Webology, Volume 7, Number 1, June, 2010

Paradigm shifts: from pre-web information systems to recent web-based contextual information retrieval

MPS Bhatia
Dept. of Computer Engineering, Netaji Subhas Institute of Technology, University of Delhi, India. E-mail: mpsbatia (at) nsit.ac.in

Akshi Kumar
Dept. of Computer Engineering, Delhi Technological University, India. E-mail: akshikumar (at) dce.ac.in

Received February 2, 2010; Accepted June 25, 2010

Abstract

As the types of user accessible data and information escalates, so does the variety of Information Retrieval (IR) practices which can match to achieve the challenges instigated. By expanding its applicability which can broaden the use, integrating technologies and methods and as long as the quest for the perfectly accurate system continues to exist it is quite possible and likely that Information Retrieval can become one of the key technology areas for current and future research and practice. This paper expounds the recent research advances in the area of Contextual Information Retrieval. It tracks and investigates the evolution of retrieval models from the pre-web (traditional) Information Retrieval paradigm and Web information retrieval to the most prominent interactive Web information retrieval field of contextual information retrieval focusing on developing models and strategies of contextual IR.

Keywords

Traditional information retrieval; Web information retrieval; Contextual information retrieval

Introduction

The field of Information Retrieval (IR) has made great strides in past years, and many industry analysts and research firms have projected a vivid future for this area. The paper discusses the evolution of retrieval models in the field of IR, with a focus the most promising applicability to future IR applications, i.e., *contextual information retrieval* (CIR). The generic Information Retrieval task can be specified as "Retrieve that amount of information which a user needs in a specific situation for solving his/her current problem" (Kuhlen, 1991). The model of IR can be defined as a set of premises and an algorithm for ranking documents with regard to a user query (Bhatia & Kumar, 2009). More formally, an IR model is a quadruple [D, Q, F, R (qi,dj)] where D is a set of logical views of documents, Q is a set of user queries, F is a framework for modeling documents and queries, and R(qi,dj) is a ranking function which associates a numeric ranking to the query qi and the document dj.

The field of Information Retrieval has become exceptionally significant in recent years due to the intriguing challenges presented in tapping the Internet and the Web as an inexhaustible source of information. The success of web search engines testifies this fact (Bhatia & Kumar, 2008a). The information seeking on the Web is fairly similar and notably different with searching in the classic information retrieval systems. The similarity is regarded in lieu that both deal with information using computer. However, the considerable dissimilarity lies in the varied information resources and potential users with diverse goals and expectations, of the Web environment which are very different with pre-web online information systems. Information resources on the Web are so heterogeneous, semi-structured, distributed, and usually time-varying unlike the information stored in pre-web (traditional) information systems which is well-organized and homogenous. Apparently, if not all, the majority of the users of pre-web (traditional) information systems have been generally experienced users like researchers and academicians with computer skills, subject knowledge, and search experience. In contrary, the Web users include almost everybody with different search behaviors and experiences, varied goals, knowledge and information needs.

The enormous growth of the Web and the increasing expectation placed by the user on the search engine to anticipate and infer his/her information needs suggest that users are becoming more and more dependent on the search engines' ranking schemes to discover information relevant to their needs. However, the "one size fits all" model of web search may limit diversity, competition, and functionality (Lawrence, 2000). Typically, users expect to find information in the top-ranked results, and more often than not they only look at the document snippets in the first few result pages and then they give up or reformulate the query. This can introduce a significant bias to their information finding process and calls for ranking schemes that take into account not only the overall page quality and relevance to the query, but also the match with the users' real search intent when they formulate the query. New search services that incorporate context, and further incorporation of context into existing search services, may increase the retrieval effectiveness, and help mitigate any negative effects of biases in access to information on the Web.

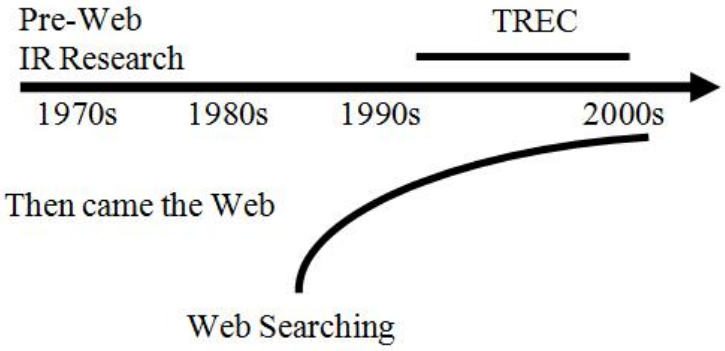
This paper focuses on the *contextual information retrieval (CIR) paradigm*, which has the primary goal to acquire a user's information seeking behavior, such as their search activities and responses, and incorporate this information into a search system. The primary goal of incorporating context is to increase the relevance of results, although other outcomes, such as effectiveness, efficiency, and subjective satisfaction, could potentially be affected as well. Thus, exploring contextual factors and modeling information seeking behavior in CIR is imperative.

Pre-web (traditional) Information Retrieval Systems and their Users

In general, not only information resources but also the users in the pre-web (traditional) systems were reasonably homogenous and conventional. The online and offline databases contained primarily structured data stored in well-organized systems. In this structured environment, each document had a specific structure and this made the storage and retrieval procedure much easier and more predictable. Also, the users were limited to some specific groups of the society mostly the academics and researchers, librarians or subject expert people. Thus the limitations of the classic IR can be summed up as:

- Individuals were ignored.
- The Corpus is predetermined.
- The Context was ignored.

Figure 1: Timeline depicting the Web emergence



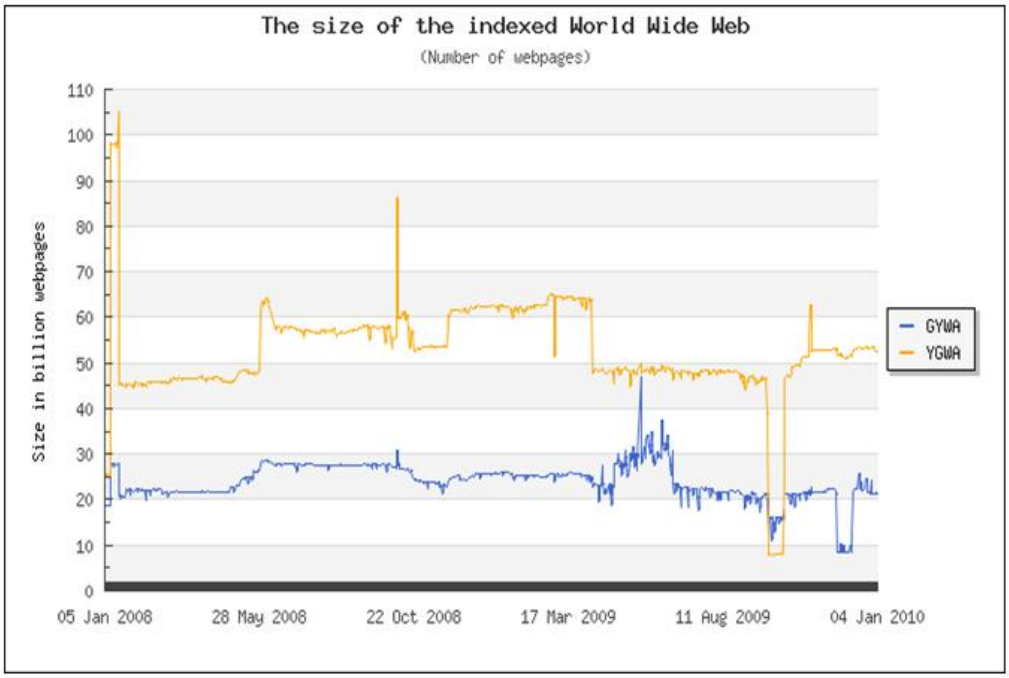
The Emergence of the Web

The Web is seemingly unlimited source of information with users from cross-section of society seeking to find information to satisfy their information needs. They require the Web to be accessible through effective and efficient information retrieval systems that deliver information need fulfillments through the retrieval of web content (Bhatia & Kumar, 2008b). The field of Information Retrieval (IR) is a long tackled the problem of finding useful information. With the prevalence of the Web, this problem compounded manifold. Web Information Retrieval (Web IR), defined as the application of theories and methodologies from IR to the Web (Bhatia & Kumar, 2008c), include problems that are a combination of challenges that stem from traditional information retrieval and challenges characterized by the nature of the Web. Specifically, the operative challenges motivating researchers in Web IR are relating either to data quality or user satisfaction. In spite of all noteworthy advances in developing more sophisticated Web IR tools, people may still encounter many problems when interacting with the Web. The paper (Bhatia & Kumar, 2008c) discusses the Web IR paradigm as a variant of classical Information Retrieval, and expounds its basics, system components, and model categories, probing the Web IR tools, tasks and performance measures.

Web IR is different from classical IR for two kinds of reasons: concepts and technologies. The following characteristics of the Web shape up the nature of Web IR and are what make it considerably different to traditional retrieval challenges (Bhatia & Kumar, 2008b):

1. *The "Abundance" of web:*
With the phenomenal growth of the Web, there is an ever increasing volume of data and information published in numerous Web pages. According to worldwidewebsize.com, According to Worldwidewebsize.com, the indexed Web contains at least 21.72 billion pages (Saturday, 12 December, 2009).

Figure 2: The Size of the indexed web in the last two years (2008-2010)



2. *Heterogeneity:*

- Information /data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.
- Much of the Web information is semi-structured due to the nested structure of HTML code, i.e., much of the Web information is linked.
- The Web is noisy: A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisement, navigational panels, and copyright notices.

3. *Dynamics:*
The freedom for anyone to publish information on the Web at anytime and anywhere implies that information on the Web is constantly changing.

4. *Duplication:*
Several studies indicate that nearly 30% of the Web's content is duplicated, mainly due to mirroring.

5. *Users Search Behavior:*
The users have different expectations and goals such as informative, transactional and navigational. They often compose short, ill-defined queries and impatiently look for the results mainly in the top 10 results.
Table 1 shows the paradigm shift from the traditional IR to the IR on the Web.

Table 1: From IR to Web IR

	Classic IR	IR on the Web
Input	Fixed Document collection	The publicly accessible Web
Goal	Retrieve documents or text with information content that is relevant to user's information need	Retrieve high quality pages that are relevant to user's need: 1. Static (files: text, audio, . . .) 2. Dynamically generated on request
Aspects	1. Processing the collection 2. Processing queries (searching)	1. Processing and representing the collection: - Gathering the static pages - "Learning" about the dynamic pages 2. Processing queries

Contextual Information Retrieval on the Web: Concept, Framework and Tools

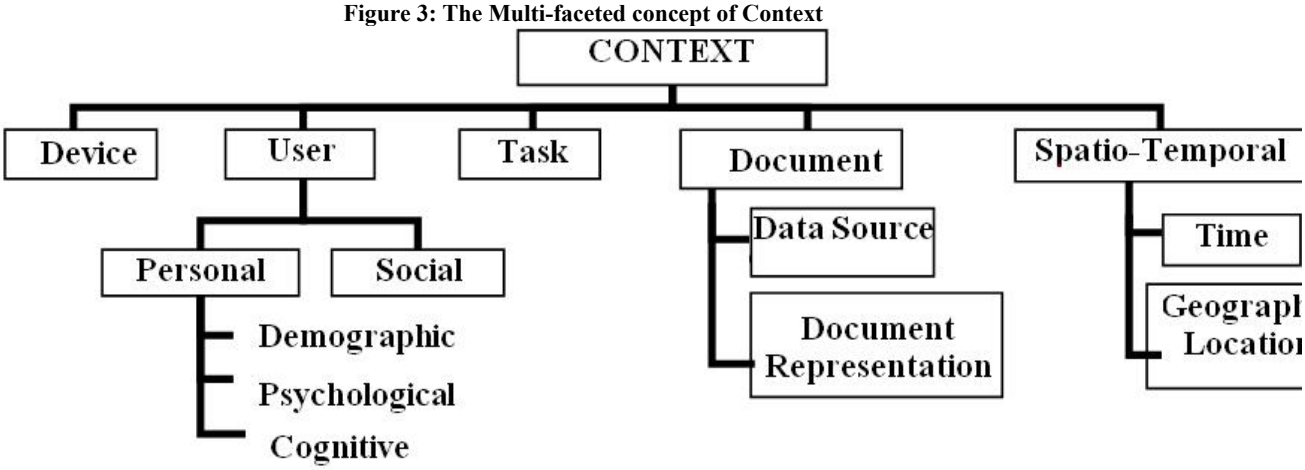
A key challenge in IR is: how to capture and how to integrate contextual information in the retrieval process in order to increase the search performances? In Allan (2002), *contextual information retrieval* is defined as "*combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for user's information needs*". Contextual IR intends to optimize the retrieval accuracy by involving two related steps: appropriately defining the context of user information needs, commonly called *search context*, and then adapting the search by taking it into account in the information selection process.

The Multi-faceted concept of Context in IR

One of the primary questions here is: Which facets of context should be considered in the retrieval process? Several studies offered context specification within and across application domains (Goker et al., 2008; Vieira et al., 2007). Figure

3 depicts the five context specific dimensions that have been explored in contextual IR literature.

Figure 3: The Multi-faceted concept of Context



The pertinent literature reviewing the multiple facets of the concept of context was beyond the scope of this paper. Although, the brief overview of the literature in this area indicates prominent and relevant research has been undertaken but highlights that the findings provided to the research community are often just some insights and proposals for the design of contextual IR with potential of several contextual factors to advance the knowledge of contextual IR is not actually exploited. Recent research advances in the area have focused on developing models and strategies for contextual IR according to the following aspects:

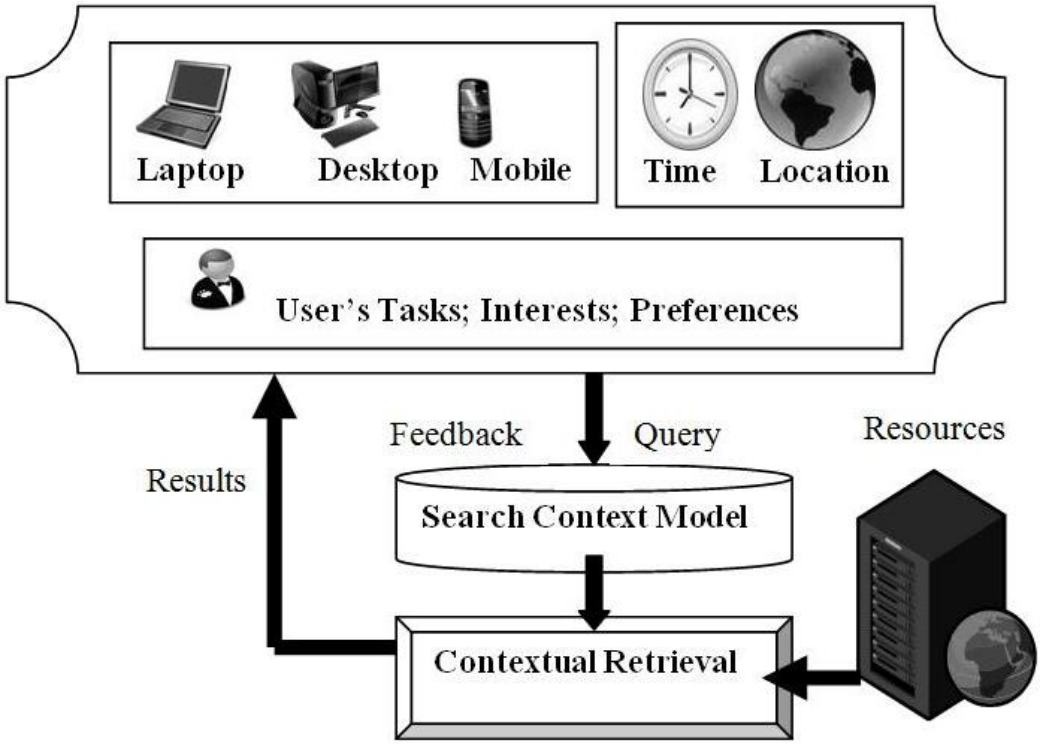
- *User's context and task*: This aspect has been addressed and explored by research studies in *personalized IR*, considered as a sub-field of contextual IR, supporting explicit/implicit representations of the user himself involved in the retrieval framework (Anand & Mobasher, 2007).
- *Device and spatio-temporal context*: This aspect has been addressed by research studies in *context-aware IR in mobile environments* (Chittaro, 2003).

The Framework for Contextual IR

An Information Retrieval System is context-aware if it exploits context data in order to deliver relevant information to the user. CIR aims at optimizing the retrieval accuracy by involving two related steps: appropriately defining the context of user information needs, commonly called *search context*, and then adapting the search by taking it into account in the information selection process. New search services that incorporate context, and further incorporation of context into existing search services, may increase the retrieval effectiveness, and help mitigate any negative effects of biases in access to information on the Web (Bhatia & Kumar, 2008a). The *contextual information retrieval paradigm* has the primary goal to acquire a user's information seeking behavior, such as their search activities and responses, and incorporate this information into a search system.

Contextual IR aims at delivering the right information to the user, in response to his query, within the right context. Numerous approaches - employing contextual user profiles, concept-based query formulation and relevance filtration and relevance feedback/suggestion - already exist today. Previous work in the area of CIR has focused on three main themes, namely, ***User Profile Modeling*** [Durrani, 1997; Joachims et al., 1997; Motomura et al., 2000; Brusilovsky, 2001; Pitkow et al., 2002; Speretta & Gauch, 2005; Teevan & Dumais, 2005), ***Query Expansion*** (Efthimiadis, 1996; Beaulieu, 1997; Xu & Croft, 2000; Lin et al., 2006) and ***Relevance Feedback*** (Rocchio, 1971; Allan, 1996; Spink & Losee, 1996; Spink, 1997; Kelly, 2004). Figure 4 presents the basic architecture of a context-aware called also Contextual IR system.

Figure 4: The Contextual IR Framework



User Profile Modeling:

Research works have focused on exploiting the sources of evidence that more precisely include approaches to build the user profile that allow learning user's context by implicitly inferring the information from the user's behavior and from external or local context sources. Several pertinent studies on Web IR systems have examined various user modeling approaches to improve the personalization of a users' Web search experience. A review of these user modeling approaches reveals that in order to construct a contextual profile these techniques utilize either user behavior or preferences. However, none of the approaches have used a combination of user behavior and preferences and do not have the capability to share a user's contextual profile information with other users, thereby potentially leading to suboptimal performance when the user needs access to information outside their original context [with an exception of WebMate (Chen & Sycara, 1998)]. While showing promise, prior conventional IR approaches employing user profile modeling have had limited success. Fundamental challenges remain, specifically:

- How to acquire, maintain and represent information about a user's interests with minimal intervention?
- How to deliver personalized search results using the user information acquired?
- How to use information acquired from various users as a knowledge base for interest communities or groups?

Query Expansion:

The individual's primary point of access in Contextual IR systems is the Query expansion, which continues to be the main approach to information seeking on the Web. The query expansion approaches attempt to expand the original search query by adding further, new or related terms. These additional terms are inserted to an existing query either by the user (Interactive query expansion, IQE), or by the retrieval system (Automatic query expansion, AQE) and intend to increase the accuracy of the search. A comparative study of the two approaches has though revealed inconclusive findings regarding the relative merits of each (Beaulieu, 1997; Koenemann & Belkin, 1996). A review of query expansion approaches has highlighted a number of attractive and promising alternatives for identifying terms to add to the original query, viz., the thesauri and concept based approaches, approaches that utilize search histories and query logs (Limbu et al., 2006). The main challenges for current query expansion techniques are:

- Which terms should be considered and included for query expansion?
- How is ranking done for the selected terms?
- Which query reformulation levels should be automatic, interactive or manual?

Relevance Feedback:

Relevance feedback is the most commonly used technique to assist in the formulation of effective query statements. The notion of relevance feedback (RF) is to take into account the results that are returned initially in response to the input query and to use information about whether or not those results are relevant to perform a new query. Relevance feedback provides a means for automatically reformulating a query to more accurately reflect a user's interests (Allan, 1996). RF has been researched extensively in interactive settings and can be exploited using either explicit or implicit feedback. The feedback information is used to construct a user's profile which is used to query, filter and return relevant information

(Limbu et al., 2006). Despite considerable research, these approaches have not been successfully implemented in web-based information retrieval (Croft et al., 2001). The main challenges faced by current RF mechanisms are:

- How to capture a user's intent supported by the information-seeking behavior and preferences and to model this information in such a way as to be able to characterize a search context that can be refined over time?
- How to facilitate building and collaboration of communities of interest for user while preserving their personal privacy?
- How to develop techniques and algorithms that combine multiple types of information to compute recommendations?

A novel context-based technique for the ad-hoc retrieval of web documents is investigated in the research (Bhatia & Kumar, 2009) which relies on a number of inter-related parameters that define the nature of the context it uses. Under the proposed Contextual Proximity Model (CPM), it is not the frequency of individual query terms that is measured, but the frequency and proximity of their co-occurrence.

Online Contextual IR tools

Recently, several popular web search engines like Google and Yahoo have integrated contextual search tools in order to adapt the results to match each user's intent and information needs. The personalized search tools integrated are the *Google's Alerts* (http://www.google.com/alerts) and *Google's Personalized Search* (enhanced version known as My Search History) by Google and Yahoo's *My Web* (http://myweb.search.yahoo.com) with the goal to allow users archive their search activity and results, and then sharing this information with other people if they choose to. The latest *Yahoo Contextual Search* (Y!Q) (Kraft et al., 2005) is designed to help people find more relevant content online using the short term search context.

Conclusion

Regarding all the likeness and disparities between web searching paradigm and pre-web search process, this paper concludes that the Web search paradigm utilizes the research methodology and approaches that belong to the pre-web era but also takes into account all new elements of the Web environment to achieve better understanding of the user's information need and to managing the huge amounts of information. In other words, while Web search research is being built upon the previous investigations, the new line of enquiry, *contextual information retrieval*, is constructing new independent area of research.

References

- Allan, J. (1996). Incremental relevance feedback for information filtering. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, Zurich, Switzerland, 270-278.
- Allan, J. (2002). Challenges in information retrieval and language modeling. *Report of a workshop held at the Center for Intelligent Information Retrieval*, University of Massachusetts, Amherst.
- Anand, S.S., & Mobasher, B. (2007). Introduction to intelligent techniques for web personalization. *ACM Transactions on Internet Technology*, 7(4), 18.
- Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation*, 53(1), 8-19.
- Bhatia, MPS., & Kumar, A. (2008a). The context-driven generation of web search. *Proceedings of Conference on Information Science Technology and Management (CISTM'08)*, 281-287.
- Bhatia, MPS., & Kumar, A. (2008b). Information retrieval and machine learning: supporting technologies for web mining research and practice. *Webology*, 5(2), article 55. Retrieved December 20, 2009, from http://www.webology.org/2008/v5n2/a55.html
- Bhatia, MPS., & Kumar, A. (2008c). A primer on the web information retrieval paradigm. *Journal of Theoretical and Applied Information Technology*, 4(7), 657-662.
- Bhatia, MPS., & Kumar A. (2009). Contextual paradigm for ad-hoc retrieval of user-centric web-data. *IET Software*, 3(4), 264-275.
- Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User Adapted Interaction*, 11, 87-110.
- Chen, L., & Sycara, K. (1998). WebMate: a personal agent for browsing and searching. *Proceedings of the International Conference on Autonomous Agents*, Minneapolis, Minnesota, USA.
- Chittaro, L. (Ed.) (2003). Human-computer interaction with mobile devices and services. *Lecture Notes in Computer Science*, Vol. 2795, Springer, Berlin.
- Croft, W.B., Cronen-Townsend, S., & Lavrenko, V. (2001). Relevance feedback and personalization: a language modeling perspective. *Proceedings of the DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, Dublin City University, Ireland.
- Durrani, Q.S. (1997). Cognitive modeling: a domain independent user modeling. *Proceedings of the IEEE International conference on System man and cybernetics*, Orlando, FL, USA.
- Efthimiadis, E.N. (1996). Query expansion. *Annual Review of Information Systems and Technology*, 31, 121-187.
- Goker, A., & Myrhaug, H. (2008). Evaluation of a mobile information system in context. *Information Process Management*, 44(1), 39-65.
- Joachims, T., Freitag, D., & Mitchell, T. (1997). WebWatcher: a tour guide for the World Wide Web. *Proceedings of the International Joint Conference on Artificial Intelligence*, Nagoya, Japan.
- Kelly, D. (2004). *Understanding implicit feedback and document preference: a naturalistic user study*. Rutgers University, New Brunswick, NJ.
- Koenemann, J., & Belkin, N.J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. *Proceedings of the SIGCHI conference on Human factors in computing systems: common*

ground. Vancouver, British Columbia, Canada, 205-212.
- Kraft, R., Maghoul, F., & Chang, C. (2005). Y!Q: contextual search at the point of inspiration. *CIKM '05: Proceedings of the 14th ACM international conference on information and knowledge management*, ACM Press, New York, NY, USA, 816-823.
- Kuhlen, R. (1991). Information and pragmatic value-adding: language games and information science. *Computers and the Humanities*, Vol. 25, 93-101.
- Lawrence, S. (2000). Context in web search. *IEEE Data Engineering Bulletin*, 23(3), 25-32.
- Motomura, Y., Yoshida, K., & Fujimoto, K. (2000). Generative user models for adaptive information retrieval. *Proceedings of the 2000 IEEE International Conference on Systems, Man, and Cybernetics*, Nashville, TN, USA.
- Limbu, D.K., Pears, R., Connor, A., & MacDonell, S. (2006). Contextual and concept-based interactive query expansion. *Proceedings of the 19th Annual Conference of the National Advisory Committee on Computing Qualifications*, Wellington, New Zealand.
- Lin, H.C., Wang, L.H., & Chen, S.M. (2006). Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques. *Expert Systems with Applications*, 31(2), 397-405.
- Pitkow, J., Schutze, H., Cass, T., Cooley, R., & et al. (2002). Personalized search. *Communications of the ACM*, 45(9), 50-55.
- Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G. (Ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*, 313-323.
- Speretta, M., & Gauch, S. (2005). Personalized search based on user search histories. *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*, pp 622-628.
- Spink, A. (1997). Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science and Technology*, 48(5), 382-394.
- Spink, A., & Losee, R.M. (1996). Feedback in information retrieval. *Annual Review of Information Science and Technology*, 31, 33-78.
- Teevan, J., & Dumais, S. (2005). Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th international SIGIR conference on research and development in information retrieval*, 449-456.
- Vieira, V., Tedesco, P., Salgado, A.C., & Brézillon, P. (2007). Investigating the specifics of contextual elements management: the cemantika approach. *Context*, 493-506.
- Xu, J., & Croft, W.B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1), 79-112.

Bibliographic information of this paper for citing:

Bhatia, MPS, & Kumar, Akshi (2010). "Paradigm shifts: from pre-web information systems to recent web-based contextual information retrieval." *Webology*, **7**(1), Article 76. Available at: http://www.webology.org/2010/v7n1/a76.html

---