were interactively re-sorted based on fuzzy membership scores with respect to user-selected concepts. In both of these systems, the re-sorting features helped to bring highly relevant documents that were buried deep in the search results to the attention of the searcher. In user studies comparing these systems to Google, we found that the interactive search results exploration features can increase user performance in finding relevant documents [4].

3. WORDBARS

The primary goals in the design of WordBars (see Figure 1) were to support users in their tasks of query refinement and search results exploration through interactive and visual features. The source of information in providing this support are the top 100 document surrogates in the search results generated for the current user query.

As the search results are obtained from the Google API [2], the frequencies of each unique term are counted within the title and snippet of each document surrogate. Common terms, as well as terms that are less than three characters long, are ignored. Porter's stemming algorithm [14] is used to reduce terms to their root forms for matching purposes, resulting in more effective frequency statistics than if exact word matches were used.

In choosing a method for visually representing this term frequency information, the goal was to provide a simple representation that allows users to browse the available terms, as well as perceive and interpret the relative frequencies of these terms in the top search results. Studies of query expansion systems have shown that users have difficulties selecting relevant terms from simple lists [18, 13]. In this work, a vertically-oriented, colour-coded histogram is used (a zoomed-in view of this histogram is shown in Figure 1). This histogram is both easy to browse, and provides additional information to the user regarding term use within the search results.

The sizes of the bars in the histogram, as well as the intensities of the colours, are used to represent the frequencies of the top 40 terms found in the search results. Using multiple visual features to represent the same data attribute provides redundant coding, and can result in an increase in the ease, speed, and accuracy in which the the users are able to perceive and interpret the information [17]. The colour scale was chosen to vary on the red-green colour channel as well as the luminance channel. Visually, this colour scale appears to be a heat scale, resulting in high frequency terms appearing hot, and low frequency terms appearing neutral or warm.

The term labels are provided to the right of each frequency bar. All the terms that are present in the query use a red font colour; all others are black. These colours allow 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, the colour distinctions can be pre-attentively processed [25], allowing the near-instant recognition of the distinction between the query terms and the other terms, as well as the relative differences in frequencies.

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 frequency histogram for the terms used in the top search results.

The methods by which users can interact with the histogram were chosen to be as simple as possible in order to reduce the learning curve associated with using the system. Single-clicks are used to select terms, which result in a re-sorting of the search results based on the term frequencies in the title and snippet (i.e., the document surrogates are sorted based on the total frequency of occurrence of the selected terms). A visual indication of which terms are selected is provided via a simple grey box placed around the

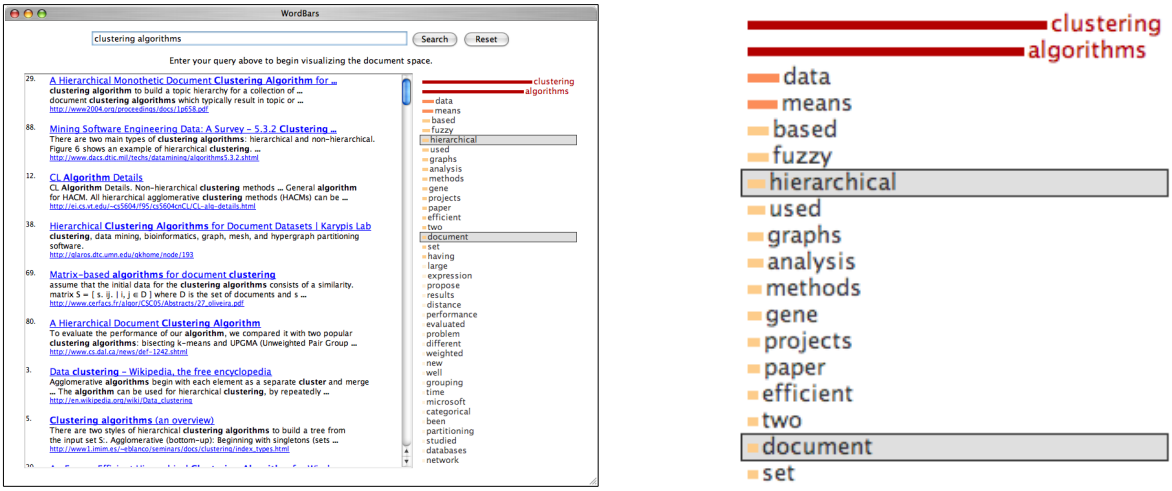


Figure 1: A screenshot of the WordBars system with a zoomed-in view of the top of the histogram. Note that the search results are re-sorted based on the terms selected in the histogram.

selected terms. Double-clicks are used to add or remove terms from the query. Clicking the search button sends the refined query to the Google API, producing a new set of search results and a new histogram of the term frequencies.

Within the document list window, the search results are displayed in a list-based representation that is similar to that used by the major search engines. The document number from the original order of the search results is included to highlight the effects of the re-sorting features. Clicking on any document title will open that document in a new window, and will change the link colour from blue to purple (as per the de-facto standard for visited links in a Web page). This allows the users to easily identify documents that have already been visited, even after the search results are subsequently re-sorted by the user.

More details on the WordBars system, along with two complete examples highlighting the effectiveness of WordBars both for vague and specific initial queries, are provided in [5]. A video of the WordBars system in action is provided on the author's Web site[1].

The main drawback to the techniques used in WordBars is that the ability to support the users in their information seeking tasks depends greatly upon the quality of the initial query. If a very poor query is provided, there will be few relevant documents returned. Exploring these documents will be of little value, and the common terms used in their titles and snippets will likely not be relevant to the user's information need. However, this lack of relevant terms in the histogram may provide an indication that the user needs to start with a better initial query.

When the initial query produces a set of search results in which at least some documents are relevant, WordBars can be very beneficial in assisting the users in their Web search tasks. The term frequency histogram provides a visual indication of the relative frequencies of the terms used in the top document surrogates from the search results. Users may interactively re-sort the search results based on the terms that are most relevant to their information need, and subsequently add these terms to their query to generate a more specific set of search results.

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

4. EXPERIMENTAL DESIGN

4.1 Method

In order to evaluate and compare the effectiveness of the features of WordBars, a study was conducted that mimics the procedure of using WordBars to first explore the search results, then attempt to build a better query, and then explore the search results further. Three understandable yet somewhat ambiguous tasks were selected and assigned to the participants in a pseudo-random order. To ensure that the system provided the same set of search results to each participant, the results of each initial search task were cached.

4.2 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. The procedures for the user study were described using this training task as an example; the participants were permitted to use and experiment with the system. The explanation and description of the research procedures took approximately 10 minutes.

Prior to beginning each target search tasks, the investigator answered any questions the participants had about the task itself. For each task, the participants were asked to perform the following steps:

1. Provide relevance scores for the top ten documents for the initial query in the order provided by the Google API using a four-point relevance scale (see Table 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

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

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.

relevant documents to the top of the list), and provide relevance scores for the top ten documents.

After completing 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 the target search tasks were completed, a post-task questionnaire was administered that included a question asking the participant to rank the features of WordBars based on their preferences. The entire procedure took approximately 60 minutes for each participant.

4.3 Tasks

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[2] test topics, and were intentionally chosen to be difficult tasks, yet understandable by a wide range of participants. It was also ensured that these search tasks were somewhat vague so that there was room for improvement by using WordBars, but not so vague as to result in very few relevant documents being returned by the underlying search engine. Although these tasks may not be representative of what people actually do when searching the Web, they do illustrate areas where current Web search engines perform poorly, and where visualization and interaction may be beneficial.

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.
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 the four-point relevance scale (see Table 1). This information was logged by the investigator, along with the time taken to complete each stage of the task. The participants were asked to only consider the document surrogates within the search results list for relevance. That a non-relevant document may appear to be relevant to a searcher when considering only the title and snippet was beyond the scope of this study.

[2]`http://trec.nist.gov/data/t14_hard.html`

Table 2: Demographic features of the participant sample.

Computer Use	10+ times per week:	100%
Computer Experience	high degree:	100%
Web Searches	10+ per week:	84%
	5-10 per week:	8%
	1-5 per week:	8%
Search Engine Preference	Google:	100%
Likelihood of Adding Terms to a Query	always:	33%
	often:	55%
	sometimes:	17%
	seldom:	0%
Likelihood of Removing Terms from a Query	always:	0%
	often:	33%
	sometimes:	50%
	seldom:	17%
Web Search Experience	high degree:	58%
	moderate degree:	42%
	low degree:	0%

5. RESULTS

5.1 Participant Demographics

Twelve computer science graduate students were recruited to participate in this study. The results from the pre-task questionnaire administered to these participants are presented in Table 2. Clearly, the participants can be considered expert users. Although this is not an accurate sample of the entire population of Web searchers, it does represent an accurate sample of "power users".

5.2 Relevance of Search Results

Although it is common to use variants of the precision metric for evaluating information retrieval systems [12], these metrics generally assume the existence of expert relevance scores for the documents in the collection being searched. Obtaining such expert scores for live Web search results is problematic, especially when the participants are permitted to modify and refine their queries.

Instead, the effectiveness of the features of WordBars are analyzed based only on the relevance scores provided by the participants. An assumption is made that all the participants provided accurate relevance judgments while using the system. In the course of the study, each participant believed they were making accurate relevance judgement choices to the best of their abilities; 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 evaluate each participant's performance in finding relevant documents, a rank comparison method was used. For each of the four sets of documents considered for relevance (original order, first re-sort, refined query, and second re-sort), the number of documents which received a score of three or four ("probably relevant" and "relevant") was counted. Each participant's relative performance was

then ranked, the idea of which is to show where the participants found the most, second most, third most, and least relevant documents (i.e., ranks 1, 2, 3, and 4, respectively). The results of this analysis are provided in Figure 2, grouped by task.

This ranked comparison method illustrates which features of WordBars were most effective in finding relevant documents. Although it does not convey how much more effective one feature is than another, neither does it allow the exceptional performance of one participant to bias the aggregate performance results.

For Task A (Figure 2a), the best results were achieved by re-sorting the search results from the original query, followed closely by the re-sorting of the search results from the refined query. Many of the participants appeared to have difficulties constructing a better query that accurately captured the assigned information need.

For Task B (Figure 2b), the participants performed marginally better than the original order of the search results by performing the re-sorting step. 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.

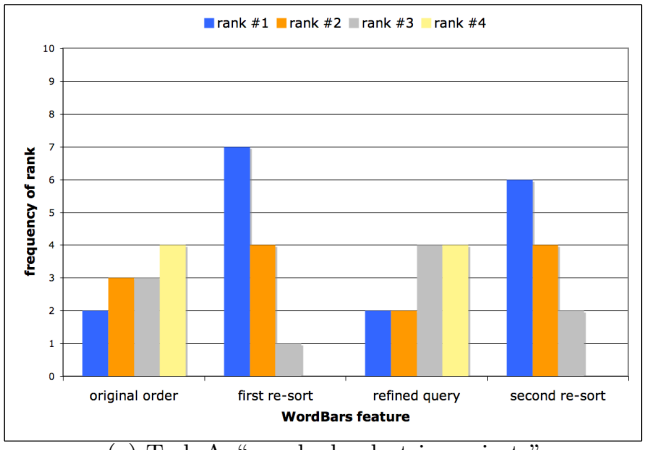
For Task C (Figure 2c), re-sorting the search results resulted in better performance than the original order, but not substantially so. Refining the query for this task produced the best results. Re-sorting the search results of this refined query resulted in a decrease in the performance compared to the refined query, but still a substantial increase over the original order. In analyzing this reduction in performance, it was found that a number of participants identified all 10 documents as relevant in the refined query search results, leaving no room for improvement from the re-sorting features.

Note that even though these results are very positive, two participants in Task A and one participant in Task C found as many or more relevant documents in the original order of the search results than when using the features of WordBars. This illustrates the fact that when users make poor choices using the features of WordBars, the results may be worse than than if the features were not used at all. These poor choices may have been due to a lack of understanding of the topic, a misunderstanding of the meaning or relevance of a specific term in the WordBars histogram, or errors in using the interface.

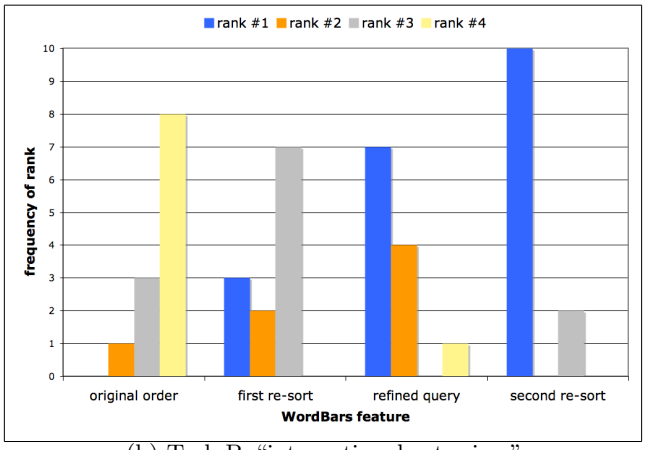
The results of pair-wise Wilcoxon signed rank tests are provided in Table 3. The increased performance of the participants over the original order proved to be statistically significant for all the features of WordBars and all tasks, except for the query refinement in Task A. For Task A, the second re-sorting of the search results was significantly better than the refined query, whereas this was not significant for the other tasks. This suggests that the re-sorting features are most effective when a poor query is used; when the query is improved sufficiently, the ability to improve upon the order of the search results is diminished. In general, these results 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.

5.3 Subjective Measures

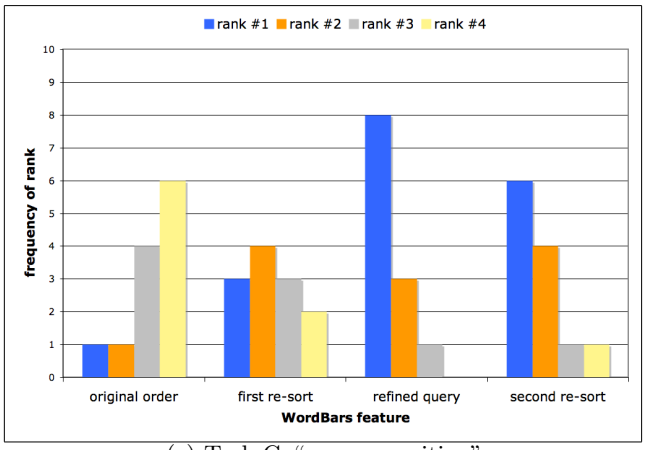
After providing the relevance scores for the top ten doc-



(a) Task A: "new hydroelectric projects"



(b) Task B: "international art crime"



(c) Task C: "euro opposition"

Figure 2: 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. Note that ties in the number of relevant documents found result in ties in the ranking.

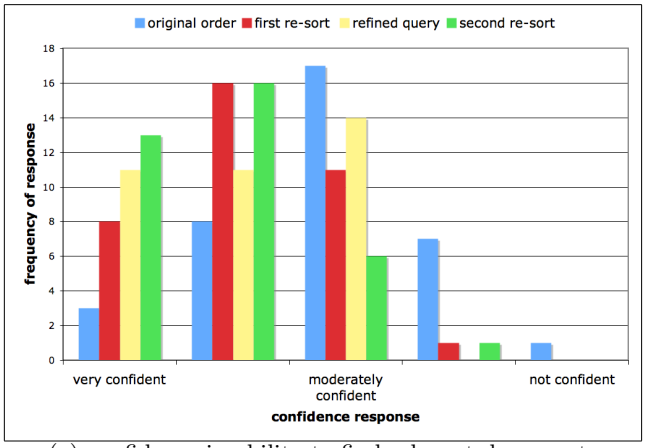
Table 3: Pair-wise Wilcoxon signed rank tests indicate the statistical significance of the features of WordBars.

Task A WordBars Feature Comparison	Pair-wise Wilcoxon Signed Rank Test
first re-sort performed better than original order	$\mathbf{Z = -2.46, p = 0.01}$
refined query performed better than original order	$Z = -0.58, p = 0.80$
second re-sort performed better than original order	$\mathbf{Z = -2.27, p = 0.02}$
second re-sort performed better than refined query	$\mathbf{Z = -2.12, p = 0.03}$
Task B **WordBars Feature Comparison**	**Pair-wise Wilcoxon** **Signed Rank Test**
first re-sort performed better than original order	$\mathbf{Z = -2.76, p = 0.01}$
refined query performed better than original order	$\mathbf{Z = -2.81, p = 0.01}$
second re-sort performed better than original order	$\mathbf{Z = -3.02, p < 0.01}$
second re-sort performed better than refined query	$Z = -0.65, p = 0.52$
Task C **WordBars Feature Comparison**	**Pair-wise Wilcoxon** **Signed Rank Test**
first re-sort performed better than original order	$\mathbf{Z = -1.98, p = 0.05}$
refined query performed better than original order	$\mathbf{Z = -2.84, p < 0.01}$
second re-sort performed better than original order	$\mathbf{Z = -2.26, p = 0.02}$
refined query performed better than second re-sort	$Z = -0.81, p = 0.42$

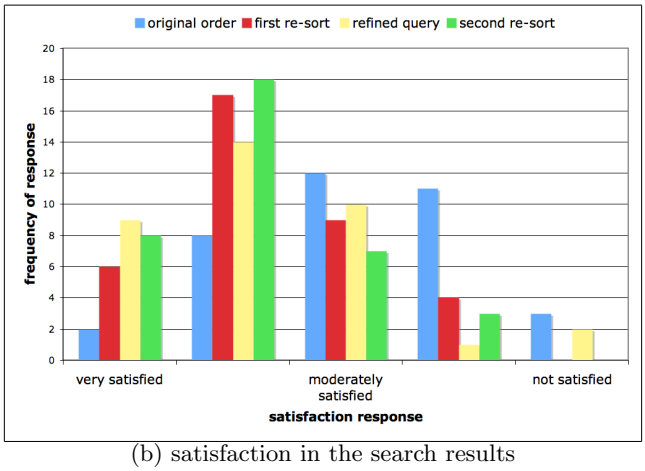
uments in each sub-task, participants completed a short in-task questionnaire to measure their subjective reactions to using the features of WordBars to find relevant documents for the assigned task. Of interest were the participants' degree of confidence in completing the task, satisfaction with the search results considered, and perceptions of ambiguity among the search results set.

For the *confidence* measure, the participants rated how confident they were in their ability to find a good set of relevant documents (Figure 3a). Whereas the confidence in the search results from the original order followed a normal distribution, the confidence from re-sorting 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 was reported for re-sorting the refined query results. While some participants reported a high degree of confidence in their refined query, many were only moderately confident. Many of these moderately confident responses were from Task A, where participants had difficulties crafting a better query.

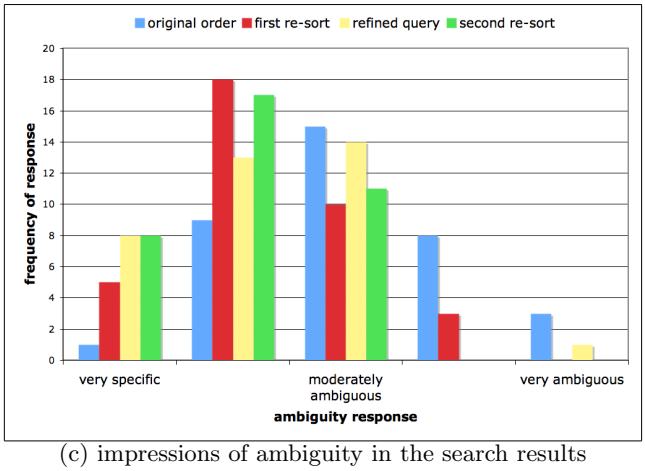
For the *satisfaction* measure, the participants rated how satisfied they were with the documents considered in the search results set (Figure 3b). Like the confidence measure, the satisfaction measure for the original order followed a normal distribution. The satisfaction measure for the three sets of search results generated from the WordBars features were



(a) confidence in ability to find relevant documents



(b) satisfaction in the search results



(c) impressions of ambiguity in the search results

Figure 3: 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.

Table 4: Friedman tests for the subjective reactions show that the differences in this data are statistically significant.

Measure	Friedman Test
confidence	$\chi^2(3) = 36.12, p < 0.001$
satisfaction	$\chi^2(3) = 29.25, p < 0.001$
ambiguity	$\chi^2(3) = 36.17, p < 0.001$

all positively skewed. While the differences in this measure among the features of WordBars was marginal; they all resulted in a higher level of satisfaction among the users than the original order of the search results.

For the *ambiguity* measure, the participants rated how ambiguous they thought the search result set was (Figure 3c). Again, the responses for the original order of the search results followed a normal distribution. For both the first re-sorting of the original search results, and the second re-sorting of the refined query search results, the ambiguity measurements were positively skewed. This indicates that many participants found the search results to be more specific when they were re-sorted. For the query refinement, the ambiguity measure showed a normal distribution with a positive skew, which illustrates that some participants were able to perform better than the original order, whereas others were not (mostly those from Task A).

The results of Friedman tests on these responses showed them to be statistically significant. These statistics are reported in Table 4.

5.4 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 the search results considered. The options were (a) the 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 4.

The original order of the search results was almost unanimously ranked last. One participant ranked the original order as the second best, and one ranked it as the third best. This result provides a clear indication that all the participants found value in the interactive query refinement and interactive search results exploration features of WordBars.

For the three sets of search results generated using the features of WordBars, the rank depended greatly upon the participants' abilities to refine their 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.

A pair-wise analysis of the results using Wilcoxon signed ranks tests showed that all the features of WordBars are

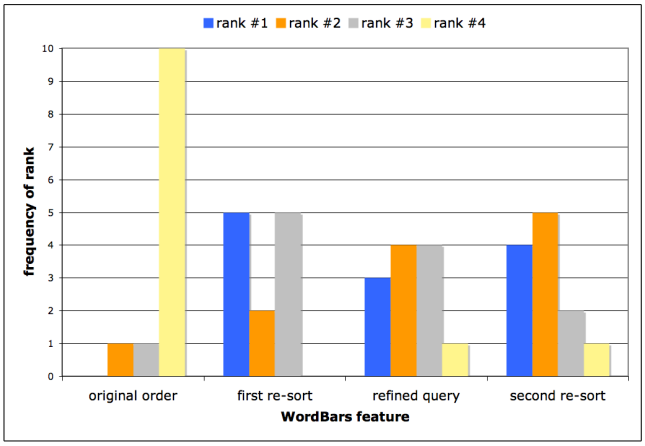


Figure 4: Preferences ranks reported by the participants for the features of WordBars. Note that the re-sorting and query refinement features were almost unanimously ranked higher (i.e., more preferable) than the original order of the search results.

Table 5: Pair-wise Wilcoxon signed ranks tests for the preference ranks show the statistical significance of the participants' preferences for the features of WordBars.

Comparison	Pair-wise Wilcoxon Signed Rank Test
first re-sort preferable to original order	$\mathbf{Z = -3.11, p < 0.01}$
refined query preferable to original order	$\mathbf{Z = -2.50, p = 0.01}$
second re-sort preferable to original order	$\mathbf{Z = -2.78, p = 0.01}$
first re-sort preferable to refined query	$Z = -0.56, p = 0.58$
second re-sort preferable to refined query	$Z = -0.58, p = 0.56$
second re-sort preferable to first re-sort	$Z = -0.12, p = 0.90$

preferable to the original order of the search results, with statistical significance (see Table 5). However, there was no clear preference between the features of WordBars themselves.

Other data was collected in the course of this study, such as the participants' prior knowledge on the topics, and the time taken to complete each task. However, this data was highly variable and was not correlated to system performance.

6. DISCUSSION

There are three features of this user study that are worth discussing in further detail: the tasks, the participants, and the relevance judgments. The three search tasks used in this study were intentionally chosen to be both easy to understand, yet somewhat ambiguous. By choosing easy to understand tasks, prior knowledge or experience in the task domain was not necessary, and had minimal influence on the participants abilities to decide document relevance. 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 ability for improvement for interactive Web search systems. 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 we have seen in this study.

The participants chosen for this user study can all be considered expert computer users, as well as intermediate to expert Web searchers. While this sample is not an accurate representation of the entire population of Web searchers, it is a sub-population (i.e., expert or power users) that can benefit from the interactivity and visualization features of WordBars to support their Web search tasks. Although our results are valid for these expert users, it is possible that we will get different results with novice to intermediate computer users.

Since live Web search results were used in this study, expert evaluations of the document surrogates were not available. As such, the performance of the participants in terms of finding relevant documents was based solely on the relevance scores provided by the participants. Our assumption that the participants were able to give accurate relevance judgments ignored the situations where the title, snippet, and URL for a document were misleading, or where the participant incorrectly interpreted the information provided. The potential for these errors can be ignored since it was possible for the participants to make them during all stages of the study. Therefore, from the perspective of the user, the assumption of accurate relevance judgments does not invalidate the results reported in this study. However, evaluating whether the users are able to accurately make relevance judgments 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, we believe that 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, then refining the query and evaluating the top ten documents, 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 take advantage of the interactive nature of WordBars; they 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. Therefore, in future research, we wish to study the benefits of interactive query refinement and interactive search results exploration under real-world Web search conditions.

A particular focus of this future work will be to study the potential utility of the visual and interactive features of WordBars for different types of searching. The study reported in this paper was designed such that each participant conducted the same set of ambiguous search tasks. In real-world use, we will be able to study a much broader range of search tasks, such as exploratory searching, targeted searching, re-finding of perviously seen information, and opportunistic searching. Future studies will also include controlled experiments that provide a more direct comparison to the top search engines, and an evaluation of methods other than term frequency for generating lists of potentially relevant terms from the top search results. A comparison between the information generated using the title and snippet versus the entire textual contents of the document will be of value, as will more extensive evaluations from an information retrieval perspective using document collections such as those provided by TREC.

7. CONCLUSIONS

The user evaluation reported in this paper illustrates the potential benefits that a visual and interactive interface to Web search such as WordBars can provide to searchers. In some cases, the participants found more relevant documents as a result of refining their queries; in others, the participants performed better by re-sorting the search results. The subjective reactions were very positive, and the participants unanimously ranked the use of the features of WordBars as superior to evaluating the search results in the original order. One of the key benefits of WordBars is the support it provides to both crafting a query and exploring the search results, within a single user interface.

The results of this study provide strong evidence in support of the fundamental hypothesis in the design of WordBars: that frequently used terms in the results of an initial search can provide valuable information to the user, both for crafting a better query as well as for re-sorting and exploring the search results. WordBars represents an example of what we believe will be the next generation of Web search interfaces: tools that focus on supporting the fundamental Web search tasks through interactive query refinement and interactive search results exploration.

The design of other interactive Web search interfaces may also benefit from the results of this study. It was shown that in many cases, participants were able to effectively make interactive selections for the term frequency histogram to improve their search performance. The visual representation of this information is an effective design feature in supporting users in interactive query refinement and interactive search results exploration tasks. The continued investigation of visual representations of information to support interactive Web search tasks will contribute greatly to the design of next-generation Web search interfaces.

8. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[2] Google. Google Web API. http://www.google.com/apis/.

[3] D. Harman. Towards interactive query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988.

[4] O. Hoeber and X. D. 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.

[5] O. Hoeber and X. D. Yang. Interactive Web information retrieval using WordBars. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[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] B. J. Jansen and U. 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.

[10] H. Joho, C. Coverson, M. Sanderson, and M. Beaulieu. Hierarchical presentation of expansion terms. In *Proceedings of the ACM Symposium on Applied Computing*, 2002.

[11] J. M. Kleinberg. Authoritative sources in an hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[12] M. Kobayashi and K. Takeda. Information retrieval on the Web. *ACM Computing Surveys*, 32(2):114–173, 2000.

[13] M. Magennis and C. 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.

[14] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[15] Y. Qiu and H. P. Frei. Concept based query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

[16] F. Radlinski and S. Dumais. Improving personalized Web search using result diversification. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

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

[18] I. 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.

[19] B. Shneiderman. *Designing the User Interface*. Addison-Wesley, 1998.

[20] 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.

[21] 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):226–234, 2001.

[22] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive Web search based on user profile construction without any effort from users. In *Proceedings of the International World Wide Web Conference*, 2004.

[23] J. Teevan, S. Dumais, and E. 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.

[24] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[25] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2004.

[26] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.