Improving Relevancy of the Search Results of Information Retrieval Systems using ROCK Clustering Techniques

Teena Rani, Ankush Goyal

***Abstract*— World wide web a rich source of information. Search engines are basically information retrieval systems that make available the information on World Wide Web. However relevancy of the information provided by IR system is an area of research. Document clustering divides documents in to meaningful groups based on similarity between them. In this paper an attempt has been made to improve the relevancy of the results retrieved from IR system by using clustering techniques. Documents in the corpus of the IR system have been clustered using ROCK clustering algorithm. The results shows that relevancy of the output of the IR system with user query is more as compared to normal IR system.**

Index Terms- Document Clustering, ROCK, Information Retrieval, Search Engines

I. INTRODUCTION

Information retrieval systems provide information for a user information need based on user query. Search engines are the IR systems with millions of web documents. Retrieval of information from such a large IR system causes a lot of problems. Usually the information provide by large IR systems is not relevant or less relevant for a user query. Page ranking algorithms tries to overcome this problem by arranging the documents retrieved according to their relevancy with the user query Clustering is a method of grouping similar type of data together. Clustering represents the data in the compact form and provides homogeneity to it. Clustering has been and is a topic of active research because from last four decades we have storage of massive collection of data and it is still growing in numbers. Initially the data is stored in the data warehouse or database heterogeneously i.e. all the data which is related or unrelated. There is a need to group the related or similar data together and the unrelated or dissimilar data separately. So clustering is the solution to the above problem because clustering plays a very important role in data mining applications such as web analysis, customer relationship management, marketing, text mining, medical diagnostics and many more. Clustering in data mining involves various approaches and clustering algorithms which help in converting heterogeneous data objects into homogeneous form. As there are very large datasets with various attributes so it turns out to be a complication for clustering algorithms. Because clustering mechanism results in the approximate clusters which contains related objects. In this paper a new approach has been explained that will use the clustering technique to improve the relevancy of the results for a user query.

Manuscript received, May, 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. LITERATURE SURVEY

Muhammad and others [1] proposed a cluster based Information Retrieval from Mathematical Markup Documents. The author concluded that the cluster based approach take less time to retrieve documents and the relevancy of the search results is high in case of clustered based IR system. Anoop Jain [2] also proposed Descriptive k-Means technique in information retrieval systems. The author concluded that the technique may suffer problems in making cluster labels. So how to give labels to cluster is also an area of research in the field of applying clustering techniques. R. Mahalakshmi [3] published a related study of clustering techniques such as K-Mean, Suffix tree and Lingo. The author finds that the Suffix Tree method is a better technique to cluster documents as compared to the other two. Poonam B.Lohiya [4] published a survey on a survey on web search result clustering and engines. This paper compares various Web clustering engines and also discusses how to evaluate their retrieval performance. [5] Mr.Keole.Ranjit et.el. Proposed how clustering techniques can be used in IR systems. Author concluded that Clustering is useful technique in the field of textual data mining.

III. PROPOSED WORK

In this paper architecture of information retrieval system has been proposed based on cluttering techniques. The architecture of the proposed system is shown in figure-1. The components of the proposed system are as follows:

User Interface – This component will accept the query from the user. It forwards the user query to the retrieval system. It shows the results retrieved from the retrieval system, back to the user.

Retrieval System – This system is the main functioning components of the IR system. It receives a query from the user interface; pass the query to the matching system. It gets

ISSN: 2278 – 7798

the results from the matching system and returns the results back to the user interface. This system may contain many sub systems such as indexer, Page ranking module etc.

Matching System – This component receives the query from the retrieval system. It also either receives the cluster that user want to search or match the user query with the appropriate cluster to get the relevant results for the query. Then it matched the user query with the documents in one or more cluster and then returns these results to the retrieval system

Clustering System – This system take all the documents in the corpus as input, Apply ROCK clustering algorithm to cluster documents in to sub groups. The performance of the proposed IR system mainly depends on how well the clustering system make cluster of the documents.

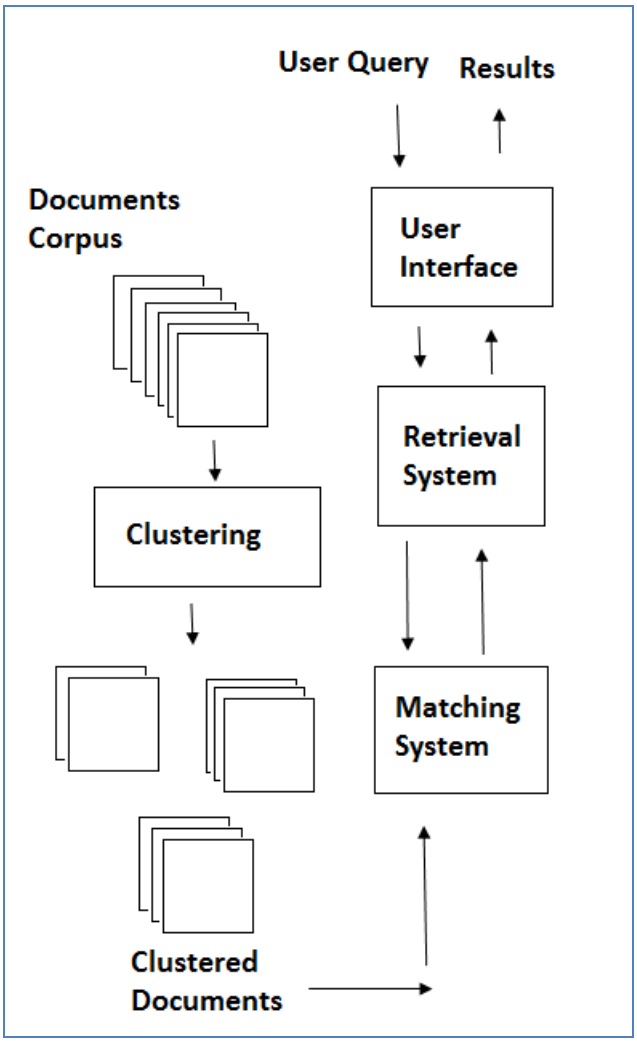

Figure-1 Proposed IR system using clustering technique.

IV. Results and Analysis

The system shown in Figure-2 has been implemented. The ROCK clustering algorithm has been used to cluster documents according to their similarity to each other. The dataset used in the implementation is as follows:

No of documents in the corpus = 15

No of clustered formed = 4

When a query is fired on the complete corpus without doing the clustering then the output is as follows:

Query 1 : Without Clustering

Let user want to search about cricket ball. But user has given the query "ball". Then user will get a list of 7 documents in which ball exists. But this list only contains 3 documents which are related to cricket ball. We compare performance of the proposed system by calculating the precision value.

Precision is a measure of ability of a system to present only relevant items.

Precision = (no of relevant items retrieved) / (total no of items retrieved)

In the case of query-1, user has received 7 documents out of which 3 documents are relevant. So the precision value of the system without clustering is as follows:

Precision = (3/7) = 42.8 %

Query 2 Without Clustering

Let user want to search about cricket rules. But user has given the query "rules". Then user will get a list of 4 documents in which ball exists. But this list only contains 2 documents which are related to cricket rules. We compare performance of the proposed system by calculating the precision value.

In the case of query-2, user has received 4 documents out of which 2 documents are relevant. So the precision value of the system without clustering is as follows:

Precision = (2/4) = 50 %

Then the same query has been fired on the IR system after applying the clustering. The results are as follows:

Query 1 : With Clustering

Let user want to search about cricket ball. User has given the query "ball" after selecting the cluster storing the documents related to cricket. Then user will get a list of 3 documents in which ball exists.

In the case of query-1, user has received 3 documents out of which 3 documents are relevant. So the precision value of the system without clustering is as follows:

Precision = (3/3) = 100 %

Query 2 : With Clustering

Let user want to search about cricket ball. User has given the query "rules" after selecting the cluster storing the documents related to cricket. Then user will get a list of 2 documents in which rules exists.

In the case of query-2, user has received 2 documents out of which 2 documents are relevant. So the precision value of the system without clustering is as follows:

Precision = (2/2) = 100 %

Table -1 is showing the precision based comparison of IR system with and without clustering.

ISSN: 2278 – 7798

Table-1 Precision based comparison of IR system with and without clustering.

Precision of IR System Without Clustering	Precision of IR System With Clustering
42.8	100

Figure-2 is showing a graph that is comparing the performance of the IR system with and with clustering. It has been observed that the relevancy of the clustering based IR system is almost double of the relevancy in case of IR system without clustering.

V. CONCLUSION AND FUTURE WORK

Clustering is useful technique to improve the performance of information retrieval systems. It reduces the no of documents which are searched for a user query. In this paper ROCK clustering technique is applied to cluster documents in an IR system. From the results it has been concluded that clustering techniques improves the performance of the IR system and it almost double the value of the precision of the IR system. In future the techniques can be applied on a larger data set having thousands of the documents.

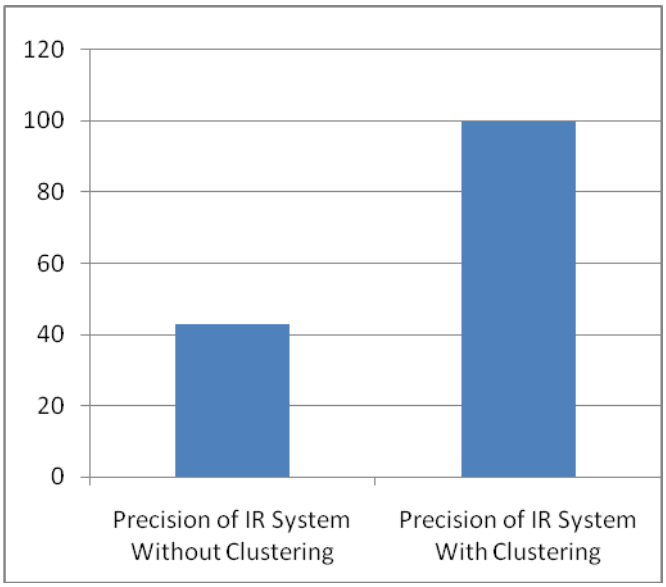

Figure-2 Precision based comparison of IR system with and without clustering.

REFERENCES

[1] Muhammad Adeel et.el. "Efficient Cluster-Based Information Retrieval from Mathematical Markup Documents", World Applied Sciences Journal 17 (5): 611-616, 2012 ISSN 1818-4952 © IDOSI Publications, 2012.

[2] Anoop Jain et.el. "Efficient Clustering Technique for Information Retrieval in Data Mining", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 6, June 2012).

[3] 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

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

[5] Mr.Keole.Ranjit et.el., "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

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