An Efficient Web Searching Algorithm for Inferring User Search Goals with Feedback Sessions

Ajith Chandran¹, Pankaj Kumar², Shyny S.T³

¹M.Tech Scholar, ²³Asst. Professor
¹³School of Computer Sciences, M.G University, Kerala, India
²Computer science Dept. FISAT, Kerala, India

Abstract-- Search Engines are the major data source of the current world. In this situation performance have very much importance. When a user submits queries to a search engine it returns millions of results. From these results users will select what they want. Sometimes they may achieve their goals from the first attempt, sometimes may not. In this paper we propose an approach to find the user search goals by analyzing search engines query logs and search history of the current user. In the first step creating a feedback session from given query and the next step is conversion of feedback session to a useful form called pseudo document. Then create the clusters of pseudo document to obtain the user search goals. In this clusters organize the result based on page ranking. Finally optimization of the number of clusters based Classified Average Precision (CAP) Evaluation.

Keywords-- Feedback Session, Pseudo Document, Average Precision, Weighted K-means clustering Algorithm.

I. INTRODUCTION

In the area of web mining, more importance is given to fast and accurate extraction of information. Query suggestions provided by the search engine will help to find the user needs. But it may cover broad topics so this may not be solution for achieve a better search result. Also same queries have different goals for different users. The analysis of user search goal improves the relevance and user satisfaction of the search engine. We propose method to analyze the user query and restructuring of the search results. This approach has following advantages. First, we can find the search goals immediately. Second Restructuring of search result. Third, it solves the problems of handling the large number of search results. Finally, user feedbacks have more importance in this method.

Many works about user search goals have been done in the past years. These works are targeted to the user search goal identification; one of the attempts is classification of user query [2]. Some other 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]. Another approach is optimization of the retrieval quality of search engine [6]. Third approach reorganizing search results, Wang et.al proposed a method organize search result by analyzing the clicked URL’s directly from user click through logs [8]. All these works will not consider any user feedbacks so these kinds of works not infer exact user search goals. Zheng Lu et.al proposed a method that using the user feedback and reorganize search result. But they are not considering the current users history in separately, also they are not mentioning the arrangement of search result inside a cluster.

In our approach, priority is given to the current user’s history; also page ranking algorithm is used to arrange the search results inside a cluster. Our method uses the user the user feedback to reduce the number of clusters. When the number of links inside a cluster is too large, it can be solved by clustering inside the cluster. Hence it will result more accurate search results. CAP evaluation [1] is used to evaluate the performance of restructured web search results. In this paper the content organized as follows: The Related works are described in the next section, in the third section detailed description of method of approaches, in the fourth section experimental results and section five concludes the paper.

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 click-through logs and session data; we are also utilizing the click-through 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.

<table>
<thead>
<tr>
<th>Search Result</th>
<th>Click Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.space.com/">www.space.com/</a></td>
<td>1</td>
</tr>
<tr>
<td>en.wikipedia.org/wiki/</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.nasa.gov/">www.nasa.gov/</a></td>
<td>2</td>
</tr>
<tr>
<td><a href="http://www.space.ca/">www.space.ca/</a></td>
<td>3</td>
</tr>
<tr>
<td><a href="http://www.isro.org/">www.isro.org/</a></td>
<td>4</td>
</tr>
<tr>
<td><a href="http://www.bbc.co.uk/space/">www.bbc.co.uk/space/</a></td>
<td>0</td>
</tr>
<tr>
<td>amazing-space.stsci.edu/</td>
<td>0</td>
</tr>
</tbody>
</table>

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. TF-IDF 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 is 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).

\[
AP = \frac{1}{N^+} \sum_{r=1}^{N^+} \text{rel}(r) \frac{R_r}{r} \tag{1}
\]

In Eq(1), N+ is the number of clicked URL’s in the cluster, r is the rank, n is the total number of URL’s in the cluster rel() is a binary function on the relevance of the given rank, R_r is the number of clicked URL’s of rank r or less. Fig. 3 shows user feedback from each cluster with click sequence and AP estimation. Here cluster 1 with AP value 0.83 and cluster 2 with AP value 1.0. AP of the session is estimated on the basis of AP value of the cluster with most number of clicked URL’s. In the case of two clusters with same number of clicked URL’s, largest value of AP is selected as AP of the session [11]. There should be a risk to avoid classifying the search results in too many clusters by error.so the Risk factor should be calculated with Eq(2).

\[
\text{Risk} = 1 - \frac{\sum_{C} c_k^2}{c^2_{\max}} , \quad n>1 \tag{2}
\]

Where C is the total number of clusters, n is the number of clicked URL’s in each selected clusters, M is the sum of clicked URL’s of all the selected clusters, c_k^2 = \frac{n(n-1)}{2} pair of clicked URL’s in each selected cluster and value of n should be greater than 1.This estimated risk value is multiplied with AP of the session will give a new value and this criteria is called Classified AP [1].
From (3) we have to obtain the user interested cluster and takes the risk of wrong selection of clusters into account.

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.

IV. EXPERIMENTAL DESIGNS

The Experimental design of our proposed method is given in fig. 4. In this figure feedback sessions of a query are first extracted from the available user click-through logs. This feedback session is mapped to pseudo document. The user search goals are inferred by clustering the pseudo document, the exact user goals are obtained in the user feedback evaluation. The obtained search result is reorganized on the basis of page ranking. Then we evaluate the performance of search result based on the CAP evaluation. The evaluated result will select the optimal search goals. The optimized search goals are again clustered and there by the sub search goals obtained. This will help to find out the user search goals.

### Cluster 1

<table>
<thead>
<tr>
<th>Search Result</th>
<th>Click Sequence</th>
<th>( rel(r) \frac{R_r}{r} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.space.com/">www.space.com/</a></td>
<td>1</td>
<td>( \frac{1}{1} )</td>
</tr>
<tr>
<td>wikipedia.org/wiki/</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.space.ca/">www.space.ca/</a></td>
<td>3</td>
<td>( \frac{2}{3} )</td>
</tr>
</tbody>
</table>

\[ AP = \frac{1}{2} \left( \frac{1}{1} + \frac{2}{3} \right) = 0.83 \]

### Cluster 2

<table>
<thead>
<tr>
<th>Search Result</th>
<th>Click Sequence</th>
<th>( rel(r) \frac{R_r}{r} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.nasa.gov/">www.nasa.gov/</a></td>
<td>2</td>
<td>( \frac{1}{1} )</td>
</tr>
<tr>
<td><a href="http://www.isro.org/">www.isro.org/</a></td>
<td>4</td>
<td>( \frac{2}{2} )</td>
</tr>
</tbody>
</table>

\[ AP = \frac{1}{2} \left( \frac{1}{1} + \frac{2}{2} \right) = 1.0 \]

AP of the session= 1.0
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.

V. 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.

ACKNOWLEDGMENT

This work was conducted in Centre for High Performance Computing lab FISAT, Angamaly, Kerala and all facilities provided by M G University, Kerala. We express our special thanks to Dr. R. Vijayakumar, Professor, School of Computer Sciences, MG University, Kerala.

REFERENCES

699-708, 2008


