Improving the Discovery of User Search Goals using Click through Data

Parth Patel[1] Jignesh Vania[2]
[1]M.E Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Engineering & Technology, Ahmedabad, Gujarat, India

Abstract— Web search applications represent user information needs by submission of query to search engine. But still the entire query submitted to search engine doesn't satisfy the user information requirements, for the reason that users may need to acquire information on diverse aspects when they submit the same query. From this discovering the numeral of dissimilar user search goals for query and depicting each goal with several keywords automatically become complicated. The suggestion and examination of user search goals can be very valuable in improving search engine importance and user knowledge. Discovering the numeral of dissimilar user search goals for query by k-means clustering with user feedback sessions. Proficiently replicate user information requirements generate a pseudo-document to map the different user feedback sessions. Clustering Pseudo documents with K means clustering result are computationally difficult and semantic similarity between the pseudo terms is also important while clustering. To conquer this problem proposed a FCM clustering algorithm to group the pseudo documents and it also measure the semantic similarity between the pseudo terms in the documents using wordnet. The FCM algorithm divides pseudo documents data for dissimilar size cluster by using fuzzy schemes. FCM selecting cluster size and central point rest on on fuzzy model. The FCM clustering algorithm it assemble quickly to a local optimum or grouping of the pseudo documents in well-organized way. Semantic similarity between the pseudo terms with Wordnet based similarity is used for comparing the similarity and diversity of pseudo terms. Lastly new result measures the clustering results with parameters like classified average precision (CAP), Voted AP (VAP), risk to avoid classifying search results and average precision (AP). It demonstrations FCM based system develop the feedback sessions outcome than the normal pseudo documents.

Key words: FCM clustering algorithm, (CAP), Voted AP (VAP)

I. INTRODUCTION

The World Wide Web has become an important source of information and services and it is very popular and interactive. The web is enormous, varied and active. As the web is growing very quickly, the users get easily lost in the hectic structure. The basic goal of a search engine is to provide useful information to the users as per their needs. Therefore, retrieving the needs of users and finding their needs have become very important. We can study user's search behaviour by search log analysis of the user from the search engine. The data can be billions of queries daily for a popular search engine. By use of client-side plug-ins large amount of browser log data also can be collected. To improve search result mining of this huge amount of search and browser log data is needed. The challenge is to create efficient and effective techniques to clean, process and model the log data. Whenever a user performs a search by firing a query and clicks a URL from the search result the contents of that page are extracted. The combination of clicked url, contents that are extracted and query of the user are stored in the server log. So when the next time user enters the query on the search engine the output is compared with the data in server log and ranking is done accordingly so that users can easily reach the goal what they are looking for. Extraction of useful information from server logs is web usage mining. It can be used to find out what people are looking for on internet. A click through data is the combination of clicked and unclicked urls from a particular search. So by use of Web Usage Mining the interesting patterns can be discovered and the needs of web based applications can be served better. Usage data extracts the behaviour of users browsing on internet. So use of click though data and search query can lead to the interest of users and the goal text of the users by applying different techniques of clustering for different types of data The ranking of pages in result are based on content and keywords. The goal of search engines is to provide relevant information to the users to cater to their needs. Hence, finding the content of the Web and retrieving the users‟ interests and needs has become increasingly important now. In web search applications, queries being submitted to search engines are to represent the information needs of users. But, sometimes queries may not be able to exactly represent users‟ specific information needs because many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same type of query. The system attempts to improve the search results by inferring user search intentions, and removing incorrect or limited information problems.

II. LITERATURE SURVEY

H-J Zeng et.al [3] proposed a query based method to cluster search results. For a given query, the rank list of documents return by a certain Web search engine, it first extracts and ranks most salient phrases as candidate cluster names, based on a regression model learned from pervious training data. Candidate clusters are formed by assigning documents to relevant salient phrases and the final cluster are generated by merging these candidate clusters. 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.

The problem of clustering investigate results has been investigate in a numeral of previous works. All of the previous work apply clustering algorithms which first group documents into similar groups according to content similarity, and produce expressive summary for clusters. Though, these summaries are often illegible which construct it difficult for Web users to recognize relevant clusters.

Zamir and Etzioni [3] introduced a Suffix Tree Clustering (STC) which first identifies sets of documents that split common phrases, and after that create clusters according to these phrases. Our applicant phrase extraction procedure is similar to STC but we supplementary calculate a number of significant properties to identify salient phrases, and make use of learning methods to rank these salient phrases. Some topic finding or text trend analysis mechanism is also related to our method. The dissimilarity is that we are specified titles and short snippets somewhat than whole documents. For the meantime, we train regression model for the ranking of cluster which is closely related to the efficiency of users' browsing. Web search engines challenge to satisfy users' information needs by standing web pages with reverence to queries. But the realism of web search is that it is frequently a procedure of querying, learning, and reformulating. A sequence of interactions among user and search engine can be essential to satisfy a solitary information need [4]. Though users query search engines in order to achieve tasks at a diversity of granularities, issue numerous query as they effort to accomplish tasks. R. Jones and K.L. Klinkner [5] learning real sessions manually labeled into hierarchical tasks, and demonstrate that timeouts, anything their length, are of incomplete utility in identifying task boundaries, achieving a greatest precision. Though, their method only identifies whether a pair of queries belongs to the same goal or mission and does not mind what the goal is in aspect.

U. Lee, Z. Liu, and J. Cho [6] study the "goal" at the back based on a user's Web query, so that this goal can be used to get better the excellence of a investigate engine's results. Preceding studies encompass mainly focused on manual query-log investigation to recognize Web query goals.

Identify the user goal automatically with no any explicit feedback from the user. User search goals represented by a number of keywords can be utilized in query suggestion [7], [8]; thus, the suggested queries can assist user to form their query more accurately. A previous exploitation of user click-through logs is to get user implicit feedback to expand training data when knowledge ranking functions in information retrieval. Adapt a recovery system to challenging groups of users and exacting collections of documents promise further improvement in retrieval quality for at least two reasons. Since physically adapting retrieval function is instance consuming or even not practical, investigate on automatic adaptation by means of machine learning is in receipt of a great deal notice.

III. PROBLEM DESCRIPTION

The way of providing exact information to the users, about what the users rifling in the web is playing a major issue. Today the basic problem is of providing the best relevant results to the users, which should be useful in a user specific way by using the user click through data efficiently. In existing systems for clustering the k means algorithm is used where cluster centres are taken approximate initially so the system is not that much efficient.

The arrangement of search result inside a cluster is not performed. The user click through data and user search logs can be used to improve the search quality of the queries fired by the users, by using feedback sessions, pseudo-documents and an accompanying algorithm. The main approach will be to classify the cluster in a user oriented way such that it improves the relevancy of the result set and the users will be able to find the most relevant queries first in a user specific way.

IV. PROPOSED SOLUTION

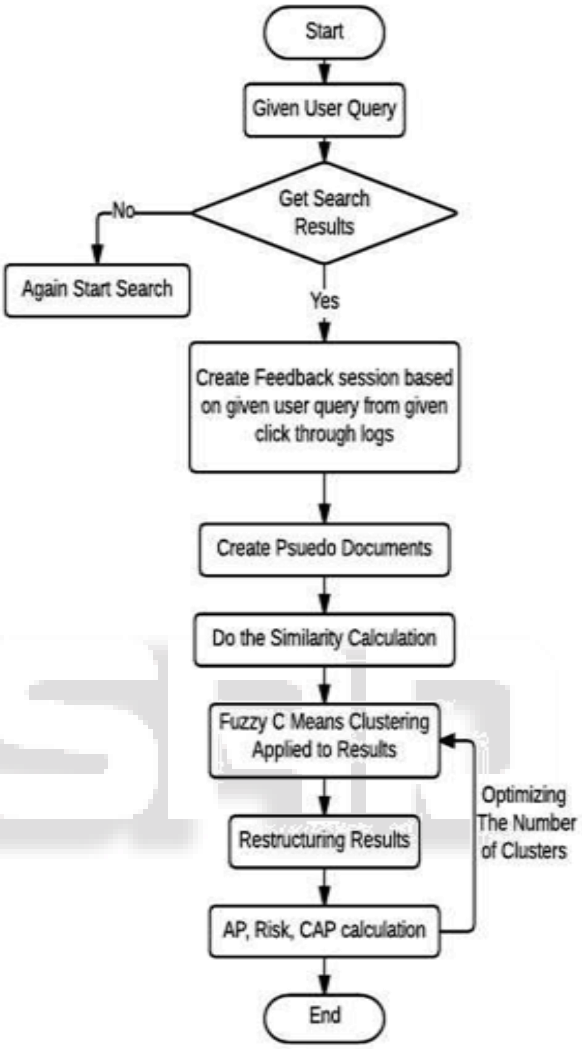

Fig. 1: Flowchart of Proposed System

Here we propose and develop the new framework

1) Creation of feedback sessions with click through mechanism
2) Preparation of pseudo documents
3) WorldNet based Similarity measure
4) Apply FCM algorithm for Clustering of pseudo documents
5) Rearrangement of Cluster Contents
6) Apply CAP approach and finally produce the optimal result content

First from a search engine for any given user query the visited links will be analysed and all the links before the last clicked link will be extracted. This is called the feedback session. In general when a user searches for any query in search engine and gets the search results he goes through analysing all the URLs from the first one and clicks the URLs he is interested in. So the URLs he clicked are relevant to him and the URLs he didn't clicked are not relevant to him. So a feedback session will contain both

clicked and unclicked URLS up to the user has analysed it which is the last clicked URL.So from the feedback session we can get the user interested and not interested URLs. From the given feedback session the titles and snippets of all URLs are extracted and on that text some processes like transforming to lower case, removing stop words, removing redundant data and stemming will be done. From this process the main keywords for each URL in a particular session are obtained. Each URL contains some keywords wich are the search goals of a particular user for the given query. After that each URL's titles represented the pseudo document is created from a feedback session. A pseudo document is a collection of keywords of a URL. So each URL will have a pseudo document created. Similarity between two pseudo-documents is calculated. WordNet is tool to measure the semantic similarity between the terms or words in the pages of documents that was selected by user .When a user given the words taken as input it finds the similarity to terms with the connections among four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The minimum unit in a WordNet is synset, which represent an exact meaning of a word. It includes the word, its clarification, and its synonyms. The specific connotation of one word under one type of POS is called a sense. Each sense of a word is in a dissimilar synset. Synsets are corresponding to senses = structures contain sets of terms with identical meanings. Each synset has a gloss that defines the concept it represents.

All pseudo documents are clustered using fuzzy c means algorithm. Fuzzy c-means (FCM) is a technique of clustering which allows one piece of pseudo documents data to belong to two or more clusters. It is the way to solve how the data with similar pseudo documents are clustered according to best semantics similarity of the pseudo terms in the documents. In this algorithm the same given data or pseudo documents does not go completely to a well definite cluster, based on the fuzzy membership function only the pseudo documents the cluster groups are formed in efficient manner with possible number of the groups at user feedback sessions.

In clusters click sequence will be used for arranging. So by these user goals the URLs are restructured into different clusters for a particular search done by a user. Inside a cluster the URLs will be arranged as per the click sequence and the unclicked URLs will be arranged as the sequence of the URLs in dataset.

Then the CAP will be calculated for our approach. To apply the evaluation criterion CAP first we need to find AP of a single session. For a single query there will be a single session which contains number of URLs that came as an outcome from search engine. The clicked URLs which are relevant to the user. The unclicked URLs which are not relevant to the user. From the AP (Average Precision) we can find the relevancy of the URLs that are arranged in the sequence.

$$AP = \frac{1}{N^+}\sum_{r=1}^{N} rel(r)\frac{R_r}{r} \quad (1)$$

As shown in (1) N^+ is total no of relevant URLs that are clicked URLs from a single session. N is total number of URLs that are upto last clicked URL from a single session. rel(r) is a binary function of given rank wich is calculated for each relevant URL. R_r is the total number of relevant or clicked URLs of total URLs from first to r.

After clustering of the content and restructuring the URLs we can introduce the VAP (Voted AP) which is AP of the different clusters. So the URLs are distributed in different classes we can find the AP of each class to find the measure of evaluation of relevancy from our method.
But VAP is not a satisfactory criterion because we have to calculate the risk of having relevant URLs in different classes. So here we introduce the risk factor which is as shown in (2)

$$Risk = \frac{\sum_{i,j=1(i<j)}^{m} d_{ij}}{C_m^2} \quad (2)$$

In (2) the every pair of clicked URLs are checked. If the both URL of a single pair are in same cluster or class then d_{ij} value will be 0. If they belong to different clusters or classes then d_{ij} value will be 1. Here m is total number of clicked or relevant URLs. C_m^2 represents the total number of URL pairs.
So now we have introduced the CAP (Classified AP) which is extended from VAP after introducing Risk.

$$CAP = VAP * (1 - Risk)^{\gamma} \quad (3)$$

Here in (3) CAP will select the value of VAP from the cluster in which the user is most interested. So from the cluster in which the most clicked URLs or relevant URLs are there the VAP is selected. γ is the parameter from which we can tune the importance of Risk in CAP. SO the CAP is the final outcome from we can measure the performance of our system which is basically relevancy factor.

From (3) we can say that if all the relevant URLs are categorized in a same cluster then the Risk value will be 0 and CAP value will be maximized. Generally higher VAP and smaller Risk will give high CAP. High CAP means better relevancy of the resulted URLs. Generally categorizing the URLs in more clusters can have chances of having relevant URLs in different clusters and Risk value will be high. Categorizing them in fewer clusters can have chances of lower Risk but lower VAP. So CAP depends on both VAP and Risk.

V. EXPERIMENTAL RESULTS

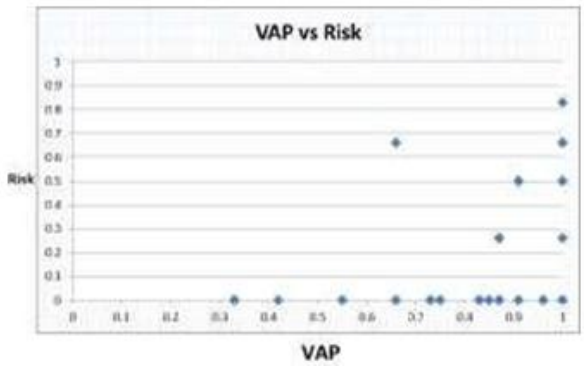

Fig. 2: Graph of VAP vs RISK

We have web search query logs of AOL search engine as dataset. This collection consists of ~2M web queries collected from ~65k users over three months. For analysis we have applied our proposed method to over 800 different queries randomly selected from the given dataset.

Representation of VAP and Risk for 50 most ambiguous queries is shown in figure. Each point represents value of VAP and Risk for a query. If the rearrangement is proper then the value of risk should be smaller and value of VAP should be higher. So from the graph we can see points are tended to right bottom corner. That shows high value of VAP and low value of Risk

Fig.4 shows the comparison chart of AP and CAP for 50 most ambiguous queries. From the chart we can see that AP is the average precision of a query which is relevancy factor of a dataset without clustering and CAP is classified average precision of a query which is relevancy factor of a dataset with clustering.

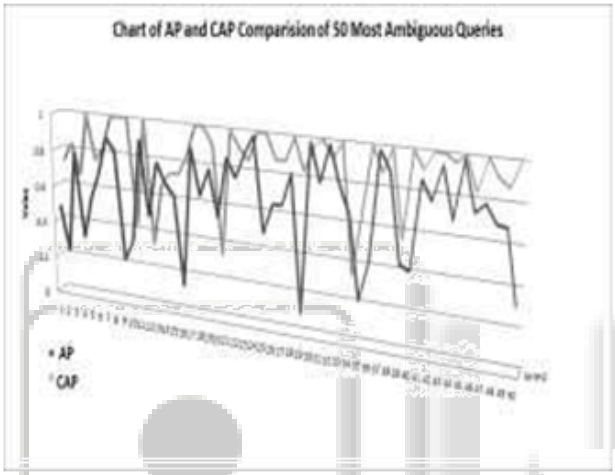

Fig. 3: Comparison of AP and CAP

So by the results we can say that by clustering the pseudo documents and then restructuring the web results give more relevancy of the search result for the given ambiguous queries. From the graph we can see that in more than 60 per cent queries CAP is higher than AP. Fig 4.4 b shows the average AP and CAP from the comparison graph.

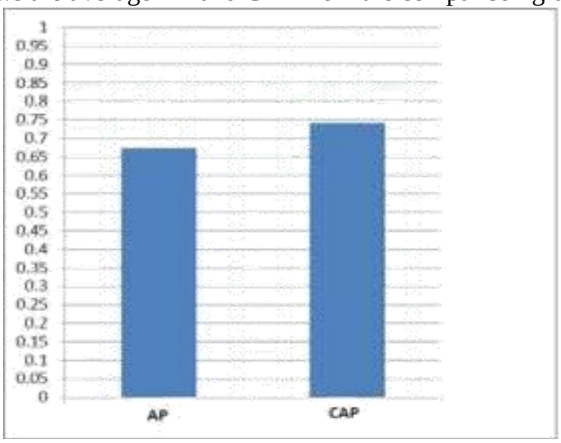

Fig. 4: Graph of AP vs CAP

VI. Conclusion and Future Work

The system can be used to improve discovery of user search goals for a query. The inferred user search goals/intents can be used for query recommendation and in keyword specifications for online advertising. The clusters 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. In future work, different other parameters can be used to optimize and fasten the process to make it more efficient.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.

[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.

[7] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data", Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[8] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, 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.502-513,2013.

