International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02| Feb -2017 www.irjet.net p-ISSN: 2395-0072

Privacy Preserving Data Mining Using Inverse Frequent ItemSet Mining Approach

Ms. Ashwini S. Chavan[1], Prof. Rahul P. Mirajkar[2]

[1]Student, Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, Kolhapur, Maharashtra, India
[2]Assistant Professor, Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, Kolhapur, Maharashtra, India

---

Abstract - *The paper presents architecture for protection of information against third party attack. Individual sensitive information is in danger with increasing technologies of data mining. A new research data mining topic, known as privacy-preserving data mining (PPDM), has been tremendously studied in recent years. Privacy preserving data mining (PPDM) aims to maintain privacy of individual data or sensitive information without sacrificing the utility of the data. Currently, privacy preserving data mining (PPDM) mainly consciousness on a way to reduce the privacy threat delivered by way of data mining operations, even as in truth, unwanted disclosure of sensitive information may manifest in the system of data collecting, data publishing, and information (i.e. The data mining effects) delivering. To this effect, paper proposes a technique called inverse frequent itemset mining approach that will help to protect sensitive information without loss of data.*

Key words: **Data mining, sensitive information, privacy preserving data mining (PPDM), inverse frequent itemset mining, protection.**

1. INTRODUCTION

Mining of data is the process of discovering interesting patterns and knowledge from large amounts of data [1]. The information collected by data mining can be very important to many applications, despite that there is another concern on focus of the privacy threats posed by data mining [2]. To address the privacy troubles in data mining, Privacy preserving data mining (PPDM) has received a top notch improvement in latest years[3][4]. The objective of PPDM is to safeguard sensitive information against the disclosure of data by maintaining its utility.

The PPDM Consideration is two-fold. First, sensitive raw data e.g. individual's ID card number, cell phone number should not be directly used for mining. Second, sensitive mining results whose disclosure will cause privacy violation should be excluded.

In PPDM process four different types of users are involved namely data provider, data collector, and data miner and decision maker. Each one has their specific role in the process. A data provider owns a few data from which precious information can be extracted. Data collector gets data from data providers so as to support the subsequent data mining operations. In order to invent useful knowledge which is expected by the decision maker, the data miner applies data mining algorithms to data obtained from data collector. A decision maker can get the data mining results directly from the data miner, or from some Information Transmitter. The focus of this topic is to achieve privacy at data miner level. *Data miner* will get data to mine from data collector to be able to in not unique layout and by means of applying one-of-a-kind data mining strategies; data miner can discover sensitive information. So venture of data miner is to hold the privacy of received result and pass the consequences to decision maker that doesn't bring about any security breach. Several studies on PPDM have been conducted [5] [6]. But none of the modern-day proposals provide privacy to unwanted disclosure of sensitive information. The paper presents a system architecture that provides privacy by the use data mining algorithms without affecting the security of sensitive information contained in the data.

2. LITERATURE SURVEY

Several theoretical approaches for privacy preserving of data have been proposed in the literature.

2.1 Related Research

B. Fung, K. Wang, R. Chen, and P. S. Yu et al [6] introduced techniques to protect the data. Data in its original form have sensitive information about person, and publishing such data will violate individual privacy. *Privacy-preserving data publishing* (PPDP) explains methods and tools for publishing useful information while preserving data privacy. The author introduces different schemes to PPDP, study the challenges in practical data publishing, and clarify the differences and requirements that distinguish PPDP from other related problems.

T. Mielikainen [7] introduces a well known technique called frequent set mining to describe binary data. However, it is an open problem how difficult it is to make opposite the frequent set mining. The author analyze the computational complexity of the problem of

finding a binary data set consistent with a known collection of frequent sets and show that in many cases the problem is computationally very difficult.

M. Malik, M. Ghazi, and R. Ali [8] introduced Privacy preserving data mining techniques to protect individual data or sensitive knowledge without giving up the utility of the data. People have become used to the privacy intrusions on their personal data and are very disinclined to share their sensitive information. This may cause to the accidental results of the data mining. So, the authors introduced present current scenario of privacy preserving data mining techniques and propose some future research directions.

S. Matwin [9] presents a brief summary and review of Privacy-preserving Data Mining (PPDM). The review of the existing approaches is built along an uncertain taxonomy of PPDM as a field. The main axes of this taxonomy specify what type of data is to be protected, and what is the possession of the data (centralized or distributed). Author discusses the relationship between PPDM and preventing unfair use of data mining techniques and round up by discussing some of the new, arising challenges before PPDM as a field.

K.Sathiyapriya and G. S. Sadasivam [10] introduced various techniques to protect the data. Data in its original form contains sensitive information. This paper introduces methods to preserving privacy of association rule mining and some hiding algorithms of association rules are evaluated. Some information is important with respect to opposing concept in organizations and individual misuses. Nowadays in order to discover the useful patterns in a great amount of data, data mining tools are used. In order to protect information, fast processing and preventing from disclosing private data to keep privacy is presented in data mining.

2.2 Existing System

The information spotted by data mining can be very vital to many applications. People have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining. Individual's privacy may be disturbed due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc. Current models and algorithms proposed for PPDM mainly focus on how to hide this sensitive information from certain mining ways. Besides the mining phase, privacy issues may also arise in the phase of data collection or data pre-processing, even in the delivery process of the mining results.

2.3 Limitations of the Existing System

Most current studies only manage to achieve privacy preserving in a statistical sense. Current studies are still in the theoretical stage. By studying different papers some observations have been drawn. Developing practical personalized anonymization methods is in urgent need. Besides, introducing personalized privacy into other types of PPDP/PPDM algorithms is also required. Therefore, it is necessary to develop new system to specify what kind of information is required and how to present, store, acquire and utilize the information.

3. PROPOSED SYSTEM

As of now, maintaining privacy of the data is a difficult task. A number of methods and techniques have been proposed for privacy preserving data mining. The theoretical studies studied earlier defined various techniques for protection of sensitive information. But there yet exist a practical implementation of a system that helps for security of sensitive information. The proposed work has following objectives,

- To define dataset and pre-processing of it.
- To transform a dataset from original state to new state to achieve hiding of data.
- To implement Inverse Frequent item set mining to hide data mining results.

The data set provided generally for mining purpose is in plain data or in relational data format. This imposes security threat on the data. To achieve Privacy in Data Mining, the data provided for mining purpose should not be delivered as it is to the data miner. The Privacy Preserving Data Publication (PPDP) is one of the techniques that provide some data transformations and avoid exposure of sensitive data. This transformed data is then provided to data miner. The mining of data produces some useful results that can be used for knowledge extraction. But it is important to hide these results for privacy reason.

The mining phase here uses apriori algorithm and finds the frequent items from transformed database. To achieve privacy for mining result some data customization is needed and that can be provided by Inverse Frequent Itemset mining process. This will avoid exposure of mining results from data miner and decision maker.

The architecture consists of four modules:

1. Data preprocessing.
2. Data transformation.
3. Data mining.
4. Data evaluation.

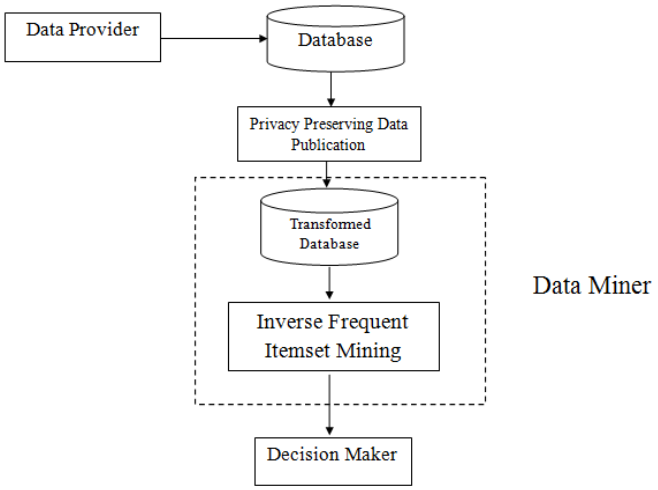

Fig -1: System Architecture

4. PROPOSED WORK

4.1 Data preprocessing

Data preprocessing is a vital important step in data mining process which deals with preparation and transformation of initial dataset. It includes basic operations as data selection, data cleaning, and data integration. Data preprocessing describes any type of processing performed on raw data and to prepare it for another processing. It modifies the data into a format that will be more easily and adequately be processed for the purpose of the user. A standard DB-Reduction algorithm for data preprocessing is implemented. This algorithm highly reduces the number of attributes and records of the data set and improves the accuracy and decreases the running time of the data mining algorithms in the later stage.

Algorithm makes use of two methods namely, select method and update method. In select method, the user can choose records from the data set. For example, suppose e a user needs particular students records residing in zipcode 416008, then by use of this select method the user can acquire those needed records, instead of moving through whole data set. This method saves the time of the user and fetches the records as needed. Next, in update method the records selected by the user in select method can be updated. For example, suppose a user needs to update a particular student record who is residing in zip code 416008, he can do this by simply making a change of zip code from 416008 to 416009. Thus, any of the record selected by the user can be updated.

4.2 Data transformation

Data transformation converts a set of data values from the data format of a one system into the data format of a data system. The goal is to transform data into forms appropriate for the data mining operations, that is, to find important features to represent the data. The data collected from data provider may contain sensitive information. To hide this information some transformation is needed. This model use privacy preserving data publishing (PPDP) technique which guarantees data utility even after data amendment. Feature selection and feature transformation are basic operations. A standard RSA (Rivest-Shamir-Adleman) algorithm is implemented for data transformation technique. For example if there are tables of student information in database, like student data and course, then the encryption can be done on single table or multiple tables. This algorithm is used for encryption and decryption of data. The data from pre-processing is encrypted so as to protect it from third party. Hence, maintaining the security of sensitive information.

4.3 Data mining

The design employs inverse data mining such as Inverse Frequent Set Mining (IFM) to produce data that cannot expose sensitive information. Reverse Data Management (RDM) which is similar to Inverse Frequent Set Mining performs operations like compute database input or modify an existing database input, in order to achieve a desired effect in the output. Aprori algorithm is implemented so as to find out the count of frequent items from the data set. For example if you have a set of 20 queries, then aprori algorithm is used to find out the count of frequent fields (items) being fired in a set of queries. By doing so, list of frequent items is separated from general data set and the remaining data is encrypted. This is termed as Inverse frequent itemset mining approach (IFM) approach. This makes data more secure from an intruder attack.

4.4 Data Evaluation

This is the last module. Data evaluation include basic operations as, identifying the truly interesting patterns which represent knowledge, and presenting mined knowledge in easy to understand fashion. The results produced in data mining module are evaluated using data evaluation module. The data obtained from data mining module is processed in this module. Not all the patterns found by data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. The final result is represented using data evaluation module.

5. SNAPSHOTS

Encryption can be single table encryption or multiple table encryptions.

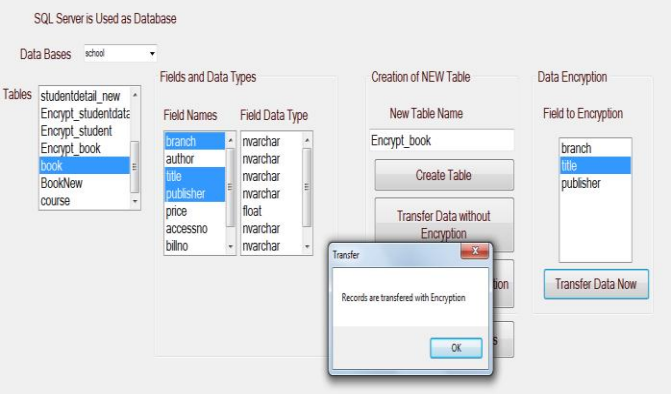

Fig -2: Single table encryption

Fig 2: shows single table encryption method. SQL server database is used as a database at backend. Database school is designed under which there are tables as book, course, student, student fees as displayed in above snapshot. User must select any one table (book) of his/her choice for further processing. After this, fieldnames is to be selected to create a new table so as to perform encryption on that table. Here fieldnames branch, title and publisher is been chosen. User must select at least two fields to create a table. She/he can also select all the fields to create a new table. After table creation (Encrypt_book) the user has to select the fields (title) from new table (Encrypt_book) on which encryption can be done. Finally we will get data in encrypted format. User can also transfer data as it is i.e. without encryption to the newly created table.

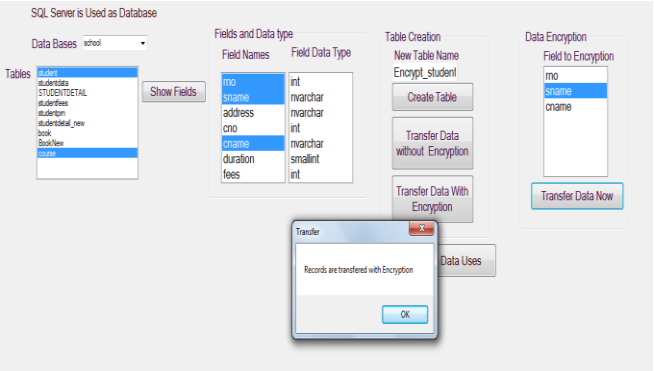

Fig -3: Multiple table encryptions

Fig 3: shows multiple table encryptions. As for single table, same process is carried out for multiple table encryptions. In this user can select more than one table and use it for encryption as shown in above snapshot. Table's course and student are selected. Fieldnames from both tables are displayed, from which user selects sname, rno and cname. Table Encrypt_student_course is created consisting of fields selected by user. At last, field that is to be encrypted is chosen and finally get message of success.

6. RESULT ANALYSIS

In this section, we calculate the time required for encryption. The encryption is done on single table and multiple tables.

6.1 Single table encryption

The encryption is performed on single table of the database and the time required for it is calculated.

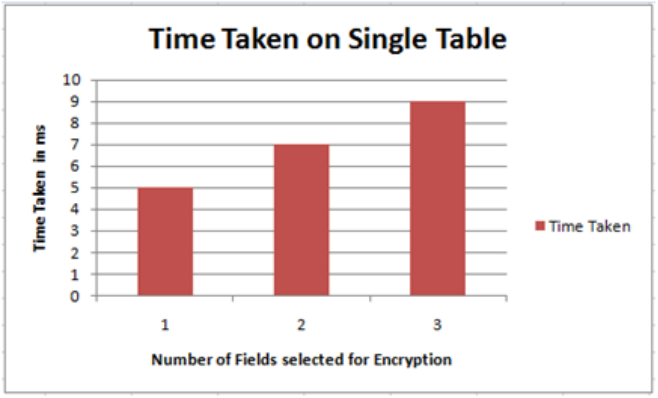

Chart -1: Single table

The time taken is in ms as shown in chart 1. The time taken for encryption of 2 fields is 5ms and that for 3 fields is 7ms and so on.

6.2 Multiple tables encryption

The encryption is performed on multiple tables of the database and the time required for it is calculated.

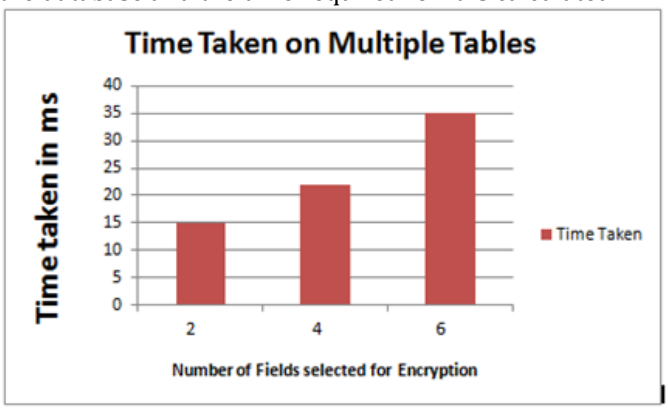

Chart -2: Multiple table

The time taken for multiple fields' encryption is more as compared to that of single table encryption.

7. CONCLUSIONS

We have made survey of different literatures and studied different techniques for maintaining privacy of the data. This paper proposes development of a system that will help to protect the sensitive information. Various algorithms are being used by the modules to protect a system against a theft. The data is protected in four steps as discussed in section IV. Inverse Frequent Itemset mining approach is used to achieve privacy. Due to this, security of data is maintained without losing the original essential information.

8. FUTURE WORK

As future work, we can implement this same system by using other data mining technologies other than apriori algorithm like FP-Growth algorithm for finding frequent items in large database. Also we can use an improved aprori algorithm. For further research we can go through additional reference paper which can helpful to look on advanced research like. Multi-Sorted Inverse Frequent Itemsets Mining: On-Going Research. It is needed that design the database more detail as per needed by emerging big data applications e.g., social network analytics. We can increase the accuracy of algorithm in spite of large database.

ACKNOWLEDGEMENT

There have been many contributors for this to take shape and authors are thankful to each of them.

REFERENCES

[1] Lei xu, Chunxio Jiang, Jian Wang, Jian Yuan, Yong Ren. Information Security in Big Data: Privacy and Data Mining, DOI 10.1109/ACCESS.2014.2362522,IEEE Access.

[2] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[3] L. Brankovic and V. Estivill-Castro, "Privacy issues in knowledge discovery and data mining," in *Australian institute of computer ethicsconference*, 1999, pp. 89–99.

[4] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *SIGMOD Rec.*, vol. 29, no. 2, pp. 439–450, 2000.

[5] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances Cryptology CRYPTO 2000*. Springer, 2000, pp. 36–54.

[6] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, p. 14, 2010

[7] T. Mielikainen, "On inverse frequent set mining," in *Workshop on Privacy Preserving Data Mining*, 2003, pp. 18–23.

[8] M. Malik, M. Ghazi, and R. Ali, "Privacy preserving data mining techniques: Current scenario and future prospects," in *Computer and Communication Technology (ICCCT), 2012 Third International Conference on*, 2012, pp. 26–32.

[9] S.Matwin,"Privacy-preserving data mining techniques: Survey and challenges," in *Discrimination and Privacy in the Information Society*. Springer, 2013, pp. 209–221...

[10] K. Sathiyapriya and G. S. Sadasivam, "A survey on privacy preserving association rule mining." *International Journal of Data Mining &Knowledge Management Process*, vol. 3, no. 2, 2013.

[11] Libing Wu, kui Gong, Yanxiang He, Xiaohua Ge Jainqun Cui, 'A Study of Improving Apriori Algorithm' 2010.

[12] Xiaohua Hu, 'DB-Reduction: A data preprocessing algorithm for data mining applications',2003

[13] A. Meliou, W. Gatterbauer, and D. Suciu, "Reverse data management," *Proceedings of the VLDB Endowment*, vol. 4, no. 12, 2011.

[14] Domenico Saccà, Edoardo Serra, and Antonio Piccolo."Multi-Sorted Inverse Frequent Itemsets Mining: On-Going Research" *AMW 2016 - Proceedings of the 10th Alberto Mendelzon International Workshop on Foundations of Data Management* (2016)