ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Implementation of QROCK Algorithm for Efficient Clustering in Library Reader Management System

Achal R. Puranik[1], Vipin M.V.[2], Santosh S Patil[3],Valarmathi B.[4]

Software Engineer, Accenture Services Pvt Ltd, Bengaluru,Karnataka, India[1]

Software Engineer, IBM India Pvt Ltd, Bengaluru, Karnataka, India[2]

Assistant Professor, School of Engineering, CUK Kalaburagi, Karnataka, India[3]

Assistant Professor, SITE, VIT Vellore, Tamil Nadu, India[4]

ABSTRACT: The data mining techniques are employed in library reader management system for providing personalized services to readers. The data used in mining for this purpose mostly will be of categorical in nature. The study was undertaken to discover an efficient categorical clustering algorithm which could replace the k-means numerical clustering method in the current mining system.

QROCK or Quick RObust Clustering using linKs [2] algorithm was found to be most suitable for categorical clustering because of its run time efficiency and simplicity. It is an agglomerative hierarchical clustering method which computes clusters by determining number of connected components in a graph and could replace many overheads in k-means clustering[8]. Categorical data itself can be given as the input which could eliminate categorical-numerical conversion for its kind. Algorithm can produce desired number of clusters unlike k-means and input overheads in the form of k-value given by the user prior to its execution have also been eliminated. This study identified that QROCK algorithm can be used for categorical clustering in library reader management system efficiently.

KEYWORDS: Categorical data, Agglomerative hierarchical clustering, Numerical clustering.

I.INTRODUCTION

Clustering of data in reader management system helps in providing services to the readers. Readers can be classified based on their reading behaviours which help them to get necessary information on books and other suggestions based on their interested domains. The datasets in the system can contain both numerical and categorical data having later in plenty. So a categorical clustering rather than numerical clustering algorithm [8] will perform well in producing clusters in such a system.

In this paper we point to QROCK algorithm which can be efficiently used to drive better results in clustering categorical data. Algorithm forms connected components of a graph based on the input data and determines the number of clusters. Initially each user is considered as an individual cluster. User attributes are chosen before clustering and similarities are measured to determine the neighbours. The neighbours who share common characteristics are grouped together to form a single cluster if they do not belong to the same. The process is carried out till no neighbours exist for any of the newly formed clusters. Various clusters formed so reflects groups of readers who share common behaviours in their attributes.

II.RELATED WORK

Clustering has been extensively studied by researchers in psychology, statistics, biology and so on. Surveys of clustering algorithms can be found in many areas. Clustering algorithms for mining large databases have been proposed but however, most of these are variants of either partitional or centroid-based hierarchical clustering. As a result,these algorithms are more suitable for clustering numeric data rather than data sets with categorical attributes. For instance,

DOI: 10.15680/IJIRCCE.2015. 0311075

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

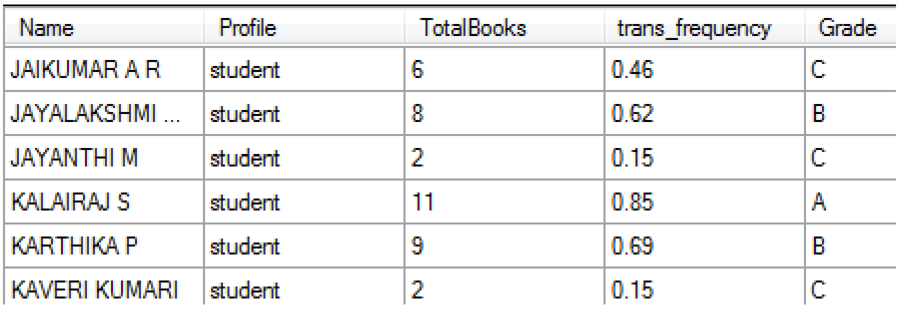
Vol. 3, Issue 11, November 2015

CLARANS employs a randomized search to find the k best cluster medoids. BIRCH, proposed first preclusters data and then uses a centroid-based hierarchical algorithm to cluster the partial clusters. The CURE algorithm uses a combination of random sampling and partition clustering to handle large databases.DBSCAN, a density-based algorithm, grows clusters by including the dense neighborhoods of points already in the cluster. This approach, however, may be prone to errors if clusters are not well-separated.

III.SYSTEM MODEL AND ASSUMPTIONS

Sample Data: Sample dataset was prepared with all the necessary information with library database as a reference. User details along with book transaction details were added to form the sample dataset.
Sample dataset was selected in such a way that clustering can be done on two attributes like designation & transactions.

Procedure: First step is to select the data and prepare it for clustering. For that, attributes of user like designation and the number of transactions done by each person were taken from the respective tables. Frequency of transactions by each user was determined by relative grading where the highest value in the frequency column was taken as the reference value. Now the number of transactions done by each user was divided with this reference value to obtain the transactional frequency .certain ranges were specified to grade the obtained frequencies, like if the frequency was above 0.6 then that user was graded 'A'. Likewise three grades A, B, C was chosen and users were put in to respective grades (Table 1).



Name	Profile	TotalBooks	trans_frequency	Grade
JAIKUMAR A R	student	6	0.46	C
JAYALAKSHMI ...	student	8	0.62	B
JAYANTHI M	student	2	0.15	C
KALAIRAJ S	student	11	0.85	A
KARTHIKA P	student	9	0.69	B
KAVERI KUMARI	student	2	0.15	C

Table 1. Selected data with attributes for clustering

Data similarities among the users were measured with the help of weighted similarity measure.
Weighted similarity measure is given by,

$$sim(T_i,T_j) = \frac{|T_i \cap T_j|}{|T_i \cap T_j| + 2 * \sum_{k \notin T_i \cap T_j} \frac{1}{D_k}}$$

Where, $T_i \cap T_j$ represent the number of similar attributes shared by two users.
D_k Represent the weight that is to be considered for each dissimilar value.
Weights can be assigned based on the priority of each attribute. Here grade attribute was given high priority.
If all the attributes are similar then the similarity is given by

$$sim(T_i,T_j) = \frac{|T_i \cap T_j|}{|T_i \cap T_j|}$$

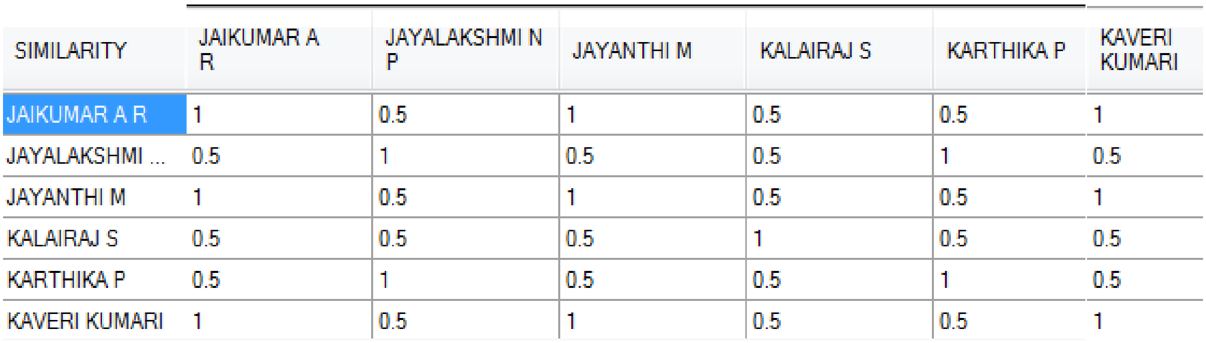
This value is taken for every set of users. One time threshold value (θ) is chosen at random (here it is 0.5) and if the weighted similarity measure for any set of users is found to be higher than the chosen threshold value then those two users are considered to be neighbours (Table 2).
A new table called neighbouring table (Table 3) was considered and the value for each cell was kept 1 wherever the users were found to be neighbours and others as 0.
After finding the neighbours, QROCK algorithm was applied to the neighbouring table to form clusters of similar characteristics.

The following are the primitives of the abstract data types used in the QROCK algorithm.
1. *initial(x)*: Creates a component that contains only x.

DOI: 10.15680/IJIRCCE.2015. 0311075

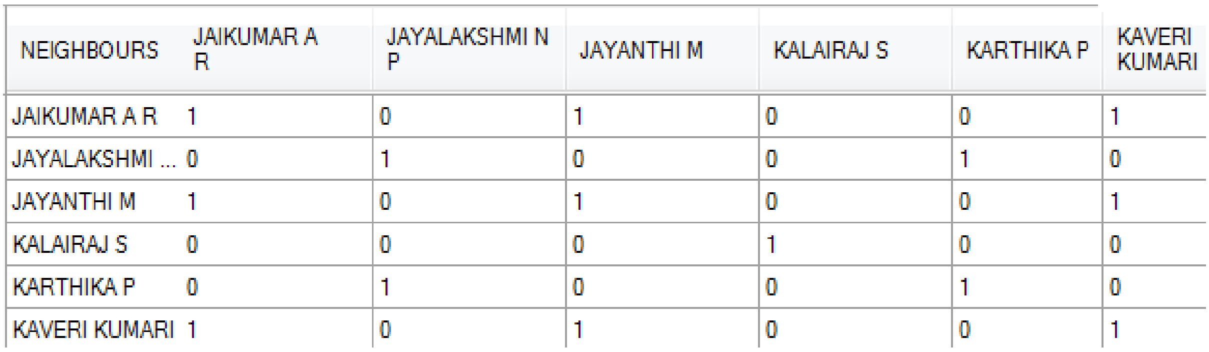
2.*find(x)*: returns the component of which x is a neighbour.

3. *merge(A,B):* takes the union of components A&B



SIMILARITY	JAIKUMAR A R	JAYALAKSHMI N P	JAYANTHI M	KALAIRAJ S	KARTHIKA P	KAVERI KUMARI
JAIKUMAR A R	1	0.5	1	0.5	0.5	1
JAYALAKSHMI ...	0.5	1	0.5	0.5	1	0.5
JAYANTHI M	1	0.5	1	0.5	0.5	1
KALAIRAJ S	0.5	0.5	0.5	1	0.5	0.5
KARTHIKA P	0.5	1	0.5	0.5	1	0.5
KAVERI KUMARI	1	0.5	1	0.5	0.5	1

Table 2.Weighted similarity measure



NEIGHBOURS	JAIKUMAR A R	JAYALAKSHMI N P	JAYANTHI M	KALAIRAJ S	KARTHIKA P	KAVERI KUMARI
JAIKUMAR A R	1	0	1	0	0	1
JAYALAKSHMI ...	0	1	0	0	1	0
JAYANTHI M	1	0	1	0	0	1
KALAIRAJ S	0	0	0	1	0	0
KARTHIKA P	0	1	0	0	1	0
KAVERI KUMARI	1	0	1	0	0	1

Table 3.List of neighbouring nodes

QROCK Algorithm

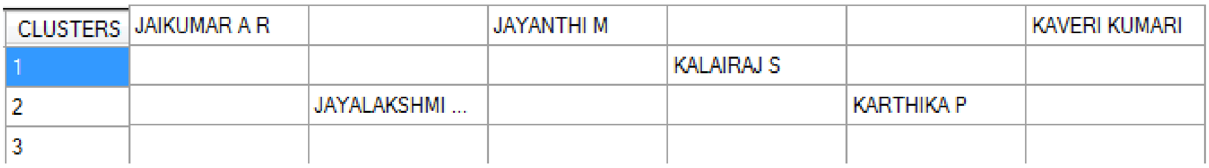
Input: A set D of data points

```
Begin
Compute neibr[i] for each iє D using θ
foreach x in D initial(x)
foreach i in D
{
    Take a fixed point x in neigr[i]
    foreach other y in neigr[i]
    {
        A = find(x)
        B = find(y)
        if A ¹ B
        thenmerge(A,B)
    }
}
End
```

Clustering efficiency is determined by the threshold value. If the value is higher we obtain tight clusters and obtain loose if vice versa

IV. RESULT AND DISCUSSION

In this paper, we proposed clustering of data for categorical attributes. The results were obtained after executing the QROCK algorithm on neighbouring table. A total of three clusters were formed which reflected clusters mainly based on 'grade' attributes (Table 4).



CLUSTERS	JAIKUMAR A R		JAYANTHI M			KAVERI KUMARI
1				KALAIRAJ S		
2		JAYALAKSHMI ...			KARTHIKA P	
3						

Table 4.Final clusters

For accuracy measure a dataset of 100 users along with their transaction details were chosen and the results were obtained and analysed. The datasets were of categorical in nature. The results obtained 3 clusters. Graph shows the distribution of data over three clusters (Figure 1).

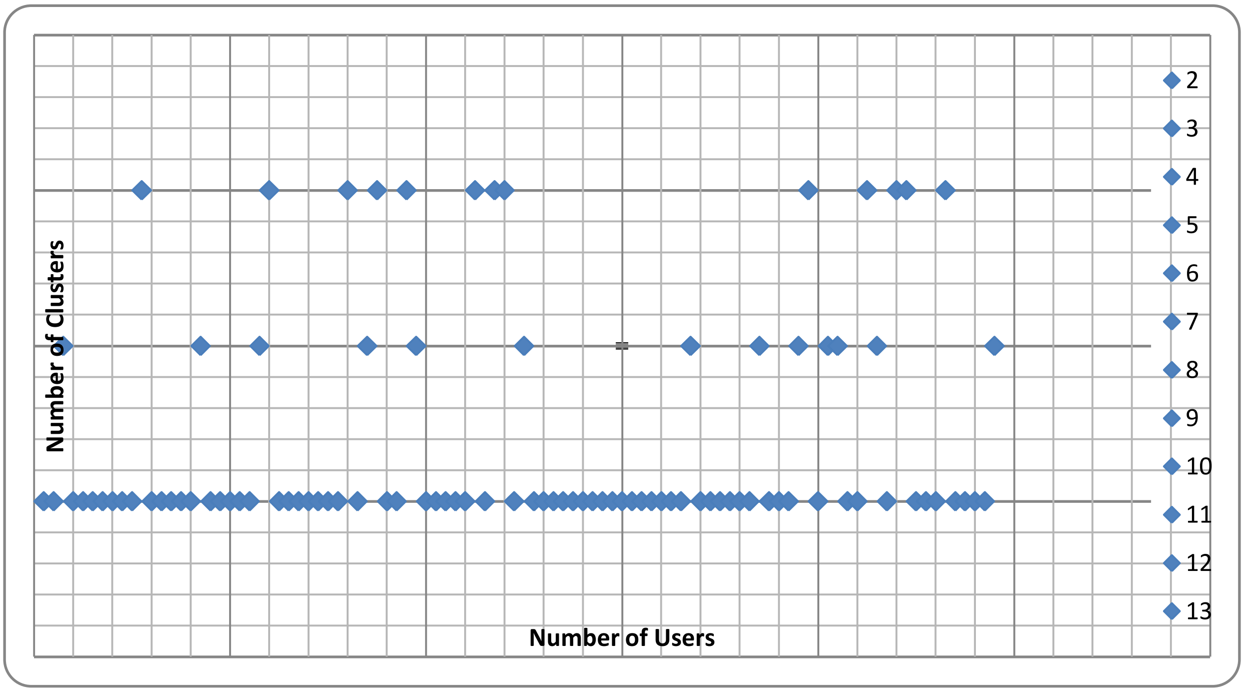


Fig.1. Distribution of clusters with QROCK algorithm

The results showed that distributed and more specified clusters could be formed in case of QROCK in categorical clustering. This point out that library information that deals mainly with categorical data can be clustered in a more efficient way with the help of QROCK algorithm.

V.CONCLUSION

This study has shown that QROCK algorithm can be chosen for the easiest clustering of categorical data. No extra computation or information is required for the execution other than the threshold value (θ), which decides the cohesion rate inside the clusters. Priority based clustering has also been proved effective in this algorithm which helps in making clusters focusing on certain attributes by giving them high weighted priority. The overall advantages and the simplicity of the algorithm have already been specified. This finding will help you in clustering categorical data in other fields as well.

REFERENCES

[1] Aravind H, C Rajgopal, K P Soman, Simple Approach to Clustering in Excel, International Journal of Computer Applications (0975 – 8887), Volume 11– No.7, December 2010

[2] K. M. Passino, M. Dutta, A. KakotiMahanta, Arun K. Pujari, QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data, http://202.41.85.117/~akpcs/akpcs/qrock.pdf.

[3] Nicholson, S. The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making. Information Technology and Libraries. 2003, 22(4):146-151

[4] Ping YU, Data Mining in Library Reader Management, 2011 International Conference on Network Computing and Information SecurityH.

[5] R. A. JARVIS AND EDWARD A. PATRICK, Clustering Using a Similarity Measure Basedon Shared Near Neighbors, IEEE TRANSACTIONS ON COMPUTERS, VOL. C-22, NO. 11, NOVEMBER 1973.

[6] S.Aranganayagi and K.Thangavel, Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure, International Journal of Information and Mathematical Sciences 6:1 2010

[7] Sudiptoguha, Rajeev Rastogi, Kyuseok Shim, ROCK : A ROBUST CLUSTERING ALGORITHM FOR CATEGORICAL DATA. Information Systems, VOL 25, no.5, pp: 345-366

[8] Teknomo, Kardi. K-Means Clustering Tutorials. http://people.revoledu.com/kardi/ tutorial/kMean/Resources.htm.

[9] Tian Zhang, Raghu Ramakrishnan, and MironLivny. Birch: An efficient data clustering method for very large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 103–114, Montreal, Canada, June 1996.