International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Volume-4, Issue-10, 2017 Regular Edition E-ISSN: 2349-7084

A Survey on Data Mining Techniques in Distributed Databases

* J. Keziya Rani

*Assistant professor, Department of Computer Science & Technology, S.K university, Anantapuramu.Andhra Pradesh, India,

Email ID: kejiyaraj@gmail.com

Available online at: http://www.ijcert.org

Received: 02/08/2017 Revised: 25/09/2017, Accepted: 18/10/2017, Published: 30/10/2017

Abstract: Data Mining is the process of discovering interesting patterns & Knowledge from large amounts of data. The data can be gathered from different sources like Databases, Data Warehouses, the web, other information Repositories. In client/server environment Distributed Databases play an important role for information processing and it is easy to foresee that their importance will rapidly grow. Because of increased growth of data sharing, distributed databases are developed. In Distributed Databases it is difficult to maintain large amounts of data in centralized databases without duplication & to maintain secured data. In distributed databases, copy of data is stored in different locations, where memory wastage is heavy. In many situations data is gathered from many places for analysis, where a privacy problem occurs. So, to avoid all this Data Mining give Distributed Data Mining(DDM) to identify correct & perfect information. In this paper I am going to concentrate on how to give security to Distributed Databases by using Data Mining Techniques like Privacy Preserving Data Mining(PPDM),Association Rule for frequent patterns, Horizontal & Vertical partitions etc.

Keywords: — Distributed Databases, Data Sharing, Centralized data, duplication, Data Mining Techniques, privacy Preserving Data Mining(PPDM), Distributed Data Mining(DDM),Association Rules, frequent patterns, Horizontal & Vertical Partitions.

---

1. Introduction

When an organization is distributed in many areas, it may choose to store all the information in a central computer called server to distribute them to all other local computers. With this it is possible to share information to all computers at one time. A Distributed Database is a single logical database that is distributed across many computers that are connected to a centralized system with network.

Various Organizations encourage the use of distributed databases because of :

- Distribution of data that belong to same business units
- Data sharing
- Data communication costs and reliability
- Database recovery
- Satisfying both transaction and analytical processing.[1]

Security problems in Distributed databases are:

- Integrity
- Centralized or Decentralized Authorization & Authentication.
- One of the main problems in Distributed Databases is where to grant system access.
- Users are granted system access from their home site or remote site.
- Success depends on Reliable Communication between different sites.
- Since different sites can give permission to grant access, the probability of unauthorized access increases.
- Distributed Database servers are arranged in network. When a request occurs to access from centralized server, all the members of the network determine whether to give permission or not.
- In Distributed Databases, the same copy of data is distributed on different sites – so there is a problem of duplication that occurs. With this integrity problems arise.

- Integrity maintenance is difficult in heterogeneous distributed databases.
- Problems occur in heterogeneous distributed databases. They are
  - Inconsistency between local integrity constraints.
  - Inconsistencies between local & global constraints etc. [2]

2. Data Mining

Data Mining should be named as "Knowledge mining from data", it means searching for knowledge in data i.e mining of knowledge from data, knowledge extraction, data/pattern analysis. The synonym for data mining is "Knowledge discovery from data" or KDD. Data Mining is an important step in discovering knowledge. The steps in Knowledge discovery process are:

1. **Data Cleaning:** To remove repeated data & inconsistent data.
2. **Data integration:** Different data sources were combined.
3. **Data Selection:** Data related to the task of analysis are retrieved from the database.
4. **Data transformation:** Data is transformed and consolidated into aggregate forms for mining.
5. **DataMining:** A process where intelligent methods are applied to extract data patterns.
6. **Pattern evaluation:** To give knowledge, evaluate the interesting patterns.

Knowledge Presentation: knowledge representation techniques are added to present mined knowledge to users. Data Mining can be applied to other forms of data, such as data streams, ordered/sequence data. Graph data, spatial data, text data, multimedia data, and the WWW.[3]. Data Mining located data is stored at different locations by using Distributed Data Mining(DDM) or Collective Data Mining(CDM).

In DDM intrusions can be launched from several different locations and targeted to many different destinations. Distributed data mining methods may be used to understand network data from several network locations to detect these distributed attacks.[3]

Privacy & Security of Data Mining: Large amounts of Data is available in electronic form on the web, powerful data mining tools have been developed to provide privacy & security of Data Mining. Many data mining applications do not touch individual private data. Examples include applications involving natural resources, the prediction of floods and droughts, astronomy, geography, biology and other scientific & engineering data. Real privacy concerns are with unconstrained access to individual records, especially access investigations and ethnicity. Privacy Sensitive data like credit card transaction records, health-care records, personal financial records, biological traits, criminal/justice etc. So, removing sensitive IDs from data may protect the privacy of most individuals.

Some of the advances in protecting privacy and data security in data mining are:

Data Security-enhancing techniques have been developed to protect data. Databases can employ a ***multilevel security model*** to classify & restrict data according to various security levels, with users permitted access to only their authorized level.

Encryption is another technique in which individual personal private data items may be encoded. This may involve blind signature(build on public key encryption), ***biometric encryption***(image of persons iris or fingerprint is used to encode his or her personal information), and ***anonymous databases***(permit the consolidation of various databases but limit to access personal private data only to those who need to know, personal data is encoded and stored at different sites).

3.Distributed Data Mining (DDM)

Data Mining in Distributed Environment means Distributed Data Mining (DDM). Distributed Data Bases plays an important role in Data Mining. Data Mining requires huge amount of data from different sites in a network. It needs lot of storage space & computational time. So to distribute the data to different sites in a flexible & scalable way DDM is necessary. Another important approach is it is necessary that all data is at centralized site, so that data is distributed from centralized site. With the centralized distribution of data to different sites there is a possibility of duplication of data, security & privacy problems. Distributed Data Mining is a powerful technology to identify the data in Distributed Data Bases. In Distributed databases data is divided into partitions called Horizontal Partition and Vertical Partition.

4.Data Mining Techniques in Distributed Databases

Data Mining Techniques in Distributed Databases are

1. Privacy preserving Distributed Data Mining (PPDM).
2. Association Rule for frequent patterns.
3. Horizontal & Vertical Partitions.

4.1 Privacy Preserving Distributed Data Mining (PPDM)

Ahmed Hafyasien[4] presented a PPDM framework. It contains three levels, at First level Transactional Database & Raw Data exist. In Second level Data Mining algorithms are developed & in Third level Rules, Patterns or Results are developed.

At Fist level Raw data is collected from different sites for analysis. So, maintenance of privacy at this stage is necessary, because data is taken from different sites & distributed for transaction purposes.[4]

At Second level Data Mining algorithms are applied for mining the data that is taken from different databases. The different process applied at this level are Modification, blocking, generalization, suppression, sampling etc. Data mining algorithms are modified for providing PPDM.[4] At Third level patterns or results are checked after gaining the knowledge. [4]

PPDM Techniques can be classified as[5]:

A)Data Modification: To provide high privacy protection, Data Modification is used to change the unique value of a database & allow the public. Methods of data modification include:

i) **Blocking**: Replacement of an existing attribute value with a “?”.

ii) **Perturbation**: Replacing attribute value by a new value (Changing a 1-value to a 0-value or adding noise).

iii)**Swapping**: Interchanging values of Individual Record.

iv) **Encryption**: Cryptographic techniques are used for encryption.

B) Data or Rule Hiding: Data Hiding means protecting sensitive data. To maintain security for people’s personal data. Rule Hiding means protecting confidential data. E.g.: Association Rule. Hiding aggregated data in the form of rules is difficult.

C) Data Distribution: Data distributed from centralized data. Centralized data can be divided as Horizontal data distribution & Vertical data distribution. Horizontal distribution refers to different sets of records exist in different places. Vertical data distribution refers to different attributes reside in different places.

D) Data Mining Algorithm: The Data Mining Algorithm for Privacy Preservation Techniques are:

i)) Clustering Algorithm

ii) Classification Data Mining Algorithm

iii) Association Rule Mining Algorithm

E) **Privacy Preserving Techniques:**

1) Heuristic bases Technique: Modifies only selected values that minimize the effective loss rather than all available values.

2) Cryptographic based Technique:

The technique includes secure multiparty computation. Cryptography based algorithms are considered for protective privacy in distributed situations by using encryption techniques.

3) Reconstruction based Techniques:

The original distribution of data is reassembled from the randomized data. Nidhi Saxena et al., [6] proposed the need for privacy preserving data mining is necessary in digital data. Because there is increased growth of data. They also stated that inaccuracy in transformed data mined or analyzed data is necessary.

Hina Vaghashia et al., [7] said that Privacy Preserving data mining provides privacy to sensitive data. It hides data, so data cannot be revealed to unauthorized people. Privacy & Security are pair parts in data mining but providing one can give ambiguity to other. Also said that still now there is no single algorithm to provide all in one like performance, cost, utility, complexity, tolerance etc.

Jaideep Vaidya & chris Clifton[8] said that the benefits of data mining & analysis are acceptable, government says they are not considerable to price the privacy of Individual. If compromised, they create some problems for individual privacy. So, it is solved neither the government give complete rights to access individual data or banned completely.

S.B.Javheri,U.V.Kulkarni[9] proposed privacy preservation in Machine Learning algorithms require time consuming encoding & decoding techniques are necessary for plain text.

A.V.Sriharsha et al.,[19] proposed Privacy Preserving technology can solve only one side of the problem. When using this technique, the non-technical difficulties faced by decision makers. The problems are Loss of valuable information, decreasing of data/service quality, increased costs & Complexity. They suggest that cross-disciplinary research is the key to remove all obstacles & conduct research with social scientists in psychology, sociology & public studies.

Nivetha.P.R et al.,[20] presented sequential pattern mining overview, in data mining it is also one of the most effective areas & for extracting sequence of Knowledge, Sequential pattern hiding method also discussed.

Gayatri Nayak et al.,[21] proposed that in Distributed Privacy Preserving Data Mining, efficiency is an important issue. They try to develop the most effective algorithm & give equal balance between different costs like computation cost, disclosure cost & communication cost.

Chin-chen chang et al.,[11] said that the development of Superior privacy-preserving algorithms to further reduce computation complexity & increase the security without sharing the data in distributed database environment.

2)Association Rules for frequent patterns:

The Increased growth of data in different organizations, data processing is a main point of Information processing. Mining of Association rules in large databases is the important task. Association rules play an important role in distributed data mining techniques. In "Market Basket Analysis" the purchasing of one item by customer is compared with another item by using association rules.Association rules are used to show the relationship between frequent data items. Association rules are frequently used in different fields like advertising, marketing and inventory mart. In Market Basket Analysis find the relationship between items that are purchased by customers[23][24].

Distributed/parallel Databases or data warehouses may store large amounts of data to be mined. Mining association rules in such databases may require lot of processing [25]. An expected solution to this problem can be a distribution system. [26]. Many large databases are distributed and with this it is possible to use more feasible distribution algorithms. Mainly use association rules is the computation of the set of large item sets in the database. Distributed computing of large item sets in distributed databases find some new problems. One may compute locally large Item sets in distributed databases easily, but a locally large item sets in distributed databases may not be globally large. Since it is high cost to give the whole data set to other distributed sites, one option is to present all the counts of all the item sets, no matter locally large or small, to other distributed sites. However, a database may contain enormous combinations of item sets, and it will involve passing a huge number of messages.

Distributed Algorithms [27]

1. Distributed association rule learning
2. Collective decision tree learning
3. Collective PCA and PCA-based clustering
4. Distributed hierarchical clustering
5. Other distributed clustering algorithms
6. Collective Bayesian network learning

The Four Parallel Algorithms are

1) Count Distribution: Count Distribution measures the frequency of a pattern inside a distributed database
2) Candidate Distribution: parallelizing the task of presenting the longer patterns of distributed databases.
3) Hybrid Distribution and Candidate Distribution: A hybrid distribution algorithm that tries to combine the strengths of the above algorithms
4) Sampling with Hybrid Count and Candidate Distribution: An algorithm that tries to only use a sample of the Distributed Databases.

In a parallel distributed data mining the main issues considered are

1. Load balancing
2. Minimizing communication
3 Overlapping communication and computation

An Apriori algorithm is widely used to find out the frequent item sets from databases. But it is not efficient in large databases because it will require more Input/Output load. Later drawback of Apriori algorithm find different algorithms to overcome the disadvantages, but those algorithms also fail to overcome their disadvantages. Hence hybrid architecture is proposed which consists of both distributed and parallel computing concepts. It combines both architectures so that it will be easy to find out frequent item set from large databases in short time.[22]. A distributed data mining algorithm FDM (Fast Distributed Mining of association rules) has been proposed by [26].

FDM Algorithm

Following are the steps for FDM algorithm

1. Initialization
2. Candidate set generation
3. Local Pruning
4. Unifying the candidate itemset
5. Computing local support
6. Broadcast the mining results

Cheung [28] proposed Fast distributed mining for association rules, for distributed databases, Apriori algorithm for association rule mining for individual databases. In [10] proposed an in secured version of FDM algorithm but algorithm violates privacy in two stages, in step four and step six. In step 4 whenever the players broadcast the item sets that are locally frequent in their private databases. In step 6 they present the sizes of the local supports of candidate item sets. Kantarioglu & Clifton [10] proposed secure implementations of these steps and in [29] describe the various implementations and proceed to analyze implementations in terms of privacy, efficiency and compare them.

Kantareioglu,M. And Clifton,C.[10] proposed to get an accurate data mining result on distributed databases & to store the private data that is acquired. They proposed a scheme to mine association rules on horizontally partititoned data called EKCS(Enhanced Kantareioglue and Clifton Scheme). Which is two-phase, Privacy-Preserving, distributed data mining scheme. It decreases the quantities of global data that are encrypted & reduces the transmission load without raising the risk of item sets leak in the first phase. In the first phase EKCS decreases the total number of items sets to encode & transmit without increasing the security risk. In the second phase, introduces two protocols for enhancing security against collusion.

3) Horizontal and Vertical Partitions:

Distributed data mining refers to the process where data is collected from multiple sources. To available data to multiple sources at a time, divide data into different partitions like[12]:

A. Horizontal Partitioning: Different sources of same database have different records containing same attributes.

B. Vertical Partitioning: Different sources of the same databases may have different attributes of the same set of records. In distributed data mining cryptography is used to achieve privacy [13, 14]. Both types of data partitions have algorithms for security. Murat Kantarcioglu, in his paper [15], analyse and discuss some privacy preservation techniques on horizontally partitioned data. Jaideep Vadiya, in his paper[16] a detailed survey performed on privacy preservation techniques on vertically partitioned data.

Some algorithms for cryptography can be applied to data on of its partition. N. Abitha et al., [13] presented changed approaches of Rail-fence algorithm and Vigenere cipher algorithm. Based on them, the changed Vigenere Cipher algorithm had more difficult procedures than changed Rail fence algorithm. On the converse encoding of records with the changed Vigenere cipher gives highly perfect results than the changed Rail fence algorithm. Hence based on the purpose, the result in hand, deadline and quantity of data sets the user can select the preferred algorithm.

The distributed data is stored on different clouds and apply different techniques for providing privacy preservation. One such approach is presented by Maria Luisa Merani, Cettina Barcellonay, Ilenia Tinnirelloy, in their paper [17] which derives and analyses two methods to secure multi-cloud data.

On clouds distributed data is in encoded form so data mining algorithms cannot be applied. Researches have been done to make data mining algorithms work on encoded data on cloud. One such work is done by Bharath K. Samanthula et al., [18]. They described k-NN algorithm to extract the data from relational database when it is encrypted.

Gayatri Nayak et al., [21] discussed different issues & Privacy Preserving Methods to distribute data & methods for handling Horizontal & Vertical Partitioned data.

Mohamed et al.,[30] proposed a privacy-preserving distributed KNN mining algorithm for horizontal partitioned data has been presented. The proposed algorithm is based on technology homomorphism and RSA encryption which is secured. No global computations at the centralized site are conducted but the KNN algorithm is computed locally for each site and local results are transferred to the centralized site to be compared. Experimental results show that PPDM has nice capability of privacy preserving, accuracy and efficiency, and relatively comparitative to classical approach.

5. Conclusion

Distributed Databases play an important role for data processing & the importance of Distributed Databases rapidly grow. When distributing data from distributed databases it is necessary to maintain unique data, to avoid duplication & to maintain security.This paper presents a brief survey on various distributed data mining techniques for Privacy Preserving Data Mining(PPDM), Association Rule for frequent patterns, Horizontal & Vertical Partitions to provide security.We concentrate on PPDM Techniques, Privacy of Individual data, to provide privacy of individual data, In providing privacy problems faced by decision makers, sequential pattern mining overview, to maintain privacy balance between different costs. So, privacy preserving algorithms reduce computation cost & increase security in distributed environment. FDM Algorithm, drawbacks of Association of frequent item sets in distributed databases.

Horizontal & Vertical Partition of distributed databases & different algorithms to maintain security are discussed. In future work, it is necessary to maintain security in distributed databases and effective algorithms need to be developed.

References

[1] Jefferey A. Hoffer. Mary B.Prescott Fred R. Mcfadden, *Modern Database Management*, 6th Ed, pearson education, PP:493-494.

[2] Dr. Mohmed Kashif Oureshi et al.IJAIR, *Security Aspects of Distributed Database*, ISSN:2278-7844,PP:201-203.

[3] Jiawei Han | Micheline Kamber | Jian Pei, *Data Mining Concepts and Techniques*, 3rd Ed, MK,PP: 5-8.

[4] Ahmed HajYasien. Thesis on "PRESERVING PRIVACY IN ASSOCIATION RULE MINING" in the Faculty of Engineering and Information Technology Griffith University June 2007.

[5] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, "State of the Art in Privacy Preserving Data Mining" Published in SIGMOD Record, 33, 2004,PP:50-57.

[6] Nidhi saxena, Priya Gupta,Onkar Singh, 2016,"A Survey on Security Techniques in Data Mining", IJASRM, Vol. 1, Issue 5, May 2016,ISSN:2455-6378,PP:159-162.

[7] Hina Vaghashia,Amit Ganatra,2015,"A Survey: Privacy Preservation Techniques in Data Mining", International Journal on Computer Applications(0975-8887),Vo.119-No.4,June 2015,PP:20-26.

[8] Jaideep vaidya, Chris Clifton,2004,"Privacy-Preserving Data Mining: Why, How, and When",

IEEE Computer Society, Security & Privacy(1540-7993/04),November/December,2004,PP:19-27.

[9] da Silva, J. C., Giannella, C., Bhargava, R., Kargupta, H., & Klusch, M. (2005). Distributed data mining and agents. *Engineering applications of artificial intelligence*, *18*(7), 791-807.

[10] Kantarioglu,M. and Clifton,C (2004) "Privacy Preserving Mining of association rules on horizontally partition data," IEEE Transactions on Knowledge & Data Engineering, Vol.16,No.9,PP:1026-1037.

[11] Chin-Chen Chang, Jich-Shan yeh & Yu-Chiang Li,2006,"Privacy-Preserving Mining of Association Rules on Distributed Databases", IJCSNS, International Journal of Computer Science & Network Security,Vol:6 No.11,PP:259-266.

[12] Aggarwal C, Yu P, "An Introduction to Privacy Preserving Data Mining", Chapter 2 in Privacy Preserving Data Mining: Models and Algorithms, Springer, NY, USA, Pg-11 to Pg-27, (2012)

[13] Abitha N, Sarada G, Manikandan G., Sairam, "A Cryptographic Approach for Achieving Privacy in Data Mining". International Conference on Circuit, Power and Computing Technologies, IEEE, (2015)

[14] Pinkas B., "Cryptographic techniques for privacy preserving data mining", SIGKDD Exploration, Vol.4 (Issue - 2), Pg-12 to Pg-19.

[15] Kantarcioglu, "A Survey of Privacy-Preserving Methods Across Horizontally Partitioned Data", Chapter 13, Privacy Preserving Data Mining: Models and Algorithms, Springer, NY, USA, Pg313 to Pg-332, (2012).

[16] Vadiya, "A Survey of Privacy-Preserving Methods Across Vertically Partitioned Data", Chapter 14 in Privacy Preserving Data Mining: Models and Algorithms, Springer, NY, USA, Pg-337 to Pg356, (2012).

[17] Merani M., Barcellonay C, Tinnirelloy I, "Multi Cloud Privacy Preserving Schemes for Linear Data Mining", IEEE, Communication and Information Systems Security, (2015).

[18] Samanthula, Elmehdwi, Jiang, "k-Nearest Neighbour Classification over Semantically Secure Encrypted Relational Data", IEEE Transactions on Knowledge and Data Engineering, (2013).

[19] A.V.Sriharsha,Dr.C.Parthasarathy,2015,"A Survey on Privacy Preserving Data Mining", Interrnational Journal of Advanced Research in Computer Science & Software Engineering, ISSN:2277 128X,Vol 5,Issue 10,October-2015,PP:631-636.

[20] Nivetha.P.R,Thamrai Selvi.K,2013,"A Surevey on Privacy Preserving Data Minging Techniques",IJCSMC,ISSN:2320-088X,Vol 2,Issue 10,October 2013,PP:166-170.

[21] Gayatri Nayak, Swagatika Devi,"A survey on Privacy Preserving Data Mining: Approaches & Techniques",IJEST, ISSN:0975-5462,Vol 3,No 3, March,2011,PP:2127-2133.

[22] Anil Vasoya, Dr.Nitin Koli,ELSEVIER,"Mining of association rules on large databases using distributed and parallel computing",1877-0509,2016,WWW.Sciencedirect.com, PP:221-230.

[23] L.Wang et al.," Efficient Mining of frequent Items sets on large uncertain databases",IEEE Transactions on Knowledge and Data Engineering, Vol.24,no.12,PP:2170-2183.Dec.2012.

[24] V.S.Tseng,"Efficent Algorithm for Mining High utility itemsets from transactional databases", IEEE Transactions on Knowledge and Data Engineering, Vol.25,no.8,PP:1772-1786.Aug.2013.

[25] Kaufmann, 1994,pp. 407-419. IEEE Tran. Knowledge and Data Eng. , vol. 8, no. 6, 1996,pp. 962-969;. Distributed Algorithm for Mining Association Rules," 31-42; (VLDB 94),

[26] J. Han , J. Pei, and Y. Yin , "Mining Frequent Patterns without Candidate Generation," Int'l. Conf. Management of Data , ACM Press, 2000,pp. 1

[27] Albert Y. Zomaya, Tarek El-Ghazawi, Ophir Frieder, "Parallel and Distributed Computing for Data Mining", IEEE Concurrency, 1999. International Journal of Computer Science and Information Technology, Volume 2, Number 2, April

[28] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. "Efficient mining of association rules in distributed databases". IEEE Trans. Knowl. Data Eng., 8(6):911–922, 1996Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[29] Tamir Tassa, "Secure Mining of Association rules in Horizontally Distributed Databases" IEEE trans. Knowledge and Data Engg. Vol. 26, no. 2, April 2014.

[30] Mohamed A.Ouda,Samesh A.Salem,Ihab A.Ali, and El-Sayed M.Saad,2012,"Privacy-Preserving Data Mining(PPDM) Method for Horizontally Partitioned Data",IJCSI, Vil.9,Issue 5,No 1, September 2012,ISSN:1694-0814,PP:339-347.

