IPHV7I3026X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .III
ISSN Online: 2319 – 9253
Print: 2319 – 9245

USER SEARCH GOAL INFERENCE AND FEEDBACK SESSION USING FUZZY MECHANISM

[#1]D.RamBabu, Assistant Professor,
[#2]TV.Naryana Rao, Professor,
[#3]A.Chandu Naik, Assistant Professor,
Dept of CSE,
SREENIDHI INSTITUTE OF SCIENCE AND TECHNOLOGY, HYD, T.S., INDIA.

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 un clicked 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. There is a high stress on search engines due to the overload of information content in the internet. Search query submitted by the user to the search engine represents the user requirements. Sometimes, particular desire of the user cannot be fulfilled by the user search query. Also, long listed search result may not be always significant to the user requirements and irrelevant documents are returned by many of the existing search engines which follow the mechanism of keyword matching. Indeed, both the users and search engine developers need to reduce the information content in the internet. In this paper, we aim to infer the user search goal by considering the clicked URLs and reorganize the web search result. We use FG-FCM based clustering for grouping the semantically similar search results which further enhances the reorganized search result.

Keywords— *Ambiguous Query, Broad-topic Query, Feedback session, Semantics,Pseudo-documents, Query classification, User search goals.*

I.INTRODUCTION

The dependency on the search engine has grown recently and the users can obtain plenty of information in the internet by submitting the query to the search engine. The requirements of the user are represented by the search query. Finding the right information when searching on search engines can be a pain for sure. Search engines present the search result to the user based on the ranking of website and not according to user interests. Thus, the result of the search engine is same for all the users though different users have different interests. For the broad-topic and ambiguous query, different users will have different search goal. For example, when the query "jaguar" is submitted to a search engine, some users may wish to find the information about the car while some others may intend to find the meaning of animal. Users' particular information needs may not be satisfied by the query given by the user. Therefore, it is required to know the exact information needs of the user. It is necessary to infer the exact user search goal for satisfying the user needs. In this paper, we aim to improve the search engine relevance by identifying the various goals of a user search query and restructuring the web search results. Inference of user search goal can also be used in recommending the list of related queries [8] for the query submitted by the user.

The user search goal has to be inferred for a search query submitted by the user based on clustering the feedback session. Feedback sessions can be represented in various ways like binary vector representation, pseudo-documents, etc. In this paper, we use pseudo-documents which contain keywords to represent the feedback session. The feedback sessions are mapped to the corresponding pseudo-documents. The semantically similar keywords are found for the given query. The search results that are semantically similar are clustered by FG-FCM clustering according to the search goal. Each cluster represents one search goal. The FG-FCM algorithm allows one piece of document data belong to two or more clusters. Also, the algorithm restructures and enhances the original search result by inferring the search goal of the user and reduces the time spent by the user in searching their information needs. By this method, user needs are satisfied. Performance of the restructured search result can be done by an evaluation criterion, Classified Average Precision (CAP).

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.

IPHV7I3026X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .III
ISSN Online: 2319 – 9253
Print: 2319 – 9245

3) The distributions of user search goals are useful in applications such as re-ranking web search results which contain different user search goals.

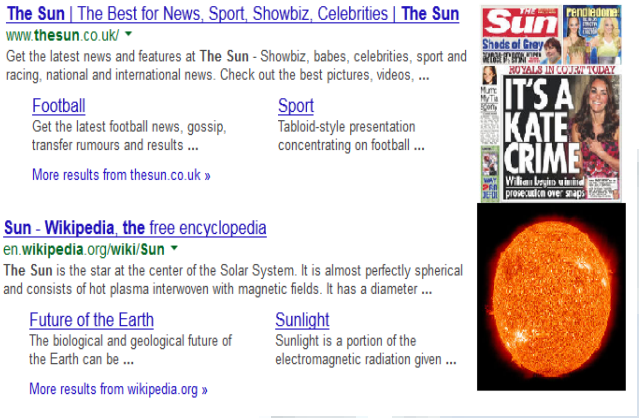

Fig. 1.Example of user search goal for the query „The sun" and its distribution

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.

II.RELATED WORK

Up to date, many works have been made to investigate on obtaining the user search goals and type of query. We examine some of the previous works to study the problem of clustering. It is important to discover different search goals of the given query to fulfill the needs of the user. Long listed search results can be restructured [2], [7], [9] according to the user requirements. Analysis of user search goals can be divided into three modules: search result reorganization, session boundary detection and query classification. In the first class, authors tried to reorganize the search results of the web. Wang and Zhai [7] analyzed the click-through logs and grouped the search result according to the clicked URLs. In second module, Jones and Klinkner [3] considered session boundaries to identify whether the queries and the goal match. In the third module, people categorized the user goal and queries into some specific classes. Lee et al. [4] categorized the user queries into "Navigational" and "Informational", and inferred the search goals automatically. The search goal can be used to improve the quality of a search engine's results. They also discussed how to automate the goal identification process. Goal-identification task was based on two types of features: user-click behavior and anchor-link distribution Li et al. [5] defined the objective of the query as "Product intent" and "Job intent" and categorized the search queries accordingly.

Today's Web search engines provide very user friendly interface. Users can submit the queries in the form of keywords similar information retrieval system. Keywords may be a simple keyword or it may be a broad-topic and anything else. Search engine lists the related queries when a query is submitted. Ricardo Baeza-Yates et al. [6] discussed that the associated queries are based on previously issued queries and are provided to the user for redirecting the search process. Semantically similar queries were also identified by the clustering process which clusters the contents stored in query log of search engine. There are many advantages of restructuring the search result according to the user search goal.

Joachims used implicit feedback to enhance the quality of search engines. He referred to click-through logs to optimize the search engine. Zheng Lu et al restructured the search result by clustering the pseudo-documents using K- means clustering [10]. But the K-means algorithm is computationally difficult to find the value of K and it does not work well with the clusters (in the original data) of different size and different density. Some of the prior works considered click-through logs as user's feedback and restructured the search results accordingly. Other works did not consider the user feedback and considered the entire search results returned by the search engine though the link was not clicked by the user. These type of works produced noisy results.

In this paper, we infer the user search goal by considering the feedback session. By this, we restructure and enhance the search result in order to satisfy the user needs. We use FG-FCM algorithm, a fast and robust algorithm for clustering the semantically similar links in the feedback sessions which provide better result than the previous works.

The approach followed in this paper for inferring the user search goal is shown in the Fig2.

2.1 User Search Query Analysis

The user search query submitted by the user has to be analyzed. The click-through logs are referred for examining the user search queries and defining the feedback sessions. The queries submitted to the search engines by the user may be a simple query or ambiguous query. It is necessary to analyze the different meanings of the ambiguous query and restructure the search result into different clusters in order to get the user needs satisfied. The search results obtained for the query submitted by the user must be collected for restructuring the search result.

2.2 Feedback Session

The first process in reorganizing the search result is the feedback session representation. Feedback session consist the list of URLs up to the URL that was clicked by the user at last in a single session. All the unclicked URLs before the last clicked URL in a single session is also included because those URLs also has been browsed and analyzed by the user. Therefore, these unclicked URLs must also be included for the feedback. From this feedback session, the clicked URLs represent what information the user entail and the unclicked URLs reflect what information the user do not require. The URLs that are present after the last clicked URL cannot be taken as a part of feedback because it is not certain whether the user have scanned those URLs or not.

Feedback session cannot be used directly for user search goal inference because it varies from that of the user click-through logs. So, it should be represented in some other forms in order to infer the user search goals efficiently. It can be represented in various forms. Binary vector representation is one of the popular ways of

IPHV7I3026X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .III
ISSN Online: 2319 – 9253
Print: 2319 – 9245

representing the feedback session. It consists of 0's and 1's where "0" represents the unclicked URL and "1" represents the clicked URL in a single session. This method cannot be used when more feedback sessions are considered because diverse feedback sessions may have unusual aspects.

The vague keywords can be used to represent the user interests for a query. But these keywords cannot be used for representing the feedback session because they are usually hidden and not expressed clearly. Therefore, pseudo-documents can be used to infer the goals of the user. The feedback sessions are mapped to the pseudo-documents. These documents can be formed by enriching those URLs present in the feedback session. Enriching the URLs can be done by adding the title and a short snippet in a small text paragraph for the same URLs.

2.3 Semantic Similarity and Fast Generalized-Fuzzy C Means Clustering

Semantics of the query submitted by the user must be analyzed and restructured accordingly. Semantically similar words can be identified by the wordnet tool. From the wordnet tool, the semantically similar words for the user search query are extracted. Then the FG-FCM clustering process begins. This algorithm is a variation of FCM algorithm which differs by adding the mathematical exponentiation to the result obtained using FCM. The number of clusters need not be specific.

The advantage of using this algorithm is that the same data element can be in more than one cluster and also the clustering process is more efficient than the existing algorithm. The titles in the feedback session are grouped based on the similarity between the semantic keywords and the titles. Also, the similarity matrix Ui,j is used for the clustering process.

The matrix consists of rows and columns where both the rows and columns represent the same titles of the search result arranged in same order. Entry in the matrix represents the similarity between the titles. Various similarity measures can be used to calculate the similarity. In this paper, cosine similarity is used. Cosine score for the titles can be computed as,

SimilarityTi,Tj = Cos (Ti,Tj) (1)

Where, Ti and Tj represent titles of the search results.

The matrix can be computed by

$$U_{ij} = \frac{1}{\sum \left[\frac{|x_i - C_j|}{|x_i - C_j|} \right]^2} \quad (2)$$

Where, Xi represents the keyword count which is the total number of the semantic keywords in each URL of the feedback session. Cj and Ck are the value of clusters which are obtained from the computation,

$$C_j = \frac{\sum U_{ij} \cdot x_i}{\sum U_{ij}} \quad (3)$$

Where, Ui,j represents the value of the similarity matrix at the position i,j. ∑ implies that the process should be repeated for the each and every title in the feedback session.

The search results are grouped based on the clustered value. Thus, the search results are restructured and reorganized into groups based on semantic similarity. By assembling the semantically similar URLs into different clusters and restructure the search result, the users' browsing experience can be improved efficiently.

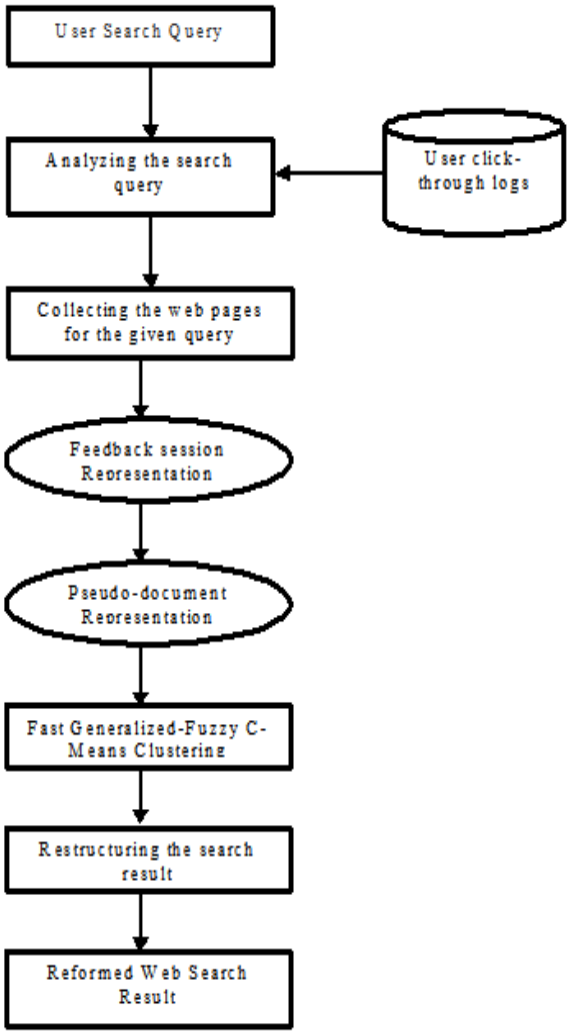

Fig.2. Search result restructuring process

This will also satisfy the requirements of the user and reduces the time spent in browsing the contents. This process will be very much useful for the ambiguous queries submitted by the user where there will be more than one meaning.

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.

IPHV7I3026X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .III
ISSN Online: 2319 – 9253
Print: 2319 – 9245

3.2 Proposed System Architecture

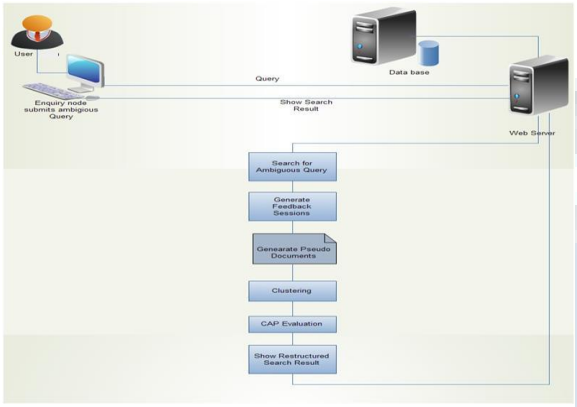

Fig. 3 Framework of approach

Fig. 3 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. 4 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. wn is the ith URL in the feedback session. The ut 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 Tui and Sui. 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:

IPHV7I3026X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .III
ISSN Online: 2319 – 9253
Print: 2319 – 9245

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. 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

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

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.

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

Where, m is the number of the clicked URLs. If ith, jth clicked URL are categorized into one class, then dij is set to 1 otherwise it will be 0.The term 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.

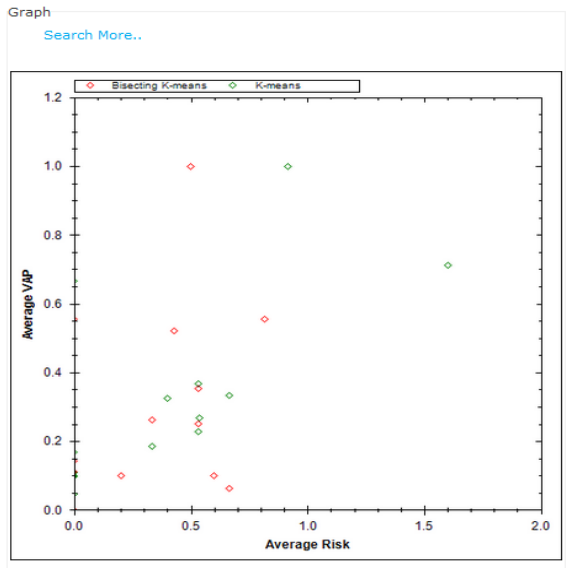

Fig. 5 The Comparison of methods. Each point represents the average Risk and VAP of a query when evaluating the performance of restructuring the search results.

The average CAPs of each query of the proposed method and previous method are shown in Fig. 6. 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.

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

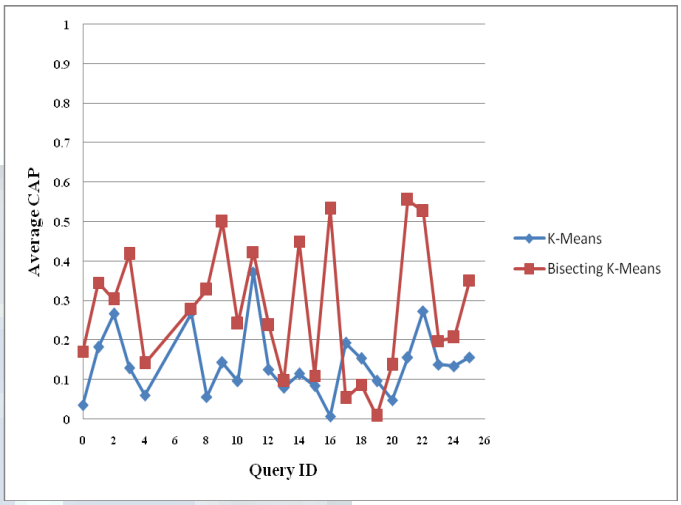

Fig. 6 The chart of CAP comparison of two methods.

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 I
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.

The method used in this paper can be used to infer the user search goal based on the feedback session. We analyze and reorganize only the search results that are obtained in the feedback session for efficient browsing. Therefore, there will not be any noisy data for restructuring the search result. Also, we consider semantically similar keywords to enhance the restructured search result. FG-FCM algorithm is used for clustering the URLs into different groups according to semantic similarity. Thus, the enhanced search result will improve the search engine relevance and satisfy the user to a greater extent. In future work, we plan to continue by investigating the feedback session not only for the single query in the form of keywords, but also for the queries submitted in the form of a sentence.

REFERENCES

[1] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACMSIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 407-416, 2000.
[2] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 145-152, 2000.
[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, pp. 699-708, 2008.
[4] U. Lee, Z. Liu and J. Cho, "Automatic Identification of User Goal sin Web Search," Proc. 14th Int'l Conf. World Wide Web, pp. 391-400, 2005.
[5] 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, pp. 339-346, 2008.
[6] B. Poblete and B.Y. Ricardo B, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web, pp. 41-50, 2008.
[7] 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 pp. 87-94, 2007.
[8] B.R. Yates, C Hurtado, and M Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology, pp. 588-596, 2004.
[9] H.J. Zeng, 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 pp. 210-217, 2004.
[10] L. Zheng, Z. Hongyuan, Y. Xiaokang, L. Weiyao and Z. Zhaohui, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, 2013.