Implementation of K-MEAN Partitioning Method in Educated Collective System

Thida Htun
Computer University, Monywa, Myanmar
thidahtun1@gmail.com

ABSTRACT

Clustering is a highly popular and widely used tool for identifying or constructing data based market pieces. In recent years, there have been growing interests in developing effective methods for searching large database based on multi-dimensional contents. Consequently, this paper intends to build up a Public Figures Civilizing (PFC) system with a common partitioning method, k-means. Most of universities make all of their duties with manual system. So this system implement computerize system for them and especially for issuing exam result. Furthermore, this information should be made accessible from consistent provision, because each developing country should promote the social standard with knowledgeable individual. For this reason, the proposed system can support the statistics of the information with two components: Client Visible Frame (CVF) module and K-means Clustering Execution (KCE) module. This paper temporarily describes the chief occupation of these two modules as a reasonable behavior.

1. INTRODUCTION

Clustering is a challenging field of research where its potential applications pose their own special requirements. Thus, cluster analysis is an imperative being activity and has been generally used in frequent applications, including pattern recognition, data analysis, image processing, and market research. As a result of clustering, one can classify dense and sparse sections and, therefore, notice overall distribution patterns and interesting correlations among data attributes. As a data mining purpose, cluster analysis can be used as a stand-alone tool to obtain close into the distribution of data, to monitor the characteristics of each cluster, and to focus on an exacting set of clusters for extra investigation.

Various clustering algorithms have been developed and are useful in many application areas. Agglomerative hierarchical clustering and partitioning based k-means clustering techniques are commonly used in many areas. Although the agglomerative clustering technique is superior to k-means clustering, the work of Steinbach et al. [2] showed that result of bisecting k-means algorithms is better than the previous two.

In our proposed system, the k-means algorithm is participated in the promotion of verifying factor. The input parameter, *k*, and partitions a set of *n* objects into *k* clusters with the intention that the resultant intracluster similarity is high but the intercluster similarity is low. Cluster dissimilarity is considered in hold to the mean value of the objects in a cluster. The social standard is a measure into the enlarged community. As a consequence, upgrading the abilities of individual is also provided by an education scheme. The level of degree in information technology is also important in those neighborhoods.

2. RELATED WORK

A cluster of data objects could be pleasured together as single group in many applications. Z. He et al. presented that the trial studies on concerning a new dissimilarity measure to the k-

modes clustering to recover the accuracy of the cluster results. It is initiated that the clustering results produced by the customized k-modes algorithm are very high in accuracy, contrasted to its original one.

There have been a lot of image retrieval systems found on image content[8]. The "Blobworld" system presented by Serge, et.al., is a representative one in CBIR systems[7]. It supports object-based queries and thus eliminates the need for the manual indexing of image content. The images in the system include the feature information represented by color and texture based on image segmentation.

Information retrieval systems attempt to achieve high performance by constructing conceptual structures to use as background knowledge [4]. The work of Sieg et. al. [1] uses concept hierarchies of a specific domain and user profile information to enhance user query by adding terms from these sources. According to their approach, a user's initial query is modified based on the user's interaction with a modular concept hierarchy.

E-H Han et al. [3] proposed an agent for exploring and categorizing documents on the Web. Their agent categorize a set of documents automatically and combine with new queries generation process and document filtering process to get most closely related documents to the initial query set. A comparison of document clustering techniques was discussed in [6]. It compared k-means and agglomerative hierarchical clustering algorithms and also showed better performance of bisecting k-means algorithm over the previous two.

3. SYSTEM DESIGN AND K_MEAN PARTITIONING

In the present day, there is an important role for promoting the talents of individual property in each country. Accordingly, the Computer Universities generating the qualified human resources enclose several responsibilities to support these necessities. To improve the assessment of each country, we need to study the investigation of human resource development based on education. Thus, we develop the public figures civilizing (PFC) system provided by common partitioning method, k_means: case study area is the examination on degree holders of a computer university. At the moment, the architecture of PFC system is shown in Figure 1.

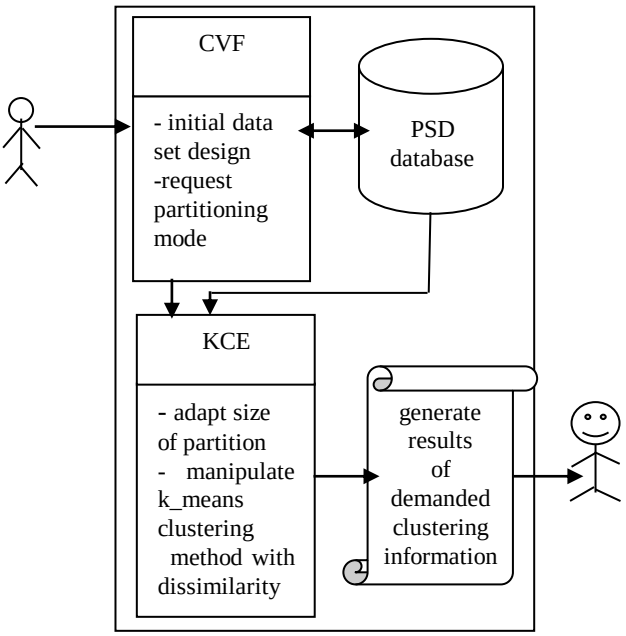

Figure 1. Overview of the system.

There are two components in this figure: Client Visible Module (CVF) and K_Means Clustering Module (KCE). Firstly, each user is allowed to initiate dataset design and a partitioning mode by means of CVF module. The dataset is then stored in PSD database carefully. KCE module manipulates k_means partitioning with dissimilarity method based on interval-scaled variables. Finally, the results of requested partitioning information are taken by every user in this system.

The k-means algorithm takes the input parameter, *k* and, partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the

mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity [6]. It is introduced as how does the k-means algorithm work.

Input : - the number of clusters, *k*
- a data set containing *n* objects.

Output : A set of *k* clusters

Method :

a) arbitrarily choose *k* objects from D as the initial cluster centers
b) repeat
c) (re) assign each objects to the cluster to which the objects is the most similar, based on the mean value of the objects in the cluster.
d) update the cluster means, i.e, calculate the mean value of the objects for each cluster.
e) ***until*** no change.

According to the above algorithm, the module of this partitioning method randomly selects *k* of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

where,

E = the sum of the square error for all objects

p = the point in space representing a given object

m_i= the mean of cluster C_i.

In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting *k* clusters as compact and separate as possible.

In our proposed system, k-means clustering and relational database extraction are played to support the user satisfactory. Firstly, we collect the inclusive information of educated persons in a particular region (Sagaing Division). The system attempts to put in order the consumer necessities to be predefined the partitioning manner. The attributes like cluster cores (district, degree, career of parents and academic year) are well supported to score the improvement of social standard based on education regulation process.

The variable *k* is assigned the value to which the user would like the objects to be partitioned into *k* clusters. Next to the k-means method, an element of the system chooses *k* objects as the *k* initial cluster centers. Every object is transmitted to a cluster based on the member of cluster core to which it is the nearest. At this point, resolving how to transmit an object into a cluster is provided with a dissimilarity value amongst variables. After that, the mean rate of each cluster is reconsidered based on the current objects in the cluster, also known as the cluster cores are updated. Using the new cluster centers, the objects are redistributed to the clusters based on which cluster core is the nearest. The iterative relocations will be completed in a stage in which no distribution of the objects in any cluster occurs.

3.1 Dissimilarity between ordinal variables

In this system, the variables are very functional for registering subjective estimations of qualities that cannot be measured independently. The degrees of holders in computer university are often enumerated in a sequential order, such as B.C.Sc/B.C.Tech,B.C.Sc(Hons:)/.C.Tech(Hons:), M.C.Sc/M.C.Tech, M.I.Sc, M.A.Sc, Ph.D (IT), Ph.D(CHT). Suppose that an ordinal variable *f* has M_f states, these ordered states define the ranking 1, …, M_f. Thus, *f* is a variable from a set of variables describing *n* objects. There are three steps to compute the dissimilarity with regard to *f*.

1) x_{if} is defined for the value of *f* with i^{th} object and *f* has M_f ordered states. Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots, M_f\}$
2) There has been a different number of states, it is often necessary to map the

range of each variable onto [0.0, 0.1] so that each variable has equal weight. This can be achieved by replacing the rank r_{if} of the i^{th} object in the f^{th} variables by

$$z_{if} = {r_{if} - 1} \Big/ {M_f - 1}$$

3) Dissimilarity can then be computed using a method of interval-scaled variables, Euclidean Distance measure, by means of z_{if} to represent the *f* value for the i^{th} object.

3.2 Types of Data in Cluster Analysis

In this section, the preprocessing of data for cluster analysis is demonstrated with n-objects, which may represent the development of public standard in a division. Thus, data matrix and dissimilarity matrix are used to support the operations on a component of the system including K-means Clustering Module (KCM). The former represents *n* objects with *p* attributes, such as name, age, gender, address, degree, and extra information of each person.

$$\begin{bmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix}$$

The latter is to store a collection of proximities that are available for all pairs of *n* objects. It is frequently characterized via an n-by-n table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ . & . & . & & \\ . & . & . & & \\ d(n,1) & d(n,2) & . & . & 0 \end{bmatrix}$$

where, d(i,j) is the précised distinction linking objects *i* and *j*. In general, d(i,j) is a nonnegative integer that is getting ready to 0 when objects *i* and *j* are extremely similar or close to each other, and turns into larger the more they diverge. Given that d(i, j)= d(j,i), and d(i,i)=0, the exceeding matrix can be accomplished.

3.3 Study Area: Statistics of Educated Person

There have been the sample data of the following Table 1 in which the constant variables are available. There are four states for ordinary, namely poor, fair, good, and excellent, that is M_f = 4.

Table 1. Sample values of educated information.

Sr.no	Degree of holders	Ordinary	District name
1	Ph.D Holders	Excellent	Monywa
2	Honous Holders	Fair	Sagaing
3	Master Holders	Good	Tamu
4	Bachelor Holders	Poor	Katha
5	Ph.D Holders	Excellent	Shwebo

Firstly, if we put back each value for ordinary by its grade, the five objects are assigned the grades 4,2,3,1 and 4, respectively. Moreover, the grading of these data must be normalized (where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0)by mapping grade 1 to 0.0, grade 2 to 0.33, grade 3 to 0.67 and grade 4 to 1.0. And then we can say, the Euclidean distance .

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots\ldots+(x_{in}-x_{jn})^2}$$

wher i=((xi1,xi2,….,xin) and j=(xj1,xj2,……,xjn)) are two n-dimensional data objects. For that reason, it is necessary to transform from data matrix into a dissimilarity matrix. Finally, the

Euclidean distance could result in the following dissimilarity matrix:

$$\begin{bmatrix} 0.0 & & & & \\ 0.67 & 0.0 & & & \\ 0.33 & 0.34 & 0.0 & & \\ 1.0 & 0.67 & 0.33 & 0.0 & \\ 0.0 & 0.67 & 0.34 & 1.0 & 0.0 \end{bmatrix}$$

4. EXPERIMENTAL RESULTS

According to the used clustering method, the requested partitioning cluster groups can be achieved as possible as in simple. Numbers of holders in Monywa district are depicted along with each degree class and their percentage are shown in figure 1. In this figure, it has 4 classes namely Ph.D degree, Master degree, Honous degree, Bechlor degree . Table 2 is used to specify the results of holders.

Table 2. Number of holders in each district.

Sr.no	District	Ph.D	Master	Honous	Bechlor
1	Monywa	10	30	40	50
2	Sagaing	9	25	30	40
3	Mawlight	8	20	20	35
4	Katha	5	16	15	25
5	Tamu	0	3	10	13
6	Shwebo	8	28	8	15
7	Kalay	6	7	7	10
8	Khanti	4	2	5	6
Total		50	131	135	194

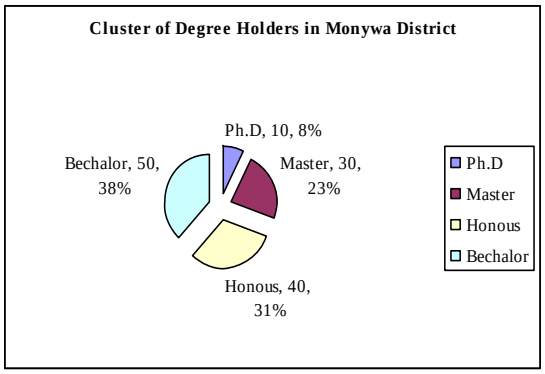

Figure 1. Cluster of degree holders.

Alternatively, this system also provides a dimension for degree holders based on occupation of their parents. In proportion to investigation, the outcome of Ph.D holders in Sagaing Division are put side by side in Figure 2 and the numbers are received by means of Table 3.

Table 3. Relationship based on occupation of parents.

Sr.no	Degree holders	Occupation of Parents			
		Staff	Farmer	other	total
1	Ph.D	7	2	1	10
2	Master	15	5	10	30
3	Honous	16	4	20	40
4	Bechalor	30	15	5	50

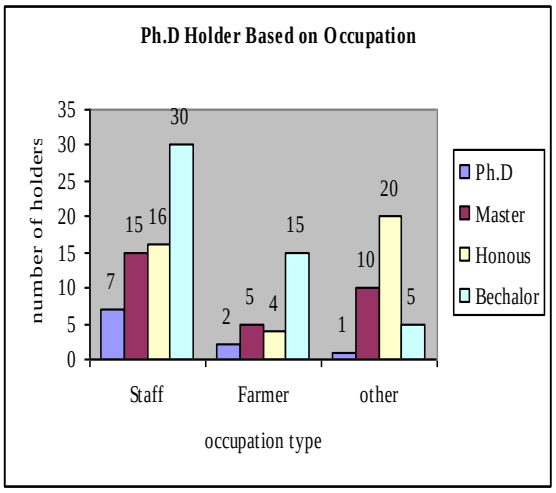

Figure 2. Ph.D holders based on occupation.

5. CONCLUSION

Cluster analysis is an important activity for a specific organization. In this paper, cluster analysis has been used in data analysis to provide the social standard with education level. In addition, K-means clustering method provides the user's requests (k_value) by recognizing these values and by producing the arrangement results automatically. On the other hand, superior partitioning methods will be used in this system and other applications will encourage these techniques. This partitioning method is not finished in accuracy comparing other ones so that it is a further work for unusual ways.

REFERENCES

[1] M. Steinbach, G. Karypis and V. Kumar, "A ComparisonofDocument ClusteringTechniques",in KDD Workshop on Text Mining,2000.

[2] Z. He, X. Xu, and S.Deng, " Improving K-Modes Algorithm Considering the Frequencies of Attribute Value in Mode", in Proceedings of the ICCA Second International Conference on Computer Applications, January 2004.

[3] S. Belongie, C. Carson, H. Greenspan, J. Malik, " Recognition of Images I Large Daabases Using a Learning Framework", Technical Report TR 97-939, U. C. Berkeley, CS Division, 1997.

[4] H. Jiawei and K. Micheline, " Data Mining concepts and Techniques", Morgan Kaufmann, 2001.

[5] A. Sieg, B. Mobasher, S. Lytinen nad R. Burke, " Using Concept Hierarchies to Enhance User Queries in Web-based Information Retrival", in Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria, February 2004.

[6] E.H. Han, D. Boley, M. Gini, R. Gross, and K. Hasting, " WebACE: A Web Agent for Document Categorization and Exploration", in Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98", May 1998.

[7]H.M. Haav and t.L.lubi, "A Survey of Concept based Information Retrieval Tools on theWeb", in Proceedings of 5th East-European Conference ADBIS*2001, Vilnius "Tecdhnika', 2001,Vol.2, pp.29-41.

[8] J.Han, Y. Kim, and S. Hwang, " Content-Based Image Retrieval Method Using Hierarchical Clustering Technique", in Proceedings of the ICCA Second International Conference on Computer Applications, January 2004.