Clustering Large Datasets with Apriori-based Algorithm and Concurrent Processing

Noppol Thangsupachai, Phichayasini Kitwatthanathawon, Supachanun Wanapu,
and Nittaya Kerdprasop, *Member, IAENG*

Abstract—**This paper presents the integrated data mining processing technique to find appropriate initial centroids in data clustering process by k-means algorithm. The processes include data cleansing, preprocessing, and finding features relation with Apriori algorithm to get appropriate features. Our clustering process compares different initial selection schemes: static selection and random selection. The calculation of SSE (Sum of Square Error) uses parallel calculation for better computational performance. We propose the Pre-KMA model that represents the processes for finding appropriate initial clustering centroids and selecting the most relevant features from large datasets. The clustering evaluation results of SSE, loop of clustering, and time of processing confirm that with the Pre-KMA model we can get better clustering result with k-means clustering methodology. The experimental result shows that calculated SSE and processing time are decreased.**

Index Terms—**Data mining, Apriori algorithm, Concurrent processing, K-means clustering**

I. Introduction

The data mining [1] is the automatic process of searching or finding useful knowledge. The process extracts data from large database with mathematics-based algorithm and statistic methodology to reveal the unknown data patterns that can be useful information. The information got from data mining process is very important knowledge that help user in decision making concerned business strategies [2]. These processes are also called Knowledge Discovery in Database (KDD) in that knowledge discovery and analysis can be performed from many information and raw data in databases [8]. The knowledge can be used in decision support system or used to predict customer's behavior or predict product sale rate in the future.

This paper studies various techniques to adapt and improve the data clustering methodology of the k-means clustering. The problems in data clustering with k-means are the selection of initial centroids that can effect to SSE (Sum of Square Error) in each cluster of data. Poor selection results in more time processing. The research has focused on the decrease of both time processing and SSE of data clustering with k-means clustering methodology.

In this paper, the main idea of data mining technique in data clustering from raw data with appropriate initial centroids selection is presented. The techniques used in this paper are Apriori algorithm for feature selection process and clustering data with k-means clustering methodology.

II. Preliminaries

Nowadays data mining technique has many ways to implement and apply to discover knowledge from raw data; this is up to the types of data analysis and usage. The data mining processes used in this paper has many integrated data mining techniques including Apriori and k-means clustering algorithms.

A. Apriori-based Algorithm

The association rules [3, 8] are one of popular data mining techniques employed by several enterprise sectors, especially in retailing business. The association rules are to be used to analyze the sale rate and sold related goods in store. Entrepreneurs can predict and arrange the shelf of products that customers usually bought together [7]. These rules represent in the format of "If…Then" rule that does not like other rules in data mining techniques for example clustering and classification. Mining for association rules have many processes and use more time processing in finding related features in data groups [9]. The result of association rule mining shows many rules which are combination of related features that users have to analysis and select usable set of features [6]. This technique quite differs from classification technique because classification methodology shows the results that are specific to some class of data.

The combination of item sets, or features, from the result of association rule mining has many patterns with several groups of items. Users have to set threshold of minimum support value to limit the result that shows only groups of item sets related to the specified criteria. The

Manuscript received December 8, 2010; revised January 20, 2011. This work was supported by grants from the Nation Research Council of Thailand (NRCT) and Suranaree University of Technology.

N. Thangsupachai is a PhD student with the Institute of Social Technology, School of Information Technology, Suranaree University of Technology, 111 University Avenue, Nakhon Ratchasima 30000, Thailand (e-mail: zoliblade@hotmail.com).

P. Kitwatthanathawon is a PhD student with the Institute of Social Technology, School of Information Technology, Suranaree University of Technology, Thailand (e-mail: ry_cher_@hotmail.com).

S. Wanapu is a PhD student with the Institute of Social Technology, School of Information Technology, Suranaree University of Technology, Nakhon Ratchasima, Thailand (e-mail: supachanun@gmail.com).

N. Kerdprasop is an associate professor with the Institute of Engineering, School of Computer Engineering, Suranaree University of Technology, Thailand (e-mail: nittaya@sut.ac.th).

results are also filtered by minimum confidence value that is to be specified by user corresponding to user’s requirement and usage. Association results include group of related features called “item set” that are considered in each frequent item set, for example, examine two related features in co-occurrence type is called two-item set.

Although the association rules are very effective to find relevant features, the method requires much time to process and analyze every possible item sets. This is due to the process that each item set will be considered and rules are to be generated in each group of item set. Thus association mining has many techniques to speed up time processing in the consideration of item sets [10]. One of those techniques is Apriori-based algorithm. Apriori is a structure to count candidate item sets efficiently. It generates candidate item sets of length k from the k-1 item sets and avoids expanding all the item set’s graph. Then it prunes the candidates which have an infrequent sub pattern. The candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates [8]. With Apriori technique the algorithm can decrease time processing in generating fewer groups of item sets and avoid infrequent candidate item sets expansion.

B. *K-means Clustering Methodology*

The data clustering [8, 10] is processing of raw data to find clusters or groups of similar data. In each cluster, members have some similarity in type of data. The principles of data clustering are finding value of score in similarity, and assigning each member to be in the same group of other members that have similar or same score.

The data mining technique in finding data clusters is different from data classification in that user does not have to specify target feature for assigning each data record to the appropriate cluster. Data clustering is thus an unsupervised learning method. The clustering method relies on the similarity measurement to automatically from groups of relevant or similar data members as visually shown in figure 1. After the clustering process, user can apply some classification algorithm to extract data pattern in each cluster for a better understanding of cluster model.

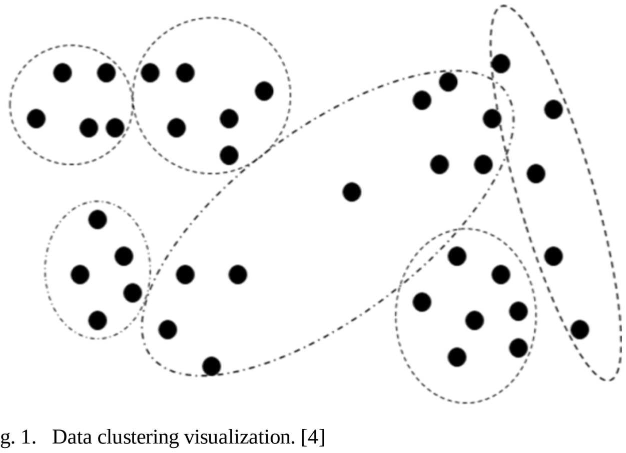

Fig. 1. Data clustering visualization. [4]

K-means clustering algorithm is the most selected technique to cluster data. K-means [7] is a nonhierarchical clustering and use looping to group data into K groups. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

The steps of k-means algorithm [6] can be summarized as the following:

1) Specify group number and select initial centroid of each group.
2) Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.
3) Calculate distance’s mean of every data member and own centroid to define new centroid in each group.
4) Repeat steps 2 and 3 until each group has stable centroid or same centroid.
5) Calculate SSE in each group to evaluate the quality of cluster result.

III. Research Methodology

This research provides Pre-KMA (Pre-Processing K-means Algorithm) model that shows the processes of preparation the suitable data, selection the good centroids, and achieving better clustering performance with concurrent processing. The processes include preprocessing steps that cleansing raw data from IPUMS data [5] and select suitable features by association analysis with Apriori-based algorithm. After that clustering subset of train data by random selection of data representatives to get initial centroids for clustering on all training data to provide good data cluster. With better initial centroids from sampled data and improved algorithm by concurrent processing, our model will get better results in SSE value and time processing.

The research processes follow the Pre-KMA model. The model includes the data set selection from IPUMS website, which is the web of Minnesota’s population center [5] to provide real data in a dedicated to collecting and distributing United States census data. These data has complicated and varied type of categories suitable for data mining task to find some knowledge patterns. These data sets will be selected for interesting subjects related to the research’s objective.

The objective of this research is to study features related to personal total incomes and other features that quite affect the incomes status. All features selected from IPUMs database will be filtered by association technique with Apriori-based algorithm to generate relevant features data sets.

The gained related feature data set will be clustered by k-means clustering technique and improved with the concurrent processing methodology. In k-means clustering, we will compare the result of clustering with clustering technique that got initial centroid from sampled data against the sequential data records without centroid selection technique. With this process we will show the SSE value affected from using initial centroids for clustering.

In the last process of Pre-KMA model we will evaluate and compare results of clustering data. Clustering results are evaluated by comparing the SSE values from clustering, time processing and number of looping in clustering process. All processes in Pre-KMA model are shown in figure 2.

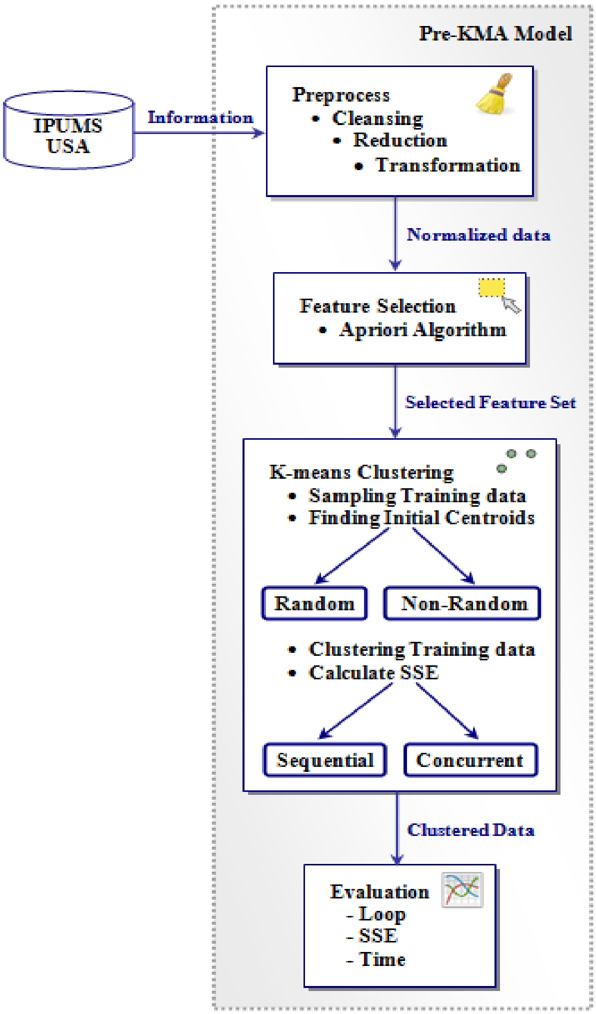

Fig. 2. Pre-Processing K-means Algorithm (Pre-KMA) model

IV. Data Sets and Implementations

A. Large Data Set Selection

This research uses large data sets from Minnesota Population Center [5]. Integrated Public Use Microdata Series (IPUMS) is a census microdata of the United States for social and economic analysis. Data sets are extracted from investigation years ranging from the year 2005-2009 that contain 12 features related to personal incomes. The selected features are: *age*, *occ*, *rooms*, *bedrooms*, *nfams*, *famsize*, *nchild*, *sex*, *marst*, *educ*, *schltype* and *inctot*, shown in table I.

TABLE I
Selected Data Features from IPUMS.

No.	Attribute	Description
1	age	Age
2	sex	Sex
3	marst	Marital status
4	rooms	Number of rooms
5	bedrooms	Number of bedrooms
6	nfams	Number of families in household
7	famsize	Number of own family members in household
8	nchild	Number of own children in the household
9	educ	Educational attainment [general version]
10	schltype	Public or private school
11	occ	Occupation
12	inctot	Total personal income

B. Data Preprocessing

The first process for data mining is data preprocessing that researchers will prepare the data sets for use in data mining processes. Researchers have to set clear criteria to filter all data sets suitable to the research objectives. The first step in data preprocessing is the data cleansing process that gets rid of noise and outlier. Then data has been reduced and transformed into the format that is appropriate for data mining software to analyze and clustering. The criteria for selecting the data records are as follows:

Criteria 1) The personal total incomes selected only range of incomes between 1,000 US$ and 100,000 US$.

Criteria 2) Age ranges between 16 and 70 years old.

Criteria 3) Feature "Number of room" will get rid of the 0 value in the data sets.

Table II shows the amount of data sets after filtering with all criteria.

TABLE II
Data Filtered with Various Criteria

Year	All data	Selected data			Percentage
		Filter criteria 1	Filter criteria 2	Filter criteria 3	
2005	69,092	45,559	39,108	39,108	56.6
2006	70,710	46,893	40,092	39,192	55.4
2007	72,092	47,477	40,476	39,642	55.0
2008	75,836	50,045	42,890	41,920	55.3
2009	77,131	50,385	42,952	42,078	54.6
Total	364,861	240,359	205,518	201,940	55.4

The data sets obtained from data preprocessing are to be clustered by k-means clustering technique. These are the results that will be used to compare with the results from clustering processes that follow the Pre-KMA model. Figure 3 shows SSE values from clustering all data sets in 5 years.

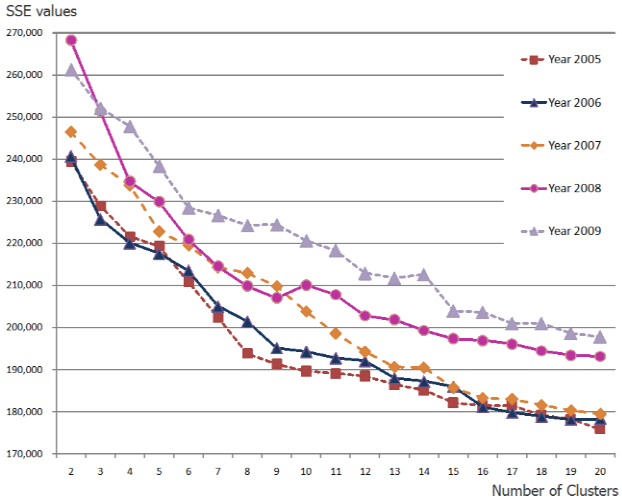

Fig. 3. SSE values from data clustering with ordinary k-means

C. *Feature Selection with Apriori-based Algorithm*

In the continuing step of Pre-KMA model is feature selection processing. The feature selection will use the association technique with Apriori-based algorithm to generate the sets of feature relation rules. With Apriori-based algorithm used to analyze and generate features that are related and affect to other features in the group, more effective action in association technique is required. We have to filter the rules that appropriate to research objective. In this research we aim at finding features that affect personal total incomes. So all features from IPUMS selection will be used to calculate the associate rules with Apriori-based to get related features for clustering. Table III shows the results of main features selected from Apriori-based technique. We then select only common features appeared in years 2005 to 2009, that is, *marst, bedrooms, nfams, sex, educ, schltype* and *inctot*

TABLE III
FEATURE SELECTION'S RESULTS FROM APRIORI ALGORITHM IN 5 YEARS

Year	Related Features in each year
2005	marst, bedrooms, nfams, sex, educ, schltype
2006	marst, bedrooms, nfams, sex, educ, schltype
2007	marst, bedrooms, nfams, sex, educ, schltype, famsize, nchild
2008	marst, bedrooms, nfams, sex, educ, schltype
2009	marst, bedrooms, nfams, sex, educ, schltype

D. *Clustering by K-means Clustering Methodology*

The main process in the Pre-KMA model is clustering selected data with the k-means clustering method. We implement the k-means clustering algorithm with the Erlang programming language. The initial centroid selection of k-means clustering is implemented with two schemes: (1) randomly select centroid points from the data set and (2) pick the first k data points as initial centroids. The effectiveness of both schemes is to be compared on the SSE values after the completion of data clustering process.

On the computation of SSE values we also implement two schemes of programming styles, that is, sequential computation and concurrent computation. Sequential computation calculates SSE values of one cluster after the others in a sequential manner, whereas concurrent computation calculates SSE values of every cluster in a parallel manner.

V. Experimental Results

The experimental results show in the following three sub-sections are the data clustering with SSE values compared in random and fixed centroid selection, amounts of looping in clustering and time processing. All results are demonstrated in as graphical comparisons.

A. *SSE Values*

The data clustering results compare the SSE values between "random" and "sequential" initial centroid selection. With the random initial centroid selection, it provides the better SSE values (the smaller SSE is the better clustering) and more efficient clustering than fixed initial centroids. The results in figure 4 also show that the SSE values decrease as the number of clusters increase.

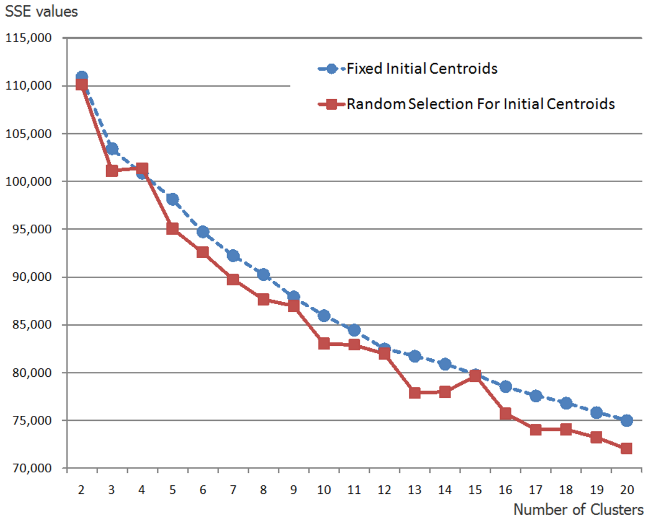

Fig. 4. Compare calculated SSE values

B. *Amount of Looping*

The random selection from train data sets to generated better initial centroids helps the clustering process in fewer amount of cluster loops. Thus the good initial centroid selection can improve the clustering process. If the last calculated centroids is the same or almost the same points with initial centroids, the amount of looping will be decreased. Figure 5 shows the amount of looping from the experimental results. The comparisons are performed with fixed versus randomly selected initial centroids.

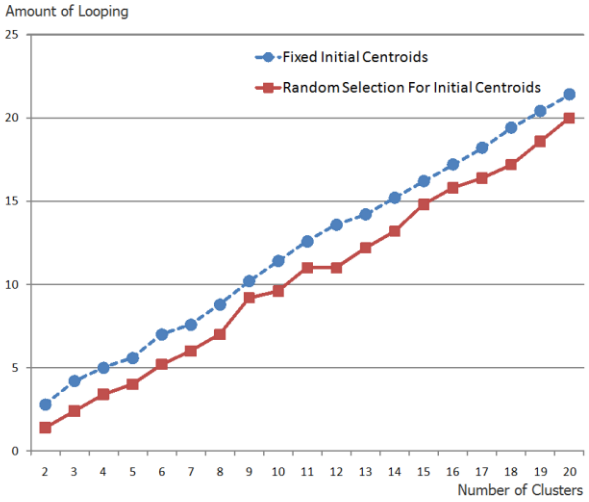

Fig. 5. Compare amount of looping

C. Processing Time

The processing time focuses on the calculation of SSE values with the concurrent processing. With concurrent processing the clustering process is to be spawned into many processes and work in parallel. Figure 6 shows time in clustering and computing the SSE values with concurrent versus sequential processing styles. Concurrent processing takes less time in 14 experiments out of 19.

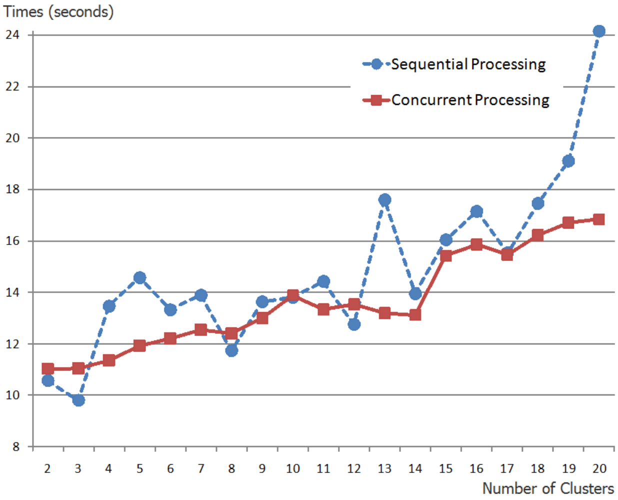

Fig. 6. Compare time processing

VI. Related Work

Dash M., et al. [11] present an effective dimensionality reduction technique, that is an essential pre-processing method to remove noisy features. In other literatures there are only few methods proposed for feature selection for clustering. Most of them are based on technique called “Wrapper” that require parameters such as number of clustering or number of features in datasets. Dash M., et al. proposed a filter method that is independent of any clustering algorithm. They used the entropy measure that is low if data has distinct clusters and high otherwise. The entropy measure is suitable for selecting the most important subset of features because it is invariant with number of dimensions and is affected only by the quality of clustering.

Sekhon J.S., et al. [12] bring up the factor that effected to accuracy and efficiency of clustering algorithm is dependent on the input data. The removing unimportant features from the dataset can help to form better clusters in lesser time. These unimportant features may be those that are redundant, or affected by noise, etc. Researchers need to consider the fact that the features finally choose, should represent the original dataset in the best possible way. In other words, the underlying structure of the original dataset should be the same as that of the dataset that contains only the selected features. They proposed a technique that selects a subset of features that best represent the entire dataset. Research experimental based on two techniques are “Distance measure and Similarity measure”. These techniques are used for pre-processing step that increase the quality of clusters generated by the underlying K-means algorithm.

Ozyer T., et al. [13] applies technique for clustering in parallel process with “Divide and conquer” approach to handle the clustering process. Researchers provide the partitioning a large dataset into subsets of manageable size based on the specifications of users to use in the clustering process; then cluster the partitions separately in parallel. The centroid of each cluster appears as the root of a tree with instances in its cluster as leaves. The clustering process is iteratively applied on the centroids with the trees growing up until get the final clustering. The conquer process is performed to get the actual intended clustering, where each instance belongs to the final cluster represented by the root of its tree.

In our research we used different techniques from other papers to get better result with simplified methods of association mining technique and parallel clustering process. In other features filtering techniques they use complicated algorithm and take more time to process and has some constraint in measuring the distance of each feature suitable for normal distribution datasets. The parallel technique provides better result in processing time for clustering that is up to the chosen datasets and partitioning technique for the parallel method in clustering.

VII. Conclusion

The data mining has many techniques available for users to apply to suitable data types and usage. From this research we present one of unsupervised data mining technique called data clustering that integrated other mining technique and concurrent processing. The preprocessing added the association rules obtained from Apriori-based algorithm in feature selection to get better feature set. In the clustering process we used random data to generate initial centroids that works better for data sets. The experimental results show that the Pre-KMA model helped data mining to process more accurate than traditional method with decrease SSE value. It is even more efficient with concurrent processing with the decrease in processing time and loop of clustering are decreased too.

This research has some improvable points in that feature selection technique proposed in this paper is appropriate for the selected data. For other data sets, the Pre-KMA model needs more experimentation and efficiency confirmation. The concurrent technique can also be improved for a better parallelization performance.

References

[1] Bertrand C., Ernest F. and Zhang H.H., *Principles and theory for data mining and machine learning*. USA: Springer, 2009, pp. 405-485.

[2] Chakrabarti S., *Data mining : know it all*. USA: Morgan Kaufmann Publishers, 2009, p. 55, p. 160.

[3] Chu W. and Li T.Y., *Foundations and advances in data mining*. USA: Springer, 2005, pp. 125-162.

[4] Han J. and Kamber M.(2001). Data Mining : concept and techniques CA: Academic Press.

[5] IPUMS. (2010, November 1). Available: http://usa.ipums.org/usa/

[6] Lawrence K.D., Kudyba S. and Klimberg R.K., *Data mining methods and applications*. USA: Auerbach Publications, 2008, pp. 83-104.

[7] Rahman H., *Data mining applications for empowering knowledge societies*. USA: Information Science Reference, 2009, pp. 43-54.

[8] Taniar D., *Data mining and knowledge discovery technologies*. USA: IGI Pub, 2008, pp. 118-142.

[9] Wang J., *Data warehousing and mining : concepts, methodologies, tools, and applications*. USA: Information Science Reference, 2009, pp. 303-335.

[10] Wu X. and Kumar V., *The top ten algorithms in data mining*. USA: CRC Press, 2009, p. 21, p. 93.

[11] Dash M., Choi K. and Scheuermann P., “Feature selection for clustering - A filter solution,” in *Proc. IEEE International Conference on Data Mining (ICDM’02)*, 2002, Maebashi: Japan, pp. 115-122.

[12] Sekhon J.S., Gophlkrishnan V. and Keong W. “Proportionate feature selection - A pre-processing step for clustering,” in *Proc. IEEE International Conference on Systems Man and Cybernetics (SMC 2008)*, Singapore, 2008, pp. 2622-2627.

[13] Ozyer T. and Alhajj R. “Parallel clustering of high dimensional data by integrating multi-objective genetic algorithm with divide and conquer,” *Applied Intelligence*, vol. 31, no. 3, pp. 318-331, Dec. 2009.