Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Optimizing Technique for Weighted Page Rank with K-Means Clustering

Supreet Kaur *,
*Student of M. Tech
Department of Computer Science Engineering,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India.

Usvir Kaur
Assistant Professor,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India.

Abstract— Clustering is an essential task in Data Mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. K-Means clustering is a clustering method in which the given data set is divided into K number of clusters. This paper is intended to give the introduction about K-means clustering and its algorithm. The experimental results of K-means clustering and its performance in case of execution time are discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution time we are using the weighted page rank with k means clustering And also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using the weighted page rank with k means clustering. K means with page rank algorithm gave results with better result set of various numbers of data-sets. In our case we are going to work on k means clustering of database with weighted page content rank algorithm

Index Terms—Clustering, K-means Clustering, Ranking method, Weighted page ranking, Execution Time.

I. INTRODUCTION[1]

In today's highly competitive business environment Clustering play an important role. As K- means Clustering is a method for making groups of the data set or the objects that are having similar properties. In this paper the II section includes the introduction part of Clustering and section III contain the related study or the literature survey about K-Means clustering algorithm with Ranking method. IV Section contains introduction about K-means Clustering algorithm and also examples of this algorithm. This Section also includes how in K-means algorithm the distance between the objects and mean is calculated and the methods of selecting initial points in K-means Clustering algorithm. Section V contains main steps in K-means clustering algorithm, then Section VI includes introduction about our proposed method Ranking Method and what is the need of Ranking method. Now section VII includes the introduction about the Methodology used for the implementation. Section VIII includes the results of both K-means clustering and Ranking Method and also shown the execution time taken by both algorithm during the clustering process With Result Set and Grpah.

II. CLUSTERING[3]

Mainly Clustering is the method which includes the grouping of similar type objects into one cluster and a cluster which includes the objects of data set is chosen in order to minimize some measure of dissimilarity. Clustering is a type of unsupervised learning not supervised learning like Classification. In clustering method, objects of the dataset are grouped into clusters, in such a way that groups are very different from each other and the objects in the same group or cluster are very similar to each other[2]. Unlike Classification, in which predefined set of classes are presented, but in Clustering there are no predefined set of classes which means that resulting clusters are not known before the execution of clustering algorithm. In this these clusters are extracted from the dataset by grouping the objects in it. Types of Clustering Algorithms[3]

a) Hierarchical Clustering Algorithm
b) K-means Clustering Algorithm
c) Density Based Clustering Algorithm
d) Self-organization maps (SOM)
e) EM clustering Algorithm

III. RANKING METHOD[4]

With regards to Clustering, ranking operations are a natural way to estimate the likelihood of the occurance of data items or the objects. So we propose evaluating ranking overall design of database for student data in order to form the clusters. So Ranking function introduce new opportunities to optimize the results of K-means clustering algorithm.

A. Need of Ranking Method

Search of relevant records or similar data search is a most popular function of database to obtain knowledge. There are certain similar records that we want to fall in one category or form one cluster. That`s why, we need to rank the more relevance student marks by a ranking method and to improve search effectiveness. In last, related answers will be returned for a given keyword query by the created index and better ranking strategy. So I have applied this Ranking method with K-means clustering method because this method is also having the property to find relevant records[5].So it is also helpful in creating clusters that are having similar properties between all data points within that cluster.

B.Weighted Page Ranking Method

Web mining technique provides the additional information through hyperlinks where different documents are connected. We can view the web as a directed labeled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph. There are
number of algorithms proposed based on link analysis. Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query[6]. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is? Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of in links and out links of the page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant.

IV. RELATED WORK[1]

The previous work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this researcher have also done analysis of K-means clustering algorithm by applying two methods, one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method applied on K-means algorithm and also compared the performance of both the methods by using graphs. The experimental results demonstrated that the proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm. But there is some sort of limitation in manner of timing sequences. So To make the algorithm time efficient we have purposed the an another algorithm which is based on k means clustering with Weighted page Ranking method.

V. PURPOSED WORK

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the relevant pages. In this paper we focused that by using Page Rank and Weighted Page Rank algorithms users may not get the required relevant documents easily, but in new algorithm Weighted Page Content Rank user can get relevant and important pages easily as it employs web structure mining and web content mining. The input parameters used in Page Rank are Back links, Weighted Page Rank uses Back links and Forward Links as Input Parameter and Weighted Page Content Rank uses Back links, Forward Link and Content as Input Parameters. As part of our future work, we are planning to carry out performance analysis of Weighted Page Content and working on finding required relevant and important pages more easily and fatly. Now to make the algorithm time efficient we have collaborate weighted page ranking with K means clustering technique and we have obtain better results as compared to previous one.

VI. METHODLOGY OF PURPOSED WORK

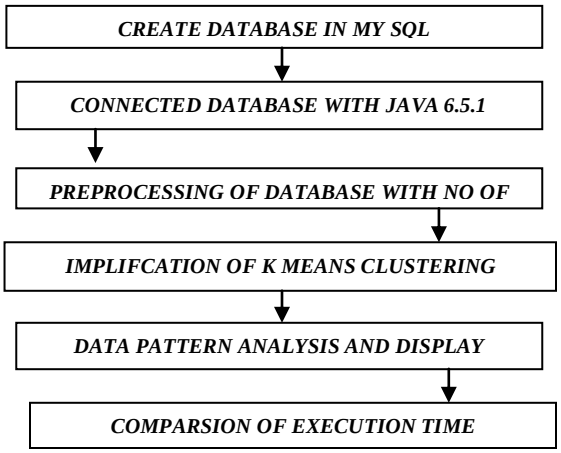

FIG. METHODOLGY USED IN PROJECT

I.) CREATE DATABASE IN MY SQL

At first end we have generated a database which include the number of records of sites which we want to cluster with records. Then we have incepted it with our data base.

ii.) CONNECTED DATABASE WITH JAVA 7.0

In the 2nd step we have used net beans software for carried out our mode of operation. we have done our working with java platform and SQL database is loaded in that and then worked down for the further operations.

iii.) PREPROCESSING OF DATABASE WITH NO OF NFORMATION RECORDS

In the third step we have carried out the preprocessing of records after loading then in main database. This Preprocessing is carried out with a clustering techniques which cluster the data base in sets of datasets which are easy for processing.

iv.) IMPLIFCATION OF K MEANS CLUSTERING TECHNIQIUE WITH WEIGHTED PAGE RANKING

Then in 4th step we have incorporated our purposed algorithm which is implementation of K means clustering algorithm with a weighted page ranking algorithm. First we implement the K means clustering algorithm in steps with Weighted page ranking algorithm in following sequential steps:

A.) STEPS OF K-MEANS CLUSTERING ALGORITHM

K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters, the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroid of each cluster is selected for clustering and then according to the chosen centriod, the data points having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroid. This algorithm consists of four steps:

1. Initialization -In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification -The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. Centroid Recalculation-Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. Convergence Condition -Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

B.) RANKING METHOD-Weighted Page Ranking Method

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is? Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of inlinks and outlinks of the page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant.

Algorithm: WPCR calculator

Input: Page P, Inlink and Outlink Weights of all backlinks of P, Query Q, d (damping factor).

Output: Rank score

Step 1: Relevance calculation:a) Find all meaningful word strings of Q (say N)

b) Find whether the N strings are occurring in P or not?

Z= Sum of frequencies of all N strings.

c) S= Set of the maximum possible strings occurring in P.

d) X= Sum of frequencies of strings in S.

e) Content Weight (CW)= X/Z

f) C= No. of query terms in P

g) D= No. of all query terms of Q while ignoring stop words.

h) Probability Weight (PW)= C/D

Step 2: Rank calculation: a) Find all back links of P (say set B).

b)PR(P)=(1-d)+d[

c) Output PR(P) i.e. the Rank score

V.) DATA PATTERN ANALYSIS AND DISPLAY

Then after the amplification of both algorithms on the data sets the data patterns are analyzed by the methods and results are presented in terms of time taken for records to be retrieved to the users. This method is to efficient in terms of time taken and comparison are done with the previous research's.

VI. Result and Comparsion of Execution Time

At last we have conclude that clusters are created in K-means clustering algorithm, using the concept of threshold value. Graph that is given below shows the number of clusters that are made on the basis of the threshold value. On the basis of the centroid the clusters are formed. This graph is made on the basis of the values x and y, which values are taken on the both axis of the graph. The Euclidean distance is calculated between both the centroid and the data points. Then weighted page ranking method is applied with this mechanism and total time taken to retrieve the records are calculated and comparisons are done.

VII. Result Parameters

We have taken three parameters to conclude our results and results sets are compared using a graph that presents the actual research work with in comparison to last one.

a. Precision: is the degree to which repeated measurements under unchanged conditions show the same results. What sort of results the algorithm retrieve are considered .This parameter consider with frequency of words in graph.

b. Recall: what sort of words are true relevant with our search and this parameter is calculated with number of words returned.

c. Time Taken to execution to retrieve Records: We have compared our result set of time taken to retrieve a records from the large database which is vast and scatted form. But we have taken a results set based on some sort of sites which are included in our database and comparison are done with previous research which was only based on K means clustering algorithm in retrieve records. The Result Set are shown in table below and compared using graph.

Table 1. Comparison of execution time

COMPARISON TABLE		
NUMBER OF LINKS	Time Taken In NS:Purposed method-Using weighted page rank with k means clustering algorithm	Time Taken In NS :Base Method-using Page Rank
10	7	9
20	16	19
30	24	27
40	30	42
50	37	48

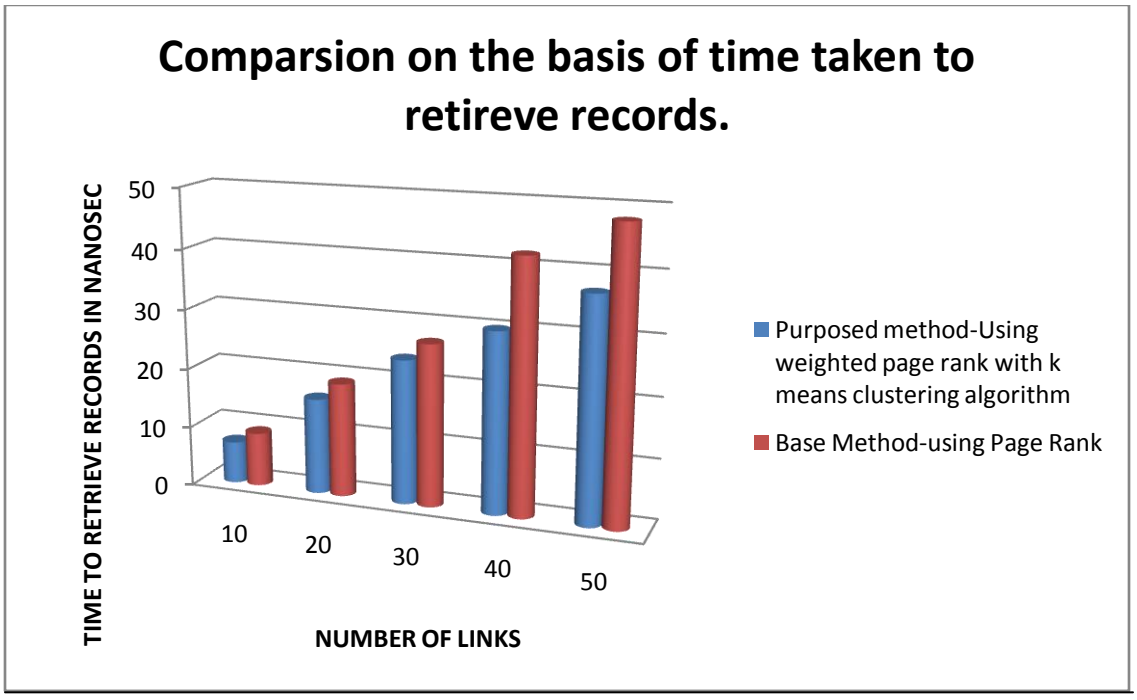

Figure: Graph Represents Result Set

VIII. Conclusion

K means with page rank algorithm gave results with better result set of various numbers of data-sets. In our case we have worked on k means clustering of database with weighted page content rank algorithm. The proposed work may benefit with

less computational time as compared to previous work as database with records are increasing day by day and there is a need of data clustering on large databases.

IX. Future Scope

Future scope shall be making the existing work more robust by doing research on various extraction algorithms as biomedical field is vast.

References

[1] Raymond Kosala and Hendrik Blockeel. "*Web Mining Research: A Survey*".

[2] Tamanna Bhatia, "*Link Analysis Algorithms For Web Mining*" ISSN :2229 - 423 (Print) |ISSN : 0976 - 8491 (Online) IJCST Vol. 2, Issue 2, June 2011.

[3] S. Brin, and L. Page, "*The Anatomy of a Large Scale Hypertextual Web Search Engine*", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[4] Wenpu Xing and Ali Ghorbani, "*Weighted PageRank Algorithm*", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[5] P Ravi Kumar, and Singh Ashutosh kumar, "*Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval*", American Journal of applied sciences, 7(6) 840-845 2010.

[6] Taher H. Haveliwala, "*Topic-Sensitive Page Rank: A Context-Sensitive Ranking Algorithms for Web Search*", IEEE transactions on Knowledge and Data Engineering Vol.15, No 4 July/August 2003.

[7] Ilsec. F. Ipsen and Steve Kirkland, "*Convergence Analysis of PageRank updating algorithm by Langville and Meyer*". Siam J. Matrixanal. Appl (Society for Industrial and Applied Mathematics) Vol. 27, No. 4, pp. 952–967 (2006).

[8] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur" *EFFICIENT K-MEANS CLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING*" *ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012*

[9] Ahamed Shafeeq B M 1 and Hareesha K S 2" *Dynamic Clustering of Data with Modified K-Means Algorithm* " *2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore*

[10] Pranit C. Patil1,Pramila M. Chawan,Prithviraj M. Chauhan3" *Extracting Information From Tables of HTML Document "* June 2012

[11] Seifedine Kadry and Ali Kalakech "*On the Improvement of Weighted Page Content Rank*" Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013

[12] Yin Ding"*Applying weighted PageRank to author citation networks" 1320 East 10th Street, Herman B Wells Library, LI025, Bloomington, IN 47405, USA*

[13] Danil Nemirovskya;b and Konstantin Avrachenkovb" *Weighted PageRank: cluster-related weights".*

