INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

INFERRING USER SEARCH GOALS FOR A QUERY BY CLUSTERING THE USER FEEDBACK SESSIONS

M. Laxmi Narasimha Reddy [1], K.Vijay Bhaskar[2], Annepaka Yadagiri [3]

[1]MVSR College of Engineering and Technology, CSE Department, Hyderabad, India
Email: reddy.musku6@gmail.com
[2]Geethanjali Colleges of Engineering and Technology, CSE Department, Hyderabad, India
Email: vijaybhaskarchamp@gmail.com
[3]Geethanjali Colleges of Engineering and Technology, CSE Department, Hyderabad, India
Email: yadagiritinkuu@gmail.com

Abstract: - Nowadays Internet is widely used by users to satisfy various information needs. However, ambiguous query/topic submitted to search engine doesn't satisfy user information needs, because different users may have different information needs on diverse aspects upon submission of same query/topic to search engine. So discovering different user search goals becomes complicated. The evaluation and depiction of user search goals can be very useful in improving search engine relevance and user knowledge. This paper proposes a novel approach for inferring user search goals by analyzing user query logs from various search engines. The proposed approach is used to discover different user search goals for a query by clustering the user feedback sessions. Feedback sessions are constructed from click through logs of various search engines. The method first generates pseudo-documents to better represent feedback sessions for clustering. Finally, clustering pseudo-documents to discover different user search goals and depict them with some keywords. Then these user search goals are used to restructure the web search results.

***Keywords*—** User searches goals, implicit feedback sessions, pseudo-documents, restructuring search results, k-means clustering.

I. INTRODUCTION

In web based search applications, user submits the query to search engine to search efficient information. The information needs of different user may differ in various aspects of query information. This becomes difficult to achieve user information needs. Sometimes ambiguous queries may not exactly represented by users so it results in

less understandable to search engine. To achieve the user specific information needs many ambiguous/uncertain queries may cover a broad topic and dissimilar users may want to get information on different aspects when they submit the same query.

To satisfy the user information needs by considering the search goals with user given query, cluster the user information needs with different search goals. Because the interference and evaluation of user search goals with query might have a numeral of advantages in improving the

Search engine significance and user knowledge. So it is necessary to collect the different user goal and retrieve the efficient information on different aspects of a query. Capturing different user search goals related to information needs changes the normal query based information retrieval. Evaluation and analysis of user search goals has many advantages as follows.

- Reorganize web search results according to user search goals by grouping search results with same information need. This can be useful to other users with different search goals to find easily what they want.
- Query recommendation by using user search goals depicted with some keywords. This can be helpful to other users to form their query more effective.
- Re ranking web search results according to different user search goals.

II. LITRATURE SURVEY

Since many years, research in web log mining has been subject of interest. Many previous works has been investigated on problem of analyzing user query logs [5], [9], [10], [12], [13]. The information in query logs has been used in many different ways, such as to infer search query intents or user goals, to classify queries, to provide context during search, to facilitate personalization, to suggest query substitutes and to identify frequently asked questions (FAQs).Effective organization of search results is critical for improving utility and relevance of any search engine. Clustering search results is an effective way to organize search results which allows a user to navigate into relevant documents quickly. Generally all existing work [3], [17]

perform clustering on a set of top ranked results to partition results into general clusters, which may contain different subtopics of the general query term. However, this clustering strategy has two deficiencies which make it not always work well. First, discovered clusters do not necessarily correspond to the interesting aspect of a topic from user-oriented perspective. Second, cluster labels are more general and not informative to identify appropriate clusters. Wang and Zhai [2] proposed approach to organize search results in user-oriented manner. They used search engines log to learn interesting aspects of similar queries and categorize search results into aspects learned. Cluster labels are generated from past query words entered by users.

III. PROPOSED SYSTEM

In this section, basic operations involved in proposed approach to discover user search goals/intents by clustering pseudo-documents are described. The flow of the proposed

System design will be as shown in Fig. 1.

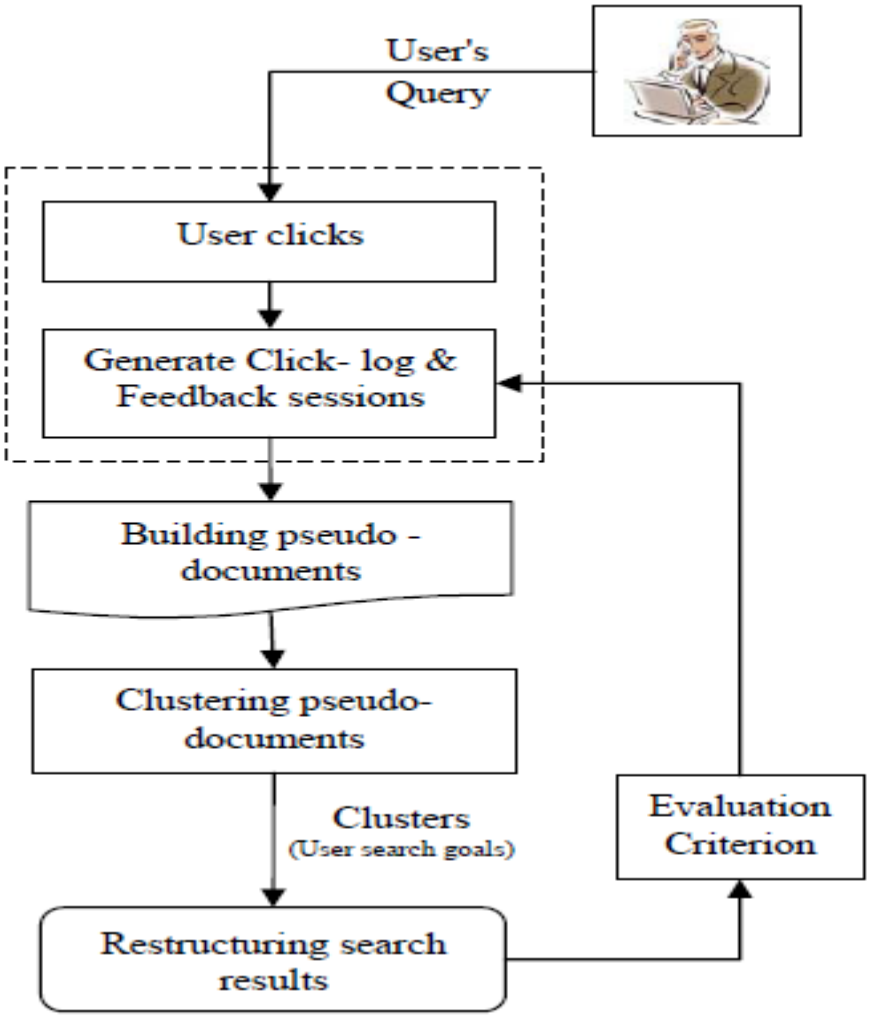

Fig. 1 Flow Diagram of Proposed System

A. Click through data

Click through data in search engines can be thought of as triplets (q,r,c) consisting of the query q, the ranking represented to the user, and the set c of links the user clicked on. Figure 1 illustrates this with an example: the user asked the query "support vector machine", received the ranking shown in Figure 1, and then clicked on the links ranked1, 3, and 7. since every query corresponds to one triplet, the amount of data that is potentially available is virtually unlimited.

Clearly, users do not click on links at random, but make a (somewhat) informed choice. While click through data is typically noisy and clicks are not "perfect" relevance judgments, the clicks are likely to convey some information. The key question is: how can this information be extracted? Before deriving a model of how click through data can be analyzed, let's first consider how it can be recorded.

B. Mapping Feedback Sessions to Pseudo-Documents

Feedback sessions are considered as users' implicit feedback. In general, a session for web search is a sequence of consecutive queries to satisfy single information and some clicked results. But to infer user search intents/goals for a particular query, single session is considered. Single session corresponds to only one query, which differs from

conservative session. The proposed feedback session consists of both clicked and unlocked URLs for a particular query in a single session and ends with last clicked URL. This shows that before last clicked URL, all the URLs are scanned and evaluated by user. Therefore, all clicked URLs and unclicked URLs before last click are considered as user feedbacks. In each feedback session clicked URL (visited link) tells users information need and unclicked URL (unvisited link) tells what users don't want. This visited link is called as positive feedback and unvisited link is called as negative feedback. There are large numbers of diverse feedback sessions in user click through log. So it is efficient to examine feedback sessions for inferring user search goals than to examine clicked URLs or search results directly.

Search Results	Click Sequence	Binary Vector
www.thesun.co.uk/	0	0
www.nineplanet.org/sol.html	1	1
www.solarview.com/eng/sun.htm	2	1
En.wikipedia.org/wiki/sun	0	0
www.thesunmagazine.org/	0	0
www.space.com/sun/	0	0

Fig: A Binary vector representation

C. Building pseudo-documents

As URLs alone are not informative enough to tell intended meaning of a submitted query. To obtain rich information, we enrich each URL with additional text content by extracting the titles and snippets of URLs appearing in feedback session. Thus, each URL in feedback session is represented by small textual content which contains its title and snippet. Then some text preprocessing is done on those textual contents, such as transforming all letters to lowercase, eliminating stop words (frequent words) and word stemming by using porter algorithm [16]. Lastly,TF-IDF [1] vector of URL's titles and snippets are formed respectively as,

Tui=[tw1,tw2,…….twn] T

Sui=[Sw1,Sw2,.....Swn]T

Where Tui and Sui are TF-IDF vectors of URLs title and snippet. ui means ith URL in the feedback sessions. twj and swj means TF-IDF value of the jth term in the URLstitle and snippet.
Fui = wtTui + wsSui = [fw1,fw2,….fwn]T Where Fui means ith URL in the feedback session wt and ws are weights of the titles and snippets.

Clustering pseudo-documents with K-means

Now next step is clustering of pseudo-documents with fuzzy k-mean clustering algorithm, the important factor is to define the distance measure between two data points as well as defining the number of clusters. Firstly representing each document using vector space model with the help of Tf-IDF value. As mentioned above the feature representation of pseudo-document is *Ffs* and similarity between two pseudo-documents is defined as below

$$Simi,j = \cos(Ffsi, Ffsj)$$

Here to cluster document, it is necessary to represent them in form of vector space model, for that here using TF-IDF value for each document. Cluster denotes user search goal i.e. intention of user and centroid of a cluster is calculated by taking average of all the vectors of the pseudo- documents in the cluster,

$$F_{centeri} = \frac{\sum_{k=1}^{C1} Ffsk}{Ci} (F_{fsk} \subset \text{Cluster } i)$$

F center i is ith cluster center and Ci is the number of pseudo-documents in the ith cluster is used represent user search goal/intent of ith cluster and *F centeri* to categorize the search results. User search goals/intents are the terms with highest values in the centre points of each cluster. These keywords can be used to suggest more meaningful labels of clusters.

D. Rearranging web search results

Reorganization of web search results are done on the basis of discovered user search goals which achieve by analyzing search results as mentioned above, inferred user search goals represents with vectors in (6) and feature representation of each URL in search result is calculated by (1) and (2) . By selecting the smallest distance between user search goal vectors and URL vectors categorizing each URL into a cluster centered with user search goals/intents. And finally rearranging links based on most visited links occur at topmost.

E. Evaluation criterion

To evaluate performance of restructured (clustered) web search results and original search results, using parameters like Average Precision (AP)

[1], Voted AP (VAP) which is AP of the class having more clicks, Risk to avoid wrong classification of search results and Classified AP (CAP). If user got correct classified results with higher AP value, this value is used to optimize the no of clusters of user search goals.

Average precision (AP): Calculated according to given user feedbacks. It is the average of precisions computed at the point of each clicked document in the ranked sequence of user feedback.

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

Where *N+* is the number of clicked documents from total retrieved documents in single user feedback session, *r* is the rank, *N* is the total number of retrieved documents, *rel()* is a

Binary function on the relevance of a given rank, and *Rr* is the number of relevant retrieved documents of rank *r* or less. Binary function on the relevance of a given rank, and *Rr* is the number of relevant retrieved documents of rank *r* or less.

***2)* Voted AP (VAP):** It is calculated for restructured search results classes i.e. different clustered results classes. It is same as AP and calculated for class which having more clicks i.e. the class user interested in.

$$VAP = \frac{1}{NC} \sum_{r=1}^{N} rel(r) \frac{R_r}{r}$$

Where NC is the number of clicked documents from the class having maximum number of clicks.

3) Risk: Sometimes VAP will always be highest value because each URL from single session is classified into the single class no matter whether users have different search goals or not. So, there should be a risk to avoid wrong classification search results into too many classes. It evaluates the normalized number of clicked URL pairs that are not in the same class.

***4*) Classified AP (CAP):** New criterion Classified AP(CAP) is extension of VAP by using above Risk. It combines AP of class having more clicks and risk of wrong classification. It is used to evaluate performance of restructured search results.

IV.CONCLUSIONS

The proposed system can be used to improve discovery of user search goals for a query by clustering user feedback sessions represented by pseudo-documents. Using proposed system, the inferred user search goals/intents can be used to restructure web search results. So, users can find exact information needed as they want very efficiently. The discovered clusters can also be used to assist users in web search.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] T. Joachims, "Evaluating Retrieval Performance Using Click through Data", Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.
[8] 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, pp.502-513,2013.
[9] 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.
[10] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.
[11] 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.
[12] 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.
[13] 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.

Authors:

M Laxmi Narasimha Reddy is presently working as Asst Professor, CSE Dept, M V S R Engineering College, Hyderabad, and Telangana, India. He received his M.Tech in CSE from JNTU; Hyderabad in the year 2010.His interest areas includes Data Mining, Networking, Software Engineering etc.

K.Vijay Bhaskar is presently working as Asst Professor, CSE Dept, Geethanjali College of Engineering and Technology, India. He received his M.Tech in CSE from JNTU, Hyderabad. His interest Research areas includes Data Mining, Networking, Software Engineering etc.

Annepaka Yadagiri is presently working as Asst Professor, CSE Dept, Geethanjali College of Engineering and Technology, India. He received his M.Tech in CSE from JNTU, Hyderabad. His interest Research areas includes Data Mining, Networking, Software Engineering etc