RESEARCH ARTICLE OPEN ACCESS

# Survey On Finding Frequent Item Set In Data Mining Using Clustering Approach

Shanmugapriya.S [1], Dr.Pothula Sujatha [2], Prettekha.A [3]
Computer Science Engineering, Pondicherry University, Pondicherry
Assistant Professor, Computer Science Engineering, Pondicherry University, Pondicherry
Computer Science Engineering, Pondicherry University, Pondicherry

**ABSTRACT**
Over the last years, finding of frequent item set of data is playing a vital role in data mining; It is an interesting branch on data mining that fully focus on sequence of pattern and events in way they dressed up. Likewise traditional techniques in data mining are static in nature; temporal data on database will affect the results while mining the dataset. Temporal data set in the sense consist of time values like current in valid time and current in transaction time. So this became more important topic with new algorithms coming up to date. There are different types of clustering algorithms involved. This paper deals with the clustering algorithms till now used for data mining.

***Keywords :—*** Data Mining, Frequent item set, Clustering Approach.

## I. INTRODUCTION

Mazaher Ghorbani et al [1] proposed that, Data mining is used to get new large sets of similar pattern or events, these needs some database and statistic techniques. The practice of automatically searching large stores of data to discover trends and events by using many algorithms by simple analysis and sophisticated mathematical models and also to segment data, evaluating the probability of future events also called Knowledge Discovery in Data (KDD). This had used first the association rules for finding frequent item set. After that classification techniques came into existence this has so many types like context based, fuzzy based etc.

Majorly all the analysis are done using the static item sets, however we need to improvise our self to temporal data. In this study paper we propose an algorithm for high quality approximation of finding large item set on temporal data. For example, we can have small transactional database of supermarket, customer bought {bread, milk} together with 20% support value means 20% of all transaction contain milk and bread together. But all these supermarket, bank etc contain time; hence called temporal databases contain time stamping information. From this we get that discovering events with time interval can be done with association rules discussed later in this, apart from this frequent item set mining having application in biological analyzing data, bug detection in software, banking, etc. We are going to take some algorithms in the sense they are association rules, clustering rules and also general algorithms to find frequent item set data and polished Apriori algorithm.

Rakesh Agrawal et al [2] proposed that, For finding valid time intervals during frequent pattern discovery, We will firstly represent it by the time cubes (TC), frequent item set are got based on giving threshold value, support and density. And item set are discovered by time intervals merged. These are explained in later in this paper.

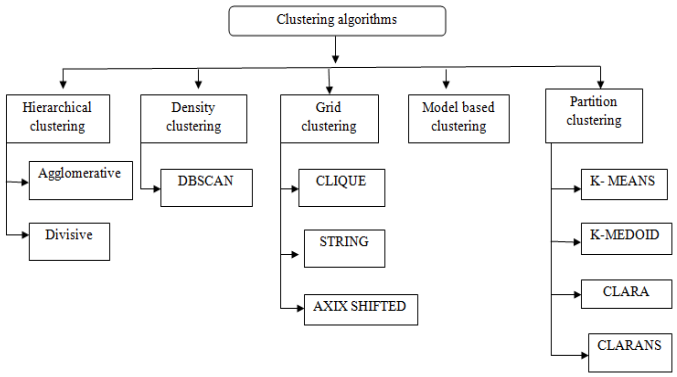


Fig. 1 clustering Algorithms

## II. CLUSTERING APPROACH

Asha Devi et al [3] said that, an unsupervised learning is a clustering mostly used in statistical data. Breaking group of objects or items into similar cluster that belongs to the same cluster. Clustering used in many applications like medical, market, business, weather forecast etc. This is used in data science involves same grouping only by grouping of data points, each need to be in same group have similar properties and features, other groups should have other similar properties. There are many clustering techniques available the mainly used are:

1. Hierarchical clustering
2. Partition based clustering
3. Model based clustering
4. Grid clustering
5. Density based clustering.

## III. PARTITION BASED CLUSTERING

Swarndeep Saket et al [4] suggest that, even there are many clustering algorithm available the mostly used is partition

algorithms the main usage is simple and effective, this will be used to divide the data points. K (partitions). The partition technique will improve the graph and realize each one cluster must consist of one object compulsory. For reducing the error and weakness of the algorithm we going to this partition technique. It has different types they are: K- means it provide.

### *A. K - MEANS ALGORITHM*

Erzhou zhu et al [5] proposed that**,** Even though there are more number of partition algorithms are there, the K- means is most popular used because of its simplicity and effectiveness. The partition technique providing k-partitions of cluster, from a dataset having x objects.

Kalpit G. Soni et al [6] proposed that, the selecting a groups for randomly to get our center point. It is represented by centroid where all average objects fall on it. When a random set of data or some observation of data is provided means the algorithm is applied then it will form some similar form of different clusters named as 'K' clusters. The center points have same length. Secondly, the distance between the points and each center will be classified and find the group which is closest for them. And based on this, we can again compute by taking mean of each group. By doing this repeatedly we will get number of iteration set and till grouping centers doesn't change much with that. It has linear complexity of O (n).

The steps can be followed as: the observations need to be taken and the initial K number of clusters is taken as with the similarities. The centroid which is nearer will be removed or taken. The following steps will be repeated until the near centroid has been caught. The result will be got as the K clusters with the minimum value. This can be explained as,

**Algorithm [6]:**

**Input:** the number of objects given: x

**Output:** based on similarities clusters got i.e.: (k).

Steps involved:

1. Select data randomly and assign centroids.
2. Continue and repeat.
3. With the similarities give data points for all clusters.
4. The centroid will be revised by getting mean of each cluster.
5. Repeat when no changes occur.

**The limitations are:**

- Not easy or predict the K value.
- While working with the global cluster it is difficult but it will be faster that hierarchical in some times.

### *B. K - MEDIODS ALGORITHM*

K- MEDIODS is a partition technique. Preeti Arora et al [7] said that, there are some limitations on the K means algorithms so to get back the disadvantages we are going to the K-MDIODS algorithm. As we said above that finding the K value is difficult in the K means, in this the Medoids is data object used here, in which the Mediods are located centrally in cluster. This is used instead of finding the K value. The K in the sense deals as a metric and variance for measuring scattering of different clusters called Euclidean distance. Where the K is an input parameter for the number of clusters used.

The K-Mediods clustering also called as PAM (Partitioning around Medoids). This has some steps to follow. They are:

Tagaram Soni Madhulatha [8], said that **algorithm**:

Step 1: initializing and select the K from the N data points called medoids (M) randomly.

Step 2: assigning all the associate data point to that close medoid that is got from step 1.

Step 3: updating each data point O for each medoid M to associate swap M and O and get the total cost by computing them. So the medoid O got which is at low cost will be selected from the configuration.

And then repeat the steps 2 and 3 until no change in medoids said **by Swarndeep Saket et al [4].**

**Limitations of K-MEDOIDS:**

- Costlier due to time complexity then by K-means.
- Use of large data set is difficult.
- Result depends on initialing parameter due to run time.

### *C. CLARA ALGORITHM*

Swarndeep Saket et al [4] said that, the extension of the K-MEDOID is CLARA algorithm. It is more efficient that the entire above algorithm used. Arpita Nagpal et al [9] said that, the CLARA used when a sample data is taken. The normal PAM algorithm used, it will select the medoid from the give data set, the data set can be randomly selected as user wish. The CLARA will be used for large data set which gives best result for the given set of data.

Raymond T. Ng et al [10] proposed that, for the better result when using the CLARA can be used will be obtained. The quality of the algorithm is taken by measuring average dissimilarity with the data set used.

**Algorithm** [10]

Step 1: Take data of 40+2K from the data set. When i=5, find the medoid using the PAM algorithm.

Step 2: For every object O in the data set, find the most similar K- medoid to O.

Step 3: now we can calculate the average dissimilarity that is obtained. If the obtained value is less means it will be used as the best medoid and can be used further.

Step 4: repeat the steps for other data.

**Limitation of the CLARA:**

1. The result will be better by seeing the size of data set used.
2. While the biased data set is used means there will be poor clustering result.

### *D. CLARANS ALGORITHM*

Swarndeep Saket et al [4] said that, CLARANS (Clustering large Application Based on Randomized Search) used to

improve the CLARA algorithm and to overcome the limitation of the K-MEDOID algorithm. There we won't able to use large data set, but the CLARANS using the large data set. This is more efficient and scalable. Raymond T. Ng et al [11] proposed that, this CLARANS not only used for finding the data points in the given data set and also used to find the polygon objects. This is one of the main memory techniques. This uses two parameters – num local and max neighbor.

**ALGORITHM**

Step 1: take two parameters of num local and max neighbor.

Step 2: select the K values or objects O using the PAM algorithm from database D.

Step 3: take the selected K objects and mark as Si, and take all other non selected objects apart.

Step 4: calculate the cost differential T for Si.

Step 5: if they obtained cost is positive means it is chosen as local optimum. Otherwise it is updated.

Step 6: next it is repeated to find the local minima.

**Limitation of CLARANS:**

1. Not safe to use the random neighbors.
2. Use of large data set is not much efficient.
3. Search in local area is not guaranteed.

### *E. MEAN SHIFT ALGORITHM*

Dorin camaniciu [12] proposed that, A technique for finding data points in dense or close area. Normally we call it as sliding window algorithm. It is an iterative approach; the main aim is to locate a center point in the group of data items. Hence we are calling this as circular sliding; this will have some radius (r) as a kernel. And with centroid (c). Mean shift algorithm is done in flat kernel. Mainly it is centroid based, hence it is noted that all region will be given as mean points, to find the centroid we need to update the candidate, we need it because to filter the duplicates near the pre processing stage.

When people using the low performance system means due to some technical issues there may be laggings in their result and efficiency. Hence to get rid of this we need the help from high level machines which is independently doing its work. Hence these using for feature extracting by controlling parameter tuning process. The main advantage of using mean shift is to get the image as sharpened. The mean shift will start its process by searching all the spaces using kernel, the tessellated region are not required by the kernel and are eliminated. The details got from the image will be needed to feature exploitation in space. Convergence yielded in the space will be made different location and using the threshold value. The bandwidth need to be calculated. All the unwanted things like noises should be preprocessed. After the process of the convergence the mean shift will take care of delineates the cluster of arbitrary (arbitrary clusters). With the independent test we can get rid of abstract cluster problem.

Bandwidth selection: there are 3 types in selecting bandwidth.

1. If there is stability in decomposition, then the large OS will taken as with the range of clusters that are obtained.
2. If the validity of the index in the cluster giving good quality of the decomposition with maximum of bandwidth gives connectivity and isolation in finding the bandwidth.
3. The user wills them able to control the kernel level bandwidth.

The mean shift using grid structure. When reduction is done the computation time is done to find the data points using kernel in the system. The advantage is it is independent algorithm used for to produce different vision algorithms. And using the calculated bandwidth the performance is analyzed. This used for mainly images both the grey and colored images can be used in this algorithm (mean shift).

## IV. DENSITY BASED CLUSTERING

Sunita Jahirabadkar et al [13] proposed that, If there given many number of points in a certain region of space many points are placed , so here, the clustering technique used to locate the similar points which are related in close way and they will be used to sketch the points which are at low density in that particular region. Hence this is called density based clustering technique.

### *F. DBSCAN SHIFT ALGORITHM*

DBSCAN algorithm will reduces lots of unnecessary query operations, by dividing data space into grid cells. It won't be used when more or high dimensional region. There is two parameter or variable to denote radius and threshold. It can be denoted as € and µ. If we need to check or find the connectivity we have to find various different objects first, like: core objects, border objects, density reachable objects, and noise objects.

**Border objects**: an object which is not surrounded by neighborhood objects like µ (minimum number of objects). And the neighborhood object can be denoted as €. The border objects are nothing but it won't do with the content or core objects as said. The border objects, core objects will be connected with the density objects. The core objects will be surrounded by the minimum number of µ i.e. objects.

**Density reachable objects**: let's consider a density region where all the objects are placed, so that each of all objects can be reached instantly and shortly. So, in an area a single object can connect with the nearby object with reachable points and that can be having dimension point denoted by we can take some value like A. where A is the dimension in core object.

**Density connected points**: In this any number of objects can be reached or connected with the use of clusters by joining similar objects into train or chain of objects. This can do with the help of core objects, the chain of clusters may be taken as C1, C2, and C3 etc.

**Noise data objects**: the density based clustering using the threshold value for finding the nearest value. All the core objects are not using the clusters and that can be reachable and having connected objects. All the reachable and connected

objects are formed using the cluster of core object; this is used to find the efficiency and quality of objects founded. These are said as the noise objects. The dense based clustering used for high dimensional data only this algorithm using this as its property. And using the K nearest for finding the neighbor distance object. This also have disadvantages like it does not deal with dense cluster with the space cluster. If there is index structure means it is very efficient to find objects.

MENG Dongxue et al [13] said that, this is used with the grid for reducing cost. And has high security and complexity. Used to find the point P from the data set D, which is a core point. If the Eps neighborhood is greater than Min point named as the seed point while doing the process. And then it is sequentially arranged.

## V. HIRARCHICAL CLUSTERING

Huiying Wang et al [14] proposed that, in this type of technique we will be used text based normally for finding the frequent item set. With this we can see whether the text given is getting overlapped or performed well. When we are using the usual methods like grid based clustering, model based clustering means it is stated that it is also efficient in nature but there is no high dimensional data in use. So we are going for the text mining with the hierarchical clustering. Here also first we use the normal apriori and then all other are used for finding the frequent item set. Normally we will go with the first of feature to be extracted and then only it is mined. Hence we need to segment the data first.

Mining the feature of the text: actually we will use the association rules for the mining the data. The association consists of for example we can take the transactional database in that we will give the support values with minimum support like we take as 50% . let us as X set of items if the X>=support value then we can take that X is frequent item set. Otherwise this is not the frequent item set. The second step is we need to give a confident value for this that can be support value>=confident value. After that the features are identified using the FTSC algorithm.

Rasmita Dash et al [15] said that, In dealing with the statistical data set we would use large dataset but it is difficult that we aim to predict something but what we got is really anything apart from that hence it is tough to predict the need data. Hence, we go for a property called **grating.** This technique used in many clustering algorithm like swarm clustering, hierarchical clustering, partition method clustering and the K means. We have seen many clustering algorithms but we would confuse that which should be used for the experiment. We will take by seeing the high performance by we can the property of grating. And also we need to see the data base and the index values. In this grating the threshold value is used. First the features are discarded i.e. by using the threshold value. There are some certain stages for getting the correct value that can be determined by validity indices. First the grating is done and second the checking process of the datasets with the valid indices took places and last with the statistical non parametric process is done to validate our result.

## VI. GRID BASED CLUSTERING

Pavel Berkhin [16] suggests that, the all previous algorithms said with the concept of connection, density and boundary that they were dealing. And the grid clustering is the next way to deal with this parameter efficiently. Here multi rectangular segments are used call as segments but there are some attributes used to numerical values for partitioning. This is combination of both hierarchical and partitioning algorithms. All the segments are used in the Bang theory.

Qiang Zhang [17] proposed that, other algorithm facing the problem of Euclidean distance so this suggesting a new rectifying technique to reduce this problem using the CES has been said. CES (clustering algorithm based on grid similarity), has a new distance metric for the distance. The neighbor will convert the projection point intersection that will speed up the clustering. This has same efficient as the DBSCAN algorithm.

[18] Said that, the procedure for grid structure is first we need to partition the space into the number of cells, and we need to calculate the density for each cell, next sorting the file happened with their density and with that find the cluster center and then traverse through neighbor cells.

### *G. STRING ALGORITHM*

This used to find different spatial clustering databases. Used with the spatial queries. All these will be divided into rectangular cell with the hierarchical structure. Explained as,

Step 1: construct the grid structure with given database and will produce the cell with hierarchical structure.

Step 2: for every cell find the confidence with the probability using the query.

Step 3: if layer is not at bottom then go to next step and find the confidence again step 3.

Step 4: if need query is got then find the region.

Step 5: repeat the process for next iteration.

Limitations of grid structure

1. Using the time complexity of O (N).

There is strong dependency on clusters.

## VII. MODEL BASED CLUSTERING

There are some disadvantages in using the hierarchical clustering and partition are not at formal model. Hence we went for the model based algorithm; Arpita Nagpal [9] et al said that, this is based on data that is generated by the probability distribution. This consists of many algorithms namely, EM (Expectation maximization), this is new extend of K means this will have weight. COMWEB this is incremental learning and has used to adjust the partition which has given as the input parameter. CLASSIT it is combination of the both above said.

Zhong et al [19] said that, the Model based is used as the Balance based clustering algorithm this is extend to the soft clustering. This using the synthesis and real data for usage of the soft clustering. There is complexity of O (N log (N)), hence this is need to be balanced.

TABLE I
COMPARISON ON VARIOUS CLUSTERING ALGORITHMS

| Parameters used | k-means | Mean shift | DBSCAN | Birch | Hierarchical clustering |
|---|---|---|---|---|---|
| **Data set** | Numerical dataset | Real data set | Numerical dataset | Numerical data set | Unlabeled data set |
| **Complexity** | O(1Kn) | O($KN^2$) | O(nlogn) | O(n) | O($N^2$) |
| **Method/Techniqueue** | KD trees | Image filtering/ discontinue supervising | Data partition | semi super secured cluster | Bottom up |
| **Advantages** | Computation is fast | Independent computation | Accelerate region queries | Take only single scan | Uses single linkage |
| **Disadvantages** | Difficult to get K value | Inappropriate window size | Quality depends on distance measure | - | Very sensitive outliers |

In this table we came to know about some parametric measures of some algorithms discussed above this paper. With this table we can compare with algorithm to find the efficient one for finding frequent item set data.

## VIII. EXPERIMENT RESULT AND EVALUATION

Tool used: **WEKA**

WEKA is a tool used in data mining for data preprocessing, clustering, classification, etc. here; we are using it for our clustering algorithms. First get environment and load any sample data set. It is very easy tool it will produce results just by selecting any of clustering algorithm available like K-means, then start it and you will get the result in output window. It will compute the clustering percentage. It also computing error in classification and show the confusion matrix.

1. K-MEANS RESULT

Number of iterations: 3

Within cluster sum of squared errors: 137.79496140158423

Initial starting points (random):

Cluster 0: 1,5.7,3.971739,3.913333,none,40,empl_contr,7.444444,4,nc

Cluster 1: 1,2,3.971739,3.913333,tc,40,ret_allw,4,0,no,11,generous,no

Missing values globally replaced with mean/mode

Final cluster centroids: Cluster#

| Attribute | Full Data | 0 | 1 |
|---|---|---|---|
| Duration | 2.1607 | 2.2533 | 1.6667 |
| Wage-increase-first-year | 3.8036 | 3.9834 | 2.8444 |
| Wage-increase-second-year | 3.9717 | 4.0209 | 3.7097 |
| Wage-increase-third-year | 3.9133 | 3.9511 | 3.7119 |
| Cost-of-living-adjustment | none | none | none |
| Working-hours | 38.0392 | 37.7541 | 39.5599 |
| Pension | empl_contr | empl_contr | none |
| Standby-pay | 7.4444 | 7.7431 | 5.8519 |
| Shift-differential | 4.871 | 5.2298 | 2.957 |
| Education-allowance | no | no | no |
| Statutory-holidays | 11.0943 | 11.237 | 10.3333 |
| Class | good | good | bad |

Time taken to build model (full training data): 0 seconds

Fig. 2 Experimental result of K-means algorithm.

Now in fig: 2 the result will be got as, Time taken to build model (full training data): 0.03 seconds. And the cluster instance got as, 57 (100%).

2. DENSITY BASED CLUSTERING RESULT

Cluster: 1 Prior probability: 0.1695

Attribute: duration

Normal Distribution. Mean = 1.6667 StdDev = 0.6667

Attribute: wage-increase-first-year

Normal Distribution. Mean = 2.8444 StdDev = 0.9335

Attribute: wage-increase-second-year

Normal Distribution. Mean = 3.7097 StdDev = 0.8495

Attribute: wage-increase-third-year

Normal Distribution. Mean = 3.7119 StdDev = 0.5699

Attribute: cost-of-living-adjustment

Discrete Estimator. Counts = 7 1 4 (Total = 12)

Attribute: working-hours

Normal Distribution. Mean = 39.5599 StdDev = 0.8234

Attribute: pension

Discrete Estimator. Counts = 7 3 2 (Total = 12)

Attribute: standby-pay

Normal Distribution. Mean = 5.8519 StdDev = 2.3171

Attribute: shift-differential

Normal Distribution. Mean = 2.957 StdDev = 1.6494

Attribute: education-allowance

Discrete Estimator. Counts = 3 8 (Total = 11)

Fig. 3 the result of density based clustering algorithm.

Here, in fig: 3, we have took 3 iterations and the result produced as with the prior probability of 0.8305 and the clustering transaction got with the 44 (75%). And we have discussed and got result of two algorithms.

## IX. CONCLUSIONS

In the growing technology data mining is playing a vital role. In this paper we have discussed some frequent item algorithms, association algorithm and some of the clustering algorithms. The existing algorithm alone would not able to tackle all the raising issues. This need to be improvised better for getting better results. That will leads to growth of new techniques.

## ACKNOWLEDGMENT

We express our deep gratitude to Dr. Pothula Sujatha our project guide for her enthusiastic encouragement and kind motivation throughout the period of study.

## REFERENCES

[1] Ghorbani, M., & Abessi, M. (2017). A New Methodology for Mining Frequent Itemsets on Temporal Data. *IEEE Transactions on Engineering Management*, *64*(4), 566–573. https://doi.org/10.1109/TEM.2017.

[2] Road, H., & Jose, S. (n.d.). Learn from Strange_chapter6.data.analysis.pdf, 487–499. https://doi.org/10.1007/BF02948845.

[3] Devi, A., & Sharma, S. (2017). Review on Analysis of Clustering Techniques in Data Mining, 36–39.

[4] J, S. S., & Pandya, S. (2016). An Overview of Partitioning Algorithms in Clustering Techniques. *International Journal of Advanced Research in Computer Engineering & Technology*, *5*(6), 2278–1323. https://doi.org/10.2135/cropsci2005.12.0453

[5] Zhu, E., & Ma, R. (2018). An effective partitional clustering algorithm based on new clustering validity index. *Applied Soft Computing Journal*, *71*, 608–621. https://doi.org/10.1016/j.asoc.2018.07.026

[6] Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research ISSN*, *13*(5), 899–906. https://doi.org/10.5120/888-1261

[7] Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia*, *78*(December 2015), 507–512. https://doi.org/10.1016/j.procs.2016.02.095

[8] Madhulatha, T. S. (2011). Comparison between K-Means and K-Medoids, 472–473.

[9] Nagpal, A., Jatain, A., & Gaur, D. (2013). Review based on data clustering algorithms, (April). https://doi.org/10.1109/CICT.2013.6558109

[10] Ng, R. T. (n.d.). NG_Han_1994.pdf, 144–155. https://doi.org/10.1.1.13.4395

[11] Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

[12] Comaniciu, D., Meer, P., & Member, S. (2002). MeanShif.pdf, *24*(5), 603–619. https://doi.org/10.1109/34.1000236

[13] Meng'Ao, L., Dongxue, M., Songyuan, G., & Shufen, L. (2015). Research and Improvement of DBSCAN Cluster Algorithm. *2015 7th International Conference on Information Technology in Medicine and Education (ITME)*, 537–540. https://doi.org/10.1109/ITME.2015.100

[14] Choi, W., & Engineering, M. S. (2008). Clustering Algorithm, (20080817), 3–6.

[15] Dash, R., & Misra, B. B. (2018). Performance analysis of clustering techniques over microarray data: A case study. *Physica A: Statistical Mechanics and Its Applications*, *493*, 162–176. https://doi.org/10.1016/j.physa.2017.10.032

[16] Berkhin, P. (n.d.). Survey of Clustering Data Mining Techniques, 1–56.

[17] Zhang, Q. (2010). A Grid Based Clustering Algorithm. *2010 International Conference on Computational Intelligence and Software Engineering*, 1–4. https://doi.org/10.1109/WICOM.2010.5600140

[18] Grid-based, S. I. (1997). Grid-based Clustering, 209–217.
[19] Shi Zhong, & Ghosh, J. (2003). Model-based clustering with soft balancing. *Third IEEE International Conference on Data Mining*, 459–466. https://doi.org/10.1109/ICDM.2003.1250953