Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Enhanced Clustering Algorithm on Academic Activities

Er. Gurpreet Singh[1], Prof. MD. Yusuf Mulge[2], Er.Roop Lal[3], Er.Amarjeet Kaur[4], Er.Nishant Pathak

[1]Research Scholar (Ph.D Computer Sc.),Pacific Academy of Higher Education & Research University. Udaipur, Rajasthan,

[2]Principal, PDM College of Engg. For women, Sarai, Aurangabad, Bahadurgarh, Haryana

[3]Assistant Professor, Department of Computer Sc. & Engg., St. Soldier Institute of Engg. & Technology, Jalandhar,

[4]Assistant Professor, Department of Computer Sc. & Engg., St. Soldier Institute of Engg. & Technology, Jalandhar,

[5]Assistant Professor, Department of Computer Sc. & Engg., St. Soldier Institute of Engg. & Technology, Jalandhar,

Abstract—Clustering analysis method is one of the main analytical methods in data mining, the method of clustering algorithm will influence the clustering results directly. This paper discusses the traditional clustering algorithms and analyzes the shortcomings of standard algorithm, such as the k-means clustering algorithm has to calculate the distance between each data object and all cluster centers in each iteration, which makes the efficiency of clustering is not high. This paper proposes an improved algorithm in order to solve this question, requiring a simple data structure to store some information[1] in every iteration, which is to be used in the next iteration. The improved method avoids computing the distance of each data object to the cluster centers repeat, saving the running time. Experimental results show that the improved method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the traditional algorithm.

Keywords: K-Means, SOM, HAC, ECA

I. INTRODUCTION

Clustering has been one of the most widely studied topics in data mining. Clustering refers to techniques for grouping similar objects in clusters. Formally, given a set of dimensional points and a function that gives the distance between two points , we are required to compute cluster centers, such that the points falling in the same cluster are similar and points that are in different cluster are dissimilar. Most of the initial clustering techniques were developed by statistics or pattern recognition communities, where the goal was to cluster a modest number of data instances. However, within the data mining community, the focus has been on clustering large datasets. [2]Developing clustering algorithms to effectively and efficiently cluster rapidly growing datasets has been identified as an important challenge.

This paper includes four parts: The second part details the k- means algorithm and shows the shortcomings of the standard algorithms like k-means, SOM, HAC The third part presents the improved clustering algorithm, the last part of this paper describes the experimental results and conclusions through experimenting with academic data sets.

II. THE K-MEANS CLUSTERING ALGORITHM

The process of k-means algorithm This part briefly describes the standard k-means algorithm. k- means is a typical clustering algorithm in data mining and which is widely used for clustering large set of datas. In 1967, MacQueen firstly proposed [3]the k-means algorithm, it was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-known cluster It is a partitioning clustering algorithm, this method is to classify the given date objects into k different clusters through the iterative, converging to a local minimum. So the results of generated clusters are compact and independent. The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance. The next phase is to take each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum. Supposing that the target object is x, xi indicates the average of cluster C_i, criterion function is defined as follows:

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

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data object and cluster center. The Euclidean distance between one vector $x=(x_1, x_2, \ldots x_n)$ and another vector $y=(y_1, y_2, \ldots y_n)$.

Steps:

1) Randomly select k data objects from dataset D as initial cluster centers.
2) Repeat;
3) Calculate the distance between each data object di (1 <= i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest cluster.
4) For each cluster j (1<=j<=k), recalculate the cluster center.
5) until no changing in the center of clusters.

III. THE KOHONAN-SOM CLUSTERING ALGORITHM

Kohonon's SOMs are a type of unsupervised learning. The goal is to discover some underlying structure of the data. However, the kind of structure we are looking for is very different than, say, PCA or vector quantization.

Kohonen's SOM is called a topology preserving map [4]because there is a topological structure imposed on the nodes in the network. A topological map is simply a mapping that preserves neighborhood relations.

In the nets we have studied so far, we have ignored the geometrical arrangements of output nodes. Each node in a given layer has been identical in that each is connected with all of the nodes in the upper and/or lower layer. We are now going to take into consideration that physical arrangement of these nodes. Nodes that are "close" together are going to interact differently than nodes that are "far" apart.

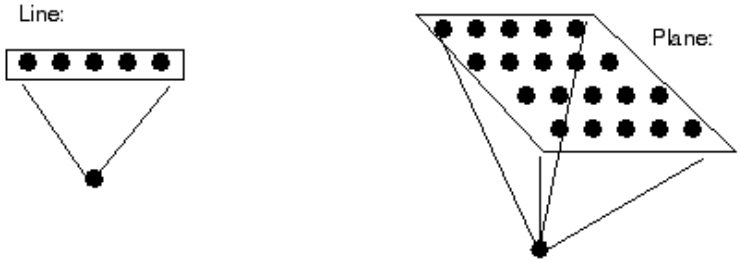

Fig: 1 The goal is to train the net so that nearby outputs correspond to nearby inputs.

The Hierarchical clustering algorithms

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC .[6] Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached.

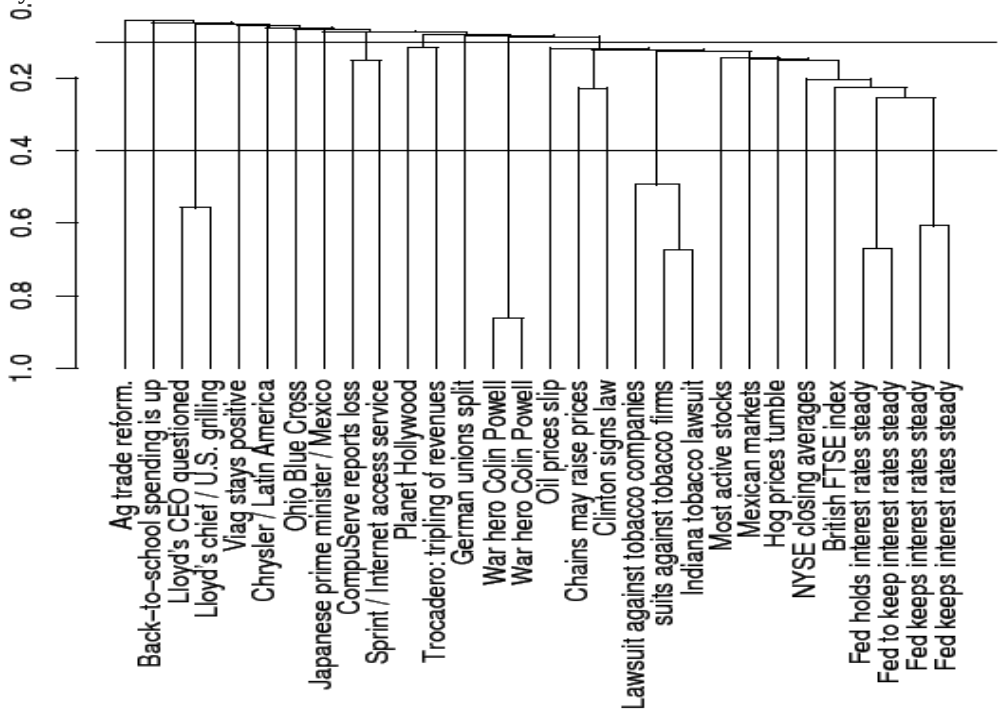

Fig 2: A dendrogram of a single-link clustering of 30 documents from Reuters-RCV1. Two possible cuts of the dendrogram are shown: at 0.4 into 24 clusters and at 0.1 into 12 clusters.


```
SIMPLEHAC(d_1, ..., d_N)
 1  for n ← 1 to N
 2  do for i ← 1 to N
 3     do C[n][i] ← SIM(d_n, d_i)
 4     I[n] ← 1  (keeps track of active clusters)
 5  A ← []  (assembles clustering as a sequence of merges)
 6  for k ← 1 to N − 1
 7  do ⟨i, m⟩ ← arg max_{⟨i,m⟩: i≠m ∧ I[i]=1 ∧ I[m]=1} C[i][m]
 8     A.APPEND(⟨i, m⟩)  (store merge)
 9     for j ← 1 to N
10     do C[i][j] ← SIM(i, m, j)
11        C[j][i] ← SIM(i, m, j)
12     I[m] ← 0  (deactivate cluster)
13  return A
```

The shortcomings of Traditional Algorithms

We can see from the above analysis of algorithms, the algorithm has to calculate the distance from each data object to every cluster center in each iteration. However, by experiments we find that it is not necessary for us to calculate that distance each time. Assuming that cluster C formed after the first j iterations, the data object x is assigned to cluster C, but in a few iterations, the data object x is still assigned to the cluster C. In this process, after several iterations, we calculate the distance from data object x to each cluster center and find that the distance to the cluster C is the smallest. So in the course of several iterations, k-means algorithm is to calculate the distance between data object x to the other cluster center, which takes up a long execution time thus affecting the efficiency of clustering.

IV. IMPROVED CLUSTERING ALGORITHM

The standard k-means algorithm needs to calculate the distance from the each date object to all the centers of k clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases. For the shortcomings of the above k-means algorithm, this paper presents an improved k-means method. The main idea of algorithm is to set two simple data structures to retain the labels of cluster and the distance of all the date objects to the nearest cluster during the each iteration, that can be used in next iteration, we calculate the distance between the current date object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in it's cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k- 1clustering centers, saving the calculative time to the k-1 cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers, and find the nearest cluster center and assign this point to the nearest cluster center. And then we separately record the label of nearest cluster center and the distance to it's center. Because in each iteration some data points still remain in the original cluster, it means that some parts of the data points will not be calculated, saving a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

V. PROPOSED ALGORITHM

The process of the improved algorithm is described as follows: Input: The number of desired clusters k, and a database $D=\{d_1, d_2, \ldots d_n\}$ containing n data objects. Output: A set of k clusters

1. Draw multiple sub-samples {SI, S2, . . . , Sj } from the original dataset.
2. Repeat step 3 for m=l to i
3. Apply combined approach for sub sample.
4. Compute centroid
5. Choose minimum of minimum distance from cluster center criteria
6. Now apply new calculation again on dataset S for k1 clusters.
7. Combine two nearest clusters into one cluster and recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

This paper proposes an improved algorithm, to obtain the initial cluster, time complexity of the improved algorithm is O(nk). Here some data points remain in the original clusters, while the others move to another clusters. If the data point retains in the original cluster, this needs O(1), else O(k). With the convergence of clustering algorithm, the number of data points moved from their cluster will reduce. If half of the data points move from their cluster, the time complexity is O(nk/2). Hence the total time complexity is O(nk). While the standard k-means clustering algorithm require O(nkt). So the proposed algorithm in this paper can effectively improve the speed of clustering and reduce the computational complexity. But the improved k-means algorithm requires the pre estimated the number of clusters, k, which is the same to the standard k-means algorithm. If you want to get to the optimal solution, you must test the different value of k.

VI. EXPERIMENTAL RESULTS

This paper selects academic data set repository of machine learning databases to test the efficiency of the improved algorithm and the standard algorithms. Two simulated experiments have been carried out to demonstrate the performance of the improved algorithm in this paper. This algorithm has also been applied to the clustering of real datasets. In two experiments, time taken for each experiment is computed. The same data set is given as input to the standard algorithm and the improved algorithm. Experiments compare improved algorithm with the standard algorithm in terms of the total execution time of clusters and their accuracy. Experimental operating system is Window 8, program language is java. This paper uses academic as the test datasets and gives a brief description of the datasets used in experiment evaluation. Table 1 shows some characteristics of the datasets.

Dataset	Number of attributes	Number of records
Academic Activities	15	3303

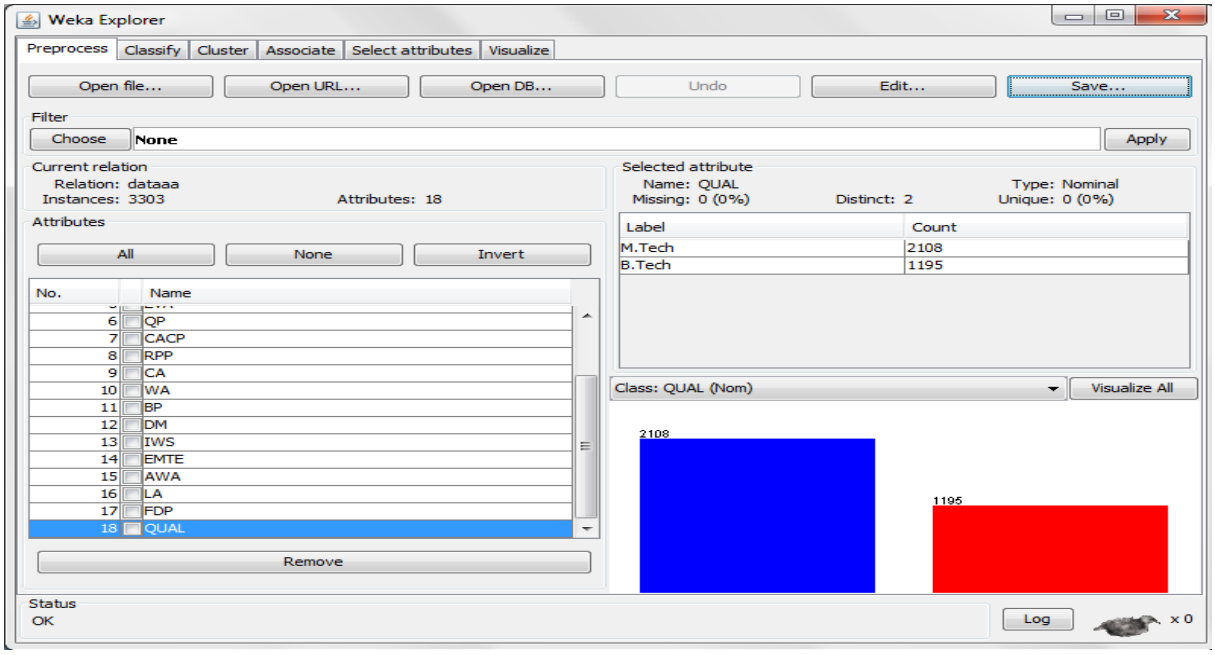

Fig 3: Display data set according to class attributes.

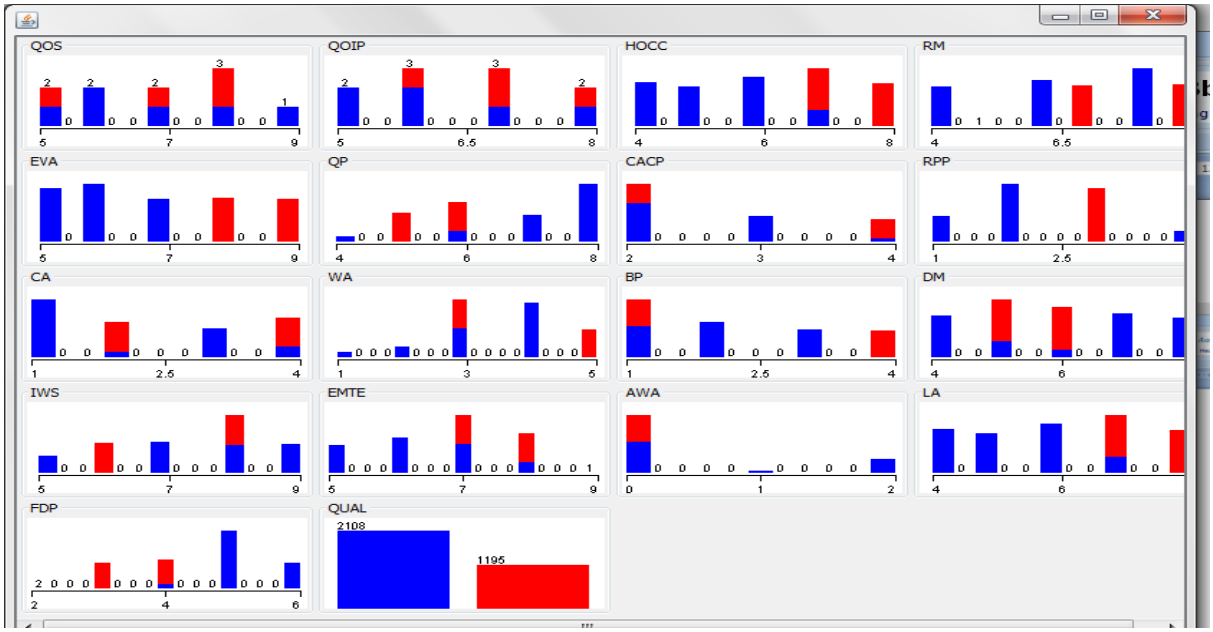

Fig 4: Display All Attributes

Table 2: Analysis between traditional and enhanced algorithm

Parameter	**SOM**	**K-Means**	**HAC**	**ECA**
Error Rate	0.8189	0.8456	0.8379	0.3672
Execution Time	297 ms	1281 ms	1341 ms	1000 ms
Accessing Time	Fast	Slow	Slow	Very fast

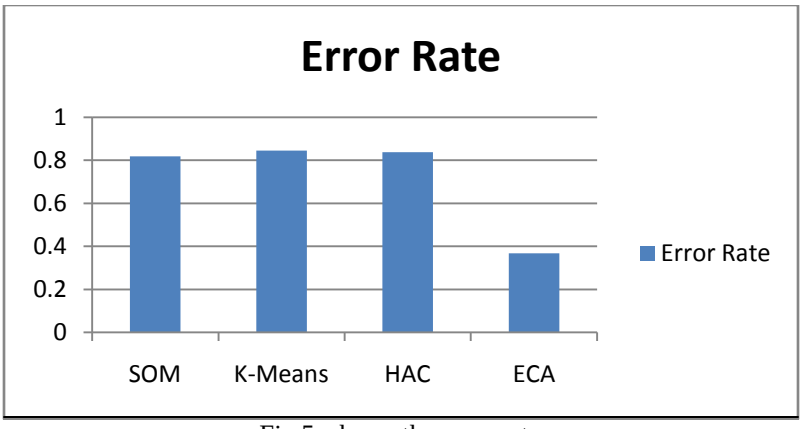

Fig 5: shows the error rate

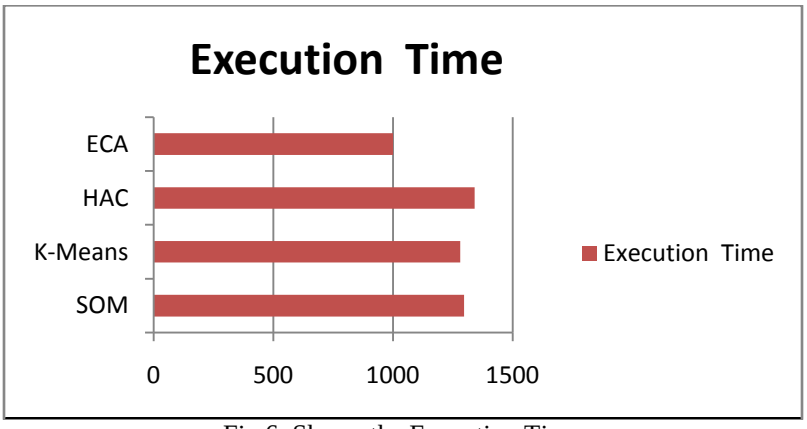

Fig 6: Shows the Execution Time

VII. CONCLUSION

K-means is a typical clustering algorithm and it is widely used for clustering large sets of data. This paper elaborates enhanced algorithm and analyses the shortcomings of the standard k-means , SOM and HAC clustering algorithm. Because the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times during every iteration, which makes the efficiency of standard clustering is not high. This paper presents a simple and efficient way for assigning data points to clusters. The proposed method in this paper ensures the entire process of clustering in O(nk) time without sacrificing the accuracy of clusters. Experimental results shows the improved algorithm can improve the execution time of k-means algorithm. So the proposed method is feasible.

REFERENCES

[1] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.

[2] Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[3] Sun Shibao, Qin Keyun,"Research on Modified k-means Data Cluster Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200– 201,July 2007.

[4] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[5] Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.

[6] Zhao YC, Song J. GDILC: A grid-based density isoline clustering algorithm. In: Zhong YX, Cui S, Yang Y, eds. Proc. of the Internet Conf. on Info-Net. Beijing: IEEE Press, 2001. 140−145. http://ieeexplore.ieee.org/iel5/7719/21161/00982709.pdf

[7] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998.

[8] K.A.Abdul Nazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm",Proceeding of the World Congress on Engineering, vol 1,london, July 2009.

[9] Fred ALN, Leitão JMN. Partitional vs hierarchical clustering using a minimum grammar complexity approach. In: Proc. of the SSPR & SPR 2000. LNCS 1876, 2000. 193−202. http://www.sigmod.org/dblp/db/conf/sspr/sspr2000.htm

[10] Gelbard R, Spiegler I. Hempel's raven paradox: A positive approach to cluster analysis. Computers and Operations Research, 2000,27(4):305−320.

[11] Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997. 146−151.

[12] Ding C, He X. K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. In: Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press, 2004. 584−589. http://www.acm.org/conferences/sac/sac2004/

[13] Hinneburg A,Keim D.An efficient approach to clustering in large multimedia databases with noise.In:Agrawal R,Stolorz PE,Piatetsky- Shapiro G,eds.Proc.of the 4th Int'l Conf.on Knowledge Discovery and Data Mining(KDD'98).New York:AAAIPress,1998.58~65.

[14] Zhang T,Ramakrishnan R,Livny M.BIRCH:An efficient data clustering method for very large databases.In:Jagadish HV,Mumick IS,eds.Proc.of the 1996 ACM SIGMOD Int'l Conf.on Management of Data.Montreal:ACM Press,1996.103~114. [15] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial- temporal data. Data & Knowledge Engineering, 2007,60(1): 208-221.

