Volume 2, Issue 11, November-2015, pp. 780-784 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

Inferring User Search Goals with Feedback Sessions using K-means clustering algorithm

[1]Dasari Amarendra, [2]Kaveti Kiran Kumar

[1]M.Tech (CSE), Department of Computer Science & Engineering, NRI Institute of Technology

[2]Assistant Professor, Department of Computer Science & Engineering, NRI Institute of Technology.

Abstract: - Recognizing or inferring client's search objective from given query is a difficult job as search engines let users to identify queries simply as a list of keywords which might refer to broad topics, to technical terminology, or even to proper nouns that can be used to guide the search procedure to the significant compilation of documents. In order needs of users are correspond to by queries submitted to search engines and different users have different search goals for a broad topic. Sometimes queries may not exactly represent the user's information needs due to the use of short queries with uncertain terms. thus to get the best results it is necessary to capture different user search goals. These user goals are nothing but information on different aspects of a query that different users want to obtain. The judgment and analysis of user search goals can be improved by the relevant result obtained from search engine and user's feedback. Here, feedback sessions are used to discover different user search goals based on series of both clicked and unclicked URL's. The pseudo-documents are generated to better represent feedback sessions which can reflect the information need of user. With this the original search results are restructured and to evaluate the performance of restructured search results, classified average precision (CAP) is used. This evaluation is used as feedback to select the optimal user search goals.

Keywords:-User search goals, feedback sessions, pseudo-documents, restructuring search results, classified average precision

1. INTRODUCTION

Web search engines attempt to satisfy user's information needs by ranking web pages with respect to queries. Web search is a process of querying, learning, and reformulating. A series of interactions between user and search engine can be necessary to satisfy a single information need. For broad queries and topics different users have different ways of representations i.e. different users have different search goals. Sometimes user specific information needs may not be represented by queries since many ambiguous queries may cover a broad topic. Therefore, it is necessary to capture different user search goals. User search goals are information on different aspects of query that user want to obtain. Inference and analysis of user search goals have advantages such as restructure the web search results according to user search goals by grouping the search results with the same search goal, user search goals represented by some keywords can be utilized in query recommendation and distribution of user search goals. There are three classes representing user search goals: 1. Query classification, 2. Search result reorganization, 3. Session boundary detection. In first class, some specific classes are predefined and query classification is performed accordingly. User goals are classified into navigational and informational. For navigational, user has particular web page in mind but for informational user's does not have particular page in mind or intends to visit multiple pages. Some other methods used for defining queries as product intent and job intent. Next method defined is tagging queries with some predefined contents to improve feature representation of queries. Disadvantages of this classification are finding suitable predefined search goal class is difficult because what user cares

about varies a lot for different queries. In second class, people try to recognize search results. First method used is learning interesting aspects of queries by analyzing the clicked URLs directly from user clickthrough logs to organize search results. Limitation of this is number of clicked URL's may be small. Another method used is analyzing the search results returned by a search engine when a query is submitted. But disadvantage of this method is feedback is not taken into account so noisy results that are not clicked by user may be analysed. In third class, aim is to detect session boundaries. This method predicts goal and mission boundaries to hierarchically segment queries logs. Limitation with this if it only identifies whether a pair of queries belong to same goal and does not care about the goal in detail. Here, aim is to discover the number of different kinds of user search goals for a query and describing each goal with some keywords. For this purpose first approach is to Cluster the feedback sessions to infer user search goals. Feedback session contains both clicked and unclicked URL's and ends with the last URL that was clicked in a session. The distributions of different search goals can be obtained after feedback sessions are clustered. Then to reflect user information needs effectively map these feedback sessions to pseudo-documents. This is nothing but the optimization method to combine the enriched URL's in a feedback session. CAP(Classified average precision) is used to evaluate the performance of user search goal inference based on restructuring web search results. Using which we can determine number of user search goals for a query.

II. RELATED WORK

Many recent works have been done to infer the user search goals. U Lee et.al proposed an automated user search goal identification method. They introduce two features to identify the search goals. They are user click behavior and anchor link distribution [2]. Some works focused on query suggestions from previous session of search engine. Cao et.al proposed a context aware query suggestion utilizing the click through log and session data [3]. Huang and Chen suggest relevant terms for user queries from similar query sessions [4]. Jones and clinker try to segment the user session hierarchically [5]. Joachim proposed a method to optimize the retrieval quality of search engines with the help user click through logs. The utilization of previous session information helps to identify the similar queries and URL's. Beeferman and Berger try to cluster similar queries and URL's [7]. Li et.al suggested a method to clarify the query intent from click through graphs [9]. Shen builds a bridging classifier to map user queries to a target category [10]. All these works are based on the clickthrough logs and session data; we are also utilizing the clickthrough log and session data for analyzing the user search goal

III. CURRENT APPROACH

In this section, we describe methods implemented in the approach; methods are mainly divided into five classes. Feedback session, pseudo document representation, clustering pseudo document, rearrangement of cluster content, evaluation with user feedback and clustering inside the cluster. Detailed description of these methods is as follows.

A. Feedback Sessions

The feedback session [1] consist of both clicked and unclicked URL's end with the last URL that was clicked in a single session. Figure 1 shows an example of feedback session represent what a user required and what he/she does not required. In the fig. 1 the value '0' indicates unclicked URL's and other values indicate click sequence of the clicked URL's

Search Results	www.isro.org	www.nasa.gov	www.space.ca	www.space.com	www.bbc.co.uk/space	En.wikipedia.oeg/wiki/space
Click Sequence	4	2	3	1	0	0

Fig 1. Feedback session

B. Pseudo Document Representations

Feedback session are unsuitable for direct use, hence some representation method is needed to describe it. Pseudo document [1] is an efficient representational method. Fig. 2 shows an illustration for mapping feedback session to pseudo documents. In the fig. 2 the URL's in the feedback session are enriched with titles and snippet, then combine the enriched URL's to form a pseudo document. After that some textual process are applied to those documents, such as removing stop words, transforming upper case to lower case letters etc. The Term Frequency- Inverse Document frequency (TF-IDF) [11] representation is used for each URL's titles and snippets.TFIDF vectors of the URL's title and snippet are multiplied with some weights. The weight value is increased by adding the important terms in the pseudo document more than one time. These important terms are obtained on the basis of number of times the words occur in the html content. It will also affect the TF-IDF vector value of the pseudo document. The resulting vector representation is used for clustering.

C. Clustering Pseudo Document

In this section, we describe how to identify user search goals from the created pseudo documents. For that we cluster pseudo documents by weighted K-means clustering. Here we don't know the exact number of search goals; hence the value of k should be different. The optimal value will be obtained through the user feedback evaluation. The each cluster can be considered as one user search goal. Each cluster contains different categories of URL's. The distribution of search results is described in the next session.

D. Distribution of Clusters and Cluster Contents

The order of distribution of clusters in the web search result can be decided on the basis of page ranking. The priority is given to the search result with highest page ranking. Based on this order, the user interface will show the search results in each cluster.

E. Evaluation with User Feedback

The optimal number of clusters are not determines yet, hence a feedback information is needed to determine the best cluster number. In order to apply the evaluation method, the single session of the user click through log can be used. The required feedback can be obtained from this single session Average Precision (AP) [11] is an evaluation criteria, which evaluated according to user feedbacks. It can be calculated using the Eq(1).

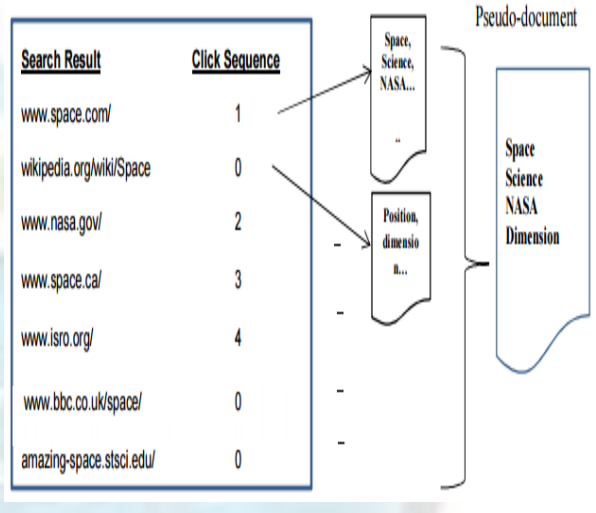

Fig 2. Evaluation with user feedback

F. Clustering Inside the Cluster

There are a large number of search results inside each cluster. It is difficult to display all the results in the clusters at the same time. So minimize the search results to a fixed value at the initial time, based on the user feedback we have to display all the search result in the cluster as sub clusters. For that we are also applying the clustering algorithm to the user selected cluster contents. This method will help to minimize problems of showing large search result in a search engine.

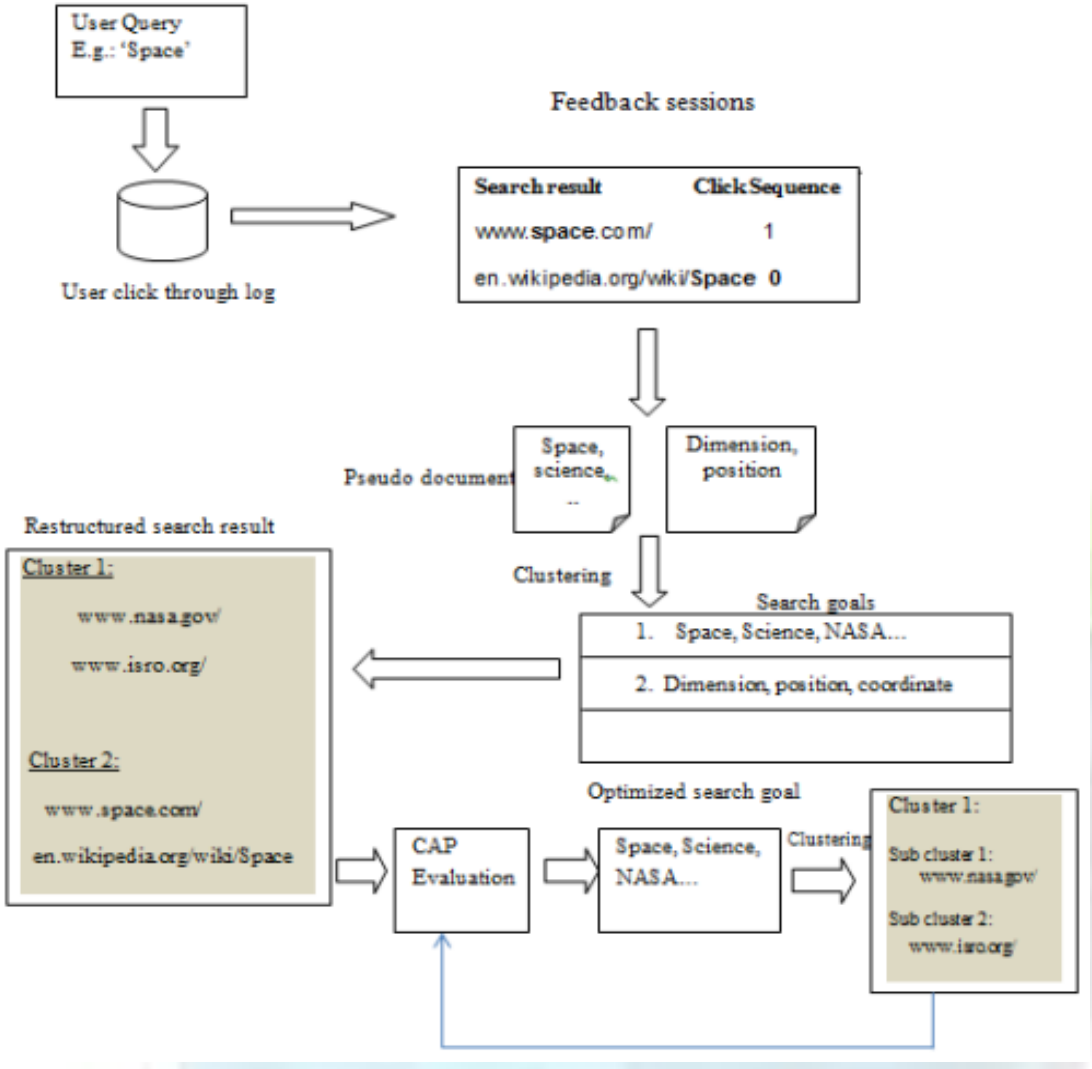

Fig 3. Search anatomy with clustering

The experiments of the proposed algorithm are done in a dataset of commercial search engine. There is some preprocess are implemented for the experimental use. The collected dataset is the click through log of the search engine collected over a period of six months. There is a large number of user queries and result in the dataset. The selected queries are moved into the database with the following attributes like session id, user query, URL, date & time, click sequence. Also the titles and snippets of all the available URLs are collected and stored in the database. Then we design a user interface for the search for the search engine. The query given in the query field is searched through the previously created database. The feedback sessions obtained are saved as documents. Now the titles and snippets corresponding to each URLs in the feedback session are find out from the database and created the pseudo document. This pseudo documents obtained are also saved into the storage device. Then, we are clustering the pseudo documents with weighted K-means clustering algorithm. The weights are assigned on the basis of number of occurrences of the words in the pseudo document. This algorithm will results the user search goals. The optimized search goals are determined through CAP evaluation method. Each clusters obtained are showing in the search engine interface. The URLs in the search results are organized on the basis of page ranking. Fig. 5 shows the restructure search result. Top ranked results only shows in the interface. Now CAP evaluation is done according to the user feedbacks obtained. The optimized clusters are selected for further clustering process. Then sub search goals are obtained on this clustering result. This will helps to infer the user

search goals from a large search results. The major advantages of our proposed methods are the following:

- User search goals can find out in less time.
- Restructuring the search result.
- It solved the problem of handling the large number of search results.
- More accurate search results will produce in sub clustering.
- User feedback has an importance in this method of approach.

IV.CONCLUSION

In this paper, we proposed a method to infer user search goals for a user given query. In the first step utilize feedback sessions to analyze user search goal and in the second step, conversion of feedback session into pseudo document. Then cluster the pseudo documents for finding the user goals. In clustering some weights are include for the important terms inside the pseudo document. It helps to focus the importance of the occurrence of the terms. After that organize cluster contents according to the page ranking. Then CAP evaluation is used to evaluate the performance of user search goal inference. If the number of resulting contents in the cluster is too large, we propose a method to cluster the contents inside the cluster. The implementation complexity of our approach is low. The running time required is short.so the search engine performance and relevance can be improved.

REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 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] 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. [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] 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.

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

[8] 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.

[9] 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.

[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 and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

