Mining of Frequent Item sets using Association Rule Mining in WEKA Environment

S.Suriya[#1], Dr.S.P.Shantharajah[#2], R.Deepalakshmi[#3]

[#1,#3]*Assistant Professor, Department of Computer Science and Engineering,*
Velammal College of Engineering and Technology, Madurai, Tamilnadu, India.
suriyas84@gmail.com,jei.deepa@gmail.com

[#2] *Professor, Department of Master of Computer Applications,*
Sona College of Technology, Salem, Tamilnadu, India.
spshantharaj@gmail.com

Abstract— **Knowledge Mining involves retrieval of potentially useful information from huge repository of data. Over a decade, research activities are carried out for identifying an effective algorithm for frequent item set mining. Apriori algorithm, a realization of frequent pattern matching, is universally adopted for reliable mining. It is based on parameters namely support and confidence. WEKA toolkit is used for experimental tool for the input datasets of the proposed algorithm. The experimental results are obtained for a Stock Market application to improve the business revenue.**

Keywords — **Association Rule Mining; Apriori; WEKA.**

I. INTRODUCTION

Knowledge mining is analyzing process of carried out over huge collection of data to identify the required information. It refers [1] to extraction of potentially useful information from large databases. The basic activities of knowledge mining are namely Data Cleaning, Data Selection, Data Transformation, Data Mining, Data Evaluation and Presentation of retrieved Knowledge (refer Fig 1: Data Mining Process). Data Cleaning involves removal of errors and inconsistencies in the data collection (from different sources) to improve the quality of data. Data Selection is concerned with picking out the relevant part for future processing. Data Transformation deals with transforming the data to appropriate forms of mining using techniques like smoothing, aggregation and normalization, etc. Data mining involves identification of interesting patterns. Data Evaluation and presentation helps in visualization of results.

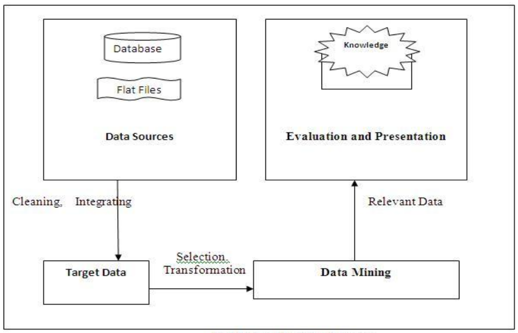


Fig 1: Data Mining Process

II. ASSOCIATION RULE MINING

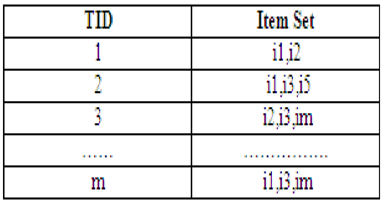
Association rule of data mining involves picking out the unknown inter-dependence of the data and finding out the rules between those items [2]. Agrawal introduced association rules for point of scale (POS) systems in supermarkets. A rule is defined as an implication of the form A=>B where A∩ B≠Ǿ. The left-hand side of the rule is called as antecedent. The right-hand side of the rule is called as consequent. For example [3] the rule { Onions, Potatoes}=>{beef} found in the sales data of a supermarket would indicate that if a customer buys Onions and potatoes together then the customer is likely to buy beef also. Such information is useful to make decisions about marketing activities. Association rules are also used in many applications including Web usage Mining, Intrusion Detection and Bio-informatics.

Support : I = { i1,i2,i3, … , im} is a collection of items. T be a collection of transactions associated with the items (refer Table1). Every transaction has an identifier TID [4]. Association rule A=>B is such that AЄI, BЄI. A is called as Premise and B is called as Conclusion. The support ,S, is defined as the proportion of transactions in the data set which contains the itemset.

Support(X=>Y) = Support (XUY) = P(XUY)

Confidence: The confidence is defined as a conditional probability
Confidence (X=>Y) = Support (XUY) / Support(X) = P(Y/X).

Table1: Transaction Database Table



TID	Item Set
1	i1,i2
2	i1,i3,i5
3	i2,i3,im
……	…………….
m	i1,i3,im

III. LITERATURE SURVEY

Jayalakshmi [5] introduced a new measure for support which does not require initially allotted weights. Item set evaluation, in classical association rule mining , by support was based on counting. A link based measure termed as w-support is used to formulate association rule mining. This system is called as “w-support link based “ because w-support can be regarded as a generalization of support . It takes weights, are not determined by assigning values to items, of transactions for evaluation. Therefore this approach is more effective than counting based measurement.

Hou Sizu [6] introduced an association analysis model which consists of seven components namely :

(i) Alarms databases – A repository of underlying effective alarms data.

(ii) Data Import – Alarm Tables are imported from alarms database to mining database for mining alarms.

(iii) Data Preprocessing – Conversion of data tables into the mining unified data format takes place.

(iv) Alarms Correlation – Alarms correlation analysis such as alarms compression, alarms filtering, alarms count, etc are used to convert and compress alarms.

(v) Sequential Pattern Mining – A sequential pattern mining algorithm is used to mine the selected unified format data in the mining database.

(vi) Post Processing – The main function of this module is to compile the results of mining into a single form like grouping, sorting and conversion for rules.

(vii) Expert Evaluation and Data Testing – Refining the mining results by adjusting the parameters for next iteration.

(viii) Rules Knowledge Warehouse – It serves as a storage of alarms correlation rules which are mined by rules engines.

This model can be used to assist network managers to position the fault quickly and accurately.

Wang [7] describes a transaction database as D=B $^{(0)}$. I , where D models transaction set, I models item set, Matrix B $^{(0)}$ is the transaction item association matrix. The elements can be defined as : to transaction I, if it associates with item j, then the corresponding element will be 1 otherwise 0.

$$D_i = \sum_{j=1}^{M} Bij^{(0)} I_j \qquad i=1,2,\ldots,N$$

where i=1,2,…,N is a transaction set.
J=1,2,...,M is a item set.

The frequency of the jth item appears in the whole transaction matrix is determined by L_j ,

$$L_j = \sum_{i=1}^{N} Bij^{(0)} .$$

Elementary operations are done to the matrix by adding the B$^{(0)}$ of the item sets to get the new matrix B$^{(1)}$ so that it is easy to identify frequent pairs of items. Later based on minimum support value, it is extended to more combination of items. At last, we could get the frequent occurring group of items satisfying minimum support. This paper realizes the formation of candidate item sets of Apriori algorithm by elementary matrix operations. It increases the efficiency of data mining.

Qing – Xiang Zhu [8] applies the apriori algorithm of association rules of data minig into tax inseption cases to accurately identify the dishonest enterprise in order to improve the efficiency and effectives of inspection. The prediction rules worked with some input information about the enterprises to which tax evasion occurred to the system.

Wei Cheng [9] describes data preprocessing based on Anomaly Detection via analysis of traffic violate type, analysis of the composition of the traffic accidents and weather , time, road type and analyse the situation of the traffic violation and drivers. Association rules was implemented in the above analysis for effective traffic management. Apriori algorithm used the database report generated from the above analysis as input for it’s working.

The major parameters for analysis are signal-to-noise (SNR) ratio, the signal level and error block rate. The improved Apriori algorithm [4] project set:

I = { { SNR1, SNR2, SNR3, ….,SNR7} , { SignalLevel1, SignalLevel2, …, SignalLevel7}, { Excellent, Good, General,Poor} }. “111 “ means { SNR >15 , Power level > -75, RS_BLER < 0.005 } then the signal quality is “Good”. Apriori algorithm proposed in the paper [4] combines the above tables and performs mining.

Tejaswi Pinniboyina [10] proposed a new coherent rules algorithm for association rules. This algorithm which is based on coherent rules allows users to mine the data without the domain knowledge. The results are comparitively outstanding over the performance of basic association rules without coherent principle.

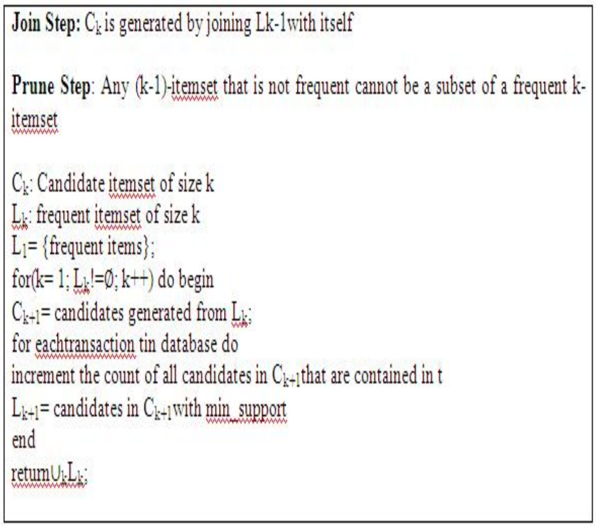
Y Jaya Babu [11] has introduced a new boolean matrix algorithm for effective spatial data mining based on associationrules. The efficiency of extracting spatial association rules was oustanding when compared to normal Apriori algorithm. This algorithm has reduced the number of times of scanning the transaction database and decreased the number of the set of candidate itemsets.

IV. APRIORI ALGORITHM

The Apriori Algorithm (refer Table2) an influential algorithm for mining frequent itemsets for boolean association rules. Frequent Itemsets: The sets of item which has minimum support (denoted by Li for ith-Itemset). Apriori Property: Any subset of frequent itemset must be frequent. Join Operation:

To find Lk, a set of candidate k-itemsets is generated by joining Lk-1with itself. The objective is to find the frequent itemsets: the sets of items that have minimum support. A subset of a frequent itemset must also be a frequent itemset and Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset) and use the frequent itemsets to generate association rules.

Table2 : Pseudocode for Apriori algorithm



```
Join Step: $C_k$ is generated by joining Lk-1with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

$C_k$: Candidate itemset of size k
$L_k$: frequent itemset of size k
$L_1$= {frequent items};
for(k= 1; $L_k$!=Ø; k++) do begin
$C_{k+1}$= candidates generated from $L_k$;
for eachtransaction tin database do
increment the count of all candidates in $C_{k+1}$that are contained in t
$L_{k+1}$= candidates in $C_{k+1}$with min_support
end
return$\cup_k L_k$;
```

V. WEKA TOOLKIT

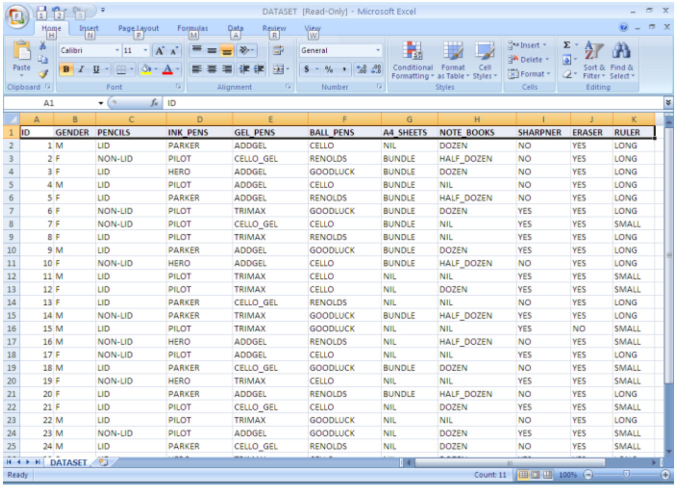
WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka [12] is an open source data mining tool. It is a free software package. The tool has 3 different interfaces namely a command line interface, an Explorer GUI and an Experimenter GUI. The GUI allows us to try out different data preparation, transformation and modeling algorithms on a dataset. The Experimenter GUI allows us to run different algorithms in a batch and compare the results. Weka contains tools for data preprocessing, classification, regression, clustering, association rules and visualization.

The key features responsible for Weka's success are:

- ➢ it provides many different algorithms for data mining and machine learning
- ➢ is is open source and freely available
- ➢ it is platform-independent
- ➢ it is easily useable by people who are not data mining specialists
- ➢ it provides flexible facilities for scripting experiments
- ➢ it has kept up-to-date, with new algorithms being added as they appear in the research literature.

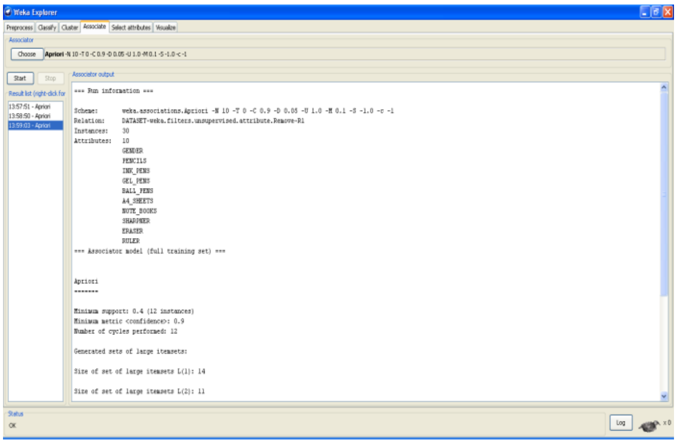
VI. EXPERIMENTAL RESULTS

Input dataset: **dataset.csv**



ID	GENDER	PENCILS	INK_PENS	GEL_PENS	BALL_PENS	A4_SHEETS	NOTE_BOOKS	SHARPNER	ERASER	RULER
1	M	LID	PARKER	ADDGEL	CELLO	NIL	DOZEN	NO	YES	LONG
2	F	NON-LID	PILOT	CELLO_GEL	RENOLDS	BUNDLE	HALF_DOZEN	NO	YES	LONG
3	F	LID	HERO	ADDGEL	GOODLUCK	BUNDLE	DOZEN	NO	YES	LONG
4	M	LID	PILOT	ADDGEL	CELLO	BUNDLE	NIL	NO	YES	LONG
5	F	LID	PARKER	ADDGEL	RENOLDS	BUNDLE	HALF_DOZEN	NO	YES	LONG
6	F	NON-LID	PILOT	TRIMAX	GOODLUCK	BUNDLE	DOZEN	YES	YES	LONG
7	F	NON-LID	PILOT	CELLO_GEL	CELLO	BUNDLE	NIL	YES	YES	SMALL
8	F	LID	PILOT	TRIMAX	RENOLDS	BUNDLE	NIL	YES	YES	LONG
9	M	LID	PARKER	ADDGEL	GOODLUCK	BUNDLE	DOZEN	YES	YES	LONG
10	F	NON-LID	HERO	ADDGEL	CELLO	BUNDLE	HALF_DOZEN	NO	YES	LONG
11	M	LID	PILOT	TRIMAX	CELLO	NIL	NIL	YES	YES	SMALL
12	F	LID	PILOT	TRIMAX	CELLO	NIL	DOZEN	YES	YES	SMALL
13	F	LID	PARKER	CELLO_GEL	RENOLDS	NIL	NIL	NO	YES	LONG
14	M	NON-LID	PARKER	TRIMAX	GOODLUCK	BUNDLE	HALF_DOZEN	YES	YES	LONG
15	M	LID	PILOT	TRIMAX	GOODLUCK	NIL	NIL	YES	NO	SMALL
16	M	NON-LID	HERO	ADDGEL	RENOLDS	NIL	HALF_DOZEN	NO	YES	SMALL
17	F	NON-LID	PILOT	ADDGEL	CELLO	NIL	NIL	YES	YES	LONG
18	M	LID	PARKER	CELLO_GEL	GOODLUCK	BUNDLE	DOZEN	NO	YES	SMALL
19	F	NON-LID	HERO	TRIMAX	CELLO	NIL	NIL	YES	YES	SMALL
20	F	LID	PARKER	ADDGEL	RENOLDS	BUNDLE	HALF_DOZEN	NO	YES	LONG
21	F	LID	PILOT	CELLO_GEL	CELLO	NIL	DOZEN	YES	YES	SMALL
22	M	LID	PILOT	TRIMAX	GOODLUCK	NIL	NIL	NO	YES	LONG
23	M	NON-LID	PILOT	ADDGEL	GOODLUCK	NIL	DOZEN	YES	YES	SMALL
24	M	LID	PARKER	CELLO_GEL	RENOLDS	NIL	DOZEN	NO	YES	SMALL

Apriori Algorithm using WEKA

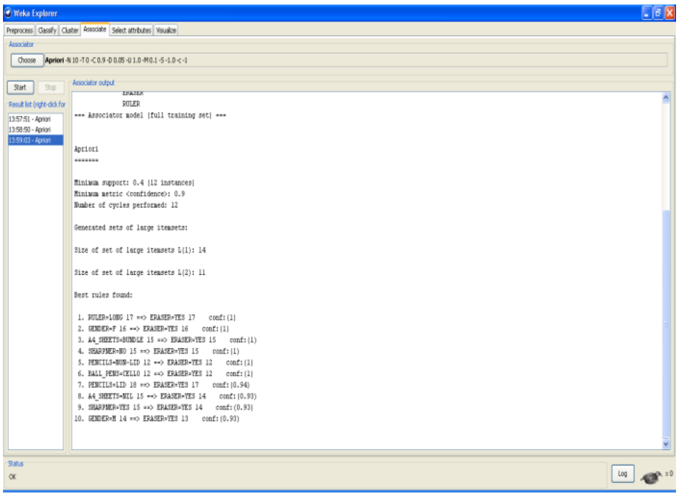


Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

13:57:51 - Apriori
13:58:50 - Apriori
13:59:03 - Apriori

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     DATASET-weka.filters.unsupervised.attribute.Remove-R1
Instances:    30
Attributes:   10
              GENDER
              PENCILS
              INK_PENS
              GEL_PENS
              BALL_PENS
              A4_SHEETS
              NOTE_BOOKS
              SHARPNER
              ERASER
              RULER
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.4 (12 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 11
```

Support and Confidence Calculation:



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

13:57:51 - Apriori
13:58:50 - Apriori
13:59:03 - Apriori

```
              ERASER
              RULER
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.4 (12 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 11

Best rules found:

 1. RULER=LONG 17 ==> ERASER=YES 17    conf:(1)
 2. GENDER=F 16 ==> ERASER=YES 16    conf:(1)
 3. A4_SHEETS=BUNDLE 15 ==> ERASER=YES 15    conf:(1)
 4. SHARPNER=NO 15 ==> ERASER=YES 15    conf:(1)
 5. PENCILS=NON-LID 12 ==> ERASER=YES 12    conf:(1)
 6. BALL_PENS=CELLO 12 ==> ERASER=YES 12    conf:(1)
 7. PENCILS=LID 18 ==> ERASER=YES 17    conf:(0.94)
 8. A4_SHEETS=NIL 15 ==> ERASER=YES 14    conf:(0.93)
 9. SHARPNER=YES 15 ==> ERASER=YES 14    conf:(0.93)
10. GENDER=M 14 ==> ERASER=YES 13    conf:(0.93)
```

VII. CONCLUSION

This paper implements association rules in a stock market field and its results were very effective for future marketing. Thus association rules prove themselves to be the most effective technique for frequent pattern matching over a decade.

ACKNOWLEDGMENT

I am successful in my research work with a great support and unshakeable confidence from (my guide) Dr.S.P.Shantharajah. I also like to thank Mrs.R.Deepalakshmi (Asst. Prof.) and Dr.P.Alli (HOD/CSE) of my institution for encouraging my research activities. I extend my heartiest thanks to my parents (S.Sundaramoorthy , R.J.Shanthi), brother (S.Suresh Kannan) and daughter (S.Shriya) for their kind co-operation.

REFERENCES

1. Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition , Morgan Kaufmann Publishers.
2. Feng Yucai, "Association Rules Incremental Updating Algorithm", Journal of Software, Sept., 1998.
3. Association Rule Learning – Wikipedia, the free encyclopedia.
4. Lei Guoping, Dai Minlu, Tan Zefu and Wang Yan, " The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rules", IEEE conference paper – project supported by the Science and Technology Research Project of Chongqing municipal education commision under contract no KJ101114 and KJ 111103, 2011
5. Jayalakshmi.S, Dr k. Nageswara Rao, "Mining Association rules for Large Transactions using New Support and Confidence Measures", Journal of Theoretical and applied Information Technology, 2005.
6. Hou Sizu, Zhang Xianfei, " Alarms Association Rules Based on Sequential Pattern Mining Algorithm", Fifth IEEE International Conference on fuzzy Systems and Knowledge Discovery, 2008.
7. Chengmin Wang, Weiqing Sun, Tieyan Zhang, Yan Zhang, " Research on Transaction item association matrix mining algorithm in large scale transaction database", IEEE sixth International Conference on FUZZY Systems and Knowledge Discovery, 2009.
8. Qing-Xiang Zhu, Li-Juan Guo, Jing Liu, Nan Xu, Wei-Xu Li, :Research of Tax Inspection cases – Choice based on Association Rules in Data Mning", Proceedings of the Eighth International IEEE conference on Machine Learning and Cybernetics, Baoding, 12 -15 July,2009
9. Wei Cheng, Xiaofeng Ji, Chunhua Han, Jianfeng Xi, "The Mining Method of the Road Traffic Illegal Data Based on Rough Sets and Association Rules", IEEE International Conference on Intelligent Computation Technology and Automation, 2010.
10. Tejaswi Pinniboyina, Navya Dhulipala2 Radha Rani Deevi, Sushma Nathani, "Mining Items From Large Database Using Coherent Rules", International Journal Of Engineering Science & Advanced Technology, Volume - 2, Special Issue - 1, 61 – 70, ISSN: 2250–3676, Jan- Feb 2012.
11. Y Jaya Babu, G J Phani Bala, Siva Rama Krishna T, "Extracting Spatial Association Rules From The Maximum Frequent Itemsets Based On Boolean Matrix", International Journal Of Engineering Science & Advanced Technology, Volume - 2, Issue - 1, 79 – 84, ISSN: 2250–3676, Jan- Feb 2012.
12. A B M Shawkat Ali, Saleh A.Wasimi, "Data Mining: Methods and Techniques", Cengage Learning India Private Limited, India.