INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

COMPARATIVE STUDY ON DISRTIBUTED DATA MINING AND ITS PRIVACY

MISS. PALLAVI A. DHANDE[1], PROF. N. S. BAND[2], PROF. H. R. DESHMUKH[3]

1. Department of Computer Science and Engineering, IBSS College of Engineering, Amravati, Maharashtra, India.
2. Professor, Department of Computer Science and Engineering, IBSS College of Engineering, Amravati, Maharashtra, India.
3. Head and Professor, Department of Computer Science and Engineering, IBSS College of Engineering, Amravati, Maharashtra, India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Distributed Data Mining (DDM) algorithms focus on one class of such distributed problem solving tasks— modeling and analysis of distributed data. This paper offers a perspective on DDM algorithms in the context of multi agents systems. Data mining can extract important knowledge from large data collections – but sometimes these collections are split among various parties. It provides a high-level survey of DDM, and then focuses on distributed clustering algorithms and some potential applications in multi-agent-based problem solving scenarios. Addresses secure mining of association rules over horizontally partitioned data. It incorporates cryptographic techniques to minimize the information shared, while adding little changes to the mining task.

Keywords: Secure Distributed Data-mining, Secure Distributed Summation, Multi agent system, Clustering, privacy.

Corresponding Author: MISS. PALLAVI A. DHANDE

Access Online On:

www.ijpret.com

How to Cite This Article:

Pallavi A. Dhande, IJPRET, 2015; Volume 3 (9): 1112-1118

PAPER-QR CODE

INTRODUCTION

This paper 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 a typical distributed environment analyzing distributed data is a non-trivial problem because of many constraints such as limited bandwidth (e.g. wireless networks), privacy-sensitive data, distributed compute nodes, only to mention a few.

The field of Distributed Data Mining (DDM) deals with these challenges in analyzing distributed data and offers many algorithmic solutions to perform different data analysis and mining operations in a fundamentally distributed manner that pays careful attention to the resource constraints. Data-mining where the data of interest is distributed across several databases and one cannot combine is centrally to perform operations on the data. More specifically, the problem under consideration is a problem sometimes called “secure multiparty computation”, or “privacy preserving data-mining”. 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. It particularly focuses on distributed clustering, a problem finding increasing number of applications in sensor networks, distributed information retrieval, and many other domains. The paper provides a detailed literature review of existing clustering algorithms in DDM (including privacy-preserving ones). In some applications the knowledge of the agents that guide reasoning and action depend on the existing domain theory. However, in many complex domains this knowledge is a result of the outcome of empirical data analysis in addition to pre-existing domain knowledge.

II. LITERATURE REVIEW

Josenildo C. da Silva proposed that, Data mining [4], [5], [6], and [7] deals with the problem of analyzing data in scalable manner. DDM is a branch of the field of data mining that offers a framework to mine distributed data paying careful attention to the distributed data and computing resources. However, most of the centralized ensemble-based method algorithms are not specifically designed to deal with stream data. There are two applications: calculating the sum of traffic across a set of networks, and detecting distributed Internet health problems. As noted above, the total Internet traffic is unknown (though it is estimated in [11]). This is a good direction for future research. Algorithms such as [8], [9], [10] deal with the limited communication issue by transmitting compact, lossy models (rather than complete specifications of the clustering's), which may be necessary for a sensor-network-based application.

The paper presents a method for performing computations on shared data without any participants revealing their secret data. This collective "intelligence" of a multi-agent system must be developed by distributed domain knowledge and analysis of distributed data observed by different agents. In a multi-agent system this knowledge is usually collective.

It seems unlikely that a centralized organization can build the type of trust required to have almost instantaneous access to traffic or performance data that they might need to analyze an arbitrary security threat. Highly optimistic traffic-growth estimates underlay the hype at the peak of the Internet boom. On the other hand, security threats act on the short term (seconds to hours). Murat Kantarcıoglu and Chris Clifton, has states in their paper that fast algorithm for distributed association rule mining is given in Cheung et. al. [12]. Their procedure for fast distributed mining of association rules (FDM) is summarized below.

1) **Candidate Sets Generation**: Generate candidate sets CGi(k) based on GLi(k−1), item sets that are supported by the Si at the (k-1)-th iteration, using the classic apriority candidate generation algorithm. Each site generates candidates based on the intersection of globally large (k-1) item sets and locally large (k-1) item sets.

2) **Broadcast Mining Results**: Each site broadcasts the local support for item sets in [iLLi(k). From this, each site is able to compute L(k). The details of the above algorithm can found in [12].

3) **Support Count Exchange**: LLi(k) are broadcast, and each site computes the local support for the items in [iLLi(k)].

4) **Local Pruning**: For each X 2 CGi(k), scan the database DBi at Si to compute X.supi. If X is locally large Si, it is included in the LLi(k) set. It is clear that if X is supported globally, it will be supported in one site.

Most of the distributed clustering algorithms are still in the domain of academic research with a few exceptions. Therefore, the scalability properties of these algorithms are mostly studied for moderately large number of nodes. Distributed clustering algorithms for this domain must address these challenges. For example, most of these distributed clustering algorithms are lot more communication efficient compared to their centralized counter parts. We would like to see secure algorithms for classification, clustering, etc. 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. For heterogeneous data, the number of choices for distributed clustering algorithms is relatively limited. However, there exist several techniques for this latter scenario. Another possibility is secure approximate data mining algorithms. Collectively a measure of confidentiality in a distributed data mining context has to quantify the difficulty for one mining peer to disclose the confidential information owned by other peers. The observation that the clustering algorithm doesn't need the exact density estimate functions but an essential approximation. We believe the need for mining of data where access is restricted by privacy concerns will increase. Examples include the "essential approximation" in this case is a sampling of points which is as coarse as possible to preserve data confidentiality while maintaining information to guide the clustering process. Knowledge discovery among intelligence services of different countries and collaboration among corporations without revealing trade secrets. The secure multi-party computation definitions from the cryptography domain may be too restrictive for our purposes. Privacy preserving data mining can be done with a reasonable increase in cost over methods that do not maintain privacy.

In summary, it is possible to mine globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. Not only must the privacy of patient records be maintained, but insurers will be unwilling to release rules pertaining only to them. The problem is that insurance companies will be concerned about sharing this data. If this rule doesn't hold globally, the insurer would like to know this – they can then try to pinpoint the problem with their policies and improve patient care. Imagine a rule indicating a high rate of complications with a particular medical procedure. The solution is efficient: The additional cost relative to previous non-secure techniques is O (number of candidate item sets sites) encryptions, and a constant increase in the number of messages. The solution is efficient: The additional cost relative to previous non-secure techniques is O (number of candidate item sets sites) encryptions, and a constant increase in the number of messages. The observations are usually time-series data sampled at a device specific rate. Therefore, collaboration with other sensor-nodes would require comparing data observed at different nodes. Distributed data mining technology offers more efficient solutions in such applications. The following discussion illustrates the power of DDM algorithms using a simple randomized technique for addressing this sensor network-related problem. Private Association Rule Mining has the basic approach outlined on except that values are passed between the local data mining sites rather than to a centralized combiner. Privacy of the data can be another reason for adopting the DDM technology. In many applications, particularly in security-related applications, data is privacy-sensitive (confidential). Previous work in privacy-preserving data mining has addressed two issues. In one, the aim is preserving customer privacy by distorting the data values [13].

III. CONCLUSION

This paper suggests that traditional centralized data mining techniques may not work well in many distributed environments where data centralization may be difficult because of limited bandwidth, privacy issues and/or the demand on response time. it is possible to mine globally valid results from distributed data without revealing information that compromises. There are many avenues for future research on this topic. Apart from obvious possibilities such as collecting statistics for each region geographic, or performance metrics between regions. Multi-agent systems are fundamentally designed for collaborative problem solving in distributed environments. Many of these application environments deal with empirical analysis and mining of data. Continued research will expand the scope of privacy-preserving data mining, and focusing all data mining methods to be applied

in situations where privacy concerns. It surveyed the data mining literature on distributed and privacy-preserving clustering algorithms. Multi-agent systems are fundamentally designed for collaborative problem solving in distributed environments. We believe the need for mining of data where access is restricted by privacy concerns will increase. It noted that while these algorithms usually perform better than their centralized counter-parts on grounds of communication efficiency and power consumption, there exist several open issues. It surveyed the data mining literature on distributed and privacy-preserving clustering algorithms. This paper underscores the possible synergy between MAS and DDM technology. The paper provides a detailed literature review of existing clustering algorithms in DDM (including privacy-preserving ones). It particularly focuses on distributed clustering, a problem finding increasing number of applications in sensor networks, distributed information retrieval, and many other domains. Implementing a method of collecting meaningful Internet wide statistics of great use to providers and researchers alike. It provides a high-level survey of DDM, and then focuses on distributed clustering algorithms and some potential applications in multi-agent-based problem solving scenarios.

IV. REFERENCES

1. Josenildo C. da Silva2, Chris Giannella et. All“Distributed Data Mining and Agents**”.**

2. Murat Kantarcıoˇglu and Chris Clifton, et All "Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data".

3. Matthew Roughan "Secure Distributed DataMining and Its Application to LargeScale Network Measurements" Matthew Roughan School of Mathematical ScienceUniversity of Adelaide SA 5005, Australia.

4. Han J. and Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufman Publishers, San Francisco, CA, 2001.

5. Hand D., Mannila H., and Smyth P. Principals of Data Mining. MIT press, Cambridge, Mass, 2001.

6. Hastie T., Tibshirani R., and Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, Berlin, Germany, 2001.

7. Witten I. and Frank E. Data Mining: Practical Machine Learning Tools and Techniques with java Implementations. Morgan Kaufman Publishers, San Fransisco, 1999.

8. Januzaj E., Kriegel H.-P., and Pfeifle M. DBDC: Density Based Distributed Clustering. In Proceedings of EDBT in Lecture Notes in Computer Science 2992, pages 88–105, 2004.

9. Merugu S. and Ghosh J. Privacy-Preserving Distributed Clustering Using Generative Models. In Proceedings of the IEEE Conference on Data Mining (ICDM), 2003.

10. Samatova N., Ostrouchov G., Geist A., and Melechko A. RACHET: An Efficient Cover- Based Merging of Clustering Hierarchies from Distributed Datasets. Distributed and Parallel Databases, 11(2):157–180, 2002.

11. A. M. Odlyzko. Internet traffic growth: Sources and implications. In B. B. Dingel, W. Weiershausen, A. K. Dutta, and K.-I. Sato, editors, Optical Transmission Systems and Equipment for WDM Networking II, volume 5247, pages 1–15. Proc. SPIE, 2003.

12. D. W.-L. Cheung, J. Han, V. Ng, A. W.-C. Fu, and Y. Fu, "A fast distributed algorithm for mining association rules," in Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (PDIS'96). Miami Beach, Florida, USA: IEEE, Dec. 1996, pp. 31–42.

13. R. Agrawal and R. Srikant, "Privacy-preserving data mining," in Proceedings of the 2000 ACM SIGMOD Conference on Management of Data. Dallas, TX: ACM, May 14-19 2000, pp. 439–

450.Available: http://doi.acm.org/10.1145/342009.335438.4) Benny Pinkas. Cryptographic techniques for privacy-preserving data mining. ACMSIGKDD Explorations Newsletter, 4(2):12–19, 2002.

14. Park B. and Kargupta H. Distributed Data Mining: Algorithms, Systems, and Applications. In The Handbook of Data Mining, edited by N. Ye, Lawrence Erlbaum Associates, pages 341–358, 2003.

15. J. C. Benaloh and M. de Mare, "One-way accumulators: A decentralized alternative to digital signatures," in Advances in Cryptology – EUROCRYPT'93, Workshop on the Theory and Application of Cryptographic Techniques, ser. Lecture Notes in Computer Science, vol. 765. Lofthus, Norway: Springer-Verlag, May 1993, pp. 274–285. [Online]. Available: http://springerlink.metapress.com/openurl.asp?genre=article&issn=03029%743&volume=765 &spage=274.