ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

A review paper on Privacy-Preserving Data Mining

Mohnish Patel[1], Prashant Richariya[2], Anurag Shrivastava[3]

[1]Mohnish Patel, NIRT, RGPV, Bhopal
[2]Prashant Richariya, NIRT, RGPV, Bhopal
[3]Anurag Shrivastava, NIRT, RGPV, Bhopal

Abstract: Data mining technology help us in extraction of useful knowledge from large data sets. The process of data collection and data dissemination may, however, result in an inherent risk of privacy threats. Some private information about individuals, businesses and organizations has to be suppressed before it is shared or published. The privacy-preserving data mining (PPDM) has thus become an important issue in current years. This paper we propose an evolutionary privacy-preserving data mining technology to find appropriate method to perform secure transactions into a database.

Keywords: data driven; artificial intelligence technology; data warehousing; threats.

I. INTRODUCTION

Data mining is a study or collection that can extract knowledgeable information from large quantities of data, and is data driven. It uses arithmetic and statistical calculations to uncover trends and correlations among the large quantities of data information stored in a database. It is a blend of artificial intelligence technology, statistics, data warehousing, and machine learning.

Data mining technology help us in extraction of useful knowledge from large data sets. The process of data collection and data dissemination may, however, result in an inherent risk of privacy threats. Some private information about individuals, businesses and organizations has to be suppressed before it is shared or published. The privacy-preserving data mining (PPDM) [1] has thus become an important issue in current years. This paper propose an evolutionary privacy-preserving data mining technology which uses data mining technique and network security cryptographic method to secure or preserver the data to find appropriate transactions to be hidden from a database [2].

Now a days, companies in business are working fast to derive a valuable competitive advantage over other businesses. A popular and fast growing technology, which can help to gain this advantage, is known as data mining. Data mining process allows a company to use the huge amount of data that it has compiled, and develop correlations and relationships among this data to help business improve efficiency in order to learn more about its customers, make better decisions and help in implementation planning.

Data mining consist of three major components Clustering or Classification, Association Rules and Sequence Analysis. The Data Mining technology can develop these analyses on its own, using commix of statistics, artificial intelligence, machine learning algorithms, and data stores. For a survey work in statistical databases refer to Adam & Wortmann (1989) and Willenborg & de Waal (2001) [3].

In our day to day life we came across unorganized data in various forms. They include books/ libraries journals, video/ audio files and unregulated text as mails, web pages and other documents. And these data can be a critical source in order to make communicative decisions. For example, for a company there is a group of people who can be determined as the paramount from among its workforce. Determining what is common among them and determining others like them would doubtlessly improve the output of the business as well company. This is the basic, on the basis of which this research was done.

The important aspect of the research was to use text mining techniques to mine the data in a group of documents and to determine what are the common characteristics among them and then to determine other documents which contains these selected characteristics. Unregulated data can be documented as any data that is not found in a database. MP3, Images, video files could be categorized as non-textual unregulated data whereas memos, email messages, word processor documents could be categorized as textual unregulated data. During this research, we concentrate on unregulated data in textual form.

II. RELATED WORK

Shweta Shrma Hitesh Gupta and Priyank Jain's proposal [17] mainly deals with information retrieval system. The area where users might be able to look for documents, information in between document or metadata from documents on the web is known as the Information retrieval.

Divya Sharma [18] provides a wide survey of different privacy preserving data mining algorithms. The author have discussed about merits and demerits of algorithm Randomization. Author purposed methods are only approximate to our goal of privacy preservation; we need to further perfect those approaches or develop some efficient methods.

SWAGATIKA DEVI[19] , provide a review of the state-of-the-art methods for privacy and analyze the representative technique for privacy preserving data mining and points out their merits and demerits. Knowledge is supremacy and the more knowledgeable we are about information break-in, we are less prone to fall prey to the evil hacker sharks of information technology.

Tipawan [20] and team discussed on the findings which could be divided into 4 topics:

(i) knowledge resource; (ii) knowledge types and/or knowledge datasets; (iii) data mining tasks; and (iv) data mining techniques and applications used in knowledge management. The describes the definition of data mining with its functionality. Then it explains knowledge management rationale and various management tools integrated in knowledge management. At last, the applications of data mining techniques are summarized and discussed.

We have presented two secure protocols for privately checking whether a k-anonymous database retains its anonymity, once a new tuple is being inserted to it. Since the proposed protocols ensure that the updated database remains K-anonymous, the results returned from a user's (or a medical researcher's) query are also k-anonymous [16]. Thus, the patient or the data provider's privacy cannot be violated from any query. As long as the database is updated properly using the proposed protocols, the user queries under our application domain are always privacy preserving.

We show that privacy and collaborative data mining can be achieved at the same time. The goal of this paper is to present technologies to solve privacy-preserving collaborative data mining [4] problems over large data sets with reasonable efficiency. The contributions of this paper contains the following: (1) a proposed definition of privacy for privacy-preserving collaborative data mining; (2) a solution for naive Bayesian classification with vertical collaboration; and (3) an efficiency analysis to show the performance scaling up with various factors.

Previous work attempted to find an optimal k-anonymization that minimizes some data distortion metric. We argue that minimizing the distortion to the training data is not relevant to the classification goal that requires extracting the structure of predication on the "future" data [5]. In this paper, we propose a k-anonymization solution for classification. Our goal is to find a k-anonymization, not necessarily optimal in the sense of minimizing data distortion, which preserves the classification structure. We conducted intensive experiments to evaluate the impact of anonymization on the classification on future data. Experiments on real-life data show that the quality of classification can be preserved even for highly restrictive anonymity requirements.

The key problem of applying geometric data [6] perturbation in multiparty collaborative mining is to securely unify multiple geometric perturbations that are preferred by different parties, respectively. We have developed three protocols for perturbation unification.

Our approach has three unique features compared to the existing approaches: 1) with geometric data perturbation, these protocols can work for many existing popular data mining algorithms, while most of other approaches are only designed for a particular mining algorithm; 2) both the two major factors: data utility and privacy guarantee are well preserved, compared to other perturbation based approaches; and 3) two of the three proposed protocols also have great scalability in terms of the number of participants, while many existing cryptographic approaches consider only two or a few more participants.

The term "privacy preserving data mining" [7] was introduced in papers Agrawal & Srikant (2000) and Lindell & Pinkas (2000). These papers considered two fundamental problems of PPDM, privacy preserving data collection and mining a dataset partitioned across several private enterprises. Agrawal and Srikant (2000) devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values; Lindell and Pinkas (2000) [8] invented a cryptographic protocol for decision tree construction over a dataset horizontally partitioned between two parties. These methods were subsequently refined and extended by many researchers worldwide. Other areas that influence the development of PPDM include cryptography and secure multiparty computation (Goldreich, 2004) [9]; (Stinson, 2006), database query auditing for disclosure detection and prevention (Kleinberg et al. 2000); (Dinur & Nissim, 2003); (Kenthapadi et al. 2005), database privacy and policy enforcement (Agrawal et al. 2002); (Aggarwal et al. 2004), database security (Castano et al. 1995) [10], and of course, specific application domains.

Encryption is a well-known technique for preserving the confidentiality of sensitive information. In comparison with the other techniques described, a strong encryption scheme can be more effective in protecting the data privacy. An encryption system normally requires that the encrypted data should be decrypted before making any operations on it. For example, if the value is hidden by a randomization-based technique, the original value will be disclosed with certain probability. If the value is encrypted using a semantic secure encryption scheme [11], the encrypted value provides no help for an attacker to find the original value. One such scheme is the homomorphic encryption scheme which was originally proposed in [12] with the aim of allowing certain computations performed on encrypted data without preliminary decryption operations. To date, there are many such systems. Homomorphic encryption is a very powerful cryptographic MAY 2008 | IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE 33 tool and has been applied in several research areas such as electronic voting, on-line auctions, etc. [14] is based on homomorphic encryption where Wright and Yang applied homomorphic encryption to the Bayesian networks induction for the case of two parties. Zhan et al., [13] proposed a cryptographic approach to tackle collaborative association rule mining among multiple parties. In this paper, we will apply homomorphic encryption [11] and digital envelope techniques to privacy-preserving data mining and use them to design privacy-oriented protocols for privacy-preserving naïve Bayesian classification problems. The preliminary idea of this paper has been published in [15].

III. PROPOSED PRIVACY PRESERVING METHODS AND TECHNIQUES

A. Randomization Method

In current privacy preserving data mining technology, the randomization method is most preferable. These methods also provide knowledge discovery and balance between privacy preservation. Here for using this method some noise is added to the data to mask the fields of records [7] which is sufficiently large so that the individual values of the records can no longer be recovered. For the implementation of randomization methods, need to implement two steps. Those are as follows: (1) data providers randomize their data and transmit randomized data to data receiver; (2) data receiver estimates original distribution of data using distribution reconstruction algorithm.

B. Encryption Method

The cryptography-based technique guarantees high level of data privacy. Encryption method resolves the [18] problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between two competitors or even between untrusted parties. That’s why privacy preserving technique needs to implement to secure the data. There are various PPDM techniques such as the method on vertically partitioned data and horizontally partitioned data. Encryption method ensures the transfer of data is secure and exact, but this method is not much efficient.

IV. RESULT AND DISCUSSION FOR FURTHER ENHANCEMENT

The goal of this paper is to present technologies to Find Appropriate Transactions to be hidden from a database with reasonable efficiency. The benefits of unregulated data is vast in textual form, for example a bank may decide whether to grant a person a loan or not based on collection of documents, i.e. the bank would need to compare the current applicant’s application with previous application and decide as to which category the applicant would belong to. Another example would be when recruiting the right person for the correct position which is all about finding the best possible match between an individual and the job. You may start the recruitment process by creating a job description. The danger in this is that if you are not aware of the functions carried out by a person performing the same job you may end up in creating a wish list that might have the possibility of turning away the right candidates and attracting the wrong ones. A better approach would be to look at the best performers in the company and use them as role models. By representing an employee by his curriculum vitae (CV) would allow us to find characteristics which are common among the star performers of the company, for an example this could be that all performers have studied in a specific university or followed a specific stream of study or even competent in specific technological area and so on. Then by looking for these common factors among the rest of the CVs we will be able to identify a handful of the correct candidates. Thus, we would need to analyze a large volume of unregulated data that falls under the category of Text Mining.

Web mining is also the important part of data mining, IN web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or regulated records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

With the rapid development of computer and information technology in the last several decades, an enormous amount of data in science and engineering has been and will continuously be generated in massive scale, either being stored in gigantic storage devices or flowing into and out of the system in the form of data streams. Moreover, such data has been made widely available, e.g., via the Internet. Such tremendous amount of data, in the order of tera- to peta-bytes, has fundamentally changed science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for new, data-intensive methods to conduct research in science and engineering. There are various research challenges in science and engineering, from the data mining perspective, with a focus on the following issues:

(1) Information network analysis; (2) Discovery, usage, and understanding of patterns and knowledge; (3) Stream data mining; (4) Mining moving object data, RFID data, and data from sensor networks; (5) Spatiotemporal and multimedia data mining; (6) Mining text, Web, and other unregulated data; (7) Data cube-oriented multidimensional online analytical mining; (8) Visual data mining; (9) Data mining by integration of sophisticated scientific and engineering domain knowledge.

V. CONCLUSION

Privacy-preserving data mining emerged in response to two equally important (and seemingly disparate) needs data analysis in order to deliver better services and ensuring the privacy rights of the data owners. Difficult as the task of addressing these needs may seem, several tangible efforts have been accomplished. In this paper, an overview of the popular approaches for doing PPDM was presented, namely: suppression, randomization, cryptography and summarization. The privacy guarantees, advantages and disadvantages of each approach were stated in order to provide a balanced view of the state of the art. Finally, the scenarios where PPDM may be used and some directions for future work were outlined.

VI. REFERENCES

[1] Alexandre Evfimievski and Tyrone Grandison, “Privacy Preserving Data Mining” at IBM Almaden Research Center, 2007.

[2] Tzung-Pei Hong, Dept. of Comput. Sci. & Inf. Eng., Nat. Univ. of Kaohsiung, Kaohsiung, Taiwan, ”Evolutionary privacy-preserving data mining”, 19-23 Sept. 2010.

[3] Adam, N. R. & Wortmann, J. C., “Security-Control Methods for Statistical Databases: A Comparative Study”, ACM Computing Surveys, Vol. 21, N. 4, pp. 515–556, 1989.

[4] Justin Zhan from Carnegie Mellon University, USA, “Privacy-Preserving Collaborative Data Mining”, 2008.

[5] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu, Fellow, IEEE, “Anonymizing Classification Data for Privacy Preservation”, may 2007.

[6] Keke Chen, Member, IEEE, and Ling Liu, Senior Member, IEEE, “Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation”, Dec-2009.

[7] Agrawal, R. & Srikant, R., “Privacy Preserving Data Mining. In Proc. of ACM SIGMOD”, Conference on Management of Data (SIGMOD’00), Dallas, TX, 2000.

[8] Lindell, Y. & Pinkas, B, “Privacy Preserving Data Mining. In Proc. of Advances in Cryptology – Crypto’00, LNCS 1880”, Springer-Verlag, pp. 20–24, 2000.

[9] Goldreich, O, “Foundations of Cryptography”, Volume I, II. Cambridge University Press., 2004.

[10] Castano, S., Fugini, M., Martella, G., & Samarati, P, “Database Security”, Addison Wesley, 456 p, 1995.

[11] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes”, In Advances in Cryptography—EUROCRYPT’99, pp. 223–238, Prague, Czech Republic, 1999.

[12] R. Rivest, L. Adleman, and M. Dertouzos, “On data banks and privacy homomorphisms”, In Foundations of Secure Computation, eds. R. A. DeMillo et al., Academic Press, pp. 169–179, 1978.

[13] R. Wright and Z. Yang, “Privacy-preserving bayesian network structure computation on distributed heterogeneous data”, In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2004.

[14] Z. Zhan, S. Matwin, and L. Chang, “Privacy-preserving collaborative association rule mining”, In 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, University of Connecticut, Storrs, CT, U.S.A., Aug. 7–10, 2005.

[15] Z. Zhan and S. Matwin, “A crypto-approach to privacypreserving data mining”, In IEEE International Workshop on Privacy Aspect of Data Mining, Hong Kong, Dec. 18–22, 2006.

[16] Er.M.R.Arun Venkatesh, Bharath University, Chennai, “Privacy-Preserving Updates to Anonymous and Confidential Database”, June-2012.

[17] Shweta Shrma Hitesh Gupta and Priyank Jain, “A Study Survey of Privacy Preserving Data Mining”, IJRICE April 2012.

[18] Divya Sharma, “A Survey on Maintaining Privacy in Data Mining”, IJERT April 2012.

[19] SWAGATIKA DEVI, “A SURVEY ON PRIVACY PRESERVING DATA MINING: APPROACHES AND TECHNIQUES”, IJEST March 2011.

[20] Tipawan Silwattananusarn and Assoc.Prof. Dr. KulthidaTuamsuk, “Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012”, IJDKP September 2012.