International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 2, Issue 5 (Sep-Oct 2014), PP. 34-36

DEVELOPMENT OF A FRAMEWORK FOR PRESERVING PRIVATE DATA IN WEB DATA MINING

Sabica Ahmad[1], Shish Ahmad[2], Jameel Ahmad[3]
Dept. CSE & IT
Integral University
Lucknow, U.P.
[1]sabicaahmad@gmail.com, [2]shish_parv@rediffmail.com, [3]jameel_integral@rediffmail.com

Abstract- The main aspire of this research work is, to develop proficient methodology to find privacy preserving association rule mining in centralized environment without infringement of any privacy constraints. The issue of privacy constraints for centralized database environment is entirely different from distributed database environment. The goal of attaining privacy in centralized database environment is, to obtain a distorted database which hides the sensitive item sets. When mining task is performed on distorted database all the sensitive rules should be hidden without any side effects. Based on heuristic approach, a new me-thodology is proposed by incorporating suggested Criteria1 and Criteria2 to identify the victim item and selecting suitable supporting transactions efficiently for sanitization purpose to hide the sensitive item sets.

Index Terms — preserving private data, frequent item sets, privacy preserving association rule mining.

I. Introduction

Data mining has been view edasa risk to privacy because of the widespread propagation of electronic data maintained by organizations. This has initiated augmented concerns about the privacy of the under-lying data .The matter of privacy plays a crucial role when several genuine people share their resources in order to obtain mutual profit but no one is interested to reveal their private data .In the process of data mining, how to determine the problem of privacy preserving has become a hot research topic in the field of data mining. Hence, privacy preserving data mining research area is evolved.

The privacy preservation data mining algorithms are generally classified into three categories namely reconstruction based, heuristic based and cryptog-raphy based

II. PRIVACY PRESERVING ASSOCIA-TION RULE MINING

We consider a method for finding privacy pre-serving association rule mining based on heuris-tic approach in centralized environment for dis-covering solution for hiding sensitive rules by fulfilling association rule hiding goals accurately or approximately.

A new method is proposed in this paper re-lated to heuristic approach to hide sensitive association rules specified by users with min-imum side effects.

The Criteria1 specifies the competent selection of victim item and Criteria2 helps to find the appropriate supporting transactions for victim item in the sanitization process to minimize side effects.

Criteria 1:

Victim item can be selected based on the follow-ing condition.

If number of times <Ai> appears in non sensitive frequent item set is greater than number of times <Aj> appears in non sensitive frequent item sets then Aj be the victim item. If number of times <Ai> appears in non sensitive frequent item set is less than number of times <Aj> appears in non sensitive frequent item sets then Ai be the victim item.

Criteria 2:

The minimum number of transactions required to hide item set is based on the value of <Ai,Aj>.supp – MinTrans +1.. For each support-ing transactions for item set <Ai,Aj>, weight is computed by using the following:

W(Tg) = No. of dependant items with victim item - number of infrequent item sets associated with victim item.

III. Proposed Framework

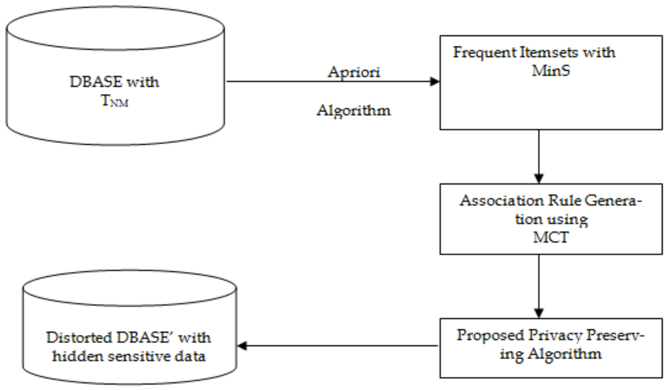
In this paper a procedure is suggested in which all the sensitive item sets whose length is greater than two are considered to find the pairs of sub patterns. From this pair only significant pair-sub patterns are considered as sensitive to hide sensitive patterns. This procedure is very significant in a way that it avoids the difficulty of forward inference attack. In order to avoid forward inference attack problem, at least one such sub pattern with length of two of the patterns should be hidden. The split pattern procedure helps to accelerate up the hiding process.

S.No.	**Symbols**	**Explanatio**
1	DBASE = $\{t_1,t_2,..t_N\}$	A original database consisting of N number of transactions
2	I =$\{i_1,i_2,\dots i_M\}$	An item set of length M
3	L_k	An item set of length k
4	T_{nm}	The nth transaction of mth item
5	S =$\{ s_1, s_2, \dots s_r\}$	Set of sensitive item sets
6	MinS	User specified Minimum support threshold
7	Supp(J)	Number of transactions supporting item set J
8	MinTrans	Based on MinS, number of transactions required to support an item set to be frequent
9	MCT	User specified Minimum confidence threshold
10	N	Size of original database, DBASE
11	F_{DBASE} =$\{L_1, L_2, L_3,\dots L_k\}$	A set consists of all frequent item sets
12	A □ B	Association rule between item sets A and B
13	F_S	The set consisting of sensitive item sets
14	F_{NS}	The Set consisting of non sensitive frequent item sets

15	F_{2S}	The set consisting of pairs determined by the procedure split pattern.
16	$<A_i,A_j>$	The sensitive item set pair
17	T_{AiAj}	Set of supporting transactions for item set $<A_i,A_j>$
18	DBASE'	Distorted database which hides all sensitive item sets.
19	Victim item	An item which is selected from the sensitive item pair which produces least side effects or no side effects when modification is done over it.
20	Victim transactions	Selected transactions to modify the victim item value.
21	MinT	A set consisting of suitable number transactions, which are to be modified to hide the sensitive item set
22	Count	Count gives number of times the victim item value has to be modified to hide sensitive item set pair.
23	$W(T_g)$	Weight for transaction T_g

Table 3.1: Symbols Used in Proposed Model

The proposed framework is depicted as below:



IV. ALGORITHM

The algorithm for the proposed model is as fol-lows:

Step 1 For a given database DBASE and set of sen-sitive item sets Fs, generate frequent item sets and store with their support values in FDBASE.

Step 2 Let the sensitive item sets are stored in Fs then the non sensitive frequent item sets are obtained by subtracting FS from FDBASE.

i.e., FNS = FDBASE - FS.

Step 3 If any item sets in FS are having more than length of two, call the procedure split pat-tern to identify the prominent pairs which are to be hidden in order to hide all the item sets whose length is greater than two.

Step 4 After step 3 a vector F2S is prepared which consists of all two pair sensitive items.

Step 5 The generated all pairs sensitive fre-quent item sets with their support values along with their supporting transactions ID's are stored in a Table TS.

Step 6 All the non sensitive frequent item sets that is F- F2S are stored along with their support values in a Table TNS.

Step 7 For each item set in F2S

If any non overlapping item set exists go to step 12.

Else the patterns <Ai,Aj><Aj,Ak> are chosen

Consider the victim item (Ai or Aj) based on the Criteria1

Step 8 Find the intersection of supporting transactions for AiAj and AjAk as follows:

TAiAjAk = TAiAj □ TAjAk

Step 9 Obtain the value for Count1 and Count2 as follows:
Count1 for AiAj = <Ai,Aj>.Supp - MinTrans + 1
Count2 for AjAk = <Aj,Ak>.Supp - MinTrans + 1

Step 10 Find minimum number of supporting transactions to be modified by applying Crite-ria2. Select smaller one from both Count1 and Count2 and many transactions are chosen from MinT and the victim item (Aj) values are re-placed with 0 values. By this, item set lower count value will be hidden. To hide the item set, which is having higher count value, Count1 – Count2 no of transactions which are not yet processed will be chosen from MinT for the process of sanitization. To protect this item set, the victim item set can be chosen based on their dependencies with the item sets in non sensitive item set FNS. Accordingly the victim item value will be replaced with zero in the selected trans-actions. After performing this, the item set which is having higher count value is also hidden.

Step 11 Modify F2S by removing the pairs <Ai, Aj> and <AjAk> from it. Go to step18.

Step 12 For the sensitive item set pair <Ai, Aj> in F2S find victim item by using criteria 1.

Step 13 After identifying the victim item, find the supporting transactions for <Ai, Aj>.

Step14 Obtain the value for Count1 and Count2 as follows:
Count1 for AiAj = <Ai, Aj>.Supp - MinTrans + 1

Step15 Select Count1 no of transactions to be modified from a set MinT obtained by the Crite-ria2.

Step 16 The value of victim item in the selected transactions is replaced with value zero.

Step 17 Update F2S by removing <Ai, Aj> from it.

Step 18 Repeat the above steps from step 7 until no more pair in the F2S to hide.

Step 19 Finally distorted database, DBASE´ is obtained in which all sensitive item sets in F2S are hidden.

Step 20 Stop the process.

V. CONCLUSION

This study has been carried out to develop method-ology in centralized as well as in distributed envi-ronment to find privacy preserving association rule mining without revealing any private data or infor-mation.

This methodology is proposed in this thesis work to hide the sensitive item sets in centralized database environment. My methodology is related to heuris-tic based approach which utilizes suggested criteria to efficiently find the victim item and its supporting transactions. The proposed methodology efficiently performs sanitization process.

REFERENCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, pp. 37-54,1996.

[2] J. Han and M.Kamber, Data Mining Con-cepts and Techniques, Elsevier 2001.

[3] R. Agrawal and R. Srikant, "Mining Sequential patterns", Proc.1995 International Conference on Data Engineering (ICDE"95), pp 3-14, Taipei, Taiwan, March 1995.

[4] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. "State-of-the-art in privacy

preserving data mining". SIGMOD Record, 33(1):50–57,2004.

[5] Ahmed HajYasien, "Preserving Privacy In Association Rule Mining", Ph D.,thesis, Griffith University, June 2007.

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

[7] Yongjian Fu, "Data mining: Tasks, tech-niques and Applications", Department of Computer Science", University of Missouri- Rolla,1997

[8] Michael Goebel, Le Gruenwald, "A Survey Of Data Mining And Knowledge Discovery Software Tools", SIGKDD Explorations, ACM SIGKDD, Vol: 1, Issue 1, pp 20- 33, June 1999.

[9] Thair Nu Phyu ,"Survey of Classifica-tion Techniques in Data Mining", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009, Vol I,IMECS-2009, Hong Kong, 2009.

[10] Yongjian Fu, Distributed data mining: Overview, University of Missouri- Rolla, 2001.

[11] R Agarwal, T Imielinski and A Swamy, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, page 207-210, 1993.

[12] R. Agrawal and R. Srikant. "Fast, algo-rithms for mining association rules in large data-bases", Proceedings of the 20th VLDB Conference Santiago, Chile, pp 487-499, 1994.

[13] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", Proc. 21st VLDB Conference, Zurich, Swizerland., 1995.

[14] Mohammed J. Zaki, "Parallel and Distrib-uted Data Mining: An Introduction", Large-Scale Parallel Data Mining Lecture Notes In Computer Science, Vol. 1759, 2000.

[15] Qinghua Zou, esley hu, Johnson,Chiu, . A pattern decomposition (PD) algorithm for finding all frequent patterns in large datasets", International Conference on Data Mining, ICDM 2001, Proceedings IEEE, 673 – 674, 2001