IJCAT International Journal of Computing and Technology, Volume 1, Issue 3, April 2014
ISSN : 2348 - 6090
www.IJCAT.org

# Improved Search Goals with Feedback Session by Using Precision Values

[1]P.Srinivasam, [2]S.Raja

[1] Assistant professor, Department of Computer Science, Muthayammal Engineering College
Namakkal, Tamil Nadu, India

[2] M.E scholar, Department of Computer Science, Muthayammal Engineering College
Namakkal, Tamil Nadu, 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 needs, because users may want to get 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. The numeral of dissimilar user search goals for query by k-means clustering is discovered by user feedback sessions. Pseudo document with k-means clustering is generated by user feedback sessions. Clustering Pseudo documents with k-means clustering results are computationally difficult and semantic similarity between the pseudo terms is also important while clustering. To conquer this problem proposed a FCM (fuzzy c means) clustering algorithm to group the pseudo documents and it also measure the semantic similarity between the pseudo terms in the documents. The FCM algorithm divides pseudo documents data for dissimilar size cluster by using fuzzy systems. FCM choosing cluster size and central point depend on fuzzy model. The FCM clustering algorithm it congregate quickly to a local optimum or grouping of the pseudo documents in well-organized way. Semantic similarity between the pseudo terms with keywords based similarity is used for comparing the similarity and diversity of pseudo terms. Finally experimental 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 shows FCM based system improve the feedback sessions outcome than the normal pseudo documents.

***Keywords* - User search goals, feedback sessions, pseudo documents, classified precision.**

## 1. Introduction

World Wide Web (WWW) is very popular and interactive. It has become an important source of information and services. The web is huge, diverse and dynamic. Extraction of interesting information from Web data has become more popular and as a result of that web mining has attracted lot of attention in recent time. Web mining is the process of discovering knowledge, such as patterns and relations, from Web data. Web mining generally has been divided into three main areas: content mining, structure mining and usage mining. Each one of these areas are associated mostly, but not exclusively, to these three predominant types of data found in the Web.

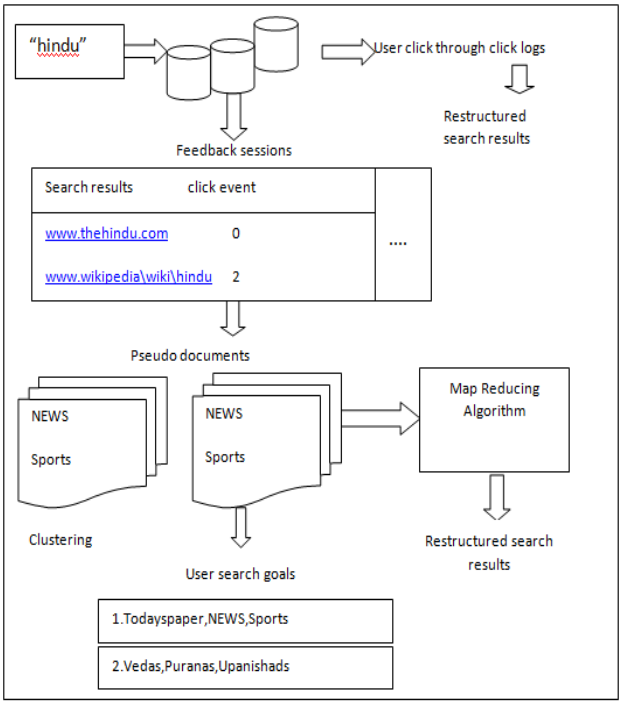


Fig: 1. Architecture Diagram

**Content:** The real data that the document was designed to give to its users. In general this data consists mainly of text and multimedia.

**Structure:** This data describes the organization of the content within the Web. This includes the organization inside a Web page, internal and external links and the website hierarchy.

**Usage:** This data describes the use of a website or search engine, reflected in the Web server's access logs, as well as in logs for specific applications.

1.1 FCM ALGORITHM

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

1. Update $U^{(k)}$ , $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

2. If $\| U^{(k+1)} - U^{(k)} \| <$ then STOP; otherwise return to Step 2.

## 2. Fuzzy C Means Clustering

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic. In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely too just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available. Any point x has a set of coefficients giving the degree of being in the $k^{th}$ cluster $w_k(x)$. With fuzzy *c*-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster. The algorithm minimizes intra-cluster variance as well, but has the same problems as *k*-means; the minimum is a local minimum, and the results depend on the initial choice of weights. Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Fuzzy C-Means is Soft K-means. Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise.

## 3. User Query

Different users may want to get information on different aspects when they submit the same query. For example, when the query "the sun" is submitted to a search engine, 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. We 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. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience.

## 4. Feedback Sessions

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. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered. The feedback session contains URL details with view details.

## 5. Pseudo Documents

The Pseudo Documents are constructed based feedback sessions. The Pseudo documents contain all keywords. The keywords related to ambiguous query. We cluster pseudo-documents by Fuzzy C means clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we set sessions to be different values and perform clustering based on different values, respectively. The optimal value will be determined through the evaluation criterion. The

pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on pseudo-documents, user search goals can then be discovered and depicted with some keywords.

## 6. CAP Calculation

In this module implement the novel evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. CAP is extended version of AP and VAP. In AP and VAP, we can't analyze the risks. If all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0; however, VAP could be very low. Generally, categorizing search results into fewer clusters will induce smaller Risk and bigger VAP, and more clusters will result in bigger Risk and smaller VAP. The proposed CAP depends on both of Risk and VAP.

## 7. Reconstructed Results

Search engines always return millions of search results, it is necessary to organize them to make it easier for users to find out what they want. Restructuring web search results is an application of inferring user search goals. We will introduce how to restructure web search results by Inferred user search goals at first. Then, the evaluation based on restructuring web search results. The original search results are restructured based on the user search goals inferred from the user search. Then, we evaluate the performance of restructuring search results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals.

## 8. Conclusion

In this project, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs.Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods.

**Acknowledgments**

I extend my heartfelt thanks to my guide **Prof. P.SRINIVASAN, M.E.,(Ph.D).,**Assistant Professor Department of Computer Science and Engineering for her exemplary guidance, constant encouragement and kind co-operation throughout the project.

## References

[1] Zheng Lu, HongyuanZha, Xiaokang Yang, Weiyao Lin, and ZhaohuiZheng," A New Algorithm for Inferring User Search Goals with Feedback Sessions" IEEE Transactions On Knowledge And Data Engineering, VOL. 25, NO. 3, MARCH 2013.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[3] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[4] 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 Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[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] 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 (CIKM '08), pp. 699-708, 2008.

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

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

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

[11] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[12] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for Web Query Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR '06), pp. 131-138, 2006.

[13] 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.

[14] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.

[15] 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.