ISSN:0975-9646

A New Approach for Inferring User Search Goals using user's implicite Feedback.

Vidya V. Hire# , Swati K. Bhavsar*

*#PG Student , *Assistance Professor*

Department of Computer Engineering (University of Pune)

Matoshri College of Engineering and Research Centre Nasik, India

***Abstract*— Different users may have dissimilar search purpose when they submit large topic and ambiguous query, to a search engine. The inference and analysis of user search goals is an important point in terms of improving performance of search engine. A novel approach is used to infer user search goals by analyzing search engine query logs. First, we propose frameworks that are used to find out different user search goals for a user submitted query. After that we use K-means algorithm for clustering the proposed pseudo documents generated from feedback session .The Feedback sessions are implement by using user click-through data and efficiently reflect the information needs of users for that query. Second step is that generation of pseudo documents from feedback session. Finally clustering of pseudo documents is done. For clustering purpose we use a algorithm which is K-means algorithm. K-means algorithm is easiest algorithm for pseudo document clustering other than other clustering algorithm. In proposed novel approach new criterion "Classified Average Precision (CAP)" is used for evaluate the performance of inferring user search goals.**

***Keywords*— Classified Average Precision, Feedback sessions, ambiguous query, click-through**

I. INTRODUCTION

An ambiguous query is a query that has a specific meaning and covers a narrow topic. Many ambiguous query cover broad topic but it is problematic, when user enters such ambiguous query to search engine. We know that different user may have different search goals when they submit the same query. Different users may want to get information about different aspects for the same query. For example, when the query "sun" is submitted to a search engine, at that time some users want to get information about a UK newspaper, while some other users want to learn about the natural knowledge of the sun. Through this paper, the basic aim is to finding the number of diverse user search goals for a query and assigning some keywords to each goal automatically. First Feedback sessions are constructed from user click-through logs that can efficiently reflect the information needs of users. In second step, we propose an approach that can generate pseudo-documents from feedback session to better represent the feedback sessions for clustering. The feedback session is a set of series of clicked and un-clicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Then, next step is that clustering of pseudo documents using simple K-Means algorithm, and depict each cluster with some keywords. The performance of restructured web search results is evaluating using parameters like classified average precision (CAP).

The rest of the paper is organized as follows: Reviews several related works is in Section II. Information related to the existing system is represented in Section III. Section IV highlights the detailed problem definition. Detail description of proposed framework is presented in Section V. Mathematical model is represented in section VI. Finally conclusion of paper is in Section VII.

II. LITRATURE SURVEY

In resent year, many works has been done for user search goals analysis. This work can be summarized into three classes: Classification of query, search result reorganization and Session boundary Detection. Some works belong to query classification [2], [3], [7], this work have been done to infer the so-called user goals or intents of a query. Some works analyze the search results returned by the search engine directly to use different query aspects [6], [5]. But query aspects without user feedback have limitations to improve search engine relevance. Some works consider user feedback and analyze the different clicked URLs of a ambiguous query present in user click-through logs directly, but the number of different clicked URLs of that ambiguous query may be not be sufficient to get ideal results. Wang and Zhai[4] proposed framework that restructure web search results according to user search goals by grouping the search results with the same search goal. for example the query "car" is clustered with some other queries, such as "car rental," "used car," "car crash," and "car audio." Some other works introduce search goals and missions to detect session boundary hierarchically [10]. But, their framework only identifies whether a pair of queries belong to the same goal. Limitation of this work is that he does not care what the goal is in detail. Other works perform analysis of the search results returned by the search engine when a query is submitted [6], [5].Here user feedback is not considered, lot off noisy search results that are not clicked by any users may be analyzed as well. Therefore, this kind of methods cannot be use for infer user search goals precisely. Other works make center of attraction is tagging queries with some predefined concepts to improve feature representation of queries [7]. Zamir et al. [7] used Suffix Tree Clustering (STC) to identify set of documents having common phrases and then form cluster according to these phrase. For clustering web documents They used documents snippets instead whole document. Generating meaningful labels for clusters is most challenging in document clustering, and this problem is solve in [8], in this work a supervised learning method is used to extract possible phrases from search result snippets and these phrases are then used to cluster web search results.

III. PROBLEM DEFINITION

To implement a framework for A New Approach to Invent User Search Goals with Feedback Session which will accept user query in the browser and will suggest a feedback based URL for an ambiguous query as well and Proposed system will also improve result of search by clustering of feedback. The basic aim is finding the number of diverse user search goals for a query and depicting each goal with some keywords automatically. First propose a novel approach to infer user search goals for a query by clustering feedback sessions. The feedback session is defined as the series of both Clicked and un-clicked URL s and ends with the last URL that was clicked in a session from user click- through logs. Then, propose a novel optimization method to map feedback sessions to pseudo documents which can efficiently reflect user information need.

IV. EXISTING SYSTEM

We define user search goals as the information on different aspects of a query that user want to obtain. Information need is a user's particular desire to obtain information to satisfy his/her requirements. Users search goals can be consider as the clusters of information needs for that query. The inference and analysis of user search goal is a very important task for improve the performance of search engine. In existing system search engine show all the URLs which contains information related to that query. Feedback session is consider here but conations all URLs (Clicked and Un-clicked) related to that that query. So all noisy URLs contains in the feedback session [6][20], so that user cannot find exact result easily. We know that different users may have different search goals when they submit any query to search engine. For example, when the query "the sun" is submitted to a search engine, at that time some users want to information of the United Kingdom newspaper, while some others want to learn the natural knowledge of the sun. So it is necessary to capture different user search goals in information retrieval. Fig 1 shows the outcome of existing system.

Fig. 1 Different user search goals and their distributions for the query "the sun" from search engine .

V. PROPOSED SYSTEM

In this section, basic operational steps involved in proposed approach to discover the user search goal by clustering pseudo-documents are described. The flow of the proposed system design will be as shown in Fig. 2.

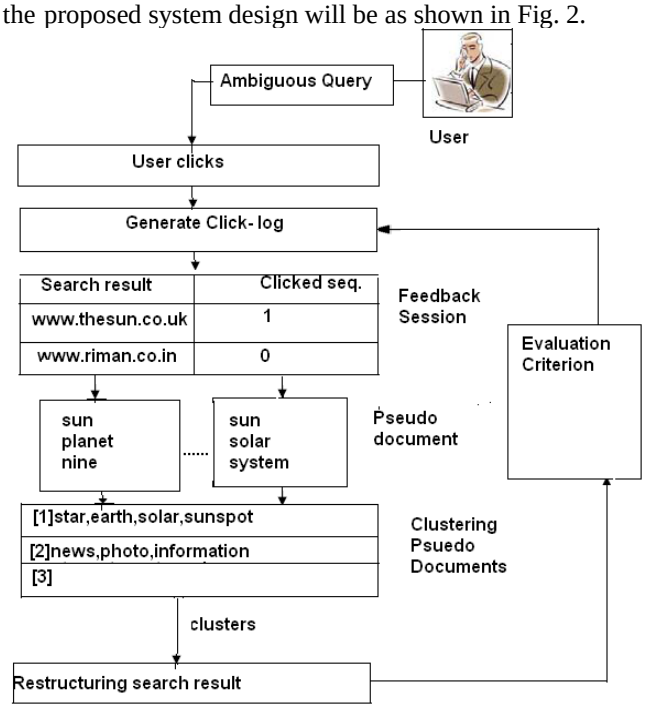

Fig. 2 Flow Diagram of Proposed System

Overall system architecture is as shown in Fig 2. Initially user enters ambiguous query to search engine, then system shows many search results is in the form of number of URLs. From this URLs user clicked some desired URLs, by using this clicked and Un-clicked data feedback system makes a feedback session. After that feedback session is mapped to pseudo document for better representation of user search goals. Finally clustering of pseudo documents is performed. At the end performance of system is calculated by using CAP evaluation criteria. Here, we describe all steps of proposed system.

A. User clicks through logs

User Click-through data log contains information related to the interactions between users and Web search engines. Whatever information search through user is stored in the user click trough logs. Construction of feedback session is performing from this user click through logs. Feedback sessions are better representation using user click through logs. It is more efficient than analyzing the user click through logs directly. For a single query each and every session is analyzed and represents the feedback session. The feedback session is totally based on a single session .An ambiguous query is that it gives more than one results. So find exact results according to the user search goal are difficult to obtain. Information related to the interactions between users and Web search is in User Click-through data log. The user click through logs contains all the user actions according to that query. User Click-through data log contains the information related to the session id, query term, position of the URL, click sequence and the URL.

B. Feedback session

The feedback sessions is discovered from every session of the user click through logs[7]. The feedback sessions consists of numbers of URLs from which some URLs are visited and some are unvisited (represented as "0").Feedback session consist of search result and click sequence. Search result shows URLs related to that ambiguous query. Fig 3 shows feedback session in single session.

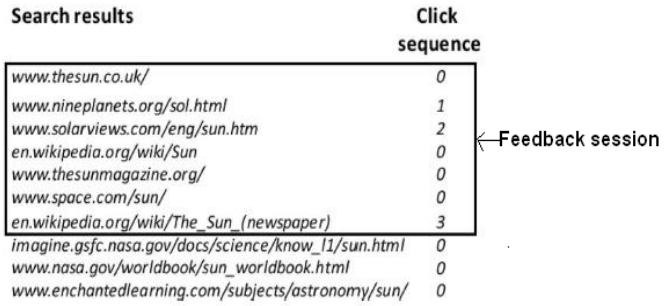

Fig. 3 Feedback session.

In Fig.3, shows single session, the left part of single session contains 10 search results of the query "the sun" and the right part is a user's click sequence from which "0" means "un-clicked." The single session includes all the 10 URLs shown in Figure.3, but the feedback session only includes the seven URLs in the upper rectangular box. Out of seven URLs three are clicked URLs and four are un-clicked .Users will scan the URLs from top to bottom one by one, we can assume that the three clicked URLs, the four un-clicked URLs shown in the rectangular box have also been browsed and evaluated by the user and they should also be a part of the user feedback. The URLs from feed-back session, clicked URLs tell what users wants and the un-clicked URLs means what users do not care about. It would be noted that the un-clicked URLs after the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not. Each and every feedback session can reflect what a user requires and what he/she does not care about. There are lots of diverse feedback sessions present in user click-through searched results and clicked URLs logs. Therefore, for inferring user search goals, it is most important to analyze the feedback sessions than to analyze the search results and clicked URLs.

C. Conversion of feedback session into pseudo document

The URLs in the feedback sessions are enriched by some format. The URLs presents in feedback session are formatted by removing the stop words and the stemming words. It is just showing the information about the whole document by some keywords. The pseudo documents contain the keywords that are retrieved from the URLs in the feedback sessions.

Building of pseudo-document has two steps.

1) Representing the URLs In the Feedback Session: In this first step, we first enrich the URLs with additional textual contents in this step we extract the titles and snippets of the each URLs appearing in the feedback session. In this way, each and every URL in a feedback session is presented as a small text paragraph that consists of its title and snippet of those URLs. After that some textual processes are perform on the text paragraphs, in which transforming all the letters to lowercases, stemming and removing stop words. Finally, each URL's title and snippet are presented in a Term Frequency-Inverse Document Frequency (TF-IDF) vector as in equation 1.

$$T_{u_i} = [t_{w_1}, t_{w_2}, \ldots\ldots\ldots\ldots\ldots\ldots, t_{w_n}]^T$$
$$S_{u_i} = [s_{w_1}, s_{w_2}, \ldots\ldots\ldots\ldots\ldots\ldots, s_{w_n}]^T \longrightarrow (1)$$

Where T_{u_i} and S_{u_i} are the TF-IDF vectors of the URL's title and snippet. In the equation (1), u_i means the i^{th} URL in the feedback session. And $w_j (1,2,\ldots\ldots\ldots\ldots n)$ is the j^{th} term appearing in the enriched URLs. In equation (1) t_{w_j} and s_{w_j} means the TF-IDF value of the j^{th} term in the URL's title and snippet, respectively. Considering that URLs' titles and snippets have different significances, we represent the enriched URL by the weighted sum of T_{u_i} ,

$$\mathbf{F}_{u_i} = \omega_t \mathbf{T}_{u_i} + \omega_s \mathbf{S}_{u_i} = [f_{w_1}, f_{w_2}, \ldots, f_{w_n}]^T, \qquad (2)$$

Where F_{u_i} means the feature representation of the i^{th} URL in the feedback session, and ω_t and ω_s are the weights of the titles and the snippets, respectively.

2) Formation of Pseudo-Document: In this step we perform optimization, in which an optimization method is used to combine both clicked and un-clicked URL's in the feedback sessions. Let F_{f_s} is an feature representation of feedback sessions and $f_{f_s}(\omega)$ be the value for term ω .Clicked and un-clicked URL's in the feedback sessions is represented by using $f_{uc_m}(m = 1,2,3.....M)$ and $f_{u\bar{c}_l}(l = 1,2,.....L)$. $f_{uc_m}(\omega)$ And $f_{u\bar{c}_l}(\omega)$ are the values of term ω in vectors. Is obtaining such as a sum of distances between F_{f_s} and each F_{uc_m} is minimized and sum of distances between F_{f_s} and each F_{uc_l} is maximized. Optimization on each dimension is
Obtained using equation (3),

$$\mathbf{F}_{fs} = [f_{fs}(w_1), f_{fs}(w_2), \ldots f_{fs}(w_n)]^T,$$
$$f_{fs}(w) = \arg\min_{f_{fs}(w)} \left\{ \sum_M [f_{fs}(w) - f_{uc_m}(w)]^2 - \lambda \sum_L [f_{fs}(w) - f_{u\bar{c}_l}(w)]^2 \right\}, f_{fs}(w) \in I_c. \longrightarrow (3)$$

D. Clustering of Pseudo Document

Next and important step is that clustering of pseudo documents generated from feedback session. Using proposed pseudo-documents, we can easily find out the user search goals. In this section, we will describe how to find user search goals and depict them with some meaningful keywords. As in (3), each feedback session is represented by a pseudo document and the feature representation of the pseudo-document is F_{f_s} .The similarity between two pseudo-documents is calculated using cosine score of $F_{f_{si}}$ and $F_{f_{sj}}$ as follows,

$$sim_{i,j} = \cos(F_{f_{si}}, F_{f_{sj}})$$

$$= \frac{F_{f_{si}} . F_{f_{sj}}}{| F_{f_{si}} || F_{f_{sj}} |} \longrightarrow (4)$$

And the distance between two feedback sessions is represented using given equation (5),

$$Dis_{i,j} = 1 - Sim_{i,j} \longrightarrow (5)$$

Clustering is performing by using k-means algorithm it is a simple and effective algorithm of clustering. We do not know the exact number of user search goals for each query, we set K to be five different values (i.e., 1; 2; . . . ; 5) and perform clustering based on these five values, respectively. After clustering all the pseudo-documents, each cluster can be considered as one user search goal that means numbers of cluster is equal to numbers of users search goals. The center point of a cluster is calculated using average of the vectors of all the pseudo-documents in the cluster, as shown in,

$$\mathbf{F}_{center_i} = \frac{\sum_{k=1}^{C_i} \mathbf{F}_{fs_k}}{C_i}, (\mathbf{F}_{fs_k} \subset Cluster\ i), \longrightarrow (6)$$

In which F_{center_i} is the i^{th} cluster's center and C_i is the number of the pseudo-documents in the i^{th} cluster. F_{center_i} is utilized to conclude the search goal of the i^{th} cluster. Finally, the terms with the highest values in the center points are used as the keywords to depict user search goals. Main advantage of using this keyword based description is that the extracted keywords can also be utilized to form a more meaningful query in query recommendation and hence we can represent user information needs more effectively.

E. Restructuring based on web search results

We know that search engines always return millions of search results, so proper organization of search result has to be required to find out the exact result of user query. Organization of search result is also required because it is necessary to organize them to make it easier for users to find out what they want. Restructuring web search results is an application of inferring user search goals. The inferred user search goals are represented by the vectors in (6) and the feature representation of each URL in the search results can be calculated in equation (1) and (2). We can classify each URL into a cluster centered by the inferred search goals. We perform categorization by choosing the smallest distance between the URL vector and user-search-goal vectors. In this way the search results can be restructured according to the inferred user search goals. The performance between
restructured (clustered) web search results and original search results is evaluated by using some parameters like Average Precision (AP), Voted AP (VAP) which is AP of the class having more clicks and Risk to avoid wrong classification of search results and Classified AP (CAP). If user got correct classified results with higher CAP value, this return value is used for optimize the no of clusters of user search goals.

VI. MATHEMATICAL MODEL

Consider user given an ambiguous query, q. When the user submits that ambiguous query to search engine then search results are obtained on the basis of that query, say
R={r1,r2,r3,r4,…,rn}
First, user will click on some of the results, is {r1, r4, r5}
And the click sequence obtained from this is given as,
{r1=1, r4=2, r5=3}. So, the clicked sequence of results
Is as follows, {r1=1, r2=0, r3=0,r4=2,r5=3,….rn=0}
From feedback session contains URL's till the last clicked URL from feedback session. These feedback sessions are represented by, {fr1, fr2,…frn}. Next step is mapping these feedback sessions to pseudo documents. By mapping feedback session to the pseudo document we find out the exact user goals. So, pseudo documents are created as, {ps1, ps2,..psn}.
Finally, cluster of these pseudo-documents is perform to find out similarity,
{ps1=sg1,
sg2,…sgn|ps2=sg1,sg2,…sgn|…|psn=sg1,sg2,..sg}
Similarity computation,in given formula
simi,j=cos(Ffsi,Ffsj)
Where, Ffs is the feature representation of feedback session. After clustering all the pseudo-documents, each cluster is considered as one user search goal.

VII. CONCLUSIONS

The proposed feedback system used to improve search engine performance, also user can easily find out search goals for a query in short time. In this approach infer user search goals by clustering the feedback sessions. Feedback sessions consist of series of both clicked and un-clicked URL's before the last click. This click and un-clicked URLs is nothing but user's implicit feedback. Then feedback sessions are mapped with pseudo-documents which is generated from feedback session to approximate goal texts in users mind. These documents enrich URL's with additional contents including titles and snippets. Based on these documents search goals can be represented with some keyword. Finally the performance between restructured (clustered) web search results and original search results is evaluated by using CAP.

ACKNOWLEDGEMENT

Miss.Vidya V. Hire is thankful to **Prof Swati K. Bhavsar Mam**, Asst. Professor, Computer Engineering Department of Matoshri College of Engineering and Research Centre, Nasik for her constant support and helping out with the preparation of this paper. Also thankful to the Principal**, Dr.G.K.Kharate** Matoshri College of Engineering and Research Centre, Nasik and **Dr.V.H.Patil**, HOD, Computer Engineering Department of Matoshri College of Engineering and Research Centre, Nasik, for being a constant source of inspiration.

REFERENCES

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

[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] 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.

[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 toCluster Web Search Results*," Proc. 27th Ann. Int'l ACM SIGIRConf. 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] D. Shen, J. Sun, Q. Yang, and Z. Chen, "*Building Bridges for WebQuery Classification*," Proc. 29th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR '06),pp. 131-138, 2006.

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

[9] T. Joachims, "*Evaluating Retrieval Performance Using Clickthrough Data*," Text Mining, J. Franke, G. Nakhaeizadeh, andm I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[10] 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