Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

CATEGORIZING WEB SEARCH RESULTS FROM SEARCH ENGINE LOGS

Darshana S. Parikh[1] and Prof. S.M. Patil[2]
[1]*Student, M.E. computer department, B.V.C.O.E. Navi Mumbai,*
[2]*Head of IT Department, B.V.C.O.E. Navi Mumbai*

Abstract- For a query, search engine provides platform for users. Different users have different search goals when they submit it to a search engine. The user search goals can be very useful in improving search engine relevance end-user experience. To increase retrieval precision, some new search engines provide manually verified answers to Frequently Asked Queries (FAQs). An underlying task is the identification of FAQs. This paper describes our attempt to cluster similar queries according to their contents as well as user logs. Our preliminary results show that the resulting clusters provide useful information for FAQ identification. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. The pseudo-documents are clustered using Fuzzy C Means, the fuzzy similarity based self- constructing algorithm. A novel optimization method is used to map feedback sessions to pseudo-documents which can efficiently reflect user information needs and finally, a new criterion "Classified Average Precision (CAP)" is used to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness.

Keywords: User Search Goal, Feedback Session, Fuzzy C Means Algorithm, Classified Average Precision

I. INTRODUCTION

It is an approach for end users to search result with their feedback session. Here, we are clustering the feedback session by using Fuzzy c-means algorithm. Also we use method to map feedback sessions to pseudo-documents which can efficiently reflect required data .Then, we evaluate the "Classified Average Precision (CAP)" of restructured web search results .Generally, and data mining is the process of analyzing data from different perspectives and summarizing it into useful information. 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. Data mining is the process of finding correlations or patterns among many 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. Prior algorithm is a methodology of association rule of data mining, is used to find out the frequently used URL.

There were Problems in information retrieval like how we represent the documents with select keywords. How document and query representations are compared to calculate the weigh. Mismatching of vocabularies .Ambiguous query .Depicting of content may be incomplete and inadequate. To discover

the number of users search goal for a query we have a framework of our approach consisting two parts. In first section, the feedback session is used where clicked and unclicked logs are recorded and scanned. Thus feedback sessions are generated. Then on the basis of this pseudo documents are generated. Then clustering has been done. Next clustering method is fuzzy-c clustering algorithm is used which will generate clustering results. After that classified average precision evaluation method is done.

II. PROPOSED SYSTEM

In this we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a novel approach to infer user search goals for a query by clustering our proposed feedback sessions. Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords

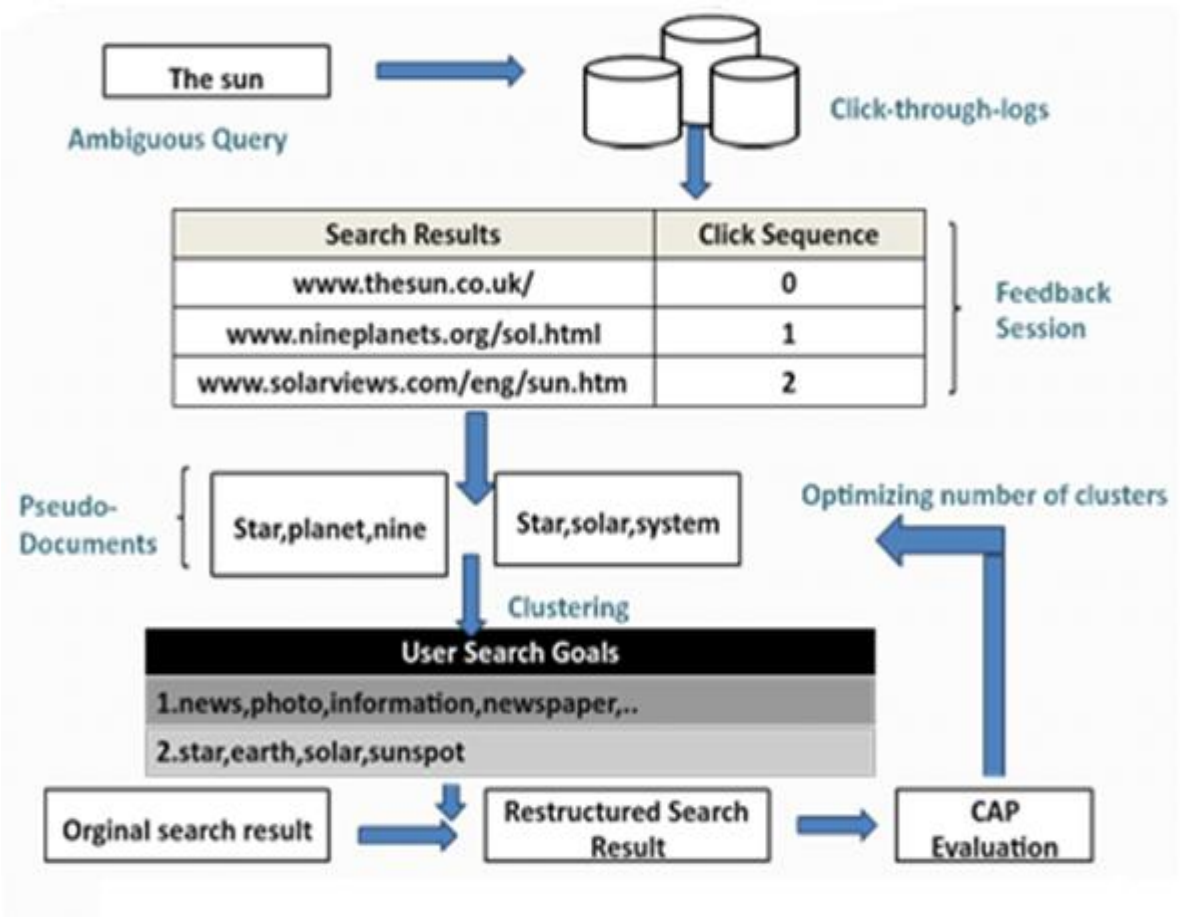

Figure 1. Architectural Framework

Framework of the approach.

- Representation of feedback sessions.
- Inferring user search goals by clustering pseudo-documents.
- Evaluation based on restructuring web search results

The framework consists of two levels.

A. The upper level

Feedback sessions are first extracted from user click-through logs. These are mapped to pseudo documents. User search goals are inferred by clustering the pseudo documents, with some keywords

B. The lower level

The original search results are restructured based on the user search goals inferred. Then evaluate the performance of restructuring search results by CAP criterion. This evaluation result is used as the feedback to select optimal number of search goals in the upper part

III. FEEDBACK SESSION

Session is used in reference to web application. It is a sequence of connection between user and server. The feedback session is combined of clicked and unclicked URLs. All URLs before the last click are scanned and analyzed by users and all these URLs are considered as feedback. Left part of the figure shows search results for the query and the right part of the figure shows sequence of user's clicks. Here '0' shows unclicked URLs. And number shows clicked URL's.

The clicked URL reflects that user wants and unclicked URL tells that user does not care.The proposed feedback session includes both clicked and unclicked URLs in a single session. The clicked URLs tell what users require. Unclicked URLs reflect what users do not care about. The unclicked URLs after the last clicked URL are not includes in the feedback session.

Sr. No	Search Result	Click Sequence
1	Nair Technologies Pvt Ltd, Navi Mumbai | Application Development ... http://www.nairtechnologies.com/	1
2	Software development navi mumbai, customize software ... - Arete http://arete.in/development.aspx	0
3	Software development - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Software_development	0
4	Ajoft Technologies: Custom Software Development Company ... http://www.ajoft.com/	2
5	Software Testing Interview Questions and Answers - IndiaBIX http://www.indiabix.com/technical/software-testing/	3
6	Free Software Testing Course - Guru99 http://www.guru99.com/software-testing.html	0
7	Software Testing Tutorial - TutorialsPoint http://www.tutorialspoint.com/software_testing/	0
8	Software testing - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Software_testing	0
9	Software Testing Fundamentals | The Basics of Software Testing for ... http://softwaretestingfundamentals.com/	1

Figure 2. Sample Feedback Session

IV. FORMING PSEUDO DOCUMENT

Mapping of feedback session to pseudo-document have two steps

A. Map feedback sessions to pseudo-documents

It is unsuitable to use the feedback sessions directly for inferring user search goals. Thus some representation is needed to describe feedback sessions in a coherent way. Binary Vector Method can be used to represent the feedback session where Clicked URL=1 Unclicked URL=0

BINARY (FEEDBACK SESSION) : 100110001

Sr. No	Search Result	Click Sequence	Binary Vector
1	Nair Technologies Pvt Ltd, Navi Mumbai | Application Development ... http://www.nairtechnologies.com/	1	1
2	Software development navi mumbai, customize software ... - Arete http://arete.in/development.aspx	0	0
3	Software development - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Software_development	0	0
4	Ajoft Technologies: Custom Software Development Company ... http://www.ajoft.com/	2	1
5	Software Testing Interview Questions and Answers - IndiaBIX http://www.indiabix.com/technical/software-testing/	3	1
6	Free Software Testing Course - Guru99 http://www.guru99.com/software-testing.html	0	0
7	Software Testing Tutorial - TutorialsPoint http://www.tutorialspoint.com/software_testing/	0	0
8	Software testing - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Software_testing	0	0
9	Software Testing Fundamentals | The Basics of Software Testing for ... http://softwaretestingfundamentals.com/	1	1

Figure 3. Binary Feedback Session

B. Representing URLS in the feedback session

First enrich the URLS with additional textual contents by extracting the titles and snippets. In this way each URL in a feedback session is represented by a small text paragraph that contains titles and snippets. Then some textual processes are implemented to those paragraphs such as: Transforming all letters, stemming, Removing stop words. Finally, each URL's title and snippet are represented by TF-IDF(Term Frequency-Inverse Document Frequency) vector.

T ui = [t w1 , t w2 , … t wn]^ T

S ui = [s w1 , s w2 , … s wn]^ T

Where Tui and sui are the TF-IDF vectors of the URL's title and snippet. Ui is the i-th URL in the feedback session Wj is the j- th term appearing in the enriched URL .Since title and snippet have different significances, we represent enriched URLs as weighted sum of Tui and Sui.

Fui = wtTui + wsSui = [fw1, fw2, .. fwn]^T

Fui indicates the importance of a term in the ith URL. Title is given a weight 2 throughout this paper. Similarly each URLs of a feedback session is represented and finally pseudo-documents are formed.

F fs =[f fs (w1), f f2(w2),… f fs (wn)]^T

F fs (w)= arg min{∑[f fs (w)- f ucm (w)]^2

-ʎ ∑[f fs (w)-~ f uc (w)]^2}

V. 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 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. FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target 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 iteration of 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.

VI. CLUSTERING PSEUDO-DOCUMENTS USING FUZZY C MEANS ALGORITHM

Each feedback session is represented by pseudo-document and the feature representation of the pseudo-document is Ffs. 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 *uij* and the cluster centers *cj* by:

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

VII. EVALUATION BASED ON RESTRUCTURING WEB SEARCH RESULTS

User search goals are not predefined; hence evaluation of its inference is a big problem. If user search goals are inferred properly, the search results can be restructured properly. Thus an evaluation method 'Classified Average precision' is proposed. It helps to select the best cluster number it is an application of inferring user search goals. The inferred ones are represented by the feature representation of each URL in the search result. Then categorize them into a cluster cantered by the inferred search goals. This is done by choosing smallest distance between URL vector and user-search goal vectors. Thus restructuring is complete. From user click through logs, we get the relevant and irrelevant feedbacks.

The Evolution of evaluation method is as follows

- Average Precision(AP)
- Voted Average Precision(VAP)
- Classified Average Precision(CAP)

NUMBER OF UNLCICKED URL : 5
NUMBER OF CLICKED URL : 4

Sr. No	Search Result	Click Sequence	Rel(r)[R/r]
1	Nair Technologies Pvt Ltd, Navi Mumbai | Application Development ... http://www.nairtechnologies.com/	1	1/1
2	Software development navi mumbai, customize software ... - Arete http://arete.in/development.aspx	0	0/2
3	Software development - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Software_development	0	0/3
4	Ajoft Technologies: Custom Software Development Company ... http://www.ajoft.com/	2	2/4
5	Software Testing Interview Questions and Answers - IndiaBIX http://www.indiabix.com/technical/software-testing/	3	3/5
6	Free Software Testing Course - Guru99 http://www.guru99.com/software-testing.html	0	0/6
7	Software Testing Tutorial - TutorialsPoint http://www.tutorialspoint.com/software_testing/	0	0/7
8	Software testing - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Software_testing	0	0/8
9	Software Testing Fundamentals | The Basics of Software Testing for ... http://softwaretestingfundamentals.com/	1	1/9

EVALUATION

CAP : 0.146604742284

Figure 4. Evaluation

VIII. 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 earch goal inference. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods. The complexity of our approach is low and our approach can be used in reality easily. For each query, the running time depends on the number of feedback sessions. The running time is usually short. In reality, our approach can discover user search goals for some popular queries offline at first. Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently.

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] J.I.Sheeba, Dr.K.Vivekanandan, "A Fuzzy Logic Based On Sentiment Classification", 2014.
[2] K.Jeyalakshmi, R.Deepa, M.Manjula, "An Efficient Clustering Sentence-Level Text Using A Novel Hierarchical Fuzzy Relational Clustering Algorithm", 2014
[3] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", 2013.
[4] Vicenc¸ Torra, "Fuzzy c-means for fuzzy hierarchical clustering", 2012.
[5] L.Suganya, Dr.B.Srinivasan, "Efficient Semantic Similarity Based Fcm For Inferring User Search Goals With Feedback Sessions", 2013
[6] Charudatt Mane, Pallavi Kulkarni, "A Novel Approach to Discover User Search Goals Using Click through Data", 2014
[7] Eugene Agichtein, Eric Brill, Susan Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information", 2006
[8] Ji-Rong Wen, Jian-Yun Nie, Hong-Jiang Zhang, "Clustering User Queries of a Search Engine", 2001.
[9] Rohini B. Mothe, V.S.Deshmukh, "A Novel Approach to Cluster Search Result based on Search Goals", 2014
[10] S.Niveditha, T.Malathi, S.R.Sivaranjhani, "Efficient Information Retrieval using Fuzzy Self Construction Algorithm", 2014
[11] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. 2006, "Learning user interaction models for predicting web search result preferences". In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06. ACM, New York, NY, USA, 3–10.
[12] T. Joachims, "Evaluating Retrieval Performance Using Click through Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physics/Springer Verlag, 2003.
[13] T. Joachims, "Optimizing Search Engines Using Click through Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.
[14] Zheng Lu, HongyuanZha, Xiaokang Yang, Weiyao Lin, and ZhaohuiZheng, 2013 "A New Algorithm for Inferring User Search Goals with Feedback Sessions" Published by the IEEE Computer Society.