# Mining E-Shoppers's Purchase Rule and Protecting Outsourced Database

S.N.Saravana Sai[1], T.Subha[2]

[1]*PG Scholar, M.E – Computer and Communication, Sri Sairam Engineering College, Chennai*

[2]*Associate Professor, Department Of Information Technology, Sri Sairam Engineering College, Chennai*

[1]saravanasai10@gmail.com

[2]subharajan@gmail.com

***Abstract-*** **Data mining is widely used in many applications. Association rule mining is an important data mining tool used to extract decision making information in business. It discovers correlations between different item sets in a transaction database. The database is the private property of the organization; the database owner may outsource the mining task to a third party service provider. While outsourcing, the database is sent to a service provider and the service provider computes and returns the association rules for the database owner; here service provider may be dishonest. In such cases, we need to satisfy two issues:-preventing dishonest party from (1)Stealing the information from database (2)corrupting the mining results. In this paper, we created a cloud based retail portal to response to dynamic needs of customer and we discuss security issues in outsourcing of association rule. Also our approach proposes a more secure encryption schema that transforms transaction non-deterministically and also uses an efficient data mining algorithm to find e-shopper's purchase rule**.

***Keywords*** **- Data mining; Association rule; Apriori algorithm; Service provider; Outsourcing.**

## I. INTRODUCTION

Due to emerging of cloud computing and its computational paradigms, the outsourcing of computing services acquires a novel relevance. Using cloud various advanced analytical services like business intelligence and knowledge discovery services can be outsourced. The organization lacking in computational resources can use this services in cost-effective way. Although this outsourcing service brings a great computational relief, there exists a serious security issue. While outsourcing the service to a third party, he has the access to the data and may learn sensitive information from it. For ex:- if a supermarket owner outsources his database for mining operation, the third party can look at the transaction and can learn which items are always co-purchased. This kind of information about organizations has to be secured and the problem of protecting this business information is referred as corporate privacy.

While outsourcing, both the original datasets and mined patterns has to be secured from third party service provider. In the case of a supermarket database the datasets are public, but the computed association rules are the property of the owner and have to be protected. Therefore protecting both the raw data and resulting association rule from the service provider is the key issue in outsourcing. There are two approaches to achieve this key issue;

1. Applying encryption function on original database and transforming it into encrypted database and outsourcing it.
2. Applying data perturbation and modifying original database. In perturbation process, database is subjected to random perturbation [4] this approach reduces the data mining efficiency. In this paper we study the problem of outsourcing the data mining task and we discuss an encryption schema which ensures corporate privacy.

## II. BACKGROUND AND DEFINITIONS

### *A. Data Mining*

Due a vast development in computer hardware technology, various powerful information processing system and vast storage capacity are developed. Due to this development, huge amount of data's collected on regular basis by many organizations. Data mining technology is used to extract fruitful information from these gigantic datasets. Data mining techniques used in retail business organization is discussed in this section.

#### *1) Market basket analysis:*

Market basket analysis is the process of analyzing the buying habit of customers to find relationship between different items in their

shopping cart. The discovery of these relationships will help a seller to develop sales strategy based on the frequently purchased items. For example, if a buyer buys bread, how likely he will buy milk on that same transaction. Also this analysis will help the seller to design a better display layout with associated items placed together, so the customers can more likely purchase them together. For these reasons the ability to predict e-purchase rule using data mining has become competitive tool for retail organizations.

*2) Association rule:*

Association rule mining finds interesting relationship among a set of large items. Since large amount of data's are collected and stored continuously in a database, many industries are interested in mining out associated items from their database. This discovery of associative relationships among huge data sets can help in decision making processes [1]. Meanwhile in determining association rules, support and confidence are the two measures that restrict whether a rule is interesting or not [2].Moreover the association rule will be considered of interest if it satisfies both a minimum support threshold and a minimum confidence threshold support [3].

**Support**: This is a measure which indicates level of dominance of each item in overall transaction

**Confidence**: This measure indicates the relationship among items (e.g. how frequently a customer buys item y if the customer buying item x)

**Example:** buy (x, "bread") →buy (x, "milk") [support = 40%, confidence = 80%] Here bread and milk are purchased simultaneously by 40% of all transactions and 80% of all consumers who buy bread also buy milk.

The common approach to find association rules follows two steps:

1. Finding the frequent itemset among a large itemsets.
2. Generating strong association rule from the frequent itemset

*B. Cloud Computing*

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network(typically the Internet).Many business organization uses the cloud services to rent application software and databases. The Service provider manages the Cloud infrastructure on which applications runs and user data's are stored on servers; which are located in remote location. The users can access the cloud-based application using a web browser or mobile apps.

The Cloud service providers usually offers services based on three fundamental models:-

1) In infrastructure as a service model, the service provider owns equipments for supporting storage, hardware, servers and networking features and the user uses this support on pay-per-use basis.

2) Platform as a Service (PaaS) provides a way to rent hardware, operating systems, storage and network capacity over the Internet fundamental models [10]

3) Infrastructure as a service (IaaS) manages all the storage services in cloud computing

Software as a Service (SaaS) is a software distribution model in which applications are hosted by a service provider and made available to customers over a network, typically the Internet.

## III. RELATED WORK

The research in privacy preserving data mining has caught much attention. A stream of past researches that focused on protecting data from third party service provider [5, 6].The main issue in this model is that private data's collected from many sources by a single collector who may not be trusted. So the data's are randomly transformed by applying data perturbation technique [4].Although this approach ensures privacy; this is not suited for performing analytical operations over it.

Another related issue is secure mining over distributed dataset. Here data on which mining has to be done is partitioned horizontally or vertically and this partitioned data's are distributed among multiple parties [5]. These multiple parties perform mining operation and share their results. This model ensures the privacy for original database but not for the mined results.

There are many privacy protecting techniques developed for statistical database, but they do not take into account required specification to data mining application. Also the perturbation techniques used affects the prediction accuracy and logical rules.

## IV. PROBLEM DESCRIPTION

This project aims to contribute the solutions to three specific problems: -

1. The problem of securely outsourcing database to third party service provider
2. Using effective algorithm to perform data mining task.
3. Providing a dynamic business environment.

Let D denotes the original database and $D^*$ denotes encrypted database. To ensure privacy guarantee for encrypted database we defines:

*A. Definition:*

We consider the data sets in a database as 2-dimensional table where row corresponds to

individual record and column corresponds to properties of that individual. The statistical knowledge are mined using this attributes, thus our problem is to protect this attribute sets.

Given a original database D we can transform it into encrypted database $D^*$, we can say $D^*$ is K-Private only if

1) Probability of each cipher item e is Prob(e)<1/k;

2) Probability of Cipher itemset E is Prob(E)<1/k.

Before outsourcing the database the user constructs K-Private cipher database $D^*$ by using encryption techniques.

## V. PROPOSED SYSTEM

To overcome the existing privacy issues our proposed model aims to devise an encryption schema to ensure the formal privacy guarantee and to validate this model over a large real life transaction database. Our approach aims on preserving both the original outsourced data and the mined results from third party. Also the decision patterns obtained from original database should be similar to one obtained from transformed database.

### *A. System Architecture*

The architecture behind our model is illustrated in fig 1.

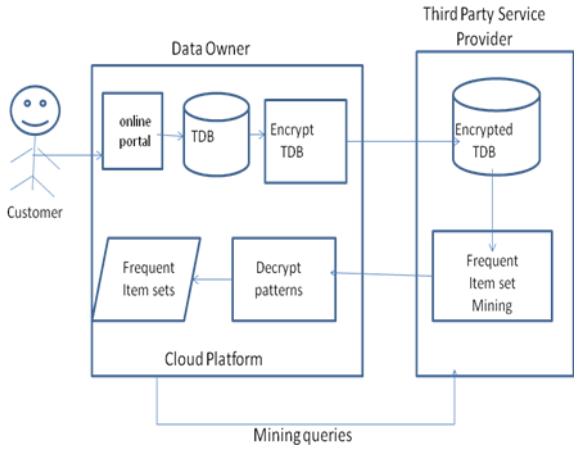


Fig. 1 System Architecture

In our model we use Business to Consumer (B2C) Online shopping portal to collect the consumers purchase itemsets. The data owner encrypts the TDB before outsourcing it to third party .The third party conducts data mining operation over encrypted TDB and returns the associated patterns (encrypted) to owner. The data owner decrypts it to extract original frequent item sets. The development modules for our system are as follows:

### *B. Creating a Cloud Framework*

In this development module we created a framework to response to the dynamic shopping needs of the customer. Using this framework the client can store their business environment in cloud platform. This cloud deployment makes the business environment easily accessible from anywhere and at any time. The customer can access cloud-based online portal through a web browser, while the business software and user's data are stored on servers at a remote location. This work serves as an initial step of developing privacy for business intelligence framework.

### *C. Data Transformation*

Outsourcing offers a great benefits like cost relief and minimizing demand resources, but a dishonest third party may (i) steal the sensitive information from database, (ii) identify frequent patterns and may reveal it to competitors.

A simple encryption technique to overcome the above security issue is of using substitution cipher [7][8]. Each transaction in database is a set of items, for ex: - $t_i$ = {jam, butter} is a transaction; here we can replace each item with an unknown symbol. So here if the transaction ti is replaced as ti = {24, 43},one cannot able to identify original items without knowing that jam is replaced as 24 and butter is replaced as 43.This idea of replacing an item by a unique symbol is known as One-to-one item mapping. But this method is vulnerable to frequent analysis and thus not suited for corporate privacy.

One possibility to enhance the security is to use one-to-n item mapping. For example in the same transaction ti = {jam, butter}, the item jam can be substituted by a set of integers (e.g. {24, 26}) and similarly butter can be substituted with a set of integers. This method ensures security but in order to perform decryption unambiguously, the item should atleast contain one unique symbol.

To overcome these security issues we developed a data transformation module which provides nondeterministic one-to-n data transformation schema. Apart from one-to-n transformation we also use fake items that are injected along with the transaction. Adding of this fake transaction won't affect the association rule mining [7].Even if the attacker knows the construction method of fake transaction, he won't able to distinguish the fake items from real, since the attacker does not have knowledge of frequency of item sets. This fake transaction has to be maintained by owner for later recovery of real supports. In order to reduce this storage overhead synopsis is used which stores all information needed on fake transaction. This synopsis is implemented by using a minimal perfect hash function [9].

### *D. Frequent itemset mining*

In our approach, the frequent pattern mining is done on the outsourced database (encrypted)

using Apriori Algorithm. Apriori algorithm is used to find frequent itemset and it is widely used in sales-purchase domain. Apriori involves an iterative level-wise search in which k-items are used to explore (k+1)-itemsets. Here initially the 1st frequent itemset is found. This is denoted as $L_1$ .$L_1$ is used to find the 2nd frequent itemset $L_2$, again this $L_2$ is used to find $L_3$ and so on, until no more frequent k-itemsets can be found. To find this $L_k$ whole database has to be inspected. The Apriori algorithm is as follows [2]

*Ck*: Candidate itemset of size k
*Lk*: frequent itemset of size k
*L1*= {frequent items};
**for**(*k*= 1; *Lk*!=∅; *k*++) **do begin**
*Ck+1*= candidates generated from *Lk*;
**for each**transaction *t*in database do
increment the count of all candidates in *Ck+1*that are contained in *t*
*Lk+1*= candidates in *Ck+1*with min_support
**end**
**return**∪*kLk*;

One example for application of Apriori algorithm is considered below [11]:

Transaction D

| TID | Items |
|---|---|
| 100 | 1,3,4 |
| 200 | 2,3,5 |
| 300 | 1,2,3,5 |
| 400 | 2,5 |

Scan D ⇒

$C_1$

| Itemset | Count |
|---|---|
| {1} | 2 |
| {2} | 3 |
| {3} | 3 |
| {4} | 1 |
| {5} | 3 |

$L_1$

| Itemset | Count |
|---|---|
| {1} | 2 |
| {2} | 3 |
| {3} | 3 |
| {5} | 3 |

$C_2$

| Itemset |
|---|
| {1,2} |
| {1,3} |
| {1,5} |
| {2,3} |
| {2,5} |
| {3,5} |

Scan D ⇒

$C_2$

| Itemset | Count |
|---|---|
| {1,2} | 1 |
| {1,3} | 2 |
| {1,5} | 1 |
| {2,3} | 2 |
| {2,5} | 3 |
| {3,5} | 2 |

$L_2$

| Itemset | Count |
|---|---|
| {1,3} | 2 |
| {2,3} | 2 |
| {2,5} | 3 |
| {3,5} | 2 |

$C_3$

| Itemset |
|---|
| {2,3,5} |

Scan D ⇒

$C_3$

| Itemset | Count |
|---|---|
| {2,3,5} | 2 |

$L_3$

| Itemset | Count |
|---|---|
| {2,3,5} | 2 |

After obtaining the support count values the measure of associated relationship can be found by using the below equation [11]:

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

Here supp_count(A U B) denotes the number of transaction that contains itemset A U B and supp_count(A) denotes number of transaction that contains itemset A.

## VI. EXPERIMENTAL ANALYSIS

We perform an analysis that demonstrates the security of an outsourced database. In our approach we implement our security schema in a large-real world database. Our experiment results shows that decision tree obtained from the original data sets and the transformed data set are very similar in terms of logical rules. All the coding works are done in java. The experiment is done on an Intel core2 duo processor with 2GB RAM running over windows platform.

## VII. CONCLUSION

In this paper the problem of privacy-preserving of associated patterns on an encrypted outsourced TDB is analyzed. Privacy preserving data mining (PPDM) is a novel research direction to preserve privacy for sensitive knowledge from disclosure. Integrating this approach with dynamic shopping framework, the true buyer's pattern can be collected and also efficient user accessibility is achieved. We proposed a sophisticated encryption technique to ensure the privacy of outsourced database. Our work extracts similar decision tree from original and transformed datasets. We haven't considered any attack models in our work. It would be interesting if our work proves robust against adversarial attack; further steps will be taken to safeguarding outsourced database from third party server attacks.

## VIII. REFERENCES


[1]. *Peltier J W,Schibmwsky J A,Schuhz D E,et al. "Interactive Psychographics:Cross-Selling in the Banking Industry" .Journal of Advertising Research, 2002, 4 (2) ,pp.7-22.*
[2]. *H. Jiawei and K. Micheline, "Data Mining: Concepts and Techniques". Morgan Kaufmann, 2001.*
[3]. *Gy_rödi, R. Gy_rödi. "Mining Association Rules in Large Databases". Proc. of Oradea EMES'02: 45-50, Oradea, Romania, 2002.*
[4]. *R. Agrawal and R. Srikant, "Privacy-preserving data mining," in Proc.*
[5]. *M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Trans. Knowledge Data Eng., vol.16, no. 9, pp. 1026–1037, Sep. 2004.*
[6]. *B. Gilburd, A. Schuster, and R. Wolff. "A new privacy model and association-rule mining algorithm for large-scale distributed environments". In VLDB, 2005.*
[7]. *R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases, SIGMOD, 1993.*
[8]. *R. Agrawal and R. Srikant. "Fast algorithms for mining association rules", VLDB, 1994.*
[9]. *T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to Algorithms" Cambridge, MA: MIT Press, 2001*
[10]. *wikipedia.org/wiki/Cloud_computing*
[11]. *D.H.Setiabudi, G.S.Budhi "Data Mining Basket Analsis using Hybrid dimension Association Rule"*

 