Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm

Amar Singh[1]
Shri Guru Granth Sahib World University, Fatehgarh Sahib, India.

Navjot Kaur[2]
Shri Guru Granth Sahib World University, Fatehgarh Sahib, India.

Abstract – The proposed work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this we have also done analysis of K-means clustering algorithm, one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method which is weighted page ranking applied on K-means algorithm, in weighted page rank algorithm mainly in links and out links are used and also compared the performance in terms of execution time of clustering. Proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

Keywords: - Data Mining; Clustering; K-means; Page rank; Weighted Page rank.

I. Introduction

CLUSTERING: It is "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In order to increase the efficiency in the database systems the number of disk accesses is to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available. Fig.1.1 shows the process of clustering of data.

Fig.1 Clustering Process

Example: In a library books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster.

K-MEANS CLUSTERING ALGORITHM: It is a well known partitioning method. In this, objects are classified as belonging to one of K-groups. The results of partitioning method are a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases. Example: A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset. K-means is a data mining algorithm which performs clustering of the data samples. As mentioned previously, clustering means the division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to cluster the database, K-means algorithm uses an iterative approach. The input in this case is the number of desired clusters and the initial means and also produces final means as output. These mentioned initial and final means are the means of clusters. If in the algorithm requirement is to produce K clusters then there will be K initial means and final means also [1].

After termination of this clustering algorithm, each object of dataset becomes a member of one cluster. The cluster is determined by searching throughout the means for the purpose to find the cluster having nearest mean to the object. Cluster with shortest distanced mean is cluster to which examined object belongs. In case of K-means algorithm, it tries to group the data items in dataset into desired number of clusters. To perform this task well it makes some iteration until some converges criteria meets. After each iteration , recently calculated means are updated such that they become closer to the final means. And at final, the algorithm converges and then stops performing iterations.

This algorithm consists of four steps:

A. *Initialization*: In this first step data set, number of clusters and the centroid that we defined for each cluster.
B. *Classification*: The distance is calculated for each data point from the centroid and the data point having minimum distance from the centroid of a cluster is assigned to that particular cluster.
C. *Centroid Recalculation*: Clusters generated previously, the centroid is again repeatly calculated means recalculation of the centroid.
D. *Convergence Condition*: Some convergence conditions are given as below:
   - Stopping when reaching a given or defined number of iterations
   - Stopping when there is no exchange of data points between the clusters
   - Stopping when a threshold value is achieved
E. 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.

WEIGHTED PAGE RANK ALGORITHM: This algorithm is an extension of Page Rank algorithm. WPR takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional Page Rank algorithm in terms of returning larger numbers of relevant pages to a given query. According to author the more popular web pages are the more linkages that other WebPages tend to have to them or are linked to by them. The proposed extended Page Rank algorithm–Weighted Page Rank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as Win(v,u) and Out(v,u), respectively. WPR supplies the most important web pages or information in front of users.

II. Related 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 algorithms users may not get the required relevant documents easily, but in new algorithm Weighted Page Rank user can get relevant and important pages easily as it employs web structure mining and web content mining. The input parameters used in Weighted Page Rank uses Back links and Forward Links as Input Parameter. 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 fastly.

III. Purposed Work

The proposed work represents ranking based method that improved K-means clustering algorithm performance and execution time. In this we have also done analysis of K-means clustering algorithm , one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method which is weighted page ranking applied on K-means algorithm and also compared the performance in terms of execution time of clustering. Proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

Weighted page rank algorithm:

A. This algorithm is an extension of Page Rank algorithm.
B. WPR takes into account the importance of both the in links and the out links of the page .
C. The extended Page Rank algorithm–Weighted Page Rank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links).
D. The popularity from the number of in links and out links is recorded as Win(v,u) and Wout (v,u), respectively. WPR supplies the most important web pages or information in front of users and distributes rank scores based on the popularity of the pages.

METHODOLOGY OF RESEARCH WORK: In this Research Work, we have used weighted page rank algorithm for ranking and for calculating executing time of the data sets.

Weighted Page Rank Algorithm is used in K-Mean clustering. Firstly we will discuss about K-Means:

Input: Link whose rank and execution time has to be calculated.

Output: Rank and execution time of entered link.

Method:

A. First read all the links one by one from the database using SQL query
B. Then extract the every link by using java http library to find the in links and out links.
C. Then apply the weighted page rank algorithm to computing the in links and out links of the every website link for calculate the rank of the website as show in fig.3.

D. Then apply the k-means algorithm to assign the clusters value to store in database.

In the methodology flowchart of K-means as shown in fig.2, our work is on 6th step i.e. where ranking method is applied. In the earlier used search engines page rank algorithm is used which is based on links, but here we used weighted page rank algorithm which basically based on the in links and out links of that webpage which results to more relevant and time consuming.

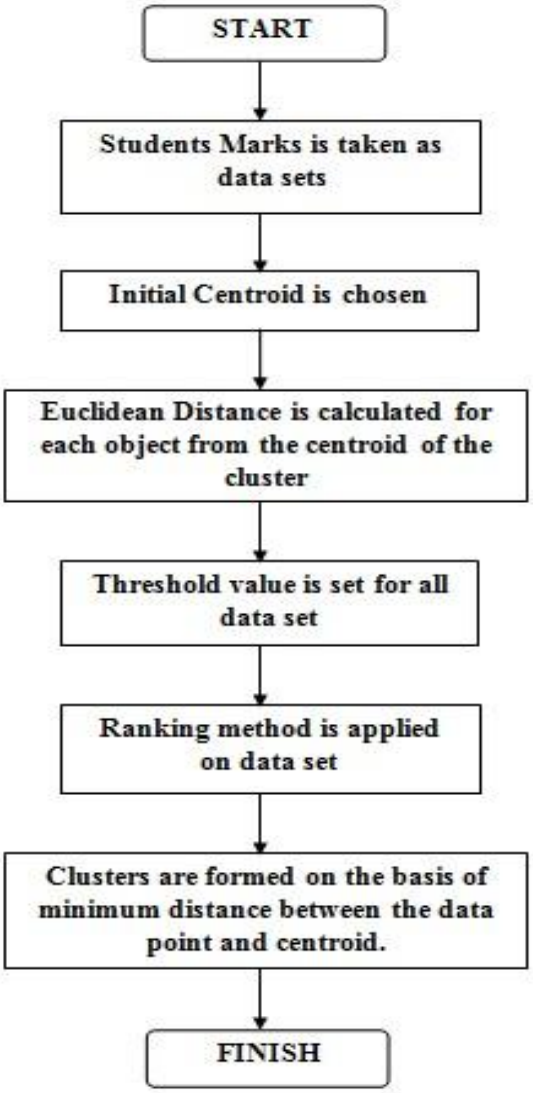

Fig.2 Methodology of K-Means algorithm

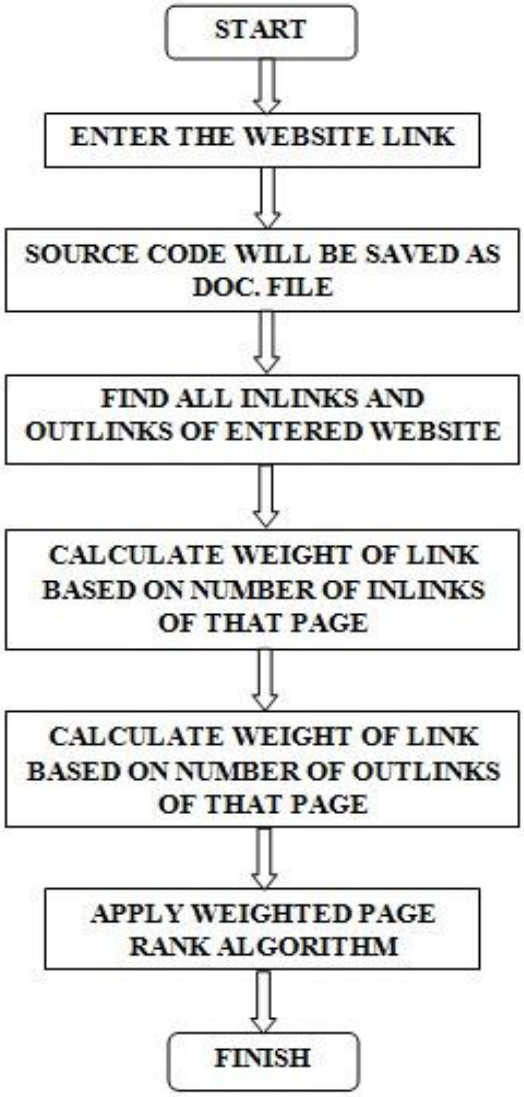

Fig.3 Methodology of WPR algorithm

STEPS OF ALGORITHM USED

A. The popularity from the number of inlinks and outlinks is recorded as Win(v,u) and Wout(v,u) , respectively.

B. Win(v,u) is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.

C. Where Iu and Ip represent the number of inlinks of page u and page p, respectively. R (v) denotes the reference page list of page v.

D. $Win(v,u) = \frac{Iu}{\sum_{p \epsilon R(v)} Ip}$

E. Wout(v,u) is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.

F. Where Ou and Op represent the number of outlinks of page u and page p, respectively. R (v) denotes the reference page list of page v.

G. $Wout(v,u) = \frac{Ou}{\sum_{p \epsilon R(v)} Op}$

H. Then apply the WPR algo. For calculating rank[23] i.e. $WPR(u) = (1-d) + d \sum_{v \epsilon B(u)} PR(v) Win(v,u) Wout(v,u)$

IV. Results and Comparison

On the basis of result sets of previous and proposed work

PREVIOUS PROJECT BASED RESULTS

Table I
Results of Page Rank

ID	LINK	PAGE RANK	TIME(ms)
1	http://www.gmail.com	91.95	150158
2	http://www.facebook.com	17.15	185204
3	http://www.w3schools.com	0.15000000000000002	203983
4	http://www.mysmartprice.com/	74.95	114760
5	http://www.funbrain.com/	28.2	173557
6	http://www.nokia.com	22.249999999999996	224739
7	http://www.w3newspapers.com	112.350000000000001	381070
8	http://www.chem4kids.com/	44.349999999999994	152819
9	http://www.wordpress.com	15.45	147913
10	http://www.pinterest.com	33.3	130607

PROPOSED PROJECT BASED RESULTS

Table II
Results of WPR

ID	LINK	WInlinks	WOutlinks	WPR	TIME(ms)
1	http://www.gmail.com	64	11	8.492238868081564	31290
2	http://www.facebook.com	115	43	84.48333333333333	75154
3	http://www.w3schools.com	265	126	122.27976878612716	108655
4	http://www.mysmartprice.com/	178	14	31.185897435897434	4590
5	http://www.funbrain.com/	103	57	58.19945652173914	43063
6	http://www.nokia.com	50	64	1.8200435930496717	96803
7	http://www.w3newspapers.com	29	320	2.498842403037296	255589
8	http://www.chem4kids.com/	241	37	79.18661417322835	29201
9	http://www.wordpress.com	173	14	31.279536290322582	30297
10	http://www.pinterest.com	98	6	90.0263157894737	19149

After calculating the results will automatically be saved in the database. Table I. shows the results taken from Page Rank algorithm by entering the URL's and Table II. Shows the results taken from the Weighted Page Rank algorithm by entering the same URL's which are used in calculating the execution time and rank of the page in Page Rank algorithm.

V. DISCUSSION

Here are the results of the previous used algorithm and our proposed algorithm. In the results we can see that the execution time for our proposed algorithm is less than the execution time of the previous algorithm used. In the graph shown in fig.4, we compare the results taken from the previous work done and our present proposed work where (1,2,3……,10) are the ID's of the entered web links and accordingly the execution time (ms) for calculating ranks are shown:

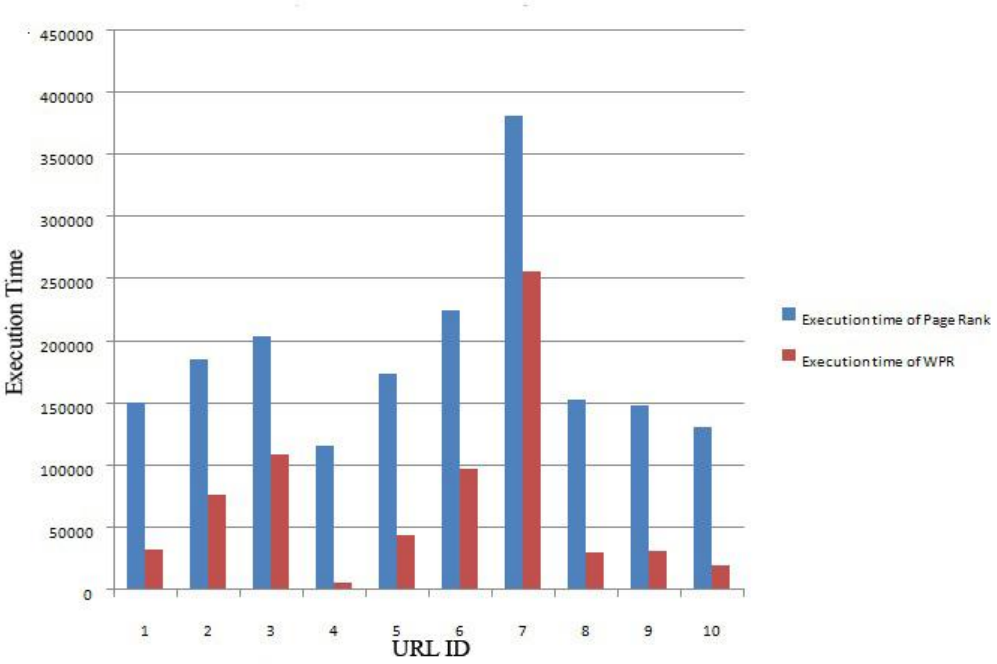

Fig.4 Graph of results comparison

From the result sets taken, we conclude that the execution time for the weighted page rank (proposed work) for calculating the rank of the page is less than the execution time for the page rank (previous work).

VI. Conclusion

In this work, we presented a cluster based ranking scheme i.e. K-means clustering over weighted page rank algorithm. Weighted Page Rank algorithm is the modified form of page rank algorithm which works on two factors, in links and out links and gives the ranks to the websites accordingly. In the previous work done, page rank algorithm is used which results to take more execution time. But we use weighted page rank algorithm with k-means which results to take less execution time as compare to the previous work done and also leads to more relevant as compared to page rank algorithm. But our project works on limited websites and cannot work on the websites which are secured will SSL layer which are not permitted to access by unknown user.

VII. Future Work

- Our project works on limited websites which are permitted to access and whose ports are open and are not secured with SSL layer, so work can be done on that websites also by taking permission.
- Can help to improve the relevancy and execution time of the search engine by using this proposed algorithm in it.

References

[1] Navjot Kaur, Jaspreet Kaur, 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.

[2] Daniel P. Huynh, "*EXPLORING AND VALIDATING DATA MINING ALGORITHMS FOR USE IN DATA ASCRIPTION",june 2008*

[3] Ahamed Shafeeq B M and Hareesha K S, *"Dynamic Clustering of Data with Modified K-means Algorithm"*, 2012 International Conference on Information and Computer Networks (ICICN 2012) ,IPCSIT vol 27(2012).

[4] Kalyani M Raval, *"Data Mining Techniques",* International journal of Advanced Research in Computer Science and Software Engineering, volume 2,Issue 10, October 2012.

[5] Khaled Alsabti,Sanjay Ranka and Vineet Singh, "*An Efficient K-Means Clustering Algorithm",*1997.

[6] Jeffrey W. Seifert, " *Data Mining: An Overview",* Congressional Research Service ˜ The Library of Congress.

[7] Rui Xu and Donald Wunsch, "*Survey of Clustering Algorithms",* IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

[8] M.Lavanya and Dr.M.usha Rani, ***"****Vision-Based Deep Web Data Extraction for Web Document Clustering",* Volume 12 Issue 5 Version 1.0 March 2012.

[9] Pranit C.Patil , Prithviraj M.Chauhan and Pramila M.Chawan, "*Extracting Information From Tables of HTML",* Volume 1, No. 4, June 2012 ISSN – 2278-1080

[10] Keerthiram Murugesan and Dr.Jun Zhang, "*CLUSTER BASED TERM WEIGHTING AND DOCUMENT RANKING MODEL",* Copyright© Keerthiram Murugesan 2011

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

[12] Raza Ali , Usman Ghani, Aasim Saeed**,** "*Data Clustering and its Applications"***.**

[13] Lawrence Page, Sergey Brin, Rajeev Motwani, and TerryWinograd, "*The PageRank Citation Ranking: Bringing Order to the Web (1998)*".

[14] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning and Gene H. Golub. "*Extrapolation Methods for Accelerating Page Rank Computations* ",2003.

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

[16] 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.

