Interactive Web Information Retrieval Using WordBars

Orland Hoeber and Xue Dong Yang

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

Abstract

It is common for web searchers to have difficulties crafting queries to fulfill their information needs. Even when they provide a good query, users often find it challenging to evaluate the results of their web searches. Sources of these problems include the lack of support for query refinement, and the static nature of the list-based representations of web search results. To address these issues, we have developed WordBars, an interactive tool for web information retrieval. WordBars visually represents the frequencies of the terms found in the first 100 document surrogates returned from the initial query. This system allows the users to interactively re-sort the search results based on the frequencies of the selected terms within the document surrogates, as well as to add and remove terms from the query, generating a new set of search results. Examples illustrate how WordBars can provide valuable support for query refinement and search results exploration, both when specific and vague initial queries are provided.

1. Introduction

Studies of web search user behaviour have shown that a large portion of web search queries consist of only one to three terms [10, 23]. 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 to refine their queries; it is up to the user to manually add or remove query terms. As a result of this lack of support, web searchers seldom make subsequent modifications to their queries [22, 23].

Even if the users are able to effectively craft a query, few consider more than three pages worth of search results [22, 23]. Spink et al. noted that "the public has a low tolerance of going in depth through what is retrieved" [23]. This low tolerance may be attributed to the static representations of web search results that are common in web search engines, and which require the 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.

In this paper, we present WordBars as a method for supporting the users in the process of interactive query refinement and interactive search results exploration. Our fundamental hypothesis in this work is that frequently used terms in the results of an initial search can provide valuable information to the user, both for interactive query expansion as well as for interactive search results re-sorting and exploration. Information visualization techniques are employed to convey the term frequency information to the users in a compact manner that can easily be interpreted and understood.

WordBars retrieves the top 100 document surrogates from the Google API [5], 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 sum of the selected 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 whenever the query is changed.

A fundamental design principle in the development of WordBars is the balance between computer automation and human control [21]. 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 automatically expand users' queries [28, 15, 26], 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 [2]. 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 interactive web information retrieval features provided by WordBars allow the users to take an active role in the information retrieval process, rather than the passive role that is common in traditional information retrieval. As such, WordBars can be classified as an information retrieval support system [30], providing support for the searchers as they browse, investigate, analyse, understand, and search a collection.

The remainder of this paper is organized as follows: A overview of the previous work on query refinement and search results re-sorting is provided in Section 2. In Section 3, an overview of the design and features of WordBars is given. Two examples for using WordBars to interactively refine a query and interactively explore the search results are provided in Section 4. The paper concludes with a discussion on the merits of WordBars in Section 5, followed by conclusions and future work in Section 6.

2. Background

2.1 Interactive Query Expansion

Query expansion is the process of adding additional terms to a user's original query, with the purpose of improving retrieval performance [4]. Although query expansion can be conducted manually by the searcher, or automatically by the information retrieval system, we focus on *interactive query expansion* which provides computer support for users to make choices which result in the expansion of their queries.

A common method for interactive query expansion is a technique known as *relevance feedback* [17], in which the users indicate the relevance and non-relevance of entire documents from the results of an initial search. This information is used to construct a new vector-based query with increased weights on the terms found in the relevant documents, and decreased weights on the terms found in the non-relevant documents.

Salton & Buckley [20] conducted an extensive evaluation of the relevance feedback techniques using a number of test collections, and showed these techniques to be quite effective. Chang & Hsu [3] clustered the initial search results, allowing the users to tag entire clusters of documents (as well as individual documents), thereby improving the user efficiency in providing the relevance feedback information.

One of the problems with applying these relevance feedback techniques directly to web searches is that the vector-based query model that is assumed in relevance feedback is not readily available for web searching. For a meta-search systems that use the Google API [5] or the Yahoo API [29], weighted, vector-based queries are not supported.

Instead, we investigate methods of analysing and processing the initial search results, and allowing the users to choose specific terms to add (or remove) from their query. This use of the data present in the top search results is often called *local analysis*. When users are able to explicitly remove terms from their query, we call this process *query refinement* rather than *query expansion* to highlight this difference.

Harman [6] provided three different lists to the user from which they could select additional terms to add to their query. The first list contained terms found in the first ten documents returned from the initial query, sorted based on various statistical techniques. The second list consisted of linguistic variations on the query terms. The third list was based on the co-occurrence of terms within the entire collection. The results reported from this work were good when users made perfect choices from the lists of available terms.

Applying the fundamentals of this technique to web search is somewhat problematic, especially for meta-search systems. Collecting the common terms used in the first ten documents returned by the initial search would require retrieving these documents from their source, introducing a delay that would not be well received by searchers that are used to near-instant response times. The utility of the list containing variations on the query terms is questionable. Generating the third list is not feasible due to the size of the collection (in the order of billions of documents).

Joho et al. [11] 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.

Our work on WordBars follows the local analysis techniques employed by Harman in the first list of terms. However, instead of collecting terms from the first ten documents returned from the initial search, we collect terms from the titles and snippets of the first 100 document surrogates. Further, we use a simple frequency statistic, rather than the statistical techniques described in Harman's work, most of which require access to the term frequencies within the entire document collection.

2.2 Search Results Re-Sorting

The re-sorting of search results based on web search personalization is a rather active research field [24, 25, 16]. These systems generally provide an automated re-sorting and filtering of the search results based on the personalized profiles of the users. 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 previous work on HotMap [7], we allowed the users to re-sort the search results based on the frequencies of the query terms within the search results. In our work on Concept Highlighter [8], we re-sorted the search results based on fuzzy membership scores with relation to user-selected concepts. In both of these systems, we found that the re-sorting features brought to the attention of the searcher highly relevant documents buried deep in the search results, and proved to be an effective method for exploring the search results.

3. WordBars

The design of WordBars is best explained with respect to three primary features: the meta-search and processing of the search results, the visual representation of the term frequency information, and the interaction features that are supported by the system. The details of these features are described in the remainder of this section.

3.1 Meta-Search and Term Frequencies

WordBars is a meta-search engine that makes use of the services of the Google API [5] to retrieve the web search results. Upon submitting a query to the system, the top 100 search results are obtained. This occurs in blocks of 10 document surrogates at at time, due to a restriction in the Google API.

As each block is retrieved, the title and snippet from each document surrogate are combined in a bag-of-words approach resulting in a document descriptor text string. Common terms, as well as terms that are less than three characters long are ignored. All other terms are reduced to their root forms using Porter's stemming algorithm [14]. The frequency of each stem in the document descriptor is counted, and this number is added to both a master vector that represents the term frequencies in the entire set of search results, and a local vector, which represents the term frequencies within the current document surrogate.

After processing each document surrogate, 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.

3.2 Visual Representation of Term Frequencies

While some previous systems have used simple textual lists to provide recommendations for additional terms to add to the query [6], providing additional information about the terms in a visual manner can be extremely beneficial. For example, Joho et al. [11] showed benefits to using a cascading menu representation of the query expansion terms. In WordBars, we opted for a simpler representation 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 sizes of the bars in the histogram, as well as the intensities of the colours, 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 [18]. The colour scale was chosen to vary both 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 colour scales used in WordBars were generated using the ColorBrewer application [1].

The term labels are provided to the right of each frequency bar. All 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 [27], 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, we assume that the most beneficial terms are those that are used most frequently within the top search results.

A grey box is 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 1 shows a visual representation of the term frequencies for a sample query.

3.3 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

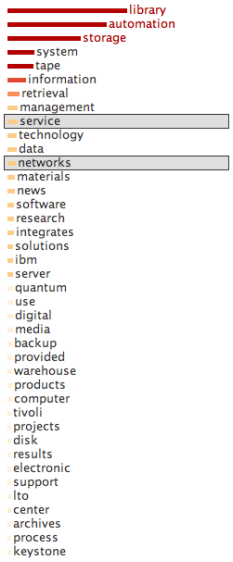

Figure 1. The visual representation of the term frequencies histogram allows the users to easily identify the frequency of the terms used in the search results, which terms are in the query (in the red font), and which terms are currently being used to sort the search results (in the grey boxes).

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, as well as a complete usage scenario, is available on the author's web site[1].

Once data begins to be displayed in the term frequency histogram, the user can interact with this interface by either single-clicking or double-clicking a term. These simple interaction methods were chosen to reduce the learning curve associated with using WordBars.

Single-clicking 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 status between selected and not selected. Selected terms are easily identified by the grey box surrounding them. This simple process allows the user to interactively explore the search results based on the terms they feel are relevant to their information needs.

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 and will retrieve the search results of the new query. Double-clicking on any term that is currently not in the query will add that term to the end of the query and will load the search results of the new 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.

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 provided by the Google API is included to highlight the effects of the re-sorting features. Clicking on any document will open that document in a new window, and will change the link colour from blue to purple (as per the defacto 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.

4. Examples

To illustrate the utility of WordBars in supporting the user's tasks of query refinement and search results exploration, we provide two examples: one with a specific initial query, and one with a vague initial query.

4.1 Specific Initial Query

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. Further, by providing a visual indication of the frequency of the terms, the users can easily interpret the relative frequency differences between terms. In addition, the user may use the term frequency histogram to re-sort the search results to further focus on a particular aspect of the information need, or even add new terms to the query, resulting in a search that is even more specific.

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 2a). 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 2b). Alternately, the user may choose to select the term "backup" to obtain a different sorting of the search results (Figure 2c). The user may decide to add this term to their

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

Figure 2. An example of exploring the search results and further refining a query when a specific query is provided as a starting point. Note the re-sorted search results in (b) and (c).

query by double-clicking on it, generating a more specific set of search results (Figure 2d).

From this example, it is easy to see the value of being able to re-sort the search results, as well as refine the query, using simple interaction features on the term frequency histogram. The ability to easily interpret the meaning of the histogram features allows the user to focus on determining the relevance of the terms in the term frequency histogram, and use this information for interactive query refinement and interactive search results exploration.

4.2 Vague Initial Query

It is common for users to provide vague queries for their web searches. This could 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. Further, web search engines provide little support to help the user improve their query.

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 a relatively low frequency of all other terms. This is due to the search results being a mixture of documents on multiple topics or sub-topics, all

Figure 3. An example of exploring the initial set of search results, and subsequently refining a query when a vague query is provided as a starting point.

of which use different terms in their descriptions.

Users of WordBars benefit from the support the system provides as they re-sort 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. The end result is that 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 the search results, indicating vagueness of the search results (Figure 3a). The user can explore the search results by selecting terms that are better descriptors of their information need, such as "hierarchical" and "documents" (Figure 3b). If the top documents are relevant, the user may choose to add these terms to the query by double-clicking on them (Figure 3c). The user may decide that some of the query terms are not very descriptive, and may choose to remove these, such as "algorithms" (Figure 3d).

Like the previous example, there is value in being able to re-sort the search results, as well as easily add and remove terms from the query. The terms presented in the term frequency 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 the query itself on this aspect (i.e, by adding the term to the query).

In both of these examples, 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 both interactively explore the search results, as well as interactively refine their queries.

5. Discussion

The first thing to note regarding the use of WordBars is that there is little ability to support the users in their information seeking 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 search results is of little value to the user. 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 list. 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 re-sort the search results based on the terms that are most relevant to their information need, or even add these terms to their query to generate a more specific set of search results.

One of the benefits of providing a list of commonly used terms in the top search results 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 provided by Nielsen as one of the primary usability principles [13]. In WordBars, this allows the users to begin with a somewhat vague query, and then use their recognition ability to add additional terms to the query, resulting in a refined query that did not require the user to remember the specific terms that are relevant to their information need.

The term frequencies in WordBars are generated from a subset of the actual document: the title of the document, 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 produce 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.

In previous work, we made use of an external knowledge base both for query refinement [9] as well as search results exploration [8]. WordBars is much more flexible, since it does not require the existence of an independent knowledge base. All the information provided to the user to support their information retrieval tasks is derived from the initial search results. This allows the users to benefit from this system, as long as their initial query includes some relevant documents.

Even when presented with a list of potential terms to add to a query, research has shown that users may still have difficulties choosing good terms from such lists [19, 12]. However, in these studies, the query terms were presented to the users in a simple list. 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. Further, the ability to re-sort the search results can allow users to see how potential query expansion terms may be used. This additional information can allow the users to make informed decisions for query expansion that would not be possible when simply considering terms in a list.

6. Conclusions and Future Work

In this paper, we have presented our work on the development of an information retrieval support system that allows the users to interactively explore web search results, as well as interactively refine their queries. Although these tasks are fundamentally different, they are supported within the same user interface, allowing the user to easily transition back and forth between them. The visual representation of the term frequencies, and the interactive nature of the support tools provided allows the users to take advantage of their intelligence and judgement abilities as they perform their information retrieval tasks using WordBars.

Through the visual representation of the term frequency histogram, the users can easily identify the relative frequencies of their query terms in the top search results, as well as the relative frequencies of other terms present in the document surrogates. Identifying terms in this list can help the user to understand the general makeup of the search results, as well as the degree of specificity of their initial query. Terms can easily be selected, resulting in a re-sorting of the search results based on the frequencies of the selected terms. This assists the users in exploring the search results. Terms can also be added or removed from the query, auto-

matically generating a new set of search results from which the user can further explore.

Although the techniques used in this work are rather simple, there is a benefit to the users for this simplicity. The interface is uncluttered, easy to learn, and simple to use. There are no complex interactions required to re-sort the search results, nor to add or remove terms from the query. The frequency statistics are calculated interactively as the search results are retrieved from the Google API, resulting in a minimal delay. These statistics are easy for the users to make sense of, and result in a meaningful ordering of the terms present in the search results. The visual display of this information allows the users to easily, quickly, and accurately perceive and interpret the term frequencies, and make use of this information through interactive query refinement and interactive search results exploration.

The current prototype system only supports a simple list of terms in the query. In future work, we wish to support more complex queries, including the ability to require or exclude specific terms in the search results. Further, the ability to personalize the search results based on terms selected through this interface will be investigated. A user evaluation of WordBars is currently in the planning stages.

References

[1] C. A. Brewer. www.colorbrewer.org, 2005.

[2] 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.

[3] C.-H. Chang and C.-C. Hsu. Enabling concept-based relevance feedback for information retrieval on the www. *IEEE Transactions on Knowledge and Data Engineering*, 11(4), 1999.

[4] E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, 31, 1996.

[5] Google. Google web api. http://www.google.com/apis/.

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

[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] O. Hoeber, X. D. Yang, and Y. Yao. Visualization support for interactive query refinement. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.

[10] 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), 2001.

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

[12] 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.

[13] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1994.

[14] M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 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] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART System - Experiments in Automatic Document Processing*. Prentice Hall, 1971.

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

[19] 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.

[20] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 1990.

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

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

[23] 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.

[24] K. Sugiyama, K. Hatano, and M. 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.

[25] 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.

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

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

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

[29] Yahoo. Yahoo search web searvices. http://developer.yahoo.com/search.

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