ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Text Mining: Pattern Extraction and Classification (Data management and Distribution)

B Sankara Babu[1], Dr. K. Rajasekhara Rao [2]

[1]Assoc. Professor, Department of Computer Science & Engineering, Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad, Telangana, India

[2]Professor & Director, Department of Computer Science & Engineering, Sri Prakash College of Engineering, Tuni Andhra Pradesh, India

ABSTRACT: Data mining refers to the process of retrieving knowledge by discovering novel and relative patterns from large datasets. Clustering and Classification are two distinct phases in data mining that work to provide an established, proven structure from a voluminous collection of facts. In this paper, our focus is to analyze clusters of documents obtained via unsupervised clustering techniques and compare the performance of classification algorithms on the documents. Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster using the k-means algorithm. Classification is a task of assigning instances to predefined classes. We have a Training set containing data that have been previously categorized, and based on this Training set the algorithms finds the category that the new data points belongs to it using the secure hashing algorithm. K-means algorithm is used for classification and SHA-256 algorithm is used for protect the data securely in digital hash code.

KEYWORDS: Data mining; Clustering and Classification; K-means algorithm; SHA-256 algorithm.

I. INTRODUCTION

Text classification is the task of automatically classifying set of documents into categories from a predefined set. This task has several applications selective distribution of information to information consumers, spam filtering, and identification of document type. Automated text classification is good because it frees organizations from the need of manually organizing document. Text classification has gained importance because of the increased amount of documents over the years. Text documents should be categorized according to its contents.

Text classification has been used in many different areas of text mining and information retrieval. Initially it was used for improving the precision and recall in information retrieval systems and finding nearest neighbors of a document. Later it has also been used for organizing the results returned by a search engine and generating hierarchical clusters of documents. Initially we applied the term frequency and inverse document frequency methods on the data and found that the results were not very satisfactory and the main reason for this was the repeated data. This provided us the motivation for trying a better classification. We applied a distance method to cluster the nearest data points by k-means algorithm. We are also trying classification to remove duplicates using SHA algorithm.

The classification of the text is traditionally done by the term frequency and the inverse document frequency, this method has lot of problems. Term frequency cannot classify the documents in the correct category because the unimportant words may appear more number of times and the document may be classified to wrong category.

Ex: 1. **Apple** leads the world in innovation with iPhone, iPad, Mac, **Apple** Watch, iOS, OS X, watchOS and more.

2. Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

3. Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne on April 1, 1976, to develop and sell personal computers.

4. Apple is the world's largest information technology company by revenue, the world's largest technology company by total assets, and the world's second-largest mobile phone manufacturer.

5. If we take apple every day, it rebuilds our immune capacity.

In the above example, the topic is regarding Apple Company but not about apple fruit.

The aim of this paper is to classify the document in the correct category. That means the document should be classified into correct type of document to which it belongs.

II. LITERATURE SURVEY

The following are the various types of feasibility studies that can be undertaken.

A. Technical Feasibility

This is concerned with specifying the equipment's and the software to satisfy the user requirements. The technical needs of the system vary considerably but might include:

- The facility to produce outputs in a given time.
- Response time under certain conditions.
- Ability to process a certain volume of transactions at a specified speed.
- Facility to communicate data to a distant location.

Technical feasibility centers on the existing computer system, hardware, software etcetera and to what extent it can support the system. In examining the technical feasibility, the configuration of the system is given more importance than the actual hardware. The configuration should provide the complete picture of the system requirements, for example how many workstations are required and how these units are interconnected so that they would operate smoothly, etcetera. The result of the Technical Feasibility Study is the basis for the documents against which dealer and manufacturer can make bids. Specific hardware and software products can then be evaluated keeping in view the logical needs.

B. Economic Feasibility

Economic analysis is the most frequently used method for evaluating the effectiveness of a new system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. If benefits outweigh costs, then the decision is made to design and implement the system. It is not done to analyze the new system. Using a Gantt Chart schedule and part chart. We assumed that the benefit of the project is greater than the cost. So if we can develop the project easily then it is used for the evaluation of the proposed. We calculate the cost/benefit analysis and we assume that the benefit is feasible so we start developing the project. It is an analysis of the cost to be incurred in the system and benefits the derivable from the system. An economic Feasibility Study should demonstrate the net benefit of the proposed course of action in the context of direct and indirect benefits and costs to the organization and to the public as a whole. It should be required for both pilot and long-term activities, plans and projects.

C. Operational Feasibility

It determines how acceptable the software is within the organization. The evaluations must then determine the general attitude and skills. Such restriction of the job will be acceptable. To the users are enough to run the proposed budget, hence the system is supposed to the feasible regarding all except of feasibility. In operational feasibility, we attempt to ensure that every user can access the system easily. We develop a menu that users can easily access and we provide shortcut keys.

We show a proper error message when any mistakes are made in the program. We provide help and a guideline menu to help the user. Changes in the ways individuals are organized into groups may then be necessary and the groups may now compete for economic resources with the needs of stabilized ones by converting a number in a file in software.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

D. Behavioral Feasibility

Normal human psychology of human beings indicates that people are resistant to change and computers are known to facilitate change. Any project formulations should consider this factor also. Before the development of the Project titled "Delhi Metro", the need to study the feasibility of the successful execution of the project was felt and thus the following factors are considered for a Feasibility Study.

- Need Analysis.
- Provide the users information pertaining to the preceding requirement.

E. Feasibility study report

The result of the Feasibility Study provides us with the following facts:

- The automated system would increase the efficiency of the system.
- The automated system would increase customer's satisfaction.
- The automated system has many requirements such as Efficiency cost effectiveness, prompt service, Reliability.

F. Disadvantages

- Less number of features in previous system.
- Difficulty to get accurate item set.
- Requires more manual effort.

G. Proposal System

In Proposed System, propose an alternative protocol for the secure computation of the union of private subsets. The proposed protocol improves upon that in terms of simplicity and efficiency as well as privacy. In particular, our protocol does not depend on commutative encryption and oblivious transfer (what simplifies it significantly and contributes towards much reduced communication and computational costs). While our solution is still not perfectly secure, it leaks excess information only to a small number (three) of possible coalitions, unlike the protocol of that discloses information also to some single players. In addition, we claim that the excess information that our protocol may leak is less sensitive than the excess information leaked by the protocol.

H. Advantages in Proposal system

- As a rising subject, data mining is playing an increasingly important role in the decision support activity of every walk of life.
- Get Efficient Item set result based on the customer request.
- Reducing Duplication by using k-Means Algorithm.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

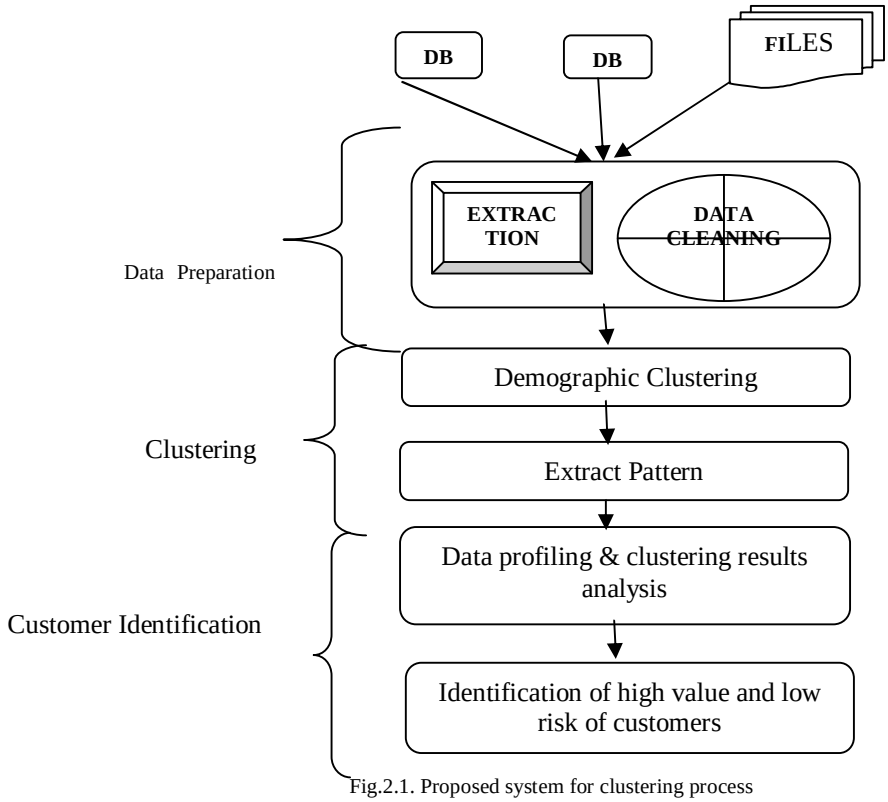

Fig.2.1. Proposed system for clustering process

Fig.2.1 explains about clustering process and pattern extraction.

III. SECURED HASH ALGORITHM (SHA-256)

SHA-256 is one of the successor hash functions to SHA-1 (collectively referred to as SHA-2), and is one of the strongest hash functions available. While SHA-1 has not been compromised in real-world conditions, SHA-256 is not much more complex to code, and has not yet been compromised in any way. The 256-bit key makes it a good partner-function for AES. It is defined in the NIST (National Institute of Standards and Technology) standard 'FIPS 180-2. NIST also provide a number of test vectors to verify correctness of implementation .A cryptographic hash (sometimes called 'digest') is a kind of 'signature' for a text or a data file. SHA1 generates an almost-unique 160-bit (20-byte) signature for a text A hash is not 'encryption' – it cannot be decrypted back to the original text (it is a 'one-way' cryptographic function, and is a fixed size for any size of source text). This makes it suitable when it is appropriate to compare 'hashed' versions of texts, as opposed to decrypting the text to obtain the original version. Such applications include stored passwords, challenge handshake authentication, and digital signatures.

Descriptions of SHA-256, SHA-384, and SHA-512: An n-bit hash is a map from arbitrary length messages to n-bit hash values. An n- bit cryptographic hash is an n-bit hash which is one-way and collision-resistant, such functions are important cryptographic primitives used for such things as digital signatures and password protection. Current popular hashes produce hash values of length n = 128 (MD4 and MD5) and n = 160 (SHA-1), and therefore can provide no more than 64 or 80 bits of security respectively..

Computation of the hash of a message begins by preparing the message: 1. Pad the message in the usual way: Suppose the length of the message M, in bits, is `. Append the bit \1" to the end of the message, and then k zero bits, where k is the smallest non-negative solution to the equation `+1+k 448 mod 512. To this append the 64-bit block which is equal to the number ` written in binary. For example, the (8-bit ASCII) message \ abs" has length 8*3 = 24 so it is padded with a one, then 448 (24 + 1) = 423 zero bits, and then its length to become the 512-bit padded message 01100001

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

01100010 01100011 1 00 0 | {z } 423 00 011000 | {z } 64 : The length of the padded message should now be a multiple of 512 bits. 2. Parse the message into N 512-bit blocks M(1); M(2) ;:::;M(N) .

The hash computation proceeds as follows: For i = 1 to N (N = number of blocks in the padded message) f Initialize registers a; b; c; d; e; f ; g; h with the (i 1)st intermediate hash value (= the initial hash value when i = 1) a H(i1) 1 b H(i1) 2 . . . h H(i1) 8 Apply the SHA-256 compression function to update registers a; b; : : : ; h For j = 0 to 63 f Compute Ch(e; f ; g), M aj(a; b; c), 0(a), 1(e), and Wj (see Definitions below) T1 h + 1(e) + Ch(e; f ; g) + Kj + Wj T2 0(a) + M aj(a; b; c) h g g f f e e d + T1 d c c b b a a T1 + T2 g Compute the ith intermediate hash value H(i) H(i) 1 a + H(i1) 1 H(i) 2 b + H(i1) 2 . . . H(i) 8 h + H(i1) 8 g H(N) = (H(N) 1 ; H(N) 2 ;:::; H(N) 8) is the hash of M.

A. K – Means Algorithm

Clustering is the process of partitioning a group of data points into a small number of clusters. For instance, the items in a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have n data points xi,i=1...n that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the square of the distance from the data points to the cluster. K-means clustering solves

$$\mathbf{argmin}c\sum_{i=1}^{k}\sum_{x\in ci} d(x,\mu i)2=\mathbf{argmin}c\sum_{i=1}^{k}\sum_{x\in ci}\|x-\mu i\|22$$

where ci is the set of points that belong to cluster i. The K-means clustering uses the Euclidean distance $d(x,\mu i)=\|x-\mu i\|2$. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters	μi= some value ,i=1,...,k
2. Attribute the closest cluster to each data point	$ci=\{j:d(xj,\mu i)\leq d(xj,\mu l),l\neq i,j=1,...,n\}$
3. Set the position of each cluster to the mean of all data points belonging to that cluster	$\mu i=1\|ci\|\sum j\in cixj,\forall i$
4. Repeat steps 2-3 until convergence	
Notation	|c|= number of elements in c

The algorithm eventually converges to a point, although it is not necessarily the minimum of the sum of squares. That is because the problem is non-convex and the algorithm is just a heuristic, converging to a local minimum. The algorithm stops when the assignments do not change from one iteration to the next.

IV. IMPLEMENTATION ISSUES AND RESULTS

Implementation is the most crucial stage in achieving a successful system and giving the user's confidence that the new system is workable and effective. Implementation of a modified application to replace an existing one. This type of conversation is relatively easy to handle, provide there are no major changes in the system. Each program is tested individually at the time of development using the data and has verified that this program linked together in the way specified in the programs specification, the computer system and its environment is tested to the satisfaction of the user. The system that has been developed is accepted and proved to be satisfactory for the user. And so the system is going to be implemented very soon. A simple operating procedure is included so that the user can understand the different functions clearly and quickly.

Initially as a first step the executable form of the application is to be created and loaded in the common server machine which is accessible to the entire user and the server is to be connected to a network. The final stage is to document the entire system which provides components and the operating procedures of the system. Implementation is the stage of

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. Implementation is the process of converting a new system design into operation. It is the phase that focuses on user training, site preparation and file conversion for installing a candidate system. The important factor that should be considered here is that the conversion should not disrupt the functioning of the organization.

Here itself we use socket programming as most. For that purpose we use a socket listener class which will listen the request comes from the client to make communication in between. Before sending the request to the server we stream the data by using a stream function. After that we need to break the data into no of part by using a split function. After that split data need to send to the corresponding.

At the end point we need to implement a buffer which needs to be collecting the split packet in a list having address field and data field. Value field store the data and address field store the address of next split. If one split in the order we don't get means some –ve ack signal need to send. For that purpose once again we need to call socket listener class and create object for that class and send the request in a form of stream. And client need to resend that corresponding packet once again .This process is repeated process for that purposes we need some source code of data.

Fig.4.1. Clustering Page

The above screenshot fig 4.1 describes about the clustering page where the user can browse and upload files. Using the clustering page we can upload up to 4 files to text documents. After the adding we can add no. of clusters. It gives the resulting clusters beside of the page It performs certain operations like find clusters, , browse files, and upload files, no. of clusters.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Fig.4.2. Clustering output

The above screenshot fig 4.2 describes about the clustering page where the user can browse and upload files. Using the clustering page we can upload up to 4 files to text documents. After the adding we can add no. of clusters and the clusters are formed with the similar type of data it gives the resulting clusters beside of the page it performs certain operations like find clusters, , browse files, and upload files, no. of clusters.

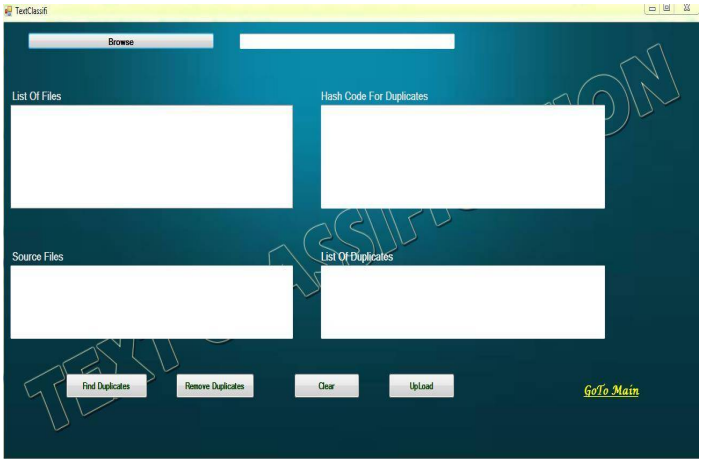

Fig.4.3. Classification page

The above screenshot fig 4.3 describes about the classification page where the user can browse and upload files. Using the classification page we can remove the duplicate files and store original files. It performs certain operations like find duplicates, remove duplicates, browse files, upload files, shows hash code with duplicates.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

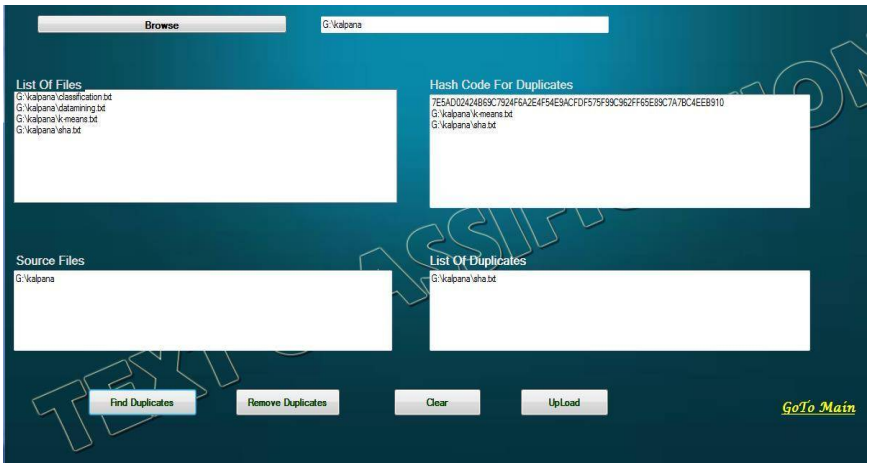

Fig.4.4. Classification upload page

The above screenshot fig 4.4 describes about the classification page where the user can browse and upload files. Using the classification page we can remove the duplicate files and store original files.

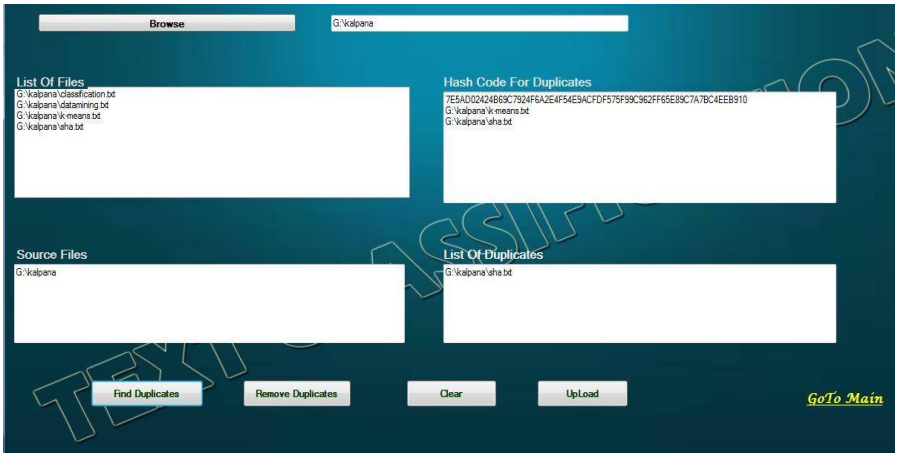

Fig.4.5. Classification Output page

The above fig 4.5 screenshot describes about the classification page where the user can browse and upload files. Using the classification page we can remove the duplicate files and store original files.

V. CONCLUSION AND FUTURE SCOPE

Data mining has come to prominence over the last two decades as a discipline in its own right which offers benefits with respect to many domains, both commercial and academic. Broadly data mining can be viewed as a application domain, as opposed to a technology. The increasing ability of institutions to collect electronic data, facilitated by advanced in computer processing, means that the desire to “mine” data is likely to expend.
The traditional text categorization is done by term frequency and that process is not enough. The other techniques should be used, the clustering and classification are introduced. The clustering helps to find the similar documents and forms into clusters with the similar type of documents using the k-means algorithm. Classification is used to remove

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

the duplicate files with the help of hash code by using the secure hashing algorithm duplicate files can be removed and original files can be stored and it classifies the documents accurately.

In terms of future scope, a variety of data mining techniques can be used by researchers to simplify customer perceptions and attitudes. Every day, every hour and every minute, terabytes of data gets generated from millions of shoppers, yet, retail managers/ business executives always grapple with relevant information that can help retailers/ researchers design strategies to generate customer loyalty.. Thus data mining can not only be applied in retailing but also can be applied in the other sectors such as banking, medicine, education, tourism, insurance and so on.

In terms of managerial and technical approach, researchers can research certain niche customer segments such as the elderly, only students, only male professionals etc. Additional sectors, such as apparel retailing, fashion products, consumer electronics, luxury brands, mobile retailing etc. can be researched. For classification of document in a correct category the Distributional Features are to be studied in more detailed manner. Along with the combination of measures like Compactness and first appearance last appearance the idf (inverse document frequency) also should be explored and idf should be used for further research and results should be analyzed.

REFERENCES

[1] Agrawal, R, Imielinski, T. and Swami, A. “Mining association rules between sets of items in large databases.” Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD’93), ACM Press, pp207-216, 1993
[2] Breiman, L., Friedman, Y., Olshen, R. and Stone, C. “Classification and Regression Trees.” Wadsworth, Belmont, CA, 1984.
[3] Fayyad, U., Piatetsky-Shapiro, H. and Smyth, P. “The KDD process for extracting useful knowledge from volumes of data.” Communications of the ACM, 39 (11), pp27 – 34, 1996
[4] Han, J., Pei, J., and Yin, Y. “Mining frequent patterns without candidate generation. “Proc. ACM SIGMOD Conference on Management of Data (SIGMOD ’00). ACM Press, pp1-12, 2000
[5] Hand, D.J. and Yu, K. “Idiot’s Bayes: Not So Stupid After All? Internet” Statist. Rev. 69, pp385- 398, 2001
[6] Hastie, T. and Tibshirani, R. “Discriminant Adaptive Nearest Neighbor Classification. “ IEEE Transaction on Pattern Analysis and Machibe Intelligence, 18(6), pp607-616, 1996
[7] Liu, B., Hsu, W. and Ma, Y. M. “Integrating classification and association rule mining.” Proc KDD-98, ACM press, pp.80-86, 1998
[8] MacQueen, J. B. “Some methods for classification and analysis of multivariate observations.” Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, University of California Press, pp281- 297, 1967
[9] Quinlan, J. R. C4.5: “Programs for Machine Learning.” Morgan Kaufmann Publishers Inc. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, 1993
[10] Yan, X. and Han, J. (2002). “gSpan: Graph-Based Substructure Pattern Mining.” Proc. IEEE International Conference on Data Mining (ICDM ’02), IEEE, pp721-724, 2002
[11] Zhang, T., Ramakrishnan, R., and Livny, M. BIRCH: “an efficient data clustering method for very large databases.” Proc. ACM SIGMOD international Conference on Management of Data, ACM Press, 1996
[12] Pham, D.T. and Afify, A.A. “Clustering techniques and their applications in engineering”. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2006
[13] A.K. Jain, M.N. Murty, and P. J. Flynn “Data clustering: a review”, ACM Computing Surveys (CSUR), Vol.31, Issue 3,, 1999.
[14] Grabmeier, J. and Rudolph, A. “Techniques of cluster algorithms in data mining” Data Mining and Knowledge Discovery, 6, 303-360, 2002
[15] Han, J. and Kamber, M. “Data Mining: Concepts and Techniques”, (Academic Press, San Diego, California, USA), 2001
[16] Jiyuan An , Jeffrey Xu Yu , Chotirat Ann Ratanamahatana , Yi-Ping Phoebe Chen “A dimensionality reduction algorithm and its application for interactive visualization”, Journal of Visual Languages and Computing, v.18 n.1, , February p.48-70, 2007