

# A Session-Based Search Engine

Smitha Sriram, Xuehua Shen, Chengxiang Zhai  
Department of Computer Science  
University of Illinois at Urbana-Champaign

## ABSTRACT

In this poster, we describe a novel session-based search engine, which puts the search in context. The search engine has a number of session-based features including expansion of the current query with user query history and clickthrough data (title and summary of clicked web pages) in the same search session and the session boundary recognition through temporal closeness and probabilistic similarity between query terms. In addition, the search engine visualizes the rank change of web pages as different queries are submitted in the same search session to help the user reformulate the query.

## Categories and Subject Descriptors

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

## General Terms

Algorithms

## Keywords

Query history, query expansion, context

## 1. INTRODUCTION

In web information retrieval, users often have to modify their queries several times before they can reach a page that meets their information need. During these interactions with the search engine, the user provides a lot of useful information to the search engine, which can be exploited to infer the user information need. For instance, if a user types “lemur” as the query, most of the pages returned by the search engine are about the animal lemur or lemur information retrieval toolkit. On the other hand, if the search engine incorporates the previous query submitted by the same user, say, “information retrieval”, the search engine can disambiguate the meaning of “lemur” and present pages that are related to lemur information retrieval toolkit. We design and implement a session-based search engine, which puts the search in context. Although

there have been several attempts at building a personalized or contextual search engine[?] or session based search engines [?][?], our search engine has the following new features:

- Incorporation of title and summary of clicked web pages and past queries in the same search session to update the query.
- Recognition of session boundary using temporal closeness and probabilistic similarity between queries.
- Visualization of rank change of each web page with different queries in the same search session.

## 2. SYSTEM DESCRIPTION

We crawl, parse and index web pages of a small domain ( Computer Science Department) and run our search engine on this text database. The search engine can operate in two modes, i.e., the session mode and tradition mode. In the session mode, the user requests that the current search activities be recorded and past queries and clickthrough information in same search session be used in the current query update. In the tradition mode, the search engine does not use any contextual information. The architecture of the search engine is illustrated in Figure ???. Arrows show the flow of information.

Figure 1: Architecture of the Search Engine

In the session mode, the search engine uses a relational database management system (MySQL in our system) to store user query history and clickthrough information(title and summary of clicked web pages). We use the IP address to differentiate the user and assume that in a short time period only one user is using this computer. The search engine is based on Lemur Toolkit and the retrieval model is KL Divergence retrieval model[?].

## 3. SESSION-BASED RETRIEVAL FEATURES

### 3.1 Session Boundary Recognition

In order to use the appropriate past queries and clickthrough information in the expansion of the current query, the search engine has to differentiate queries among different sessions. In addition, sometimes during a search session, the user will submit a query that is motivated from a different information need. In this case, using past queries in a session solely based on time will introduce irrelevant terms and affect the retrieval performance. In order to recognize the end of a previous session and the beginning of the current

session, we propose a method using the Jensen-Shannon(JS) Divergence to compute the similarity between two terms. JS Divergence compares the probability distribution of two terms in a query over the set of web pages in the collection. This quantified difference gives us the similarity measure between two queries, thus aiding in boundary recognition between two sessions (we choose a threshold and believe the queries are from the same session if the similarity is above the threshold). The similarity between two queries is computed as follows.

$$\sum_i \sum_j JS(t_i, t_j)$$

where  $t_i$  and  $t_j$  are terms from query  $q_1$  and  $q_2$  respectively.

### 3.2 Session-Based Query Updating

If  $q_1, q_2, \dots, q_k$  are the queries submitted by the user in a particular session and the current query is  $q_k$ . The search engine will expand the current query  $q_k$  using  $q_1, q_2, \dots, q_{k-1}$  to form a new query  $q'$ , which provides more clue about the user information need. We propose a query history based query model [?] using the maximum likelihood estimate and a decaying factor to reduce the importance of terms far away from the current query with respect to time. The probability of word  $w$  according to this model is given as follows.

$$p(w|q') = p(w|q_1, \dots, q_k) \propto \frac{1}{k} \sum_{i=1}^k \frac{c(w, q_i)}{|q_i|} (1 - \alpha)^{k-i} \alpha$$

where  $\alpha$  is the decay factor to reduce the importance of past queries and  $k-i$  is the distance of the query from the current query.

Another important piece of information that is captured during a session is the title and summary of web pages clicked by the user. We assume that a web page is clicked because the user has read the title and summary of the web page and found that the document is useful. We augment our query model as follows.

$$p(w|q) \propto \beta p(w|q') + \frac{1 - \beta}{k - 1} \sum_{i=1}^{k-1} \frac{c(w, s_i)}{|s_i|} (1 - \alpha)^{k-i} \alpha$$

where  $s_i$  is the combination of the title and summary of clicked web page for query  $q_i$  and  $\beta$  is the weight given to past and current query model.  $\alpha$  is the decay factor to reduce the importance of past titles and summaries.

### 3.3 Visualization of Web Page Rank Change

A new user interface feature of our system is the ability to visualize the rank change of web pages during the current session. After receiving the user query, the search engine returns a list of top ranked documents based on the session-based retrieval algorithm. Along with the title and summary of web pages, users can view the rank change for each page as they change their queries. As the session progresses, they will be able to see how their query modifications have affected the ranking of each web page which helps them reformulate their query more accurately.

## 4. EVALUATION

No existing data is available for testing context-based search engine. Evaluation of the effectiveness of our search engine needs a large scale of quantitative user study. The design of such an experiment need consider factors such as accuracy of context provided by the query updating algorithm and session boundary recognition

algorithm. So far this kind of evaluation has not been done because of the limitation of resources. We did some study for AP data [?], which shows the improvement using query history to expand the query. Some preliminary study for the department web search has been done, which shows that the search engine performs better than the current department search engine, which has no session based retrieval. A specific information need example is information on the master program offered by the department. The following queries were submitted consecutively - *graduate program, masters*. If we do not capture the previous query (graduate program) submitted by the user, neither of search engines will return correct web page. But if we incorporate the previous query, the correct web page will be ranked on the top. Another example is to find course web site using mere description of the courses. Since most students do not remember the course number of most courses, they tend to search using keywords describing the course or the name of the professor who teaches this course. When the search engine modifies the query using the previous query as contextual cues, the user will obtain the relevant web page in the top ranked list. In both cases, it takes students lesser time to find relevant web pages using our session-based search engine.

## 5. CONCLUSION AND FUTURE WORK

In this poster, we describe a novel session-based search engine using past queries and clickthrough information in the same session. We also propose an algorithm to recognize the session boundary in order to prevent unrelated queries from interfering in web page ranking. Along with the title and summary of each web page, the search engine visualizes the rank change of each web page with different queries in the same search session.

We have only conducted a preliminary evaluation of the system. It is important to conduct a thorough quantitative evaluation. We will also study the efficiency of the search engine and test whether the session-based search engine can be deployed in a large domain.

## 6. REFERENCES

- [1] X. Shen, C. Zhai *Exploiting Query History for Document Ranking in Interactive Information Retrieval*. SIGIR Toronto, Canada, 2003.
- [2] C. Huang, L. Chien, Y Oyang *Relevant term suggestion in interactive web search based on contextual information in query session logs*. Journal of the American Society for Information Science and Technology Volume 54 , Issue 7, May 2003.
- [3] S. Dumais, E. Cutrell, JJ Cadiz, G. Jancke, R. Sarin, D. C. Robbins *Stuff I've Seen: A System for Personal Information Retrieval and Re-Use*. SIGIR Toronto, Canada, 2003.
- [4] C. Zhai and J. Lafferty *Model-based Feedback in KL Divergence Retrieval Model*. Proceedings of the 10th International Conference on Information and Knowledge Management(CIKM)", 2001.