Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2
© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PATTERN MINING TECHNIQUES FOR PRIVACY PRESERVING TRANSACTIONAL DATASETS USING ATTRIBUTE IMPACT MATRIX

VIJAYKUMAR, DR.T.CHRISTOPHER
Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India.
Head, Assistant professor of computer science, Government Arts College, Udumalpet, Tamil Nadu, India.
Email: Vijayakumar123phd@gmail.com

ABSTRACT

Privacy preservation being studied well in recent days of data publishing where the sensitive items have to be hidden without spoiling the originality of data. We propose a new approach for privacy preservation of data items while publishing transactional data sets. The proposed approach generates transactional patterns using input data set, for each item set I to N , the proposed approach generates pattern set Ps. From generated pattern set ps, we compute support and profit values to select most frequent pattern set Mps. Identified pattern set is used to compute the attribute impact matrix , which represent the sanitized data set where the sensitive items are represented with impact measure of different number of item sets. Published data could be used to infer some knowledge but they could not back track the personal information from the sanitized data base. The proposed method is a simpler one which reduces the time and space complexity.

Key Terms: *Privacy Preservation, Knowledge Hiding, Transactional Data Set, Attribute Impact Matrix.*

1. INTRODUCTION

The most business intelligence is extracted from large transactional data sets where the user information and other purchase details and other useful information's are used to extract business intelligence. The transactional data sets are nothing but the information about the purchase history of many users and the size of data set and point has no boundary. For example , a customer shops at a super market and a log can be generated about his purchase like 'name', 'address','mobile','food items','cosmetics'and etc.. This kind of logs could be generated and maintained by the organizations which will be shared by many business organization. Suppose if an marketing company looks for a exact location to market his product then they can use the transactional data set to identify the location where his product has focus or he can mine any information about his product or his competitor. The transactional data set contains many useful customer information and used for many business processes.

Privacy preservation is the process of hiding customer information like name, mobile no and addresses from any third party. The customer information has to be kept secret with the organization and the organization has no right to give it to others which leads to information leak. Sometimes, the transactional data set may contains information about some purchase details about a hot product like 'Viagra' and 'Pregnancy Test Card'. These information has to be avoided while giving transactional data set to others , because the end user can easily identify who purchased the other products using these details.

Sometimes few product manufacturers would like to offer another product of different manufacturer but may oscillate in selection of a product. In that situation they need a complete transactional set. But the data owner will not be ready to expose the complete data where there are personal information about their customers. This is where sanitization becomes necessary, so that the privacy preservation has to be done on the transactional data set before handover the data set to other organization.

Knowledge hiding is the process of hiding sensitive information from large data set without tampering the originality of the data set. While sanitizing the originality of the data set has to be retained and also should be useful for other to make some inference from the data set published.

Attribute impact measure is the value of an attribute which participates in all item sets. For example an attribute or item may present or selected in some n item set and may not get selected in others. The attribute impact measure represent the deapthness of an item at all item sets. This also shows the importance of attribute or item which has to be hidden while publishing. Consider, for instance, the example of a large retail company which sells thousands of different products, and has numerous daily purchase transactions. The large amount of transactional data may contain customer spending patterns and trends that are essential for marketing and planning purposes.

	Soap	Toothpaste	Horlicks	Cream	Pregnancy Test	Viagra
Arun	×		×			
Siva	×		×			x
Rathna		×		×	×	
Kumar		×	×			
Raja	×		×	×		x

Fig1: Original Purchase Data Set

The table 1, shows the transaction set of a purchase data set where there exist number of records of purchase done by various customers. From the table, the items pregnancy test and Viagra are the most sensitive and private items which has to be hidden while publishing the data set. If the organization exposes the private information to other end user then the trustworthy of the organization will become questionable. So that the organizations have to release the data set without disclosing this information also with originality. To overcome this there must be a suitable method which should not disclose the privacy items and personal information, so that an attacker could not be able to identify or infer others purchase and so on. Also in case of business point of view the people's identity should not be visible; otherwise the product goal will not reach all the people.

2. RELATED WORKS

There are many approaches has been discussed in the literature to preserve the personal information of users. We discuss few of them here for understanding the problem statement.

An Efficient Method for Knowledge Hiding Through Database Extension [2], propose a new solution by integrating the advantages of both these techniques with the view of minimizing information loss and privacy loss. By making use of cryptographic techniques to store sensitive data and providing access to the stored data based on an individual's role, we ensure that the data is safe from privacy breaches. The trade-off between data utility and data safety of our proposed method will be assessed.

A generalized Framework of Privacy Preservation in Distributed Data mining for [4] Unstructured Data Environment [4], proposes a solution to this problem by managing unstructured data in to structured data using legacy system and distributed data partitioned method for gives distributed data for mining multi text documents. This frame work gives the testing of the similarities among text documents and privacy preserving meta data hiding technique, which are explored in text mining.

A Fuzzy Approach for Privacy Preserving in Data Mining [5], addresses the problem of Privacy Preserving in Data Mining by transforming the attributes to fuzzy attributes. Due to fuzzification, exact value cannot be predicted thus maintaining individual privacy, and also better accuracy of mining results were achieved.

Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Multiple R. H. S. Items [6], propose two algorithms, ADSRRC (Advanced Decrease Support of R. H. S. items of Rule Cluster) and RRLR (Remove and Reinsert L. H. S. of Rule), for hiding sensitive association rules. Both algorithms are developed to overcome limitations of existing rule hiding algorithm DSRRC (Decrease Support of R. H. S. items of Rule Cluster). Algorithm ADSRRC overcomes limitation of multiple sorting in database as well as it selects transaction to be modified based on different criteria than DSRRC algorithm. Algorithm RRLR overcomes limitation of hiding rules having multiple R. H. S. items. Experimental results show that both proposed algorithms outperform DSRRC in terms of side effects generated and data quality in most cases.

Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining [7], present a novel hiding-missing-artificial utility (HMAU) algorithm is proposed to hide sensitive itemsets through transaction deletion. The transaction with the maximal ratio of sensitive to nonsensitive one is thus selected to be entirely deleted. Three side effects of hiding failures, missing itemsets, and artificial itemsets are considered to evaluate whether the transactions are required to be deleted for hiding sensitive itemsets. Three weights are also assigned as the importance to three factors, which can be set according to the requirement of users.

Hiding Sensitive Association Rules without Altering the Support of Sensitive Item [9], uses the data distortion technique where the position of the sensitive items is altered but its support is never changed. The size of the database remains the same. It uses the idea of representative rules to prune the rules first and then hides the sensitive rules. Advantage of this approach is that it hides maximum number of rules however, the existing approaches fail to hide all the desired rules, which are supposed to be hidden in minimum number of passes.

An Efficient Algorithm for Sequential Pattern Mining with Privacy Preservation [10], presents an index-based algorithm named SSAPP for exploring frequent sequential patterns in a distributed environment with privacy preservation. The SSAPP algorithm uses an equivalent form of a sequential pattern to reduce the number of cryptographic operations, such as decryption and encryption. In order to improve the efficiency of sequential pattern mining, the SSAPP algorithm keeps track of patterns in a tree data structure called SS-Tree. This tree is used to compress and represent sequences from a sequence database. Moreover, a SS-Tree allows one to obtain frequent sequential patterns without generation of candidate sequences.

A Novel Community Detection Algorithm for Privacy Preservation in Social Networks [11], presents a novel method for community detection with the assumption of privacy preservation is proposed. In the proposed approach is like hierarchical clustering, nodes are divided alliteratively based on learning automata (LA). A set of LA can find min-cut of a graph as two communities for each iteration. Simulation results on standard datasets of social network have revealed a relative improvement in comparison with alternative methods.

Identity-Based Privacy Preservation Framework over u-Healthcare System [12], proposes an identity-based privacy preservation framework over u-healthcare systems. Our framework is based on the concepts of identity-based cryptography and non-interactive key agreement scheme using bilinear pairing. The proposed framework achieves authentication, patient anonymity, un-traceability, patient data privacy and session key secrecy, and resistance against impersonation and replay attacks.

All the above methods have the problem of anonymity and suffers with mining original information from sanitized data. We propose a new sanitization approach for privacy preservation using attribute impact matrix which explained earlier in this chapter.

3. PROPOSED METHOD

The proposed approach generates patterns and selects the most frequent and effective patterns using support and confident values. From the selected patterns a set of items which have more impact on all item set are identified using attribute impact measure. The final sanitized data set is generated using the attribute impact factor computed.

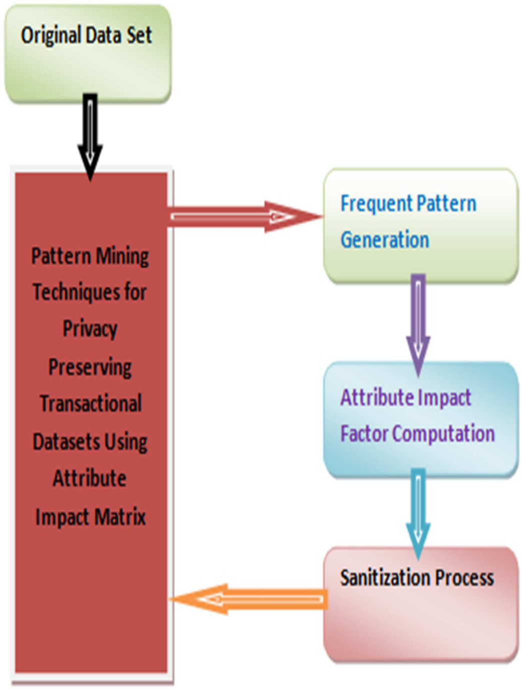

Original Data Set
Pattern Mining Techniques for Privacy Preserving Transactional Datasets Using Attribute Impact Matrix
Frequent Pattern Generation
Attribute Impact Factor Computation
Sanitization Process

Figure2: Proposed method architecture.

3.1 Frequent Pattern Generation:

The frequent pattern is generated on a given transactional data set T_s , where N specifies the number of attributes and TN specifies the total number of transactions available. Initially the number of attributes which forms the whole transaction set is identified and we generate combinatory of patterns set Ps. The combinatory of pattern set is computed according to the possible combinations which can be formed. Unlike generalized frequent pattern mining algorithm the support and count values are computed for the identified patterns in the pattern set P_s. Using computed support and count values the pattern P_i which has support value greater than support threshold S_t will be selected for sanitization process and other will be omitted.

Algorithm:

```
Input: Data set Ts.
Output: Pattern set OPs.
step1: Compute Total number of transactions Tn =$\sum Ts$
Step2: Identify set of attributes Ats = $\int_1^{Tn} \sum Attr \nexists Ats$
Step3: Compute possible patterns Ps = $\int_1^{N} \forall(Tsi) \nexists Ps$
Step6: for each pattern $P_i$ from $P_s$
Compute count = ø($P_i$£($P_s$)).
        Ø- Number of pattern $p_i$ contained in $p_s$.
        Compute support = count/TN.
        If support > ST then
        Add $P_i$ to Output pattern set Ops = $\sum Ps(i) \cup \int Psj(support > ST)$
        End.
        End
Step7: stop
```

3.2 Attribute Impact Matrix Computation:

The attribute impact matrix specifies the importance of the attribute or how well the attribute participated in all item sets computed. We compute the impact factor of every item of transaction set at each level of item sets from 1 to n, where n is the maximum number of item set can be formed. The attribute impact factor is formed using total number of transactions the item appeared and how many number of transactions it present in particular item set pattern. By computing all impact factors a generic value will be computed to represent the impact value of the item at all the levels.

Algorithm:

```
Input: Transaction Data set Ts, Pattern set Ops.
Output: Attribute impact matrix IMPF, CIV.
Step1: initialize IMPF set
Step2: compute total number of transactions Tn.
Step3: for each item $I_i$ from attribute set Ats
for each item set count Is
        compute number of occurrences Noc = $\int_1^{Tn} \sum(Ti \in Ii)$
        compute impact factor IMPF(i,j) = $\frac{Noc}{Tn}$
        end
        end
Step4: for each item I from Attribute set Ats
        compute cumulative impact value $CIV_{(i)} = \frac{\int_1^{size(Ats)} \sum IMPF(i,j)}{size(Ats)}$
        end
Step5: stop.
```

3.3 Sanitization Process:

The computed cumulative impact value and impact factor matrix is used to generate sanitized data set. From available impact value we identify set of attributes which has low impact value is identified which represents the sensitive item and has to be hidden. Once the item which is sensitive is identified then the value at the sensitive item is represented with a set of impact factor values and used for data publishing. This shows the data set as original and the end user can infer required knowledge from the sanitized data set.

Algorithm:

```
Input: Transaction data set Ts, IMPF, CIV
Output: Sanitized data set STs.
step1: for each attribute $A_i$ from Ts
        if CIV of Ai < CIT then //CIT- Cumulative impact threshold
                Ts(Ai) = $(\forall(IMPF(Ai))$
        end
        end.
Step2: stop.
```

4. RESULTS AND DISCUSSION

The proposed method has produced efficient results in sanitization and produced good results. The proposed method generates frequent pattern and from the pattern generated cumulative impact value and impact factor is computed to identify sensitive item. The proposed method retained the originality and also it preserves the private information. At this stage even if we declare the names with the sanitized data set the user cannot identify which record belongs to one and it is not possible to identify the purchase pattern of the user.

Table1: Shows The Original Data Set

Names	Soap	Tooth Paste	Horlicks	Cream	Pregnancy Test
Siva	1	1	1	1	0
radha	0	1	1	1	1
Ram	1	1	0	1	0
Rajes	1	1	1	1	0
Saran	1	1	0	1	1
Kumar	1	1	1	1	0
Shela	1	1	0	0	0
Selva	1	1	0	0	0
Sivasankar	1	1	0	0	0
rohini	0	1	1	0	1

Table3: Shows The Support Count Values For Different Pattern.

Pattern Type					Count/Support
Soap	Tooth Paste	Horlicks	Cream	Preg. Test	0/0.0
Soap	Tooth Paste				7/0.7
Soap	Tooth Paste	Horlicks			3/0.3
Soap	Tooth Paste	Horlicks	Cream		3/0.3
Soap	P.T	Horlicks	Cream		0/0.0
Soap	Tooth Paste	Horlicks	P.T		0/0.0
Soap	Tooth Paste	P.T	Cream		1/0.1
Tooth Paste	Horlicks	Cream	Preg. Test		1/0.1
Tooth Paste	Horlicks				5/0.5
Tooth Paste	Horlicks	Cream			5/0.4
Horlicks	Cream	Preg. Test			1/0.1
Horlicks	Cream				4/0.4
Cream	Preg. Test				2/0.4
Soap	Horlicks				3/0.3
Soap	Cream				5/0.5
Soap	P.T				1/0.1
Tooth Paste	Cream				6/0.6
Tooth Paste	P.T				3/0.3

Horlicks	Soap				3/0.3
Horlicks	Cream				4/0.4
Horlicks	P.T				2/0.4

Table4: List Of Pattern Set With Count And Support Which Is Greater Than Threshold

Pattern Type					Count/Support
Soap	Tooth Paste				7/0.7
Soap	Tooth Paste	Horlicks			3/0.3
Soap	Tooth Paste	Horlicks	Cream		3/0.3
Tooth Paste	Horlicks				5/0.5
Tooth Paste	Horlicks	Cream			5/0.4
Horlicks	Cream				4/0.4
Soap	Horlicks				3/0.3
Soap	Cream				5/0.5
Tooth Paste	Cream				6/0.6
Tooth Paste	P.T				3/0.3
Horlicks	Soap				3/0.3
Horlicks	Cream				4/0.4

Table 5: Shows The Computed Cumulative Impact Value And Impact Factor Of Different Item Set.

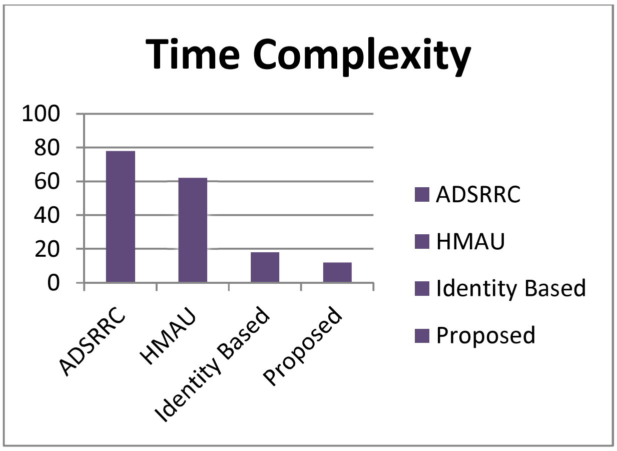
Pattern Type				Pregnancy Test		
				Support count	Impact factor	CIV
Soap	Tooth Paste			7/0.7	1/0.0 2/0.1	0.3/5=0.06<0.1
Soap	Tooth Paste	Horlicks		3/0.3	3/0.2	
Soap	Tooth Paste	Horlicks	Cream	3/0.3	4/0.0 5/0.0	
Tooth Paste	Horlicks			5/0.5		
Tooth Paste	Horlicks	Cream		5/0.4		
Horlicks	Cream			4/0.4		
Soap	Horlicks			3/0.3		
Soap	Cream			5/0.5		
Tooth Paste	Cream			6/0.6		
Tooth Paste	P.T			3/0.3		
Horlicks	Soap			3/0.3		
Horlicks	Cream			4/0.4		

The table 5 shows the result of computed impact factor at all item sets and impact values.

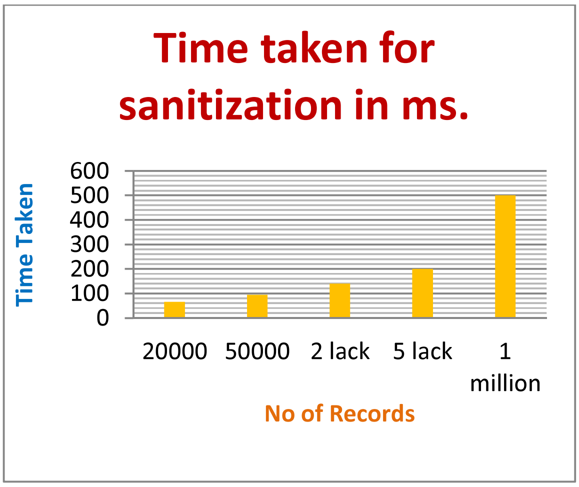
Table 6: Result Of Proposed System

Names	Soap	Tooth Paste	Horlicks	Cream	Pregnancy Test
Siva	1	1	1	1	1/0.0 2/0.1
radha	0	1	1	1	3/0.2
Ram	1	1	0	1	4/0.0 5/0.0
Rajes	1	1	1	1	
Saran	1	1	0	1	
Kumar	1	1	1	1	
Shela	1	1	0	0	
Selva	1	1	0	0	
Sivasankar	1	1	0	0	
rohini	0	1	1	0	

The table 6, shows the result produced by the proposed method and the pregnancy test has been hidden with the set of impact factors and been published.



Graph1: Shows The Time Complexity Between Other Methods.



Graph2 : Shows The Overall Time Taken For Sanitization Process .

5. CONCLUSION

We proposed a new sanitization approach for data publishing with privacy preservation based on impact matrix. The proposed method computes frequent pattern using support and count values. Based on computed pattern the impact factor of each attribute is computed using which cumulative impact value is computed. The sensitive item is identified using CIV based on which the proposed method generates sanitized data set for publication. The sanitized data set maintains the originality of the data and the end user can infer any information from that. The proposed method has produced efficient results with less time complexity.

REFERENCES

[1] Gokce, Sensitive knowledge hiding application , IEEE , Conference on Electrical, Electronics and Computer Engineering (ELECO), Page(s): 558 – 562, 2010.

[2] Murugeswari.s, An Efficient Method for Knowledge Hiding Through Database Extension, IEEE conference on Recent Trends in Information, Telecommunication and Computing (ITC), Page(s): 342 – 344, 2010.

[3] Shikha Sharma, An Extended Method for Privacy Preserving Association Rule Mining, Volume International Journal of Advanced Research in Computer Science and Software Engineering, 2, Issue 10, October 2012.

[4] V. Thavavel, A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012

[5] M Sridhar and Raveendra B Babu. A Fuzzy Approach for Privacy Preserving in Data Mining. International Journal of Computer Applications 57(18):1-5, November 2012.

[6] Komal Shah, Amit Thakkar and Amit Ganatra. Article: Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Multiple R.H.S. Items. International Journal of Computer Applications 45(1):1-7, May 2012.

[7] Chun Wei Lin, Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining, The Scientific World Journal Volume 2014 (2014).

[8] 8 T. Hong, C. Lin, C. Chang, and S. Wang, “Hiding sensitive itemsets by inserting dummy transactions,” in Proceedings of the IEEE International Conference on Granular Computing (GrC '11), pp. 246–249, November 2011.

[9] Dhyanendra Jain , Hiding Sensitive Association Rules without Altering the Support of Sensitive Item, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[10] Marcin Gorawski, An Efficient Algorithm for Sequential Pattern Mining with Privacy Preservation, Advances in Systems Science Advances in Intelligent Systems and Computing Volume 240, 2014, pp 151-161.

[11] Fatemeh Amiri, A Novel Community Detection Algorithm for Privacy Preservation in Social Networks, Intelligent Informatics Advances in Intelligent Systems and Computing Volume 182, 2013, pp 443-450.

[12] Kambombo Mtonga, Identity-Based Privacy Preservation Framework over u-Healthcare System, Multimedia and Ubiquitous Engineering Lecture Notes in Electrical Engineering Volume 240, 2013, pp 203-210.

[13] Murugeswari.s, An Efficient Method for Knowledge Hiding Through Database Extension, IEEE conference on Recent Trends in Information, Telecommunication and Computing (ITC), Page(s): 342 – 344, 2010.

[14] Abul O, MotifHider: A knowledge hiding approach to sequence masking , IEEE Conference on Computer and Information Sciences, pages 171-176 ,2009.

[15] Aris Gkoulalas-Divanis, Vassilios S. Verykios, "A Hybrid Approach to Frequent Itemset Hiding," ictai, vol. 1, pp.297-304, 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.1 (ICTAI 2007), 2007