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 5, May 2016**

# Implementing the Privacy Preservation Data Mining Based on Association Sharing Technique

Yamini M. Babnekar, Dr. Sheetal S. Dhande

Student, Dept. of CSE, Sipna College of Engineering and Technology, Amravati, India

Associate Professor, Dept. of CSE, Sipna College of Engineering and Technology, Amravati, India

**ABSTRACT -** The PPDM systems that currently exist makes the data secure using varied mechanisms. Cryptography is the most commonly used mechanism to achieve data integrity. To incorporate cryptographic techniques key generation and key exchange is an integral function to be achieved where in adversaries can benefit if improper techniques are adopted that are mainly used for data sharing. To overcome this drawback, in this paper we have implemented a Key Distribution-Less Privacy Preserving Data Mining (KDLPPDM) system. The adoption of the optimal data mining technique is also critical and must facilitate accurate analysis with the use of limited data. Limited work is carried out to study the effect of the varied data mining techniques in PPDM systems. Along with the KDLPPDM approach for efficient sharing of data, we have use the Association rule sharing model that perform collaborative filtering and make use of frequent Itemset Mining technique for finding out association frequency of data files present. And at the time of actual sharing of associated file the secret key once generate for the individual user needs to be entered properly for maintaining privacy of data mining.

**KEYWORDS:** Privacy Preserving Data mining (PPDM), Association rule, KDLPPDM, Keyword frequency, Key generation.

## I. INTRODUCTION

In business or corporate houses and government bodies possess certain framework or infrastructures for maintaining huge data collections for analyzing and processing it [1]. The information extracted from its local or confined databases are of sufficient for accomplishing or facilitating the expected results. Therefore, such shortcomings do require a platform or system that could effectively collect the huge distributed data and can perform the data mining to get the expected information that could be analyzed efficiently and precisely. Such scenarios put forth the need for privacy preservation in data mining (PPDM) systems [2]. The major objectives of the PPDM systems is to maintain the data integrity of the data published and to achieve efficient data mining results. There are many application areas that includes mainly secure data sharing track rebels or terrorists, security agencies or intelligence bureau and any other private office information sharing between their employees as in their intranets.

Data mining is capable of analyzing vast amount of information within a minimum amount of time. On the other hand, the excessive processing power of intelligent algorithms puts the sensitive and confidential information that resides in large and distributed data stores at risk. Providing solutions to database security problems combines several techniques and mechanisms. An organization may have data at different sensitivity levels [3]. This data is made available only to those with appropriate rights. Simply restricting access to sensitive data does not ensure complete sensitive data protection. Based on the knowledge of semantics of the application, the user may infer sensitive data items from non-sensitive data. Association rule mining is a technique in data mining that identifies the regularities found in large volume of data [4]. Such a technique may identify and reveal hidden information that is private for an individual or organization.

To know the data mining where association rule sharing is important consider the example of passenger information required at the time of traveling. In order to preserve privacy, passenger information records can be de-identified before the records are shared [5] [6]. This can be accomplished by deleting unique identity fields, such as name and passport number from the dataset. However, even if this information is deleted, there are still other kinds of

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 5, May 2016**

information like personal or behavioral information. It includes date of birth, zip code, gender, and number of children, number of calls, and number of accounts. When this information is linked with other available datasets, it could potentially identify subjects. To avoid such exposure of sensitive information, algorithm for privacy preservation in association rule mining becomes a must.

In the current business environment, its success is defined by collaboration, team efforts and partnership, rather than lonely spectacular individual efforts in isolation. So the collaboration becomes especially important as it brings mutual benefit for its users [7]. Sometimes, such collaboration even occurs among competitors, or among companies that have conflict of interests, but the collaborators are aware that the benefit brought by such collaboration will give them an advantage over other competitors. For this kind of collaboration and associative sharing, data's privacy becomes extremely important: all the parties of the collaboration promise to provide their private and associative data to all parties, but neither of them wants each other or any third party to learn much about their private data [8]. One of the major problems that accompany with the huge collection or repository of data is confidentiality. The need for privacy is sometimes due number of users in different offices and sharing important documents on the network.

Large number of research papers are available in this field, each tackling the problem in different angle using different techniques. Most of the methods result in information loss and side-effects [9]. To overcome the privacy preserving problem for extracting the useful knowledge and address the use of efficient data mining algorithms. To preserve privacy the use of the Commutative RSA cryptographic algorithm is considered. The Association rule sharing model of data mining algorithm is considered for mining in the Key Generation and Key Formulation system [11]. The KDLPPDM system discussed in this paper considers no key exchange to establish the Commutative RSA algorithm which makes it robust even in the presence of adversaries. The remaining paper is organized as Section II gives the literature study related to PPDM. Section III defines the objectives behind the implementation of KDLPPDM. Section IV explains the complete working of our proposed model, along with stepwise workflow. Section V, is the result and discussion section in which the result obtained by our system is studied to check the proper working of model. Finally, Section VI, concluded the paper.

## II. LITERATURE REVIEW

Dehkordi et.al [12] proposed a novel method for privacy preserving association rule mining based on genetic algorithms. It also makes sure that no normal rules are falsely hidden (lost rules) and no extra fake rules (ghost rules) are mistakenly mined after the rule hiding process using genetic algorithm. The algorithm sanitizes both rule and itemset with minimal side effects by introducing new hiding strategies.

S. Vijayarani et. al [13] uses tabu search optimization technique to modify the sensitive items for hiding the sensitive association rules. This approach has the advantage of modifying the sensitive rules accurately without affecting the non-sensitive rules and no false rules are generated. The disadvantage is that it needs several iterations for selecting the optimal transaction for modification. By developing new fitness functions and applying other optimization techniques the number of iterations can be minimized.

Duraisamy et al. [14] proposes an algorithm to minimally modify the database such that no sensitive rules containing sensitive items on the right hand side of the rule will be discovered. The time complexity is reduced because of clustering the sensitive rules and updating database only after all the sensitive rules are hidden. It also modifies minimum numbers of transactions and the alteration in the transactions are stopped when the confidence of the sensitive rules are reduced than the minimum confidence. But the algorithm can hide only sensitive rules with single antecedent and consequent and with the sensitive item in the consequent.

The paper proposed by Assaf Schuster et al. [15] presents a cryptographic privacy-preserving association rule mining algorithm in which all of the cryptographic primitives involve only pairs of participants. The advantage of this algorithm is its scalability and the disadvantage is that, a rule cannot be found correct before the algorithm gathers information from k resources. Thus, candidate generation occurs more slowly and hence the delay in the convergence of the recall. The amount of manager consultation messages is also high.

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 5, May 2016**

## II. OBJECTIVES BEHIND KDLPPDM

- The major objectives of the Privacy preservation Data Mining systems is to maintain the data integrity of the data published and to achieve efficient data mining results.
- For achieving the Privacy preservation for sharing data we have to develop a security mechanism that does not distribute our authentication information i.e. here KDLPPDM.
- The major goal of the collaborative data mining is to work jointly with different users having individual features.
- In the process of data mining the generation of association rule refers towards investigating the inter-relationships among numerous data.
- Association sharing of files should be done along with providing security mechanism, so that only valid user can get valid information all the times.

## III. WORKFLOW OF PROPOSED SYSTEM

The figure 1 below shows the complete working of our proposed and implemented system that works on the platform of privacy preserving data mining i.e. PPDM. The system has two main roles one for User and other for Administrator. As we are focusing on the secure sharing of data here the authorization of any new user is done by the administrator of that organization. The administrator has the authority of granting permission to any user or denying access of any user. Remaining all the facilities are for User of our system. For the security purpose we have perform Authentication using password and Authorization by granting permission of administrator. All registered and authorized user can access their home page containing all the functionality and facility that they can perform. All the user have their allotted space in which they can store or share their files, and makes the data filtered according to the number of users. For finding association we are performing data mining with some techniques, this is explained completely in below figure. Along with this for preserving privacy we have use the Key Generation and key formulation mechanism that satisfies the criteria of PPDM.

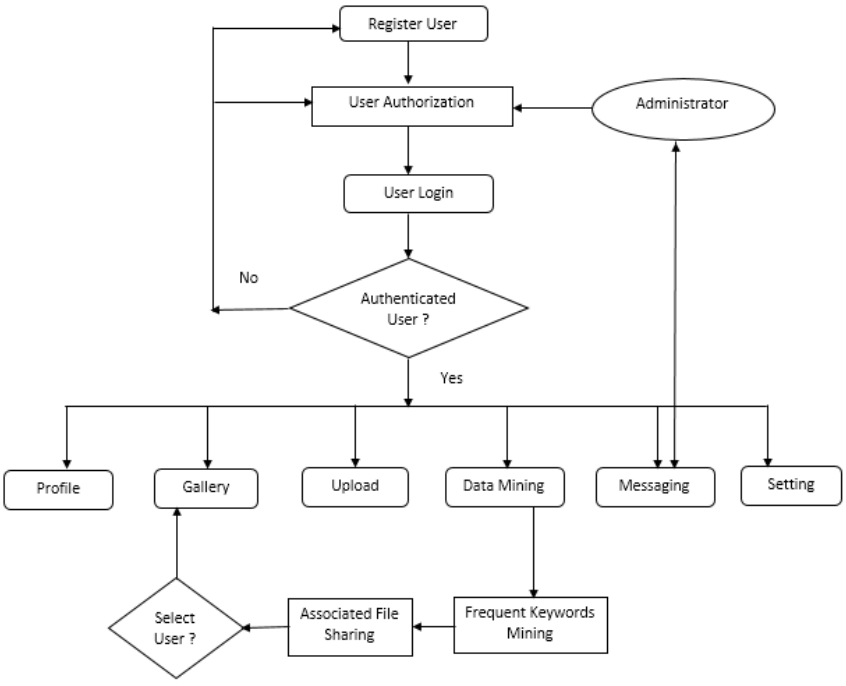


Figure 1: Data Flow Diagram of working System

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 5, May 2016**

- **Working of KDLPPDM**

As stated earlier here, we are performing Data mining for Association sharing of documents. For this, we have develop the model for sharing the associated document in the collaborative manner. The complete working of this association sharing model is shown in the above figure 2 below.

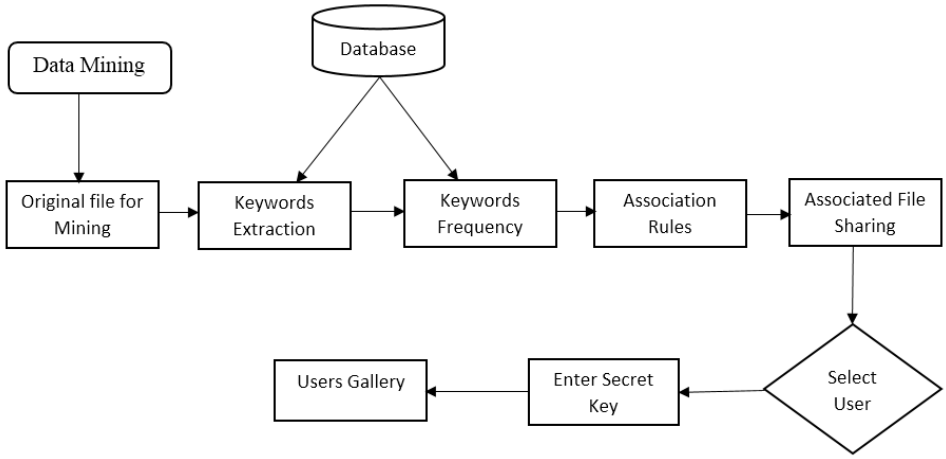


Figure 2: Working of Data mining and Association rule sharing process

The step wise working of the system is as follows:

At the time of user registration, if the registration is successful he can get some space in which he is able to store all his important documents in particular folders according to their file type.
**Step 1:** If anyone from his organization wants some document related to the same domain that you are belongs to, so for finding out association we are using Data mining techniques.

**Step 2:** Here the we choose the file related to same domain and use the technique of frequent Itemsets mining, but here we are mining the frequent keywords

**Step 3:** Now the frequency of the important words related to that domain is find out which calculate the maximum occurrences of the word.

**Step 4:** After that the most associated document according to the maximum occurrence of the keywords is suggested to us from all available documents that are belong to same domain in Association Rules model.

**Step 5:** As we get the most associated document that user wants in his respective domain we have given the functionality of sharing that document to that respective user.

**Step 6:** For security purpose here the sharing of document is not possible until the login user can enter this secret secret key which is assigned to him at the time of his registration.

**Step 7:** When user enter his secret key and share that file to respective user who wants, the file will displayed in the recipients gallery. When he opens that file he is able to read the contents of that file from the same interface.

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 5, May 2016

## IV. RESULT AND DISCUSSION

In this section, we present the experimental results of Association Rule sharing model that is performing the extraction of frequently occurring keywords form files that we have selecting according to category for sharing. After performing the mining of frequent occurrence of keyword the association result is generated that suggest the sharing of files. This framework is more useful for evaluating association sharing when experimenting with large number of files. The output form our information system should accomplish in the following graphs and charts.

### *A. Keyword Frequency Bar Statics*

The First Graph shows the statics of the number of keywords occurring in the files. The X-axis shows the files selected from the same categories and Y-axis shows the Number of keywords occur in that particular file. The above graph shows that currently there are four files of same category is selected and the vertical statistic shows that the second file name abstract has largest occurrence of 147 keyword cloud that is mainly associated with that category and remaining file have different number of occurrences of keyword cloud.

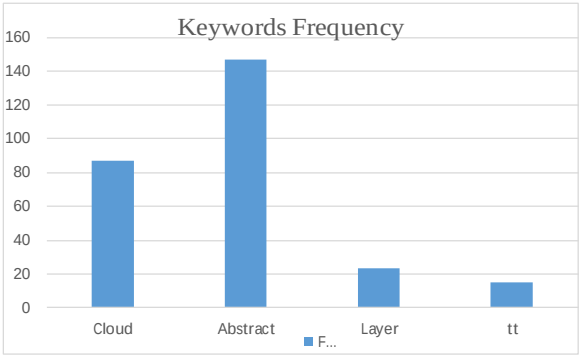


Figure 3: Bar Graph of File showing the frequency of largest occurrence keyword

### *B. Statics of Association result generated*

The Second Line Graph shows the statics related to the file which is largely associated to the file we have passed. This graph shows that the largest occurrence of keyword cloud is present in the file name ‘pp.txt’ this suggest that this file is mostly associated with the file name ‘abstrct.txt’ form the category of cloud. The below graph on the horizontal scale shows that currently there are four file related to the category of cloud and vertical statistic shows that pp.txt file is largely associated that the file selected as form the category of cloud.

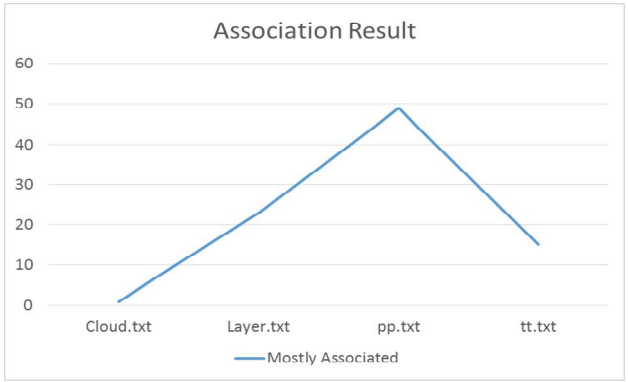


Figure 4: Bar Graph Result generated after finding out association

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 5, May 2016**

## V. CONCLUSION

Here, we considering the problem of privacy preserving collaborative data mining with frequent sharing of associated data among users. In particular, we study how multiple parties to jointly conduct association rule mining on private data and shared file by finding keyword frequency from that original file. We provided an efficient association rule sharing procedure to carry out such a computation. In order to securely collecting necessary statistical measures from data of multiple shared files, we have developed a mechanism for finding important keywords and their frequency. After finding largest frequency we can able to get largest occurrences and share file which is most associated. For providing Privacy to our shared files we have develop the mechanism as Key distribution less privacy preserving (KDLPPDM) model, in which the secret key is provided to the user at the time of its registration process. And for secure sharing of associated documents, it is necessary for all the users to enter their secret key. In this way, we have develop the model of Key distribution less privacy preservation and collaborative filtering for associative sharing of documents. In future, we will extend our method to deal with data sets to find out the association. We will also apply our technique to other data mining computations, such as privacy-preserving clustering. We will try to make extension to sharing of document in the intranet circle, so that document makes more useful for more number of users.

## REFERENCES

[1] S KumaraSwamy, Manjula S H, K R Venugopal, Iyengar S S, L M Patnaik, "Association Rule Sharing Model for Privacy Preservation and Collaborative Data Mining Efficiency", *Proceedings of 2014 RAECS UIET Panjab University Chandigarh*, 06 – 08 March, 2014.
[2] W. Du and Z. Zhan. Building decision tree classifier on private data. In *Workshop on Privacy, Security, and Data Mining at The 2002 IEEE International Conference on Data Mining (ICDM'02)*, Maebashi City, Japan, December 9 2002.
[3] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology - Crypto2000, Lecture Notes in Computer Science*, volume 1880, 2000.
[4] Malik, M.B. Ghazi, M.A. ; Ali, R., "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", *Computer and Communication Technology (ICCCT)*, Third International Conference on 23-25 Nov. 2012
[5] Data mining Articles [Online] available:
http://www.dataminingarticles.com/info/data-mining-introduction/
[6] Association rules (in data mining) [Online]
Available:http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining
[7] Amit Kumar, "A review on Privacy Preservation and Collaborative Data Mining", Council for Innovative Research Peer Review Research Publishing System, *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY*, [Online] available: http://cirworld.org/journals/index.php/ijct/article/view/5217/pdf_642
[9] Chen, K. and L. Liu, 2009. Privacy-preserving multiparty collaborative mining with geometric data perturbation. IEEE Trans. Parallel Distribut. Syst., 20: 1764-1776. DOI: 10.1109/TPDS.2009.26.
[10] www.research.ibm.com/journal/sj/444/niblett.html.
[11] K. Sathiyapriya and Dr. G. Sudha Sadasivam, "SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, Vol.3, No.2, March 2013.
[12] Mohammad Naderi Dehkordi, Kambiz Badie, Ahmad Khadem Zadeh, " A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms", *Journal of software*, vol. 4, no.6, August 2009
[13] S. Vijayarani, A. Tamilarasi, R. SeethaLakshmi, "Tabu Search based Association Rule Hiding", International Journal of Computer Applications 19(1):12-18, April 2011.
[14] Dr. Duraiswamy. K, Dr. Manjula. D, and Maheswari. N "A New Approach to Sensitive Rule Hiding", *ccsenet journal*, vol 1, No. 3, August 2008, 107-111.
[15] Tirumala prasad B, Dr. MHM Krishna Prasad, "Distributed Count Association Rule Mining Algorithm", *International Journal of Computer Trends and Technology*, July to Aug Issue 2011,pp.280-284.