Computer and Information Science

www.ccsenet.org/journal.html

Vol. 1, No. 3

August 2008

A Novel Sanitization Approach for Privacy Preserving Utility Itemset Mining

R.R.Rajalaxmi (Corresponding author)

Computer Science and Engineering

Kongu Engineering College

Erode – 638 052, Perundurai, TamilNadu, India

E-mail: rrr_kec@yahoo.co.in

A.M.Natarajan

Bannari Amman Institute of Technology

Sathy, TamilNadu, India

Abstract

Data mining plays a vital role in today's information world wherein it has been widely applied in various business organizations. The current trend in business collaboration demands the need to share data or mined results to gain mutual benefit. However it has also raised a potential threat of revealing sensitive information when releasing data. Data sanitization is the process to conceal the sensitive itemsets present in the source database with appropriate modifications and release the modified database. The problem of finding an optimal solution for the sanitization process which minimizes the non-sensitive patterns lost is NP-hard. Recent researches in data sanitization approaches hide the sensitive itemsets by reducing the support of the itemsets which considers only the presence or absence of itemsets. However in real world scenario the transactions contain the purchased quantities of the items with their unit price. Hence it is essential to consider the utility of itemsets in the source database. In order to address this utility mining model was introduced to find high utility itemsets. In this paper, we focus primarily on protecting privacy in utility mining. Here we consider the utility of the itemsets and propose a novel approach for sanitization such that minimal changes are made to the database with minimum number of non-sensitive itemsets removed from the database.

Keywords: Privacy Preserving Data mining, Frequent Itemset mining, Association Rule Mining, Utility mining, Data sanitization

1. Introduction

Large volumes of detailed personal data are regularly collected. Such data include shopping habits, criminal records, medical history, and credit records, among others (Brankovic L and Estivill-Castro V. 1999). These data can be analyzed by applications which make use of data mining techniques. Hence such data is an important asset to business organizations and governments for decision making processes and also to offer social benefits, such as medical research, crime reduction, national security, etc. (Jefferies P. 2000). On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly (Culnan M.J, 1993, Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni. 2004).

With the conventional data analysis methods there is a limited threat to privacy. Also these techniques mainly present the results based on the mathematical characteristics associated with the data. Making use of such techniques may not reveal some interesting patterns which are hidden in the data. By using appropriate data mining techniques it is possible to explore the hidden patterns. But the threat to privacy becomes real since data mining techniques are able to derive highly sensitive knowledge from unclassified data which is not even known to database holders (Elisa Bertino, Ravi Sandhu, 2005). In order to overcome this issue the data owners may decide not to share or release such data for analysis provided they should make a compromise for exploring hidden knowledge (Estivill-Castro V and Brankovic L, 1999, Atallah M., Bertino E., Elmagarmid A., Ibrahim M., and Verykios V. 1999).

Privacy preservation in data mining is the emerging research area which addresses the different ways to protect the sensitive knowledge discovery (Agrawal R and Srikant.2000, Verykios V.S., Bertino E., Fovino I.N., Provenza L.P.,

Saygin Y., and Theodoridis Y. 2004). It is mandatory to gain the benefit of data mining and as well as maintaining privacy. This paper focuses on the privacy preservation in utility mining. The rest of the paper is organized as follows: Section 2 discusses the background and related work in privacy preserving data mining. Section 3 gives an overview of the utility mining. In section 4 we describe the problem to be solved. Section 5 explains the proposed approach for privacy preservation in utility mining. Section 6 discusses the various experimental results carried and the detailed discussion on the analysis of results. Finally we conclude the paper in Section 7.

2. Background and Related Work

2.1 Frequent pattern Mining

Let I = $\{i_1, i_2, i_3, \ldots, i_n\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The occurrence frequency or support of an itemset is the number of transactions that contain the itemset (Han J and Kamber M. 2006). If the relative support of an itemset I satisfies a prespecified minimum support threshold, then I is a frequent itemset.

2.2 Privacy preservation of frequent itemsets

Since frequent itemset mining is a preliminary step in the association rule mining, most of the researches have addressed the privacy preservation of frequent itemsets with respect to association rule mining. A heuristic approach is presented in (Charu C. Aggarwal, Jan Pei, Bo Zhang, 2006) to hide the sensitive association rules against adversarial data mining. Randomization is an approach used to protect the discovery of association rules (Brankovic L and Estivill-Castro V. 1999). Additive Perturbation (Muralidhar K.,Parsa R. and Sarathy R. 1999) is used to provide security to the databases which does not reveal sensitive information. To hide restrictive patterns which are generated after the mining process a simple and effective way is to decrease their support in a given database (Agrawal D and Aggarwal C.C.2001, Alexandre Evfimievski,Ramakrishnan Srikant,Rakesh Agrawal,Johannes Gehrke. 2002, Evfimievski A, Srikant R, Agrawal R, and Gehrke J. 2002, Rizvi S. J. and Haritsa J. R. 1994). Data sanitization is the process of altering the transactions and was introduced in (Atallah M., Bertino E., Elmagarmid A., Ibrahim M., and Verykios V.. 1999). To do so, a small number of transactions have to be modified by deleting one or more items from them or even changing items in transactions, i.e., adding noise to the data. However this work relies on boolean association rules. The authors also proved that the optimal sanitization problem is NP-hard. On the other hand, the approach must hold the following restrictions: (1) the impact on the data in the database should be minimal and (2) an appropriate balance between privacy and knowledge discovery must be guaranteed. A set of new strategies and algorithms were proposed (Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni. 2004) for hiding sensitive knowledge from data by reducing the support and confidence of rules which specify how significant they are. The sensitive transactions are modified by removing some items, or inserting new items depending on the hiding strategy. However the proposed strategies do not consider the undesired side effects of hiding the sensitive rules. To overcome the limitations the authors in (Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen. January 2007) proposed a new approach which classifies the valid modifications for hiding sensitive rules and represents each class of the modifications by three attributes. The work in (Oliveira S. R. M. and Za¨ıane O. R. 2002, Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen. 2007)," concentrates on hiding the frequent itemsets which specifies a new framework for the hiding process. But the authors in (ZhihuiWang, Wei Wang, Baile Shi, S. H. Boey. 2006), proposed a novel approach by incorporating the hiding process in the frequent itemset mining algorithm itself so that it does not generate sensitive frequent itemsets.

The earlier work of association rule mining or frequent itemset mining considers either the presence or absence of items in the transaction. Hence privacy preservation is performed by either decreasing the confidence or support. However in real world the transactions made by the customer consist of quantities of items purchased with unit price of the items. In order to address privacy preservation in such databases a new set of strategies are required. Our work differs from the earlier work of frequent itemset hiding wherein we consider the utility of the items in the transactions.

3. Utility Mining

Earlier studies in frequent itemset mining focus in generating frequent itemsets by considering the presence or absence of items in the transactions. But in practical applications the transactions contain the purchased quantities of the items. In order to address this utility mining model was introduced in (Hong Yao, Howard J. Hamilton, and Cory J. Butz. 2004). Utility of an item specifies how useful an item is. The main goal in utility mining is to find the itemsets which yield high utility. In traditional association rule mining the utility of an item is either 0 or 1 (Rakesh Agrawal Ramakrishnan Srikant. 1994).

Let $I = \{i1, i2, \ldots im\}$ be a set of items, D be a transaction database, and UT<I,U> be a utility table, where U is a subset of the real numbers that reflect the utilities of the items. The *utility mining problem* is to discover all itemsets in a

transaction database D with utility values higher than the *minimum utility threshold*, given a utility table *UT.* We use the definition of a set of terms that leads to the formal definition of utility mining problem (Hong Yao, Howard J. Hamilton, and Cory J. Butz. 2004). Utility mining is to find all the itemsets whose utility values are beyond a user specified threshold called MUT. An itemset X is a *high utility itemset* if $u(X) \geq MUT$, where $X \subseteq I$ and MUT is the minimum utility threshold, otherwise, it is a *low utility itemset*.

4. Problem Formulation

In this work, our goal is to hide a group of interesting patterns which contains highly sensitive knowledge. We refer to these interesting patterns as sensitive patterns, and we define them as follows:

Definition: Let D is a transactional database, P be a set of all interesting patterns that can be mined from D, and P_s be a set of sensitive patterns that need to be hidden according to some security policies. A set of patterns, denoted by P_s, is said to be sensitive if $P_s \subset P$

The specific problem addressed in this paper can be stated as follows:

Given a transaction database, utility database, MUT, and a set of sensitive high utility itemsets (P_s), how can we modify the database such that using the same MUT, the set of non-sensitive high utility itemsets in the modified database can still be mined?

5. Conflict based Sanitization Approach

To solve this problem, we propose a novel approach called Conflict based Utility Itemset Sanitization (CUIS) that strategically modifies the database to decrease the utility of the sensitive itemsets. The approach is iteratively applied in a greedy fashion until the utility of each sensitive itemset falls below the threshold MUT while the number of non-sensitive itemsets with utility smaller than MUT is minimized. For each sensitive transaction the conflict degree is computed as the number of sensitive items that the transaction is supporting. It selects the transactions with maximum conflict degree for modification. The approach CUIS is described as follows:

Input: the collection P of high utility itemsets , the set P_s of sensitive high utility itemsets., Minimum Utility Threshold MUT .

Output: safe and sharable Database D'

Steps:

$D' = D$;

For each sensitive itemset p in P_s

1. Identify the set of transactions T_s supported by P_s
2. Find the difference of the sensitive itemset p as

   $Diff = U(p) - MUT$

   While ($Diff > 0$) do
3. Choose transaction $T \subset T_s$ such that the conflict degree is maximal

   Find the item $i_{max} \in T_s$ which has maximum transaction utility
4. if ($U(i_{max}, T) < Diff$)

   a. $O(i_{max}, T) = 0$

   b. $Diff = Diff - U(i_{max}, T)$

   else

   c. $O(i_{max}, T) = O(i_{max}, T) - ((diff / S(i_{max}))$

   d. $Diff = 0$

The set of transactions supported by the sensitive itemsets are identified (Step 1). For each of the private itemset the utility difference is computed (Step 2). Selecting the transaction with the maximum conflict degree can reduce the number of non-sensitive itemsets deleted in the sanitization process and the victim item which has maximum utility is selected for modification (Step 3). The process is repeated until the utility of the sensitive itemset falls below the MUT(Step 4).

6. Experimental Analysis

To measure the effectiveness, we adopt the set of metrics proposed in (Oliveira S. R. M. and Za¨ıane O. R. 2002) in terms of information loss and non-sensitive patterns removed as a side effect of the transformation process. The performance measures are specified as follows:

Misses Cost: It denotes the percentage of legitimate patterns that are not discovered from *D'*

$$MC = \frac{\#\sim Ps(D) - \#\sim Ps(D')}{\#\sim Ps(D)} \quad (1)$$

where $\#\sim P_s(X)$ denotes the number of non-sensitive patterns in the database X.

Original and Sanitized database Difference: It denotes the difference between the original (*D*) and sanitized database (*D'*)

$$diff(D, D') = \frac{|D - D'|}{|D|} \quad (2)$$

To measure the effectiveness of the algorithm, experiments were conducted on the synthetic datasets generated using IBM synthetic data generator (*http://www.almaden.ibm.com/software/quest/Resources/index.shtml*, 2003). It is highly advanced and considers typical properties of real transactional databases such as high frequencies of some itemsets, mean length of transactions, etc. This generator can produce only a binary form of transactional databases. Therefore, internal and external utilities were generated separately from the log-normal distribution in range [1, . . . , 10] (internal) and [1, . . . , 20] (external). Table 1 lists the characteristics of the datasets used in the experiment.

The experiments are conducted on a PC with Pentium IV processor (2 GHz) having 256MB main memory, running Windows XP. The effectiveness of the algorithm is studied based on the following condition: we varied the number of sensitive itemsets to hide and the disclosure threshold is zero. Based on this condition, no sensitive itemsets is disclosed from the sanitized database. Given the minimum utility threshold and an original dataset, the high utility itemsets are generated using the two-phase algorithm proposed in (Ying Liu, Wei-keng Liao, and Alok Choudhary. 2005). A certain number of sensitive itemsets were randomly selected from a set of high utility itemsets.

A standard method of frequent itemset hiding approach reduces the support of the itemsets either by deleting the items or the transactions that support it (Evfimievski A, Srikant R, Agrawal R, and Gehrke J. 2002). In order to compare our results we decrease the utility of the itemsets in High Utility Itemset Sanitization (HUIS) approach by choosing the victim item for modification such that it yields maximum profit in that transaction. However the proposed approach (CUIS) selects transactions based on the number of sensitive itemsets that are supported by it and shows promising results in terms of misses cost and the difference between the original and sanitized database.

Fig. 1-6 shows the summary of the sanitizing algorithm based on varying size of the database with the performance measurements. We fixed the minimum utility threshold as 0.1% and |Ps|=60. Fig.1-6 shows that CUIS outperforms HUIS and attains lowest miss cost and difference in almost all the datasets.

7. Conclusion

Privacy becomes an important factor in data mining so that sensitive information is not revealed after mining. However data quality is important such that no false information is released provided privacy is not jeopardized. In this paper we proposed a novel approach for preserving privacy in sensitive high utility itemset mining. The various experimental results show that the approach applies minimum number of changes to the database and minimal amount of non-sensitive itemsets are missed which is the ultimate aim of data sanitization. Future work has to be carried over to develop optimal algorithms for data sanitization.

References

Agrawal D and Aggarwal C.C (2001). "*On the Design and Quantification of Privacy Preserving Data Mining Algorithms*", In Proc. of ACM SIGMOD/PODS, pages 247–255, Santa Barbara, CA.

Agrawal R and Srikant (2000). R, "*Privacy Preserving Data Mining*,", Proc. ACM SIGMOD Conf. on Management of Data, pp. 439-450.

Alexandre Evfimievski,Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke. (2002). "*Privacy preserving mining of association rules*", Proceedings of SIGKDD.

Atallah M., Bertino E., Elmagarmid A., Ibrahim M., and Verykios V.. (1999). *Disclosure Limitation of Sensitive Rules*. In Proc. of IEEE Knowledge and Data Engineering Workshop, pages 45–52, Chicago, Illinois.

Brankovic L and Estivill-Castro V. (1999). *Privacy Issues in Knowledge Discovery and Data Mining*. In Proc. of Australian Institute of Computer Ethics Conference (AICEC99), Melbourne, Victoria, Australia,.

Charu C. Aggarwal, Jan Pei, Bo Zhang, (2006). "*On Privacy Preservation against Adversarial Data Mining*", Proceedings of Knowledge Discovery in Databases

Culnan M.J, (1993), "*How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information*", MIS Quarterly, 17(3): 341–363.

Elisa Bertino, Ravi Sandhu, (2005), "*Database Security—Concepts, Approaches, and Challenges*", IEEE transactions on dependable and secure computing, vol.2, no. 1.

Estivill-Castro Vand Brankovic L, (1999). "*Data Swapping: Balancing Privacy Against Precision in Mining for Logic Rules*", In Proc. of Data Warehousing and Knowledge Discovery (DaWaK-99), pages 389–398, Florence, Italy.

Evfimievski A, Srikant R, Agrawal R, and Gehrke J. (2002). "*Privacy Preserving Mining of Association Rules*", In Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pages 217–228, Edmonton, AB, Canada.

Han J and Kamber M. (2006). "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, San Francisco, CA.

Hong Yao, Howard J. Hamilton, and Cory J. Butz. (2004). "*A Foundational Approach to Mining Itemset Utilities from Databases*", In Proceedings of the 4th SIAM International Conference on Data Mining.

Jefferies P. (2000). "*Multimedia, Cyberspace & Ethics*", In Proc. of International Conference on Information Visualization, pages 99–104, London, England.

Muralidhar K.,Parsa R. and Sarathy R. (1999). "*A General Additive data Perturbation method for Database Security* ",Management Science,45(10):1399-1415.

Oliveira S. R. M. and Zaiane O. R. (2002). "*Privacy Preserving Frequent Itemset Mining*." In Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining, pages 43–54, Japan.

Rakesh Agrawal Ramakrishnan Srikant. (1994). "*Fast algorithms for mining association rules*", Proceedings of the 20th VLDB Conference.

Rizvi S. J. and Haritsa J. R. (1994). "*Privacy-Preserving Association Rule Mining*". In Proc. of the 28th International Conference on Very Large Data Bases, x(2002), Hong Kong, China.

Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni. (April 2004). "*Association rule hiding*", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 4.

Verykios V.S., Bertino E., Fovino I.N., Provenza L.P., Saygin Y., and Theodoridis Y. (2004). "*State-of-the-Art in Privacy Preserving Data Mining"* ACM SIGMOD Record, vol. 3, no. 1, pp. 50-57.

Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen. (January 2007). "*Hiding sensitive association rules with limited side effects*", IEEE Transactions on Knowledge and Data Engineering, Vol.19, No.1.

ZhihuiWang, Wei Wang, Baile Shi, S. H. Boey. (2006). "*Preserving Private Knowledge in Frequent Pattern Mining",* Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06).

Ying Liu, Wei-keng Liao, and Alok Choudhary. (2005). "*A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets"*, PAKDD 2005, LNAI 3518, pp. 689 – 695.

URL *http://www.almaden.ibm.com/software/quest/Resources/index.shtml*, 2003.

Table 1. Summary of the datasets used in the experiment.

Database name	|D|	|I|	|L|	|T|
T10I10L5D1K	1K	10	5	10
T10I10L5D2K	2K	10	5	10
T20I10L5D4K	4K	10	5	20

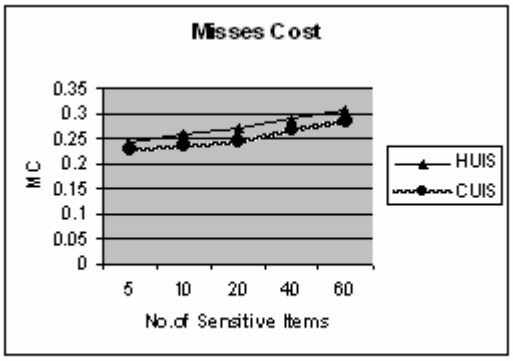

Figure 1. Misses cost for |Ps|=60 and T10I10L5D1K

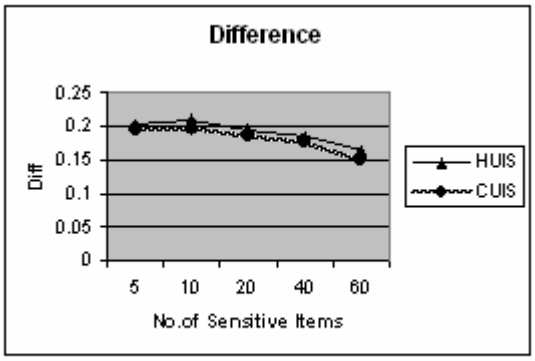

Figure 2. Difference for |Ps|=60 and T10I10L5D1K

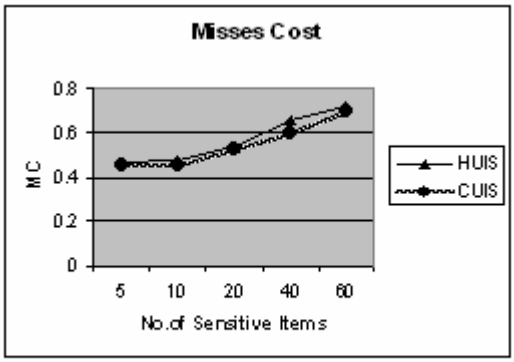

Figure 3. Misses cost for |Ps|=60 and T10I10L5D2K

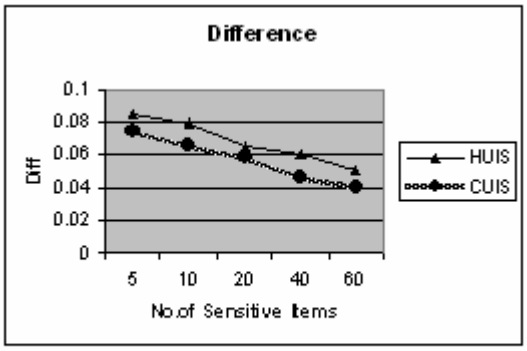

Figure 4. Difference for |Ps|=60 and T10I10L5D2K

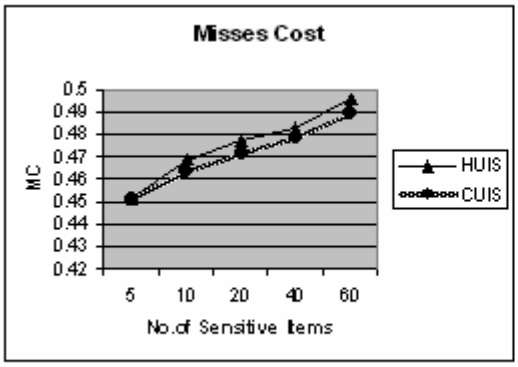

Figure 5. Misses cost for |Ps|=60 and T20I10L5D4K

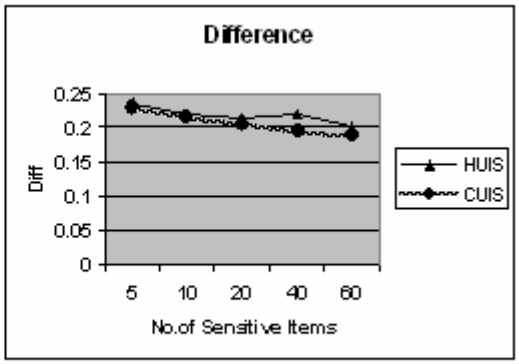

Figure 6. Difference for |Ps|=60 and T20I10L5D4K