

Inferring User Search using Feedback Session and Semantic Search

Deshmukh N. M.
MBES's College of
Engineering, Ambajogai,
Maharashtra.

Patil B. M.
MBES's College of
Engineering, Ambajogai,
Maharashtra.

Chandode V. M.
MBES's College of
Engineering, Ambajogai,
Maharashtra.

ABSTRACT

Different users may have diverse search goals for ambiguous queries when they put forward it to a search engine. Search engine relevance and user experience are enhanced by the analysis of user search goals. In this work, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. First, goals of particular queries by clustering feedback session are found which depends upon user click through logs. It can efficiently replicate goals of users. Secondly, to estimate query texts in user minds, clustered feedback session is mapped to pseudo documents. Then 'CAP' (Classified Average Precision) is used to appraise performance of user search goals. Finally, a new criterion is "Semantic code" is proposed in which one can find out a particular aim of user's search.

Keywords

User search goals, feedback sessions, semantic search and pseudo-documents.

1. INTRODUCTION

In web search applications, user queries are submitted to search engines to symbolize the information requests for users. On the other hand, many times the queries may not exactly give users specific information needs. For example, the query "The Apple" is submitted to search engine, some users may want to search Apple products like iphones, beats, tv, watches, Mac book , iPods etc. Other user may want to know about company, CEOs, founder and some other may want to know about or may search for good quality of apple fruits. Therefore it is necessary to capture diverse user search goals in information retrieval. User search goals can be taken as the clusters of information needs for a query. In this work, number of different user goals for a query is discovered and expressed each goal with some keywords mechanically [2]. First of all, a new way to infer user search goals of query is given by clustering projected feedback sessions. The feedback session is differentiated as the series of evenly clicked and unclicked URLs that ends with the most recent URL that was clicked in a session from user click through logs. Then a novel semantic search method is proposed to generate more relevant results. Then an optimization technique to plot feedback sessions to pseudo documents which can proficiently give user information requirements. At last these pseudo documents are clustered and attach them some keywords. Since an assessment of clustering is also a significant difficulty, a novel evaluation criterion Classified Average Precision (CAP) is proposed to evaluate the performance of the restructured web search results. In brief this work has three main contributions as:

- A structure to infer different user search goals of a query by using semantic search are given and clustering feedback sessions. Clustering feedback

sessions is more efficient than clustering search results or clicked URLs directly is shown here. Furthermore, the distributions of more than one user search goals can be obtained conveniently after feedback sessions are clustered.

- A novel optimization technique is proposed to merge the enriched URLs in a feedback session to form a pseudo-document, which can effectively give the needed information to user. Thus, users goal in detail are found.
- A new measure CAP is proposed to evaluate the performance of user search goal inference based on restructuring web search results. Thus, anyone can determine the number of user search goals for a query.

Overview of system architecture is given in Figure 1.

2. RELATED WORK

Many works has been done until now on inferring user search results [14], [15], [16]. But there work only belongs to query classification. In some kind of works query search results directly retrieved to analyze the search goals of users [6], [20]. However, there are many limitations in trying to know the goals of the users query without feedback session.

- 1) Wang and Zhai worked on this [19]. They clustered goals of the queries and learned goals of similar queries, which can solve the problem partially. There work took user feedback in consideration and analyze the clicked URLs of a particular query directly in user click through logs. However, the problem is number of clicked URLs is not that big to get ideal result through that user click through logs, if anyone will try to search one query into a query cluster, which will not give us expected output. One will have to search it in clusters of similar kind of queries.
- 2) In the year 2002 gave a method to automatically optimization of a data from user click trough log. As per T. Joachims a good data is that which is mostly related to the clicked data that is it must be relevant data. Prior to the use of user click-through logs is to get user implicit feedback to expand training data when understanding rank functions in information recovery. Much kind of works has been done by Thorsten Joachims on in which way implicit feedback is used to improve the quality of information retrieval [9], [10], [11]. In this new approach feedback sessions as user implicit feedback is used and implied a new optimization technique which combines both clicked and unclicked URL's in feedback sessions to discover what users really needed and what they don't need.

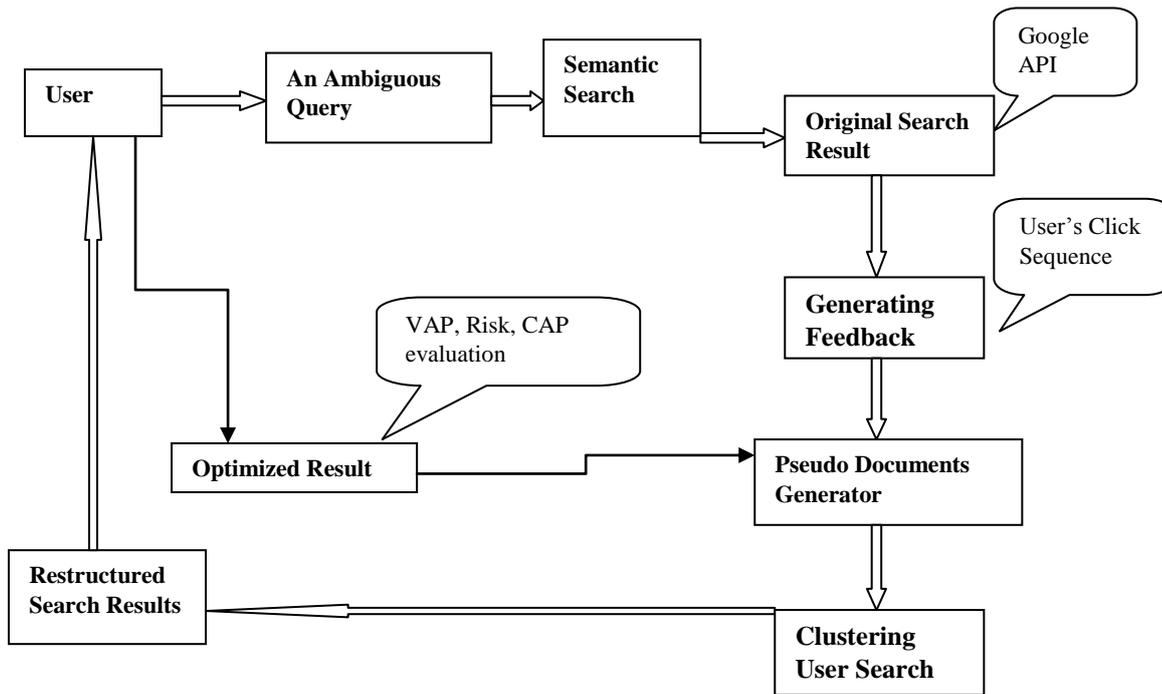


Fig. 1: System Architecture

- 3) Restructuring web search results of a particular query is one of the applications of inferring the user search goals. There are many correlated works especially on organizing the search results of a query [7], [19], [21].
- 4) Q-C He, Z.chen, H-J Zeng, W-Y ma, J ma (4) these people made research on restructuring search results of a particular query.

This approach has following 4 steps:

- Fetch the search result of a query.
- Calculation of phrase characteristic and parsing of text.
- Expression ranking or phrase ranking
- Post refinement.

3. SEMANTIC SEARCH

A semantic word is accustomed to get better search exactness by understanding the searcher's aim and contextual meaning of expressions as they become visible in the searchable data space on web or in a closed system, to generate more related results. Semantic search system consider many points including context of seek out, locality, goal, difference of words synonyms, widespread and particular queries, identical concept and natural language queries to provide relevant search results. It is more efficient to use semantic or the science of meaning in language rather than using ranking algorithms such as Google's page rank to predict relevancy, semantic search uses produce highly relevant search results [17]. In most conditions the goal is to deliver the information queried by a user rather than have a user sort through a list of loosely related keyword results. Considering a real world example of semantic search as if you are working on computer and someone ask you "Do you have windows there?" Here windows mean Microsoft window operating system and if someone is new in office and asking for office window, space he /she has different meaning by window but having same question.

Search engines can't distinguish the goals based on context because it does not know the context. But in case of human beings, they can realize the queries which are based on contexts. Semantic search in search engine is the way in which search engine would provide related search results based on contextual meaning of search.

4. CLUSTERING OF PSEUDO DOCUMENTS

By computing similarity between two pseudo-documents is clustering is done. One can cluster pseudo-documents by K-means clustering which is simple and effective. As no one know the accurate number of user search goals for each query, K is set to five different values, (i.e., 1; 2; . . . ; 5) and carry out clustering based on these five values. The optimal results will be found through the evaluation criterion after clustering all the pseudo-documents, each cluster can be considered as one user search goal.

Feedback session

Click Sequence

http://en.m.wikipedia.org/wiki/The_Apple	0
m.imdb.com/title	0
http://www.apple.com	1
http://www.apple.com/watch	2
http://m.youtube.com	0
applecider.co.uk	0
www.cnbc.com	3
iphone.appleinsider.com	4
http://www.cnet.com	0

The center point of a cluster is computed as the middling of the vectors of all the pseudo-documents in the cluster. Lastly, the terms with the maximum values in the center points are used as the keywords to represent user search goals. Observe that one more advantage of using this keyword based description is that the extracted keywords can also be utilized to outline a more meaningful query in query suggestion and thus can give user information requirements more efficiently. Furthermore, since one can get much number of the feedback sessions in each cluster, the helpful distributions of user search goals can be achieved simultaneously.

5. RESTRUCTURING SEARCH RESULTS

To apply the evaluation method for large data, the single sessions in user click through logs are used to reduce manual work. By user click-through logs, one can find out implicitly relevant feedbacks, here “clicked” is taken into account as relevant and “un-clicked” means irrelevant. A probable evaluation criterion is the average precision (AP) which uses a particular way of evaluation according to user implied feedbacks. AP is the precisions’ average which is calculated at the each relevant document’s point in the ranked sequence, as given below:

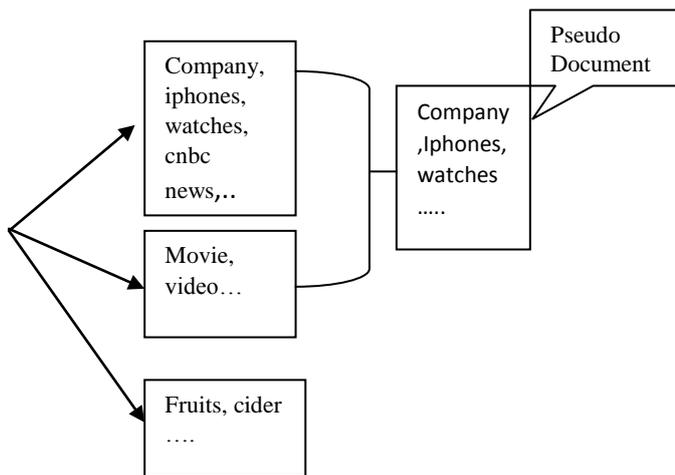


Fig. 2: Illustration for mapping feedback session to Pseudo Document.

$$AP = \frac{1}{N^+} \sum_{r=1}^N rel(r) \frac{R_r}{r} \quad (5.1)$$

Where, N is the total number of retrieved documents, N^+ is the number of relevant (or clicked) documents in the retrieved ones, r is the rank $rel()$ is a binary function on the relevance of a given rank, and R_r is the number of relevant retrieved document of rank r or less. As, AP is not appropriate for evaluating the restructured or clustered search results. The new criterion for evaluating restructured results is “Voted AP (VAP)” which is the AP of the class including more clicks namely votes. The single session URLs are restructured into two classes depending upon clicks URL. VAP is useful when there is single search goal. If there are many search goals occurs with VAP therefore risk is proposed as follows:

$$Risk = \frac{\sum_{i,j=1(i < j)}^m d_{ij}}{C_m^2} \quad (5.2)$$

It calculates the number of clicked URL pairs that are not in the same class, where m is the number of the clicked URLs.

If the pair of the i_{th} clicked URL and the j_{th} clicked URL is not categorized into one class d_{ij} will be 1; otherwise, it will be 0 based on risk, VAP is extended by introducing a new criterion “Classified AP,” as below

$$CAP = VAP * (1 - Risk)^v \quad (5.3)$$

From given equation, one can see that CAP selects the AP of the class that user is interested in and takes the risk of wrong classification in consideration. And v is used to adjust the influence of Risk on CAP. The mean and the variance of the optimal v are 0.697 and 0.005, respectively. Thus, v set to be 0.7. Finally, CAP is used to estimate the presentation of restructuring search results.

6. OBJECT EVALUATION AND COMPARISON

Here, objective evaluation of this search goal inference method and the comparison with old method that is like Google search is given. This proposed method clusters feedback sessions to infer user search goals. In order to show that when inferring user search goals, clustering proposed feedback sessions are more proficient than clustering search results and clicked URLs directly. In order to further compare this method with the old method, 5 most ambiguous queries are tested. The results are shown in fig 3, 4 and 5. In fig.3 and 4 VAP and risk comparison between two methods is given.

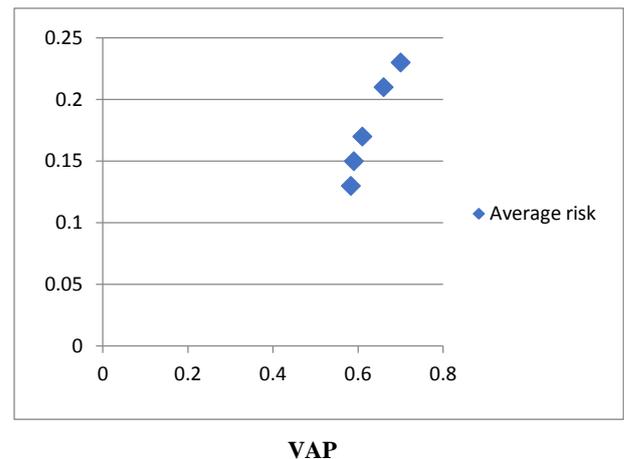
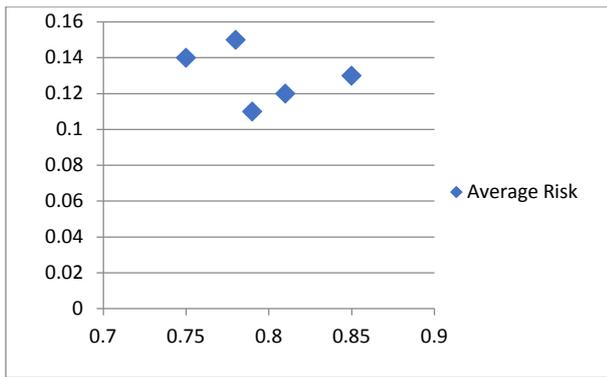


Fig. 3: VAP and Risk of Old Method like Google API for Five Ambiguous Queries

Table 1: VAP and Risk of Old Method for Five Ambiguous Queries

QUERY NO.	PROPOSED METHOD	PREVIOUS METHOD
1	0.583	0.13
2	0.61	0.17
3	0.66	0.21
4	0.7	0.23
5	0.59	0.15



VAP

Fig.4: VAP and Risk of Proposed Method for Five Ambiguous Queries

Table 2: VAP and Risk of Proposed Method for Five Ambiguous Queries

QUERY NO.	PROPOSED METHOD	PREVIOUS METHOD
1	0.75	0.14
2	0.81	0.12
3	0.78	0.15
4	0.85	0.13
5	0.79	0.11

In fig. 5 the CAPs of each query of the two methods are given. It is obvious that this method has the highest CAP for all ambiguous queries. CAP for first five queries is given.

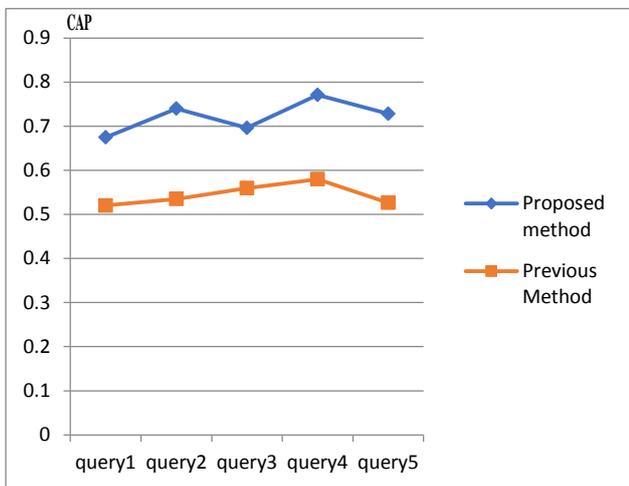


Fig. 5: CAP for Five Ambiguous Queries.

Table 3: CAP for Five Ambiguous Queries.

CAP	PROPOSED METHOD	PREVIOUS METHOD
QUERY 1	0.6748	0.52
QUERY 2	0.74	0.535
QUERY 3	0.6961	0.5596
QUERY 4	0.771	0.58
QUERY 5	0.7281	0.5265

7. PERFORMANCE ANALYSIS

Here some spontaneous explanation showing why clustering feedback sessions and pseudo documents are better than the old method is given. With the introduction of feedback sessions, everyone will have a lot of advantages. Some advantages are summarized as follows:

- 1) Feedback sessions can be considered as a process of resampling. Without resampling, there could be many noisy URLs in the search results, which are hardly ever clicked by users. If search results are clustered with these noisy ones, the performance of clustering will disgrace greatly. Feedback sessions “resample” the URLs and remove those noisy ones. Therefore, this method is much better than old method
- 2) Feedback session is a meaningful mixture of several URLs. So, it can give user information requirement more precisely and there are plenty of feedback sessions to be analyzed.

In following screenshots how the clustering in feedback session is performed for ambiguous query is shown:

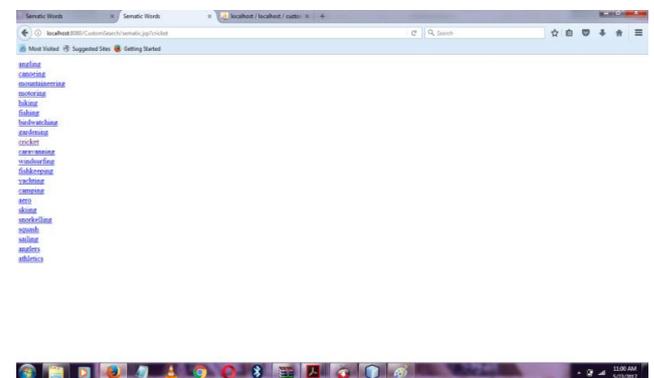


Fig.6: Snapshot of Semantic Words Are Generated.

In first screenshot semantic words for a query cricket is given and in second screenshot original results of the query after searching are got. By seeing this, one can say that semantic words are more efficient. Then particular URLs are selected, here ESPNcricinfo-wikipedia. After submitting this query in next search feedback session shows only related queries, as shown in last screenshot.

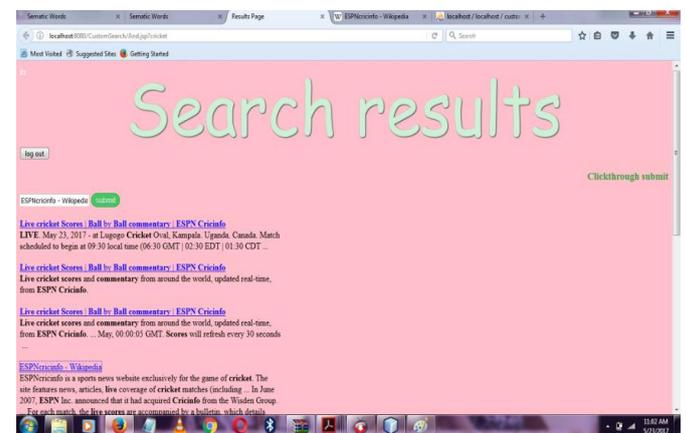


Fig.7: Snapshot of Original Result

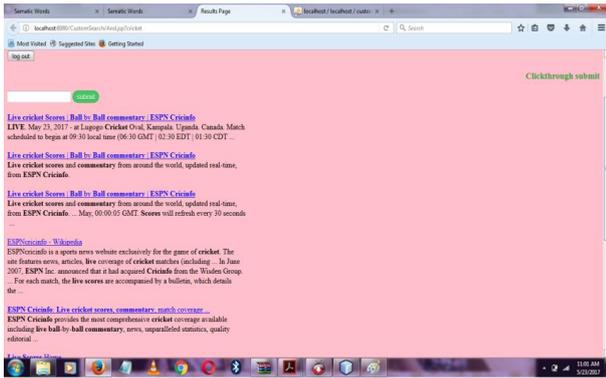


Fig.8: Snapshot of Restructured Result.

8. CONCLUSION AND FUTURE SCOPE

In this work, a new method has been projected to infer user search goals for an uncertainty by using clustering search results directly. First of all, semantic words which gives exactness is generated and then feedback session which helps in finding out user's particular aim. Second, feedback sessions are mapped to pseudo documents to find out exact aim in user's mind. At last, a new criterion CAP (Classified Average Precision) is formulated to estimate the performance of user search goal inference. The difficulty of this approach is low and it can be used in actuality. Performance can be improved in future by analyzing feedback session with improved method.

9. REFERENCES

- [1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin and Zhaohui Zheng, A new algorithm inferring user search goals using feedback session, IEEE transaction on knowledge and data engineering, vol. 25, no. 3, march 2013
- [2] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.
- [3] R. Baeza-Yates, C. Hurtado, and M. Mendoza "Query Recommendation using query login search Engines." Proc. Int'l Conf. Current Trend in Database Technology (EDBT' 04), pp. 2004.
- [4] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine QueryLog," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD'00), pp.407-416, 2000.
- [5] S. Beitzel, E. Jensen, A. Chowdhury, and O Frieder "Varying Approaches to Topical we Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR'07), pp. 783-784, 2007.
- [6] 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 Data Mining (SIGKDD'08), pp. 875-883, 2008
- [7] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 2000.

- [8] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp.2003.
- [9] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003
- [10] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.
- [11] T. Joachims, L. Granka, B. Pang, H Hembrooke, and G. Gay, "Accurately Interpreting Click through Data as Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 05), pp. 154-161, 2005.
- [12] R. Jones and K.L. Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 699-00
- [13] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 387-396, 2006.
- [14] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005.
- [15] X. Li, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346, 2008.
- [16] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf Artificial Intelligence (IJCAI '07), pp. 2007.
- [17] B. Poblete and B.-Y Ricardo, "Query-Sets Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 2008.
- [18] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for Web Query Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 131-138, 2006.
- [19] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
- [20] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.
- [21] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.