ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

A Review: Data Mining Concepts and Privacy Preservation

Shailja Agnihotri

Assistant Professor, Dept. of I.T., GGDSD College, Sector 32, Chandigarh, India

ABSTRACT: The primary purpose of data mining is the extraction and the discovery of useful information from the dataset. The data is stored in the data warehouse and extracted using data mining techniques. The extracted information is useful in every domain namely banking, medicine, for finding criminal records or some driving information etc. The datasets or the databases consists of huge information and also provides large volume of inherent valuable information for the prediction of various factors may be performance, growth etc. The data is being stored and extracted efficiently but it is not the sole concern. The privacy and security of data is the key issue which can be further investigated and researched upon. PPDM i.e. Privacy Preserving Data Mining includes the techniques to save the data or the sensitive information from various risks. It leads to another term PPDP i.e. Privacy Preserving Data Publishing which includes the statistical processing of the sanitized data and protection of the privacy of individual records. This paper focuses on the privacy of information in data mining. It also lists the various cryptographic and non-cryptographic techniques to protect the privacy of the extracted data.

KEYWORDS: data mining, PPDP, PPDM, clustering, KDD, cryptography

I. Introduction

Data Mining is the data driven technique to study and collect relevant and useful information from the large volume of data available with any organization. It basically uses the calculations may be statistical or arithmetic to predict the trends for the data stored in the datasets. The paper focuses on the basic concepts of data mining as well as the privacy concern related with the discovery of knowledge from the databases. The various cryptographic and non-cryptographic approaches can be used. Basically, data mining is the combination of three components-

- Classification- it includes defining different groups to be considered for the application of data mining techniques.
- Clustering- It includes identifying the groups sharing the common characteristics.
- Association Rules- includes the identification of the relationships between the onetime events like shopping bag contents of some customer at the shop.
- Sequence Analysis- includes the relationships but specified over a time period.
- Forecasting- It includes predicting patterns and future values in the datasets.

The data in the databases exist in many forms. Multimedia databases also exist. The data related with audio, video or images is termed as non regulated non textual data and the text data is termed as non regulated textual data. This paper discusses the privacy issues and provides the solution for the non regulated textual data. The business application area faces the valuable competition with their competitors. The data mining components may be unified together for better mining results. Security issues may include the intrusion detection etc or the potential hazards when adversary also has the data mining capabilities.

DOI: 10.15680/IJIRCCE.2016. 0408025

II. DIFFERENCE BETWEEN KDD AND DATA MINING

There exist an overlap between the KDD i.e. Knowledge Discovery in Databases and Data Mining. Actually data mining is part of KDD. KDD includes the tools and techniques for extracting useful and meaningful information from the databases which exist in the datasets. Data mining involves the implementation of the specific algorithm exists in KDD process.

There exists various models for data mining such as CRISP-DM, Six Sigma, SEMMA, Scientific Model, Hybrid Data model etc. Data mining is the convergence of the various domains such as Artificial Intelligence, Machine Learning, Visualization, Databases, Statistics etc.

III. DATA MINING KEY ISSUES

- User Interaction issue
- Mining of different knowledge types issue
- Maintaining multiple level abstraction issue
- Guiding discovery process through knowledge issue
- Query Language support for Data Mining issue
- Presenting the discovered meaningful results issue
- Noisy data and data cleaning process issue
- Algorithms performance issues
- Data mining in multimedia databases issue
- Mining knowledge from distributed and heterogeneous databases issue
- Privacy and security of the extracted information issue

The important concern of this paper is the privacy issue. Data mining is playing a significant role in vast number of applications. It has been proved helpful in finding information related to terror attacks even and also confronting various computer attacks. But it also poses some privacy threats such as privacy issue in distributed environment. In distributed environment, two or more parties agree to generate a common output through the application of data mining techniques but they reserve the privacy of their shared input datasets. The values of their respective datasets are made confidential. For implementing data mining in such a scenario, they may agree upon some trusted party which will convert their input values and provide the secure and privacy maintained results. But here the trusted party may be the semi-honest one or the malicious party. Semi-honest trusted party tries to find out the additional results but the privacy is maintained and in malicious party case, the additional information is intentionally fetched, manipulated and privacy may be leaked. However, there are many issues which need the attention for maintaining privacy and security of data mining results.

IV. APPLICATIONS

- Data collection and surveys: the companies or the political parties' collects data or conduct surveys for predicting many new patterns which would prove meaningful to them in future. If the respondents are assured of their data privacy they would be more willing to participate. For such types of data, the randomization, a privacy preserving technique can be applied.

- Monitoring for emergencies: Public security or National security can also be looked through data mining. Terrorist attacks, disease outbreaks or other disasters can be detected in advance. But for the effective detection, the system may have to collect commercial, personal or sensor data. The collection of such type of data highlights the privacy issue.

- Product traceability: with the advancement in technology, the product will be manufactured with the RFID tags. These distributed RFID traces can be mined and will help in detection of any inefficiencies or any criminal activity regarding products etc. It will have to go through high level of detailed process and may require the use of highly valuable and sensitive data.

- Medical research: Health records are considered as the highly sensitive type of data. Sharing or availability of patient records history or information may pose various privacy issues. Here, PPDM can be used over vertically partitioned data for maintaining the privacy.

- Social networks: Here most of the data is of low quality and also the data is personal and confidential. In such type of application area, the PPDM techniques are useful for maintaining the private data privacy.

V. Various Cryptographic and Non-cryptographic Techniques

The two types of techniques can be used for maintaining the privacy i.e.

- Cryptography techniques
- Non-cryptography techniques

Cryptography techniques result in maintaining the privacy and also loss of some information. These techniques are useful in the areas where there is collaboration of multiple parties for sharing their respective non sensitive input data. The data may be distributed horizontally or vertically among the parties and there are various cryptographic algorithms which support the maintenance of privacy. Some of the cryptography methods are:

- Secure multiparty communication: these involve the use of encryption protocol. These work for horizontally or vertically partitioned datasets. These are safe, secure and trust-worthy but hard to implement as complexity grows.

- Asymmetric Ciphers: This technique uses two keys, a public key and a private key. The RSA algorithm can be implemented under this scheme. In public key transmission there is high security as the key is not be revealed. There also exist some other fast encryption methods under public key encryption

Non- Cryptographic Techniques:
These include various methods which do not include the cryptographic measures. Some of them are discussed as under:

- Randomization: It is one of the techniques which is based on data distortion i.e. it works on adding some noise to the existing data. This technique maintains the balance between the privacy and the knowledge discovery. This method is easy to implement and simple.

- Anonymization: This includes the implementation of generalization and suppression methods for the generation of individual records.

- Summarization: This method develops the summary of the data under PPDM and also hides the individual records and results are through the aggregate queries. It may be the extension of Randomization technique. This technique is somewhat new as compared to other techniques.

- Suppression: This technique states that the privacy can be maintained by suppressing the sensitive data before the implementation of the technique. Specific attributes of the database can be suppressed to extract the results. For instance, identity of any person may be suppressed for the processing of some records.

VI. Literature Review

[1] Justin Zhan discusses that data collection is the major step in the data mining process. But collection of data from different parties is difficult. This paper lays stress on privacy preservation and multi party collaborative data mining process. The paper illustrates the privacy preserving process using Bayesian classification. Using this, it provided the solution for preserving the privacy of the data.

[2] Fayyad et al. tell the concepts of data mining and the KDD process. They provided an overview of both the concepts and also the relation between the both. The paper tells about the various data mining applications, data mining techniques, challenges and the future directions for the research in the particular field. It further states that there is need of formulation of tools and techniques for extracting useful data from the continuously growing digital data. The major problem is the conversion of low level data into other forms that may be more compact, abstract and useful.

[3] Dileep Kumar, and Vishnu Swaroop defined the data mining as process of mining for inherent, initially unknown and useful information for huge databases by applying various knowledge discovery techniques. They considered privacy and security as the two main anxieties among the parties. This paper focuses on these key issues and the use of laws and regulations.

[4] Murat Kantarcioglu and Chris Clifton tell that data mining is the process which can extract useful, meaningful data from large data sets. Problems are faced when there is involvement of multiple parties, who want to share and extract data. The paper focuses on secure mining association rules over horizontally partitioned data. The methods mentioned use cryptographic techniques to maintain the privacy. They believe that the need for data mining techniques application, where access is restricted by privacy concerns will increase.

[5] Amit Kumar Jha and Divakar Singh conducted a survey for the cloud computing establishment and security issues in cloud computing. Cloud computing is considered as the emerging area for the research perspective. They listed the various current solutions for maintaining privacy in the cloud computing. The survey discusses the cloud computing benefits, characteristics and issues regarding security. The key for maintaining privacy is to separate the sensitive data from the non-sensitive data followed by the encryption of the sensitive data.

[6] Ashish Chouhan and Dr.Anju Singh tell that there are number of algorithms exist for maintaining privacy but the real problem lies with the multi party involvement in the data mining process. There is a need to shield the priviledged data. The paper presents the review of the progressive strategies for controlling privacy. The techniques mentioned are K-anonymity and distributed privacy preserving technique. The high dimensional knowledge sets are discussed.

[7] Alexandre Evfimievski and Tyrone Grandison lay focus on PPDM i.e. Privacy Preserving Data Mining techniques. It has been referred as the area of data mining that safeguards the relevant and sensitive information from the unwanted disclosure. Privacy preservation should also be applied on the individual records also as applied on datasets in aggregation. The paper lists and explains various PPDM approaches. Advantages and disadvantages of various approaches are also defined and also the application areas are explained where privacy of information is the key issue.

[8] Divya Sharma explained the data mining as the process of discovery of new patterns from huge datasets. The goal of data mining is to extract the information which is useful and meaningful for the parties involved in the process. The paper provides a survey of various privacy preserving data mining algorithms. One of the different techniques i.e. Randomization is explained in more detail.

[9] Swagatika Devi tells that privacy preserving data mining is the very popular as the huge datasets are growing in number and also the parties involved in the data mining process. The paper discusses the various techniques like k-anonymization and also distributed privacy preserving data mining technique. It also explains the available state-of-the-art methods for maintaining the privacy with their respective merits and demerits. Present problems are also highlighted.

[10] S.R.M. Oliveira and O.R. Zaiane address the problem of transactional databases for protecting the sensitive knowledge. They have introduced a new algorithm i.e. one scan algorithm which maintains the accuracy and privacy in association rule mining.

[11] M. Kantarcıoglu and C. Clifton explain that the data mining is useful in extracting the vital information from the huge datasets but these datasets may be distributed among multiple parties. The well known issue of privacy may restrict the parties from sharing of their data, this paper focuses on secure data mining of data over the horizontally partitioned data. The cryptographic techniques are also discussed.

[12] R. Agrawal and R. Srikant primarily focus on the issue of data mining. They suggested that the accurate models can be developed to access the precise information in case of data mining. They have listed the variations between the actual and the original records. By using some reconstructed distributions, the accuracy with respect to the original data can be achieved.

VII. CONCLUSION

The various data mining techniques are being used to extract information rather useful information from the large datasets. Since so many years, data mining has been proved as the research area but this paper focused mainly on privacy issue. PPDM ensures the efficient data analysis and protection of privacy of the information of the users. This paper also discussed the two party and multiparty cases. The cryptographic and non-cryptographic techniques for protection of data also discussed. Various approaches are available still no algorithm or technique has been proved efficient for the privacy or security of data. More work or research needs to be done. May be the PPDM can be used more efficiently. For the knowledge discovery the more unified approach can be followed which would provide the single step for the data mining. The components of data mining clustering or classification, association rules and visualization may be combined. The combination would prove the better unified approach for implementing single step data mining process. Also the oblivious transfer protocol may be improved upon to improve the performance. Although this paper focuses on textual databases but new improved data mining techniques are also required for the multimedia databases which include the audio, video or image data. The more scalable algorithms need to be developed to achieve more efficiency and accuracy with the surety of privacy. Existing PPDM approaches are not so flexible yet. There is a requirement to explore the limits of PPDM.

REFERENCES

[1] Justin Zhan from Carnegie Mellon University, USA, —Privacy- Preserving Collaborative Data Mining, 2008.

[2] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, American Association for Artificial Intelligence, 1996.

[3] Dileep Kumar, and Vishnu Swaroop. " Data Security and Privacy in Data Mining:Research Issues & Preparation."

[4] Murat Kantarcioglu, Chris Clifton, "Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, 2004

[5] Amit Kumar Jha and Divakar Singh ,"A Survey of Cloud Computing Service and Privacy Issues," Advances in Computer Science and Information Technology (ACSIT), Volume 1, Number 2; November, 2014

[6] Ashish Chouhan and Dr.Anju Singh ,"Privacy Preserving Data Mining: A Survey on Anonymity," International Journal of Electrical, Electronics and Computer Engineering 2015

[7] Alexandre Evfimievski and Tyrone Grandison, —Privacy Preserving Data Mining‖ at IBM Almaden Research Center, 2007

[8] Divya Sharma, —A Survey on Maintaining Privacy in Data Mining, IJERT April 2012.

[9] Swagatika Devi, - A Survey On Privacy Preserving Data Mining: Approaches and Techniques, IJEST March 2011

[10] S.R.M. Oliveira and O.R. Zaiane, “Protecting Sensitive Knowledge by Data Sanitization,” Proc. 3rd IEEE Int’l Conf. Data Mining (ICDM 03), IEEE CS Press, 2003, pp.613–616.

[11] M. Kantarcıoglu and C. Clifton, “Privacy-Preserving Distributed Mining of Association Rules on Horizon-tally partitioned Data,” IEEE Trans. Knowledge Data Eng., vol. 16, no. 9, 2004, pp. 1026–103

[12] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” Proc. ACM SIGMOD Conf. Management of Data, ACM Press, 2000, pp. 439–450; http://doi.acm.org/10.1145/342009.335438.

BIOGRAPHY

Shailja Agnihotri is an Assistant Professor in the Information Technology Department, GGDSD College, Sector 32, Chandigarh, India. She received Master of Science in Information Technnology (MSc(IT)) degree in 2005 from Panjab University, Chandigarh, India.