ISSN: 2320 - 8791
www.ijreat.org

Improve Search Engine Relevance with Filter session

Addlin Shinney R[1], Saravana Kumar T[2] and Roslin Mary M[3]

[1]**Information Technology, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, 628215, India**

[2]**Information Technology, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, 628215, India**

[3]**Information Technology, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, 628215, India**

Abstract

Nowadays every individual are accessing web to find out the needed information. Usage of web rapidly increases and for that the search engine results should be reorganized for easier user interaction For web search user have different goals for different query while submitting into the search engine. Inference of user goal for a query helps to improve search engine relevance by analyzing query logs. In this paper a framework is proposed to identify different user search goal for a query by clustering filter session. Filter session is generated based on user clicked logs that reflect user needs. Feedback contains both clicked and un-clicked URL's Second, propose a new method to generate precise text to appropriate representation of filter session for clustering. Mapping filters session to precise text to find goal text in user mind. Finally propose a novel Criterion "Assorted Average Precision" to determine the performance of inferring user search goals.

Keywords: *User search goal, Filter session, precise text, restructuring search results, Assorted Average Precision.*

1. Introduction

In web application, user has to submit their query to the search engine. Search engine list out the results related to that particular query. Different user's wanted to get different aspects of information for a same query. For example, user submits query as "apple", some user want to view about the apple product and some user want to know about the nutrients of the fruit apple. So it is important to know the user search goal for a query. User search goal is defined as an information need of the user .Information needed by the user is based on their desire and that should satisfy their need. Lot of advantages is there for analyzing user search goal .Search goals are represented by the cluster of information needs. Advantages are summed as follows. First, reorganizing search results based on the user search goal and similar search results are placed on the same cluster; by this type of representation user can easily find the information they want. Second,Keywords are used to represent different user search goals. Keywords are used for query recommendation and the indicated query helps the user to find their queries more precisely. Third, re-ranking web search results based on the distribution of user search goals for a query.

Many works are investigated for its usefulness. That is categorized into three major classifications. Query categorization, restructure search results, and task boundary identification. In the first class, user tries to infer search goal by defining some classes and perform query categorization. [1] To identify user search goal query is classified into "informational' and "navigational". For informational type of query user refers many pages to get their information needs. For navigational type of query user have a predefined webpage in their mind and want to visit that particular page. They have proven that majority of the query have a predictable goal. They analyze whether and how predictable goals depicted. The purpose of human study is to check for the feasibility of automatic identification of user goal. Two features are there to automatic identification of user search goal. That is user –click behavior and the next one is anchor-link distribution. If the goal of the query is navigational, then most of the past users clicked on the single page. If the goal of the query is informational the most of the users clicked on multiple pages. So to find out the user goal they introduce click distribution. This click distribution captures how frequently the users clicked on multiple answers. Then they normalized the click distribution to find out the user goal. For anchor-link distribution there is an anchor text. To determine the goal for the navigational query, extract the destination

of the extracted URLs and find user goal. For navigational query extracted URLs majority focused on single website. For informational query, extracted URLs mainly focused on multiple pages. From multiple pages user goal has to be finding out by anchor list distribution. The result majority clicked by the user is predicted as user goal. Limitation is finding a predefined search classes are not get easily achieved.

Second class; restructure search results [2] finding interesting aspects for a query from the user click-through logs. Based on the logs the resulting web pages are reorganized. For given a query interesting aspects are determined and did text classification that automatically classify search results. For that the text classifier has to train offline, and then it will classify web pages on the fly. For doing test classifier they did three steps. The first one is data set. During the training phase web pages with known labels are used to train the classifier. Training purposes collects the data sets and train the classifier. The second one is pre-processing, here they extract plain text from web pages and indicate each keyword, figure, title with a vector. The search engine returns the result that contains a short description of the results. The user who needs the entire content can download the whole document. This method is time consuming. For each result it generates a short summary of document. The last one is classification; Support Vector Mechanism (SVM) was used as a classifier. It is very fast and effective in text classification. This SVM has a hyper plane that divides the category of search result. This classifier maximizes the margin between each category and user can easily identified their needs. User interface, it accepts the keyword given by the user and passed it to the search engine and it parsed the returned pages. The search result is categorized in hierarchical order. Under each category related web pages are organized. The category is expanded based on user interest. By clicking the title the results are showed in another window. Drawback is while analyzing the logs the noisy results are also get analyzed. So this method did not find the user goal more precisely.

In the third class, [3] try to find the task boundaries to predict the task goals and mission of the query. In that work session are manually labeled and the task are categorized into hierarchical way. First for automatic segmentation find the task and subtask for the query. If task and subtask were easily found then performance of search engine can be evaluated. For that elapsed time is fixed and based on that manually identify goal. Limitation is this method doesn't care about user goal in detail and for involved task get interrupt then to model the task is difficult.

In this paper, we try to find out different type of user search goal and representing each goal with a keyword automatically. For that first we proposed a method to identify user search goal by clustering filter sessions. Feedback is nothing but the combination of both clicked and un-clicked URL of the resulting web pages and that session is ended with the last URL that was clicked on that session. Then we propose a method to map the filter session to the precise text that exactly represents user need. At last cluster this document to find the user search goal and represent each cluster with keywords. To evaluate the clustering we propose a Assorted Average Precision (AAP) to determine the performance of the reorganized search results. This new criterion helps to reflect the user information needs. The overall methodology of our project is given below.

Sum up of our work as follows:

- We propose a framework to find out the user search goal for a query. This is done by clustering filter session. This method is more efficient than clustering web search results and clicked URLs. After filter sessions are clustered we can obtain distribution of different user search goals.
- We propose a new method to combine the URLs in the filter session. Based on that pseudo- document are generated which exactly reflect the information need of the user.
- We propose an AAP to examine the performance of user search goal. Thus we can determine different user search goals for a single query.

The rest of the paper organized as follows: The model of our approach is in section 2.The filter session and its representation in section 3. Section 4 find user goal by clustering precise text. Section 5 reviews several related works. Section 6 concludes the paper.

2. Model of Our Approach

Fig 1 shows the model of our work. First the original web search result. To reorganize the resulting search result first the filter session are taken from user clicked logs and then map to precise text. Depicting each with a keyword and finally based on that resulting web pages are restructured.

ISSN: 2320 - 8791
www.ijreat.org

3. Filter session and its Representation

In the section, describes the filter session and precise text to represent filter sessions. This precise text will exactly helps to predict the information need of the user.

3.1 Filter sessions

Basically, a session is a series of consecutive queries to satisfy single information need. In this paper we try to infer search goal for a specific query. Single session have a single query is introduced. Meanwhile the filter session we considered is a single session and further we can extend this paper to the whole session.

The filter session contains both clicked and un clicked URLs and finish the session with the last URL clicked in the single session. Before the final click all URLs are scanned by the user. So we consider both the clicked and un-clicked URLs before the last click on the session. Fig.2 illustrates the single filter session. The left section shows the 8 search results for the query "apple". The right section shows the click sequence and "0" shows un-clicked. From the fig, single session has 8 URLs and only 6 URLs are considered for filter session. The filter session is indicated by the red color rectangular box. The six URLs consist of 3 clicked and 3 un-clicked URLs. Generally user will scan all the URLs of the search result web page and reasonably the 3 URL in the rectangular box also evaluated by the user.

SEARCH RESULTS	CLICK SEQUENCE
www.apple.com/in/	0
en.wikipedia.org/wiki/apple_Inc	1
en.wikipedia.org/wiki/apple	0
www.forbes.com/companies/apple	2
www.fruits.com/apple	0
www.bloomberg/quote/apple	3
bgr.com/apple	0
Techcrunch.com/apple	0

Fig. 2. A filter session for a single session

Each filter session that reflects user needs and user don't care about the information. There is plenty of filter session in user logs. This method is more efficient to infer user search goal compared with search logs.

3.2 Generate Precise Text

Based on user click logs and queries, filter session varies a lot. So inferring user search goal only by referring filter session is not suitable. Some representation method is needed to represent filter session in a coherent way. To represent filter session there are many ways.

SEARCH RESULTS	Click Frequency	Binary Vector
www.apple.com/in/	0	0
en.wikipedia.org/wiki/apple_Inc	1	1
en.wikipedia.org/wiki/apple	0	0
www.forbes.com/companies/apple	2	1
www.fruits.com/apple	0	0
www.bloomberg/quote/apple	3	1

Fig. 3. The binary vector method

Fig. 3. Shows the binary vector representation of URLs returned for the query "apple". For binary vector clicked URLs are represented by "1" and un-clicked URLs are denoted by "0". This filter session has the binary vector [010101]. Different filter session has different number of URLs and based on that the filter session also gets changed.

Binary vector is not so efficient to predict user search goals. So it is not suitable to use binary vector representation and we need to introduce new method to represent filter session.

For a single query, user has a vague representation of keywords. Using the keyword to check whether the retrieved document satisfy their needs. Key words are named as "target text". These target texts reflect the information need of the user but not done in explicit way. So we introduce a precise text that helps to identify user information. A new method is introduced to map the filter session to precise text. It includes two steps to build the precise text showed in Fig.4.

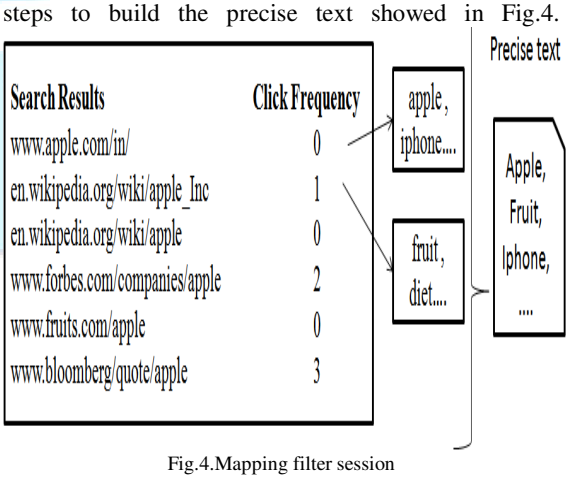

Fig.4.Mapping filter session

They are described in the following way:

1. Presenting the URLs in the filter session. In the first step we first select the URLs and enrich it by adding text content that are taken from the filter session. In the

same way each URLs in the filter session are described by small text summary. That summary contains both tittles and snippet. Then we have to apply some text process to the text summary. The text processes are transforming all the letters to lowercases, stemming and removing stop words. The title and snippet represented by Inverse Document Frequency. The feature representation of the URLs in the filter session and for the title, snippets assign weights. For weights of snippets is set to 1 initially. Then we specify titles are more significant than the snippets. Therefore we assign weight of the title is 2 more than the snippets.

2. Generating precise text based on URL representation. To find the feature representation of the filter session considers both clicked and un-clicked URLs. While doing search user skip some URLs because that is similar to that of the previous one. In those situations the un-clicked URLs wrongly reduce the weight of the precise text. Our method solve the problem in three cases: The first case is the ideal case that one term appears in all the clicked URLs and not appear in the un-clicked URLs. In the case people won't prefer the un-clicked URLs, because it doesn't have important information. The second case is defined as general case. In this case the term appears in clicked and subset of un-clicked URLs. User skips this because of duplication. Skipping doesn't affect this case. The third class is the bad case. In this case the term appears in both the clicked and all the un-clicked URLs. People skip because of duplication, hence this method assign reasonable weight. Then we find the user search goal based on the precise text.

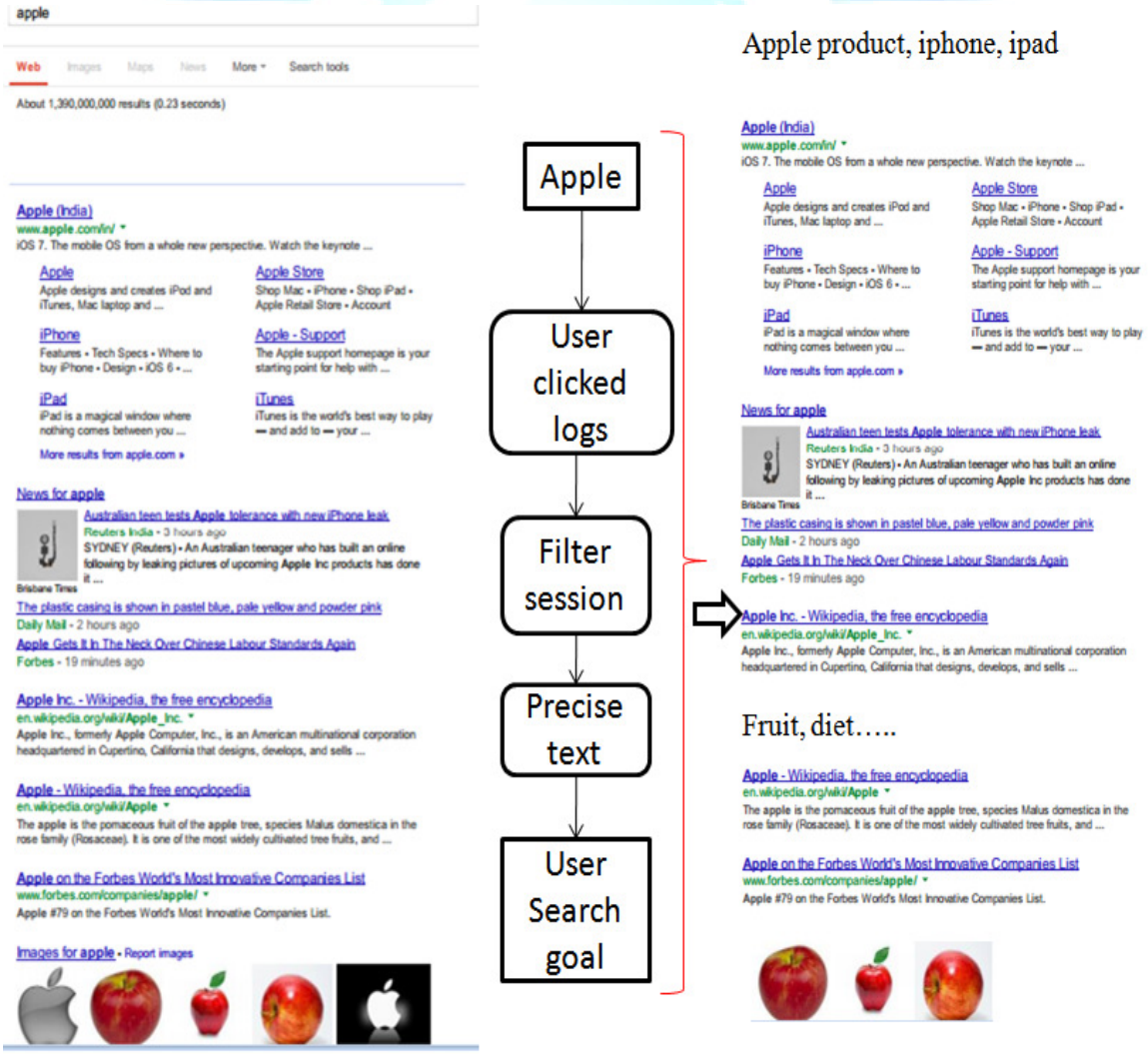

Fig.1. The model of our approach

Precise text is discovered next we find the different goal text for each query. Based on these 3 cases we assign weights foe the title and snippets present in the precise text.

4. Find User Search by Clustering Precise Text

In this section, find user search goal and depict them with more meaningful keywords. We cluster precise text by k-means method. This k-means clustering method very simple and effective. Since we don't know the exact number of user search goal for each query. So we consider k=5, five different values and perform clustering based on the five values. Each cluster has a user search goal and the center point of the cluster has to find. Finally the value with the highest point is considered as the keyword to depict user search goal. Another advantage of using keyword is, it helps the user to form the query in query recommendation and represent information in efficient way. Finding useful distribution for the user search goals and we obtain the number of filter sessions.

5. Evaluate Web Search Results

This type of evaluation is a big problem because user goals are not yet predicted. Correct numbers of clusters are not determined yet. The filter information is needed to find the best cluster number. So we have to develop a new metrics to evaluate the performance. First user goal has to be determined and based on that the web pages are reorganized. If user goal evaluated correctly then we can easily restructure the resulting web page. It is the one of the application finding user search goal. A method Assorted Average Precision (APP) helps to find the performance of web search results. We also describe the new method to find the correct number of clusters.

5.1 Reorganize Search Results

Always search engines return millions of search results. So it is important to organize the returned search results and make the user easily find the information that they need, feel them convenient to use the search engine. Reorganize search result is an application to find user search goal, then based on the restructured search result the performance is evaluated.

5.2 Evaluation Measure

Single session is considered to minimize the manual work. Because from user clicked logs we get relevant and irrelevant feedback. The clicked URLs are relevant and un-clicked URLs are irrelevant. An evaluation based on the user feedback, ranked the relevant document. This is also an unsatisfactory, so we have to avoid the risk while classifying search results.

6. Related Work

Recent years work is focused to find user search or user intents. But they focused on query classification. [4] Another work on the search result directly to restructure web pages to find different query aspects. This method had a limitation that to improve search engine. [5] For navigational type of query user have a predefined webpage in their mind and want to visit that particular page. They have proven that majority of the query have a predictable goal. They analyze whether and how predictable goals depicted. The purpose of human study is to check for the feasibility of automatic identification of user goal. Two features are there to automatic identification of user search goal. That is user –click behavior and the next one is anchor-link distribution. If the goal of the query is navigational, then most of the past users clicked on the single page. If the goal of the query is informational the most of the users clicked on multiple pages. So to find out the user goal they introduce click distribution. This click distribution captures how frequently the users clicked on multiple answers. Then they normalized the click distribution to find out the user goal. For anchor-link distribution there is an anchor text. To determine the goal for the navigational query, extract the destination of the extracted URLs and find user goal. For navigational query extracted URLs majority focused on single website. For informational query, extracted URLs mainly focused on multiple pages. From multiple pages user goal has to be finding out by anchor list distribution. Some work focused on directly on user-clicked logs. Based on the logs the resulting web pages are reorganized. For given a query interesting aspects are determined and did text classification that automatically classify search results. For that the text classifier has to train offline, and then it will classify web pages on the fly. For doing test classifier they did three steps. The first one is data set. During the training phase web pages with known labels are used to train the classifier. But it is not a good idea because different clicked URLs are there and from that getting an ideal result is very difficult. Another work on interesting aspects that solve some problem in part. However this method is not suitable for user search goal on a single query. For example for the query "car" it is clustered into "car rental", "used car", "car crash", "car audio" . From this method different aspects can be learned. But under the category "used car" [6] it also have some category and it is difficult to learn. Another method is to find the session manually and label in

hierarchical segment. In that work session are manually labeled and the task are categorized into hierarchical way. First for automatic segmentation find the task and subtask for the query. If task and subtask were easily found then performance of search engine can be evaluated. For that elapsed time is fixed and based on that manually identify goal. For that the session goal and mission has to find out. For involved task, if any interrupt is there model the task is very difficult. The utilization of click-logs is to obtain user needs to maximize training data. Its application is restructuring web search results.

7. Conclusion

In this paper, a new approach is introduced to find user search goal. That is done by clustering filter session represented by precise text. First we introduced a filter session to find the user search goal. For the case of considering search results or user-logs our method is efficient. Comparing with all the methods we explained in the above section our method effectively does restructuring search result web pages. This filter session effectively reflect user needs. Second, we map filter session to precise text to appropriate target text in user mind. The precise text is formed by adding some title and snippets. Based on that user target text is identified and depicting each of them with a keyword. Finally the reorganized search results are evaluated for its performance.

References

[1]. Allen, R. B., Two digital library interfaces that exploit hierarchical structure. In Proceedings of DAGS95: Electronic Publishing and the Information Superhighway (1995).

[2]. Chakrabarti, S., Dom, B., Agrawal, R., and Raghavan, P. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. The VLDB Journal 7,(1998).

[3]T.Joachims. Optimizing search engines using clickthrough data. In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2002.

[4] M. Just and P. Carpenter. A theory of reading: From eye fixations to comprehension. Psychological Review, 87:329–354, 1980.

[5] B.J. Jansen and U. Pooch. A review of Web searching studies and a framework for future research. J. of the American Society of Information Science and Technology, 52(3):235 { 246, 2001.

[6] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In Proceedings of ACM SIGKDD '00, 2000.

[7] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[8] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 41-50, 2008.