Survey of Clustering Techniques for Information Retrieval in Data Mining

Teena Rani, Ankush Goyal

***Abstract*— Relevancy of search results returned by the search engine for a user query is an important research area for many scholars. Most of the users are not able to accurately fire their query for an information need. Further because of the large data set search engines return a very large amount of data for a user query. It is the responsibility of the page ranking algorithm to arrange this result data according to its relevancy to user query. But user always find a large amount of unrelated information in the result list. Clustering is a useful data mining tool to handle this situation. The data set at the information retrieval system can be clustered using any of the clustering algorithm such as K-means, ROCK etc. In this paper a brief review of all the attempts made in improving the relevancy of search results using clustering techniques in recent years has been given.**

Index Terms : Information Retrieval, Clustering, Data Mining, K-Mean, ROCK.

I. INTRODUCTION

Search engines becomes a very common source of knowledge in recent years. People rely on search engines or information retrieval system for their information need now a days. Search engines keep on crawling the web to collect information which is available on world wide web. As the size of world wide web is keep on growing day and night so search engines stores millions of web pages in their repository. To extract the information for a user query search engines depends upon indexer module. Further a page ranking module arrange the pages returned for the user query by using page rank algorithms. The efficiency of information retrieval system mainly depends upon the performance of the page ranking algorithm. Clustering is a data mining technique which pus the related documents in separate distinct clusters. The application of clustering techniques to improve the performance of information retrieval system is analyzed by many authors in recent years[1][2][3].

In recent years web clustering search engines has been proposed to overcome the problem of ambiguity in information retrieval systems. These systems categories and cluster search results. Users can select clusters for their query and then they will get relevant information back for their query. The importance of data mining and knowledge discovery is increasing in the area of information retrieval systems. Text mining is field which deals in gathering meaningful information from textual data. It include the techniques like text categorization and text clustering. So if information retrieval systems will cluster the documents according to their similarity then more relevant results can be returned by these systems for a user query.

This paper is divided into three sections. Section I is gives a brief introduction about information retrieval systems and clustering techniques. Section II describes the literature reviewed and outcomes of the review with relative merit and demerit of the existing work. Section III gives conclusion and future scope of this survey.

Manuscript received March , 2015.

Teena Rani, M.Tech. student, Department of Computer Science and Engineering, Shri Ram College of Engineering and Management, Palwal, Haryana, India.

Ankush Goyal, Assistant Professor, Department of Computer Science and Engineering, Shri Ram College of Engineering and Management, Palwal, Haryana, India.

II. RELATED STUDY

[5] Keole.Ranjit.K et.el. analyse the problem of getting large amount of information for a user query from web search engines. Author concluded that it is very difficult for a user to find the exact information that satisfy his/her information need from this large information. Even performance of most of the page ranking algorithm is not good enough to solve the problem. Author suggest that clustering is a useful tool to solve this problem. Clustering divide [5] data objects into many groups based on their similarity. If these clusters are used to fire a query then the data retrieved by the search engines will be of good quality and relevant to the user information need. In this article [5] author explains how web mining techniques such as clustering can be useful in information retrieval systems. Author analyze many clustering algorithms such as K-Mean, CURE, ROCK, FUZZY clustering etc and their relative merits and demerits to cluster documents.

[6] R.Mahalaxmi et.el. gives a relative study on K-Mean, Suffix Tree and LINGO clustering algorithms. Author study these three algorithms to enhance the performance of web search engines. Author analyze and compare the performance of three algorithms on the basis of many parameters such as purity, normalized information, index and F measures. Further a table comparing various constraints for these three algorithms has been given which is shown in Table 1.

Table-1 Various constraint of Lingo, STC and K-Mean algorithms for keyword system and data source e-tools search. Author concluded that each algorithm has its own merits and demerits. Lingo gives high cluster diversity. In Lingo number of clusters produced are more as compared to the other STC and K-Mean algorithm. The scalability is high in STC as compared to Lingo and K-mean algorithm.

ISSN: 2278 – 7798

Table-1 Comparison of Lingo, STC and K-Mean algorithms

Parameters	LINGO	STC	K-MEANS
No Of Documents Retrieved	93	94	77
No. Of Clusters	20	15	7
% Of Overlapping	15.05	58	0
Avg Document Relevancy To Clusters	78.79	46.39	45.08
% Of Relevant Labels	75	58.33	50
Recall	85.37	79.27	79.41
Precision	75.27	79.79	70.13
purity	75.26882	0.740532	77.92208
Random index	70.34	73.78	70.52
F-Measure(3)	44.44444	44.18162	41.37931
F-Measure(5)	41.6	41.354	38.73103

[8] Poonam B.Lohiya give an idea of web clustering engines that organize search results by topics. Author highlight main characteristics of many clustering engines and compare their retrieval performance. To facilitate user in firing accurate query for his/her information need, search engines can group search results by topic. The user does not have to formulate his/her query but select a already existing query from a popup list. Thus user specify his/her information need accurately. This type of search engine has many important features such as fast subtopic retrieval, topic exploration and user can alleviate information overlook. Author proposed the architecture for such type of search engines which is shown in figure-1.

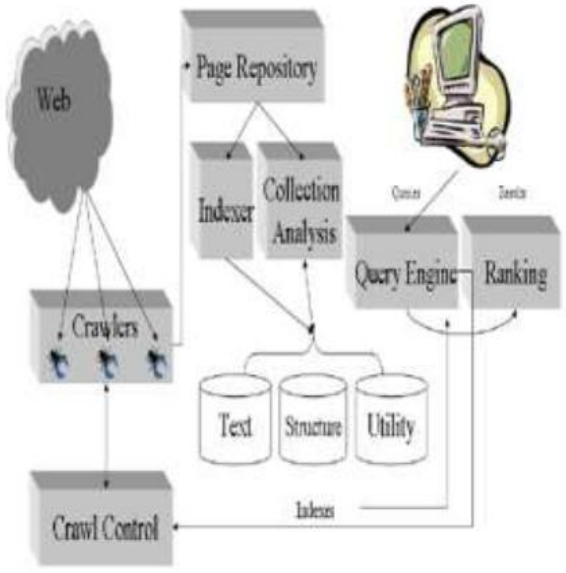

Figure-1 Architecture of web search engine with collection analysis.

Author concluded that users put short and ambiguous queries and thus get search results which are less relevant to his/her information need. Clustering techniques is a candidate solution to this problem. A future work is needed to cluster documents and assigning labels to clusters. The coherence of clusters is also an important issue in this field.

[7] Muhammad Adeel et.el. analyze and compare clustering techniques such as K-Mean, Agglomerative Hierarchical Clustering and Kohonen self organizing maps in clustering document to facilitate information retrieval systems. Author compare the three algorithms on the basis of many parameters. Author concluded that K-Mean K-mean is the fastest clustering technique and self organizing map is the slowest among all three techniques. [3] S.M.Jagatheesan et.el. developed a fuzzy based clustering algorithm for information retrieval system. Author concluded that FUZZY rule work in efficient manner on considering with the feature extraction based on the processing time and overlapping clusters. Experiment results shows that the cluster information obtained gives the frequency of word on the given dataset. Author suggest that this work can be extended to produce the fuzzy clustering on the basis of centrality measures.

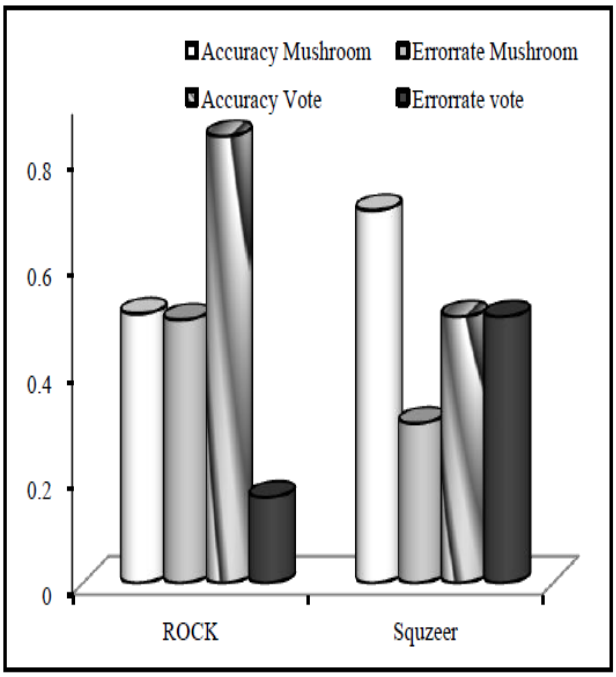

Figure-2 Comparison of clustering techniques ROCK and SQUEEZER for data set mushroom and vote.

[4] S.Anitha Elavarasi et.el. publish a survey on clustering algorithms and similarity measures for categorical data and compare almost ten different clustering algorithms. Author mainly compare accuracy and error rate for four clustering algorithms which are K-Mean, Fuzzy K-Modes, ROCK and Squeezer. Author mainly compare two algorithms ROCK and squeezer. Author take two data sets namely mushroom and vote. The comparison of performance of ROCK and SQUEEZER is given in Figure-2.

From the Figure-2 it is concluded that Squeezer works better for mushroom data set and ROCK work better for vote data set.

ISSN: 2278 – 7798

[1] Manjot Kaur et.el. analyze web document clustering approaches using K-Means algorithm. The performance og K-Mean algorithm depends upon choice of initial clustering centers which are chosen randomly most of the time. Authors propose a technique how to find better initial cancroids and to provide an efficient way of assigning initial data points to clusters that will reduce the time complexity. Author concluded that the proposed algorithm provide better results for various data sets. But the value of k; the number of clusters will be given as input value to K-Mean which is a problem area for using K-Mean algorithm. [2] Anoop jain et.el proposed a new clustering algorithm for information retrieval in data mining. The performance and scalability of proposed algorithm is more than the existing algorithms such as K-Mean.

III. CONCLUSION

After exploring the literature it is concluded that web search engines and information retrieval systems provide large volume of data for given user query. Finding useful information from this large data is again a problem for the user and attracts researchers to solve this problem. Data mining technique such as clustering is a useful technique to solve this problem. The performance of different clustering techniques such as K-Mean, ROCK, Squeezer, self organizing map is analyzed in many articles. In future work can be done in applying any of the clustering technique to facilitate the information retrieval systems so that user will get the desired information easily from these systems.

REFERENCES

[1] [1] Manjot kaur and Navjot Kaur, " Web Document clustering approaches using K-Mean algorithm". IJARCSSE, Volume 3, Issue 5, May 2013.

[2] [2] Anoop Jain, Aruna Bajpai and Manish Kumar Rohila, " Efficient Clustering Technique for Information Retrieval in Data Mining". International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 6, June 2012.

[3] [3] S.M. Jagatheesan and V. Thiagarasu, "Development of Fuzzy based categorical Text Clustering Algorithm for Information Retrieval". International Journal of Innovative Research in Computer, Vol. 2, Issue 1, January 2014.

[4] [4] S. Anitha Elavarasi and J. Akilandeswari , " SURVEY ON CLUSTERING ALGORITHM AND SIMILARITY MEASURE FOR CATEGORICAL DATA". ICTACT JOURNAL ON SOFT COMPUTING, JANUARY 2014, VOLUME: 04, ISSUE: 02.

[5] [5] Keole.Ranjit R and Dr.Karde.Pravin.P, "Information Retrieval From Web Document Using Clustering Techniques". International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 3 March 2013 Page No. 759-764.

[6] [6] R.Mahalakshmi, V.Lakshmi Praba, "A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO". International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-6, August 2013.

[7] [7] Muhammad Adeel et.el, "Efficient Cluster-Based Information Retrievalfrom Mathematical Markup Documents". World Applied Sciences Journal 17 (5): 611-616, 2012.

[8] [8] Poonam B.Lohiya, "A Survey On Web Search Result Clustering And Engines". International Journal of Science, Engineering and Technology Research (IJSETR)Volume 2, Issue 2, February 2013 ISSN: 2278 – 7798.

Teena Rani is currently perusing Master of Technology from Shri Ram College of engineering and management, Palwal, Haryana, India. She has completed B.Tech from lingayas university in INFORMATION TECHNOLOGY in the year 2009. Her research area included information retrieval, document clustering and web mining.

Ankush Goyal is currently working as an assistant professor in the department of computer science and engineering at Shri Ram college of engineering and management. He has guided many UG and PG students in their project and dissertation. His area of research include genetic algorithms, information retrieval and web mining.

ISSN: 2278 – 7798