International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 1, Feb 2013, 47-58

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

IMPLEMENTATION OF DATA MINING IN ONLINE SHOPPING SYSTEM USING TANAGRA TOOL

VISHAL JAIN[1], GAGANDEEP SINGH NARULA[2] & MAYANK SINGH[3]

[1,3]Lingaya's University, Faridabad, Haryana, India

[2]G.T.B.I.T, New Delhi, India

ABSTRACT

Data Mining is a technology that is used for identifying patterns and ways from large quantities of data or other repositories. This technology works in a way that it adopts data integration method to generate Data Warehouse. Then with the help of algorithm, it extracts useful information. Data Mining is powerful technology that is widely used in various applications like E-Commerce, Educational System, Remote Sensing, Online shopping system etc. Here we deal with Online shopping processes i.e. it is concerned with developing new methods to discover knowledge from online store database. Database is taken from any online shopping site. Since large amount of data is available in Online Shopping System, there is need to collect appropriate data which employs use of various data mining technologies. In this paper, we put light on analyzing reviews of customers' purchasing different items of different brands. Outcomes of results are presented by analysis of TANAGRA tool

KEYWORDS: Data Mining, Classification, Clustering, Association Rule, TANAGRA

INTRODUCTION

A huge collection of data is present on any Online Shopping website that includes information about various products belonging to different categories. Analyzing them manually can lead to wastage of time and thus in order to save time and improves accuracy, we have used concept of data mining in Online Shopping System. In Online Shopping System, we are given database of many customers with their corresponding products purchased; we could identify between loyalty customers and normal customers. All this can be done through various data mining tasks like Classification, Association Rule etc. These tasks execute transactions in shopping database automatically in less time. These tasks help in identifying customer behaviors, improve customer quality service and provide good transportation facilities.

Data Mining is also proven useful in field of Pharmacy. It helps in DNA Analysis. In this, data mining tasks compare frequently genes patterns of patients and uses visualization effects to evaluate results.

Data Mining System architecture has following components described as below:-

- **Data Source or Database**: - It may be data source, database and word documents of Online Database System. In database, data may be cleansed, updated and modified.
- **Database Server**: - It fetches instructions or data from given database according to customer request.
- **Data Mining Engine**: - It is most important component of Data Mining System. It contains modules or algorithms for performing mining tasks like Classification, Association Rule, and Clustering.
- **Local Model Generator**: - After using several algorithms for classification and clustering, this component generates local model consisting of patterns and modules related to customer's query.

- **Final Model Generator**: - Depending on implementation method used for generation of local model, it provides interaction among customers and system by specifying data mining query or task. It leads to interpretation of data mining results evaluated to customers. It uses various visualization and GUI strategies at this step. All these components are arranged in a diagram termed as Data Mining Framework System as shown in figure 1.

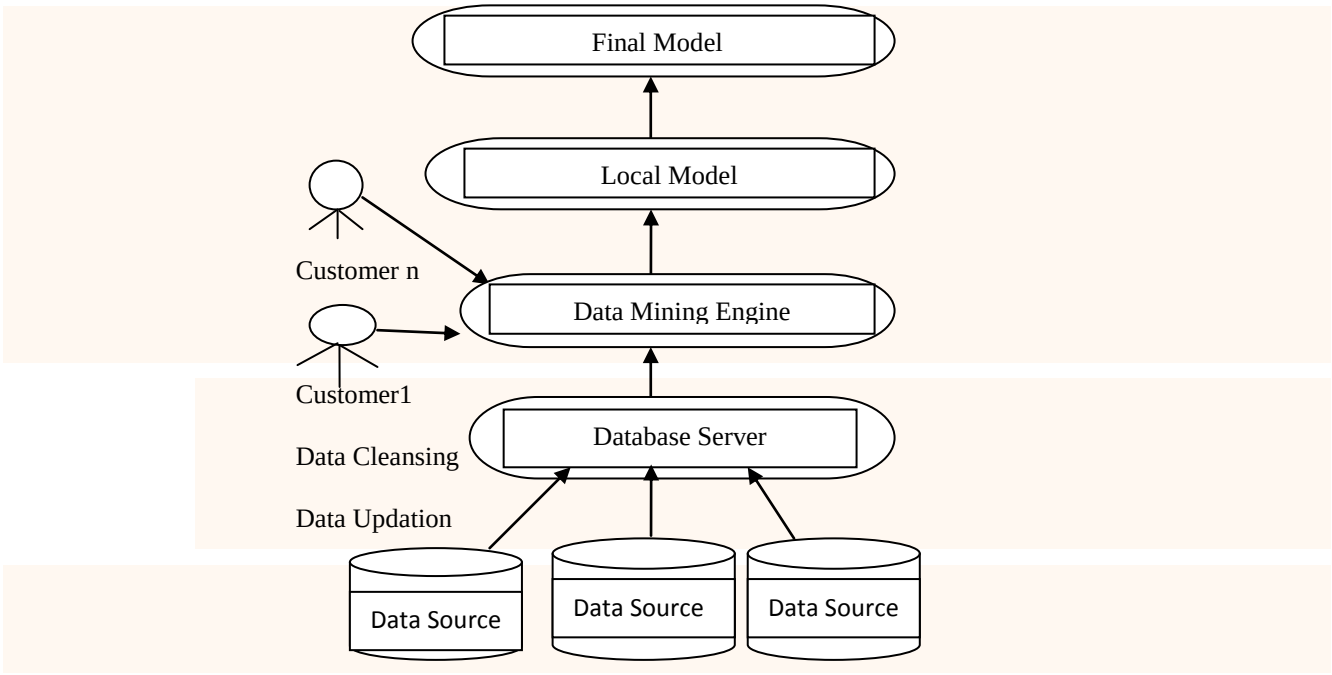


Figure 1: Data Mining Framework [1]

APPLICATION OF DATA MINING

Data Mining is widely used in various areas like Credit Card Fraud Detection System, Health Insurance, Security System and Sensor Management, Online shopping system, Distributed Clustering and many more. In this section, we will use data mining in online shopping system as providing best deals and offers to customers, relationship between customers viewing products and purchasing products. It also deals with classification of customers on basis of their reviews about purchased products.

As there are huge number of products available online and to choose required product from them along with its relationship, we have use different technologies. Some of Data Mining technologies are listed below:

- Association Rule
- Classification of various products
- Clustering

Association Rule

Represented as: - "If Then" rules. If is called Antecedent, Then is called consequent [2].

They are used to show relationship among various data items. To implement these rules, we have one of Data Mining algorithm called APRIORI algorithm [3]. This algorithm is one of finest approaches to find frequent item sets from transaction database and derive association rules. If item sets are obtained, then they are used to generate association rule. Support (S) and Confidence (C) are normal methods used to generate association rules. Support for S….T is defined as

percent of transaction in database that contains combination of S and T. Confidence for S…..T is defined as ratio of percent of transaction containing S and T to the percent of transaction containing S only.

Apriori algorithm works as follows:

Let frequent item sets (item sets that have minimum support) = Fk containing concepts Ck where size of item sets=k

- It first scans the database and searches for frequent item sets and count for each item.
- Then, it compares item sets with minimum support required.
- It then repeats the following steps to extract all item sets.
- Generate Ck+1 candidate of frequent item sets of size K+1. Note these item sets have been generated from item sets of size k.
- Scan the database as above.
- Add the item sets that satisfies minimum support requirement.

Implementation of Apriori algorithm is shown below. Consider a dataset with following items as shown below:

Assume minimum support = 2(i.e. minimum support = 2/5 = 40%).

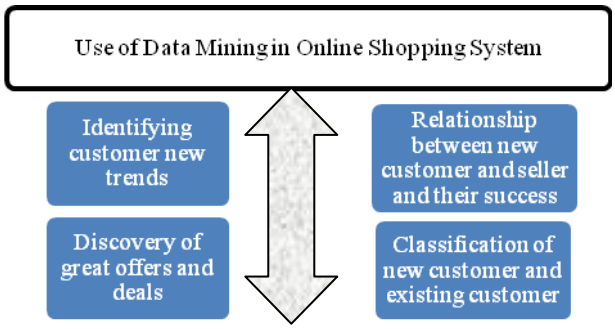
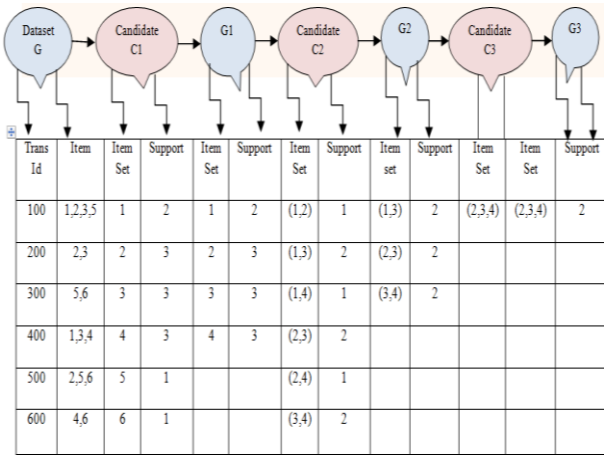


Figure 2: Benefits of Data Mining



Trans Id	Item	Item Set	Support	Item Set	Support	Item Set	Support	Item set	Support	Item Set	Item Set	Support
100	1,2,3,5	1	2	1	2	(1,2)	1	(1,3)	2	(2,3,4)	(2,3,4)	2
200	2,3	2	3	2	3	(1,3)	2	(2,3)	2			
300	5,6	3	3	3	3	(1,4)	1	(3,4)	2			
400	1,3,4	4	3	4	3	(2,3)	2					
500	2,5,6	5	1			(2,4)	1					
600	4,6	6	1			(3,4)	2					

Figure 3: Implementation of Apriori Algorithm

Classification

Here we have to classify our products according to customer reviews and feedback. Classification is defined as data mining task that maps data into groups and classes. *It is supervised technique.* In this technology, we will design our model.

Model Construction

It has set of predefined classes and each record is assured to belong to predefined class. Data Mining is done by use of data set which follows classification algorithms to generate classification rule.

Clustering

It is defined as technology to extract data on basis of groups. It arranges similar objects in one group and different objects in other group. *It is unsupervised technique.* In case of online shopping system, cluster has been used to group customer according to their reviews like in our online database clustering can be used to group those customers who have same reviews about different products.

Clustering in TANAGRA requires use of various algorithms like *CT (Clustering Tree), K-Means, EM-Clustering* and many more. Some of the results of algorithms are implemented in further section

Tanagra Tool

Tanagra: Tool for Research purposes in Data Mining

Author: Ricco Rakotomalala

Version: 1.4.45

Tanagra is free, open source, user friendly software developed for students and researchers to mine their data. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and database areas. It is programmed in Pascal language. It has various components like Data Visualization, Statistics, Clustering, Association, Factorial Analysis and many more. With the help of TANAGRA, we can visualize our data. Data Visualization includes Viewing Dataset, plotting values on graphs and scatter plot. It is used to show relationship among attributes in 2D axes. TANAGRA includes basic clustering algorithm like K-Means, EM- Clustering etc. Our database is shown in table 1.

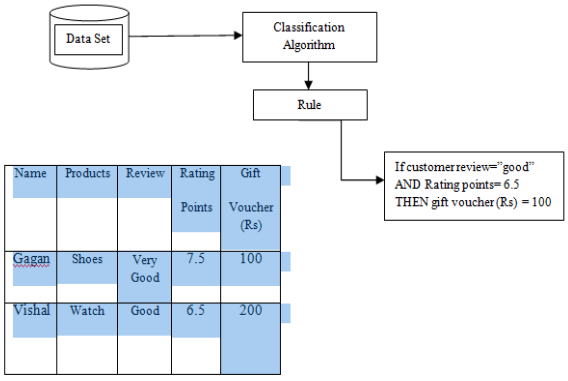


Figure 4: Model Construction

Table 1: Customer Dataset

Name of Customer	Product	Brand	Price (Rs)	Discount %	Customer Reviews	Rating Points	Gift Voucher (Rs)
Sunil	Shoes	Adidas	3500	20	Good	6.5	250
Gagan	Laptop	Hp	27500	15	Very good	8.5	450
Vishal	Watch	Diesel	4500	10	Good	6.5	250
Rajiv	Phone	Apple	45000	NA	Excellent	9.5	1000
Tarun	Bag	Reebok	2000	10	Average	5.5	150

Importing Data to Tanagra

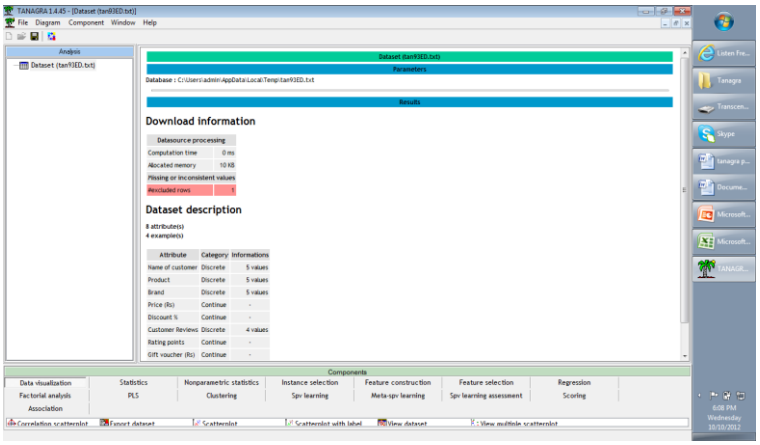


Figure 5: Tanagra 1.4.45 with Customer Dataset

Types of Attributes in Tanagra

- Discrete attributes: - Attributes having Non numerical values ex. name.
- Continuous attributes: - Attributes having numerical values ex. Roll no, age.
- Target attributes: - Attribute on which result is found (i.e. axis attributes).
- Input attributes: - Resulting values on the graph.

Data Visualization in Tanagra

Data Visualization includes:

- View dataset
- Scatter plot with Label
- Correlation Scatter plot
- View Multiple Scatter plot

Here we have used "*Scatter plot with label*" function to show relationship between two variables namely Price (Rs) and rating points and analyzed them using Tanagra. For using this function, a precondition is fulfilled:

- Two or more continuous attributes must be available in dataset.
- There should be at least one discrete attribute for showing identification tag from a discrete attribute.

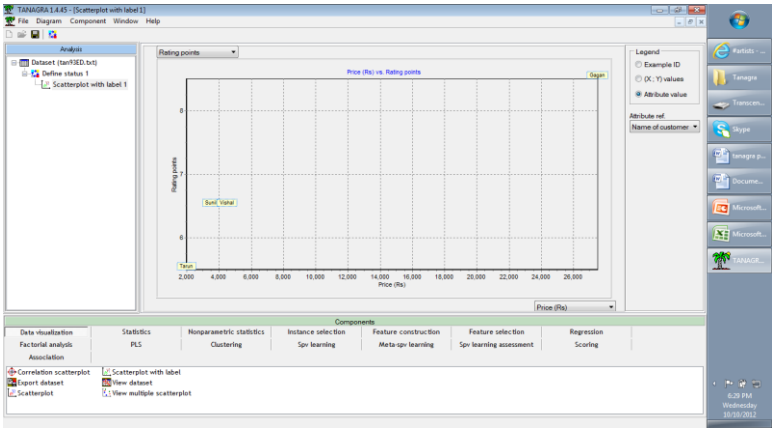


Figure 6: Scatter Plot with Label

Another function defined under Data Visualization is "Correlation Scatter plot". Although both functions perform same task of showing relationship among variables. But Correlation Scatter plot is better due to its following aspects:

- It is used when customer wants to compare its results statistically. E.g. if customer wants to get details of discount (%) available on all deals without mentioning product specifications.
- This function shows relationship between continuous attributes only which are chosen as Input and Target attributes on basis of predetermined condition.
- There must be at least three continuous attributes available in dataset.

Here Price (Rs) and Discount (%) are Input attributes whereas Gift Voucher (Rs) and Rating Points are Target attributes. This function visualizes Target attributes with Input attributes on scatter plot.

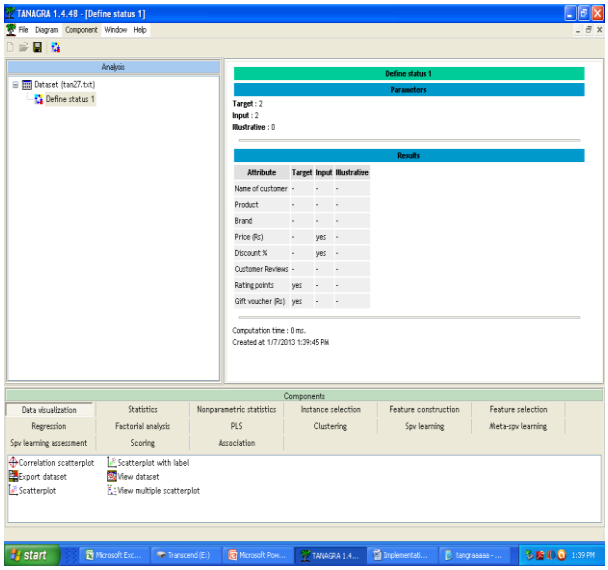


Figure 7: Defining Customer Dataset for Correlation Scatter Plot Function

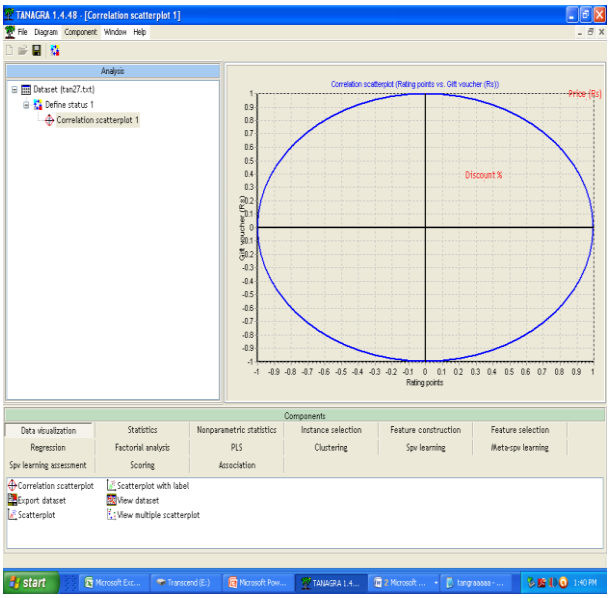


Figure 8: Results of Correlation Scatter Plot Function

It is seen from above screen shot that Discount (%) is computed using values of Gift voucher (Rs) and Rating Points. Less costly Gift vouchers (Rs) have more Rating Points.

Clustering in Tanagra

We have defined status 2 for implementing concept of Clustering and its algorithms. While defining status for clustering, it must be noted to take care of attributes because they are different in Data Visualization. There are many algorithms included in Clustering technique as follows:

- EM- Clustering
- Clustering tree (CT)
- K-Means
- Hierarchical Clustering Analysis (HCA)
- Kohonen – SOM
- Learning Vector Quantizer (LVQ)
- VARKMeans

Here, we have shown one of clustering algorithms named Clustering Tree (CT) using this tool. While performing clustering, our Input attributes are Name of Customer, Product and Brand whereas Target attributes are Price (in Rs), Discount (%) and Rating points.

Clustering Tree which is written as CT is one of Clustering Algorithms. It can handle more than one continuous attributes. It generates a tree classifying customers according to their reviews like in our database Sunil and Vishal have same reviews so they are categorized in one group.

For implementing CT algorithm, conditions should be fulfilled:

- Two or more continuous attributes must be available in dataset.
- Out of these continuous attributes, one must be at least one target attribute and one input attribute.

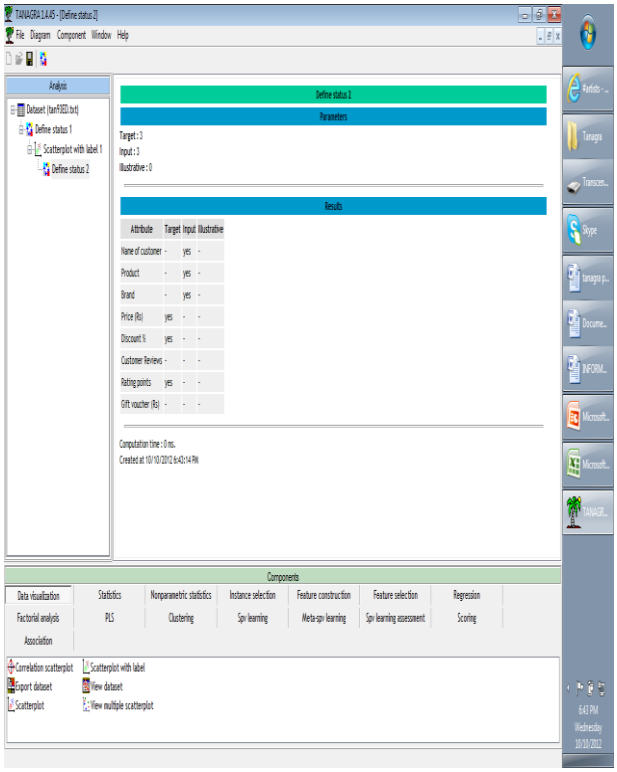


Figure 9: Defining Customer Dataset 2

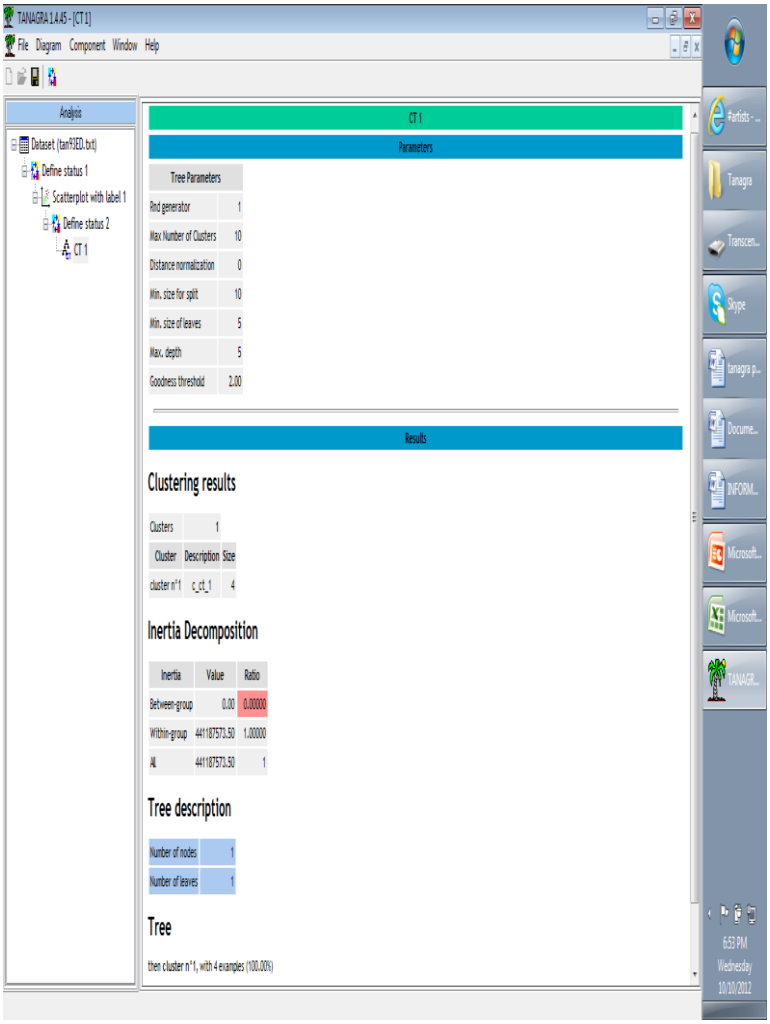


Figure 10: Results of CT Algorithm

Association in Tanagra

Association Rule learning is defined as method or procedure that is conducted to discover relations among variables in a given database. We have already discussed Association Rule in Section2 of this paper.

One of the finest algorithm of Data Mining which is called as Apriori algorithm has control on association rule. The practical implementation of this algorithm is shown with the help of TANAGRA. Association includes:

- A priori
- A priori MR
- A priori PT
- Spv Assoc Rule
- Spv Assoc Tree

Here we have used “A priori” function for performing association whose precondition must be fulfilled.

- There must be at least two discrete attributes available in dataset.
- It can also take into account binary 0/1 continuous attributes.

While performing association, our discrete attributes may be Name of Customer, Product, Brand and Customer Reviews.

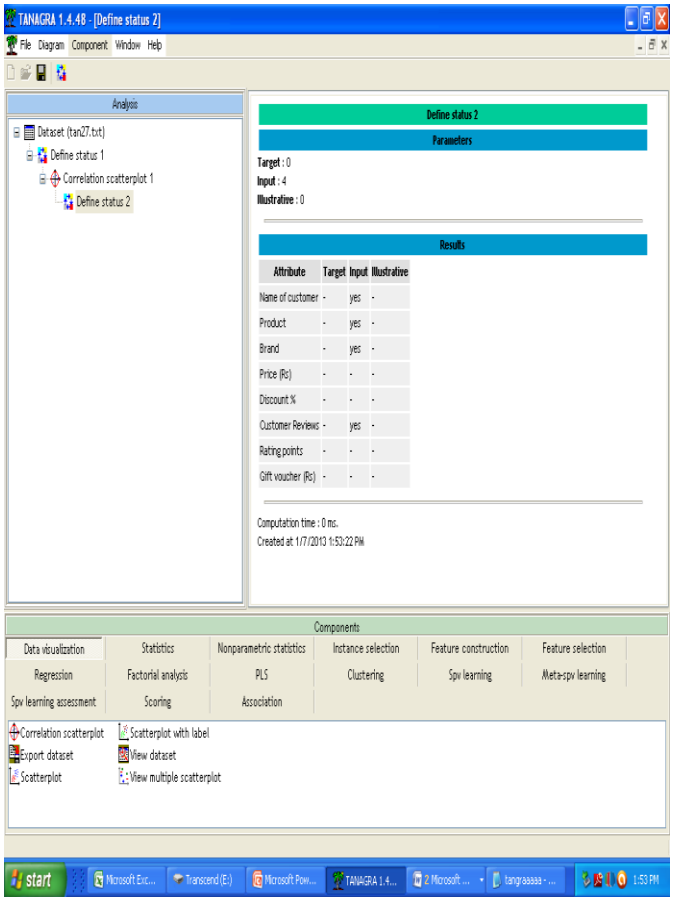


Figure 11: Defining Customer Dataset for Association Function

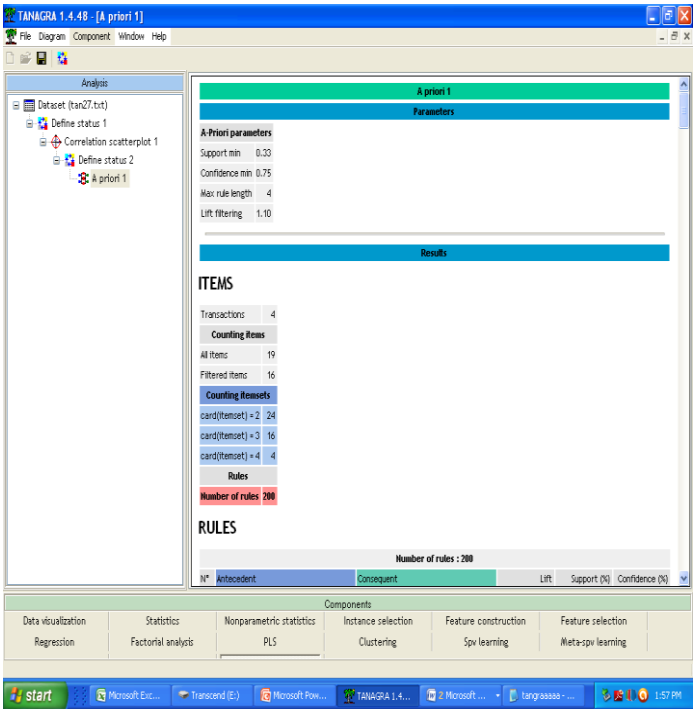


Figure 12: Results of a Priori Function

Use of TANAGRA determines A priori parameters without using any mathematical formulae. Parameters include Min Support, Min Confidence, and Rule length. It is evident from above screen shot.

CONCLUSIONS

This paper highlights the brief introduction about Data Mining and its various applications. Since data mining is a technology that has its roots in various applications like Remote Sensing, Security Systems, Distributed Clustering and many more. We are concerned about application of data mining in an Online Shopping System where we are performing various mining tasks like Association Rule, Clustering and Classification on customer dataset.

It describes Association Rule with the help of Apriori algorithm that is used for generating frequent item sets and describes relationship among customers and sellers and their success. We applied data mining techniques to discover knowledge. This paper also revolves around the concept of data mining tool which is used for research purposes named TANAGRA. We have used version 1.4.45 here. It is free, open source and user friendly software developed for students and researchers for mining data. This software helps to visualize customer dataset and classifies different customers according to their reviews regarding products. TANAGRA uses basic clustering algorithms like CT, K-Means, and EM Clustering etc for differentiating between various customers. Each one of this knowledge can be used to identify customer records.

REFERENCES

1. Accessible from Mr.Lobo L.M.R.J, Sunita B Aher, “Data Mining in Educational System using WEKA”, “International Conference on Emerging Technology Trends (ICETT)”, 2001.
2. Accessible from Sonali Agarwal, Neera Singh, Dr. G.N. Pandey, “Implementation of Data Mining and Data Warehouse in E-Governance”, “International Journal of Computer Applications (IJCA) (0975-8887), Vol.9-No.4,” November 2010.
3. Sheikh, L Tanveer B. and Hamdani, "Interesting Measures for Mining Association Rules", “IEEE-INMIC Conference”, December 2004.
4. Parthasarathy, S.Zaki, Li.W, “Parallel Data Mining for association rules”, “Knowledge and Information systems”, pages 1-29.
5. Mrs. P. Nancy, Dr. R. Geetha Ramani, “A Comparison on Data Mining Algorithms in Classification of Social Network Data”, “International Journal of Computer Applications (IJCA) (0975-8887) Vol.32-No.8”, October 2011.
6. Parthasarathy, S.Zaki, “Clustering homogenous datasets”, “In PDKK”, pages 566-574.
7. Bharati M Ramager, “Data Mining techniques and Applications”, “International Journal of Computer Science and Engineering Vol. 8”, December 2009.
8. Sunita B Aher, Mr. LOBO L.M.R.J, “Data Mining in Educational System using WEKA”, “International Conference on Emerging Technology Trends (ICETT) 2011, Proceedings published by International Journal of Computer Applications (IJCA).
9. Kifaya, “Evaluation using Associative Classification and Clustering”, “Communications of the IBIMA vol. 11 IISN “, pages 1943-7765.
10. Rakesh Aggarwal, Tomasz Imielinski, Arun Swami, “Mining Association Rules between Sets of Items in Large Databases”, “In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data”, pages 654-659.

11. N Sivaram, K.Ramar, “Applicability of Clustering and Classification Algorithms for Recruitment Data Mining”, “International Journal of Computer Applications (IJCA), Vol.4, No-5”, March 2010.
12. Ferenc Bodon, “A fast Apriori implementation”, “In Proceedings of the IEEE ICDM Workshop on Frequent Item set Mining Implementations”, 2003.
13. Tan Pang-Ning, “An Introduction to Data Mining”. “Pearson Education”, 2007.
14. U.K. Pandey, and S.Pal, “Data Mining: A Prediction of performer using Classification”, “International Journal of Computer Science and Information Technology (IJCSIT), Vol.2(2), IISN:0975-9646”, pages 686-690, 2011.
15. K Saravana Kumar, R. Manicka Chezian, “A Survey on Association Rule Mining using Apriori Algorithm”, “International Journal of Computer Applications (IJCA), Vol.45-Number 5”, 2012.
16. Amirmahadi Mohammadighavam, Neda Rajabpour, Ali Naserasadi, “A Survey on Data Mining Approaches”, “International Journal of Computer Applications (IJCA) (0975-8887) Vol.36-Number 6”, 2011.
17. Dhanashree S. Deshpande, “A Survey on Web Data Mining Applications”, “International Journal of Computer Applications (IJCA), ETCSIT- Number 3”, 2012.