Exploiting Personal Search History to Improve Search Accuracy

Xuehua Shen
Department of Computer Science
University of Illinois at Urbana-Champaign

Bin Tan
Department of Computer Science
University of Illinois at Urbana-Champaign

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

ABSTRACT

Personal search history is an important type of personal information, from which we can learn a user's interests and information needs, thus improve the search service for the user. In this paper, we describe our recent work on User-Centered Adaptive Information Retrieval (UCAIR), which aims at capturing personal search history with a client-side search agent and exploiting the history information to help a user optimize search results.

We propose a decision theoretic framework and develop techniques for implicit user modeling based on a user's personal search history. We propose several context-sensitive retrieval algorithms based on statistical language models to combine the personal search history with the current query for better ranking of documents. Using these techniques, we have developed an intelligent client-side web search agent, i.e., the UCAIR search agent, which can automatically capture a user's personal search history, store it on the local disk, and exploit it to provide personalized search.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Human Factors

Keywords

personal search history, personalized search, user context, user model

1. INTRODUCTION

With low cost, high capacity disks available, it becomes feasible to store almost all personal information including text, audio and video on the user's local hard disk [4]. Several applications, such as Google desktop search and Stuff I've Seen [2] have recently been developed to help a user find or refind information on personal computers. Indeed, research of Personal Information Management (PIM), which intends to realize Vannevar Bush's Memex [1] vision of turning personal information collection, storage, sharing and organization into an important aspect of the life of everyone with a computer [14], has been attracting much attention in recent years.

Search is an important activity in many people's daily lives, and personal search history is the log of all search activities of a user. During the interaction with a retrieval system, an individual user naturally generates valuable personal search history data, such as the submitted queries and clicked web pages. From such personal search history, we can learn much about the user. For example, we may infer what topic is interesting to the user and what products the user is interested in purchasing. Search is also strongly correlated with many other user activities such as news reading and document editing; the personal search history can thus also provide traces of the individual's other activities. More importantly, however, personal search history provides direct clues about a user's interests and information needs, thus can be exploited to improve the search service for a user.

Consider the ambiguous query "IR applications", in which "IR" can stand for "information retrieval" or "infrared". Just based on the query text, it is impossible to determine what the user is actually interested in. However, any of the following additional information in the personal search history could potentially help determine the intended meaning of the acronym "IR" in the query: (1) The previous query that the user entered is "Web search engines". (2) The user was browsing an information retrieval conference homepage before entering this query.

Therefore, personal search history is an important part of personal information and should be properly stored, managed, and exploited. However, to our best knowledge, there has so far been little work on managing and exploiting this important personal information. In this paper, we present our work on the User-Centered Adaptive Information Retrieval (UCAIR) project and discuss some future research directions in exploiting a user's personal search history.

The UCAIR project (http://sifaka.cs.uiuc.edu/ir/ucair/) aims at capturing and exploiting personal search history on the client side to improve retrieval accuracy for a specific user. In the rest of the paper, we will present some of our work on this project, including (1) a general decision-theoretic framework for modeling user context information and supporting context-sensitive retrieval, (2) specific retrieval models for exploiting search context to improve search accuracy based on statistical language models, and (3) a client-side search agent, UCAIR search agent. The UCAIR search agent captures and stores personal search history on a user's local computer. Unlike Google's personalized search service, which stores the personal search history at the server side, the UCAIR search agent stores the data at the client side to give the user full

control and reduce the user's privacy concerns. Moreover, the UCAIR search agent can also capture additional personal data such as personal browsing history, which would be hard to capture from the server side. Furthermore, the UCAIR search agent can potentially accommodate collaborative search.

2. A DECISION-THEORETIC FRAMEWORK FOR OPTIMAL INTERACTIVE RETRIEVAL

In order to exploit personal search history to improve search accuracy in a general way, we view information retrieval (IR) as a decision optimization problem and propose a formal decision-theoretic framework based on Bayesian decision theory for optimizing interactive retrieval [10].

Information retrieval is inherently an interactive process, in which a user would iteratively reformulate queries and view documents [3, 5]. In interactive IR, a user interacts with the retrieval system through an "action dialogue", in which the system responds to each user action with some system actions. For example, the user's action may be submitting a query and the system's response may be returning a list of ten document summaries. In general, the space of user actions and system responses and their granularities would depend on the interface of a particular retrieval system.

In principle, every action of the user can potentially provide new evidence to help the system better infer the user's information need. Thus in order to respond optimally, the system should use *all* the evidence collected so far about the user when choosing a response. When viewed in this way, most existing search engines are clearly non-optimal. For example, if a user has viewed some documents on the first page of search results, when the user clicks on the "Next Page" link to fetch more results, an existing retrieval system would still return the next page of results retrieved based on the original query without considering the new evidence that a particular result has been viewed by the user.

We propose to optimize retrieval performance by adapting system responses based on *every* action that a user has taken, and cast the optimization problem as a decision task. Specifically, at any time, the system would attempt to do two tasks: (1) updating user model updating: monitor any useful evidence from the user regarding his information need and update the user model as soon as such evidence is available; (2) improving search results: immediately rerank all the documents that the user has not yet seen as soon as the user model is updated. We emphasize eager updating and reranking, which makes our work quite different from any existing work.

Note that in a traditional retrieval framework, the retrieval problem is often taken as matching a query with documents and ranking documents according to their relevance values. As a result, the whole retrieval process is a simple independent cycle of "query" and "result display". In the proposed new retrieval framework, the user's search context plays an important role and the inferred implicit user model is exploited immediately to benefit the user. The new retrieval framework is thus fundamentally different from the traditional framework, and is inherently more general.

In [10], we formalized these ideas and proposed a formal decision theoretic framework for optimizing retrieval performance through implicit user modeling in interactive information retrieval. Using this framework, we could derive language models for exploiting both short-term and long-term search history to improve search accuracy.

3. LANGUAGE MODELS FOR EXPLOITING PERSONAL SEARCH HISTORY

3.1 Exploiting short-term search history

For an "ad hoc" information retrieval task, which only lasts for a short period of time, once the information need is satisfied, the user would generally no longer be interested in such information. For example, a user may be looking for information about used cars in order to buy one, but once the user has bought a car, he/she is generally no longer interested in such information. In such cases, the long-term personal search history collected over a long period of time is unlikely to be very useful, but the short-term search history can be much more useful.

In [8, 9], we studied how to construct and update a user model based on the *immediate* search context and implicit feedback information and use the model to improve the accuracy of ad hoc retrieval. In order to *maximally* benefit the user of a retrieval system through implicit user modeling, we perform "eager implicit feedback". That is, as soon as we observe any new piece of evidence from the user, we would update the system's belief about the user's information need and respond with improved retrieval results based on the updated user model.

We further propose specific techniques to capture and exploit two types of implicit feedback information: (1) identifying related immediately preceding query and using the query and the corresponding search results to select appropriate terms to expand the current query, and (2) exploiting the viewed document summaries to immediately rerank any documents that have not yet been seen by the user. We proposed several context-sensitive retrieval algorithms based on statistical language models to combine the preceding queries and clicked document summaries with the current query for better ranking of documents. We use the TREC AP data to create a test collection with search context information, and quantitatively evaluate our models using this test set. Experiment results show that using short-term personal search history, especially the clicked document summaries, can improve retrieval performance substantially [8].

3.2 Exploiting long-term search history

Several existing studies [13, 11, 7] have explored various kinds of user context information to improve search results. However, no work has systematically explored how to exploit the long-term personal search history to improve search results. Work on exploiting short-term context has simply ignored the long-term context, while some work that does consider long-term search history (e.g., [13]) has only studied the usefulness of such history information, leaving many research questions unanswered (e.g., how to combine click-through information with the search results and how to optimize the weights on the history data). The evaluation in most previous work also tends to rely on having users judge some pre-designed topics.

We systematically studied how to exploit long-term search history to improve search accuracy [12]. Long-term search history includes not only the very useful short-term context but also other useful information in remote context. For example, a user may have some relatively stable long-term interests, and in such a case, we will have recurring queries in the history that can clearly be exploited to improve the results for the current query. Even if the current query is a fresh query that has not be seen before, the history information may still be useful because it may contain some related queries. At least, such long-term search history of a user would be useful for learning his general preferences (e.g., computer science terms occurring more frequently than medical terms in the history), which can be potentially used to improve the results for the current query.

Mining and exploiting the personal long-term search history is

a challenging task. Although the long-term search history clearly contains useful information that can help improve the results for the current query, the history also has a lot of noise and it is not immediately clear how we can extract the most useful information and at the same time avoid introducing noise or distracting information. In this sense, the problem we study is more challenging than the one involving only immediate search history which is less noisy. We propose mixture models to represent a user's information need and apply statistical language models to mine the search history of a user for relevant context information. We propose several retrieval algorithms to improve document ranking by combining such relevant context information with the current query to obtain an improved estimate of the query language model [12].

We collect the personal long-term search history from a group of users' daily web search activities. We quantitatively evaluate these algorithms and study the effects of several parameters such as the time cutoff and importance of several components of search history. We also analyze the different behavior of our algorithms on fresh and recurring queries. Experiment results show that the proposed search history mining algorithms can effectively personalize search results and improve retrieval accuracy for both recurring and fresh queries, though recurring queries naturally benefit more from history mining. We find that more recent history information tends to be more beneficial than remote history information and using the entire history achieves the best performance. We also find clicked document summaries to be more useful than other history information [12].

4. UCAIR SEARCH AGENT

In this section, we present a client-side web search agent called UCAIR, in which we implement some of the methods discussed in the previous sections for performing personalized search by capturing and exploiting personal search history. UCAIR is a web browser plug-in [1] that acts as a proxy for web search engines. Currently, it is only implemented for Internet Explorer interacting with Google or Yahoo, but it is a matter of engineering to make it run on other web browsers and interact with other search engines.

The issue of privacy is a primary obstacle for deploying any real world applications involving serious user modeling, such as personalized search. For this reason, UCAIR is strictly running as a client-side search agent, as opposed to a server-side application. This way, the captured user information always resides on the computer that the user is using, thus the user does not need to release any information to the outside. Client-side personalization also allows the system to easily observe a lot of user information that may not be easily available to a server. Furthermore, performing personalized search on the client-side is more scalable than on the server-side, since the overhead of computation and storage is distributed among clients.

As shown in Figure 1, the UCAIR toolbar has three major components: (1) the (implicit) user modeling module captures a user's search history information, including the submitted queries and any clicked search results and infers search session boundaries; (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.

When a user submits a query through the UCAIR, the query is stored in a log file on the local disk, and at the same time sent to the search engine. UCAIR then retrieves the search result pages from the search engine. Before the results are presented to the user, the session boundary detection algorithm is invoked to decide whether the current query and the previous query belong to the same information session, and if so, a query expansion algorithm is executed to select query terms from the previous query and the previous search results to do query expansion. The search results using a possibly expanded query are then presented to the user. When the user clicks on any search result, the unseen results are reranked according to the learned implicit user model based on the clicked search result. When the user clicks on "Next Page" to fetch more results, those results can be further reranked based on any additional information about the user.

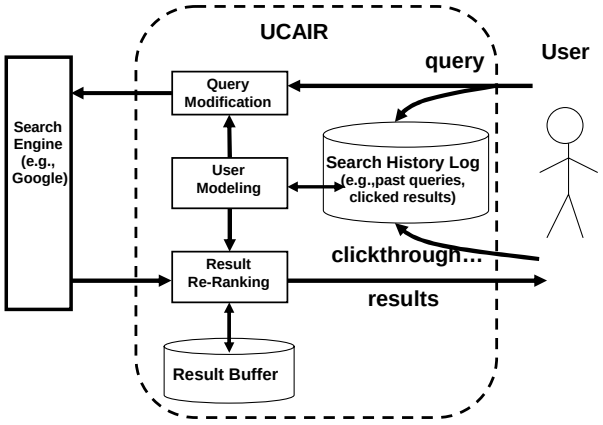

Figure 1: UCAIR architecture

In UCAIR, we consider four basic user actions: (1) submitting a keyword query, (2) viewing a document, (3) clicking on the "Back" button, and (4) clicking the "Next Page" link on a result page. For each of these four actions, the system responds with, respectively, (1) generating a ranked list of results by sending a possibly expanded query to a search engine, (2) updating the information need model, (3) reranking the unseen results on the current result page based on the current model, and (4) reranking the unseen pages and generating the next page of results based on the current model.

Behind these responses, there are three basic tasks: (1) decide whether the previous query is related to the current query and if so expand the current query with useful terms from the previous query or the results of the previous query, (2) update the information need model based on a newly clicked document summary, and (3) rerank a set of unseen documents based on the current model.

In Figure 2, we show how the UCAIR search agent can successfully disambiguate an ambiguous query "jaguar" by exploiting a viewed document summary. In this case, the initial retrieval results using "jaguar" (shown on the left side) contain two results about the Jaguar cars followed by two results about the Jaguar software. However, after the user views the web page content of the second result (about "Jaguar car") and returns to the search result page by clicking "Back" button, UCAIR automatically nominates two new search results about Jaguar cars (shown on the right side), while the original two results about Jaguar software are pushed down on the list (unseen from the picture). Quantitative evaluation on web search [9] show that our search agent can improve search accuracy over Google.

5. FUTURE WORK

[1] UCAIR is available at: http://sifaka.cs.uiuc.edu/ir/ucair/download.html

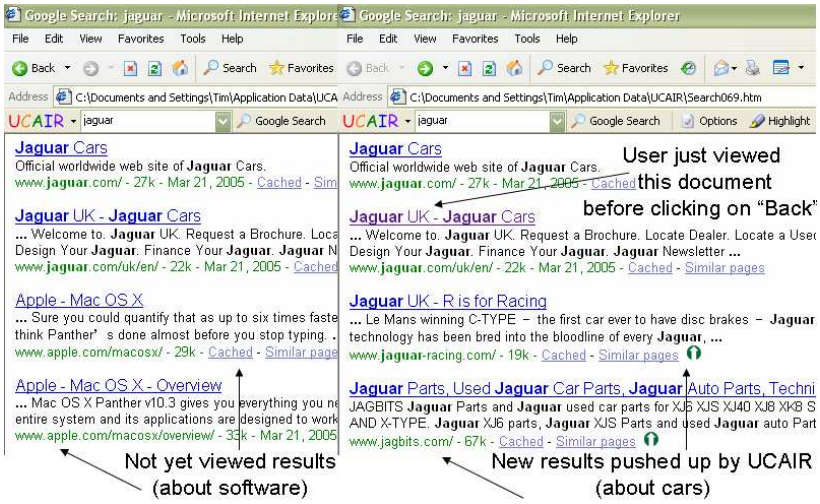

Figure 2: Screen shots for result reranking

We believe that exploiting personal search history to learn a user's interests and information needs and to improve search accuracy for the user is an important and promising direction. The UCAIR project has been going for only about 2 years, but our preliminary studies and results already led to useful algorithms and a useful system. Our current work can be extended in several ways:

First, we will study algorithms to automatically decide whether a specific query can achieve better performance using the personal search history or not. Using personal search history does not always improve the retrieval accuracy. We will study how a retrieval system can decide whether the personal search history can be exploited to improve retrieval accuracy.

Second, we will take a broader view of user context. So far, we have only exploited the personal search history for the personalized search. However, personal information is much more than that. For example, a lot of useful information such as emails, calender items and WORD documents is stored in the personal computer. All such personal information can potentially be exploited to improve retrieval accuracy. There is some work such as [13] in this direction. We will use machine learning techniques and mixture language models to mine useful information from these noisy contexts, which otherwise may not help improve the retrieval accuracy [6].

Third, we will go beyond one person's personal search history and study how to exploit the search history of a group of users to perform collaborative search. Through some initial study, we find that a group of users such as peers in a research group often share some similar information needs. We will study whether we can leverage the search history from other similar people to improve retrieval accuracy. How to preserve privacy in collaborative search is a very important issue that clearly needs to be addressed.

6. REFERENCES

[1] V. Bush. As we may think. *The Atlantic Monthly*, 1945.

[2] S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of SIGIR 2003*, pages 72–79, 2003.

[3] L. Finkelstein, E. Gabrilovich, Y. Matias, et al. Placing search in context: The concept revisited. In *Proceedings of WWW 2001*, 2001.

[4] J. Gemmell, G. Bell, R. Lueder, S. M. Drucker, and C. Wong. Mylifebits: Fulfilling the memex vision. In *Proceedings of ACM Multimedia 2002*, pages 235–238, 2002.

[5] P. Ingwersen and N. Belkin. Information retrieval in context – IRiX. *SIGIR Forum*, 38(2), 2004.

[6] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proceedings of SIGIR 2004*, 2004.

[7] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proceedings of SIGKDD 2005*, 2005.

[8] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of SIGIR 2005*, pages 43–50, 2005.

[9] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proceedings of CIKM 2005*, 2005.

[10] X. Shen, B. Tan, and C. Zhai. UCAIR toolbar: A personalized search toolbar (Demo). In *Proceedings of SIGIR 2005*, page 681, 2005.

[11] 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.

[12] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of SIGKDD 2006*, 2006.

[13] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activites. In *Proceedings of SIGIR 2005*, 2005.

[14] J. Teevan, W. Jones, and B. B. Bederson. Special issue on personal information management. *Communications of the ACM.*, 49(1):40–43, 2006.