International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 04
February 2016

Secure Web Data Analysis in Vertical Partitioned Database

Priyanka Pandey[1] & Javed Akhtar Khan[2]

[1,2]Department of Computer Science and Engineering, Takshshila Institute of Engineering and Technology, Jabalpur, M.P., India

pandeypriyanka906@gmail.com; javedaktarkhan@takshshila.org

Abstract:

Data mining is to extract or mine data from large amounts of database. However, data is often collected by several different sites. Among these, association rule mining has wide applications to discover interesting relationships among attributes. In many organizations the database may exist in centralized or in distributed environment. In distributed environment, database may be partitioned in different ways such as horizontally partitioned, vertically partitioned and mixed mode. The paper presents privacy preserving data mining algorithms operating over vertically partitioned data. In this paper, we represent the communication among the sites and data transformation between these sites with the help of algorithm in distributed environment.

Keywords— Data Mining; Distributed Database; Privacy preserving protocols; Negative Association rules mining; Web Database.

1. Introduction

The main aim of data mining [1] [2] [3] technology is to explore hidden information from large databases. Many data mining techniques are exist such as association rule mining, clustering, classification and so on are well known and have wide applications in the real world. The issue of privacy arises in two situations namely centralized and distributed environment [4] [5] [6]. In centralized environment, database is available in single location and the multiple users are allowed to access the database .The main aim of privacy preserving data mining in this situation is to perform the mining process by hiding sensitive data/information from users. In distributed environment, the database is available across multiple sites and the main aim of privacy preserving data mining in this environment is to find the global mining results by preserving the individual sites private data/information. Every site can access the global results [7] [8] [9], which are useful for analysis.

1.1 Database

A collection of related data, information and related pieces [10] [11] of data representing /capturing the information about a real-world enterprise or part of an enterprise. An Example University Database: Data about students, faculty, courses, research-laboratories, course registration/enrollment etc.

1.2 Distributed Database

A distributed database is a database in which storage devices [12] [13] are not all attached to a common processing unit such as the CPU, controlled by a distributed database management system (together sometimes called a distributed database system). In distributed database environment, the database among different sites can be partitioned as horizontally, vertically and mixed mode. Many privacy preserving data mining algorithms have been proposed for different partitioning methods in order to find the global mining results by satisfying the privacy constraints. In this paper, privacy preserving association rule mining for n number of vertically partitioned databases at n sites along with data mine, find the actual support when data transmitted between these sites in distributed environment.

2. Proposed Work

Due to the increased demand for knowledge discovery in all industrial domains [14] [15] [16], it is necessary to store all the raw data and to provide useful patterns with respective to the user needs. Generally, the storage of all raw data will be done in a database maintained by concerned organizations. Data mining techniques are available to retrieve useful information from large database. Prediction and description are the two fundamental goals of data mining. To full fill these goals many data mining techniques exists such as association rules, classification, clustering and so on. Among these, association rule has wide applications to discover interesting relationship among attributes in large databases. Association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets.

In this process, we consider the centralized database (Figure1) are divided into distributed database DB1, DB2,…..DBn and their own key values k1,k2,……..kn, or Select N number of sites each having their own database DB1, DB2,…., DBn . Each site calculates their frequent items set and negative support value. The proposed algorithm shows the process of encryption and decryption from all the sites presented in the distributed environments.

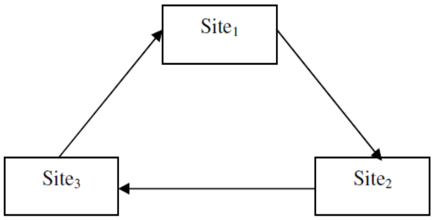


Figure1: Shows the Distributed Architecture

Proposed Algorithm

Encryption Process

Here number of database server is more than two.

Step1: Take the centralized Database (Web Database)
Step2: Convert into the vertically partitioned distributed database (N Number of datasets)
Step3: Calculate the support count of each database.
Step4: Calculate the support and confidence.
Step5: Calculate the negative support and negative confidence.
Step6: calculate partial support and partial confidence.
Step7: Add their own private key in all partial support and partial confidence.
Step8: Divided the partial support, partial confidence and partial lift into the three different values.
Step9: converted partial support and partial confidence values into the ASCII value and compute the matrix Y.
Step10: Take the transpose of the matrix (Y^T).

Step11: convert ASCII code matrix (Y^T) into the binary format.

Step12: consider our own secret key(X matrix)

Step 13: covert the X matrix into binary format

Step 14: perform Exclusive-or between X and Y.

Step15: The resultant matrix is the encrypted format of plain text stored into the associative memory.

Setp16: The resultant matrix is sanded to the protocol initiator Server.

Decryption Process

Step 1: consider the resultant matrix M

Step 2: compute transpose of M matrix as MT matrix

Step 3: convert matrix MT into binary format

Step 4: consider our own private key X

Step 5: covert matrix X into binary number format

Step 6: perform exclusive-or operation between MT and X

Step 7: The resultant matrix is converted to the ASCII code and finally we have the original text.

Step8: After receiving all the original values from the different database, the protocol initiator takes the step for data analysis by calculating Global support and global confidence.

Step9: after that the protocol initiator broadcast the results to all the database server admin presents in the distributed environments.

In this proposed work shall use the apache web server log file of a financial web site user having more than 16 million hits in a month’s time [11]. A web server log contains the following information: Number of Hits, Number of Visitors, Visitor Referring Website, Visitor Referral Website, Time and Duration, Path Analysis, Visitor IP Address, Browser Type and Cookies. Data of a typical web server is shown in following sample data of the first raw of the log file:

```
10.32.1.43 - - [10/Nov/2013:00:07:00] "GET /flower_store/product.screen?product_id=FL-DLH-02 HTTP/1.1" 200 10901 "http://mystore.splunk.com/flower_store/category.screen?category_id=GIFTS&JSESSIONID=SD7SL1FF9ADFF2" "Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.0.10) Gecko/20070223 CentOS/1.5.0.10-0.1.el4.centos Firefox/1.5.0.10" 4361 3217
```

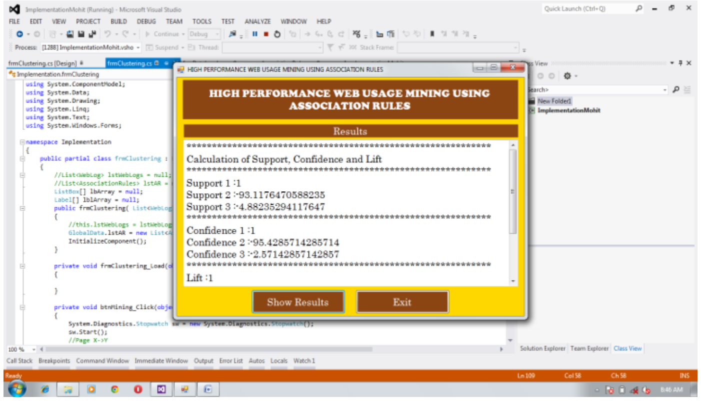
This work shall use the basic steps of the web usage mining [12] with modifications to improve the performance and accuracy of the extracted information. For improvements, work will be done in following modified steps [13]:

1. Data Collection – From a server log of a website
2. Data Pre-processing – Formatting data from the server log obtained. The server log contains huge information in each record which is separated and stored in lists with rows and columns. Some redundant information is removed and data rows are ordered on the basis of the date and time of hit.
3. Pattern Discovery – In this step I will apply the following (but not to max) association rules:
   I. If Webpage X is hit then Webpage Y is also hit by the user
   II. If Webpage X is hit then user moves away from the site by clicking an external link
   III. If a user comes on site once then he visits it again
4. Pattern Analysis – From the association rules applied in step 3 data shall be analysed to find the support and

confidence for each rule and filtering of records shall be done on basis of low support or confidence. Knowledge will be produced for the web admin in readable format for future use.

5. Result Generation – Each phase will be measured for time taken and time complexity to show that the performance of the proposed system is high. Comparatives shall also be produced to show the accuracy of the produced patterns in respect of the base work.

Proposed work will be implemented using C#.NET Windows and Server Log Dataset Extracted from a live server from a financial website. The work will be done in following steps: first step is Interface Development, in which C#.net windows/JAVA Swing forms will be used to develop the interface. Various relevant navigational buttons will be included in it which allows traversing with validations. Second step is Loading Server Log Data Set in which, Server Log dataset is a collection of server messages generated when a server starts and stops. The details such as login, logout, system shutdown, errors and other information's will be stored in it by the server machine during the normal course of operations on the server. These will be used for clustering of message using association rules. Third step is Data Pre-processing, in data pre processing file handling will be used for loading the server log file in project and will be applied data pre-processing on the loaded data. In data pre-processing stopping and stemming will be applied to filter data for important words. Stopping will be used to remove all special characters, remove any words with length less than or equal to 3 characters and removing any words which are prepositions. Stemming will be applied to remove words with having similar meaning and different tense of English. Fourth step is negative Association Rule Mining, in this phase; various characteristics of the server log file will be used for creating negative association rules. These rules shall be applied to create clusters. And last phase analysis phase in which we analyze our results with the help of graphs; Results will be calculated by using the clusters and formulas for calculation of performance and accuracy of the system. Graphs will be drawn using the various results as calculated for all accuracy, precision, recall and specificity. The proposed algorithm of this paper is shows the steps, how we are analyzing our web data. And figure2, shows the process of loading of web data, process of data cleaning after the loading of web data is done, data analyzing after applying the negative association rule mining, data analysis and the final output of our process, in which its shows the negative support count, negative confidence and negative lift or importance.


ImplementationMohit (Running) - Microsoft Visual Studio
FILE EDIT VIEW PROJECT BUILD DEBUG TEAM TOOLS TEST ANALYZE WINDOW HELP
Continue Debug
Process: [1288] ImplementationMohit.vsho
Suspend Thread: Stack Frame:
frmClustering.cs [Design]
frmClustering.cs
Implementation.frmClustering
100 %
Call Stack Breakpoints Command Window Immediate Window Output Error List Autos Locals Watch 1
Ready
Ln 109
Solution Explorer Team Explorer Class View
New Folder1
ImplementationMohit

```
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Windows.Forms;

namespace Implementation
{
    public partial class frmClustering :
    {
        //List<WebLog> lstWebLogs = null;
        //List<AssociationRules> lstAR =
        ListBox[] lbArray = null;
        Label[] lblArray = null;
        public frmClustering( List<WebLog
        {
            //this.lstWebLogs = lstWebLog
            GlobalData.lstAR = new List<A
            InitializeComponent();
        }

        private void frmClustering_Load(o
        {

        }

        private void btnMining_Click(obje
        {
            System.Diagnostics.Stopwatch sw = new System.Diagnostics.Stopwatch();
            sw.Start();
            //Page X->Y
```

HIGH PERFORMANCE WEB USAGE MINING USING ASSOCIATION RULES

HIGH PERFORMANCE WEB USAGE MINING USING ASSOCIATION RULES

Results

```
**************************************************************
Calculation of Support, Confidence and Lift
**************************************************************
Support 1 :1
Support 2 :-93.1176470588235
Support 3 :-4.88235294117647
**************************************************************
Confidence 1 :1
Confidence 2 :-95.4285714285714
Confidence 3 :-2.57142857142857
**************************************************************
Lift :1
```

Figure 2: Shows the final output of our process

Knowledge has been produced for the web admin in readable format for future use.

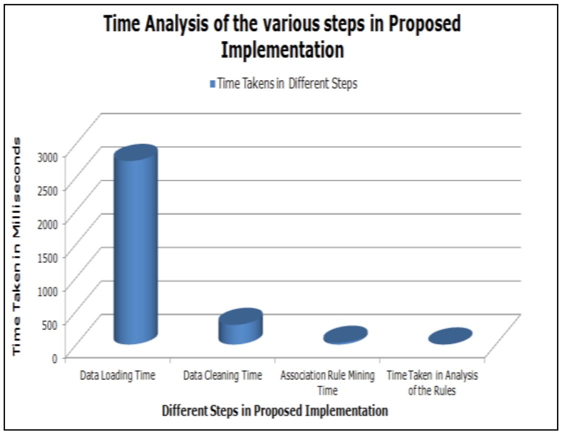


Figure 3: Time Analysis of various steps in Proposed Implementation

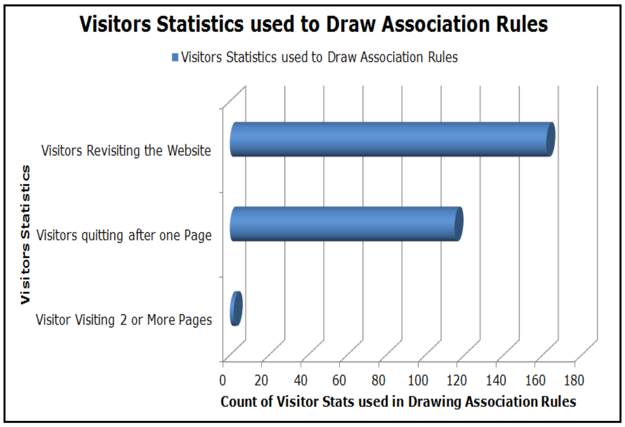


Figure 4: Visitors Statistics used to Draw Association Rules

Inference: The above graph is a clear indicative of the efficiency of the proposed system. The maximum time has been taken in loading and pre-processing of data. Time taken in drawing association rules is very less than the above two times. Inference: The above graph shows the visitor statistics for all the visitors which are grouped on the basis of the association rules applied. These stats are also in conformance with the value of confidence which is indicative that first rule in not providing very confident results.

Conclusion

The difficulty of preserving privacy in association rule mining when the database is distributed vertically partitioned database among n (n>2) number of sites when no trusted party is considered. In this paper, cryptography algorithm by using the key values and matrix transpose is adopted to enhance the privacy further. The proposed replica capably to find global frequent item sets even when no site can be treated as trusted. The trusted party initiates the process and prepares the merged list. All the sites computes the partial supports and total supports for all the item sets in the merged list using the cryptography technique and based on these results finally trusted party finds actual frequent item sets. And after comparing the result of these, we find output that data leakage with trusted party is more as compare to without trusted party so privacy also without trusted party is more as compare to with trusted party. Future task is to use all the privacy preserving technique and compare them according to their complexity and reduce the complexity, with zero percentage of data leakage.

References

[1] Ming-Syan Chen, Jiawei Han,Yu, P.S. (1996), Data mining: an overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering, Vol. 8 No. 6, pp 866 –883.

[2] A.C Yao(1986), How to generate and exchange secrets, In proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pp 162-167.

[3] Y Lindell and B pinkas (2000), Privacy preserving data mining, In Proc. O CRYPTO'00, pp36-54. Springer-Verlag2000. [4] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin,

and Michael Y.Zhu (2003), Tools for privacy preserving distributed data mining, SIGKDD Explorations, Vol. 4, No.2 pp 1-7.

[5] M. Kantarcioglu and C. Clifto (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. In IEEE Transactions on Knowledge and Data Engineering Journal, volume 16(9), pp. 1026-1037.

[6] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. (2004),State-of- the-art in privacy preserving data mining, SIGMOD Record, 33(1):50–57.

[7] Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza (2005), A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Data Mining and Knowledge Discovery, Vol. 11, 121–154.

[8] Chin-Chen Chang, Jieh-Shan Yeh, and Yu-Chiang Li (2006), Privacy-Preserving Mining of Association Rules on Distributed Databases, IJCSNS International Journal of Computer Science and Network Security, Vol.6 No.11.

[9] Alex Gurevich, Ehud Gudes (2006), Privacy preserving data mining algorithms without the use of secure computation or perturbation, 10th international database Engineering and Applications Symposium IDEAS06 IEEE.

[10] Mahmoud Hussein, Ashraf El-Sisi, and Nabil Ismail (2008), Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base, I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II,LNAI 5178, pp. 607–616, 2008.© Springer-Verlag Berlin Heidelberg.

[11] Jian Wang, Yongcheng Luo, Yan Zhao, Jiajin Le(2009), A Survey on Privacy Preserving Data Mining, First International Workshop on Database Technology and Applications, pp. 111-114.

[12] Josh Cohen Benaloh. Secret sharing homomorphisms: Keeping shares of a secret. In A.M. Odlyzko, editor, Advances in Cryptography - CRYPTO86: Proceedings, volume 263, pages 251–260. Springer-Verlag, Lecture Notes in Computer Science, 1986.

[13] C. L. Blake and C. J. Merz. UCI repository of machine learning databases,1998.

[14] M. Blum and S. Goldwasser. An efficient probabilistic public-key encryption that hides all partial information. In R. Blakely, editor, Advances in Cryptology {Crypto 84 Proceedings. Springer-Verlag, 1984.

[15] Paul S. Bradley and Usama M. Fayyad. Re_ning initial points for K-Means clustering. In Proceedings of the Fifteenth International Conference on Machine Learning, pages 91{99. Morgan Kaufmann, San Francisco, CA, 1998.

[16] Christian Cachin. Efficient private bidding and auctions with an oblivious third party. In Proceedings of the Sixth ACM conference on Computer and communications security, pages 120–127, Kent Ridge Digital Labs, Singapore, 1999.ACM Press.

[17] Philip Chan. An Extensible Meta-Learning Approach for Scalable and Accurate Inductive Learning. PhD thesis, Department of Computer Science, Columbia University, New York, NY, 1996.