AN ENHANCED TECHNIQUE FOR TEXT RETRIEVAL IN WEB SEARCH

S. Aarif Ahamed[1], V.Venkateshwaradevi[2], B.A.Vishnu Priya[1]

[1]Assistant Professor, Department of Computer Science and Engineering,
M.A.M School of Engineering, Trichy, India.

[2]Assistant Professor, Department of Computer Science and Engineering,
Saranathan Engineering College, Trichy, India.

ABSTRACT

The web is a medium for accessing a great variety of information stored in various locations. As data on the web grows rapidly it may lead to several problems such as increased difficulty of finding relevant information. When a user submits a query to the search engine, it must be able to retrieve information according to the user's intention. But search engine retrieves the list of pages ranked based on their similarity to the query. Sometimes the results are not according to users' interests, because many relevant terms may be absent from queries and words may be ambiguous. Therefore, the results produced by the search engine are not satisfactory to fulfill the user query request. In order to solve this ambiguity, the proposed system aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. First, the given user query is pre-processed to find the root word. Then, a feedback session has been introduced to infer user search goals for a query. Each feedback sessions are mapped to pseudo-documents using keywords which can efficiently reflect user information needs. Agglomerative Hierarchical clustering techniques is used to cluster the pseudo - documents. Improved HITS algorithm is used to rank cluster results relevant for a particular topic. The web results are restructured. Finally, we introduced user editable browser that allow the user to perform editing operations such as deletion and emphasis while browsing the search results.

Index Terms :- user search goal, clicked URL, un clicked URL, VO – pair, personalization, feedback session, pseudo – documents, goal text, keyword, clustering, HITS, user editable browser.

1. INTRODUCTION

With the fast growth of the Web, a user can obtain abundant information easily by submitting a query to a search engine. Many existing search engines use keyword matching as the search mechanism, which usually causes the situation that a large number of non-relevant documents containing query terms are founded out, and the user will make strenuous efforts to browse these non-relevant documents. Thus, it is not simple to find out the real user goal from such short queries. For example, when users submit a query, "Michael Jackson". The search engine provides documents containing the query terms "Michael Jackson". But, the real user goal in their mind may be that "I want to download Michael Jackson's music". The snippet is sorted at about the 25th rank so that users may take lots of time to browse the search-result snippets one by one and finally find this satisfying snippet. The purpose of user goal identification focuses on understanding "what the intension of the user is as he/she submits a query," and further dealing with the problem of query ambiguity to response more suitable search results for fitting user needs. Another example, when the query "the sun" is submitted to a search engine [19], some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun. Therefore, it is necessary and potential to capture different user search goals in information retrieval. Define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user's particular desire to obtain information to satisfy his/her need. Broder and Rose and Levinson defined three coarse classes of user goals, including Navigational, Informational and Transactional/Resource.

1. **Navigational:** When users submit navigational queries, they tend to find certain specific sites in their mind. For example, the situation may be that a user forgets the URL of a specific site or want to find the site of a company. Therefore, this kind of queries is classified into Navigational class.

2. **Informational:** When users submit informational queries, they may tend to find relevant information to the query. For example, when users submit the query, "Michael Jackson", they possibly want to find his albums or lyrics.

3. **Resource:** Users may want to find some software or non-textual resources. For example, the query "FireFox" may be submitted when a user wants to find this web browser software.

The rest of the paper is organized as follows. Section 2 reviews various techniques for effective inferring user search goals. The existing system is presented in Section 3. The proposed system is described in Section 4. Section 5 concludes the paper.

2. RELATED WORKS

2.1 Identifying User Goals from Web Search Results

Yao-Sheng Chang et al [17] propose a novel probabilistic inference model which effectively employs syntactic features to discover a variety of confined user goals by utilizing Web search results. Analyze the user goals in the viewpoint of Natural Language Processing (NLP). Assume that the user goal should be expressed with the form of a hidden sentence in his/her mind. In general, a typical sentence includes a subject (S), a verb (V), and an object (O). Also, assume that the subject of the hidden sentence in user mind is the user himself/herself and the combined pair of the verb and object is called VO - pair. On the basis of VO - pairs, potential user goal can be represented. For example when users submit a query “Michael Jackson”, predict that the hidden sentence in the user mind is “I want to download Michael Jackson’s music,” and the potential user goal is the VO - pairs “download music” (verb + object). The object can be regarded as the noun after the verb. There are four types of VO – pairs “Vc+Na”, “Vc+Nc”, “Vd+Na”, and “Vd+Nc”, where Vc (active transitive verb), Vd (ditransitive verb), Va (active intransitive verb), Ve (activetransitive verb with sentential object), Na(common noun) and Nc (location noun) [17].

2.1.1 Basic probabilistic inference model

Given a query *q*, try to identify the user goal (VO-pair) *g* from the Web search results. First, submit it to a search engine like Google and get a set *S* of snippets from it. Then, predict some probable user goals *g* with respect to each retrieved snippet related to *q* [6]. Through a goal *g* appearing with high probability in each snippet, obtain a user goal with the highest probability value in that snippet. The probability of *g* is calculated as

$$p(g|q) = \sum_{si\epsilon S} P(si|q)\,P(g|si,q) \quad (1)$$

where *si* is one snippet of *S*. Because the query *q* should be covered in the snippet *si*, Equation (1) can be approximated as

$$p(g|q) = \sum_{si\;\epsilon\;S} P(si \mid q)\,P(g \mid si)(2)$$

2.2 An Overview of Personalization in Web Search

The above mentioned technique has some drawbacks as follows: More challenges to adopt VO – pair classes to certain languages and time Consuming to identify VO – pair. In order to overcome these drawbacks Indu Chawla [8] introduces Web search personalization algorithms to improve the Web search experience by using an individual’s data e.g. user’s domain of interest, preferences, query history, browser history etc [8]. Using these factors they extract the results that are the most relevant to that individual. Personalization can be broadly categorized in two types: context oriented and individual oriented [8]. Context oriented personalization include factors like the nature of information available, the information currently being examined, the applications in use, when, and so on. Individual oriented personalization uses user interests, query history, browser history, pages visited etc.

2.3 Context Oriented Personalization

Users do not enter all the terms relevant to what they want to search. So the search query is small and in current information retrieval approaches, documents are retrieved only according to the terms specified in the query. Context can be used to improve the search results as what people are looking for. One way to identify relevant context terms is to simply ask the searcher to provide them as part of their query or to complete a preferences form by selecting from among a variety of basic categories. The information can also be used for query modification and search engine uses these terms to re-rank a limited set of search results and promoting results that relate to these categories [8].

2.3.1 Individual Oriented Personalization

Individual oriented personalization uses user profile to infer user search goals. User’s profiles can be constructed either by using information explicitly given by the user or implicitly gathered by the user such as asking users to fill out registration forms or to specify the Web page categories of their interests. Users have to modify their preferences by themselves if their interests change. User profiles may include demographic information, interests or preferences of either a group of users or a single person. Individual user’s profile is built only from topics of interest to the user. There are short term and long term interests for an individual user. Short-term profiles represent the user’s current interests whereas long-term profiles indicate interests that are not subject to frequent changes over time. The goal of user profiling is to collect information about the subjects in which a user is interested, and the length of time over which they have exhibited this interest, in order to improve the quality of information access and infer user’s intentions. The user profile was represented by a hierarchical category tree and the corresponding keywords associated with each category [8]. The user profile was automatically learned from the user’s search history. Even though the Web search personalization algorithm improve the Web search experience by using an individual’s data, but has some drawbacks as follows: In Context oriented personalization expecting searchers to provide context information explicitly as part of their search is not ideal. Many users are simply unwilling to provide this type of additional information and even asking for it can lead to frustration certainly asking the user for anything close to personal information is liable to alienate many users because of privacy concerns. And in the second option where the user can choose from among the categories provided, the problem is to know about the users’ interest for displaying the categories to the user. Moreover, searchers often do not have enough knowledge available to them to explicitly express such context information even if they were inclined to do so; In Individual oriented personalization a single user profile or model can contain a too large variety of different topics so that new queries can be incorrectly biased and also users are becoming more concerned about threats to privacy in the online environment.

2.3 Query recommendation using query logs in search engines

Baeza et al [1] present an algorithm to recommend related queries to a query submitted to a search engine. The related queries are based in previously issued queries, and can be issued by the user to the search engine to tune or redirect the search process. The method proposed is based on a query clustering process in which groups of semantically similar queries are identified. The clustering process uses the content of historical preferences of users registered in the query log of the search engine. The method not only discovers the related queries, but also ranks them according to a relevance criterion. The clustering process is based on a term-weight vector representation of queries, obtained from the aggregation of the term-weight vectors of the clicked URL's for the query. Semantically similar queries may not share query-terms but they do share terms in the documents selected by users. Thus the framework avoids the problems of comparing and clustering sparse collection of vectors, in which semantically similar queries are difficult to find, a problem that appears in previous work on query clustering. Further, query vectors can be clustered and manipulated similarly to traditional document vectors.

Ranking the queries according to two criteria:

1. The similarity of the queries to the input query (query submitted to the search engine).

2. The support, which measures how much the answers of the query have attracted the attention of users.

The combination of measures (1) and (2) defines the interest of a recommended query.

2.3.1 Discovering Related Queries

A single query (list of terms) may be submitted to the search engine several times, and each submission of the query induces a different query session, which consists of a query, along with the URLs clicked in its answer.

QuerySession : = (query, (clickedURL)*)

A more detailed notion of query session may consider the rank of each clicked URL and the answer page in which the URL appears, among other data that can be considered for improved versions of the algorithm.

The query recommender algorithm operates in the **following steps:**

1. Queries along with the text of their clicked URL's extracted from the Web log are clustered. This is a preprocessing phase of the algorithm that can be conducted at periodical and regular intervals.
2. Given an input query (i.e., a query submitted to the search engine) first find the cluster to which the input query belongs. Then compute a rank score for each query in the cluster.
3. Finally, the related queries are returned ordered according to their rank score.

The rank score of a related query measures its interest and is obtained by combining the following notions:

1. **Similarity of the query**. The similarity of the query to the input query.
2. **Support of the query**. This is a measure of how relevant is the query in the cluster. Measure the support of the query as the fraction of the documents returned by the query that captured the attention of users (clicked documents). It is estimated from the query log as well.

2.3.2 Query Clustering

Query Similarity

To compute the similarity of two queries, first build a term-weight vector for each query. Vocabulary is the set of all different words in the clicked URLs. Stopwords (frequent words) are eliminated from the vocabulary considered. Each term is weighted according to the number of occurrences and the number of clicks of the documents in which the term appears. Given a query q, and a URL u, let Pop(q, u) be the popularity of u (fraction of clicks) in the answers of q. Let Tf(t, u) be the number of occurrences of term t in URL u. We define a vector representation for q, where q[i] is the ith component of the vector associated to the ith term of the vocabulary (all different words), as follows:

$$q[i] = \sum_{URL\ u} \frac{Pop(q,u)*Tf(ti,u)}{\max t\ TF(t,u)} \qquad (3$$

where the sum ranges over all clicked URLs. Query recommendation algorithm considers only clicks URL's but not clicked that are similar to the term weight representation of the input query

3. EXISTING SYSTEM

Zheng Lu et al [19] [16] extended the previously discussed algorithm in an effective manner. This section presents an approach in order to infer user search goals by organizing search results by aspect learned from user click through logs.

Given an input query the general procedure of the approach is:

All the feedback sessions related to the given query will be extracted from user click- through logs. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. The clicked URLs tell what users require and the unclicked URLs reflect what users do not care about. Then, map each feedback sessions to pseudo-documents using keywords which can efficiently reflect user information needs.

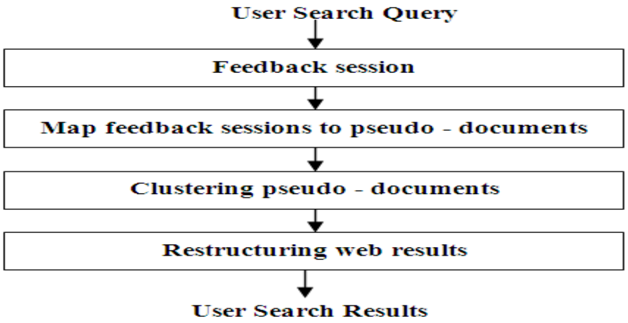
User Search Query
Feedback session
Map feedback sessions to pseudo - documents
Clustering pseudo - documents
Restructuring web results
User Search Results

Figure 3.1 Dataflow diagram for the existing system

By using k- means clustering techniques to cluster the pseudo-documents to infer user search goals. Finally restructure the web search results inferring user search goals. Then classified average precision (CAP) has been introduced to evaluate the performance of the restructured web search results. Figure 3.1 represents the data flow diagram for the existing system.

4. PROPOSED SYSTEM

The above mentioned technique has some drawbacks as follows:

- Generate noisy and redundant search results.
- Cluster labels generated are not informative enough to allow the user to identify the right search result.
- Do not generate the search results if the query has been entered for the first time.

In order to overcome these drawbacks we had extended the previously discussed approach in an effective manner. Given an input query the general procedure of the approach is:

Since the query contains parts of speech and special characters which are not required for analysis as they do not truly reflect the relevance of a search result. If this query is used for analysis, it may give inconsistent and inaccurate results. Therefore the user query will be pre-processed to identify the root words. All the feedback sessions related to the given query will be extracted from user click- through logs. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. The clicked URLs tell what users require and the unclicked URLs reflect what users do not care about. Then, map each feedback sessions to pseudo-documents using keywords which can efficiently reflect user information needs. K-mean clustering is too simple and it performs poorly for large set of data. And also it requires prior knowledge of number of clusters to be generated. Therefore, instead of using k-mean clustering we had used Agglomerative Hierarchical Clustering to cluster the pseudo – documents. Name of the algorithm refers to its way of working, as it creates hierarchical results in an "agglomerative" or "bottom-up" way, i.e. by merging smaller groups into larger ones. Algorithm takes as input a matrix of pairwise similarities between objects. In case of documents this matrix is created by calculating all pairwise similarities between documents using cosine similarity. It returns a binary tree of clusters, called dendrogram.

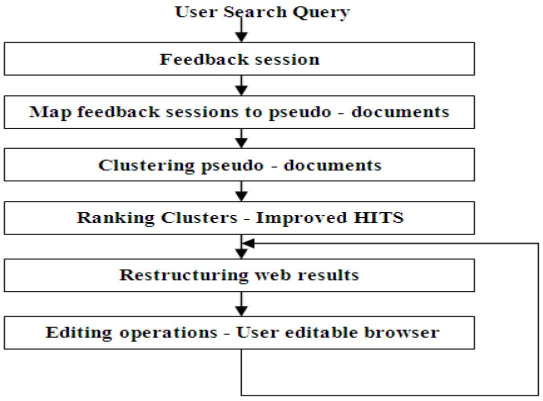
User Search Query
Feedback session
Map feedback sessions to pseudo - documents
Clustering pseudo - documents
Ranking Clusters - Improved HITS
Restructuring web results
Editing operations - User editable browser

Figure 4.1 Dataflow diagram for proposed system

Improved HITS (Hyperlink – Induced Topic Search) algorithm is applied to the clusters generated by Agglomerative Hierarchical Clustering to rank cluster results relevant for a particular topic. Ranking is done by assigning relevance weight to the cluster results. Restructure the web search results based on result generated by Improved HITS algorithm. Finally, we introduce user editable browser to perform editing operations such as deletion and emphasis while browsing the search results. When the user deletes a part of the search result, the system degrades search results which include the deleted term or sentence. When the user emphasizes a part of the search result, the system upgrades the search results which include the emphasized term or sentence. The system reranks the search results according to the user intention and shows the reranked results to the user. In this way, the user can easily obtain the optimized search result for the given query. Figure 4.1 represents the data flow diagram for the proposed system.

5. CONCLUSION

This paper aims to discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. First, the given user query is pre-processed to find the root word. Then, a feedback session has been introduced to infer user search goals for a query. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Each feedback sessions are mapped to pseudo-documents using keywords which can efficiently reflect user information needs. By using Agglomerative Hierarchical clustering techniques to cluster the pseudo - documents. Improved HITS(Hyperlink – Induced Topic Search) algorithm is used to rank cluster results relevant for a particular topic. The web results are restructured. Finally, user editable browser has been introduced allow the user to perform editing operations such as deletion and emphasis while browsing the search results. This paper can be extended to support query recommendation there by suggesting queries that can helps the user to form queries more precisely.

6. ACKNOWLEDGMENTS

We thanks to the light, our god allah, who guided us through the way. We have taken efforts in this paper. However, it would not have been possible without his kind support and help of many individuals and organizations. I wish to avail myself of this opportunity, express a sense of gratitude and love to my friends and my beloved parents for their manual support, strength, help and for everything.

References

[1] Baeza -Yates .R, Hurtado. C , and Mendoza. M, “QUERY RECOMMENDATION USING QUERY LOGS IN SEARCH ENGINES,” Proc. Int’l Conf. Current Trends in Database Technology (EDBT ’04), pp. 588-596, 2004. (Conference Style)

[2] Beeferman.D and Berger. A, “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. (Conference Style)

[3] Beitzel. S, Jensen .E., Chowdhury. A, and Frieder. O, "VARYING APPROACHES TO TOPICAL WEB QUERY CLASSIFICATION," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007. (Conference Style)

[4] Cao.H , Jiang.D, Pei.J, He.Q, Liao.Z, Chen.E, and Li.H, "CONTEXT-AWARE QUERY SUGGESTION BY MINING CLICK-THROUGH," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008. (Conference Style)

[5] Chen .H and Dumais. S, "BRINGING ORDER TO THE WEB: AUTOMATICALLY CATEGORIZING SEARCH RESULTS," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000. (Conference Style)

[6] Haibo Yu , Tsunenori Mine, Makoto Amamiya," TOWARDS USER INTENT BASED SEARCHING" 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11 (Conference Style)

[7] Huang C.-K, Chien L.-F, and Oyang Y.-J, "RELEVANT TERM SUGGESTION IN INTERACTIVE WEB SEARCH BASED ON CONTEXTUAL INFORMATION IN QUERY SESSION LOGS," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003. (Journal Style)

[8] Indu Chawla., "AN OVERVIEW OF PERSONALIZATION IN WEB SEARCH. In IEEE'2010. (Journal Style)

[9] Joachims .T , "EVALUATING RETRIEVAL PERFORMANCE USING CLICKTHROUGH DATA," Text Mining, J. Franke, G. Nakhaeizadeh, and Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003. (Journal Style)

[10] Joachims .T, "OPTIMIZING SEARCH ENGINES USING CLICKTHROUGH DATA," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002. (Conference Style)

[11] Joachims .T, "OPTIMIZING SEARCH ENGINES USING CLICKTHROUGH DATA". (Journal Style)

[12] Lee.U, Liu.Z, and Cho. J, "AUTOMATIC IDENTIFICATION OF USER GOALS IN WEB SEARCH," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005. (Conference Style)

[13] Poblete.B and Ricardo .B.-Y, "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. (Conference Style)

[14] Shital C. Patil, Prof. R.R. Keole, " WEB USAGE MINING AND WEBCONTENT MINING – A COMBINE APPROACH FOR ENHANCING SEARCH RESULT DELIVERY" ISSN:2277 128X volume 3, Issue 10, Oct' 2013. (Journal Style)

[15] Takehiro Yamamoto, Satoshi Nakmura, and Kutsumi Tanaka," AN EDITABLE BROWSER FOR RERANKING WEB SEARCH RESULTS". (Journal Style)

[16] Wang .X and Zhai. C.-X, "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. (Conference Style)

[17] Yao-Sheng Chang, Kuan-Yu He, Scott Yu, Wen-Hsiang Lu., "IDENTIFYING USER GOALS FROM WEB SEARCH RESULTS" (Journal Style)

[18] Zeng .H.-J, He. Q.-C, Chen .Z, Ma. W.-Y, and Ma. J, "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. (Conference Style)

[19] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, and Zhaohui Zhengr, "A NEW ALGORITHM FOR INFERRING USER SEARCH GOALS WITH FEEDBACK SESSIONS," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, March 2013. (Journal Style)