ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Inferring User Search Goals for a Query by Clustering the Feedback Sessions

Shaik.Masthan
Post Graduate Student,
Department of CSE,
KSRM College.

K.Srinivasa Rao
Associate Professor,
Department of CSE,
KSRM College.

ABSTRACT:

Different users may have different search goals when they submit it to a search engine. If we know the user search goals means we can easily improve their searching and user experience. In this paper, we propose a novel approach to infer user search goals by analyzing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs.

1. Introduction:

In web search applications, queries are submitted to searchengines to represent the information needs of users.However, sometimes queries may not exactly representusers' specific information needs since many ambiguousqueries may cover a broad topic and different users maywant to get information on different aspects when theysubmit the same query. For example, when the query "thesun" is submitted to a search engine, some users want tolocate the homepage of a United Kingdom newspaper,while some others want to learn the natural knowledge ofthe sun, as shown in Fig. 1. Therefore, it is necessary andpotential to capture different user search goals in information retrieval. We define user search goals as the information on different aspects of a query that user groups want toobtain. Information need is a user's particular desire toobtain information to satisfy his/her need.

The inference and analysis of user search goalscan have a lot of advantages in improving search engineelevance and user experience. Some advantages aresummarized as follows. First, we can restructure web search results according to user search goalsby grouping the search results with the same search goal;thus, users with different search goals can easily find whatthey want. Second, user search goals represented by somekeywords can be utilized in query recommendation; thus, the suggested queries can help users to form theirqueries more precisely. Third, the distributions of usersearch goals can also be useful in applications such asreran king web search results that contain different usersearch goals.

In this paper, we aim at discovering the number ofdiverse user search goals for a query and depicting eachgoal with some keywords automatically. We first propose anovel approach to infer user search goals for a query byclustering our proposed feedback sessions. The feedbacksession is defined as the series of both clicked andunclicked URLs and ends with the last URL that wasclicked in a session from user click-through logs. Then, wepropose a novel optimization method to map feedbacksessions to pseudo-documents which can efficiently reflectuser information needs. At last, we cluster these pseudodocuments to infer user search goals and depict them withsome keywords. Since the evaluation of clustering is alsoan important problem, we also propose a novel evaluationcriterion classified average precision (CAP) to evaluate theperformance of the restructured web search results. Wealso demonstrate that the proposed evaluation criterion canhelp us to optimize the parameter in the clustering methodwhen inferring user search goals.To sum up, our work has three major contributions asfollows: We propose a framework to infer different user searchgoals for a query by clustering feedback sessions. Wedemonstrate that clustering feedback sessions is moreefficient than clustering search results or clickedURLs directly.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered. We propose a novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail. We propose a new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus, we can determine the number of user search goals for a query.

2. Literature Survey:

2.1. Context-Aware Query Suggestion by Mining Click-Through and Session Data:

Query suggestion plays an important role in improving the usability of search engines. Although some recently proposed methods can make meaningful query suggestions by mining query patterns from search logs, none of them are context-aware – they do not take into account the immediately preceding queries as context in query suggestion. We test our approach on a large-scale search log of a commercial search engine containing 1.8 billion search queries, 2.6 billion clicks, and 840 million query sessions. The experimental results clearly show that our approach outperforms two baseline methods in both coverage and quality of suggestions. In this paper, we proposed a novel approach to query suggestion using click-through and session data. Unlike previous methods, our approach considers not only the current query but also the recent queries in the same session to provide more meaningful suggestions. Moreover, we group similar queries into concepts and provide suggestions based on the concepts. The experimental results on a large-scale data containing billions of queries and URLs clearly show our approach outperforms two baselines in both coverage and quality.

2.2. Bringing Order to the Web: Automatically Categorizing Search Results:

We developed and evaluated a user interface that organizes search results into a hierarchical category structure. Support Vector Machine classifiers were built offline using manually classified web pages.

This approach has the advantage of leveraging known and consistent category information to assist the user in quickly focusing in on task-relevant information. In our current interface we have a "NotCategorized" group at the bottom. In our experiment 5-40% of the results for each query were NotCategorized, but few of the answers were in the NotCategorized group. We hope to deploy our system more widely to look at this issue by getting a large sample of typical user queries. This would also allow us to explore a wider range of user tasks in addition to the known-item scenario we used. We chose to order categories by the number of matches and within each category to order the pages by search rank. Our text classification algorithms can easily handle thousands of categories, and we may have to move beyond our simple display heuristics for such cases.

2.3. Optimizing Search Engines Using Click-throughData:

This paper presents an approach to automatically optimizing the retrieval quality of search engines using click-through data. Intuitively, a good information retrieval system should present relevant documents high in the ranking, with less relevant documents following below. While previous approaches to learning retrieval functions from examples exist, they typically require training data generated from relevance judgments by experts. This makes them difficult and expensive to apply. The goal of this paper is to develop a method that utilizes click-through data for training, namely the query-log of the search engine in connection with the log of links the users clicked on in the presented ranking. Such click-through data is available in abundance and can be recorded at very low cost. Taking a Support Vector Machine (SVM) approach, this paper presents a method for learning retrieval functions. From a theoretical perspective, this method is shown to be well-founded in a risk minimization framework. Furthermore, it is shown to be feasible even for large sets of queries and features.

2.4. Accurately Interpreting Click-through Data as Implicit Feedback:

This paper examines the reliability of implicit feedback generated from click-through data in WWW search.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Analyzing the users' decision process using eyetracking and comparing implicit feedback against manual relevance judgments, we conclude that clicks are informative but biased. While this makes the interpretation of clicks as absolute relevance judgments difficult, we show that relative preferences derived from clicks are reasonably accurate on average.We presented the first comprehensive study addressing the reliability of implicit feedback for WWW search engines that combines detailed evidence about the users' decision process as derived from eyetracking, with a comparison against explicit relevance judgments.Furthermore, we are exploring relative feedback from clicks not only for results within a single query, but spanning a chain of related queries.

2.5. Generating QuerySubstitutions:

We have shown that we are able to generate highly relevant query substitutions. Further work includes building a semantic classifier, to predict the semantic class of the rewriting. With such a classifier we would be able to focus on the targeted subtypes of rewriting, such as spelling variants, synonyms, or topically related terms. To improve our algorithm, we can also take inspiration from machine translation techniques. Query rewriting can be viewed as a machine translation problem, where the source language is the language of user search queries, and the target language is the language of the application (for instance advertiser language in the case of sponsored search). In order to generalize our work to any application, we also need to work on introducing a language model, so that in the absence of filtering with the list of sponsored queries, we avoid producing nonsensical queries. In addition, with the algorithm in operation we could learn a new ranking function using click information for labels.

2.6. Automatic Identification of User Goalsin Web Search:

There have been recent interests in studying the "goal" behind a user's Web query, so that this goal can be used to improve the quality of a search engine's results. Previous studies have mainly focused on using manual query-log investigation to identify Web query goals. In this paper we study whether and how we can automate this goal-identification process.

We first present our results from a human subject study that strongly indicates the feasibility of automatic query-goal identification. We then propose two types of features for the goal-identification task: user-click behavior and anchor-link distribution. Our experimental evaluation shows that by combining these features we can correctly identify the goals for 90% of the queries studied.

3. Modules:

1.LOGIN:

A user can log in to a system to obtain access and can then log out or log off when the access is no longer needed.

2.USER SEARCH LOGS:

The user enters the queries to the search engine. The queries are maintained as a log and the results will be produced based on the keywords.

3.FEEDBACK SESSIONS:

The feedback sessions is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. We combine the enriched URL's in a feedback sessions to form a pseudo document..

4.PSEUDO DOCUMENTS:

The feedback sessions vary a lot for different clicks through and queries,it is not suitable to directly use the feedback sessions some method id needed to represent the feedbacks in a more efficient way.

» Represent the URL in the feedback session.

» Forming pseudo documents based on URL representations.

5.CLUSTERING THE PSEUDO DOCUMENTS:

The Pseudo documents are clustered into K means clustering .It performs clustering based on the five values.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The terms with the highest values in the center points are used as the keywords to depict user search goals. The clustering is the process based on a term-weight vector representation of queries.We do rank the suggested queries based on two criteria's:

» The similarity of the queries to input query (the query submitted to the search engine)

» The support which measures how much the answers of the query have attracted the user's attention.

6. FINAL RESTRUCTURED RESULTS:

The results are restructured based on the evaluation of web search goals..This approach Is called CAP(Classified Average Precision).Search engines will returns millions of search results so I is necessary to organize them to make it easier for users to find what they want. The user search goals are represented as the vectors. So we perform categorization by choosing the smallest distance between the URL vector anduser-search –goal vectors. By this way the results can be restructured according to the inferred user search goals.

Fig.1 HOME SCREEN

Click on Administrator to login as admin:

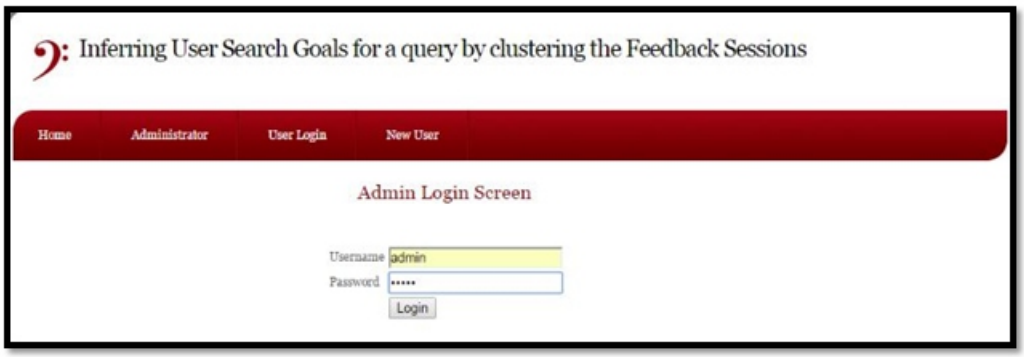

Fig.2LOGIN AS ADMIN

Click on Load Dataset to load the dataset into our application.After successfully loading the dataset:
(In our application we are using AOL Dataset (small_dataset.txt) will be loaded)

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Fig.3 DATASET LOADING

Click on new user to register as a new user:

Fig.4 REGISTERING A NEW USER

After successful registration, click on User login to login as a registered user:

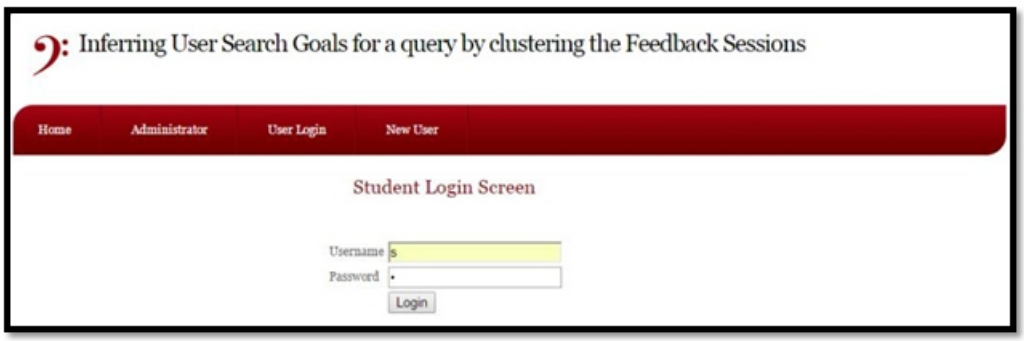

Fig.5 LOGIN SCREEN

Click on search for searching:

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

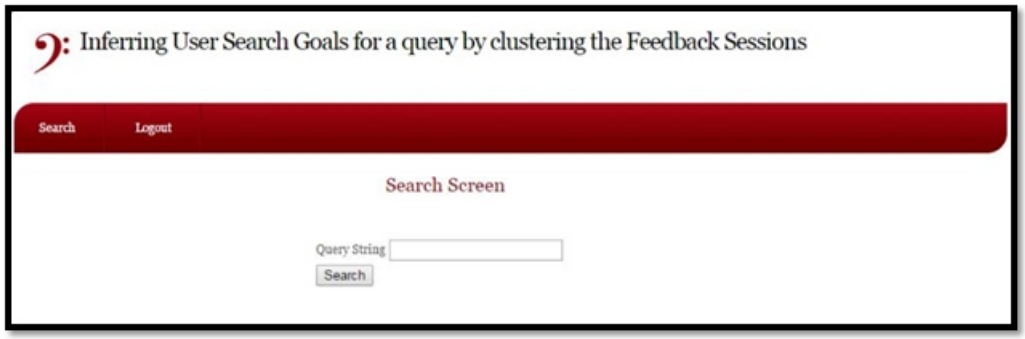

Fig.6 USER SEARCH SCREEN

Fig.7 ENTER THE QUERY FOR SEARCHING

Here we are searching for the query 'cbsnews':
Searched results
In the above result 1.0 is the similarity match with the searched query.

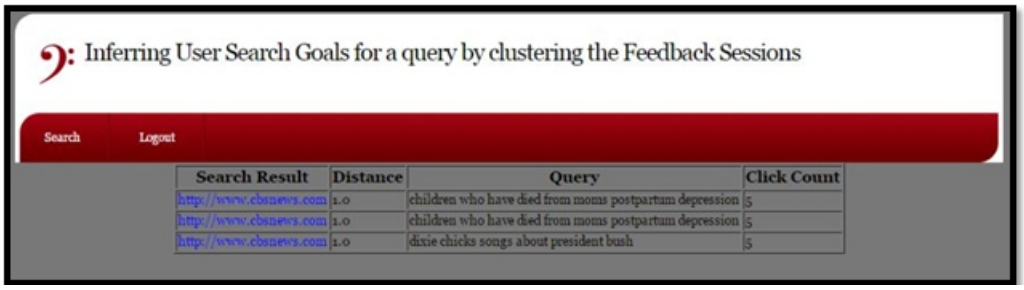

Fig.8 SEARCHED RESULTS

4. Conclusions:

In this paper, a novel approach has been proposed to inferuser search goals for a query by clustering its feedbacksessions represented by pseudo-documents. First, weintroduce feedback sessions to be analyzed to infer usersearch goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the lastclick are considered as user implicit feedbacks and takeninto account to construct feedback sessions. Therefore,feedback sessions can reflect user information needs moreefficiently. Second, we map feedback sessions to pseudodocuments to approximate goal texts in user minds. Thepseudo-documents can enrich the URLs with additionaltextual contents including the titles and snippets. Based onthese pseudo-documents, user search goals can then bediscovered and depicted with some keywords. Finally, anew criterion CAP is formulated to evaluate the performance of user search goal inference. Experimental resultson user click-through logs from a commercial search enginedemonstrate the effectiveness of our proposed methods.The complexity of our approach is low and our approachcan be used in reality easily. For each query, the running timedepends on the number of feedback sessions. However, thedimension ofFfsin (3) and (5) is not very high. Therefore,the running time is usually short.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

In reality, our approachcan discover user search goals for some popular queriesoffline at first. Then, when users submit one of the queries,the search engine can return the results that are categorizedinto different groups according to user search goals online.Thus, users can find what they want conveniently.

Reference:

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines,"Proc. Int'l Conf.Current Trends in Database Technology (EDBT '04),pp. 588-596,2004.

[3] D. Beeferman and A. Berger, "Agglomerative Clustering of aSearch Engine Query Log,"Proc. Sixth ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '00),pp. 407-416,2000.

[4] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "VaryingApproaches to Topical Web Query Classification,"Proc. 30th Ann.Int'l ACM SIGIR Conf. Research and Development (SIGIR '07),pp. 783-784, 2007.

[5] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li,"Context-Aware Query Suggestion by Mining Click-Through,"Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and DataMining (SIGKDD '08),pp. 875-883, 2008.

[6] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results,"Proc. SIGCHI Conf. HumanFactors in Computing Systems (SIGCHI '00),pp. 145-152, 2000.

[7] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant TermSuggestion in Interactive Web Search Based on ContextualInformation in Query Session Logs," J. Am. Soc. for InformationScience and Technology, vol.54, no. 7, pp. 638-649, 2003.

[8] T. Joachims, "Evaluating Retrieval Performance Using Click-through Data," Text Mining,J. Franke, G. Nakhaeizadeh, andI. Renz, eds., pp. 79-96, Physica/ Springer Verlag, 2003.

[9] T. Joachims, "Optimizing Search Engines Using Click-throughData,"Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discoveryand Data Mining (SIGKDD '02),pp. 133-142, 2002.

[10] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay,"Accurately Interpreting Clickthrough Data as Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research andDevelopment in Information Retrieval (SIGIR '05),pp. 154-161, 2005.