ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Solving Management Problems Using RSTDB a Frequent Pattern Mining Technique

Vaibhav Kant Singh

Assistant Professor, Department of Computer Science & Engineering, Institute of Technology,

Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chhattisgarh, India

Abstract: We are living in a highly competitive environment where time to make critical decision is getting lesser and lesser. Data Mining which is the exploration and extraction of meaningful information from the large source of data generated as a result of various data processing activities is capable to help the higher management authorities to make important decisions on time. Data mining which is basically a step in Knowledge Discovery from Database process uses several techniques to implement the above objective. Association rule mining is one of the data mining techniques used, which require the mining of frequent data sets in the transaction database. Thus, Data Mining basically helps in creation of decision support system for higher managers. In this paper we have proposed the use of RSTDB algorithm for the extraction of Information from large databases for solving management problem.

Keywords: I Data-mining, Knowledge Discovery from Database, Association rule, Frequent Pattern Mining.

I. INTRODUCTION

With the advent of Computing technology there are two main areas that are highly affected, the areas are:-

1. Business Data Processing
2. Scientific Computing

With the decreasing cost of the computer hardware and large capacity to store huge amount of data , along with the interfaces and platforms available to generate and store large amount of data even small enterprises afford to have Giga & Tera Bytes of data stored in their office systems generated as a result of the various data processing activities. We know that we are living in an era of Information Technology where processing is mainly done on information rather than raw data.

We know that anything and everything that is present is data so. Each and every activity that is performed generates some data. The data generated is stored in data bases in form of records. The higher manager can take the benefit of the generated records which are result of various activities prevailing if the record is transformed into some form that can be evaluated in a short duration of time. The above problem of transforming the raw data in report form is done by a process called Knowledge Discovery from Database (KDD).

Knowledge Discovery from Database (KDD)

The KDD process involves the following 6 steps.

STEP I : Selection
STEP II : Preprocessing
STEPIII: Transformation
STEPIV: Data Mining
STEP V : Evaluation & Interpretation
STEP VI: Data Visualization

The KDD is a very vital process by which the higher management officials are provided with the reports generated after the data mining step in the KDD process. The reports are in form that gives concise representation of the large data base. The reports are generated in form patterns that are easily recognized by the eyes. The reports are generated from the result of the data mining step in KDD.Thus, data mining is a very vital step in data mining.

Data Mining

Discovering relations that connects variables in a database is the subject of data mining. The data mining system self-learns from the previous history of the investigated system, formulating and testing hypothesis about rules which systems obey. When concise and valuable knowledge about the system of interest is discovered, it can and should be interpreted into some decision support system, which helps the manager to make wise and informed business decisions.

Data Mining Systems

Depending upon how the data mining system has utilized the relational database the data mining system are of three types:-

1. They may not use it at all.
2. Loosely coupled
3. Tightly coupled

A majority of data mining system do not use relational database at all and have its own memory and storage management. In case of loosely coupled systems DBMS is used only for storage and retrieval purpose. In tightly coupled approach, the portions of the application programs are selectively pushed to the database system to perform the necessary computation. Data are stored in the database and all processing is done at the database end.

DM Techniques

Researchers identify two fundamental goals of data mining:

1. Prediction
2. Description

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, and description focuses on finding patterns describing the data and the subsequent presentation for user interpretation. Some of the data mining techniques are:-

1. Association rule Mining
2. Clustering
3. Classification
4. Frequent Episodes
5. Deviation detection
6. Neural Network
7. Genetic algorithm
8. Rough Set
9. Support Vector Machines

Association Mining
An association rule is an expression of the form X=>Y, where X and Y are the sets of items. The intuitive meaning of such a rule is that the transaction of the database which contains X tends to contain Y.

Clustering
Clustering is a method of grouping data into different groups, so that the data in each group share similar trends and patterns. Clustering constitutes a major class of data mining algorithms. The algorithm attempts to automatically partition the data space into a set of regions or clusters, to which the examples in the table are assigned, either deterministically or probability-wise.

Classification
Classification involves finding rules that partition the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classification analyzes the training data set and constructs a model based on the class label to the future unlabelled records.

Frequent Episodes
Frequent Episodes are sequence of events that occur frequently, close to each other and are extracted from the time sequences. How close it has to be to consider it as frequent is domain dependent.

Deviation Detection
Deviation detection is to identify outlying points in a particular data set , and explain whether they are due to noise or other impurities being present in the data or due to trivial reasons.

Neural Network
Neural Networks are a new paradigm in computing, which involves developing mathematical structures with ability to learn. The methods are result of academic attempts to model the nervous system learning.

Genetic Algorithm
Genetic algorithms are a relatively new computing paradigm, inspired by "Darwin's Theory of Evolution". A population of individuals, each representing a possible solution to a problem, is initially created at random. Then pairs of individuals combine (Crossover) to produce off-spring for the next-generation. A mutation process is also used to randomly modify the genetic structure of some members of each new generation. The algorithm runs to generate solutions for successive generations.

Rough Sets Techniques
The rough sets theory has recently become a popular theory in the field of data mining. The theory, introduced by Pawlak in the early 1980s, provides a formal framework for the automated transformation of data into knowledge.

Support Vector Machines (SVM)
The SVM is based on statistical learning theory and is increasingly becoming useful in data mining. The main idea is to non-linearly map the data set into a high dimensional feature space and use a linear discriminator to classify the data.

II. FREQUENT PATTERN MINING

The frequent pattern mining problem was first introduced by R. Agrawal, as mining association rules between sets of items. Let I= $\{i_1,\ldots\ldots, i_m\}$ be a set of items. An itemset X $\subseteq$ I is a subset of items. Hereafter, we write itemsets as X= $i_{j1}\ldots\ldots i_{jn}$. Particularly, an itemset with an l item is called an l-itemset. A transaction T= (tid,X) is a tuples where tid is a transaction-id and X is an itemset. A transaction T= (tid, X) is said to contain itemset Y, if Y $\subseteq$ X. A transaction database TDB is a set of transactions.

Given a Transaction database TDB a support threshold min sup and a confidence threshold min conf , the problem of association rule mining is to find the complete set of association rules that have support and confidence no less than the user-specified thresholds, respectively.

Association rule mining can be divided into two steps. First ,frequent patterns with respect to support threshold min sup are mined. Second, association rules are generated with respect to confidence threshold min conf. As shown in many studies, the first step, mining frequent patterns, is significantly more costly in terms of time than the rule generation step.

As we shall see later, frequent pattern mining is not only used in association rule mining .Instead, frequent pattern mining is the basis for many data mining tasks, such as sequential pattern mining and associative classification. It also has broad applications, such as basket data analysis, cross-marketing, catalog design, sale campaign analysis, web log analysis etc.

III. LITERATURE SURVEY

In [1] the new algorithm RSTDB is proposed for frequent pattern mining. The approach is briefly described along with small description of its implementation. In [2] the application of RSTDB is discussed and is compared with the traditional approach for frequent pattern mining. The scope of RSTDB is discussed in the paper. In [3] RSTDB algorithm and some other cache conscious efforts for

DOI 10.17148/IJARCCE.2015.4860

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

frequent pattern mining are discussed. In [4] RSTDB is discussed from the point of view of candidate generation and test view. In [5] RSTDB approach is elaborated with implementation interfaces. In [6] both candidate generation and test and divide and conquer are compared for frequent pattern mining. Also conclusion was drawn that pattern growth methods are efficient in maximum situations. In [7] the tree projection algorithm for identification of frequent itemsets is given. In [8] how to mine association rules between sets of items in large databases is given. In [9] an alternative approach for finding frequent patterns is given. The approach employs technique other than the candidate generation technique. In [10] algorithms which can accelerate the identification of association rules are proposed. In [11] how to mine association rules from long patterns present in databases is shown. In [12] how mining can be performed in sequential patterns is elaborated. In [13] we have made a comparison between RSTDB and FP-Tree algorithm.

IV. APRIORI ALGORITHM

Apriori algorithm consists of two phases namely:-

1. Candidate Generation
2. Pruning

The Candidate-Generation is described below:-

gen_cand_itemsets with the given L_{k-1} as follows
$C_k = \phi$
For all itemset $l_1 \epsilon L_{k-1}$ do
For all itemset $l_2 \epsilon L_{k-1}$ do
If $l_1[]=l_2[1] \wedge l_1[2]= l_2[2] \wedge \ldots\ldots \wedge l_1[k-1]< l_2[k-1]$
Then c= $l_1[1], l_1[2] \ldots . l_1[k-1], l_2[k-1]$
$C_k = C_k \cup \{ c \}$

The Pruning Algorithm is described below:-

Prune(C_k)
For all c C_k
For all (k-1) subsets d of c do
If d $\notin L_{k-1}$
Then $C_k = C_k \backslash \{c\}$

Here,
k is the number of passes required.
L_{k-1} is the frequent itemset.
C_k is the candidate itemset.

The Apriori algorithm is the most basic Algorithm used for frequent itemset mining. We know that Data Mining is basically the fourth step involved in the KDD process. Thus, data mining will take as input the output of the third stage and will provide output that will be input for the evaluation & interpretation stage{In this stage reports are generated}.The input to data mining stage is basically the processed data i.e. the input to data mining stage is the transformed data or information. In other words Transaction database is database generated from the raw database. The Apriori algorithm is put into the transaction database to generate the final frequent set, which is the set of frequent items that is having occurrence more than a prescribed threshold value called the support value.

Shortcomings of the Above Algorithm

The Algorithm suffers from two major shortcomings:-

1. It is costly to handle large number of candidate sets.
2. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching.

V. PROPOSED WORK RSTDB

RSTDB algorithm is influenced from the Apriori algorithm with some cases where it yields better result by the presence of heuristic function in the algorithm structure. RSTDB like Apriori is simple to implement and could be easily incorporated for deriving frequent item-set from large database. The frequent dataset could be later on utilized for extraction of information for solving various issues.
In the current paper RSTDB is proposed as an approach for solving management problems in large organizations. As the amount of data transaction going on in large organization is large. Thus each organization is having large amount of data, this data which is going to take space in memory and is not analysed. RSTDB could be an approach utilizing which various management problems could be solved.

RSTDB Algorithm

Step 1
Calculate the size of each transaction in the Transaction Database.

Step 2
Evaluate the transaction set having maximum size.

Step 3
Check for the transaction set size having frequency or support value more than the given threshold value. Set this transaction size as the maximum value up to which scanning & candidate generation step has to proceed. This will be the maximum value of k up to which iteration has to be done.
Value HeuFn [no. of items, TDB size] = Max k
Step 4 & Step 5 iterates until k = Max k
Value of k lies between 1 and Max k
After the end of this step we are having values of L and Max k.If value of k=1 no further scanning of database is done.

Step 4
Candidate Generation
gen_cand_itemsets with the given L_{k-1} as follows
$C_k = \varnothing$
for all item set $I_1 \in L_{k-1}$ do
for all item set $I_2 \in L_{k-1}$ do
if $I_1[1]= I_2[1] \wedge I_1[2]= I_2[2] \wedge \ldots \wedge I_1[k-1]< I_2[k-1]$
then c= $I_1[1], I_1[2] \ldots . I_1[k-1], I_2[k-1]$
$C_k = C_k \cup \{c\}$

DOI 10.17148/IJARCCE.2015.4860

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Step 5
Candidate Prune Step
Prune (C_k)
for all c $\in$ C_k
for all (k-1) subsets d of c do
If d $\notin$ L_{k-1}
then $C_k = C_k \backslash \{c\}$

After this step each time the value of L_k is updated with all the candidates of C_k having support values either greater than or equal to minimum threshold.

After the iteration completes the Union of all the L_k is done to obtain L_{final}.

Here,
I is the item set present in L_{k-1} of the TDB.
L_k is a set of candidate k item set having support greater than threshold
k represents database scan number.
C_k is a superset of L_k.

VI. CONCLUSION

As RSTDB provide good result in certain situation when compared from Apriori algorithm, which is the most basic and simplest approach for finding frequent patterns. Thus taking the scenario under consideration we can say that RSTDB which is an Association rule mining approach i.e. a Data mining technique can be used for solving management issues by preparation of reports on the frequently arrived patterns.

REFERENCES

[1] V.K. Singh, V. Shah, Y.K. Jain, A. Shukla, A.S. Thoke, V.K. Singh, C. Dule and V. Parganiha, "Proposing an efficient method for frequent pattern mining," ICCSS-08, Published in the Proceeding and Journal of WASET, ISSN: 2070-3724, pp. 384-389, Bangkok, Thailand, Dec. 17-19, 2008.
[2] V.K. Singh and V.K. Singh, "Minimizing space time complexity by RSTDB a new method for frequent pattern mining," IHCI 2009, Sponsored by IEEE UP Chapter, Proceedings of the First International Conference on Intelligent Human Computer Interaction, Published by Springer, ISBN: 978-81-8489-203-1, pp. 361-371, Indian Institute of Information Technology, Allahabad, Jan. 20-23, 2009.
[3] V.K. Singh, "RSTDB and cache conscious techniques for frequent pattern mining," CERA-09, Proceeding of Fourth International Conference on Computer applications in Electrical Engineering, pp. 433-436, Indian Institute of Technology, Roorkee, India, Feb 19-21, 2010.
[4] V.K. Singh and V.K. Singh, " RSTDB a new candidate generation and test algorithm for frequent pattern mining" CNC-2010, ACEEE and IEEE, IEEE Communication Society, Washington DC, Proceeding of International Conference on Advances in Communication Network and Computing, Published by ACM DL, ISBN 978-0-7695-4209-6, pp. 416-418, Calicut, Kerala, India, 4-5 Oct. 2010.
[5] V.K. Singh and V.K. Singh, "RSTDB a candidate generation and test approach," Published in International Journal Research Digest, ISSN: 0973-6387, Bilaspur, (C.G.), India, Oct-Dec 2010.
[6] V.K. Singh and V. Shah, "Minimizing space time complexity in frequent pattern mining by reducing TDB scanning and using pattern growth methods,"(Periodical style—Accepted for publication), Chhattisgarh Journal of Science and Technology, ISSN: 0973-7219, to be published.
[7] R. Agarwal, C. Aggarwal and V.V.V. Prasad, "A tree projection algorithm for generation of frequent itemsets," Journal of Parallel and Distributed Computing, vol 61, pp. 350-371, 2001
[8] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," (SIGMOD'93), Proceeding ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, DC, 1993.
[9] J. Han, J. Pei and Y. Yin, "Mining frequent patterns without candidate generation," SIGMOD' 00, Proceeding of 2000 ACM SIGMOD, International Conference on Management of Data, pp. 1-12, Dallas, TX, 2000.
[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," VLDB' 94, pp. 487-499, 1994.
[11] R.J. Bayardo, "Efficiently mining long patterns from databases," SIGMOD'98, pp. 3-14, 1998.
[12] R. Agrawal and R. Srikant, "Mining sequential patterns," ICDE'95, pp. 3-14, 1995.
[13] V.K. Singh, "Comparing Proposed Test Algorithm RSTDB with FP-Tree Growth Method for Frequent Pattern Mining," Published in Aryabhatt Journal of Mathematics and Informatics, ISSN 0975-7139, vol 5, issue 1, pp. 137-140June 3 2013

BIOGRAPHY

Vaibhav Kant Singh is M.Tech. (CSE), GATE Qualified (CSE), B.E. with Honours (CSE), MISTE, Assistant Professor in the Department of Computer Science & Engineering, Institute of Technology, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, (C.G.), India. He is having around 8 years of teaching experience. He is having 5 Papers in International Journal, 9 Papers in International Conference and around 18 Papers in National Conference. His papers are published in ACM, Springer, WASET, AIP etc. He has taught various subjects like Artificial Intelligence, Artificial Neural Network, Data Mining, Network Programming, Compiler Design, Fuzzy Logic, Java, Design and Analysis of Algorithm, Operating System, Parallel Computing, Internet Fundamentals and Application, C++etc.