International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
ISSN (Online) 2349-6967
Volume 2 , Issue 1(Jan-Feb 2015), PP97-104

PERFORMING DATA MINING IN (SRMS) THROUGH VERTICAL APPROACH WITH ASSOCIATION RULES

Mr. Ambarish S. Durani(1)
MTech (IIIrd Sem),Vidharbha Institute of Technology, Nagpur (ambarish007.durani@gmail.com)

Miss. Rashmi B. Sune(2)

Lecturer, Department of Computer, SFS School,
Badnera, sune.rashmi@gmail.com

Abstract:- This system technique is used for efficient data mining in SRMS (Student Records Management System) through vertical approach with association rules in distributed databases. The current leading technique is that of Kantarcioglu and Clifton [1]. In this system I deal with two challenges or issues, one that computes the union of private subsets that each of the interacting users hold, and another that tests the inclusion of an element held by one user in a subset held by another. The existing system uses different techniques for data mining purpose like Apriori algorithm. The Fast Distributed Mining (FDM) algorithm of Cheung et al. [2], which is an unsecured distributed version of the Apriori algorithm. Proposed system offers enhanced privacy and data mining with respect to the Encryption techniques and Association rule with Fp-Growth Algorithm in private cloud (system contains different files of subjects with respect to their branches). Due to this above techniques the expected effect on this system is that, it is simpler and more efficient in terms of communication cost and combinational cost. Due to these techniques it will affect the parameter like time consumption for execution, length of the code is decrease, find the data fast, extracting hidden predictive information from large databases and the efficiency of this proposed system should increase by the 20%.

I. INTRODUCTION

This propose system technique is used for efficient data mining in SRMS (Student Records Management System) through vertical approach with association rules in distributed databases. The current leading technique is that of Kantarcioglu and Clifton. This proposed system is ready to implements two methods, one that computes the union of private subsets that each of the interacting users hold, and another that tests the inclusion of an element held by one user in a subset held by another. Today in organizations, the developments in the transaction processing technology requires that, amount and rate of data capture should match the speed of processing of the data into information which can be utilized for decision making. A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data that is required for decision making process. Data mining involves the use of various data analysis tools to discover new facts, valid patterns and relationships in large data sets. Data mining also includes analysis and prediction for the data. Data mining helps in extracting meaningful new patterns that cannot be found just by querying or processing data or metadata in the data warehouse. This paper includes need for data warehousing and data mining, how data warehousing and mining helps decision making systems, Knowledge Discovery process and various techniques involve in data mining.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data Mining is a process of extracting hidden predictive information from large databases. It is a powerful new technology to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. For a commercial business, the discovery of previously unknown statistical patterns or trends can provide valuable insight into the function and environment of their organization. Data-mining techniques can generally be grouped into two categories: predictive method and descriptive method. Descriptive method: It a method of finding human interpretable patterns that describe the data. Data mining in this case is useful to group together similar documents returned by search engine according to their context. Predictive method: In this method, we can use some variables to predict unknown or future values of other variable. It is used to predict whether a newly arrived customer will spend more than 100$ at a department store.

Data-mining techniques:

The following list describes many data-mining techniques in use today. Each of these techniques exists in several variations and can be applied to one or more of the categories above.

- Regression modeling—This technique applies standard statistics to data to prove or disprove a hypothesis. One example of this is linear regression, in which variables are measured against a standard or target variable path over time. A second example is logistic regression, where the probability of an event is predicted based on known values in correlation with the occurrence of prior similar events.
- Visualization—This technique builds multidimensional graphs to allow a data analyst to decipher trends, patterns, or relationships.
- Correlation—This technique identifies relationships between two or more variables in a data group.
- Variance analysis—This is a statistical technique to identify differences in mean values between a target or known variable and nondependent variables or variable groups.
- Discriminate analysis—This is a classification technique used to identify or “discriminate” the factors leading to membership within a grouping.
- Forecasting—Forecasting techniques predict variable outcomes based on the known outcomes of past events.
- Cluster analysis—This technique reduces data instances to cluster groupings and then analyzes the attributes displayed by each group.
- Decision trees—Decision trees separate data based on sets of rules that can be described in “if-then-else” language.
- Neural networks—Neural networks are data models that are meant to simulate cognitive functions. These techniques “learn” with each iteration through the data, allowing for greater flexibility in the discovery of patterns and trends.

Distributed database is a technique which is used to collect the data from different location, i.e. the data is not store in only one system it is distributed through various system. It consists of several computers that do not share a memory or a clock. The computers communicate with each other by exchanging messages over a communication network and each computer has its own memory and runs its own operating system. The advantages of the Distributed operating system is, Resource sharing, Enhance performance, Improved reliability and availability and Modular expandability. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. The main objective is to minimize data with encryption and decryption techniques through association rule. The mining of data is performed with the association rule. The data is gathered /stored in encrypted format in under the specific file format. Administrator performed the union of the collected encrypted data and stored in one main table. It is simpler and is significantly more efficient in terms of combinational cost and communication cost.

II. The aim of this study is covered:

1) Provide an overview of existing techniques that can be used for extracting of useful information from databases.

2) Provide a feature classification technique that identifies important aspects to study knowledge discovery.

3) Investigate existing knowledge discovery and data mining software tools using the proposed feature classification scheme.

II. PROBLEM STATEMENT

Previous work in privacy preserving data mining has considered two related settings. One, in which the data owner and the data miner are two different entities, and another, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold. In the first setting, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonymizing the data prior to its release. The main approach in this context is to apply data perturbation. The idea is that the perturbed data can be used to infer general trends in the data, without revealing original record information.

In the second setting, the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners. This is a problem of secure multiparty computation.

The existing research problems that this study suggests is the implementation of the techniques presented here to the problem of distributed association rule mining in the vertical setting, the problem of mining generalized association rules, and the problem of subgroup discovery in horizontally partitioned data.

III. LITERATURE SURVEY / RELATED WORK

1. Secure Mining of Association Rules in Horizontally Distributed Databases by Tamir Tassa
He proposes a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton. Our protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm of Cheung et al., which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms — one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol in. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

2. Security in Outsourcing of Association Rule Mining by W. K. Wong & David W. Cheung
Outsourcing association rule mining to an outside service provider brings several important benefits to the data owner. These include (i) relief from the high mining cost, (ii) minimization of demands in resources, and (iii) effective centralized mining for multiple distributed owners. On the other hand, security is an issue; the service provider should be prevented from accessing the actual data since (i) the data may be associated with private information, (ii) the frequency analysis is meant to be used solely by the owner. This paper proposes substitution cipher techniques in the encryption of transactional data for outsourcing association rule mining. After identifying the non-trivial threats to a straightforward one-to-one item mapping substitution cipher, we propose a more secure encryption scheme based on a one-to-n item mapping that transforms transactions non-deterministically, yet guarantees correct decryption. We develop an effective and efficient encryption algorithm based on this method. Our algorithm performs a single pass over the database and thus is suitable for applications in which data owners send streams of transactions to the service provider. A comprehensive cryptanalysis study is carried out. The results show that our technique is highly secure with a low data transformation cost.

3. An Overview of Secure Mining of Association Rules in Horizontally Distributed Databases by Ms. Sonal Patil, Mr. Harshad S. Patil

In this paper, propose a protocol for secure mining of association rules in horizontally distributed databases. Now a day the current leading protocol is Kantarcioglu and Clifton. This protocol is based on the Fast Distributed Mining (FDM) algorithm which is an unsecured distributed version of the Apriori algorithm. The main ingredients in this protocol are two novel secure multi-party algorithms 1. That computes the union of private subsets that each of the interacting players hold, and 2. Tests the inclusion of an element held by one player in a subset held by another. In this protocol offers enhanced privacy with respect to the other one. Differences in this protocol, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost. The existing system uses different techniques for data mining purpose like Apriori algorithm. The Fast Distributed Mining (FDM) algorithm of Cheung et al. [2], which is an unsecured distributed version of the Apriori algorithm.Data mining is not particularly new — statisticians have used similar manual approaches to review data and provide business projections for many years. Changes in data mining techniques, however, have enabled organizations to collect, analyze, and access data in new ways. The first change occurred in the area of basic data collection. Before companies made the transition from ledgers and other paper-based records to computer-based systems, managers had to wait for staff to put the pieces together to know how well the business was performing or how current performance periods compared with previous periods. As companies started collecting and saving basic data in computers, they were able to start answering detailed questions quicker and with more ease.

That goal defines a problem of secure multi-user computation. In such problems, there are N users that hold private inputs of data, $x1, \ldots, xN$, and they wish to securely compute $y = f(x1, \ldots, xN)$ for some public function f [1]. If there existed a trusted third party, the users could submit to him their inputs and he would perform the function evaluation and send to them the resulting output. In the absence of such a trusted third party, it is needed to devise techniques that the users can run on their own in order to arrive at the required output y. The next goal is to secure the inputs of each user. If the both are combined together (data mining and Secure) the third party involvement is avoided. In our problem, the inputs are the partial databases, and the required output is the list of association rules that hold in the unified database with support and confidence no smaller than the given thresholds s and c, respectively. They can be applied only to small inputs and functions which are realizable by simple circuits. For mining of data and encryption/decryption different techniques are available. Like for data mining K-means algorithms, Apriori Algorithm Fast Distributed Mining and for encryption/Decryption RSA, DES.

Association Rule Mining: In data mining, association rule Learning is a popular and well researched method for discovering interesting relations between variables in large databases. It analyzes and present strong rules discovered in databases using different measures of interestingness. Based on the concept of Strong, rules, Agrawal et al., introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

Privacy protection is an important issue in data mining. Businesses usually do not want to share their own private (statistical) information with service providers[32] .A stream of past research has focused on protecting privacy against third-party players in distributed data mining [7,8]. These studies are not directly related to the problem, since our interest is to protect data and results against a service provider, who alone should perform the mining task. Different data mining models (such as association-rule mining, decision tree classification, etc.) have different security requirements. Specialized approaches for the protection of sensitive information under different models have been designed [9, 4]. In our case, we need a simple scheme that enables a third-party miner to find association rules in a transformed database while preventing it from accessing the private information in the original database. The perturbation technique proposed in [6] can be adapted for our problem.

However, this solution returns only approximate results. Also, the miner knows the exact number of itemsets (and their frequencies) that will be used by the owner (i.e., the actual support of the perturbed itemsets). This can be considered a breach of privacy. On the other hand, our data encryption technique ensures the accuracy of the results and at the same time hides from the miner the original number of items and the exact frequencies of the itemsets. Substitution cipher is a well-known method that is used in the encryption of plain text. Each letter in the text is replaced by another letter. Although the number of possible substitutions (i.e., letter orderings) is very large (26!), the encryption is in fact not difficult to break if a dictionary of words with their expected frequencies is available [10]. Due to the extreme size of the search space, adversaries usually resort to local search techniques to break the cipher. Genetic algorithms [35] have been widely used in cryptanalysis to attack security schemes [5, 10]. In order to validate the security of our encryption scheme, we evaluate its resiliency to attacks by a genetic algorithm

In recent years the sizes of databases has increased rapidly. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data. The term Data Mining, or Knowledge Discovery in Databases, has been adopted for a field of research dealing with the automatic discovery of implicit information or knowledge within databases.

Mining Algorithm Strategies

The association rules are considered interesting if they satisfy both a *minimum support* threshold and a *minimum confidence* threshold. A more formal definition is the following . Let X = {i1, i2, …, im} be a set of items. Let *D*, the task- elevant data, be a set of database transactions where each transaction *T* is a set of items such that $T \subseteq X$. Each transaction is associated with an identifier, called TID. Let *A* be a set of items. A transaction *T* is said to contain *A* if and only if $A \subseteq T$. An association rule is implication of the form $A \Rightarrow B$, where $A \subset X$, B $\subset X$, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set *D* with *support s*, where *s* is the percentage of transactions in *D* that contain $A \cup B$ (i.e., both *A* and *B*). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has *confidence c* in the transaction set *D* if *c* is the percentage of transactions in *D* containing *A* that also contain *B*. This is taken to be the conditional probability, *P(B|A)*. That is,

$$support(A \Rightarrow B) = P(A \cup B) \qquad (1)$$

$$confidence(A \Rightarrow B) = P(B \mid A) \qquad (2)$$

The definition of a frequent pattern relies on the following considerations.A set of items is referred to as an itemset (pattern). An itemset that contains *k* items is a *k*-itemset.

For example the set {*name*, *semester*} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of itemset. An itemset satisfies *minimum support* if the occurrence frequency of the itemset is greater than or equal to the product of *minimum support* and the total number of transactions in D. The number of transactions required for the itemset to satisfy *minimum support* is therefore referred to as the *minimum support count*. If an itemset satisfie*s minimum support*, then it is a *frequent itemset* (*frequent pattern*). The most common approach to finding association rules is to break up the problem into two parts are

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.

2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence [11].

Encryption Algorithm Strategies

Cloud Searchable Strong Encryption (SSE) encryption delivers the benefits of the cloud, while assuring security and compliance for your most sensitive information. Cloud makes it easy to protect any type of data with standards-based encryption that only you can unlock – because no one else can access your encryption keys. And Cloud does this without disabling your applications – maintaining business-critical functions while keeping your data fully protected. The cloud data and services reside in massively scalable data centers and can be accessed everywhere. The growth of the cloud users has unfortunately been accompanied with a growth in malicious activity in the cloud. More and more vulnerabilities are discovered, and nearly every day, new security advisories are published. We propose a simple data protection model where data is encrypted using Advanced Encryption Standard (AES) before it is launched in the cloud, thus ensuring data confidentiality and security.

IV. METHODOLOGY / PROPOSED WORK

The SRMS contain different methodology to perform different works.

1. Data Mining

2. Vertical Approach for finding the required data.

3. Association Rule

The SRMS contain different modules to perform different works. The specific use of every module is described as follows.

Modules:

Login Module:

Homepage is entry point of the system. In this module Login control is used for authenticating the user as well as Administrator. New user is adding with the registration control use by admin.

Private Cloud Module:

The Static information related to different branches is stored at cloud side. **Ex**. Student ID, Name, Faculty Name, ID) and Marks are stored in to cloud with their respective format as marks is considered as important entity

Step 1: Select semester for what user wants to store/add a data for different students.

Step 2: Data is stored in cloud with the help of encryption techniques for security purpose, and to avoid unauthorized access of the data. All the data from different sources is gathered or saved in one single folder which is located in cloud.

Administrator Work Module:

The vital work of the admin is to provide privileges' for different uses who logged in to their private cloud. Administrator has all the access for managing the user from the different branches. And he can perform the other operation as, Add user, students, exams, semester, and subject allocations. Admin collect all the encrypted data then it performs the union on the entire data with applying fp-growth association rule for mining purpose. Mining data is decrypted and served at the main server.

Process log:

For calculating the communication cost and combinational cost one "process.log" file maintain for storing the all cost it contains

- Total Communication cost
- Total Combinational Cost

One "Log file" folder is used for the maintains the Users record.

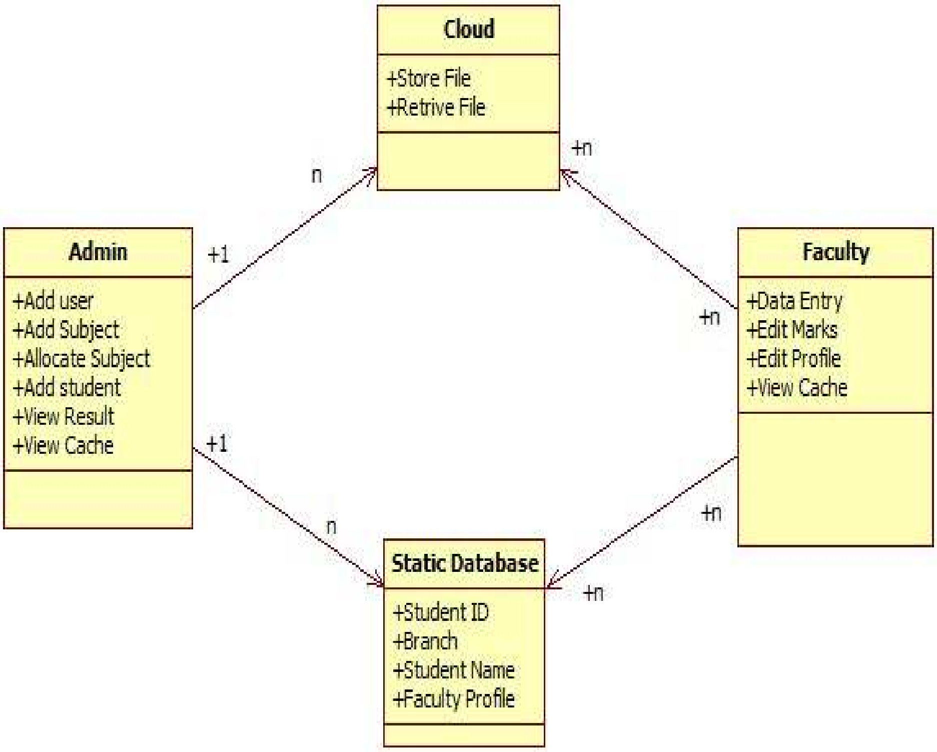

Fig. 1

V. REFERENCES

[1]. Tamir Tassa.“Secure Mining of Association Rules in Horizontally Distributed Databases,” In *IEEE Transactions On Knowledge And Data Engineering,*

[2]. R. Fagin, M. Naor, and P. Winkler.”Comparing Information Without Leaking It,” *Communications of the ACM*, 39:77–85, 1996

[3]. Murat Kantarcıoˇglu and Chris Clifton,” Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,” in *Ieee Transactions On Knowledge And Data Engineering, To Appear, 29 Jan. 2003; revised 11 Jun. 2003, accepted 1 Jul. 2003.*

[4]. D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu, “Efficient mining of association rules in distributed databases,” *IEEE Trans. Knowledge Data Eng.*, vol. 8, no. 6, pp. 911–922, Dec. 1996.

[5]. D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Santa Barbara, California, USA: ACM, May 21-23 2001, pp. 247–255. [Online]. Available: http: //doi.acm.org/10.1145/375551.375602

[6]. Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology – CRYPTO 2000*. Springer-Verlag, Aug. 20-24 2000, pp. 36–54. [Online]. Available: http://link.springer.de/link/ service/series/0558/bibs/1880/18800036.htm

[7]. A. C. Yao, "How to generate and exchange secrets," in *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*. IEEE, 1986, pp. 162–167.