International Journal of Latest Research In Engineering and Computing (IJLREC)
Volume 1, Issue 1 : Page No.40-49, September-October 2013
www.ijlrec.com

Extending Database for Hiding Sensitive Frequent Data Items

Mrs.P.Rubini[1], Dr.D.Tensing[2]
[1]Research Scholar, Dept of CSE, Anna University, Chennai, Tamilnadu, India
[2]Professor & Director, School of Civil Engg, Karunya University, Coimbatore, India

Abstract: In this paper, we propose a novel, exact border-based approach that provides an optimal solution for the hiding of sensitive frequent item sets by 1) minimally extending the original database by a synthetically generated database part-the database extension, 2) formulating the creation of the database extension as a constraint satisfaction problem, 3) mapping the constraint satisfaction problem to an equivalent binary integer programming problem, 4) exploiting underutilized synthetic transactions to proportionally increase the support of non sensitive item sets, 5) minimally relaxing the constraint satisfaction problem to provide an approximate solution close to the optimal one when an ideal solution does not exist, and 6) using a partitioning in the universe of the items to increase the efficiency of the proposed hiding algorithm. Extending the original database for sensitive item set hiding is proved to provide optimal solutions to an extended set of hiding problems compared to previous approaches and to provide solutions of higher quality. Moreover, the application of binary integer programming enables the simultaneous hiding of the sensitive item sets and thus allows for the identification of globally optimal solutions

Key words: knowledge hiding, association rule mining, binary integer programming.

INTRODUCTION

Advances in data collection, processing, and analysis, along with privacy concerns regarding the misuse of the induced knowledge from this data, soon brought into existence the field of privacy preserving data mining. Simple de-identification of the data prior to its mining is insufficient to guarantee a privacy-aware outcome since intelligent analysis of the data, through inference based attacks, may reveal sensitive patterns that were unknown to the database owner before mining the data. Thus, compliance to privacy regulations requires the incorporation of advanced and sophisticated solutions. This paper concentrates on a subfield of privacy preserving data mining, known as "Data hiding. "To motivate our discussion, we present the following scenario..Let us suppose that we are negotiating a deal with Dedtrees Paper Company, as purchasing directors of Big Mart, a large supermarket chain. They offer their products with a reduced price if we agree to provide them access to our database of customer purchases. We accept the deal and Dedtrees starts mining our data. By using an association rule mining tool, they find that people who purchase skim milk also purchase Green paper. Dedtrees now runs a coupon marketing campaign saying that "you can get 50 cents off skim milk with every purchase of a Dedtrees product." This campaign cuts heavily into the sales of Green paper, which increases the prices to us, based on the lower sales. During our next negotiation with Dedtrees, we find out that with reduced competition, they are unwilling to offer us a low price. Finally, we start to lose business to our competitors, who were able to negotiate a better deal with Green paper. This scenario demonstrates the urging need to prevent disclosure of both confidential personal information from summarized data and of sensitive knowledge that can be mined from this data. For this reason, algorithms need to be devised that can effectively protect the sensitive knowledge from being mined.

In the following sections, we present a novel approach that strategically performs sensitive frequent item set hiding based on a new notion of hybrid database generation. This approach broadens the regular process of data sanitization by applying an extension to the original database instead of either modifying existing transactions (directly or through the application of transformations) or rebuilding the data set from scratch to accommodate knowledge hiding. The extended portion of the data set contains a set of carefully crafted transactions that achieve to lower the importance of the sensitive patterns to a degree that they become uninteresting from the perspective of the data mining algorithm, while minimally affecting the importance of the non-sensitive ones.

The hiding process is guided by the need to maximize the data utility of the sanitized database by introducing the least possible amount of side effects, such as

1) The hiding of non-sensitive patterns or
2) The production of frequent patterns that were not existent in the initial data set (ghost item sets).

The released database, which consists of the initial part (original database) and the extended part (database extension), can guarantee the protection of the sensitive knowledge, when mined at the same or higher support as the one used in the original database. Therefore, to protect the sensitive knowledge, the security administrator should appropriately set the minimum support threshold to a value that is lower than the usually used. Threshold in a typical mining scenario concerning the data at hand. In accordance to the majority of the knowledge hiding approaches, this work aims at creating a sanitized version of the original data set that can safely be released to the public. The rationale is that such a data set will be useful to allow for the extraction of various types of knowledge patterns and to provide the opportunity to its owner to safely make use of this data at his or her own discretion. The approach introduced in this paper is exact in nature; provided that a hiding solution that causes no side effects in the sanitized database exists, the proposed algorithm is guaranteed to find it. On the contrary, when an exact solution is impossible, the algorithm identifies an approximate solution that is close to the optimal one. To accomplish the hiding task, the proposed approach administers the sanitization part by formulating a Constraint Satisfaction Problem (CSP) and by solving it through Binary Integer Programming (BIP). The measure of distance is used to formulate the criterion that will drive the optimization process to the optimal solution. Through a set of experiments, we demonstrate the effectiveness of this approach toward identifying optimal hiding solutions bearing no side effects.

RELATED WORK

The presented methodology lies between the fields of frequent item set hiding and synthetic database generation (examined in the context of privacy preservation). To the best of our knowledge, apart from ongoing research work regarding an additive model for sensitive item set hiding, this approach is the first to facilitate knowledge hiding through the extension of the database. Extending the original database to accommodate knowledge hiding can be considered as a bridging between the item set hiding and the synthetic database generation approaches. In what follows, we review some of the fundamental related work in both research directions.

KNOWLEDGE HIDING FORMULATION

This section provides the necessary background regarding sensitive item set hiding and sets out the problem at hand, as well as the proposed hiding methodology.

A. Hiding Methodology:

To properly introduce the hiding methodology, one needs to consider the existence of three databases, all depicted in binary format. They are defined as follows: Database Do is the original transaction database that, when mined at a certain support threshold msup, leads to the disclosure of some sensitive knowledge in the form of sensitive frequent patterns. This sensitive knowledge needs to be protected. . Database Dx is a minimal extension of Do that is created by the hiding algorithm during the sanitization process, in order to facilitate knowledge hiding. Database D is the union of database Do and the applied extension Dx and corresponds to the sanitized outcome that can be safely released. Suppose that database Do consist of N transactions. By performing frequent item set mining in Do, using a support threshold msup set by the owner of the data, a set of frequent patterns are discovered (denoted hereon as FDo), among which a subset S contains patterns that are sensitive from the owner's perspective. The goal of the hiding algorithm is to create a minimal extension to the original database Do in a way that the final, sanitized database D protects the sensitive item sets from disclosure. The database extension can by itself be considered as a new database Dx, since it consists of a set of transactions in the same space of items I as the ones of Do. Among alternative hiding solutions that may exist, the target of the proposed algorithm is to protect the sensitive knowledge, while minimally affecting the nonsensitive item sets appearing in FDo. This means that all the nonsensitive item sets in FDo should continue to appear as frequent among the mined patterns from D, when performing frequent item set mining using the same or a higher threshold. The hiding of a sensitive item set is equivalent to a degradation of its statistical significance, in terms of support, in the result database. The proposed algorithm first applies border revision to identify the revised borders for D, then computes the minimal size for the extension Dx and, by using the item sets of the revised borders, defines a CSP that is solved using BIP. In this way, all possible assignments of item sets to the transactions of Dx are examined and the optimal assignment is bound to be found.

Although the properties of the produced CSP allow for an acceptable runtime of the hiding algorithm, there are cases in which the partitioning approach of Section 6 becomes useful to accommodate for very large problem sizes.

	a	b	c	d	e	f
$\mathcal{D}_\mathcal{O}$	1	1	0	0	0	1
	1	1	1	1	0	0
	1	0	1	0	0	1
	1	0	0	0	0	0
	0	1	0	0	1	0
	1	1	1	1	1	0
	0	0	0	1	0	0
	1	1	1	0	1	0
	0	1	1	0	0	0
	1	0	1	1	1	0
$\mathcal{D}_\mathcal{X}$	1	0	0	0	0	0
	1	0	1	1	0	0
	1	0	1	1	0	0
	1	1	0	0	0	0

TABLE 1

Sanitized Database D as a Mixture of the Original Database Do and the Applied Extension Dx

B. *A Running Example*

Suppose we are provided with database Do in Table 1.Applying frequent item set mining in Do using mfreq = 0:3 leads to the set of large item sets FDo appearing in the upper part of Table 2. Among these item sets, let S ={e,ae,bc} denote the sensitive knowledge that has to be protected. The proposed hiding algorithm aims at the creation of a database extension Dx to Do (see Table 1) that allows the hiding of the sensitive knowledge, while keeping the nonsensitive patterns frequent in the sanitized outcome.

Table 1 summarizes the target of the hiding algorithm. The unions of the two data sets Do and Dx corresponds to the sanitized outcome D that can be safely released. Thus, the primary goal of the hiding algorithm is to construct the privacy-aware extension Dx such that 1) it contains the least amount of transactions needed to ensure the proper hiding of the sensitive knowledge in Do and 2) it introduces no side effects in the hiding process. As we can observe in the lower part of Table 2, all the sensitive item sets of Do along with their supersets are infrequent in D (shown under the dashed line), while the entire nonsensitive item sets of Do remain frequent. Since Do is extended, in order to ensure that the nonsensitive patterns will remain frequent in D, the hiding algorithm needs to appropriately increase their support in the sanitized database.

Frequent itemset in $\mathcal{D}_\mathcal{O}$	Support
{a}	7
{b}, {c}	6
{ac}	5
{d}, {e}, {ab}, {bc}	4
{ad}, {ae}, {be}, {cd}, {ce}, {abc}, {acd}, {ace}	3
Frequent itemset in $\mathcal{D}$	**Support**
{a}	11
{c}	8
{b}, {ac}	7
{d}	6
{ab}, {ad}, {cd}, {acd}	5
~~{e}~~, ~~{bc}~~	4
~~{ae}~~, ~~{be}~~, ~~{ce}~~, ~~{abc}~~, ~~{ace}~~	3

TABLE 2
Frequent Item Sets for DO and DX at msup = 3

I. MAIN ISSUES PERTAINING TO THE HIDING METHODOLOGY

The proposed hiding solution creates a sanitized database D that corresponds to a mixture of the original transactions in Do and a set of synthetic transactions, artificially created to prohibit the leakage of sensitive knowledge. For security reasons, all the transactions in D are assumed to be randomly ordered so that it is difficult for an adversary to distinguish between the real ones and those that were added by the hiding algorithm to secure the sensitive knowledge. There are several issues of major

importance, involving the hiding methodology, that need to be examined. To continue, let P denote the size of database D, N is the size of database Do, and Q the size of the extension Dx.

A. Size of the Database Extension

Since database Do is extended by Dx to construct database D, an initial and very important step in the hiding process is the computation of the size of Dx. A lower bound on this value can be established based on the sensitive item set in S, which has the highest support (breaking ties arbitrarily). The rationale here is given as follows: by identifying the sensitive item set with the highest support, one can safely decide upon the minimum number of transactions that must not support this item set in Dx, so that it becomes infrequent in D. This number, theoretically, is sufficient to allow the hiding of all the other item sets participating in S and all its supersets, and corresponds to the minimum number of transactions that Dx must have to properly secure the sensitive knowledge. Theorem 1 demonstrates how this lower bound Q is established.

Theorem 1:

Let IM € S such that for all I € S it holds that sup(IM,Do) ≥ sup(I,Do). Then, the minimum size of Dx to allow the hiding of the sensitive item sets from S in D is equal to

Q= [(sup(IM,Do)/mfreq) -N]+1

Proof*:*

We only need to prove that any item set I € S will become hidden in D if and only if Q > (sup(I,Do)/mfreq) -N, provided that I is not supported in Dx.Since item set I must be infrequent in database D, the following condition holds:

Sup(I,D) < msup ⟶ sup(I,Do)+ sup(I,Dx)< mfreq.P

Moreover, since sup(I,Dx)≥0 and (by construction) P=N+Q, we have that

Sup(I,Do) < mfreq.(N+Q)⟶ sup(I,Do)<N+Q.

The last inequality was relaxed by removing term sup(I,Dx). Since it consists of the summation of nonnegative terms and a nonnegative term was removed, the inequality will continue to hold. Its holding proves the holding of (1) since the sensitive item set with the highest support will require the largest amount of transactions (not supporting it) in the extension Dx in order to be properly hidden. The lower bound of the number of the necessary transactions for Dx will thus equal the floor value of sup(IM,Do)/mfreq-N plus one. Moreover, as expected, item sets having lower support than IM may be supported by some transactions of database Dx , as long as they are infrequent in D. provides the absolute minimum size of Dx to accommodate for the sensitive knowledge hiding. However, as is presented later on, this lower bound may, under certain circumstances, be insufficient to allow for the identification of an optimal solution, even if such a solution exists. This situation may occur if, for instance, the number of transactions returned by (1) is too small to allow for consistency among the different requirements imposed upon the status (frequent versus infrequent) of the various item sets appearing in D.

B. Exact and Ideal Solutions

Having identified the size Q of database Dx, the next step is to properly construct these transactions to facilitate knowledge hiding. Since the actual values of all the items in the database extension are unknown at this point, the hiding algorithm represents them with binary variables that will be instantiated later on in the process. In what follows, let Uqm be the binary variable corresponding to the mth item of transaction Tq € Dx(q€[1,Q],m€[1,M]), when Do is in the sanitization process. Under this formulation, the goal of the hiding algorithm becomes to optimally adjust all the binary variables involved in all the transactions of Dx in order to hide the sensitive item sets, while minimally affecting the nonsensitive ones in a way that they remain frequent in the sanitized outcome. This is the notion of an exact solution.

Definition 1 (feasible/exact/approximate solution):

A solution to the hiding of the sensitive knowledge in Do is considered as feasible if it achieves to hide the sensitive patterns. Any feasible solution, introducing no side effects in the hiding process, is called exact. Finally, any non exact feasible solution is called approximate.

In a typical hiding scenario, distinct feasible solutions are of different quality. Thus, an optimization criterion needs to be incorporated in the hiding strategy to guide the algorithm to the best possible among all the feasible solutions. The metric of distance is applied to quantify the notion of "harm" caused to the original data set by the sanitization process. In the context of this work, the distance between data base Do and its sanitized version D is measured based on the extension Dx as follows:

$$\text{Dist}(Do,D) = \sum_{q\in[1,Q],m\in[1,M]} uqm$$

As one can observe, the minimum impact of D can be quantified as the minimum distance between Do and D. Thus, the objective of the hiding algorithm becomes to appropriately set the uqm variables such that the sensitive knowledge is hidden, while distance is minimized. An interesting property of the distance measure is that it allows the hiding algorithm to ensure high quality in the sanitized database D and to identify the ideal solution, if one exists. The notion of an ideal solution is presented in Definition 3. Based on the notion of distance and the size of the extension Dx, the database quality is defined as follows:

Definition 2 (database quality):

Given the sanitized database D, its original version Do, and the produced extension Dx , the quality of database D is measured both in the size of Dx and in the number of binary variables set to "1" in the transactions of Dx (i.e., the distance metric). In both cases, lower values correspond to better solutions. Through (1), the hiding algorithm is capable of identifying the lower bound in the size of Dx that is necessary to accommodate for hiding of the sensitive knowledge in Do.

Definition 3 (ideal solution):

A solution to the hiding of the sensitive item sets is considered as ideal if it has the minimum distance among all the existing exact solutions and is obtained through the minimum expansion of Dx. In that sense, ideal is a solution that is both minimal (with respect to distance and size of extension) and exact.

C. *The Revision of the Borders:*

The concept of border revision provides the underlying mechanism for the specification of the values of the uqm variables to 1s or 0s, in a way that minimizes the impact on D. The rationale behind this process is that hiding of a set of item sets corresponds to a movement of the original borderline in the lattice that separates the frequent item sets from their infrequent counterparts (see Fig. 1), such that the sensitive item sets lie below the revised borderline. There are four possible scenarios involving the status of each item set I prior and after the application of border revision.

- C1. Item set I was frequent in Do and remains frequent in D.
- C2. Item set I was infrequent in Do and is infrequent in D.
- C3. Item set I was frequent in Do and became Infrequent in D.
- C4. Item set I was infrequent in Do and became frequent in D.

Since the borders are revised to accommodate for an exact solution, the revised hyper plane (Fig. 1b) is designed to be ideal in the sense that it excludes only the sensitive item sets and their supersets from the set of frequent patterns in D, leaving the rest of the item sets in their previous status as in database Do. To properly define the notion of an ideally revised border (hereon called "revised border"), we first need to introduce two sets related to the set of sensitive item sets S.

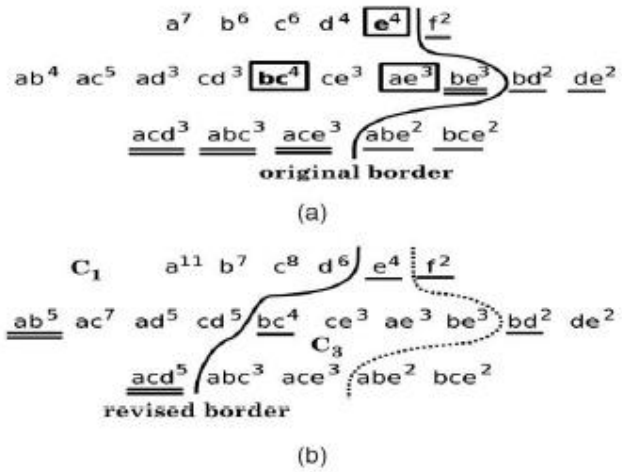


Fig. 1. An item set lattice demonstrating (a) the original border and the sensitive item sets, and (b) the revised border.

D. *Problem Size Reduction:*

To enforce the computed revised border and identify the exact hiding solution, a mechanism is needed to regulate the status (frequent versus infrequent) of all the item sets in D. Let C be the minimal set of border item sets used to regulate the values of the various uqm variables in Dx . Moreover, suppose that I ∈ C is an item set, whose behavior we want to regulate in D. Then, item set I will be frequent in D if and only if sup(I,Do)+sup(I,Dx)≥ mfreq*(N+Q) corresponds to the minimum number of times that an item set I has to appear in the extension Dx to remain frequent in D. On the other hand, provides the maximum number of times that an item set I has to appear in Dx to be infrequent in database D. To identify an exact solution to the hiding

problem, every possible item set in P, according to its position in the lattice—with respect to the revised border—must satisfy above. However, the complexity of solving the entire system of the 2^M -1 inequalities is well known to be NP-hard .Therefore, one should restrict the problem to capture only a small subset of these inequalities, thus leading to a problem size that is computationally manageable. The proposed problem formulation achieves this by reducing the number of the participating inequalities that need to be satisfied. Even more, by carefully selecting the item sets of set C, the hiding algorithm ensures that the exact same solution to the one of solving the entire system of inequalities is attained. This is accomplished by exploiting cover relations existing among the item sets in the lattice due to the monotonicity of support.

E. *Handling of Sub optimality:*

Since an exact solution may not always be feasible, the hiding algorithm should be capable of identifying good approximate solutions. There are two possible scenarios that may lead to nonexistence of an exact solution. Under the first scenario, Do itself does not allow for an optimal solution due to the various supports of the participating item sets. Under the second scenario, database Do is capable of providing an exact solution, but the size of the database extension is insufficient to satisfy all the required inequalities of this solution. To tackle the first case, the hiding algorithm assigns different degrees of importance to different inequalities. To be more precise, while it is crucial to ensure that holds for all sensitive item sets in D, thus they are properly protected from disclosure, satisfaction of an item set rests in the discretion of ensuring the minimal possible impact of the sanitization process to Do. This inherent difference in the significance of the two inequalities, along with the fact that solving the system of all inequalities of the form always leads to a feasible solution (i.e., for any database Do), allows the relaxation of the problem, when needed, and the identification of a good approximate solution. To overcome the second issue, the hiding algorithm incorporates the use of a safety margin threshold, which produces a further expansion of Dx by a certain number of transactions. These transactions must be added to the ones computed. The introduction of a safety margin can be justified as follows: Since the lower bound on the size of database Dx , it is possible that the artificially created transactions are too few to accommodate for the proper hiding of knowledge. This situation may occur due to conflicting constraints imposed by the various item sets regarding their status in D. These constraints require more transactions (or to be more precise, more item modifications) in order to be met. Thus, a proper safety margin will allow the algorithm to identify an exact solution if such a solution exists. Moreover, as is demonstrated in Section 5.4, the additional extension of Dx, due to the incorporation of the safety margin, can be restricted to the necessary degree. A portion of transactions in Dx is selected and removed at a later point, thus reducing its size and allowing an exact solution. Therefore, the only side effect of the use of the safety margin in the hiding process is inflation in the number of constraints and associated binary variables in the problem formulation, leading to a minuscule overhead in the runtime of the hiding algorithm.

a	b	c	d	e	f
1	1	1	1	1	0
0	1	0	0	1	0
1	0	1	1	1	0
0	0	0	0	0	0

TABLE 4 - Database DX after the Solution of the CSP

II. HYBRID SOLUTION METHODOLOGY

In the following sections, we present a way to minimize the problem size by regulating the status of only an essential portion of the item sets from P. Moreover, we propose a solution to the size of the extension Dx and formulate the hiding process as a CSP that is solved by using BIP. Finally, the critical issues of how to ensure validity of transactions in D and how to handle sub optimality in the non exact solutions are addressed.

A. *Adjusting the Size of the Extension:*

Equation provides the absolute minimum number of transactions that need to be added in Dx , to allow for the proper hiding of the sensitive item sets of Do. However, this lower bound can, under certain circumstances, be insufficient to allow for the identification of an exact solution, even if one exists.

To circumvent this problem, one needs to expand the size Q of Dx as determined by a certain number of transactions. A threshold, called safety margin (denoted hereon as SM), is incorporated for this purpose. Safety margins can be either predefined or be computed dynamically, based on particular
properties of database Do and/or other parameters regarding the hiding process. In any case, the target of using a safety margin is to ensure that an adequate number of transactions participate in Dx , thus an exact solution (if one exists) will not be lost due to the small size of the extension.

Since for each transaction in Dx, M new binary variables are introduced that need to be tuned when solving the system of inequalities from C, one would ideally want to identify a sufficiently large number of transactions for Dx (that allow for an exact solution), while this number be as low as possible to avoid unnecessary variables and constraints participating in the hiding process. Supposing that the value of Q is adjusted and a sufficiently large safety margin is used, the methodology minimizes the size of Dx after the sanitization process to allow for an ideal solution.

B. *Formulation and Solution of the CSP:*

A CSP is defined by a set of variables and a set of constraints, where each variable has a nonempty domain of potential values. The constraints involve a subset of the variables and specify the allowable combinations of values that these variables can attain. An assignment that does not violate the set of constraints is called "consistent." A solution of a CSP is a complete assignment of values to the variables that satisfies all the constraints. Since in this work all variables involved are binary in nature, the produced CSP is solved by using a technique called BIP that transforms it to an optimization problem. To avoid the high degree2 of constraints, the application of a Constraints Degree Reduction (CDR) approach is essential. This approach relies on the binary nature of the variables to linearize all the nonlinear constraints. Linearization does not lead to any information loss; its only side effect is an increase in the number of variables and constraints in the system. On the other hand, the resulting inequalities are simple in nature and allow for fast solutions, thus adhere for an efficient solution of the entire CSP. The proposed CSP formulation (as an optimization process) is depicted in Fig. 2, while Fig. 3 demonstrates the CDR approach that is enforced. Supposing that DX is large enough and that database DO allows for an exact solution, the above hiding formulation is capable of identifying it. However, database DO may not always allow for an exact solution. An approach for dealing with sub optimality, while the following section examines the issue of validity in the transactions of DX and presents an algorithm for removing unnecessary transactions, existing due to the use of the safety margin.

$$\textbf{minimize}\left(\sum_{q\in[1,Q+\mathrm{SM}],m\in[1,M]} u_{qm}\right)$$

$$\textit{subject to}\begin{cases}\sum_{q=1}^{Q+\mathrm{SM}}\prod_{i_m\in I}u_{qm} < thr, \forall I\in \mathcal{B}d^{-}(\mathcal{F}'_{\mathcal{D}})\\ \sum_{q=1}^{Q+\mathrm{SM}}\prod_{i_m\in I}u_{qm} \geq thr, \forall I\in \mathcal{B}d^{+}(\mathcal{F}'_{\mathcal{D}})\end{cases}$$

$$\text{where}\quad thr = mfreq\cdot(N+Q+\mathrm{SM}) - \sup(I,\mathcal{D}_{\mathcal{O}})$$

\

Fig. 2. The CSP formulation as an optimization process.\

$$\textbf{Replace All}\quad\left(\sum_{T_q\in\mathcal{D}_{\mathcal{X}}}\Psi_s \gtreqless thr, \Psi_s = \prod_{i_m\in T_q} u_{qm}\right)$$

$$\textbf{With}$$

$$\forall i\begin{cases} c_1: & \Psi_s \leq u_{q1}\\ c_2: & \Psi_s \leq u_{q2}\\ \vdots & \\ c_Z: & \Psi_s \leq u_{qm}\\ & \Psi_s \geq u_{q1}+u_{q2}+\ldots+u_{qm}-|Z|+1\end{cases}$$

$$\textbf{And}$$

$$\sum_s \Psi_s \gtreqless thr$$

$$\text{where}\quad \Psi_s\in\{0,1\}$$

Fig. 3. The CDR approach

$$
\mathcal{B}d^{+}(\mathcal{F}'_{\mathcal{D}}) \begin{cases} \{ab\} : u_{11}u_{12} + u_{21}u_{22} + u_{31}u_{32} + \\ \qquad + u_{41}u_{42} \geq 0.2 \\ \{acd\} : u_{11}u_{13}u_{14} + u_{21}u_{23}u_{24} + \\ \qquad + u_{31}u_{33}u_{34} + u_{41}u_{43}u_{44} \geq 1.2 \end{cases}
$$

$$
\mathcal{B}d^{-}(\mathcal{F}'_{\mathcal{D}}) \begin{cases} \{e\} : u_{15} + u_{25} + u_{35} + u_{45} < 0.2 \\ \{f\} : u_{16} + u_{26} + u_{36} + u_{46} < 2.2 \\ \{bc\} : u_{12}u_{13} + u_{22}u_{23} + u_{32}u_{33} + \\ \qquad + u_{42}u_{43} < 0.2 \\ \{bd\} : u_{12}u_{14} + u_{22}u_{24} + u_{32}u_{34} + \\ \qquad + u_{42}u_{44} < 2.2 \end{cases}
$$

$$
T_q \in \mathcal{D}_{\mathcal{X}} \begin{cases} u_{11} + u_{12} + u_{13} + u_{14} + u_{15} + u_{16} \geq 1 \\ u_{21} + u_{22} + u_{23} + u_{24} + u_{25} + u_{26} \geq 1 \\ u_{31} + u_{32} + u_{33} + u_{34} + u_{35} + u_{36} \geq 1 \\ u_{41} + u_{42} + u_{43} + u_{44} + u_{45} + u_{46} \geq 1 \end{cases}
$$

Fig. 4. The constraints in the CSP of the running example.

RESULTS

Original Data

Hiding Data

FUTURE ENHANCEMENT:

First, develop and design a sound sanitization algorithm performed on the set of frequent item sets with support counts. The input of the algorithm is: a set of frequent item sets with support counts *FS* discovered from a real database, a set of association rules *R* derived from *FS*, and a subset of sensitive rules *R*B*h*B. The output of the algorithm is a set of frequent item sets with support counts from which we can just derive the set of rules *R*-*R*B*h*B. The algorithm itself should take into the following considerations:

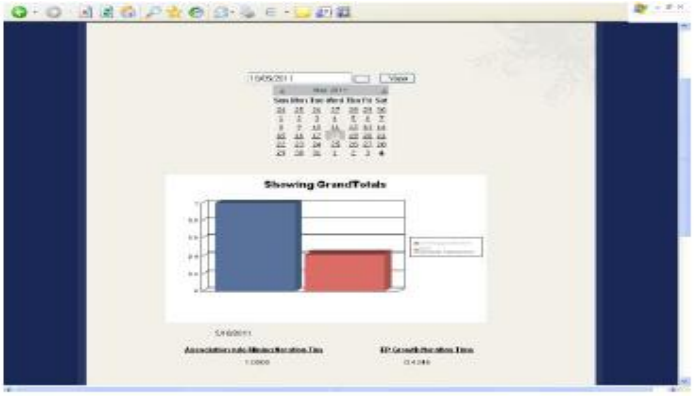
1) Ideally, the support and confidence of the rules in *R*-*R*B*h*B should remain unchanged as much as possible;

2) The algorithm should be able to select appropriate hiding strategies according to different kinds of correlations among the rules in *R* and *R*B*h*B, and

3) It should provide a security mechanism of preventing rule-based reasoning, that is, deal with the case in which sensitive rules can be reasoned from non-sensitive rules.

Second, investigate how to restrict the number of transactions in the new released database. Our current work of the FP-tree-based inverse frequent set mining did not restrict and control the number transactions in the new generated database. As an important characteristic of transaction database, the number of transactions will directly affect the support of a rule. The number of transactions that our current algorithms output is previously unknown, making the support of a rule is uncertain although the support count of the corresponding item set is certain.

Third, develop an integrated secure association rule mining tool which can conceal (protect) privacy data & sensitive association rules contained in the data simultaneously. In privacy preserving data sharing context, both the sensitive data and rules contained in the data need hiding (we call sensitive data hiding in database DHD and sensitive rules hiding in database KHD). However, currently, DHD and KHD techniques are always investigated separately, and there is still lack of a tool integrating both DHD and KHD techniques. Development of such a tool is significant and imperative under this situation.



CONCLUSIONS

In this paper, we have presented a novel, exact border based approach to sensitive knowledge hiding, through the introduction of a minimal extension to the original database. By exploiting the revised borders as well as the cover relationships among the item sets, we were able to minimize the set of item sets participating in the CSP, which provides a solution to the sensitive knowledge hiding. The attained solution is identical to the one of solving the CSP involving the whole set of item sets. The proposed methodology is capable of identifying an ideal solution whenever one exists, or approximate the exact solution, otherwise. In this work, we provided insight on various topics, such as the minimum expansion of the original database, the validation of the constructed transactions, and the treatment of sub optimality in solutions. A partitioning approach was introduced to improve the scalability of the algorithm. Finally, through experiments we demonstrated that the solutions of this algorithm are typically of higher quality than those produced by other state-of-the-art approaches.

REFERENCES

1. R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining,"
2. C.C. Aggarwal and P.S. Yu, Privacy Preserving Data Mining: Models and Algorithms (Advances in Database Systems).
3. V.S. Verykios, A.K. Emagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," IEEE Trans. Knowledge and Data Eng.,
4. A. Gkoulalas-Divanis and V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," Proc. ACM Conf. Information and Knowledge Management (CIKM '06), pp. 748-757,
5. G.V. Moustakides and V.S. Verykios, personal comm., 2006.712 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009
6. Y.-H. Wu, C.-M. Chiang, and A.L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 29-42, Jan. 2007.
7. X. Sun and P.S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," Proc. Fifth IEEE Int'l Conf. Data Mining, pp. 426-433, 2005.
8. X. Sun and P.S. Yu, "Hiding Sensitive Frequent Itemsets by a Border-Based Approach," Computing Science and Eng., vol. 1, no. 1, pp. 74-94, 2007

