USER SEARCH GOAL INFERENCE AND FEEDBACK SESSION USING FAST GENERALIZED – FUZZY C MEANS ALGORITHM

R.Priyanka [1], N. Rajkumar[2]

[1]*PG Scholar, Department of Software Engineering (M.E.), Sri Ramakrishna Engineering College, Coimbatore, India*
[2]*Head of the department, Department of Software Engineering (M.E.), Sri Ramakrishna Engineering College, Coimbatore, India*

There is a high stress on search engines due to the overload of information content in the internet. Search query submitted by the user to the search engine represents the user requirements. Sometimes, particular desire of the user cannot be fulfilled by the user search query. Also, long listed search result may not be always significant to the user requirements and irrelevant documents are returned by many of the existing search engines which follow the mechanism of keyword matching. Indeed, both the users and search engine developers need to reduce the information content in the internet. In this paper, we aim to infer the user search goal by considering the clicked URLs and reorganize the web search result. We use FG-FCM based clustering for grouping the semantically similar search results which further enhances the reorganized search result.

Keywords— *Ambiguous Query, Broad-topic Query, Feedback session, Semantics.*

----------***----------

1. INTRODUCTION

The dependency on the search engine has grown recently and the users can obtain plenty of information in the internet by submitting the query to the search engine. The requirements of the user are represented by the search query. Finding the right information when searching on search engines can be a pain for sure. Search engines present the search result to the user based on the ranking of website and not according to user interests. Thus, the result of the search engine is same for all the users though different users have different interests. For the broad-topic and ambiguous query, different users will have different search goal. For example, when the query "jaguar" is submitted to a search engine, some users may wish to find the information about the car while some others may intend to find the meaning of animal. Users' particular information needs may not be satisfied by the query given by the user. Therefore, it is required to know the exact information needs of the user. It is necessary to infer the exact user search goal for satisfying the user needs. In this paper, we aim to improve the search engine relevance by identifying the various goals of a user search query and restructuring the web search results. Inference of user search goal can also be used in recommending the list of related queries [8] for the query submitted by the user.

The user search goal has to be inferred for a search query submitted by the user based on clustering the feedback session. Feedback sessions can be represented in various ways like binary vector representation, pseudo-documents, etc. In this paper, we use pseudo-documents which contain keywords to represent the feedback session. The feedback sessions are mapped to the corresponding pseudo-documents. The semantically similar keywords are found for the given query. The search results that are semantically similar are clustered by FG-FCM clustering according to the search goal. Each cluster represents one search goal. The FG-FCM algorithm allows one piece of document data belong to two or more clusters. Also, the algorithm restructures and enhances the original search result by inferring the search goal of the user and reduces the time spent by the user in searching their information needs. By this method, user needs are satisfied. Performance of the restructured search result can be done by an evaluation criterion, Classified Average Precision (CAP).

2. RELATED WORKS

Up to date, many works have been made to investigate on obtaining the user search goals and type of query. We examine some of the previous works to study the problem of clustering. It is important to discover different search goals of the given query to fulfill the needs of the user. Long listed search results can be restructured [2], [7], [9] according to the user requirements. Analysis of user search goals can be divided into three modules: search result reorganization, session boundary detection and query classification. In the first class, authors tried to reorganize the search results of the web. Wang and Zhai [7] analyzed the click-through logs and grouped the search result according to the clicked URLs. In second module, Jones and Klinkner [3] considered session boundaries to identify whether the queries and the goal match. In the third module, people categorized the user goal and queries into some specific classes. Lee et al. [4] categorized the user queries into "Navigational" and "Informational", and inferred the search goals automatically. The search goal can be used to improve the quality of a search engine's results. They also

discussed how to automate the goal identification process. Goal-identification task was based on two types of features: user-click behavior and anchor-link distribution Li et al. [5] defined the objective of the query as "Product intent" and "Job intent" and categorized the search queries accordingly.

Today's Web search engines provide very user friendly interface. Users can submit the queries in the form of keywords similar information retrieval system. Keywords may be a simple keyword or it may be a broad-topic and anything else. Search engine lists the related queries when a query is submitted. Ricardo Baeza-Yates et al. [6] discussed that the associated queries are based on previously issued queries and are provided to the user for redirecting the search process. Semantically similar queries were also identified by the clustering process which clusters the contents stored in query log of search engine. There are many advantages of restructuring the search result according to the user search goal.

Joachims used implicit feedback to enhance the quality of search engines. He referred to click-through logs to optimize the search engine. Zheng Lu et al restructured the search result by clustering the pseudo-documents using K- means clustering [10]. But the K-means algorithm is computationally difficult to find the value of K and it does not work well with the clusters (in the original data) of different size and different density. Some of the prior works considered click-through logs as user's feedback and restructured the search results accordingly. Other works did not consider the user feedback and considered the entire search results returned by the search engine though the link was not clicked by the user. These type of works produced noisy results.

In this paper, we infer the user search goal by considering the feedback session. By this, we restructure and enhance the search result in order to satisfy the user needs. We use FG-FCM algorithm, a fast and robust algorithm for clustering the semantically similar links in the feedback sessions which provide better result than the previous works

3. RESEARCH METHODOLOHY

The approach followed in this paper for inferring the user search goal is shown in the Fig 1.

3.1 User Search Query Analysis

The user search query submitted by the user has to be analyzed. The click-through logs are referred for examining the user search queries and defining the feedback sessions. The queries submitted to the search engines by the user may be a simple query or ambiguous query. It is necessary to analyze the different meanings of the ambiguous query and restructure the search result into different clusters in order to get the user needs satisfied. The search results obtained for the query submitted by the user must be collected for restructuring the search result.

3.2 Feedback Session

The first process in reorganizing the search result is the feedback session representation. Feedback session consist the list of URLs up to the URL that was clicked by the user at last in a single session. All the unclicked URLs before the last clicked URL in a single session is also included because those URLs also has been browsed and analyzed by the user. Therefore, these unclicked URLs must also be included for the feedback. From this feedback session, the clicked URLs represent what information the user entail and the unclicked URLs reflect what information the user do not require. The URLs that are present after the last clicked URL cannot be taken as a part of feedback because it is not certain whether the user have scanned those URLs or not.

Feedback session cannot be used directly for user search goal inference because it varies from that of the user click-through logs. So, it should be represented in some other forms in order to infer the user search goals efficiently. It can be represented in various forms. Binary vector representation is one of the popular ways of representing the feedback session. It consists of 0's and 1's where "0" represents the unclicked URL and "1" represents the clicked URL in a single session. This method cannot be used when more feedback sessions are considered because diverse feedback sessions may have unusual aspects.

The vague keywords can be used to represent the user interests for a query. But these keywords cannot be used for representing the feedback session because they are usually hidden and not expressed clearly. Therefore, pseudo-documents can be used to infer the goals of the user. The feedback sessions are mapped to the pseudo-documents. These documents can be formed by enriching those URLs present in the feedback session. Enriching the URLs can be done by adding the title and a short snippet in a small text paragraph for the same URLs.

3.3 Semantic Similarity and Fast Generalized-Fuzzy C Means Clustering

Semantics of the query submitted by the user must be analyzed and restructured accordingly. Semantically similar words can be identified by the wordnet tool. From the wordnet tool, the semantically similar words for the user search query are extracted. Then the FG-FCM clustering process begins. This algorithm is a variation of FCM algorithm which differs by adding the mathematical exponentiation to the result obtained using FCM. The number of clusters need not be specific.

The advantage of using this algorithm is that the same data element can be in more than one cluster and also the clustering process is more efficient than the existing algorithm. The titles in the feedback session are grouped based on the similarity between the semantic keywords and the titles. Also, the similarity matrix $U_{i,j}$ is used for the clustering process.

The matrix consists of rows and columns where both the rows and columns represent the same titles of the search result arranged in same order. Entry in the matrix represents the similarity between the titles. Various similarity measures can be used to calculate the similarity. In this paper, cosine similarity is used. Cosine score for the titles can be computed as,

$$\text{Similarity}_{Ti,Tj} = \text{Cos}(Ti,Tj) \quad (1)$$

Where, Ti and Tj represent titles of the search results.

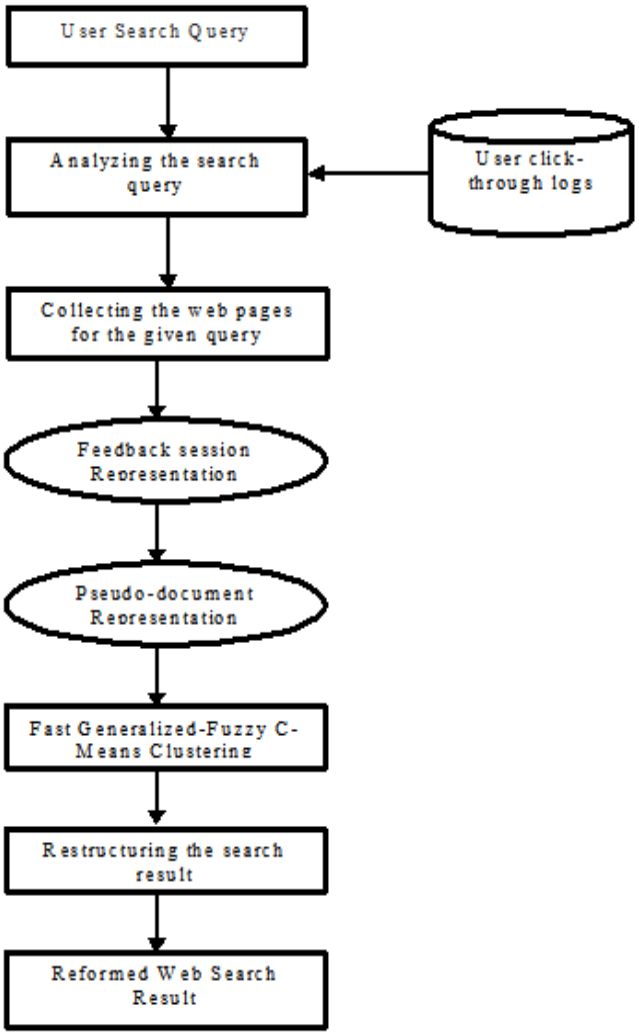

Fig.1. Search result restructuring process

The matrix can be computed by

$$U_{ij} = \frac{1}{\sum\left[\frac{|x_i - C_j|}{|x_i - C_j|}\right]^2} \quad (2)$$

Where, X_i represents the keyword count which is the total number of the semantic keywords in each URL of the feedback session. C_j and C_k are the value of clusters which are obtained from the computation,

$$C_j = \frac{\sum U_{ij} \cdot x_i}{\sum U_{ij}} \quad (3)$$

Where, $U_{i,j}$ represents the value of the similarity matrix at the position i,j. $\sum$ implies that the process should be repeated for the each and every title in the feedback session.

The search results are grouped based on the clustered value. Thus, the search results are restructured and reorganized into groups based on semantic similarity. By assembling the semantically similar URLs into different clusters and restructure the search result, the users' browsing experience can be improved efficiently. This will also satisfy the requirements of the user and reduces the time spent in browsing the contents. This process will be very much useful for the ambiguous queries submitted by the user where there will be more than one meaning.

4. PERFORMANCE EVALUATION

The performance of the approach we have followed can be evaluated from the Classified Average Precision (CAP) criterion. It can be computed using Voted Average Precision (VAP) and risk as,

$$CAP = VAP \times (1 - Risk)^{\gamma} \quad (4)$$

Where, γ is a parameter to adjust the value of risk. Risk can be calculated as

$$Risk = \frac{\sum_{i,j=1(i<j)}^{m} d_{ij}}{C_m^2} \quad (5)$$

Risk factor is added to avoid error. The value $d_{i,j}$ can be either 1 or 0. The value of $d_{i,j}$ is 0 if the pair of i^{th} clicked URL and j^{th} clicked URL are classified into same class else the value of $d_{i,j}$ is 1. C_m^2 is the count of the URL pairs that are clicked in a single session. The Voted Average Precision (VAP) can be calculated from Average Precision (AP). It can be calculated as

$$AP = \frac{1}{N^{\pm}} \sum_{r=1}^{N} rel(r) \frac{R_r}{r} \qquad (6)$$

Where, N^{+} is the total documents that are clicked in the retrieved search result, r is the rank of the URL, N is the total count of the retrieved search result, rel () is the binary function which is computed for the value of r. R_r is the total count of the retrieved search results with the rank r or less than that.

In Voted Average Precision (VAP), votes are referred as clicks. It can be calculated after reorganizing the search results into different clusters and use the AP equation to calculate the "Voted Average Precision". The cluster with more number of URLs is taken into account for calculating VAP. Suppose, number of clicked URLs in both the clusters is same, then the higher value of AP among the clusters is chosen as VAP.

The comparison between K-Means and FG-FCM algorithm is shown in the figure.

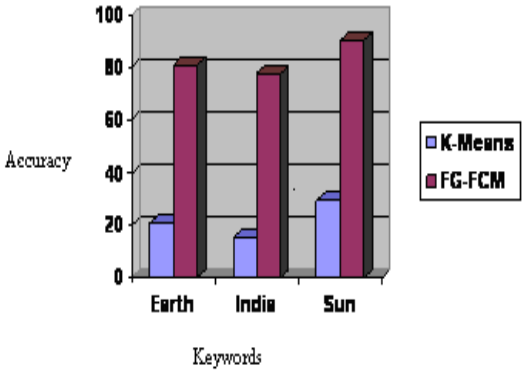

Fig 1 FG-FCM vs. K-Means

5. CONCLUSIONS AND FUTURE WORK

The method used in this paper can be used to infer the user search goal based on the feedback session. We analyze and reorganize only the search results that are obtained in the feedback session for efficient browsing. Therefore, there will not be any noisy data for restructuring the search result. Also, we consider semantically similar keywords to enhance the restructured search result. FG-FCM algorithm is used for clustering the URLs into different groups according to semantic similarity. Thus, the enhanced search result will improve the search engine relevance and satisfy the user to a greater extent. In future work, we plan to continue by investigating the feedback session not only for the single query in the form of keywords, but also for the queries submitted in the form of a sentence

ACKNOWLEDGEMENTS

I would like to thank Dr.N.Rajkumar for giving his innovative ideas, valuable comments and suggestions which led to improvise the presentation quality of the paper and for the successful completion of the work.

REFERENCES

[1] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACMSIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 407-416, 2000.

[2] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 145-152, 2000.

[3] R. Jones and K.L., Klinkner "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management, pp. 699-708, 2008.

[4] U. Lee, Z. Liu and J. Cho, "Automatic Identification of User Goal sin Web Search," Proc. 14th Int'l Conf. World Wide Web, pp. 391-400, 2005.

[5] X. Li, Y.Y. Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 339-346, 2008.

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

[7] 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 pp. 87-94, 2007.

[8] B.R. Yates, C Hurtado, and M Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology, pp. 588-596, 2004.

[9] H.J. Zeng, 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 pp. 210-217, 2004.

[10] L. Zheng, Z. Hongyuan, Y. Xiaokang, L. Weiyao and Z. Zhaohui, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, 2013.

BIOGRAPHIES

Dr.N. Rajkumar obtained his Bachelor's degree in Computer Science and Engineering from Madurai Kamaraj University in 1991 and His Masters in Engg. the same stream in the 1995 from Jadavpur university, Kolkata. He has completed his Masters in Business Administration from IGNOU in the year 2003. His doctorate is in the field of Data Mining, which he completed in 2005 from PSG College of Technology, Coimbatore. He is currently the Head of the department of Computer Science and Engineering at Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu. He has served in the field of education for over 20 years at various Technical Institutions. He has been instrumental in the conduct of 30 short-term courses and has also attended 20 courses conducted by other institution and organizations. He has authored for 2 books for the benefit of the student community in Networking and Computer Servicing. He has published as many as 50 papers in International Journals, Conferences and at the National level in his area of expertise namely Data Mining, Networking and Parallel computing respectively. He has guided 100- Project Scholar's to-date. His E-mail id is nrk29@rediffmail.com

R.Priyanka received her B.E CSE from Hindusthan college of Engineering and Technology, Coimbatore affiliated to Anna University, Chennai in 2012. At present, She is pursuing her M.E Software Engineering in Sri Ramakrishna Engineering College affiliated to Anna University Chennai. Her E-mail id is iamlavs38@gmail.com