International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 5, Number 1 (2015), pp. 19-26

http://www. irphouse.com

Enumerating User Search Goals with Clustered Feedback Sessions

Jyoti Kumari[1] , Teppala Satyanarayana[2], Monalisa Lenka[3], Priyanka Yadav[3], B. Srinivasa Rao[4]

1M.Tech Student, Dept. of CST, GITAM University, Visakhapatnam, A.P, India
2M.Tech Student, Dept. of CSE, MIRACLE College, Visakhapatnam, A.P, India
3M.Tech Student, Dept. of IT, GITAM University, Visakhapatnam, A.P, India
4 Assistant Professor, Dept. of CSE, GITAM University, Visakhapatnam, A.P, India

ABSTRACT

In web search applications, queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. An approach to Infer user search goals by analyzing search engine query logs. A framework is proposed to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click through logs and efficiently reflect the information needs of users. An approach for generating the pseudo documents for better representation of feedback sessions. Finally, we propose a new criterion "Classified Average Precision" to evaluate the performance of inferring user search goals. The main aim is to provide web search results based on the user feedback. This user feedback is very useful to improve the search engine.

Index Terms—User search goals, feedback sessions, pseudo-documents, restructuring search results, K-mean clustering and classified average precision (CAP).

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 re-

sults 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. For example, when the query “the sun” is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun, as shown in Fig. 1. Therefore, it is necessary to discover different user information search goals. User information need is to desire and obtain the information to satisfy the needs of each user.

Figure 1: Example of different user objectives for” the sun” 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.
- Reranking web search results according to different user search goals.

User search goal analysis is important to optimize searchengine and effective query results organization. Whenquery is submitted to search engine, the returned web pagesof search results are analyzed [3], [4]. Since it does notconsider user feed-

back, many unuseful and noisy searchresults that are not clicked by user may be analyzed. Thismay degrade the search goals discovery. X. Wang and C-X.Zhai [2] learns interesting aspects of similar query/topicfrom web search logs which consists clicked web pagesURLs and organize search results accordingly. Theirapproach may results in limitation, as the different clickedURLs for a query/topic may be small in number. There aremany works [11], [12] which classify queries into somepredefined specific classes and try to find out query intentsand user goals. However, different queries have differentsearch goals and finding precise, suitable predefined searchgoal classes may be difficult and sometimes impossible tocategorize.

OVERVIEW:

The rest of the paper is organized as follows: Section II contains literature review about related work. Section III contains description of the system design. Finally paper is concluded in the Section IV.

II. LITERATURE REVIEW

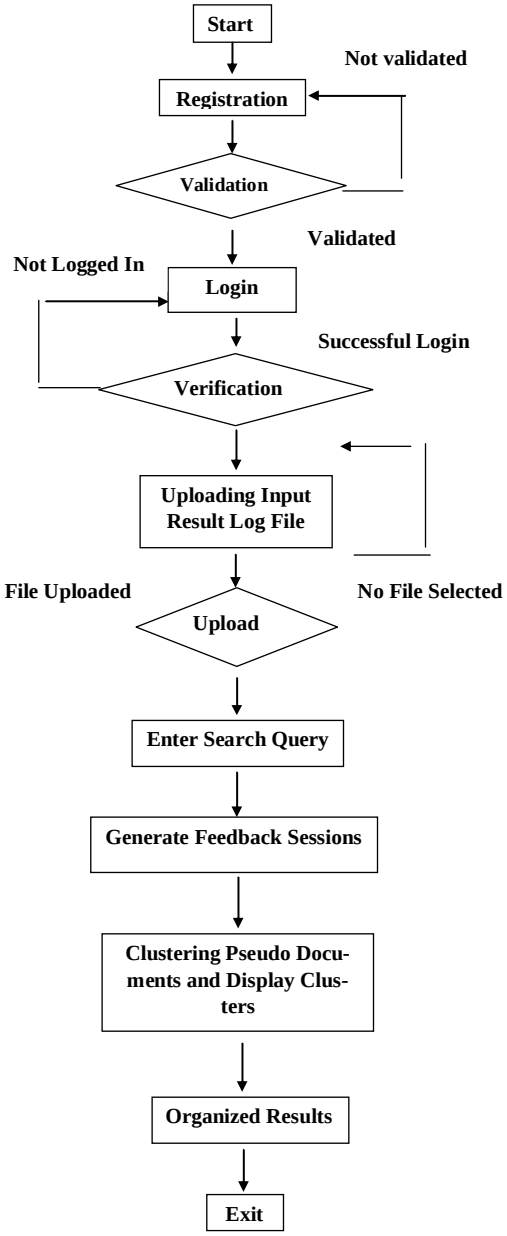
Enumerating user search goals can be very useful in improving search engine relevance and user experience.Since many years, research in web log mining has been subject of interest. Many previous works has been investigated on problem of analyzing user query logs [5], [8], [9], [11].

- Lu et al.[1] proposed a novel approach to infer user search goals by analyzing search engine query logs. First, they have proposed a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, they have proposed a novel approach to generate pseudo-documents to represent the feedback sessions. Finally, they have proposed a new criterion “Classified Average Precision (CAP)” to evaluate the performance of inferring user search goals.
- Wang and Zhai [2] proposed approach to organize search results in user-oriented manner. They showed that log-based method can consistently outperform cluster-based method and improve over the ranking baseline, especially when the queries are difficult or the search results are diverse.
- H-J Zeng et.al [3] proposed a query based method to cluster search results. But this method only produces the result with higher level of the documents only and it doesn’t make the results for all search based user goals.
- H. Chen and S. Dumais [4] developed a user interface that organizes web search results into hierarchical categories. This approach has advantage of known category labels information, for classifying new items into the category structure and to help user to quickly focus on task relevant information.
- T. Joachims [5] proposed an approach to automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking. Taking support vector machine (SVM) approach, for learning ranking functions in information retrieval.

III. PROPOSED SYSTEM

In this paper, we aim at 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 a query by clustering our proposed feedback sessions. Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Then, we cluster these pseudo documents to infer user search goals and depict them with some keywords. At last, we organize the result according to the weight i.e. number of clicks the URL has got.

System Flowchart:



I. Representing Clickthrough data

In web search environment, there are many abundant queries and user clicks. User clicks represent implicit relevance feedback. In this framework, user clicks are recorded in user clickthrough data. User uses clickthrough data stored in user logs to simulate user experience in web search. In general, when query is issued, the user usually scans links to documents in a result list from first to last.Clearly, the user clicks on the links to the documents thatlook relevant of informed choice and skips other documents.

Therefore, the proposed approach utilize user click asrelevance judgments to evaluate search precision sinceclickthrough data can be collected at low cost, it is possibleto do large scale evaluation under this framework.

1) Constructing Feedback sessions:

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 unclicked 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 whatusers do not 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 clickthrough log. So it is efficient to examine feedback sessions for inferring user search goals than to examine clicked URLs or search results directly.

II. Building pseudo-documents using feedback sessions.

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,

$T_{ui} = [t_{w1};t_{w2};...;t_{wn}]^T;$

$S_{ui} = [s_{w1};s_{w2};...;s_{wn}]^T;$

where $T_{ui}$and $S_{ui}$are TF-IDF vectors of URL's title and snippet, respectively. ui is ith URL in feedback session. W_j is the jth term in the enriched URL. The t_{wj} ands_{wj}denotes jthterm in the URL's title and snippet respectively. Feature representation F_{ui}, of ith enriched URL is weighted sum ofT_{ui} andS_{ui}.

$$F_{ui} = w_t T_{ui} + w_s S_{ui} = [f_{w1}, f_{w2}, \ldots\ldots, f_{wn}]^T$$

where w_t and w_s are weights of title and snippet respectively. Each term of F_{ui}, denotes importance of term in ith URL.

In order to obtain feature representation of a feedback session, optimization method is used to merge feature representations of each clicked and unclicked enriched URLs in the feedback session. Let F_{fs}be feature representation of a feedback session, and are feature representation of clicked and unclicked URLs respectively and$F_{fs(w)}$is value for term w. F_{fs}should be such that sum of distance between $F_{fs}$and each is minimized and sum of distance between $F_{fs}$and is maximized.

$$F_{fs} = [f_{fs}(w_1), f_{fs}(w_2), \ldots.., f_{fs}(w_n)]^T$$

Each feedback session is represented by F_{fs}. This is nothing but pseudo-document which is used for discovering user intents or search goals. These pseudo-documents contain what user requires and what do not, which is used to learn interesting aspects of a query.

III. Clustering pseudo-documents with K-means

In order to cluster pseudo-documents with k-means, the important factor is to define the distance measure between two data points and defining the number of clusters. There are two variations of distance measures, one is derived from cosine based similarity and the other is derived from Jaccard similarity coefficient. The feature representation of pseudo-document is F_{fs}. The similarity between two pseudo-documents is defined as below:

$$Sim_{i,j} = \cos(F_{fsi}, F_{fsj}) =$$

OR, (4)

$$Sim_{i,j} = \cos(F_{fsi}, F_{fsj}) =$$

And the distance two feedback sessions i.e. pseudodocuments is

$$Dist_{i,j} = 1 - Sim_{i,j} \quad (5)$$

K-means algorithm is used to cluster pseudodocuments because of its simplicity and effectiveness. K-means clustering results in good quality performance for document clustering. As a prior number of user search goals for a query are unknown so we have chosen arbitrary value for k initially (i.e. 1, 2, 3, 4, 5). Then, perform clustering on these five different values. The optimal value for k is determined by evaluation criterion.

After clustering all pseudo-documents, each cluster denotes user search goal i.e. intention of user. Centroid of a cluster is calculated by taking average of all the vectors of the pseudo- documents in the cluster,

$$F_{centeri} = \quad (6)$$

where$F_{centeri}$is ith cluster center and $C_i$is the number of pseudo-documents in the i^{th} cluster. $F_{centeri}$is used represent user search goal/intent of i^{th} cluster and to categorize the search results. User search goals/intents depicted with the terms with highest values in the center points of each cluster. These depicted keywords can be used to suggest more meaningful and precise query.

IV. Restructuring web search results

Web search results are reorganized on the basis ofdiscovered user search

goals/intents. As inferred user search goals are depicted with vectors in (6) and feature representation of each URL in search result is calculated by (1) and (2). Then categorize each URL into a cluster centered with user search goals/intents by selecting smallest distance between user search goal vectors and URL vectors.

V. Evaluation criterion

The performance of restructured (clustered) web search results and original search results is evaluated by using parameters such as:

1) ***Average precision (AP):***

It is calculated according to given user feedbacks. AP is the average of precisions computed at the point of each clicked document in the ranked sequence of user feedback.

$$AP= 1 / N^{+}\sum rel(r)\ R_r \quad r, \quad r=1$$

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.

2) Voted AP (VAP):

We calculate this for restructuredsearch results classes i.e. different clustered results classes. It is similar to AP and calculated for class which having more clicks i.e. the class in which user is interested.

$$VAP = 1/NC$$

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

3) Risk:

Sometimes VAP will be the highest value as each URL from single session is classified into the single class which is independent of user search goals. So, there should be a risk to avoid wrong classification search results into too many classes. It enumerates the normalized number of clicked URL pairs which are not in the same class.

$$Risk =$$

where m is number of clicked URLs

Also, $d_{ij} = 0$, if pair of clicked URLs belongs to same class $d_{ij} = 1$, otherwise.

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 enumerate performance of restructured search results.

$$CAP = VAP \times (1\text{-}Risk)^{\square}$$

where γ is normalizing factor used to adjust influence of Risk on CAP.

IV. CONCLUSION

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.

REFERENCE:

[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*"- IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, PP. 502-513, MARCH 2013.

[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.