International Journal of Advance Research in Science and Engineering
Vol. No. 5, Issue No. 08 , August 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SECURE WEB MINING

Jayalakshmi[1], Dr.M.Mohammed Ismail[2], P. Rizwan Ahmed[3]

[1] *Research Scholar, Computer Science, Mazharul Uloom College, Ambur, Tamil Nadu, (India)*

[2,3] *Associate Professor & Head , Research Department of Computer Science, Mazharul Uloom College, Ambur, Tamil Nadu, (India)*

ABSTRACT

Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Data mining and data warehousing go hand-in-hand: most tools operate by gathering all data into a central site, then running an algorithm against that data. However, privacy concerns can prevent building a centralized warehouse – data may be distributed among several custodians, none of which are allowed to transfer their data to another site. This project addresses the problem of computing association rules within such a scenario. We assume homogeneous databases: All sites have the same schema, but each site has information on different entities. The goal is to produce association rules that hold globally, while limiting the information shared about each site. In this project, a new protocol is proposed which combines the advantages of the existing approaches to perform privacy preserving in distributed mining of association rules. Both the privacy and performance characteristics of the proposed protocol are studied and compared with the mining and cryptographic approaches.

Keywords: Data Mining, Web Mining

I. OBJECTIVE

Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set. Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Data sharing can bring a lot of advantages for research and business collaboration. However, large repositories of data contain private data and sensitive rules that must be preserved before published. Motivated by the multiple conflicting requirements of data sharing, privacy preserving and knowledge discovery, privacy preserving data mining (PPDM) has become a research hotspot in data mining and database security fields. Two problems are addressed in PPDM: one is the protection of private data; another is the protection of sensitive rules (knowledge) contained in the data.

International Journal of Advance Research in Science and Engineering
Vol. No. 5, Issue No. 08 , August 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

II. DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

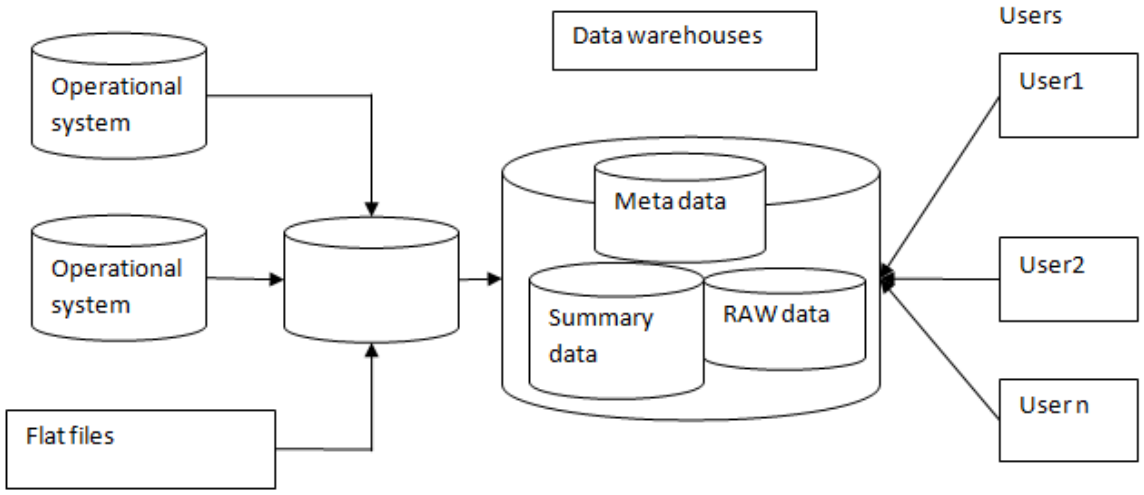

Fig 1: Process of Data Mining

2.1 Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- Nonoperational data, such as industry sales, forecast data, and macro economic data
- Meta data - data about the data itself, such as logical database design or data dictionary definitions

2.2 Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

2.3 Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

III. DATA WAREHOUSES

In computing, a data warehouse (DW or DWH) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from multiple disparate sources. Data warehouses store current as well as historical data and are commonly used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The data stored in the warehouse are uploaded from the operational systems (such as marketing, sales etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before they are used in the DW for reporting.

The typical ETL-based data warehouse uses staging, integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups often called dimensions and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.

Data mining elements

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new

(unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

3.1 Data Mining Techniques

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine those data mining techniques with example to have a good overview of them.

3.2 Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

3.3 Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that “given all past records of employees who left the company, predict which current employees are probably to leave in the future.” In this case, we divide the employee’s records into two groups that are “leave” and “stay”. And then we can ask our data mining software to classify the employees into each group.

IV. CLUSTERING

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or

one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

4.1 Privacy-Preserving Data Mining on Data Grids in the Presence of Malicious Participants

Distributed data mining allows data to be shared without compromising privacy. On the one hand, data mining techniques have been shown to be a leading tool for data analysis, and as such they are likely to satisfy researchers' needs as an interface to the data stored in a grid. On the other hand, the models produced by data mining tools are statistical and thus satisfy the privacy concerns of the data owners. As a result, different HMOs can choose to reveal their databases not for direct reading but rather to a distributed data mining algorithm that will execute at the different sites and produce a statistical model of the combined database. That the algorithm produces statistics still does not guarantee privacy: an HMO also has to make certain that the data mining algorithm itself does not leak information. For instance, an algorithm in which each HMO computes its mortality rate and then sends it to a polling station which computes the global statistics would not meet this criterion because the polling station would be informed of the mortality rate for each HMO. This calls for a specific type of distributed data mining algorithm that is privacy-preserving. The idea there is to perturb the data by adding random transactions to the database. These perturbations hide the original data, but average out in the statistics.

V.CONCLUSION

Cryptographic tools can enable data mining that would otherwise be prevented due to security concerns. We have given procedures to mine distributed association rules on vertically partitioned data. We have shown that distributed association rule mining can be done efficiently under reasonable security assumptions. We believe the need for mining of data where access is restricted by privacy concerns will increase. Examples include knowledge discovery among bank services of different datasets. Another possibility is secure *approximate* data mining algorithms. Allowing error in the results may enable more efficient algorithms that maintain the desired level of security. In summary, it is possible to mine globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. Such privacy preserving data mining can be done with a reasonable increase in cost over methods that do not maintain privacy

REFERENCES

[1]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2]. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD Conference*, pages 439–450, 2000.

[3]. D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In *STOC*, pages 503–513, 1990.

[4]. M. Bellare, R. Canetti, and H. Krawczyk. Keying hash functions for message authentication. In *Crypto*, pages 1–15, 1996.

[5]. A. Ben-David, N. Nisan, and B. Pinkas. FairplayMP - A system for secure multi-party computation. In *CCS*, pages 257–266, 2008.

[6]. J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secret secret. In *Crypto*, pages 251–260, 1986.

[7]. J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In *ASIACRYPT*, pages 236–252, 2005.

[8]. D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In *PDIS*, pages 31– 42, 1996.

[9]. D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. *IEEE Trans. Knowl. Data Eng.*, 8(6):911–922, 1996.

[10]. T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Transactions on Information Theory*, 31:469– 472, 1985.