ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Clustering of Web Search Result Based on C-Means Algorithm and User Search Recommendation

Prof. M.M. Siddiqui, Mrudula Mahajan, Nisarg Kadam, Swwapnil Kambley, Sonali Chaudhari

Dept. of Computer, MES College of Engineering, Savitribai Phule Pune University, Pune, India

ABSTRACT: Clustering of web search results or web document clustering; has become a very interestingresearchareaamongacademicandscientificcommunitiesinvolvedininformation retrieval (IR) and web search. To obtain good results in web document, clustering the algorithms must meet the following specific requirements: Automatically define the number of clusters to be created; generate relevant clusters for the user and assign the documents to appropriate clusters; define labels or names for the clusters that are easily understood by users; handle overlapping clusters (this means thatdocumentscanbelongtomultipleclusters);handleshortinputdatadescriptions (documentsnippets);reducethehigh-dimensionthatispresentedinthemanagement of document collections; handle the processing time (the algorithm must be able to work with snippets and not only with the full text of the document); and handle the noise that is very common in the collection ofdocuments. We proposed a method for personalized web search. Personalized web search is any action taken to optimize the search result according to user's individual preferences. Different information retrieval techniques have been widely used to reduce access latency problem of the internet. Traffic behaviour analysis methods do not depend on the packets payload, which means that they can work with encrypted network communication protocols.

KEYWORDS: Sorting and searching; Query suggestion; Pattern matching; Clustering; Web log analysis;Image Processing; Information retrieval

I. INTRODUCTION

As the amount of information on the Web rapidly increases, it creates many new challenges for Web search. When the same query is submitted by different users, a typical search engine returns the same result, regardless of who submitted the query. This may not be suitable for users with different information needs. For example, for the query apple, some users may be interested in documents dealing with apple as fruit, while some other users may want documents related to Apple computers. One waytodisambiguatethewordsinaqueryistoassociateasmallsetofcategorieswith the query. For example, if the category cooking or the category fruit is associated with the query apple, then the user's intention becomes clear. Current search engines such as Google or Yahoo! have hierarchies of categories to help users to specify their intentions. The use of hierarchical categories such as the Library of Congress Classification is also common among librarians.

II. RELATED WORK

Clustering of Web Search Results based on an Iterative Fuzzy C-meansAlgorithmandBayesianInformation Criterion: Theclusteringofwebsearchhasbecomeaveryinterestingresearcharea among academic and scientific communities involved in information retrieval. Clustering of web search result systems, also called Web Clustering Engines, seek to increase the coverage of documents presented for the user to review, while reducing the time spent reviewing them. Several algorithms for web document clustering already exist, but results show there is room for more to be done. This paper introduces a new description-centric algorithm for clustering of web results called IFCWR. IFCWR initially selects a maximum estimated number of clustersusingForge'sstrategy,thenititerativelymergesclustersuntilresults cannot

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

be improved. Every merge operation implies the execution ofFuzzy C-Means for clustering results of web search and the calculus of Bayesian Information Criterion for automatically evaluating the best solution and number of clusters.

• Personalized Web Search Engine using Dynamic User Profile and ClusteringTechniques: Internet is large interconnection of small networks that is commonly known as World Wide Web. The amount of information's available on internet in digital form is very huge and growing at exponential rate following Moore's law. So, it's makes difficult to find exact search result according to user preferences. In this paper, we proposed a method for personalized web search. Personalized web search is any action taken to optimize the search result according to user's individual preferences. Different information retrieval techniques have been widely used to reduce access latency problem of the internet. This paper comprised and focuses different techniques for efficient personalized web search and also suggests the techniques for personalized web search according to the merits and demerits of various available techniques.

• PersonalizedWebSearchforImprovingRetrievalEffectiveness: Current Web search engines are built to serve all users, independent of the special needs of any individual user. Personalization of Web search is to carry out retrieval for each user incorporating his/her interests. We propose a novel technique to learn user profiles from users search histories. The user profiles are then used to improve retrieval effectiveness in Web search. A user profile and a general profile are learned from the user's search history and a category hierarchy, respectively. These two profiles are combined to map a user query into a set of categories which represent the user's search intention and serve as a context to disambiguate the words in the user's query. Web search is conducted based on boththeuserqueryandthesetofcategories. Severalprofilelearningand category mapping algorithms and a fusion algorithm are provided and evaluated. Experimental results indicate that our technique to personalize Web search is both effective and efficient. Specifically, we provide a strategy to: 1. model and gather the user's search history, 2. construct a user profile based on the search history and construct a general profile based on the ODP (Open Directory Project 1) category hierarchy, 3. deduce appropriate categories for each user query based on the user's profile and the general profile, and 4. Improve Web search effectiveness by using these categories as a context for each query.

• CollaborativeFuzzyClusteringfromMultipleWeightedViews: Clustering with multi view data is becoming a hot topic in data mining, pattern recognition, and machine learning. In order to realize an effective multi view clustering, two issues must be addressed, namely, how to combinetheclusteringresultfromeachviewandhowtoidentifytheimportanceofeachview.
Inthispaper,basedonanewlyproposedobjective function which explicitly incorporates two penalty terms, a basic multi viewfuzzyclusteringalgorithm,calledcollaborativefuzzyc-means(CoFCM), is firstly proposed. It is then extended into its weighted view version, called weighted view collaborative fuzzy c-means (WV-Co-FCM), by identifying the importance of each view.

III. PROPOSED ALGORITHM

Fuzzy c-Means Algorithm:
The fuzzy c-means (FCM) algorithm is a clustering algorithm developed by Dunn, and later on improved by Bezdek. It is useful when the required number of clusters are pre-determined; thus, the algorithm tries to put each of the data points to one of the clusters. What makes FCM different is that it does not decide the absolute membership of a data point to a given cluster; instead, it calculates the likelihood (the degree of membership) that a data point will belong to that cluster. Hence, depending on the accuracy of the clustering that is required in practice, appropriate tolerance measures can be put in place. Since the absolute membership is not calculated, FCM can be extremely fast because the number of iterations required to achieve a specific clustering exercise corresponds to the required accuracy.
In each iteration of the FCM algorithm, the following objective function '*J*' is minimised:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} \delta_{ij} ||x_i - c_j||^2$$

Here, n is the number of data points, c is the number of clusters required, c_j is the centre vector for cluster j, and δ_{ij} is the degree of membership for the i^{th} data point x_i in cluster i. Note that, in each iteration, the algorithm maintains a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

centre vector for each of the clusters. These data-points are calculated as the weighted average of the data-points, where the weights are given by the degrees of membership.

IV. PSEUDO CODE

Step 1: Give any user query as an input
Step 2: Fetch Google Search Result
Step 3: Calculate TF-IDF
Step 4: Apply C-means clustering algorithm: We apply fuzzy c-means clustering algorithm on collected objects for making the customized window on browsers for user search quickly.
Step 5: User Behavioural Analysis -

- Content-based = It is calculating on the basis of user search term, mouse event listener etc.
- Collaborative filtering = In this method Item based and user based filtering use on the basis of Search term with following attribute –
    1. Search Term Category
    2. User Click Stream
    3. Clustering results

Step 6: Customised window generation
Step 7: End

V. SIMULATION RESULTS

Clustering of web search result systems, also called Web Clustering Engines, seek to increase the coverage of documents presented for the user to review, while reducing the time spent reviewing them.

Memory-based recommender systems with m users and n items typically require O(mn) space to store the rating information. In itembasedcollaborativefiltering(CF)algorithms,thefeaturevectorofeach item has lengthm,and I ttakesO(m)timetocomputethesimilaritybetween two items using the Pearson or cosine distances.

Clustering of web search result systems, also called Web Clustering Engines, seek to increase the coverage of documents presented for the user to review, while reducing the time spent reviewing them.

We proposed a framework for personalized web search which uses a dynamic user profile to automatically update user profile and collaborativefilteringforconsideringrecommendationwhichhelpstoretrieve search result and relevant document to user according to its need and preferences by diagnosing its web search behavior according to previous search history.

User profile is used to represent user's interest and to infer their intention to user new queries. User profile can be created in two modes:

i) manually by user
ii) automatic profile generation using user search histories. Then user query is matched with related category, where it belongs to stored local database. There are different clustering algorithms that can be used to categorize the local database in different module so that user queries can be further matched with its profile and related category to show efficient search result.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

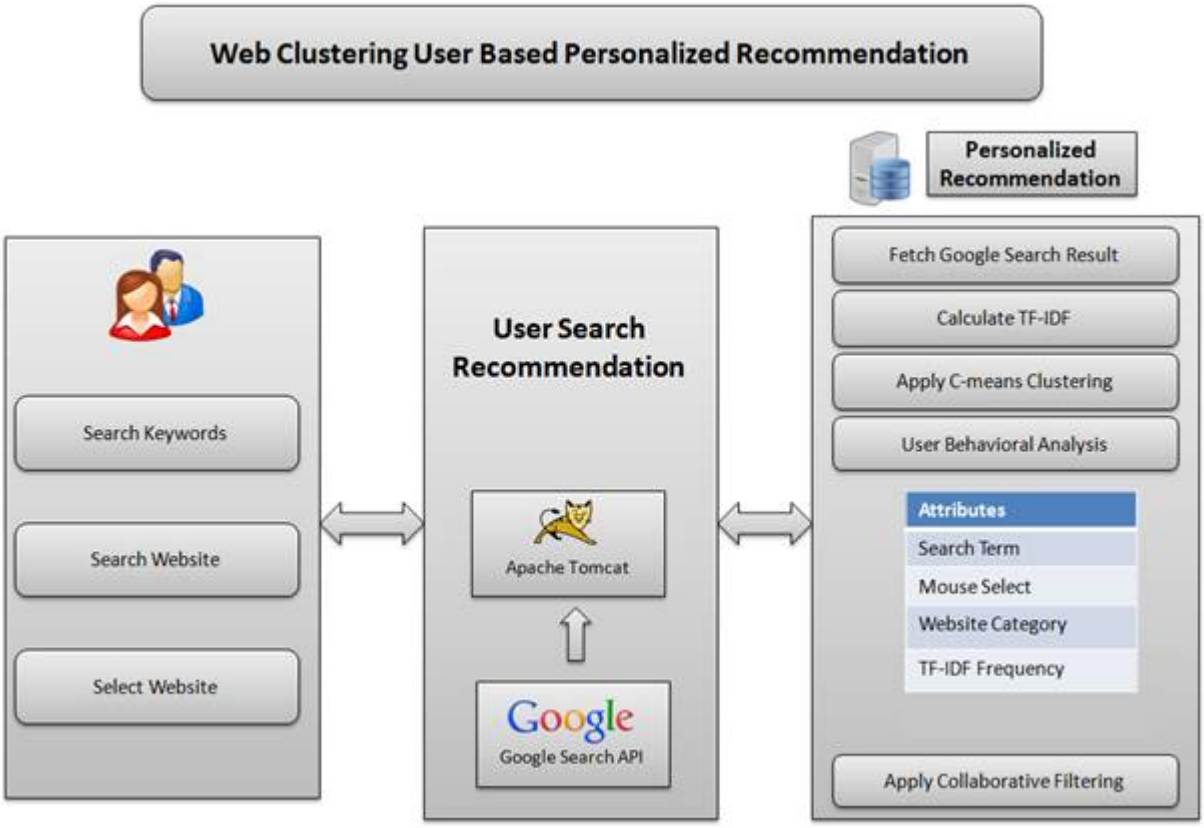

Fig.1. System Architecture

Fig.1.represents proper system component and mechanism of the working of Software. It works in three phases:

- Front End
- Google Search API
- Back End

Fig.2. Searched Query result

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

VI. CONCLUSION AND FUTURE WORK

Personalized web search is any action taken to optimize the search result according to user'sindividualpreferences. Also,thedifferentinformationretrievaltechniqueshavebeenwidelyusedtoreduceaccess latency problem of the internet.

REFERENCES

1. C. Carpineto and G. Romano, "Optimal meta search results clustering," presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland, 2010.
2. C. Carpineto, et al., "A survey of Web clustering engines," ACM Comput. Surv., vol. 41, pp. 1-38, 2009.
3. R. Baeza-Yates, A. and B. Ribeiro-Neto, Modern Information Retrieval: Addison Wesley Longman Publishing Co., Inc., 1999.
4. C. Carpineto, et al., "Evaluating subtopic retrieval methods: Clustering versus diversification of search results," Information Processing & Management, vol. 48, pp. 358-373, 2012.
5. Z. Oren and E. Oren, "Web document clustering: a feasibility demonstration," presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, 1998.
6. K. Hammouda, "Web Mining: Clustering Web Documents A Preliminary Review," ed, 2001, pp. 1-13.
7. Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short- and long-term behavior on search personalization.
8. Anoj Kumar; Mohd. Ashraf, "Personalized web search engine using dynamic user profile and clustering techniques".
9. K. Selvakumar; S. Sendhilkumar, Challenges and recent trends in personalizedWebsearch: A survey
10. Rong Hu; Wanchun Dou; Xiaoqing Frank Liu; Jianxun Liu, "PersonalizedSearching for Web Service Using User Interests"
11. J. Lai; B. Soh, "Personalized Web search results with profile comparisons"
12. M. R. Sumalatha; V. Vaidehi; A. Kannan; S. Anandhi, "Information Retrieval using Semantic Web Browse -Personalized and Categorical Web Search"
13. Masomeh Azimzadeh, Reza Badie, Mohammad Mehdi Esnaashari, "A review on web search engine's automatic evaluation methods and how to select the evaluation method"

BIOGRAPHY

Prof. M. M. Siddiqui, Assistant Professor,Dept. of Computer Engg., MES College of Engineering, Pune
Mrudula Mahajan, Student, BE Computer Engineering, MES College of Engineering, Pune
Nisarg Kadam, Student, BE Computer Engineering, MES College of Engineering, Pune
Swwapnil Kambley, Student, BE Computer Engineering, MES College of Engineering, Pune
Sonali Chaudhari, Student, BE Computer Engineering, MES College of Engineering, Pune