

# User-Centered Adaptive Information Retrieval

Xuehua Shen  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
xshen@cs.uiuc.edu

## ABSTRACT

Information retrieval systems are critical for overcoming information overload. A major deficiency of existing retrieval systems is that they generally lack user modeling and are not adaptive to individual users; information about the actual user and search context is largely ignored. For example, a tourist and a programmer may use the same word “java” to search for different information, but the current retrieval systems would return the same results.

In the proposed research, I will study the User-Centered Adaptive Information Retrieval (UCAIR), which aims at capturing and exploiting user context in the retrieval process. I propose a decision theoretic framework and develop techniques for implicit user modeling in information retrieval to improve retrieval accuracy. In the proposed new retrieval paradigm, the user’s search context plays an important role and the inferred implicit user model is exploited immediately to benefit the user. I also propose several context-sensitive retrieval algorithms based on statistical language models to combine user context information with the current query for better ranking of documents. Using these techniques, an intelligent client-side web search agent will be developed, which can perform eager implicit feedback, e.g., query expansion based on previous queries and immediate result reranking based on clickthrough information. I will study how to use TREC data set to create a test collection with search context information for quantitatively evaluating the proposed algorithms. I will also design and conduct the user study to further investigate the effectiveness of these algorithms in real applications.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

## General Terms

Algorithms, Experimentation, Human Factors

## Keywords

interactive retrieval, personalized search, user context, user model

## 1. INTRODUCTION

Although many information retrieval systems (e.g., web search engines and digital library systems) have been successfully deployed, the current retrieval systems are far from optimal. A major deficiency of existing retrieval systems is that they generally lack user modeling and are not adaptive to individual users [13]. This inherent non-optimality is seen clearly in the following two cases: (1) Different users may use exactly the same query (e.g., “Java”) to search for different information (e.g., the Java island in Indonesia or the Java programming language), but existing IR systems return the same results for these users. Without considering the actual user, it is impossible to know which sense “Java” refers to in a query. (2) A user’s information needs may change over time. The same user may use “Java” sometimes to mean the Java island in Indonesia and some other times to mean the programming language. Without recognizing the search context, it would be again impossible to recognize the correct sense.

Thus using user context information about the user and the query is necessary for improving the retrieval performance. Indeed, context-sensitive information retrieval essentially boils down to capturing and exploiting related user context information of a query to improve search accuracy.

|          | Short term (dynamic)        | Long term (static)  |
|----------|-----------------------------|---------------------|
| Implicit | immediately viewed document | past query log      |
| Explicit | judged relevant documents   | occupation, hobbies |

Table 1: Typology and examples of user context

As shown in Table 1, many kinds of user context information can be potentially exploited [7]. Explicit context consists of information given by a user explicitly, whereas implicit context refers to any context information naturally available while a user interacts with a retrieval system. Relevance feedback [15] can be considered as a way for a user to provide more context of search explicitly and is known to be effective for improving retrieval accuracy. While explicit context information is more reliable than implicit context, it is often not available to us because it requires extra effort from the user. Implicit context information is thus more interesting to exploit.

For this reason, *implicit feedback* has attracted much attention recently [4, 21, 18, 11]. In general, the retrieval results using the user’s initial query may not be satisfactory; often, the user would need to revise the query to improve the retrieval/ranking accuracy [6]. For a complex or difficult information need, the user may need to modify his query and view ranked documents with many iterations before the information need is completely satisfied. In such an interactive retrieval scenario, the information nat-

urally available to the retrieval system is more than just the current user query and the document collection – in general, all the interaction history can be available to the retrieval system, including past queries, information about which documents the user has chosen to view, and even how a user has read a document (e.g., which part of a document the user spends a lot of time in reading). We define implicit feedback broadly as exploiting all such naturally available interaction history to improve retrieval results.

While there are some existing work in the context-sensitive IR, it is far from a solved problem. In the proposed research, I will study the User-Centered Adaptive Information Retrieval (UCAIR), which aims at capturing and exploiting such implicit context at the client side to improve retrieval accuracy for a specific user. The proposed research includes the following thrusts: (1) A general decision-theoretic framework for context-sensitive retrieval will be developed for modeling text and user context information. (2) Specific retrieval models for exploiting search context based on statistical language model will be proposed to improve retrieval accuracy. (3) A client-side search agent will be built for context-sensitive search to evaluate the usability of the proposed algorithms. (4) Evaluation methodology of context-sensitive IR will be investigated, which includes TREC-style evaluation and the user study in real applications such as web search.

The rest of the paper is organized as follows. In Section 2, the related work in this area is presented. In Section 3, my current work is described, including the general framework for context-sensitive retrieval, specific context-sensitive retrieval models, a prototype of a client-side intelligent search agent and evaluation methodology of context-sensitive information retrieval. In Section 4, the future work is presented. In Section 5, I describe some issues which I am particularly interested in the discussion at the doctoral consortium.

## 2. RELATED WORK

In a recent workshop in 2002 about challenges in information retrieval and language modeling [1], personalized and contextual search is considered as one of the two grand challenges in information retrieval. There are some studies about user context, especially implicit feedback (see [12] for a bibliography of implicit feedback).

In [2], a special web browser *Curious Browser* is developed to record user actions on web pages (implicit feedback) including dwelling time, mouse click, mouse movement, scrolling and elapsed time and user explicit rating of web pages (relevance feedback). The results show that the dwelling time on a page, amount of scrolling on a page and the combination of time and scrolling have a strong correlation with explicit relevance ratings while the individual scrolling methods and mouse clicks are ineffective in predicting explicit interests. In [5], implicit feedback, particularly the combination of clickthrough, dwelling time and how a user exit a result or end a search session, is also found to be associated with user explicit rating. However, in [11], the effect of the task on the display time and potential impact of this relationship on the effectiveness of display time as the implicit feedback is studied. The results show that there is no general direct relationship between display time and usefulness. Moreover, the display time depends on the specific tasks and specific users.

In [19], authors use desktop search index as the user profile for personalized search. They consider the user profile as the implicit feedback and incorporate them into the ranking of web search results. They test different combination of corpus representation, user representation and document representation. It is found that the combination of personalized search ranking and original web ranking can achieve better results than the original ranking.

Following the work of studying how clickthrough data can be interpreted as implicit feedback [10], user search logs are partitioned into the query chains, from which relative relevance information is extracted to learn a better ranking formula of a library search system [14]. It is found that using evidence of query chains that is present in search engine logs can learn a better ranking formula compared with the traditional ranking formula and the ranking formula simply considering relative relevance from query logs [9].

In [18], authors use web browsing history in past  $N$  days for personalized search. They partition the browsing history data into three categories according to the time stamp, i.e., persistent data (before today), today data (today but before the current session) and current session data. They found that the performance of using web browsing history is competitive with that using relevance feedback.

In [21], from the unobtrusively tracked user interaction with the search system, several implicit feedback models (e.g., binary vote model and Jeffrey's condition model) are constructed based on the different weights of document representations (e.g. title and query dependent summary of relevant documents and top-ranked sentences extracted from top-ranked documents) and relevance path. The terms in the implicit feedback models are used to do query expansion.

The proposed retrieval framework integrates implicit user modeling with the interactive retrieval process, while the previous work either studies implicit user modeling separately from retrieval [2] or only studies specific retrieval models for exploiting implicit feedback to better match a query with documents [21]. The proposed research will also extend the existing research in several ways as mentioned in Section 1.

## 3. ACHIEVEMENTS SO FAR

### 3.1 A decision-theoretic framework for context-sensitive IR

To exploit context for personalized search in a general way, the retrieval problem is viewed as a decision problem, in which all contextual information and the normally available query and documents should be considered together to optimize the retrieval decision. In general, in response to every user action, the system would choose an optimal system action to take. For example, a user's action may be submitting a query and the system's response may be returning a list of 10 document summaries.

An advantage of treating retrieval generally as a decision-making problem is that we may also treat a user's viewing a document in the search results as a user action, to which the system can respond with updating its own user model about the user's information need. Although, in this case, such a response does not affect the user immediately, we may imagine that after the user views the document and returns to see more search results, the system can choose to rerank any unseen search results based on the updated user model. Indeed, to bring maximum benefit of context to the user, we would like to exploit context as soon as it is available and respond immediately based on any new piece of context information.

I propose a decision-theoretic framework for optimizing interactive information retrieval based on eager user model updating [17], in which the system responds to every user action by choosing some system action to optimize a utility function. Specifically, as soon as we observe any new piece of evidence from the user, the system would attempt to perform two tasks: (1) compute the current user model to update its belief about the user's information need (2) choose a response that minimizes a loss function. For example, immediately after the user views a document, we could use the knowledge that the viewed document summary is probably relevant

to rerank the unseen results so as to minimize a loss function that favors a decision to rank relevant documents above irrelevant ones.

In the traditional retrieval paradigm, the retrieval problem is cast as matching a query with documents and rank documents according to their relevance values. As a result, the whole retrieval process is a simple *independent* cycle of “query submission” and “result display”, which is inadequate for exploiting context. The decision-theoretic framework I propose generalizes this traditional retrieval paradigm and allows us to exploit the user’s search context in a quite general way.

### 3.2 Language models for context-sensitive IR

When instantiating the general decision-theoretic framework described above with specific retrieval methods, we obtain specific retrieval models that can rank documents based on search context. As a case study, I propose several different language models for using implicit feedback information in the same search session to improve retrieval accuracy in interactive information retrieval [16].

I use the KL-divergence retrieval model [22] as a basis and propose to treat context-sensitive retrieval generally as estimating a query language model based on the current query and any search context information. I proposed and tested several statistical language models to incorporate query and clickthrough history into the KL-divergence model, including linear interpolation with fixed coefficients, Bayesian interpolation, Online Bayesian updating and Batch Bayesian updating.

In general, the experiment results show that using implicit feedback information, especially the clickthrough data, can effectively and efficiently improve retrieval performance without requiring additional effort from the user at all [16].

### 3.3 A Context-Sensitive IR system – UCAIR Search Agent

A client-side search agent (called UCAIR) embedded in a web browser has been developed, which can capture a user’s search context and perform implicit feedback [17]. As shown in Figure 1, the UCAIR search agent has 3 major components: (1) The (implicit) user modeling module captures a user’s search context and history information and infers search session boundaries. Currently the user modeling module captures the submitted queries and any clicked search results. (2) The query modification module selectively improves the query formulation according to the current user model. (3) The result re-ranking module immediately re-ranks any unseen search results whenever the user model is updated. The UCAIR search agent incorporates models and algorithms proposed in Section 3.2 to dynamically rerank the search results to reflect the most updated knowledge of the user’s information need whenever any new piece of implicit feedback becomes available.

I chose to do context-sensitive IR at the client side instead of the server side as it has three remarkable advantages. First, the user does not need to worry about privacy infringement, which is a big concern for personalized search [20]. Second, a richer category of user interactions such as mouse movement can be easily captured for implicit feedback. Third, the computation needed for personalization and the storage of the user profile are both done at the client side, so the server is not burdened [8].

Specific techniques are implemented to capture and exploit two types of implicit feedback information: (1) identifying any related immediately preceding query and using the query and its corresponding search results to select appropriate terms to expand the current query, and (2) exploiting the viewed document summaries to dynamically rerank any document that has not yet been seen by the user.

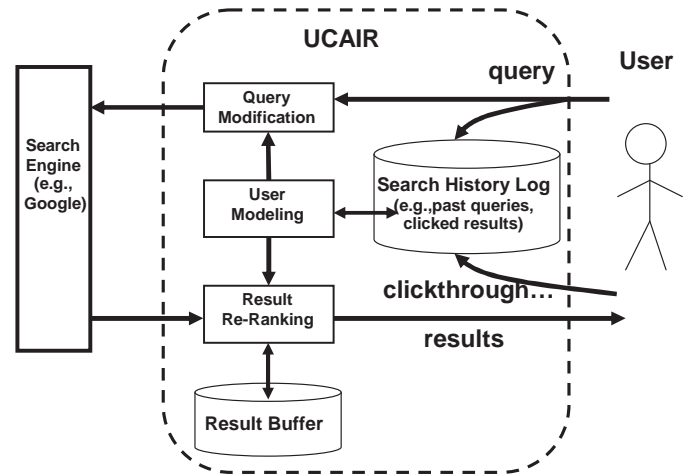


Figure 1: UCAIR architecture

User studies show that the UCAIR search agent improves performance over a popular search engine (Google), on which UCAIR search agent is built.

### 3.4 Evaluation of Context-Sensitive IR

Evaluation of context-sensitive IR poses special challenges due to the difficulty in collecting appropriate user interaction data and cleanly identifying baseline methods. For example, one challenge in evaluating implicit feedback algorithms is that there does not exist any suitable test collection for evaluation. In my study, I used the TREC AP data to create a test collection with implicit feedback information that can be used to quantitatively evaluate implicit feedback algorithms. To the best of my knowledge, this is the first test set for implicit feedback [16].

When evaluating the UCAIR search agent, I conducted a user study involving 6 people. The participants are asked to do a web search on selected query topics from TREC 2004 Terabyte track and TREC 2003 Web track topic distillation task and then make relevance judgments of the search results. By comparing our ranking that incorporates context information and Google’s original ranking, we can see whether the use of context information is beneficial. Such a method [17] can be applicable to evaluating similar context-sensitive retrieval systems.

## 4. FUTURE WORK

I plan to do my thesis research as follows.

### 4.1 Framework and Retrieval Models

The general decision-theoretic framework for context-sensitive IR will be further studied. I will also further explore retrieval models to incorporate useful context information into the retrieval process. For example, we may treat a clicked document summary differently depending on whether the current query is a generalization or refinement of the previous query. So far, I have only explored some very simple language models for incorporating implicit feedback information. Another interesting research question is how to optimize some parameters in the context-sensitive retrieval algorithms. Currently, algorithm parameters are pre-specified according to previous experiments, which is apparently not optimal for all retrieval problems. Some optimization methods such as EM algorithm may be applied to learning optimal parameter setting.

## 4.2 Capturing and Exploiting User Context

I will study other important user interactions. At the client side, UCAIR search agent can capture and exploit many other user actions such as mouse movement and dwelling time on a document, which may have strong correlation with the document's relevance [2, 5]. It is interesting to study how to model the non-textual context information dwelling time in retrieval models. I plan to implement the proposed algorithms in the UCAIR search agent and evaluate the effectiveness and usability of the algorithms by conducting the user study.

## 4.3 Non-intrusive Personalization

I will study non-intrusive personalization. Some existing research find that sometimes personalized search is intrusive, i.e., sometimes personalization can hurt the retrieval performance. Thus it is interesting to study whether we can propose an algorithm to do personalization in the non-intrusive way. Three research questions need to be addressed: 1) how do we define and measure the intrusiveness? 2) whether can we propose an algorithm to predict the retrieval performance of personalized search? 3) whether can we propose a personalized search algorithm which utilizes appropriate user context information and at the same time guarantees the retrieval performance? In [3], clarity score is proposed to measure the ambiguity of the query. Clarity of a query is defined as the KL divergence between the collection distribution and the query distribution. In a similar way, we may measure whether a query is good for the personalization or not using statistical language model. The clues of possible good personalization include very diversified search results and some search results are similar with the user models while other search results are not.

## 4.4 Long-term User Context

I will study whether long-term user context can be exploited to improve retrieval performance. So far, my work has only studied the short-term search context. However, there is some evidence such as [18, 5] which shows that the long-term user context can also be useful for improving retrieval accuracy. I plan to focus on whether we can extract *relevant* context information from the long-term user context and how to model the extracted relevant context information into retrieval models.

## 5. ISSUES TO DISCUSS

For the extension of my current work, I am interested in exploring the possible improvement including the retrieval framework proposed in [17], the context-sensitive retrieval models proposed in [16] and the intelligent search agent [17]. I am also interested in discussing the evaluation methods and evaluation metrics for the context-sensitive information retrieval.

For the future research direction mentioned in Section 4, I am particularly interested in seeking suggestions about the possible research direction(s). For example, which directions are more interesting to explore? Whether are there some similar works that I should study? What research questions should I address in these directions? I am also interested in the suggestion about new directions which I should explore but do not mention in Section 4.

## 6. REFERENCES

- [1] J. Allan et al. Challenges in information retrieval and language modeling: Report of a workshop held at the center for intelligent information retrieval, University of Massachusetts at Amherst, September 2002. *SIGIR Forum*, 37(1):31–47, 2003.
- [2] M. Claypool, P. Le, M. Waseda, and D. Brown. Implicit interest indicators. In *Proceedings of Intelligent User Interfaces 2001*, pages 33–40, 2001.
- [3] W. B. Croft, S. Cronen-Townsend, and V. Larvrenko. Relevance feedback and personalization: A language modeling perspective. In *Proceedings of Second DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.
- [4] L. Finkelstein, E. Gabrilovich, Y. Matias, et al. Placing search in context: The concept revisited. In *Proceedings of WWW 2001*, 2001.
- [5] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transaction of Information System*, 23(2):147–168, 2005.
- [6] C.-K. Huang, L.-F. Chien, and Y.-J. Oyang. Query session based term suggestion for interactive web search. In *Proceedings of WWW 2001*, 2001.
- [7] P. Ingwersen and N. Belkin. Information retrieval in context – IRiX. *SIGIR Forum*, 38(2), 2004.
- [8] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of WWW 2003*, pages 271–279, 2003.
- [9] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of SIGKDD 2002*, pages 133–142, 2002.
- [10] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR 2005*, 2005.
- [11] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proceedings of SIGIR 2004*, 2004.
- [12] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, 2003.
- [13] G. Nunberg. As google goes, so goes the nation. *New York Times*, May 2003.
- [14] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proceedings of SIGKDD 2005*, 2005.
- [15] J. J. Rocchio. *Relevance Feedback Information Retrieval*, pages 313–323. Prentice-Hall, 1971.
- [16] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of SIGIR 2005*, pages 43–50, 2005.
- [17] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proceedings of CIKM 2005*, 2005.
- [18] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of WWW 2004*, pages 675–684, 2004.
- [19] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of SIGIR 2005*, 2005.
- [20] E. Volokh. Personalization and privacy. *Communications of the ACM*, 43(8):84–88, 2000.
- [21] R. W. White, J. M. Jose, C. J. van Rijsbergen, and I. Ruthven. A simulated study of implicit feedback models. In *Proceedings of ECIR 2004*, pages 311–326, 2004.
- [22] C. Zhai and J. Lafferty. Model-based feedback in KL divergence retrieval model. In *Proceedings of the CIKM 2001*, pages 403–410, 2001.

## **Motivation for Attending Doctoral Consortium**

The Ph.D. study is a long journey of doing research and pursuing knowledge. Currently I am writing the thesis proposal and plan to defend the thesis next year.

Besides the discussion of my research work with my advisor and colleagues, I like to communicate my research with people and listen to their comments and suggestions. Doctorial Consortium of SIGIR conference provides a perfect opportunity for me to discuss my research idea with experienced IR researchers and other doctoral students.

My thesis work includes the user models and the formal retrieval models. The consortium provides a special forum to discuss my research idea with experts in both areas together. Moreover, I have very similar research interests with some researchers attending previous consortia. Thus I believe that I can learn from the researchers and students in the consortium, which will help me pursue the high-quality research.

The doctoral consortium will also provide me an opportunity to discuss the research of other doctoral students, through which I can potentially help other people improve their work.

## **Statement from my advisor Professor ChengXiang Zhai**

Xuehua Shen has been working on information retrieval with me for about 3 years. His main research topic is personalized search and has made good progress in this direction. He is currently writing his dissertation proposal and is expected to finish it soon (in a month or so). Thus the timing for him to attend SIGIR 06 Doctoral Consortium is perfect.

Personalized search poses challenges in several sub-areas in information retrieval, including user modeling, context-sensitive retrieval models, machine learning, user studies, and evaluation in general. While Xuehua has already made good progress in his thesis research (including a SIGIR paper, CIKM paper, and a prototype personalized search agent), to finish his thesis, he still needs to address large-scale user studies, non-intrusive personalization, and personalized search result organization. Many research questions are quite unstructured, thus he will be able to benefit a lot from discussing his thesis research with experienced IR researchers and many other students. Feedback from all these related areas would significantly help him in further finishing his Ph.D. thesis on personalized search.

Xuehua will pursue an academia career. So opportunities for him to have one-on-one interactions with experienced IR researchers will also be quite beneficial for him.

I strongly recommend Xuehua Shen to attend SIGIR 06 Doctoral Consortium.