From Frequent Itemset To Knowledge Discovery in Crime Database

P. PRABAKARAN, Dr. K. RAMESHKUMAR
Research Scholar, Research Supervisor
Computer Science Department,
Mass college of arts and science, Kumbakonam, India

Abstract : Data mining or knowledge discovery is the computer assisted process of digging through analyzing enormous sets of data and then extracting the meaning of the data. It is used to extract information from large voluminous databases. Data mining is the important task in knowledge discovery process and also well established field in computer science. In recent years data mining is applied into both commercial and scientific purposes.

IndexTerms **- KDD, Decision tree, SPM, Apriori, ARM.**

I. INTRODUCTION

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from collection of data. It is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. The areas like marketing, telecommunication, manufacturing and fraud detection can be applied to KDD. Data Mining is "the computational process of discovering patterns in large data sets" for the goal of "extracting information from a data set and transform it into an understandable structure for further use".

The tools in data mining predicts future trends and behaviors, making businesses to make proactive, make knowledge driven decisions etc. It also offered automated, prospective analyses which move beyond the analyses of past events provided by retrospective tools typical decision support systems. Software can be used for patterns in large batches of data, to know about customers in case of business, to develop effective marketing strategies, to increase sales and decrease costs. It depends on effective data collection, storing data in warehousing as well as processing.

II. DATA MINING IN CRIME PATTERN

Data Mining technique can be applied to various area. Every day the crime rate is increasing and making huge causes to the people in the society. Crime investigation plays a major role in the police system. Police stations are using the system of storing and retrieving the criminal data and subsequent reporting. Since there are numerous records are increasing day by day it is very difficult to find out the suspects from among the huge data. It is a great task for the police department to detect and prevent crimes and criminals. The survival of crime in a community and eradication of it is a task to the society because of its dangerous impact on the development of its whole. Actually it paves way to massive waste of enforcement energy and huge economic loss. Hence, with the modern techniques in the field of criminology and the science of criminal behavior, constant pains are taken to classify recognized classification of crimes and criminals to provide a balanced platform for punishment of various categories of criminals.

III. FREQUENT PATTERN MINING

Frequent Patterns are itemset that is present in a dataset with frequency number less than the fixed support threshold. For example, the items like mobile phone and memory card appear frequently together in a dataset are called as frequent itemset. The sequential pattern is one which appears sequentially in the database. Buying first bread, vegetables and then cheese for making sandwich occurs frequently in the supermarket. The various structural forms such as sub-graphs, sub-trees or sub-lattices combined with itemset are called as substructure. The structural patterns are formed when the substructure occurs frequently in a graph database.

III.1. Sequential pattern mining

Sequential Pattern Mining mines statistically relevant patterns between data, where the values are delivered in sequence. The time series mining is closely related to the sequential pattern mining since the values are presumed to be discrete. But in some cases it is considered as a different activity. In structured data mining, it plays a special role. Traditionally there are several computational problems are identified within this area. It comprises of creating well structured database and indexes for sequential ordered information, mining the frequent patterns, analyzing and comparing sequences for similarity and also retrieving missing sequence numbers.

III.2. Frequent tree mining

Frequent Tree Mining is based on a property which states that every non-empty sub tree of a frequent tree pattern is also frequent. Hence, they expand the candidate-generation and test approach to undertake the mining process. The technique for frequent tree pattern mining is resourceful and scalable when the patterns are not too difficult. On the other hand, if there are many difficult patterns in the data set, there can be a vast number of candidates required to be generated and tested. This may cause to degrade the performance.

III.3. Apriori algorithm

Apriori algorithm follows "Bottom up" approach where the frequent subsets are extended one item at a time. Each set of data has a number of items called a transaction. The output of Apriori is set of rules that tell how often items are contained in sets of data. Apriori algorithm is otherwise called as candidate generation and test approach and possesses a property called as Apriori property. The algorithm specifies that all non empty subsets of a frequent itemset must be frequent.

Apriori Algorithm

Apriori algorithm carries two concepts 1) Self Join and 2) Pruning. It applies level-wise search where k-itemset are used to find (K+1) itemset.

Step 1: First, the set of frequent one itemset is identified which is marked as C1.

Step 2: The second step is calculating the support. It denotes the occurrence of the item in the database and it needs to check the entire database.

Step 3: The third step is the pruning step. It carried out on C1 in which the items are compared with the minimum support factor. The items which satisfy the minimum support criteria are only considered for the subsequent processes which are denoted by L1.

Step 4: The fourth step is the candidate set generation. It carries out two itemset generation and marked as C2.

Step 5: The next step is to scan the database again to calculate the support of the two itemset. As per the minimum support the generated candidate sets are tested and only the itemset which satisfies the minimum support criteria are further used for 3-itemset candidate set generation.

Step 6: Steps from 1 to 5 are repeated till there is no frequent itemset than can be generated.

The following example helps to understand the concept used by the Apriori. Table 3.3.1 exhibits a transactional database possessing four transactions. TID is a unique identification applied to every transaction.

Table 3.3.1 Transaction Database

TID	Items
T001	A,C,D
T002	B,C,E
T003	A,B,C,E
T004	B,E

While running the 1st step of scanning, the database recognizes the number of event for a particular item. After the first step C1 is derived, which is shown in Table 3.3.2.

Table 3.3.2 C1 Candidate set

Itemset	Support
{A}	2
{B}	3
{C}	3

Table 3.3.3 L1 Itemset

Itemset	Support
{A}	2
{B}	3

{D}	1
{E}	3

→

{C}	3
{E}	3

The subsequent step is pruning the itemset. In this step, the support count of an itemset is compared with the minimum support. For further processing the itemset which satisfies the minimum support is considered. In the above table 3.3.2, let us consider minimum support as 2, we will get L1 from this step. Table 3.3.3 shows the result of pruning. When the candidate generation step is done, all possible but unique two itemset candidates are generated.

Apriori algorithm

Input: Datasets X, MinSupp S

Output: Y(X, S)

$D_1 := \{\{a\} \mid a \in I\}$

k := 1

while $D_k \neq \{\}$ do
for all transactions (TID, i) $\in$X do
for all candidate sets $Z \in D_k$ do
if $Z \subseteq i$ then
increment Z support by 1
end if
end for
end for
$B_k := \{Z \in D_k \mid Z \text{ support} \geq S\}$
$D_{k+1} := \{\}$
for all Z, $P \in B_k$
such that Z[i] = P[i]
for $1 \leq i \leq k-1$, and Z[k] < P[k] do
$i := Z \cup \{P[k]\}$
if $\forall K \subset L; |K| = k : K \in B_k$ then
add i to D_{k+1}
end if
end for
increment k by 1
end while

IV. ASSOCIATION RULE MINING

Association Rule Mining (ARM) aims to extract interesting correlations, frequent patterns, associations [92] or casual structures among sets of items in the transactional database. The patterns discovered with this technique can be represented in the form of association rules. ARM plays an important role in analyzing and predicting consumer behavior. They are significant in product grouping, shopping, basket data analysis, store layout and catalog design. Generation of association rules are carried out through evaluating the facts for frequent If-then patterns [15] [211] and the most important relationships are identified by using the criteria support and confidence. Support is a measure that indicates how frequently the items in the transaction appear in the database. Confidence is a measure that indicates the number of times If-then statements are found to be true.

V. CRIME PATTERN MINING

Crime Pattern analysis is a generic term for a number of related analytical disciplines such as crime, crime trend analysis, hot spot analysis and general profile analysis. Above all, crime pattern analysis noticed the linkages between crimes and other forms of offences to reveal similarities and differences. It can help the crime investigation teams to reveal the relationships between crimes and other committed offences quickly.

VI. VERTICAL LAYOUT APPROACH

VI.1. Eclat Algorithm

Eclat stands for Equivalence CLAss Transformation algorithm is the novel and primary algorithm to produce every frequent itemset in a depth first manner. During vertical layout storage, the computation of support count makes easy by just intersecting the covers of two of its subsets that gives the set itself. This method is efficiently used inside the Apriori algorithm by the Eclat algorithm. But this method is not probable for all cases because the total size of all covers at a certain iteration of the local set generation procedure could go beyond the main memory limits. Eclat algorithm is much suitable only for bigger datasets.

VI.2. Extension of Eclat Algorithm

1. Diffset with Eclat (dEclat) algorithm.
2. Hybrid algorithm, combination of Apriori and Eclat
3. Vertical Itemset Partitioning for Efficient Rule Extraction (VIPER) algorithm.

VII. EXTENSION OF FREQUENT PATTERN

VII.1. Maximal Frequent Pattern [131]

The maximal candidate set is a superset of the maximal patterns. In contrast GenMax maintains only the current known maximal patterns for pruning. MaxMiner [20] is another algorithm for finding the maximal elements. It uses efficient pruning techniques to quickly narrow the search. MaxMiner employs a breadth-first traversal of the search space; it reduces database scanning by employing a look ahead pruning strategy. It also employs item (re)ordering heuristic to increase the effectiveness of superset-frequency pruning. Since MaxMiner uses the original horizontal database format, it can perform the same number of passes over a database as Apriori does.

Maximal Frequent Pattern mining algorithms are:

1. Pincer-Search
2. MaxMiner

VII.2. Surprising Pattern

Surprising pattern is based on the number of bits in which a basket sequence can be encoded under a carefully chosen coding scheme. In this scheme it is inexpensive to encode sequences of itemset, that have steady, hence likely to be well-known, correlation between items. Conversely, a sequence with large code length hints at a possibly surprising correlation.

VII.3. Sequential pattern

Sequential Pattern Mining is a part of data mining which finds out statistically related patterns between data examples where the values are delivered in a sequence. It is usually assumed that the values are distinct, and thus time series mining is closely related, but usually treated as a different activity. Sequential Pattern Mining is a special case of structured data mining.

Some of the Sequential Pattern Mining algorithms are:

1. Generalized algorithm
2. Vertical format based algorithm

VII.4. Structural pattern

Frequent substructures are more essential patterns that can be identified in a group of graphs. Nowadays studies have developed enormous frequent substructure mining techniques. The below mentioned are some of the structural patterns mining algorithms:

1. Apriori based Graph Mining (AGM) algorithm
2. Pattern Growth based graph pattern mining algorithm
3. Frequent Sub Graph Mining (FSG) algorithm
4. MoFa
5. Frequent Tree Mining algorithm
6. Pattern Matching Tree mining algorithm
7. Fast Frequent Sub graph mining algorithm
8. Spanning tree based maximal graph mINing (SPIN)

VIII. CONCLUSION

The ultimate goal of data mining is the prediction of human behavior, and is by far the most common business application; however this can easily be modeled to meet the objective of detection and deterrence of criminals. These and many more application have

demonstrated that, rather than requiring a human to attempt to deal with hundreds of descriptive attributes, data mining allows the automatic analysis of databases and the recognition of important trends and behavioral patterns. The proposed work demands the real world live data for analysis purpose in order to give valuable, hidden, potential patterns from the large amount of data. In proposed work the data is distributed at different locations so the concept of web mining is used properly. Web mining means to extract the valuable, important pattern or knowledge from data which is distributed at remote locations. For mining frequent items and itemsets on stream data, started with sticky sampling and lossy counting algorithms for approximate frequency counts over data streams. These counter based methods were followed by tree based algorithms to mine frequent patterns one of the most popular data structure used here is FP-Tree.

REFERENCES

[1] U.M. Fayyad and R, Uthurusamy, "Evolving data mining into solutions for insights.Communications of the ACM"2002.

[2] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: An Overview and Case Studies"

[3] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin Michael Chau, "Crime Data Mining: A General Framework and Some Examples"

[4] Data Mining In Time Series Databases (2004) by M. Last., A. Kandel, and H. Bunke

[5] Buntine, W. 1996. Graphical Models for Discovering Knowledge. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. PiatetskyShapiro, P. Smyth, and R. Uthurusamy, 59–82.