Web Information Retrieval Support Systems: The Future of Web Search

Orland Hoeber
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada A1B 3X5
hoeber@cs.mun.ca

Abstract

The primary focus of Web Information Retrieval Support Systems (WIRSS) are to address the aspects of search that consider the specific needs and goals of the individuals conducting Web searches. WIRSS moves beyond the traditional focus of automated searching within digital collections, applying intelligent methods and Web-based technologies to assist users in specifying their information needs, evaluating and exploring search results, and managing the information they find. The goal of this paper is to provide an overview of some of the key issues, challenges, and opportunities in WIRSS research.

1. Introduction

The traditional definition of information retrieval focuses on automated searching within digital collections. The goal is to find all the relevant documents, while selecting as few of the non-relevant documents as possible [1, 26]. One drawback of this definition is that there is little acknowledgment of the activities users may wish to perform with the information retrieval system. The underlying assumption is that users are able to accurately describe their information needs to the system, and that providing ranked lists of documents will satisfy the users' needs.

Two key aspects of conducting Web searches, crafting queries and evaluating search results, are inherently human tasks. Searchers have mental models of the information needs they wish to satisfy, and draw upon their knowledge of other related concepts as they seek documents that may be relevant to these goals. When using a Web search engine to assist in fulfilling an information seeking task, a balance must be struck between computer automation and human control [21]. It is the human decision-making aspect of Web search that Web Information Retrieval Support Systems (WIRSS) aim to enhance and promote. Such systems are especially valuable for Web search tasks that are ambiguous or exploratory in nature [9, 10].

While traditional Web search systems focus on search and browsing functionalities, WIRSSs focus on the functionalities that support user-centric tasks that are performed with the Web search system [34]. In order for users to extract useful information from the Web, and be able to make effective use of this information, users must take an active role in the tasks associated with Web search. These tasks include crafting and refining queries; browsing, filtering, investigating, and exploring search results sets; and analyzing, understanding, organizing, and saving retrieved documents.

Fundamentally, this is a change in the design assumptions for Web search systems; moving the focus of Web search from the documents being searched to the tasks that people need to perform. We believe this philosophical shift will mark the move towards next-generation Web search systems, and a transition from information retrieval to knowledge retrieval.

In this paper, we expand upon the prior foundation for WIRSS research [34, 33]. We focus on the key issues, challenges, and benefits of research on WIRSS. The discussion is broken into five sections that deal with the breadth of topics associated with WIRSS research.

2. Web

When people today think of information retrieval, text searching, and document searching, they commonly think of Web search engines. The use of Web search engines have become common place among Web users, and are increasingly being used in all aspects of society [16]. Nielsen reported that 88% of Web users start with a Web search engine when provided with a task to complete using the Web [19]. This is in support of earlier studies that reported nearly 85% of Web users find new Web pages using search engines [5].

Although the Web can be considered a single distributed document collection, this collection has features that make many of the traditional approaches to information retrieval

impossible or impractical to implement [31]. These features include the size of the Web (billions of documents), the diversity of the collection as a whole (documents available on virtually any topic), and the potential diversity of individual documents (single documents may discuss multiple distinct topics).

The focus of much research on Web search has been to address these challenges from the perspective of the underlying data, as well as to exploit the connected nature of the Web [2, 20]. The success of these research activities has lead to a common perception that Web search is a solved problem. However, as little research has addressed the human aspect of Web search, we don't subscribe to this belief.

Yao noted that "the next evolution of retrieval systems is to move from IRS [information retrieval systems] to information retrieval support systems (IRSS)" [34]. Extending this into the Web domain, the next evolution of Web search is to move from Web information retrieval systems (WIRS) to Web information retrieval support systems (WIRSS). In this evolutionary step, the focus changes from the indexing and query matching of today's Web search engines, to supporting the fundamental knowledge-generating activities of the users.

An interesting area for future WIRSS research activities is to take advantage of the features of the Web as a means for supporting searcher goals. For example, in some circumstances, a searcher may wish to be exposed to a broad range of topics relevant to a given query. A WIRSS could exploit the diversity of the collection and select documents that represent the breadth of topics that match the given query. The searcher could then use the system to browse the documents and ultimately focus on a particular topic of interest. Although this may resemble systems that cluster Web search results (e.g., [27, 35]), the focus in the design of such a system is on exploring the breadth of documents available, rather than organizing the top documents in the search results set.

3. Information

The primary information present in Web search activities are the documents themselves, along with the document surrogates (titles, snippets, and URLs) commonly provided in Web search results. Although other information is also used by the algorithmic approach to Web search (e.g., link structures, document types, term metrics, etc.), from the user's perspective the only information they are exposed to are the list of search results and the documents to which they point.

The key aspect in WIRSS is to support searchers in finding useful information and knowledge from Web resources [34]. However, a fundamental issue that seems to get little attention in the research literature on Web search is how the information that supports the Web search activities gets presented to, and ultimately used by the searchers. It seems that the list-based representation used by the top Web search engines has become so common-place that there is little if any discussion on whether these simple interfaces are providing adequate support to the users.

There are two facets of the information perspective of WIRSS that are promising avenues for further research: personalization and information visualization. Fundamentally, the personalization of Web search deals with the modeling of searchers' interests, and then using these models to affect the outcomes of their future Web search activities. In particular, this area of research deals with using machine learning algorithms to generate searcher profiles based on the information they have found useful in the past. These profiles can the be used to filter, re-order, or categorize the search results. Combining them into group profiles can result in more robust systems when the information that is available is incomplete. Although a number of Web search personalization methods have been developed in recent years [3, 17, 25, 29], there remain many opportunities for further research in this domain.

Information visualization techniques address the challenges of representing aspects of Web searches to the useres in order to promote their understanding of the underlying information. Fundamentally, information visualization is a technique for creating interactive graphical representations of abstract data or concepts [28]. Moreover, information visualization promotes a cognitive activity in which users are able to gain understanding or insight into the data being graphically displayed by taking advantage of human visual information processing capabilities [23].

Our research has focused on providing visual representations of Web search, in support of both query refinement [11, 12] and search results exploration activities [6, 7, 8, 9, 10]. There are a myriad of opportunities for further research that addresses the issues of information overload during Web search activities. Such visual WIRSSs can be designed to take advantage of aspects of the human visual processing systems to interactively convey information about the search processes to the users, allowing them to interactively manipulate features of their search activities.

4. Retrieval

The underlying Web search engine is a critical aspect in any WIRSS. Existing Web search engines primarily focus on the indexing of documents, the matching of queries to the indexes, and providing a list-based representation of the search results set. The main focus has been on supporting the traditional tasks of retrieval and browsing [1].

In recent years, many of the top search engines have made portions of their core search engine functionality accessible to the public. Both Google [4] and Yahoo [30] pro-

vide comprehensive access to their underlying search engines, with the only technical restriction being the number of queries that can be submitted per day. Although this has made WIRSS research and development easier, transitioning research prototypes into publicly accessible systems remains difficult.

Although the retrieval aspect of Web search has received a significant amount of attention, both in the research community and through the commercial activities by the top search engines, there remains further avenues for research. In particular, supporting the features of WIRSSs directly within the internal data structures of the Web search engine will result in substantial performance and stability improvements. Natural language processing within the underlying search engine, as well as support for query-by-example and weighted queries, may also be beneficial for the support of query specification and refinement within WIRSSs.

5 Support

As noted previously, traditional approaches to Web search have not adequately addressed the complexity of information, instead providing only simple text boxes for entering queries and simple list-based representations of search results. An outcome of these simple interfaces is that people use them in a simple manner: queries commonly contain only one or two terms [13, 24], and people seldom venture beyond the third page of the search results [22, 24]. While Web search engines may be able to perform well under these conditions when the searcher is seeking to fulfill simple targeted search operations, more complex searches are not well supported. The focus of WIRSS research is to move beyond the simple functionality provided in these interfaces, supporting the searchers at a deeper, task-oriented level.

An important aspect of any information system is the ability to provide information seeking functions which assist users in defining and articulating their problems, and finding solutions to these problems [18]. As noted by Yao and Yao, "the lack of consideration of the diversity, background, and intentions of users affects the performance of IR systems" [32].

WIRSSs focus on supporting the user-oriented aspects of Web search, including activities such as *investigating*, *analyzing*, *organizing*, *filtering*, *understanding*, *saving*, *sharing*, *modifying*, *manipulating*, *summarizing*. Ultimately, research in WIRSS should support the searcher as they perform one or more of these activities, with the ultimate goal of improving the human aspects of Web search.

Clustering is often cited as a useful method for organizing search results, supporting users in their tasks of investigating the search results set. Examples include Vivisimo [27] and Grouper [35]. Recent efforts have attempted to provide consistent cluster naming in order to promote topic learning [15]. Opportunities exist to extend these techniques, and develop new techniques, to support the broad range searcher activities.

In order judge the utility of such approaches, a better understanding of searcher tasks, goals, and activities is needed [34]. From a research perspective, not only do we need to study how people currently use Web search engines to satisfy their information needs, but also how effectively they are able to learn and use the features of new WIRSS prototypes. As a result, there is a strong need for human-computer interaction research within the domain of WIRSS, both as snapshot case studies and in longitudinal settings.

6. Systems

From a systems perspective, most Web search engines have traditionally operated as stand-alone applications, designed to be used to provide an answer and then discarded. Recent advancements have added the ability for systems to remember past search activities, system preferences, and personalized content [14]. These new features, along with the availability of APIs, have begun to address the need for the extension and integration of existing systems [33].

Within WIRSS research activities, we must strive towards building systems and frameworks that can be combined together. Designing systems to communicate with one another can allow one system that provides a visual representation of Web search results to be combined with a system that supports the personalization of Web search results, for example. The end result of such a systems-level approach to inter-operability will be the ability to more readily construct a system that supports the breadth of activities a searcher may wish to perform. Inspiration for such research should be taken from the APIs provided by Google [4] and Yahoo [30].

7. Conclusion

As the amount of information on the Web continues to grow, search engines will continue to be the primary method by which people find information. The advances that Web search companies make in their algorithms and infrastructure has and will continue to allow them to index the Web as it grows, yet still return the results of a search in fractions of a second. Other advances in Web search will include indexing the "deep Web" and improving the ability to deduce the potential relevance of documents.

One aspect that will have a significant impact on the utility of Web search engines of the future will be the support provided for the users as they conduct their search activities. As people become more accustomed to searching and using

the Web, they will begin to demand more powerful tools to support their search needs. In this paper, we have provided an overview of some of the key areas of WIRSS research, providing a vision for what we believe to be the future of Web search: Web information retrieval support systems.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[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] D. Bueno, R. Conejo, and A. David. Metiorew: An objective oriented content based and collaborative recommending system. In *Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia*, pages 310–320, 2001.

[4] Google. Google web API. `http://www.google.com/apis/`, 2005.

[5] Graphics, Visualization, & Usability Center. GVU's 10th WWW user survey. `http://www.gvu.gatech.edu/user_surveys/survey-1998-10/`, 1998.

[6] O. Hoeber. Exploring Web search results by visually specifying utility functions. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.

[7] 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.

[8] 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.

[9] O. Hoeber and X. D. Yang. Evaluating the effectiveness of term frequency histograms for supporting interactive Web search tasks. In *Proceedings of the ACM Conference on Desigining Interactive Systems*, 2008.

[10] O. Hoeber and X. D. Yang. Evaluating WordBars in exploratory Web search scenarios. *Information Processing and Management*, 44(2):485–510, 2008.

[11] 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.

[12] O. Hoeber, X. D. Yang, and Y. Yao. VisiQ: Supporting visual and interactive query refinement. *Web Intelligence and Agent Systems: An International Journal*, 5(3):311–329, 2007.

[13] 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.

[14] S. Kamvar and M. Mayer. Personally speaking. `http://googleblog.blogspot.com/2007/02/personally-speaking.html`, 2007.

[15] B. Kules, J. Kustanowitz, and B. Shneiderman. Categorizing Web search results into meaningful and stable categories using fast-feature techniques. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, 2006.

[16] S. Lawrence and C. L. Giles. Accessibility of information on the Web. *Nature*, 400:107–109, 1999.

[17] Z. Ma, G. Pant, and O. R. L. Sheng. Interest-based personalized search. *ACM Transactions on Information Systems*, 25(1), 2007.

[18] G. Marchionini. Interfaces for end-user information seeking. *Journal of the American Society for Information Science*, 43(2):156–163, 1992.

[19] J. Nielsen. When search engines become answer engines. `http://www.useit.com/alertbox/20040816.html`, August 2004.

[20] E. M. Rasmussen. Indexing and retrieval for the Web. *Annual Review of Information Science and Technology*, 37(1):91–124, 2003.

[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):6–12, 1999.

[23] R. Spence. *Information Visualization: Design for Interaction*. Pearson Education, 2nd edition, 2007.

[24] 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.

[25] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive Web search based on user profile construction without any effort from users. In *Proceedings of the World Wide Web Conference*, pages 675–684, 2004.

[26] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[27] Vivisimo. Vivisimo search engine. `http://www.vivisimo.com/`, 2005.

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

[29] S. Wedig and O. Madani. A large-scale analysis of query logs for assessing personalization opportunities. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 742–747, 2006.

[30] Yahoo! Yahoo! search Web services. `http://developer.yahoo.com/search/`, 2005.

[31] K. Yang. Information retrieval on the Web. *Annual Review of Information Science and Technology*, 39(1):33 – 80, 2005.

[32] J. Yao and Y. Yao. Information granulation for web based information retrieval support systems. In *Proceedings of SPIE*, volume 5098, pages 138–146, 2003.

[33] J. Yao and Y. Yao. Web-based information retrieval support systems: building research tools for scientists in the new information age. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, pages 570–573, 2003.

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

[35] O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to Web search results. In *Proceedings of the Eighth International World Wide Web Conference*, 1999.