Efficient Page Ranking Using ROCK Clustering Algorithm

Neelam Gupta
M.Tech Student, KITE, Jaipur

Abstract - Today is a Data Deluge age. Therefore there is a huge need of mining the selected data from the data warehouses and the databases. It comprises of data which is subject oriented, integrated, time-variant and non volatile on which enormous number of queries can be requested by the users. There may be a practicability that the requested queries belong to same subject area. However, it is quite troublesome and time consuming to provide the response even in case of similar queries. Hence the solution to this problem is to make use of the concept of clustering the data for the purpose of identifying group of documents belonging to same subject areas. ROCK algorithm is the one of the best suited algorithm for clustering categorical data. In this paper ROCK clustering algorithm is being proposed to cluster page repository of search engine. So this paper approaches to cluster the data in such a way that query response time is decreased by searching the clusters obtained instead of whole database or data warehouse.

Keywords: Search Engine; Page Rank; Web Crawler, Rock Clustering.

I. INTRODUCTION

As the world is growing tremendously, so the data is growing in the databases. Today is a Data Deluge age. Data in the data warehouses is stored like wealth which is very precious to for its users. Therefore there is a huge need of mining the selected data from the data warehouses and the databases. Initially, with the advent of computers we started collecting massive information, but after that it created chaos in dealing with the stored data. Then data mining came into existence, it actually allows users to get only the important and relevant information or data from the large data warehouses. We call data mining as a process of discovering knowledge from the databases because these days it's very important to know about the recent trends and patterns going on in the market. Data mining [11] is done to search the valuable data form data warehouses just like mining the rocks to search the valuable ores. There may be different types of data that can be mined from the data warehouse, flat files, relational databases, transactional databases, multimedia databases, time series databases etc. There are different algorithms and techniques that are applied to retrieve the meaningful and relevant data. This helps in analyzing the data in various ways so that proper decisions can be made according to the prevailing trends in the market.

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. For example, if there is a large dataset and we want to have a look on the similar objects with compaction, then clustering is the well suited technique because it merges the related data together and categorizes the data into separate clusters. Each cluster has its own identity and some unique features are associated with each cluster which helps in differentiating one cluster with another. There are varieties of algorithms which meet the clustering requirements and are successfully applied to real-time datasets.

II. PROBLEM INTRODUCTION

Clustering is a data mining technique of grouping similar type of data or queries together which helps in identifying similar subject areas. The major problem is to identify heterogeneous subject areas where frequent queries are asked. There are number of agglomerative clustering algorithms which are used to cluster the data. The problem with these algorithms is that they make use of distance measures to calculate similarity. So the best suited algorithm for clustering the categorical data is Robust Clustering Using Links (ROCK) [1] algorithm as it does not use distance

measures instead it uses Jacquard coefficient to find the similarity between the data objects to classify the clusters. The mechanism to classify clusters based on the similarity measure shall be used over a given set of data. This method will make clusters of the data corresponding to different subject areas so that a prior knowledge about similarity can be maintained which in turn will help to discover accurate and consistent clusters and will reduce the query response time. The main objective of our work is to implement ROCK and to decrease the query response time by searching the documents in the resulted clusters instead of searching the whole database. This technique actually reduces the searching time of documents from the database.

2.1 Limitation of Traditional Clustering Algorithms

Many hierarchical clustering algorithms are not well suited when it comes to categorical datasets. Experiments showed that the distance measures cannot lead to high-quality clusters when clustering categorical data. Also, most clustering algorithms merge most similarity points in a single cluster at each step and this "localized" approach is prone to errors. Various centroid based hierarchical clustering algorithms merge the data points until a desired number of clusters are obtained, however distances between the centroids of different clusters are poor and they does not provide the good quality clusters. Initially the points are merged in different clusters but when the clustering process progress the situation gets worse because of the ripple effect (as the size of the cluster increases, the number of the attributes also increases but their mean value decreases. So it becomes quite tough to find the difference between the two points which differs in the context of attributes. So there are many clustering algorithms which cannot be used for clustering the categorical data, they are only meant to cluster the Euclidean space data. As there is enormous amount of categorical data present on World Wide Web, so it becomes mandatory to group the data so that it can be used in taking some decisions and user can only get the data he is interested in.

III. PROPOSED APPROACH

ROCK algorithm is the best suited algorithm for clustering categorical data because it may use Jacquard or Cosine similarity coefficients to find out the similarity between the two data points and moreover it uses the idea of links to determine the neighbours. It is difficult to manage and handle the large chunks of data; therefore clustering can help grouping them in order. What we have observed in general is that the task of finding or searching some document out of the large amount of data is cumbersome. Also, the response time of searching the document is very high due high scale searching among the data. So our approach is to cluster the data in order to divide the data into some groups with similar features and hence to decrease the query response time by searching the clusters obtained instead of whole database or data warehouse.

This project works in two steps:-

1. Clustering of data by ROCK algorithm and to store the clusters.
2. Reducing the query response or query search time by providing the results from the obtained clusters instead of the database.

3.1 Applying ROCK algorithm to the dataset

To implement the ROCK algorithm we have taken a categorical dataset i.e. some documents containing the data. We have taken different type of data with different attributes as it will check the functionality and applicability of this algorithm. This dataset contains documents related to various conferences and journals (national and international). As this data is heterogeneous in nature so we took this challenge to make some clusters out of it which will decrease the complexity of this dataset. We wanted to get the clusters of similar data. The aim was to decrease the intra-cluster similarity and to increase the inter-cluster similarity. Similarity between the two documents is calculated by using Jacquard coefficient. The similarity values lies in between 0 and 1. So we set the threshold value to 0.4 so that we can get the value of adjacency matrix. The values above 0.4 will be converted to 1 and the value below 0.4 will be converted to 0. This means the higher the similarity value, the more is the relation between the documents and lesser value of similarity denotes the some or no relation between the two documents. Jacquard coefficient finds out the similarity between each and every documents present in the dataset. The next step is to calculate the adjacency matrix that is by converting the lesser value to 0 and higher value to 1. After getting the adjacency matrix, our aim is to find the link matrix. Link matrix can be obtained by multiplying the adjacency matrix with itself. Clustering points based only on the closeness or similarity between them is not strong enough to distinguish two not so well-separated clusters because it is possible for points in different clusters to be neighbors. The link-based approach adopts a global approach to the clustering problem. It captures the global knowledge of neighboring data points into the relationship between individual pairs of points. Since the ROCK clustering algorithm utilizes the information about links between points when making decisions on the points to be merged into a single cluster, it is very robust. Goodness measure; while performing clustering the motive of using goodness measure is – to maximize the criterion function and to identify the best pair of clusters to be merged at each step of ROCK. This is an iterative step because

clusters are merged according to the goodness measure values. At every iteration the value of goodness measure increases and the more clusters are merged. ROCK works on agglomerative bottom up approach so there are major chances that only a single cluster is obtained in the end, so to avoid this consequence we have applied a threshold value, after which the merging process of clusters stops and the desired number of clusters can be obtained.

3.2 Proposed Algorithm:-

Algorithm: Create Initial Clusters

Input: Set of documents

Output: Initial Document Clustered

For all documents in the dataset do
{
Take a document from the document set.
List of tokens = Parse the document and apply normalization on words to get list of tokens
Create a keywords list of most frequent tokens of the document.
Create a cluster with that single document and create a keywords list of most frequent tokens of the document.
Add this initial cluster to the cluster repository.
}

Algorithm: Refine clusters using ROCK

Input: A set of clusters

Output: A set of refined clusters

For all clusters do
{
Cluster1 = Take a cluster from the Cluster Repository
For all documents in that cluster do
{
Docuemt1 = take a document from the cluster
Calculate the goodness measure of this document with all other clusters using ROCK
Find document cluster relation weight = goodness measure
If (document cluster relation weight > relation threshold)
{
Add the docmeunt1 in the Cluster1 with the assigned document cluster relation weight
Update the list of related words of Cluster1 by adding some keywords of document1 (if needed)
}
}

IV. CONCLUSION

In this paper we conclude that general users put short, ambiguous queries which can't specify the actual information need of the users. Clustering is the best possible solution for this problem, it facilitates quick browsing throughout the search result. To improve the search result clustering, First, more work needs to be done to improve the quality of the cluster labels and the coherence of the cluster structure. Second, the incrementality, because the web pages change very frequently and because new pages are always added to the web. Third, the fact that very often a web page relates to more than one subject should also be considered and lead to algorithms that allow for overlapping clusters. Fourth, Inconsistency is another problem. The contents of a cluster do not always correspond to the label and the navigation through the cluster sub hierarchies does not necessarily lead to more specific results. Fifth, advanced visualization techniques might be used to provide better overviews and guide the interaction with clustered results.

REFERENCES

[1] Oren Zamir and Oren Etzioni. Document Clustering: A Feasibility Demonstration.Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval, 1998, pp 46-54.

[2] C. Benincasa, A. Calden, E. Hanlon, M. Kindzerske, K. Law, E. Lam,J Rhoades, I. Roy,M. atz, E.Valentine and N. Whitaker, "Page Rank Algorithm"2006, http://www.math.umass.edu/~law/Research/PageRank/Google.pdf.

[3] Q. Tan, P. Mitra, C. Lee Giles, „Designing Clustering-Based Web Crawling Policies for Search Engine Crawlers□, Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, New York 2007, pp.535-544
[4] Sudipto Guho, Rajeev Rastogi, Kyuseok Shim, ROCK(clustering algorithm for categorical attribute)
[5] [ALSS95] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, and Kyuseok Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In Proc. of the VLDB Conference, Zurich, Switzerland, September 1995.
[6] [CLR90] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. Introduction to Algorithms. The MIT Press, Massachusetts, 1990.
[7] [CW87] Donald Coppersmith a d Shmuel Winograd. Matrix multiplication via arithmetic progressions. In Proc. of the 19th Annual ACM Symposium on Theory of Computing, 1987.
[8] [DH73] Richard O. Duda and Peter E. Hard. Pattern Classi_cation and Scene Analysis. A Wiley-Interscience Publication, New York, 1973.
[9] [EKSX96] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial database with noise. In Int'l Conference on Knowledge Discovery in Databases and Data Mining (KDD-96), Portland, Oregon, August 1996.
[10] [EKX95] Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. A database interface for clustering in large spatial databases. In Int'l Conference on Knowledge Discovery in Databases and Data Mining (KDD-95), Montreal, Canada, August 1995.
[11] [GRS98] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. CURE: A clustering algorithm for large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, May 1998.
[12] [HKKM97] Eui-Hong Han, George Karypis, Vipin Kumar, and Bamshad Mobasher. Clustering based on association rule hypergraphs. In 1997 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, June 1997.
[13] [JD88] Anil K. Jain and Richard C. Dubes. Algorithms for Clustering Data. Prentice Hall, Englewood Cli_s, New Jersey, 1988.