ISSN:0975-9646

A Compressive Survey on Restructuring User Search Results by Using Feedback Session

Shinde Sonali Bhaskar

Dept.of Computer Networking
Flora Institute of Technology
Pune,India

***Abstract*— this internet search engine relevance may be enhanced by means of considering end user search goal. In addition to the individual search engine optimization experience is usually increased through inferring individual search goals. This paper proposes a novel approach to infer user search goals by analyzing search engine query logs known as feedback session. First 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 efficiently. Second a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. In the proposed technique we are going to implement Semantic Clustering algorithm to improve the results as compare to other techniques. Finally, a new criterion "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals.**

***Index Terms*— feedback sessions, pseudo documents, classified average precision, user search goals.**

.

1. INTRODUCTION

Information present on internet is increasing vastly and the tool which is used to access information is known as web search engine. While searching data on internet user queries are submitted to web search Engine to gain information on internet and by using queries user search goals can be analyzed. But sometimes ambiguous queries may have a large amount of unneeded data. As well as two different user may want different type of data when they submit same type of query. For e.g. flora when user submits query some user wants to get information about flora the natural language about flora. But some user wants to get information about Flora institute of technology. Therefore it's very necessary to cover different user search goals in information retrieval. So what this user search goal? User search goal can be defined as the information which user wants on different dimensions of a query and that will be clustered data on different dimensions of query. So the suggested queries can be used to have queries near to the results..

2. RELATED WORK

As inferring and analyzing user search goals is very essential issue to have search engine relevance and to improve search results regarding to query. A lot of work has been investigated. It is as below.

I. Early Information Retrieval

[2] This paper mainly discuss about the early web based models. IST web search engines were known as information retrieval based engines based on different models. Key difference between IST and today's technology are in IST collections of web pages were composed (not documents) and the pages had to be crawled. The collections were much larger, i.e. each query word retrieved too many documents. The approaches which were used are Page ranking and using index terms. Index terms were simple to refer a query and ranking sorted documents according to degree of similarity. But the drawback was that it didn't satisfied user goals and assumed independence of index terms.

II. Agglomerative Clustering of a Search Engine by Query Log

[3] This paper has proposed a new method for data mining. Data will be a collected from User transactions using a search engine. A cluster of same type of queries and URLs similar to particular queries will be discovered. In every log user query to particular search engine and URL selected by the user among the retrieved data from the search engine is collected. The data set is viewed by bipartite graph, having vertices on one side and on the other side to URL's corresponding to the respective queries. In this agglomerative clustering algorithm is used. But some of the drawbacks which occurs by this techniques are-

1. Big problem occur regarding how to combine content ignorant and content aware clustering

2. This system suffers from the unpopular URLs which were not clicked by users.

III. Automatic Hierarchical Segmentation of Search Topics

[4] In this paper Context Aware query suggestion is used. The usability feature can be improved by using query suggestion while data surfing .Now a days there are various techniques those who are making query suggestion but none of them provide context aware query suggestion. While query suggestion immediately preceding query are not taken as context. in this paper while query suggestion 2 steps are used.

1. Off-line step
2. Online query suggestion step.

In offline query suggestion data sparseness is traced as well as query summarization is made by clustering on clicks of

bipartite graph. And concept sequence suffix tree is constructed from click through data.

Next step is Online query suggestion in which query sequence is submitted by user is mapped to a sequence of concepts.

By context in the concept sequence suffix tree queries to the user can be suggested in context aware manner. In this technique not only the current query but also recent queries in the same session are considered. The main thing is similar queries are combined into concepts and concepts are used to provide suggestion.

IV. Clustering User Queries of a Search Engine

[5] In this paper user queries are clustered of a Search Engine and which is used to increase retrieval precision level. Now a day's some of the search engines have the frequently asked queries with the manually verified answers. One of the most important tasks is to identify FAQs. The clustering of the queries is made with their content as well as user log .and the resulting cluster show the information which is very useful and essential to identify FAQ.

Now a day's information on internet increased with a huge manner. Yet people are not getting satisfied with the provided links and existing search engine performance. Because existing search engine respond to a query with thousands of links and thousands of documents, In order to have more accurate answers and more specific answers with the user queries now a day's lots of search engine have FAQs with manually verified answers. for e.g. ask.com, www.ask.com.in this system they try to understand user queries and then suggest some similar queries which is asked by users a lots of time and for which queries a search engine have a accurate and verified answer. where as in a older system documents are retrieved in the basis of keyword matching. In which documents are retrieved by matching keywords. In which lots of information is unnecessary. And many of them are not relevant to question.

V. Automatic Identification of user search goals in web search

[6]In this paper user search goals are considered as Navigational and informational. And queries are categorized into these classes. In this technique first query Intent is tried to learn from clicked Graphs and then the queries are classified according to the specified Intents and all of the other work focus on tagging queries with some previously defined concepts so that feature representation of the queries would be improved.

VI. Bringing Order to the web

[7]In this paper the novel approach which was used is Automatic Categorizing of Searching Results. In which web pages are represented and categorized in various categories. Also a user interface is used by which organized web search results got. Here algorithm which was used was Text classification. And categorizing search results interface was sounds good as compared to documents with the ranking list interface. And with this technique results which has got are 5-10% faster. Also focus on items from the various categories instead of goes through all the results. But this approach covered the full range of web content. But the problem was with all user queries does not matched with the web content. The experiments which have conducted, from that 5-40% results were under 'Not Categorized'.

VII. Generating Query Substitutions

[8]In Query substitution the new query is generated and replaced by the user's original query .in this paper modifications were made this technique uses modification based on substitution of query. The newly generated queries are related closely to the original queries and terms. Query substitution method is generally contrasted with the two methods, query expansion and query relaxation. Query expansion deals with pseudo-relevance feedback. Which is expensive process and do not have any aim. The query relaxation can be used by Boolean/TF-IDF retrieval and effects on query specificity. As compared to other two techniques Query substitution gives more specific results. Experiments which are taken mostly show that this method increases coverage area of web and effectiveness in the searching of web information

VIII. Learning to cluster web search Results

[9] In this paper unsupervised clustering converted to supervise clustering. At first the documents are clustered by extracting the phrases and which are used as candidate cluster. And lastly final clusters are made by merging all these candidate clusters. At last in the clustering algorithm is used to for clustering candidate cluster. While clustering snippets are taken as input for clustering instead of whole document. So there are various advantages of this system-

1. By this technique cluster data can be easily identified by using shorter cluster names.

2. According to score clusters are ranked.

IX. Learning Query Intent from Regularized Click Graphs

[10] In this paper query intent classifier is improved by using click graphs. This method is essential for user interface query classification. Improvement in feature representation is the goal of existing systems. But the proposed system is completely focused on enriching feature representation. The main purpose of this system is to increase training data by using clicked graph. Unlabeled queries can be understood from the labeled queries. As well as learning is regularized with graphs. Advantages of this system are-

1. Effectiveness can be increased by using by expanding training data.

2. By this classification performance is improved.

3. Feature Representation is enriched.

Comparison of all of the above techniques is as provided in table 1 which represents various techniques for inferring user search goal.

3. PROPOSED MODEL

As there are lots of improved results we get from the framework used in [1]. So we are going to consider that

technique and for better results which uses Semantic clustering algorithm. In this approach various user search goals are identified by feedback sessions clustering. Feedback session can be defined as Query log for the entered query which contains clicked URLS as well as UN clicked URLS and which ends with the URL which was clicked for last time. Instead of clustering of user search results it is more convenient to cluster feedback sessions. From feedback sessions pseudo-documents are generated and taken for clustering to represent feedback sessions in a better way. It will result into improvement of efficiency of analysis of user search goals. And lastly a new approach is used known as CAP (Classified Average Precision) for evaluation of inferred user search goals to improve the results. The process of model as shown in fig.1

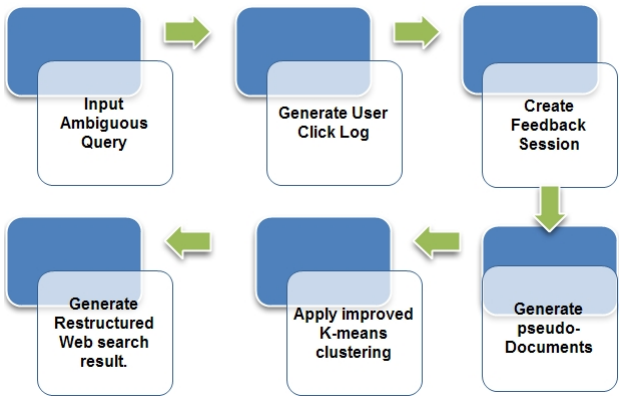


Figure 1.Framework of our Approach

Fig 1 shows framework of proposed approach which is divided into two parts in the first part feedback sessions are generated from user clicked query log.

1. Feedback session can be defined a session for a particular query which contains a record of clicked and UNclicked URLs for a particular query and which ends with a last clicked URL. When we consider a last clicked URL it shows that all the URLs are scanned for that session. Each feedback session can result into user's interest in the URLs.

2. Then from the feedback session pseudo-documents are generated. Pseudo-documents can be generated by performing two steps

1. Representing URL in the feedback session

- Firstly URLs in the feedback session are enriched with extraction of titles and snippets of the returned URLs in the feedback session.
- By this way each URL is represented with small text paragraph which consist title and snippets.
- Then textual processes are used on them like stemming, removing stop words. And finally each URLs title and snippet is represented by term frequency and inverse frequency.

2. Forming pseudo-documents based on URL representation.

In this step from the enriched URLs pseudo documents are formed.

And lastly user search goals are analyzed from clustering of this pseudo-documents.in the referred technique k-means clustering algorithm is used .Results can be improved by implementing Semantic clustering algorithm from the various clustering algorithm.

4. SEMANTIC CLUSTERING FOR CONCEPT DETECTION

Basic Concept:

Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of synonymy. From this statement, we can conclude that since the term-based approach, the clustering might also suffer from achieving high conceptualization.

We also can state that a phrase-based or sequential pattern based clustering can be effective for finding conceptually relevant cluster than the term-based clustering because we get a good conceptualization within a phrase or within an ordered set of terms than a single term.

Proposed Algorithm:

Module Description:

1) **Pre-Processing:**
   Steps: a. Stop word Removal
   b. Non Word Removal
   c. Stemming
   Here in our proposed algorithm we require a pre-processed dataset with maintaining the sentences as it is.
2) **Finding Sequential Patterns:**
   Here we propose a simple and effective technique to find Sequential Pattern within a document.

Let we divide a file in paragraph. Now for each term within each paragraph we can calculate the co-occurrence of a term with its **very next** term.

Assumption:

Let we take a min_support count as 2 because if a pair of words is co-occurred for 2 or more than 2 times, we can say that it is a sequential or an ordered pattern.

For example:

Global warm. Global income.

Global income. Global warm. // taken after applying stemming

(Global, worm) = 2

(Global, income) = 2

Here we can see that 'global warming' and 'global income' both are different concepts and they are sequential in nature.

Now further we find the pairs having a term that is common in them Such as,

Global warm. Global income.

Global income. Warm increase.

Here (global, warm) and (warm, increase) have a common term 'warm'.

Now taking union of these to patterns we get a new pattern (global, warm, increase).

We then find such patterns within the whole document and also their frequencies or weights.

Here in our proposed algorithm, we limit our patterns' length up to 3 to avoid excessive looping. And we can analyze that patterns' of length 3 are much and more sufficient for conceptualization.

Now let we have a dataset which contains these types of term-sets.

$D_0, D_1, D_2 \ldots$ Dn. Where n= Total no. of Documents

Algorithm's Steps:

For i = 0 to n-1

Do

For j = i+1 to n-1

Do

Sim_{ij} := CosineSim(D_i,D_j)

End for

End for

Sim_{max} := max{ Sim_{ij} } // Similarity of D_i and D_j which is maximum than any other pair of document.

cluster := 0

$C_{cluster}$:= {D_i, D_j } // where Sim_{ij} (D_i,D_j) = Sim_{max}

$C_{cluster}$->centroid := Sim_{max}

For each D_r // D_r are the remaining document i.e r != i and r != j

Do

For each $C_{cluster}$

Do

Sim_{mean} := $\sum$ (CosineSim(D_r, $D_{cluster,\ k}$)) / $|C_{cluster}|$ ..// $|C_{cluster}|$ is total documents in Cluster

If Sim_{mean} >= $C_{cluster}$ ->centroid

$C_{cluster}$:= $C_{cluster}$ U D_r

$C_{cluster}$->centroid := recalculate centroid($C_{cluster}$)

Else

cluster= cluster+1

$C_{cluster}$:= {D_r}

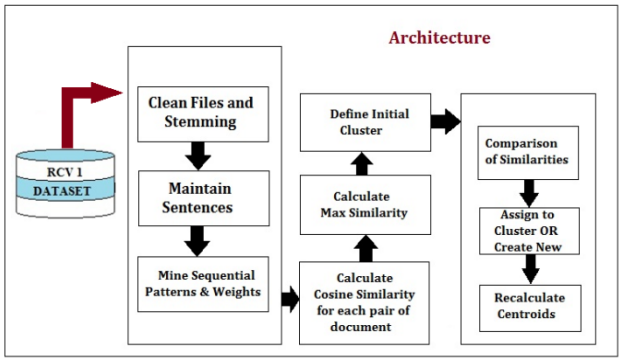
$C_{cluster}$->centroid := calculate centroid($C_{cluster}$)

End for

End for

Here in our proposed algorithm, we limit our patterns' length up to 3 to avoid excessive looping. And we can analyze that patterns' of length 3 are much and more sufficient for conceptualization.

ARCHITECTURE



Advantages of proposed Algorithm:

1) As previous clustering algorithms suffer from polysemy and synonymy, the proposed algorithm is able to cluster the documents which are conceptually similar.
2) Semantics are preserved as we use sequential or ordered term set.
3) Semantics can be further maintained by increasing length of term set.
4) Small and feasible to implement.

Limitations

1) Requires a little lengthy preprocessing.
2) Number of clusters is not user defined. As number of sparse documents increases the number of cluster may increase.

In advance we do not know about the actual value of user search goal different values are tried and the optimal value will be determined by the feedback of CAP which is used to improve results. In the bottom part restructuring of original search result is made on the basis of user search goals inferred from the first part. Then evaluation of performance of restructured web search results is performed by the CAP and lastly the evaluation result will be used as a feedback to the first part which helps to select the optimal number of user search goals.

This technique has a lot of advantages and they are as given below.

1. The search results with the same search goal will be grouped together. So that user can easily find what type of information they want.
2. Web search engine is restructured.
3. Pseudo documents can be used in query suggestion and recommendation.

5. CONCLUSION

This paper mainly compares the various existing techniques in inferring user search goals. The time complexity of proposed approach is going to reduce as compared with other technique. And by using this technique we are going to increase the efficiency of inferring user search goal as well as to fulfill user information need by providing a well-structured web search result.

REFERENCES-

[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 UserSearch Goals with Feedback Sessions, IEEE transactions onknowledge and data engineering, 2013.

[2] R.Baeza-Yates and B.Ribeiro-Neto, Modern InformationRetrieval, ACM Press, 1999.

[3] Chen. H and S. Dumais, "Bringing Order to the Web:

[4] Automatically Categorizing Search Results," Proc. SIGCHIConf. Human Factors in Computing Systems (SIGCHI '00), pp.145-152, 2000.

[5] Joachims, "Optimizing Search Engines UsingClickthrough Data," Proc. Eighth ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

[6] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma,"Learning to Cluster Web Search Results," Proc. 27th Ann. Int'lACM SIGIR Conf. Research and Development in InformationRetrieval (SIGIR '04), pp. 210-217, 2004.

[7] U. Lee, Z. Liu, and J. Cho, "Automatic Identification ofUser Goals in Web Search," Proc. 14th Int'l Conf. World WideWeb (WWW '05), pp. 391-400, 2005.

[8] R. Jones, B. Rey, O. Madani, and W. Greiner, "GeneratingQuery Substitutions," Proc. 15th Int'l Conf. World Wide Web(WWW '06), pp. 387-396, 2006.

[9] Wang and C.-X Zhai, "Learn from Web Search Logs toOrganize Search Results," Proc. 30th Ann. Int'l ACM SIGIRConf. Research and Development in Information Retrieval(SIGIR '07), pp. 87-94, 2007.

[10] Li, Y.-Y Wang, and A. Acero, "Learning Query Intentfrom Regularized Click Graphs," Proc. 31st Ann. Int'l ACMSIGIR Conf. esearch and Development in Information Retrieval(SIGIR '08), pp. 339-346, 2008.

[11] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li,"Context-Aware Query Suggestion by Mining Click-Through,"Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discoveryand Data Mining (SIGKDD '08), pp. 875-883, 2008.