OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A Novel Approach for User Search Results Using Feedback Sessions

Miss. T.Yogeshwari[1], Mr. S. Balamurugan[2]
[1, 2](PG Student, Assistant Professor, Sri Manakula Vinayagar Engineering College, Pondicherry-605106)

Abstract: *In present scenario user search results using Fuzzy c-means algorithm focuses queries are submitted to search engines to represent the information needs of users. The proposed feedbacks sessions are clustered by data are bound to each cluster by means of a membership function. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Pseudo-documents are generated to better understand the clustered feedbacks. Fuzzy C-means clustering algorithm is used to cluster the feedbacks. Clustering the feedbacks can effectively reflect the user needs. Fuzzy c-means algorithm uses the reciprocal of distances to decide the cluster centers. Ranking model is used to provide ranks to the URL based on the user search feedbacks. Evaluate the performance using "Classified Average Precision (CAP)" for user search results.*

Keywords: *Fuzzy c means algorithm, member function, feedback sessions, pseudo documents, classified average precision.*

I. Introduction

It is a novel approach for user search result with their feedback session. First, we have to cluster the feedback session by using Fuzzy c-means algorithm. Second, a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Third, evaluate the CAP of restructured web search results.Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining is the process of choosing, discovering, and exhibiting huge volumes of data to determine unknown patterns or associations useful to the data analyst. The objectives of data mining can be classified into two tasks: description and prediction. While the purpose of description is to mine understandable forms and relations from data, the goal of prediction is to forecast one or more variables of interest.

Clustering is the most important concept used here. Clustering analyzes data objects without consulting a known class label. The objects are grouped or clustered based on the principle of maximizing the intra class similarity and minimizing the inter class similarity. Apriori algorithm is a methodology of association rule of data mining, is used to findout the frequently used URL.

II. Feedback Session

The proposed feedback session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. The clicked URLs tell what users require and the unclicked URLs reflect what users do not care about. It is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

First, we are extracting the titles and snippets of the returned URLs appearing in the feedback session. Each URL in a feedback session is represented by a small text paragraph that consists of its title and snippet.Each URL's title and snippet are represented by a Term Frequency-Inverse Document Frequency (TF-IDF) vector, respectively, as in

$$\mathbf{T_{ui}=[t_{w1};t_{w2};...;t_{wn}]^T;}$$
$$\mathbf{S_{ui}=[s_{w1};s_{w2};...;s_{wn}]^T;}$$

whereT_{ui}andS_{ui} are the TF-IDF vectors of the URL's title and snippet, respectively. ui means theith URL in the feedback session. And w_j(j=1;2;...;n) is the j^{th} term appearing in the enriched URLs.t_{wj} and s_{wj} represent the TF-IDF value of the j^{th} term in the URL's title and snippet, respectively.

The distributions of different user search goals can be obtained conveniently after feedback sessions are clustered. A novel optimization method isused to combine the enriched URLs in a feedback session to form a pseudo-document, which can effectively reflect the information need of a user. We infer the user goals by clustering, feedback sessions are proposed. Clustering the feedbacks can effectively reflect the user needs.

III. Forming Pseudo Document

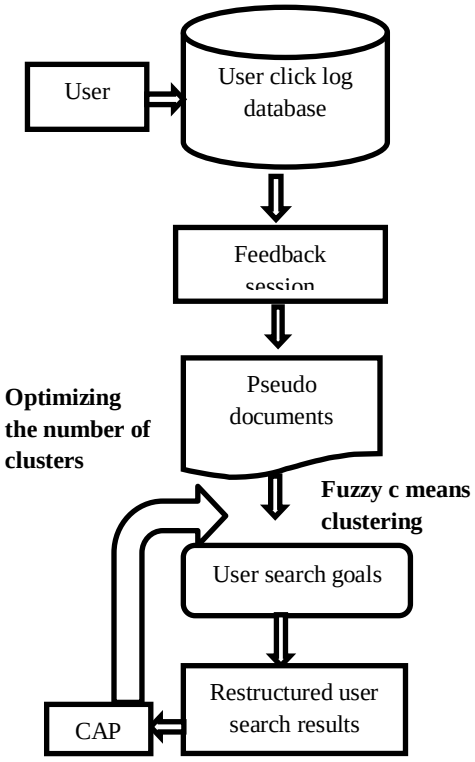
We propose an optimization method to combineboth clicked and unclicked URLs in the feedback session. Let F_{fs} be the feature representation of a feedback session and f_{fs} (w) is the value for the term w. Let F_{ucm}(m=1;2,...;M) and F_{ucl}(l=1;2, ...;L) be the feature representations of the clicked and unclicked URLs in this feedbacksession, respectively. Let f_{ucm} (w) andf_{ucl}(w)be the values for the term win the vectors. We want to obtain such a F_{fs} that the sum of the distances between F_{fs} and each F_{ucm} is minimized and the sum of the distances betweenF_{fs} and each F_{ucl} is maximized.

$$F_{fs}= \left[f_{fs}(w1); f_{fs}(w2); ... f_{fs}(wn)^T; \right]$$

$$F_{fs}(w)=\arg\min_{ffs(w)} \sum \left[f_{fs}(w) - \left[f_{ucm}(w)^{\,2} - \lambda^2 \right] \right]$$

$$\sum f_{fs}\left[(w) - f_{ucl}(w)^{\,2} \right] ; f_{fs}(w) \in Ic$$

Let I_c be the interval $\left[\mu f_{uc}(w) - \sigma f_{uc}(w),\ \mu f_{uc}(w) + \sigma f_{uc}(w) \right]$ and I_{ucl} be the interval $\left[\mu f_{ucl}(w) - \sigma f_{ucl}(w),\ \mu f_{ucl}(w) + \sigma f_{ucl}(w) \right]$,where $\mu f_{uc}(w)$ and $\sigma f_{uc}(w)$ represent the mean and mean square error of f_{uc} (w)respectively, and $\mu f_{ucl}(w)$and $\sigma f_{ucl}(w)$ represent the mean and mean square error of $f_{ucl}(w)$,respectively.Even if people skip some unclicked URLs because of duplication. Each dimension of F_{fs} indicates the importance of a term in this feedback session. F_{fs} is the pseudo-document that we want to introduce. It reflects what users desire and what they do not care about. It can be used to approximate the goal texts in user mind.



IV. Fuzzy C Means Algorithm

In Fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership

The Fuzzy C-Means algorithm (FCM) is used in the areas like computational geometry, data compression and vector quantization, pattern recognition and pattern classification. Fuzzy C-Mean (FCM) is an unsupervised clustering algorithm that has been applied to wide range of problems involving feature analysis, clustering and classifier design.

The main features of that algorithm were the (i) use of a fuzzy local similarity measure, (ii) shielding of the algorithm from noise-related hypersensitivities.FCM clustering techniques are based on fuzzy behavior and they provide a technique which is natural for producing a clustering where membership weights have a natural interpretation but not probabilistic at all.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.

FCM clustering which constitute theoldest component of software computing, are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster.

FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, andtarget recognition.More the data is near to the cluster center more is its membership towards the particular cluster center. The basic idea of fuzzy c-means is to find a fuzzy pseudo-partition to minimize the cost function.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.Fuzzy c-means algorithm uses the reciprocal of distances to decide the cluster centers.

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula. The FCM algorithm converges to a local minimum of the c-means functional. Hence, different initializations may lead to different results.The minimization of the c-means functional represents a nonlinear optimization problem that can be solved by using a variety of methods, including iterative minimization, simulated annealing or genetic algorithms.

The Algorithm Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Time complexity of FCM is **O (ndc^{2}i).**

V. Clustering Pseudo-Documents Using Fuzzy C Means Algorithm

Each feedback session is represented bypseudo-document and the feature representation of the pseudo-document is F_{fs}. We cluster pseudo-documents by Fuzzy c-means clustering which is simple and effective Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{d_{ij}}{d_{ik}}\right]^{\frac{2}{m-1}}}$$

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

$\forall j=1, 2 \ldots c$. This iteration will stop when

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| < \varepsilon \right\}$$

whereis a termination criterion between 0 and 1, whereas *k* is the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

FCM clustering is an iterative process. The process stops when the maximum number of iterations is reached, or when
the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.

5.1 STEPS

1) Randomly select *'c'* cluster centers.
2) calculate the fuzzy membership $'\mu_{ij}'$ using:

$$u_{ij} = \frac{1}{\sum_{k-1}^{c} \left[\frac{d_{ij}}{d_{ik}} \right]^{\frac{2}{m-1}}}$$

3) compute the fuzzy centers $'v_j'$ using:

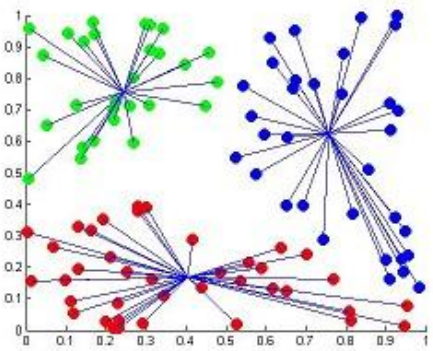
$$v_j = \frac{\sum_{i-1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i-1}^{N} u_{ij}^{m}}$$

$\forall j=1,2 \ldots c$.

4) Repeat step 2) and 3) until the minimum 'J' value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.

Where,

k' is the iteration step.
β' is the termination criterion between [0, 1].
'$U = (\mu_{ij})_{n*c}$' is the fuzzy membership matrix.
'J' is the objective function.



FCM is also called as Fuzzy ISODATA. FCM employs fuzzy partitioning such that a data point can belong to all groups which different membership grades between 0 and 1.

5.2 Parameters of the FCM algorithm

Before using the FCM algorithm, the following parameters must be specified:

- the number of clusters, *c*,
- the fuzziness exponent, *m*,
- The termination tolerance, *ε*.
- norm-inducing matrix, A

Norm inducing matrixes are 3 types. They are

- Euclidean norm
- diagonal norm
- Mahalanobis norm

After clustering all the pseudo-documents, each cluster can be considered as one user search goal.

VI. Evaluating Cap (Classified Average Precision)

CAP (classified Average Precision) is used toevaluate the performance of user search goal inference based on restructuring web search results. A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks. AP is the average of precisions computed at the point of each relevant document in the ranked sequence,

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

Where N^{+} is the number of relevant (or clicked) documents in the retrieved ones, r is the rank, N is the total number of retrieved documents, rel() is a binary function on the relevance of a given rank, and R_r is the number of relevant retrieved documents of rank r or less. “Voted AP (VAP)”

which is the AP of the class including more clicks namely votes. There should be a riskto avoid classifying search results into too many classes by error. We propose the risk as follows

$$Risk = \frac{\sum^{m}_{i};j=1(i<j)\, d_{ij}}{C^{2}_{m}}$$

It calculates the normalized number of clicked URL pairs that are not in the same class, where m is the number of the clicked URLs. If the pair of the i^{th} clicked URL and the j^{th} clicked URL are not categorized into one class, d_{ij} will be 1; otherwise, it will be 0. $C^{2}_{m}=m(m-1)/2$ is the total number of the clicked URL pairs.

We can further extend VAP by introducing the above Risk and propose a new criterion “Classified AP,” as shown below

CAP = VAP × (1-Risk)$^{\gamma}$is used to adjust the influence of Risk on CAP, which can be learned from training data.

VII. Conclusion

In this paper, a novel approach has been proposed to user search results for a query by clustering its feedback sessions represented by pseudo-documents. Clustering feedback sessions are more efficient than clustering search results or clicked URL’s directly. A new criterion called classified average Precision is used to evaluate the performance of restructured web search results. In this paper , we used Fuzzy c means clustering which constitute the oldest component of software computing, are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. The execution time of FCM clustering algorithm for arbitrary data points depends only on the number of clusters and not on the data points. The distance between data points and some shape of the distribution, has the effect on the performance and behavior of the algorithm. Gives best result for overlapped data set and comparatively better then k-means algorithm.

REFRENCES

[1] Gayathri A, Nandhakumar C, Gokulavani M, Santhamani V,” Inferring User Goals Using Customer Feedback and Analyzing Customer Behavior”, International Journal of Computer Applications Technology and Research Volume 3– Issue 2, 125 - 129, 2014.

[2] T. Velmurugan, T.Santhanam,” Implementation of Fuzzy C-Means Clustering Algorithm for Arbitrary Data Points”, International Conference OnSystemics, Cybernetics And Informatics

[3] ZhengLu,HongyuanZha, XiaokangYang,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.

[4] SoumiGhosh , Sanjay Kumar Dubey , "Comparative Analysis of K-Means and Fuzzy C Means Algorithms", ((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[5] "A New Algorithm for Clustering Search Results" GIANSALVATOREMECCA, SALVATORERAUNICH, ALESSANDROP APPALARDO Dipartimento di Matematica e InformaticaUniversitàdella Basilicata Potenza – Italy.

[6] 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..

[7] 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.

[8] D.Kavitha, K.M.Subramanian, Dr.K.Venkatachalam,"SURVEY ON INFERRING USER SEARCH GOAL USING FEEDBACK SESSION", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 2, Issue 12, December 2013.

[9] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building bridges for web query classification," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006, pp. 131–138.

[10] Charudatt Mane, PallaviKulkarni," A Novel Approach to Discover User Search Goals Using Clickthrough Data", Charudatt Mane et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 20-24.

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