PROJECT REPORT ON

Detection of Sensitive Items in Market Basket Database using Association Rule Mining for Privacy Preserving

MASTER OF TECHNOLOGY (Mtech)

(DEPARTMENT OF COMPUTER SCIENCE)

SUBMITTED TO-

Dr. Dharmveer Singh Rajput

SUBMITTED BY-

Rishi Bhardwaj(10503869)

Raghav Gupta(9910103575)

TABLE OF CONTENT

ABSTRACT

Mining Association rules on large data sets has received considerable attention in recent years. Association rules are useful for determining correlations between attributes of a relation and have applications in marketing, financial, and retail sectors. Association rules in market basket database represent the shopping behaviour of customers. The association information may reveal trade secrets. Association rule hiding in privacy preserving data mining hides sensitive rules containing sensitive items. In this paper, a new method is proposed to detect the sensitive items for hiding sensitive association rules. This proposed method finds the frequent item sets and generates the association rules.

CHAPTER-1

INTRODUCTION

1.1 **OVERVIEW :**

Data mining is an essential technology to extract patterns or knowledge from large repositories of data. Association rules in market basket database represent the shopping behaviour of customers. The association information may reveal trade secrets. It must be hidden before publishing. Association rule hiding in privacy reserving data mining hides sensitive rules containing sensitive items. Data mining is a knowledge discovery process of analysing large databases to find useful patterns. It has numerous applications such as marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, etc. The knowledge discovered from a database can be expressed in patterns such as decision trees, clusters or association rules. Retail shops are interested in associations between different items that customer place in the shopping basket. By discovering interesting association rules, an organization can identify underlying patterns useful in planning of new store layout, new product assortments and which product to put on promotion. The knowledge discovered can be beneficial to business competitors. In order to preserve their competitive edge some partners may hesitate to disclose sensitive information and also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners.

Privacy preserving data mining is a broad research area for protecting sensitive data or knowledge. There have been two types of privacy concerning data mining. The first type of privacy, called output privacy, is that preventing the mining output from malicious inference attacks. The second type of privacy, input privacy, is that sanitizing the raw data itself before performing mining. The framework of privacy preserving data mining is shown in Fig 1.

MARKET BASKET ANALYSIS USING ASSOCIATION RULE MINING STRATEGY:

The term market basket analysis using association rule learning in the retail business refers to the research that provides the retailer with information to understand the purchasing behaviour of a customer. This association information will enable the retailer to understand the customer's needs and rewrite the store's layout accordingly, develop cross-promotional programs, or even capture new customers. An example for this was when one super market
chain discovered in its analysis that customers who bought diapers often bought beer as well, have put the diapers close to beer coolers, and their sales increased dramatically.

1.2 APPLICATION:

Data mining is a knowledge discovery process of analysing large databases to find useful patterns. Items that customer place in the shopping basket. By discovering interesting association rules, an organization can identify underlying patterns useful in planning of new store layout, new product assortments and which product to put on promotion.

The knowledge discovered can be beneficial to business competitors. In order to preserve their competitive edge some partners may hesitate to disclose sensitive information and also to prevent data mining techniques from discovering sensitive knowledge which is not even known to the database owners.

1.3 MAJOR METHOD: The initial method used was apriori algorithm.

A PRIORI ALGORITHM

The algorithm is based on the a *priori hypothesis*, viz., an itemset can only be frequent if all its subsets are also frequent. This information makes possible to reduce the search space efficiently when discovering the frequent itemsets, because using this information reduces the number of candidates. The Apriori algorithm is a level-wise method, which means that it discovers the *k*-itemsets during k^{th} database scan.

The algorithms for discovering large itemsets make multiple passes over the data. In the first pass, the algorithm counts the support of individual items and determines which of them are large, i.e. have minimum support. In each Subsequent pass, it starts with a seed set of itemsets found to be large in the previous pass. This seed set is used for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, it determines which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new large itemsets are found. Many algorithms for generating association rules were presented over time. Some well known algorithms are Apriori, Eclat and FP-Growth. Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support.

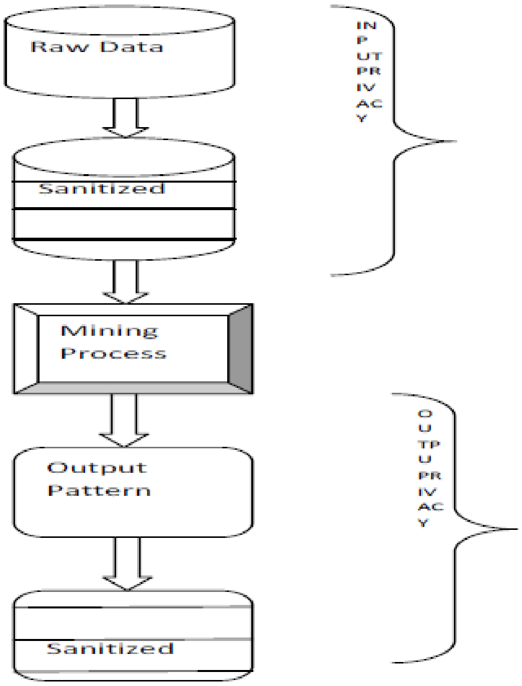

Figure 1. Framework of Privacy Preserving Data Mining

CHAPTER-2

LITERATURE SURVEY

2.1 RELATED WORK:

[1], the authors hide the sensitive knowledge in association rule mining, by selecting the sensitive rules manually.

[2] Authors introduced a heuristic approach to hide restrictive association rules that requires the sensitive items. The sensitive items have to be identified manually and it has to be removed from its transactions.
[3], the authors selected the sensitive attributes and hide the attributes by randomization process.

2.2 GAPS IN LITERATURE:

Data mining represents a wide range of tools and techniques to extract useful information which can contain sensitive information from a large collection of data. Data should be manipulated or distorted in such a way that information cannot be discovered through data mining techniques. Sensitive information has to be protected against unauthorized access.

Privacy preserving data mining considers the problem of data mining. Association rule hiding is one of the privacy preserving techniques for hiding sensitive association rules. For hiding sensitive association information in market basket database the sensitive items are given as input. Detection of sensitive items manually in market basket database containing thousands of transactions takes more computational time. So there is a need to find a method to detect the sensitive items.

2.3 OBJECTIVES OF RESEARCH:

1. MARKET BASKET ANALYSIS USING ASSOCIATION RULE MINING STRATEGY

The term market basket analysis using association rule learning in the retail business refers to the research that provides the retailer with information to understand the purchasing behaviour of a customer. This association information will enable the retailer to understand the customer's needs and rewrite the store's layout accordingly develop cross-promotional programs, or even capture new customers.

2. ASSOCIATION RULE MINING ALGORITHMS

The algorithms for discovering large itemsets make multiple passes over the data. In the first pass, the algorithm counts the support of individual items and determines which of them are large, i.e. have minimum support. In each subsequent pass, it starts with a seed set of itemsets found to be large in the previous pass.

This seed set is used for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, it determines which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new large itemsets are found.

A PRIORI ALGORITHM

The algorithm is based on the a *priori hypothesis*, viz., an itemset can only be frequent if all its subsets are also frequent. This information makes possible to reduce the search space efficiently when discovering the frequent itemsets, because using this information reduces the number of candidates. The Apriori algorithm is a level-wise method, which means that it discovers the *k*-itemsets during k^{th} database scan.

CHAPTER 3

PROPOSED METHOD

3.1 OVERVIEW: The proposed method first finds the number of occurrences of individual items. Using maximum number of frequent items, association rules are generated based on minimum support and minimum confidence. This approach uses the concept of representative rules. Having the frequent items on LHS of the rule, its consequents are combined in RHS. The consequents are selected as sensitive items. More number of sensitive items are selected by varying minimum support and minimum confidence, which are user specified values.

3.2 PROPOSED ALGORITHM :

Input:

(1) Database D
(2) Minimum support: min_supp
(3) Minimum confidence: min_conf

Output:

Sensitive item(s) for Sensitive Association rule hiding.

Algorithm:

1. Find the frequencies of individual items and store it in F.

2. Generate the association rules from D using minimum support and minimum confidence.

3. Select the single antecedent and consequent rules.

4. **{**

5. If F is empty then EXIT;

6. Select all the rules with min_supp containing f and
store in U//f on LHS

```
7. Repeat {
8. Select all the rules from U with same LHS
9. Join RHS of selected rules and store in R; //make them representative rules
10. }Until (U is empty);
11. Selected RHS will be the sensitive item(s).
```

Maximum occurrences of items are selected by user specified value in terms of percentage. The proposed algorithm first tries to find the frequent items and the association rules are generated using the Apriori algorithm [1]. These rules are used to form representative association rules having frequent items on LHS. The selected RHS are considered as sensitive items. The selected sensitive items are used to hide sensitive rules.
RR is a least set of rules that allow deducing all association rules without accessing a database.

Experimental Result:-

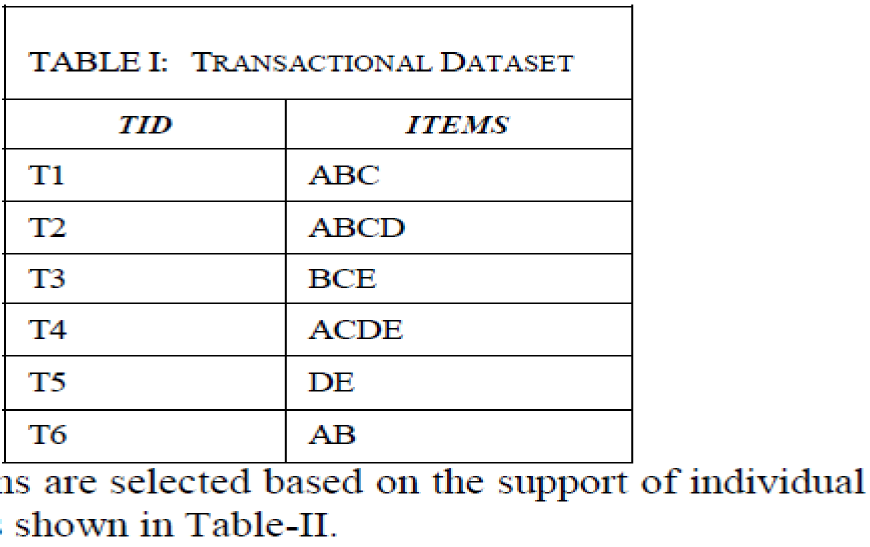

TABLE I: TRANSACTIONAL DATASET

TID	*ITEMS*
T1	ABC
T2	ABCD
T3	BCE
T4	ACDE
T5	DE
T6	AB

Frequent items are selected based on the support of individual items. This is shown in Table-II.

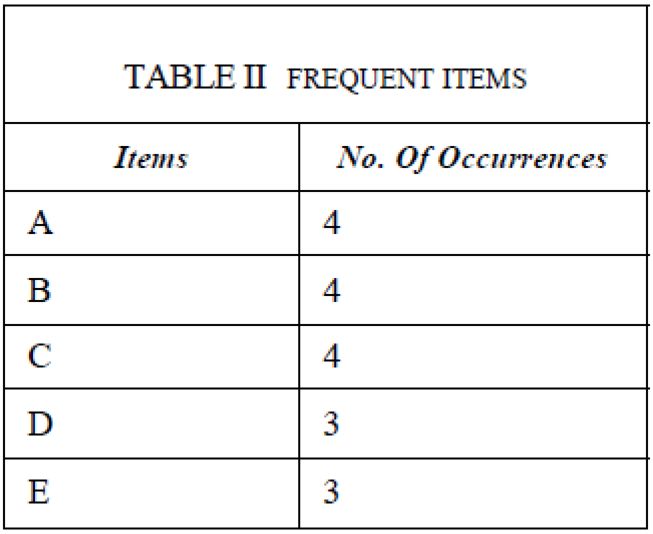

TABLE II FREQUENT ITEMS	
Items	*No. Of Occurrences*
A	4
B	4
C	4
D	3
E	3

For the Dataset given in Table - I min_supp is taken as 33% and a min_conf is taken as 70 % and frequent item F={A}.

Association rules are generated using apriori algorithm. The rules containing single antecedent and consequent are selected. All the rules containing ‘A’ either in RHS or LHS are chosen and represent them in representative rule format. Association rules are shown in Table III.

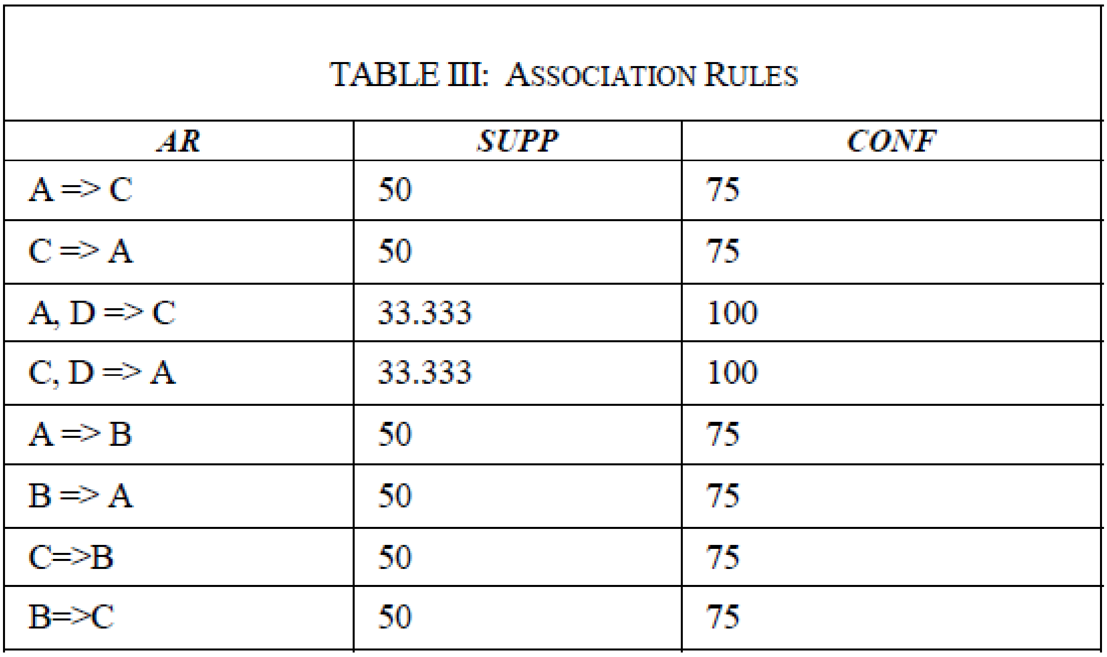

TABLE III: ASSOCIATION RULES

AR	*SUPP*	*CONF*
A => C	50	75
C => A	50	75
A, D => C	33.333	100
C, D => A	33.333	100
A => B	50	75
B => A	50	75
C=>B	50	75
B=>C	50	75

From these association rules, the rules containing frequent item A with single antecedent and consequent are selected. They are A=>C(50,75), C=>A(50,75), A,D=>C(33.33,100), C,D=> A(33,100), A=>B(50,75), B=>A(50, 75). From this rules set, select the rules that can be represented in the form of representative rules. Like the rules A=>B and A=>C can be represented as

A=> BC

In the above representative rules B and C are considered as sensitive items. Similarly the frequent items B and C are taken and representative association rules are formed. From the representative association rules, sensitive item A is detected. Hence, for the given transactional data set A, B and C are detected as sensitive items. More number of sensitive items can be selected by varying minimum support and minimum confidence of frequent items.

Conclusion:-

Market basket analysis is an essential activity to promote business. Buyer behavior patterns are discovered through market basket analysis. The knowledge discovered may reveal trade secrets also. The data owner may lose his business secrets. Privacy preserving data mining hides certain sensitive information, so that they cannot be discovered through data mining techniques. Association rule hiding algorithm sanitizes the database such that no sensitive association rule is derived from it. For hiding sensitive association rules, sensitive items are selected manually. In market basket database the selection of sensitive items takes more computational time. The proposed algorithm detects the sensitive items using a set of representative association rules. It is experimented with sample data sets. Further research work is in progress to hide the sensitive association rules using the selected sensitive items.

REFERENCES

[1] Agrawal R, Imielnski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Washington DC

[2] Agrawal R, Srikant R (2000) Privacy preserving data mining. In ACMSIGMOD Conference on Management Of Data, Dallas, Texas, pp 439–4501.

[3] Clifton C, Marks D (1996) Security and privacy implications of datamining. In: SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery.

[4] Dasseni E, Verykios V, Elmagarmid A, Bertino E (2001) Hiding association rules by using confidence and support. In: Proceedings of 4th Information Hiding Workshop, Pittsburgh, PA, pp 369– 383.

[5] Evfimievski A (2002) Randomization in privacy preserving data mining. SIGKDD Explorations 4(2), Issue 2:43–48.

[6]Oliveira S, Zaiane O (2003 a) Algorithms for balancing privacy and knowledge discovery in association rule mining. In: Proceedings of7th International Database Engineering and Applications Symposium (IDEAS03), Hong Kong

[7] Oliveira S, Zaiane O (2003 b) Protecting sensitive knowledge by data sanitization. In: Proceedings of IEEE International Conference on Data Mining.

[8] Wenliang Du , Zhijun Zhan(2003)Using Randomized Response Techniques for Privacy Preserving Data mining.SIGKDD'03

[9] Dr. K.Duraisamy, N.Maheswari(2008) Identification of Sensitive Items in Privacy Preserving - Association Rule Mining,CCSE journal.