Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 6, June 2014, pg.184 – 192

RESEARCH ARTICLE

Improved Accuracy and User Satisfaction by Inferring User Search Goals based on Feedback Sessions

Ms. S. S. Jadhav[1], Prof. N. D. Kale[2]

[1] Department of Computer Engineering, University of Pune, India

[2] Department of Computer Engineering, University of Pune, India

[1] sheetaljadhav132@yahoo.com; [2] navnath1577@yahoo.co.in

Abstract— User search goals can be defined as information on various aspects of query that user want to obtain and it can be considered as the collection of information needs for a query. Different users may have different search goals in their mind when they pass ambiguous query to a search engine. Thus, it is important to infer and analyze user search goals to improve the performance of a search engine and user experience. By clustering the proposed feedback sessions, we infer different user search goals for a query. The feedback session is combination of both clicked and unclicked URLs and this feedback session is mapped to the pseudo-documents to better represent the information needs of user. These pseudo-documents are clustered using bisecting K-means clustering algorithm which produces better results than K-means clustering algorithm and reduces computation time. Finally, Classified Average Precision (CAP) evaluation criterion is used to evaluate the performance of system. In this way, the proposed system can infer user search goals efficiently and satisfy information needs of user. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

Keywords— Click through data, Classified average precision, Feedback sessions, Pseudo-documents, Query classification, User search goals

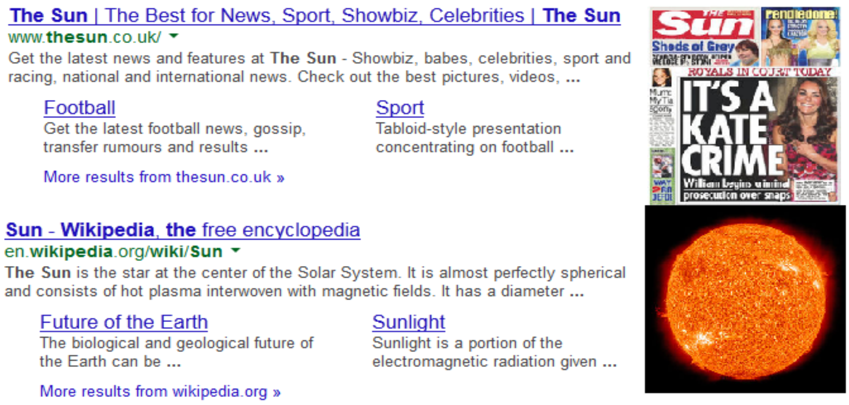
I. INTRODUCTION

Now day’s web search is more booming area of research. There are so many efficient methods already presented by different authors, every method claiming their efficiency in their own ways.

This area is basically defined by the uses search goals. However, sometimes queries entered by the user may not exactly represent information needs since many ambiguous queries cover large topics and various users may want to get information on different subject when they submit the same query. For example, when the user submit ‘The sun’ as query to

Google search engine , some users may want to get information related to United Kingdom newspaper, while other users want to get the natural knowledge of the sun, as shown in Fig. 1. So, it is important and necessary to find out different search goals in information retrieval. User search goals can be defined as information on various aspects of query that user want to obtain. User search goals can be considered as the collection of information needs for a query. Finding appropriate user search goals and performing its analysis have many of advantages in enhancing performance of search engine relevance and user experience. Some advantages are summarized as follows:

1) We can restructure web search results according to user search goals. In this, search results are grouped together with the same search goal. Thus, users with different search goals can find what information they want.
2) User search goals which are represented by the keywords can be used in query recommendation; thus, the users can take help of the suggested queries to form their queries more precisely.
3) The distributions of user search goals are useful in applications such as re-ranking web search results which contain different user search goals.



The Sun | The Best for News, Sport, Showbiz, Celebrities | The Sun
www.thesun.co.uk/
Get the latest news and features at The Sun - Showbiz, babes, celebrities, sport and racing, national and international news. Check out the best pictures, videos, ...

Football
Get the latest football news, gossip, transfer rumours and results ...

Sport
Tabloid-style presentation concentrating on football ...

More results from thesun.co.uk »

Sun - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Sun
The Sun is the star at the center of the Solar System. It is almost perfectly spherical and consists of hot plasma interwoven with magnetic fields. It has a diameter ...

Future of the Earth
The biological and geological future of the Earth can be ...

Sunlight
Sunlight is a portion of the electromagnetic radiation given ...

More results from wikipedia.org »

Fig. 1.Example of user search goal for the query ‘The sun’ and its distribution

In this paper, we give solution to discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a novel approach to infer user search goals for query by clustering our proposed feedback sessions. The feedback session is defined as the series of both clicked and unclicked URLs and ends with last URL that was clicked in a session from user clickthrough logs. Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords. Since the evaluation of clustering is also an important problem, we also propose a novel evaluation criterion classified average precision (CAP) to evaluate performance of the restructured web search results.

In next section II we are presenting existing systems that worked on user search methods. In section III, the proposed approach is represented. In section IV we are presenting the experimental results and analysis. Last section concludes the paper.

II. EXISTING SYSTEM

In this section we can compare the different methods those are presented to solve the problem of user search goals.

- In [2], Lee et al. study whether and how they can automate goal-identification process. They first present their results from a human subject study that strongly indicates the feasibility of automatic query-goal identification. They stated that majority of queries have a predictable goal. Taxonomy of query goals based on two types: Navigational queries and Informational queries. Two features are used for the prediction of user goal:

  1. Past user-click behavior:

  If a query is navigational, users will primarily click on the result that the user has in mind. Therefore, by Observing the past user-click behavior on the query, we can identify the goal.

  2. Anchor-link distribution:

  If users associate particular query with a particular website then most of the links that contain the anchor will point to that particular website. Hence by observing the destinations of the links with the query keyword as the anchor, we can identify the potential goal of the query.

  Their experimental evaluation shows that by combining these features they an correctly identify the goals for 90% of the queries.
- R. Jones and K.L. Klinkner[3], defined session boundaries and automatic hierarchical segmentation of search topics. In this approach, analysis of typical timeouts used to divide query streams into sessions and the hierarchical analysis of user search tasks into short-term goal and long-term missions is done.

  This method only identifies whether a pair of queries belong to the same goal or mission but does not care about what the goal is in detail.

- In [4], Wang and Zhai learn interesting aspects of queries by analyzing the clicked URLs directly from user click-through logs to organize search results. However, this method has limitation since the number of different clicked URLs of a query may be small. In this paper, they propose to address these two deficiencies by (1) learning "interesting aspects" of a topic from Web search logs and organizing search results accordingly; and (2) generating more meaningful cluster labels using past query words entered by users. They evaluate their proposed method on a commercial search engine log data. Compared with the traditional methods of clustering search results, our method can give better result organization and more meaningful labels.

- In [5], [6] J.Zheg, H.Zhen analyze the search results and returned by the search engine when a query is submitted. Since user feedback is not considered, several noisy search results that are not clicked by any users may be analyzed as well. In this paper, they reformalize the clustering problem as a salient phrase ranking problem. Given a query and the ranked list of documents (typically a list of titles and snippets) returned by a certain Web search engine, their method first extracts and ranks salient phrases as candidate cluster names, based on a regression model learned from human labeled training data. The documents are assigned to relevant salient phrases to form candidate clusters, and the final clusters are generated by merging these candidate clusters

- In [7], Li et al. define query intents as “Product intent” and “Job intent” and they try to classify queries according to the defined intents. Other works focus on tagging queries with some pre-defined concept to improve feature representation of queries [8]. However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical. In the second class, people try to reorganize search results.

III. PROPOSED SYSTEM

3.1 Problem Definition

To discover the number of diverse user search goals for a query and depict each goal with some keywords automatically. The evaluation of user search goal inference is a big problem, since user search goals are not predefined and there is no ground truth. Previous work has not proposed a suitable approach for this problem.

3.2 Proposed System Architecture

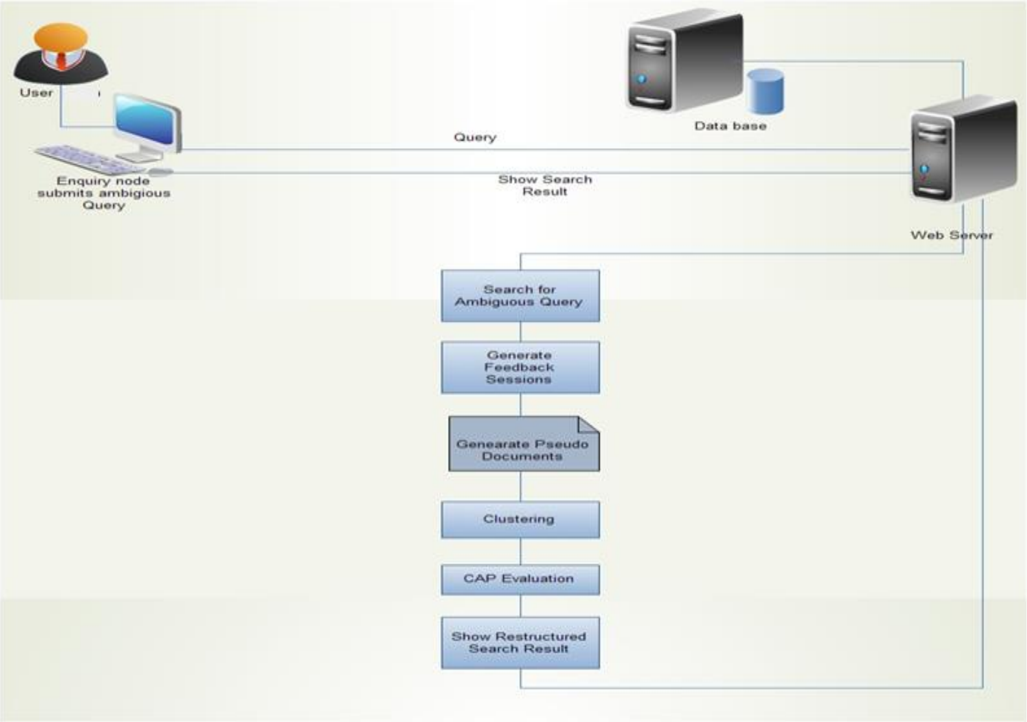

User
Data base
Query
Show Search Result
Enquiry node submits ambigious Query
Web Server
Search for Ambiguous Query
Generate Feedback Sessions
Genearate Pseudo Documents
Clustering
CAP Evaluation
Show Restructured Search Result

Fig. 2 Framework of approach

Fig. 2 shows the framework of our approach. Our framework consists of two parts. In the first part, all the feedback sessions of a query are extracted from user click-through logs and converted to the pseudo-documents. Then, user search goals are inferred by performing the clustering on these pseudo-documents. Each goal is depicted with some keywords. As the exact number of user search goals are not known in advance, several values are tried and the optimal value will be calculated.

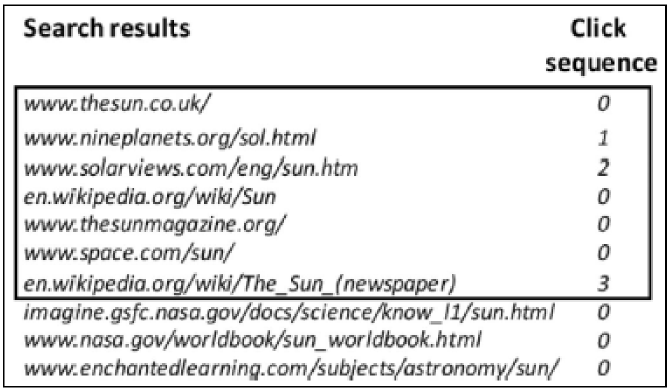
In the second part, the original search results are rearranged based on the user search goals inferred from the first part. Then, the performance of restructured search result is evaluated by evaluation criterion CAP and final evaluation result will be used as the feedback to get the optimal number of user search goals in the first part.

Major parts of the system are discussed as follows:

1) Feedback session:

Generally, a session is used in reference to web applications. It is sequence of interaction between server and the user. The feedback session is combination of both clicked and unclicked URLs and this session stops with the last URL clicked by user. It is important that before the last click, all the URLs are scanned and analyzed by users. Thus, both the clicked and unclicked URLs before the last click are considered as a part of the user feedback. Fig. 3 shows a feedback session and a single session.

In Fig. 3, the left part shows 10 search results for the query and the right part shows sequence for user clicks .Here "0" shows unclicked URLs. The single session is composed of all 10 URLs in Fig. 3, but the feedback session is consisting of seven URLs in the rectangular box.



Search results	Click sequence
www.thesun.co.uk/	0
www.nineplanets.org/sol.html	1
www.solarviews.com/eng/sun.htm	2
en.wikipedia.org/wiki/Sun	0
www.thesunmagazine.org/	0
www.space.com/sun/	0
en.wikipedia.org/wiki/The_Sun_(newspaper)	3
imagine.gsfc.nasa.gov/docs/science/know_l1/sun.html	0
www.nasa.gov/worldbook/sun_worldbook.html	0
www.enchantedlearning.com/subjects/astronomy/sun/	0

Fig. 3 Feedback session in single session for the query 'The sun'

These seven URLs again composed of three clicked URLs and four unclicked URLs. Inside the feedback session, the clicked URLs reflect what user wants and the unclicked URL tells what users do not care. It is important that the unclicked URLs after the last clicked URL should not be considered as the part of feedback sessions.

2) Mapping of Feedback Sessions to Pseudo-Documents:

Mapping of feedback session to Pseudo-document includes two steps.

a) Representing the URLs in the feedback session

In the first step, we extract titles and snippets of URLs appearing in the feedback session. Then textual processes are implemented on snippet and titles like converting all the letters to lowercase, steaming and removing stop words.

We use Term Frequency-Inverse Document Frequency (TF-IDF) vector [1] to represent each URL's title and snippet, respectively,

$$T_{ui} = (t_{w1}, t_{w2}, \ldots\ldots, t_{wn})^T$$
$$S_{ui} = (s_{w1}, s_{w2}, \ldots\ldots, s_{wn})^T$$

Where T_{ui} and S_{ui} are the TF-IDF vectors of the URL's title and snippet, respectively. ui is the ith URL in the feedback session. The wn is term appearing in the URL. Here, a “term” is nothing but word or a number in the document collections.

$$F_{ui} = w_t T_{ui} + w_s S_{ui}$$

F_{ui} is Feature representation of the ith URL in the feedback session which is weighted sum of T_{ui} and S_{ui}. w_t and w_s are the weights of the titles and the snippets, respectively.

b) Forming pseudo-document based on URL representations

In the second step, we form pseudo-document based on URLs representation. This is done by combining the clicked and unclicked URLs. Once pseudo document is created we can infer search goals effectively.

3) Clustering the Pseudo-documents:

One of the most popular clustering methods used today is the K-means clustering algorithm. However, it has been reported that the bisecting K-means algorithm, an augmented variant of the original K-means algorithm, produces better clustering results than the standard K-means. The bisecting K-means simply repeats standard K-means clustering where k is fixed. In our paper, we are using bisecting K-mean algorithm which will produces better clustering results.

4) Classified Average Precision(CAP)evaluation method

We apply CAP method to evaluate the results and restructure the web results. We can obtain an implicit relevance feedback, namely “clicked” which means relevant and “unclicked” means irrelevant. Average precision (AP) [1] evaluates as per user implicit feedbacks. AP is calculated as:

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

Where, N^+ is the number of clicked documents. r is the rank, N is the total number of documents that are retrieved, rel() is a binary function. R_r is the number of clicked retrieved documents of rank r or less. AP is not best solution for evaluating clustered searching results. Thus we use new criterion “Classified AP,” as

$$CAP = VAP * (1 - \mathrm{Risk})^{\gamma}$$

Where, “Voted AP (VAP)” is the AP of the class including more clicks. Risk is used to avoid classification of search results into too many classes. γ is used to adjust the influence of Risk. And it is given as

$$\mathrm{Risk} = \frac{\sum_{\mathrm{i,j=1(i<j)}}^{\mathrm{m}} \mathrm{d\,ij}}{\mathrm{C_m}}$$

Where, m is the number of the clicked URLs. If ith, jth clicked URL are categorized into one class, then d_{ij} is set to 1 otherwise it will be 0.The term C_m is total number of the clicked URL pairs.

IV. Result Analysis

The data set that we used is based on the clickthrough logs from a commercial search engine collected over a period of two months. As shown in Fig. , we compare proposed method with previous existing method. Risk and VAP are used to evaluate the performance of restructuring search results together. Each point in Fig. 4 represents the average Risk and VAP of a query. If the search results of a query are restructured properly, Risk should be small and VAP should be high and the point should tend to be at the top left corner. We can see that the points of our method are closer to the top left corner comparatively.

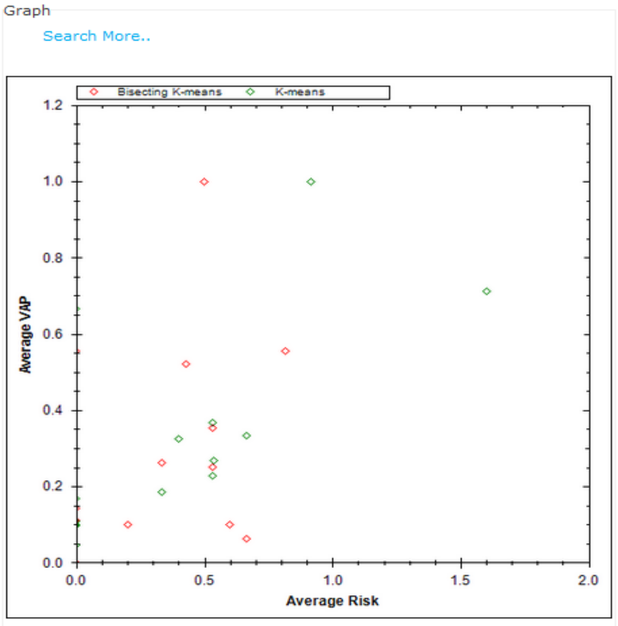


Fig. 4 The Comparison of methods. Each point represents the average Risk and VAP of a query when evaluating the performance of restructuring the search results.

The average CAPs of each query of the proposed method and previous method are shown in Fig. 5. It is obvious that our method usually has the highest average CAP. Previous method had used K-Means algorithm where as we are using Bisecting K-Means algorithm. Clustering results of proposed method are better than previous method.

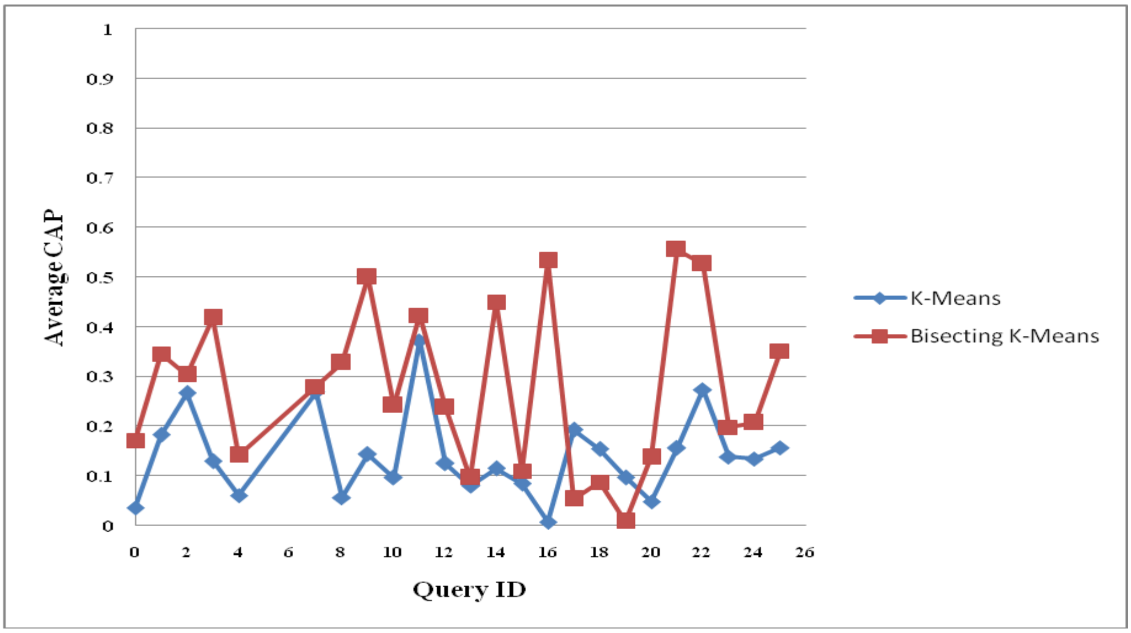


Fig. 5 The chart of CAP comparison of two methods.

We compute the mean average VAP, Risk, and CAP of queries as shown in table I . We can see that the mean average CAP of our method is the highest than previous method. The results of previous method are lower than ours due to the lack of user feedbacks.

TABLE II
CAP COMPARISON OF METHODS

Method	**Mean Average VAP**	**Mean average CAP**
Our proposed Method	0.822	0.645
Previous Method	0.755	0.563

V. CONCLUSIONS

The proposed method focuses on inferring the user search goals by performing clustering on feedback session represented by pseudo-documents. Feedback sessions can reflect user information needs more efficiently. This system helps to the user to reduce their extra efforts while gathering information using search engine. The proposed system can be used to improve discovery of user search goals for a similar query.

This approach satisfies information needs of the user as well as saves lot of time to search ambiguous query. By using this approach we get efficient and correct search results for the query. As we are using bisecting K-means algorithm, it reduces computation time and gives better clustering results.

Proposed approach has low complexity and can be used in reality. The running time of query depends on the number of feedback session and thus it is usually short. In reality, this approach can identify user search goals with some keywords automatically. When users submit the queries, restructured results are returned by the search engines that are categorized into different groups as per the user search goals. Thus, users can find information related to query conveniently without any extra efforts.

ACKNOWLEDGMENT

I express my special thanks with gratitude and great respect to my valuable guide Prof. N.D.Kale for his keen interest, fruitful suggestions & valuable guidance with motivation & constant encouragement to complete my dissertation work. I am also thankful to him for his great patience, constructive criticism & useful suggestion apart from invaluable guidance.

I would like to express my deep sense of gratitude and sincere thanks to the H.O.D Prof. Y.B.Gurav, for their support and providing access to data/documents/processes needed during the project.

REFERENCES

[1] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", vol. 25, no. 3, 2013.

[2] 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.

[3] 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. 699-708, 2008

[4] 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.

[5] 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.

[6] 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. 145-152, 2000.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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. 638-649, 2003.