Maintaining Data Confidentiality in Association Rule Mining in Distributed Environment

Mr.Abhishek P. Bangar[1] Mr.Deokate G.S.[2] Mr.Kahate S.A[3]
[1,2,3]Department of Computer Engineering
[1,2,3]SPCOE. Dumbarwadi

Abstract— The data in real world applications is distributed at multiple locations, and the owner of the databases may be different people. Thus to perform mining task, the data needs to be kept at central location which causes threat to the privacy of corporate data. Hence the key challenge is to applying mining on distributed source data with preserving privacy of corporate data. The system addresses the problem of incrementally mining frequent itemsets in dynamic environment. The assumption made here is that, after initial mining the source undergoes into small changes in each time. The privacy of data should not be threatened by an adversary i.e. the miner and target database owner should not be able to recover original data from transformed data.

Key words: Association Rules, Frequent Pattern Mining, Security, Integrity, Protection

I. Introduction

Data mining technology which reveals patterns in large databases could compromise the information that an individual or an organization regards as private. The aim of privacy-preserving data mining is to find the right balance between maximizing analysis results (that are useful for the common good) and keeping the inferences that disclose private information about organizations or individuals at a minimum.

The amount of data kept in computer files is growing at a phenomenal rate. It is estimated that the amount of data in the world is doubling every 20 months [otIC98]. At the same time, the users of these data are expecting more sophisticated information. Simple structured languages (like SQL) are not adequate to support these increasing demands for information. Data mining attempts to solve the problem. Data mining is often defined as the process of discovering meaningful, new correlation patterns and trends through non-trivial extraction of implicit, previously unknown information from large amount of data stored in repositories using pat- tern recognition as well as statistical and mathematical techniques [FPSSU96]. A SQL query is usually stated or written to retrieve specific data while data miners might not even be exactly sure of what they require. So, the output of a SQL query is usually a subset of the database; whereas the output of a data mining query is an analysis of the contents of the database Data mining can be used to classify data into predefined classes (classification), or to partition a set of pat- terns into disjoint and homogeneous groups (clustering), or to identify frequent patterns in the data, in the form of dependencies among concepts-attributes (associations). The focus in this thesis will be on the associations.

In general, data mining promises to discover unknown information. If the data is personal or corporate data, data mining offers the potential to reveal what others regard as private. This is more apparent as Internet technology gives the opportunity for data users to share or obtain data about individuals or corporations. In some cases, it may be of mutual benefit for two corporations (usually competitors) to share their data for an analysis task. However, they would like to ensure their own data remains private. In other words, there is a need to protect private knowledge during a data mining process. This problem is called Privacy Preserving Data Mining (PPDM).

Privacy-preserving data mining studies techniques for meeting the potentially conflicting goals of respecting individual rights and allowing legitimate organizations to collect and mine massive data sets. Association rule mining finds interesting associations and/or correlation relation-ships among large sets of data items [AIS93]. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is Market Basket Analysis.

A Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. The challenge here is: how can we mine the data across the distributed sources securely or without either party disclosing its data to the others? Most of the algorithms developed in this field do not take privacy into account because the focus is on efficiency.

A simple approach to mining private data over multiple sources is to run existing data mining tools at each site independently and combine the results [Cha96] [PC00]. However, this approach failed to give valid results for the following reasons: Values for a single entity may be split across sources. Data mining at individual sites will be unable to detect cross-site correlations. The same item may be duplicated at different sites, and will be over- weighted in the results. Data at a single site is likely to be from a homogeneous population. Important geographic or demographic distinctions between that population and others cannot be seen on a single site. There is a tradeoff between privacy and efficiency in privacy-preserving problems. The solutions may be even impractical when complete privacy protection is required. In real world applications, controlled and limited information disclosure is usually acceptable.

II. Review of Mining in Distributed Environment

Many Authors represent many techniques for Association rules, data mining, mining methods and algorithms, security, integrity, and protection etc.

In [1] W. K. Wong, D. W. Cheung, E. Hung, and H. Liu have presented Protecting privacy in incremental maintenance for distributed association rule mining, Distributed association rule mining algorithms are used to discover important knowledge from databases. Privacy concerns can prevent parties from sharing the data. They design new algorithm to solve traditional mining problems without disclosing (original or derived) information of their own data to other parties.

In[2] M. Ahluwalia, A. Gangopadhyay, and Z. Chen have presented “Preserving Privacy in Mining Quantitative Association Rules”. The author of this paper introduces a new method for hiding sensitive quantitative association rules based on the concept of genetic algorithm. Genetic algorithm is employed to find the interesting quantitative rules from the given data set and weighing mechanism is used in that paper to identify the transactions for data perturbation, thereby reducing number of modifications to the database and preserving the interesting non sensitive rules. The main purpose of author is fully support the security of the database and to maintain the utility and certainty of mined rules at highest level.

In [3] M. Ahluwalia, A. Gangopadhyay, Z. Chen, and Y. Yesha, have presented " Tar-get-Based Privacy Preserving Association Rule Mining”, Author of this paper consider a special case in association rule mining where mining is conducted by a third party over data located at a central location that is updated from several source locations. The data at the central location is at rest while that flowing in through source locations is in motion.author impose some limitations in that paper to the central target location tracks and privatizes changes and a third party mines the data incrementally.

In [4] S. Rizvi and J. R. Haritsa, have presented "Maintaining Data Privacy in Association Rule Mining,", Author of this paper investigate with respect to mining association rules, That the users can be encouraged to provide correct information by ensuring that the mining process cannot, with any reasonable degree of certainty, violate their privacy. Author of this paper present a scheme, based on probabilistic distortion of user data, that can simultaneously provide a high degree of privacy to the user and retain a high level of accuracy in the mining results. In that paper the performance of the scheme is validated against representative real and synthetic datasets.

In [5] A. Evfimevski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Data Mining is extraction of rich knowledge from poor information that it is in database. These databases consist of sensitive and insensitive data or knowledge. Sensitive data are those data which consist of secret and important information and whose owners don’t want it to be leaked. Privacy preservation is also required for secure transformation of personal data. In this paper author have presented different types of association rules to be mined and how different privacy preserving techniques and algorithms for different levels of mining can be applied on them to protect the privacy of data with less information loss and high accuracy.

In [6] W. Cheung and O. R. Zaïane, "Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint," The propose of a novel data structure called CATS Tree. CATS Tree extends the idea of FP-Tree to improve storage compression and allow frequent pattern mining without generation of candidate item sets. Author proposed algorithms enable frequent pattern mining with different supports without rebuilding the tree structure. The algorithms allow mining with a single pass over the database as well as efficient insertion or deletion of transactions at any time.

In [7] X. Li and S. Sarkar, have presented "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining," A frequently used disclosure protection method is data perturbation. When used for data mining, it is desirable that perturbation preserves statistical relationships between attributes, while providing adequate protection for individual confidential data. To achieve this goal, Author propose a kd tree based perturbation method, which recursively partitions a data set into smaller subsets such that data records within each subset are more homogeneous after each partition. The confidential data in each final subset are then perturbed using the subset average.

In[8] C. C. Aggarwal and P. S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Privacy preserving data mining has become an important problem because of the large amount of personal data which is tracked by many business applications. In many cases, users are unwilling to provide personal information unless the privacy of sensitive information is guaranteed. In this paper, Author propose a new framework for privacy preserving data mining of multi-dimensional data a new and flexible approach for privacy preserving data mining which does not require new problem-specific algorithms, since it maps the original data set into a new anonymized data set. This anonymized data closely matches the characteristics of the original data including the correlations among the different dimensions.

In [9] V. Ganti, J. Gehrke, and R. Ramakrishnan, "DEMON: Mining and monitoring evolving data," The input data to a data mining process resides in a large data warehouse whose data is kept up-to-date through periodic or occasional addition and deletion of blocks of data. In this paper, Author consider a dynamic environment that evolves through systematic addition or deletion of blocks of data. They introduce a new dimension, called the data span dimension, which allows user-defined selections of a temporal subset of the database. Taking this new degree of freedom into account, They describe efficient model maintenance algorithms for frequent itemsets and clusters. They then describe a generic algorithm that takes any traditional incremental model maintenance algorithm and transforms it into an algorithm that allows restrictions on the data span dimension.

In [10] W. Cheung and O. R. Zaïane, "Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint,".They propose a novel data structure called CATS Tree. CATS Tree extends the idea of FP-Tree to improve storage compression and allow frequent pattern mining without generation of candidate item sets. The proposed algorithms enable frequent pattern mining with different supports without rebuilding the tree structure.

In [11] X. Xiao and Y. Tao, "m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets," This paper remedies the drawback which is none of the existing solutions supports re-publication of the microdata , after it has been updated with insertions and deletions. They reveal the characteristics of the re-publication problem that invalidate the conventional approaches leveraging k-anonymity and l-diversity. Based on rigorous theoretical analysis, They develop a new generalization principle m-invariance that effectively limits the risk of privacy disclosure in re-publication.

III. CONCLUSION

Frequent Pattern Mining is an initial component in many applications and opinion mining frameworks. We compared between standard and focused web crawlers to understand which one is better and apply it in our opinion mining framework in a future work.

ACKNOWLEDGMENT

I would be thankful to my guide assistant professor Mr. Deokate G.D. here for help when I have some troubles in paper writing. I will also thanks to Mr. Kahate S.A. (Assistant Professor, SPCOE, Dumbarwadi) and my other faculty members and class mates for their concern and support both in study and life.

REFERENCES

[1] W. K. Wong, D. W. Cheung, E. Hung, and H. Liu, "Protecting privacy in incremental maintenance for distributed association rule mining," PAKDD'08: Proceedings of the 12th Pacific Asia conference on Advances in knowledge discovery and data mining, 2008.

[2] M. Ahluwalia, A. Gangopadhyay, and Z. Chen, "Preserving Privacy in Mining Quantitative Association Rules," International Journal of Information Security and Privacy, 2010.

[3] M. Ahluwalia, A. Gangopadhyay, Z. Chen, and Y. Yesha, "Target Based Privacy Preserving Association Rule Mining," presented at 26th ACM Symposium on Applied Computing, TaiChung, Taiwan, 2011.

[4] S. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," VLDB, pp. 682-693, 2002.

[5] A. Evfimevski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02), pp. 217 - 228, 2002.

[6] W. Cheung and O. R. Zaïane, "Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint,"In Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS'03), pp. 111-116, 2003.

[7] X. Li and S. Sarkar, "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 18, pp. 1278-1283, 2006.

[8] C. C. Aggarwal and P. S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," 9th International Conference on Extending Database Technology, pp. 183-199, 2004.

[9] V. Ganti, J. Gehrke, and R. Ramakrishnan, "DEMON: Mining and monitoring evolving data," Proc. 16th Int. Conf. Data Engineering, San Diego, CA, pp. 439-448, 2000.

[10] W. Cheung and O. R. Zaïane, "Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint,"In Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS'03), pp. 111-116, 2003.

[11] X. Xiao and Y. Tao, "m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets," Proceedings of ACM International Conference on Management of Data (SIGMOD), pp. 689-700, 2007.