A Novel Approach to Cluster Search Result based on Search Goals

Rohini B. Mothe
(M.E. Student) Department of Computer Engg.
STES'S SKNCOE,
Pune, Maharastra, India

V.S.Deshmukh
ASST..Professor Department of Computer Engg
STES'S SKNCOE,
Pune, Maharastra, India

ABSTRACT

In present days world wide web provides a platform for users to satisfy their information needs, for this purpose search engine tools are commonly used. Available search engine give result for a particular query in the form of flat rank list, which works well for non-ambiguous query.But,in case of ambiguous query which having multiple aspects the flat rank list not works well. So in such cases reorganization of search result is necessary. In this paper, proposed a method which reorganizes search result by analyzing user's implicit feedback. Based upon this feedback doing text processing, enriching each url by combination of title and snippet ,and mapping these data to Pseudo-document. Pseudo-document contain set of keywords which are different aspects of query. And then performing clustering on these pseudo-document using fuzzy k-mean clustering. And these clusters contain links which are most relevant to each other. Also rearranging results based upon most visited links such that it should occur at topmost. And this reorganization will increase the performance and evaluation of search engine. And the cluster labels.

General Terms

Clustering Algorithm.

Keywords

Fuzzy k-means clustering, Implicit feedback, Pseudo-documents, User search goals.

1. INTRODUCTION

Available search engine tools works well for a non ambiguous query which doesn't have broad meaning .But in the case of ambiguous query which is having multiple aspects, where different users have different aspects for same query, these tools not provide user's interested result, as these tools provide results in the form of flat rank list. Consider a scenario, when user submits a query "Sun" to search engine, some users are interested to know information about Sunflower and some users want to know information about technology and some users may interested in solar system. To provide search results according user's interested aspects ,it is essential to find out different search goal text and to reorganize search results on basis user search goal using Fuzzy k-mean clustering to get user its interested aspects.

Evaluation of user search goal plays an important role and it might have a numeral of advantages, one of its advantages is enhancing the search engine performance and user knowledge. Evaluating different user search goals related to information needs changes the normal query based information retrieval and to improve utility of search engine, it is necessary to collect the different user goal as well as retrieve the efficient information on different aspects of a query.

For effective reorganization of search results it is necessary to analysis of search results, which is also used to optimize search engine. When submitting query to search engine, the returned web pages of search results are analyzed [7], [8]. But analyzing of search results without considering user feedback, many unwanted and noisy search result that are unclicked by user may get analyzed, which is time consuming and may degrade the search goals discovery. Learning interesting aspects of similar query/topic from web search logs which consists clicked web pages URLs and organize search results accordingly this approach present in[7] by Wang and C-X. Zhai. Deficiency of these approach results in limitation, as the different clicked URLs for a query may be small in number. In [14], [4] here they used query classification approach where classify queries into some predefined classes and try to find out query intents and user goals. But in case of non ambiguous query having multiple aspects for each predefining a class and such classes for each aspect of query and for such multiple queries is critical job. Predefining classes may be difficult and sometimes impossible to categorize.

So clustering of search result is an efficient way to organize search result in systematic and useful way .And it is an good approach to get user its interested document easily. In this approach, our aim is to discover different user search goals for a query and depict each search goal with some keywords automatically which used as labels of clusters. To discover the user interested information automatically, Firstly collecting feedback session by analyzing search engine log data. Afterwards, mapping feedback sessions to documents known as pseudo-documents by using text processing methods, which reflects user information needs. These pseudo-documents contains keywords which are user search goals. Finally, clustering of pseudo-documents done by using Fuzzy K- means clustering algorithm for inferring user search goals and depicting them with some meaningful keywords. So user search goals plays an important role to restructure the web search results.

2. RELATED WORK

Due to advantages of clustering web search results lots of work has been done in this area. Many previous works has been investigated on problem of analyzing user query logs [13], [9], [13], [4], [6]. The information present in (search) query logs can be used in multiple purposes, such as to infer search query intents or user goals, to classify queries, to provide personalization based on search results, also for suggesting query substitutes. To enhance utility as well as relevance of any search engine, effective organization of search results is necessary and which is critical .One of the advantage of clustering is it allows a user to navigate into relevant documents quickly which is the best way .Presently all existing work [7], perform clustering on a set of top ranked results generated by search engine, to partition generated results into general clusters, which may contain different subtopics of the general query term. But, this strategy of clustering has two deficiencies which make it not always work well. First, resultant clusters do not necessarily correspond to the interesting aspect of a topic from user-oriented perspective. Second, cluster labels are more general and not informative to identify appropriate clusters. Wang and Zhai [3] proposed approach to organize search results in user-oriented manner. In this strategy they have used search engines log to learn interesting aspects of similar queries and categorize search results into aspects learned. Cluster labels are generated by using past query words entered by users.

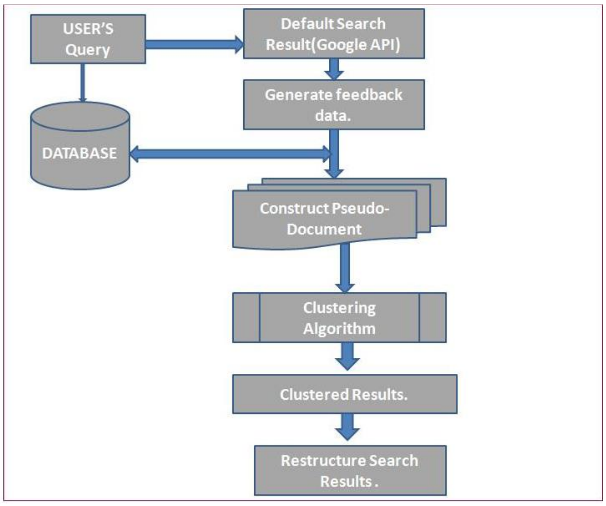
USER'S Query
Default Search Result(Google API)
Generate feedback data.
DATABASE
Construct Pseudo-Document
Clustering Algorithm
Clustered Results.
Restructure Search Results .

Figure 5.1 :System Architecture.

In [1] by Zheng Lu, Hongyuan Zha proposed approach which is based on feedback session. Here they considered both click and unclick link. Here Single session related to only one query. By analysing user feedback they construct feedback session. On the basis of this feedback session they perform clustering using fuzzy k-mean algorithm.

In[18]by Wang and Zai clustered queries and learnedaspects of these similar queries .In [3] Zheng lu,Hongyuan zha,Weiyao lin and Zhaohui zheng in this they have used feedback session in which considering user feedback and on basis of that generating results. Some works [16] based on search goals and mission to detect session boundary hierarchically. It identifies whether a pair of queries belong to same goal or mission.In [2],[9],[7] by Thorsten Joachims did many works to improve retrieval by using implicit feedback.

In [17] by Xin ye Li proposed an improved k-means algorithm by using a few of related and unrelated feedback to guide clustering Web retrieval result. The improved algorithm first selected initial cluster centroid based on feedback messages, then during the clustering process, it removed large unrelated documents which increased the clustering speed and optimized the clustering result. During the clustering process, the centroids of clusters including unrelated documents needn't be modified in order to avoid noise influence. Experiment result illustrate that this algorithm is superior to the traditional k-means algorithm.

In [7] by H-J Zeng proposed method to cluster search results which is a query based. In this 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. Clusters are formed by assigning documents to relevant salient phrases known as candidate clusters and by merging these candidate clusters this the final cluster are generated . This method only produces the result with higher level of the documents only [7].

As stated by H. Chen and S. Dumais [8] in this they organize web search results into hierarchical categories. For classifying search results they used Automatic text classification technique (SVM classifier) .Advantage of known category labels information, for classifying new items into the category structure and to help user to quickly focus on task relevant information [8].

3. PROBLEM STATEMENT

Effective way to reorganize search results is clustering of web search result. Here in this approach reorganizing search results truly based on user search goals. These search goals represents user's interested aspect. Discover the number of user search goal for a query based upon these keywords and using fuzzy k-mean clustering algorithm, forms the cluster which contain one label which is one of the aspect of query and that cluster contain links related to each other and label .And rearrange in such way that top most visited links should occur at topmost.

. These will be added when the publications are assembled.

4. PROPOSED APPROACH

In this section, describing proposed approach in which reorganization of search results can be done using search goal and fuzzy K-mean algorithm. Flow of proposed approach in figure 5.1.

As shown in figure 5.1. If the feddback data is not present in databse for query then using google api showing result same as google.When for a query get user's implicit feedback then mapping these feedback data to pseudo-document which contain set of keywords.Finally using fuzzy k-mean clustering algorithm clustering pseudo-documents for that query.then by using cosine similarity and Euclidian distance mapping similarity between documents.Finally reaarnging links such that most visited links should occur at topmost.

Constructing pseudo-documents:

Every URL present in feedback data is combination of its title and snippet which is small textual content and URLs alone are not so much informative, snippet which present with that URL contain important information which are useful to achieve intended meaning of a submitted query. To enriching information, here enriching each URL by extracting the titles and snippets of URLs stored in feedback session. Then afterwards text pre-processing is done on those textual contents, such as removing stop words, transforming all letters to lowercase, word stemming by using porter algorithm [16]. Finally, TF-IDF [8] vector of URL's titles and snippets are formed respectively as:

$$T_{ui} = [t_{w1}, t_{w2}\ldots\ldots\ldots\ldots\ldots\ldots\ldots \, twn]^{T}$$

$$S_{ui}=[S_{w1}, S_{w2},\ldots\ldots\ldots\ldots\ldots \, S_{wn}]^{T} \qquad (1)$$

Here $\mathbf{T_{ui}}$ and $\mathbf{S_{ui}}$ are TF-IDF vectors of URL’s title and snippet, respectively. u_i is ith URL in feedback session. where Wj is the jth term present in the enriched URL. And the term $\mathbf{t}_{wj}$ and $\mathbf{s}_{wj}$ denotes jth term in the URL’s title and snippet respectively. Here in this approach enriching of URL known as Feature representation of that URL.Feature representation of F_{ui}, of iith enriched URL is weighted sum of T_{ui} and

$$F_{ui}=w_1T_{ui}+ w_2S_{ui}=[f_{w1},f_{w2}\ldots\ldots\ldots f_{wn}]^{T} \qquad (2)$$

where $\mathbf{w}_t$ and $\mathbf{w}_s$ are weights of title and snippet respectively. Each term of $\mathbf{F}_{ui}$, represents importance of term in ith URL.Optimization method is used to merge feature representations of each clicked and unclicked enriched URLs in the feedback for obtaining feature representation of a feedback, optimization method is used. Let F_{fs} be feature representation of a feedback session, $\boldsymbol{F_{ucm}}$and F_{uci} are feature representation of clicked and unclicked URLs respectively and F_{fs} is value for term F_{fs}.and it should be such that sum of distance between F_{fs} and $F_{\mathbf{uc}i}$ each is minimized and sum of distance between F_{fs} and $F_{\mathbf{uc}i}$ is maximized.

$$F_{fs}=[ff_{s(w1)}, ff_{s(w2)}\ldots\ldots\ldots\ldots ff_{s(wn)},]^{T} \qquad .(3)$$

Feedback is represented by F_{fs}. This is nothing but pseudo-document which is used for discovering user intents or search goals. These pseudo-documents contain what user requires and what do not, which is used to learn interesting aspects of a query.

C: Clustering pseudo-documents with Improved K-means:

Now next step is clustering of pseudo-documents with fuzzy k-mean clustering algorithm, the important factor is to define the distance measure between two data points as well as defining the number of clusters. Firstly representing each document using vector space model with the help of Tf-IDF value.As mentioned above the feature representation of pseudo-document is F_{fs} and similarity between two pseudo-documents is defined as below:

$$Sim_{i,j} =\cos (F_{fsi}, F_{fsj}) \qquad (4).$$

Here to cluster document ,it is necessary to represent them in form of vector space model,for thathere using TF-IDF value for each docunment.

Cluster denotes user search goal i.e. intention of user and centroid of a cluster is calculated by taking average of all the vectors of the pseudo- documents in the cluster,

$$F_{centeri}=\frac{\sum_{k=1}^{C1} Ffsk}{Ci}(F_{fsk} \subset \text{Cluster i}) \qquad (6)$$

$F_{center\ i}$ is ith cluster center and C_i is the number of pseudo-documents in the ith cluster is used represent user search goal/intent of ith cluster and $F_{centeri}$ to categorize the search results. User search goals/intents are the terms with highest values in the centre points of each cluster. These keywords can be used to suggest more meaningful labels of clusters.

D. Rearranging web search results

Reorganization of web search results are done on the basis of discovered user search goals which achieve by analyzing search results as mentioned above, inferred user search goals represents with vectors in (6) and feature representation of each URL in search result is calculated by (1) and (2) . By selecting the smallest distance between user search goal vectors and URL vectors categorizing each URL into a cluster centered with user search goals/intents.And finally rearranging links based on most visited links occur at topmost.

E. Evaluation criterion

To evaluate performance of restructured (clustered) web search results and original search results , using parameters like Average Precision (AP) [1], Voted AP (VAP) which is AP of the class having more clicks, Risk to avoid wrong classification of search results and Classified AP (CAP). If user got correct classified results with higher AP value, this value is used to optimize the no of clusters of user search goals.

1) Average precision (AP): Calculated according to given user feedbacks. It is the average of precisions computed at the point of each clicked document in the ranked sequence of user feedback.[1]

$$AP=\frac{1}{N+}\sum_{r=1}^{n} \mathrm{rel}\,(r)\,\frac{Rr}{r}$$

.N+ denotes the number of clicked documents from total retrieved documents in single user feedback session, 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 Rr is the number of relevant retrieved documents of rank r or less.

2) Voted AP (VAP): VAP is calculated for restructured search results classes i.e. different clustered results classes. It is same as AP and calculated for class which having more clicks i.e. the class user interested in.

$$VAP=\frac{1}{NC}\sum_{r=1}^{n} \mathrm{rel}\,(r)\,\frac{Rr}{r}$$

where N_C is the number of clicked documents from the class having maximum number of clicks.[1]

3) Risk: At sometimes VAP will always be highest value because each URL from single session is classified into the single class no matter whether users have different search goals or not. So, there should be a risk to avoid wrong classification search results into too many classes. It evaluates the normalized number of clicked URL pairs that are not in the same class

$$Risk=\frac{\sum_{i,j=1(i<j)}^{m} dij}{Cm2}$$

where m is number of clicked URLs and d_{ij} is 0 if pair of clicked URLs belongs to same class otherwise d_{ij} is 1.[1]

4) Classified AP (CAP): New criterion Classified AP (CAP) is extension of VAP by using above Risk. It combines AP of class having more clicks and risk of wrong classification. It is used to evaluate performance of restructured search results.

$$CAP=VAP^{*}\,(1-\mathrm{Risk})^{\gamma}$$

where γ is normalizing factor used to adjust influence of Risk on CAP. Generally, categorizing search results into less 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.[1]

5. RESULTS AND DISCUSSIONS

Here as mentioned in [1],with the help of CAP,AP,VAP parameter to check performance of proposed system. Here dataset is real time data that is user feedback .The following graph shows the comparison between proposed method and previous method. Following graphs shows results for 50 queries .X-axis represent query ID and Y-axis Risk, CAP parameter. As shown in figure 6.1 shows comparison between proposed method and old method[1] based on risk parameter. Proposed method shows less risk value. And in figure 6.2 proposed method shows highest value for CAP parameter.And in figure 6.3 shows highest value for vap or proposed syatem than old method.In [1] mention that the system has best performance if it it has less risk value and highest,VAP CAP value. Based upon these graph we can show that proposed method has best results as compare to [1] old method.

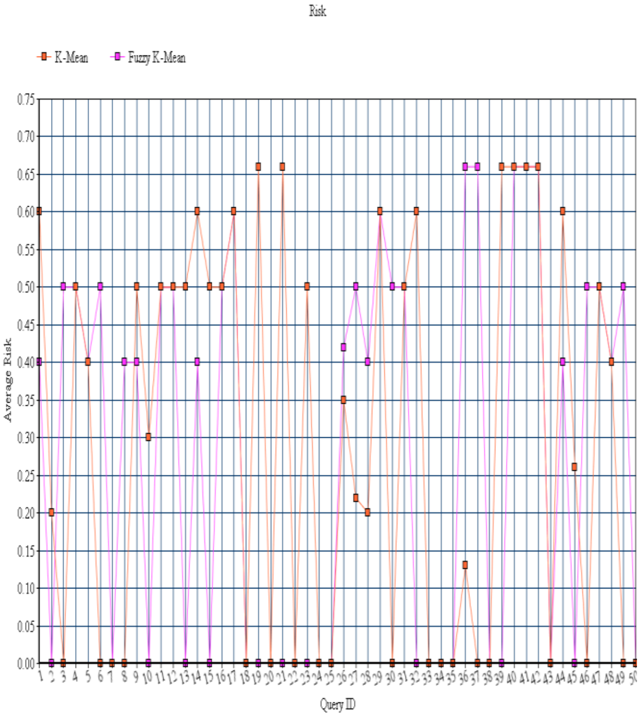

Fig:6.1 Risk Based Parameter Comparison

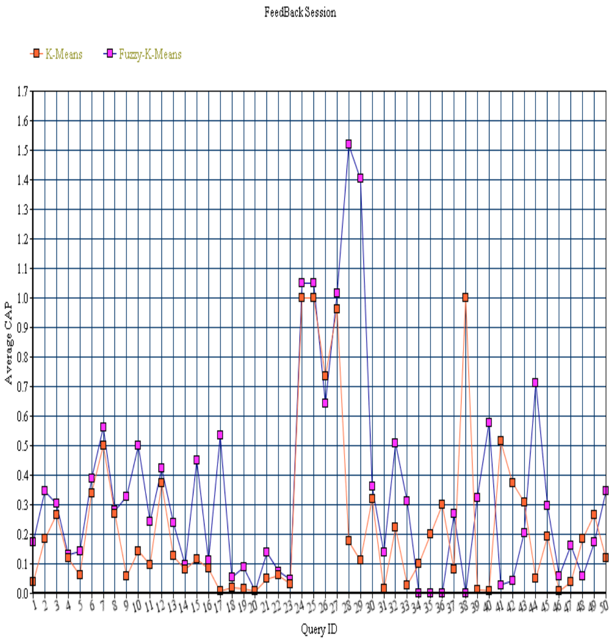

Fig:6.2 CAP Based Parameter Comparison

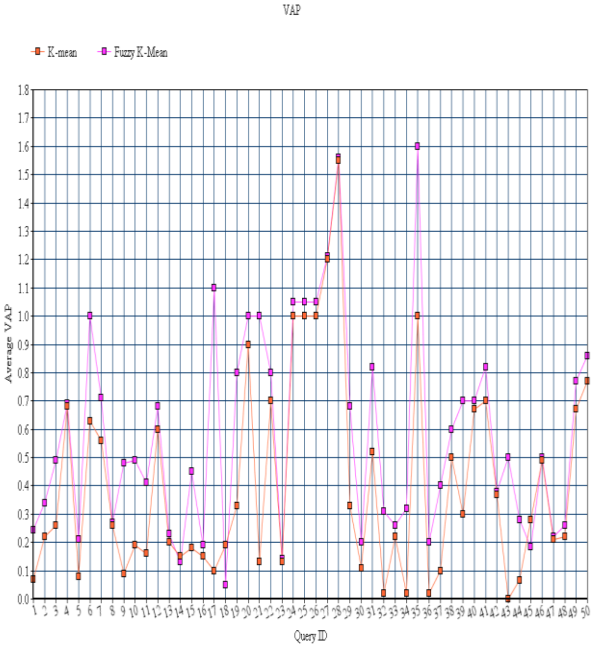

Fig:6.3 CAP Based Parameter Comparison

6. CONCLUSION

This paper proposed an approach to automatically reorganize search result .This approach completely based user's implicit feedback data.and fuzzy K-mean clustering algorithm, By using clustering algorithm forms the cluster whose center will predict clusters label which will nothing but user's interested aspects. And different clusters of a query show the different aspect of query and contain relevant document. And finally rearranging links such that most visited links occur at topmost. Future work will be to collaborate query classification and serach result combination so that user will get more classified results.

7. REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, ZhaohuiZheng, "*A New Algorithm for Inferring User Search Goals withFeedback Sessions*", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, pp.502-513,2013.

[2] R. Jones and K.L. Klinkner, "Beyond the Session Timeout:Automatic Hierarchical Segmentation of Search Topics in QueryLogs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 699-708, 2008.

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

[4] 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.

[5] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay,"Accurately Interpreting Click through Data as Implicit Feedback,"Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Developmentin Information Retrieval (SIGIR '05), pp. 154-161, 2005.

[6] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," *P*roc. Int'lConf. Current Trends in Database Technology (EDBT '04), pp.588-596, 2004.

[7] 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 SIGIRConf. Research and Development in Information Retrieval(SIGIR '04), pp. 210-217, 2004.

[8] 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.

[9] T. Joachims, "Optimizing Search Engines Using ClickthroughData," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

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

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

[12] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

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

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

[15] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[16] O. Zamir and O. Etzioni. "*Grouper:* A dynamic clustering interface to web search results. Computer Networks", 31(11-16), pp.1361-1374, 1999.

[17] Xinye Li "An improved method in clustering Web retrieval result based on relevance feedback", Computer Science and Service System (CSSS), IEEE International Conference ,pp. 3000 - 3003,2011.