Privacy Preserving In Data Mining Using Sanitization algorithm

Arun T. Maheshwari[1], *Jasmine Jha*[2]
[1]*P.G. Student, Department of Computer Engineering, L.J Institute of Engineering and Technology, Ahmedabad, Gujarat, India*
[2]*Assistant Professor, Department of Computer Engineering, L.J Institute of Engineering and Technology, Ahmedabad, Gujarat, India*

ABSTRACT

Data mining is a process of analysing data from different perspectives and summarizing it into useful information.Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. However, the sharing of data has also raised a number of ethical issues. Some such issues include those of privacy, data security, and intellectual property rights. A number of algorithmic techniques have been designed for privacy-preserving data mining due to the wide proliferation of sensitive information on the internet. In this paper, we have proposed an tabular approach in which the output of data mining applications is sanitized for privacy-preservation purposes.

Keyword:*Privacy Preserving Data Mining, Document Sanitization.*

1. INTRODUCTION

Data mining is one of the core processes in knowledge discovery of databases [1]. Data mining research deals with the extraction of potentially useful information from large collections of data. Privacy issues are further exacerbated, now that the World Wide Web makes it easy for the new data to be automatically collected and added to databases. Data mining, with its promise to evidently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. The primary task in data mining is the development of models about aggregated data. Repositories of data contain sensitive information which must be protected against unauthorized access [1]. The protection of the confidentiality of this information has been a long-term goal for the database security research community and the government statistical agencies. In data mining, lattice based technique is usually used to evaluate how relevant a word in a document [3].

In this paper, an algorithm for data sanitization is proposed which is based on tabular form instead of lattice structure to efficiently hide the sensitive item sets with minimal side effects and also to be used for distributed data.

The rest of this paper is organized as follows. Section-II gives background about Sanitization techniques. Section-III describes proposed system, Section-IV shows Experimental result and Section-V covers conclusion and future work

2. RELATED WORK

Data mining is a powerful concept for data analysis and process of discovery interesting pattern from the huge amount of data, data stored in various databases such as data warehouse, World Wide Web, external sources [2]. Data mining tools aims to find useful patterns from large amount of data.Sanitization of a document involves removing sensitive information from the document, so that it may be distributed to a broader audience. Such sanitization is needed while declassifying documents involving sensitive or confidential information such as corporate emails, intelligence reports, medical records.

Data Sanitization Techniques [2][6]

Nulling Out:

Simply deleting a column of data by replacing it with NULL values is an effective way of ensuring that it is not in appropriately visible in test environments. Unfortunately it is also one of the least desirable options from a test database standpoint. Usually the test teams need to work on the data or at least a realistic approximation of it.

Masking Data:

Masking data means replacing certain fields with a Mask character (such as an X). This effectively disguises the data content while preserving the same formatting on front end screens and reports

Substitution:

This technique consists of randomly replacing the contents of a column of data with information that looks similar but is completely unrelated to the real details.

Shuffling Records:

Shuffling is similar to substitution except that the substitution data is derived from the column itself. Essentially the data in a column is randomly moved between rows until there is no longer any reasonable correlation with the remaining information in the row.

Number Variance:

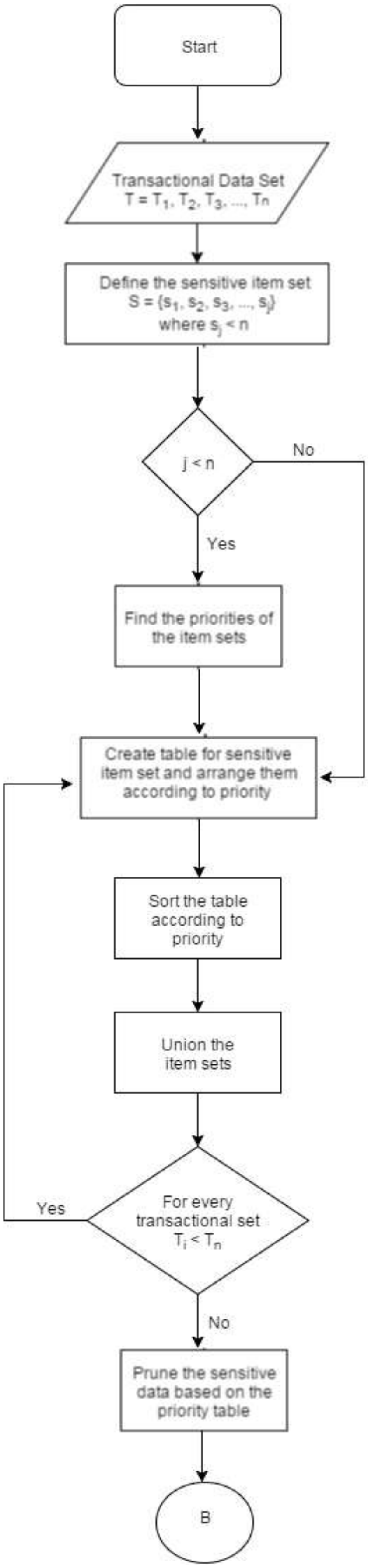
The Number Variance technique is useful on numeric data. Simply put, the algorithm involves modifying each number value in a column by some random percentage of its real value. This technique has the nice advantage of providing a reasonable disguise for the numeric data while still keeping the range and distribution of values in the column within viable limits.

3. PROPOSED WORK

With increasingly large available data on the internet, its confidentiality is at risk. Privacy Preservation data mining has emerged to address one of the side effects of data mining Technology. The threat to individual’s confidential information through data mining can get affected. There is an urgent need to be able to infer some mechanism to avoid the projection of all the sensitive information. Alteration of data, filtering of the data, blocking of the data are some of the approaches. Given specific rules to be hidden, the techniques involve is to hide only the given sensitive data

In this paper, we have proposed a sanitization algorithm that efficiently hides the sensitive information from database.

The flowchart for the proposed system is as follows:


Start
Transactional Data Set
T = T1, T2, T3, ..., Tn
Define the sensitive item set
S = {s1, s2, s3, ..., sj}
where sj < n
j < n
No
Yes
Find the priorities of the item sets
Create table for sensitive item set and arrange them according to priority
Sort the table according to priority
Union the item sets
For every transactional set
Ti < Tn
Yes
No
Prune the sensitive data based on the priority table
B

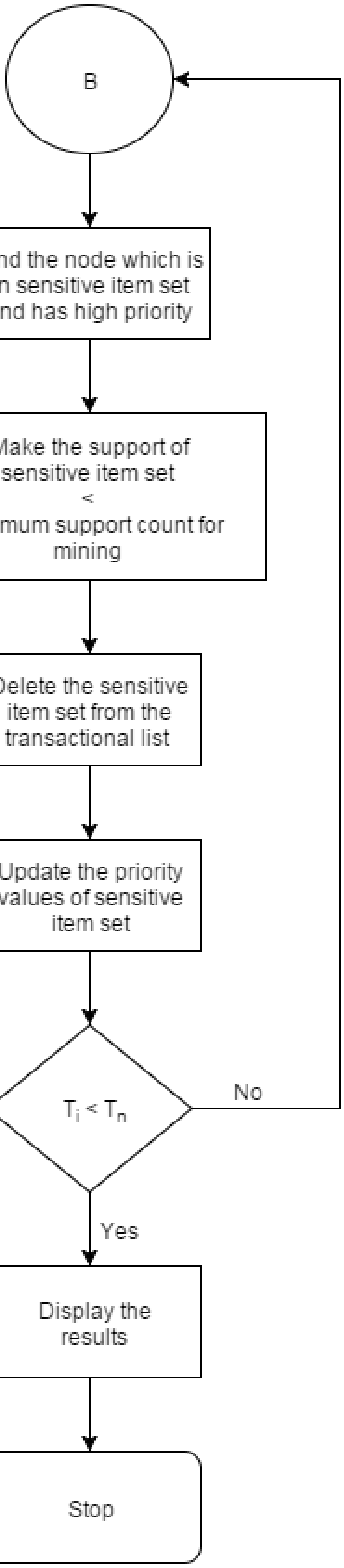

B
Find the node which is in sensitive item set and has high priority
Make the support of sensitive item set < Minimum support count for mining
Delete the sensitive item set from the transactional list
Update the priority values of sensitive item set
$T_i < T_n$
No
Yes
Display the results
Stop

Fig-1: System Flow

The proposed solution steps are as follows:

Input: A Transactional data set is taken for the input
SET T = {T1, T2,T3,T4,...,Tn}

Output: Sensitive Item sets are sanitized.

Step1: Define the sensitive item setS= {s1,s2,...,sj}, sj should be less than Tn

Step2: Find out priorities of the item sets

```
if j<n then
for i = T0 to Tn do
P(i) = |p(n)-s*n+1|
return (Ti,Pi)
end for
end IF
```

Step3: Create table for sensitive item set and their priority.

3-1: Arrange table based on the priority and individual transaction
3-2: Set the itemset and union if itemset occurs in union.
3-3: Repeat the step 3 for all transactional item set T

Step4: Prune the Sensitive data based on their priority from the table generated in step 3

4-1: Find the node which is in sensitive itemset and also have the high priority
4-2: make the support of sensitive item set less than minimum support count for mining.
4-3: delete the sensitive item set from the transactional list.
4-4: Update the priority values of sensitive item set.
4-5: Repeat step 4 until data sets have no sensitive item set.

Below Figure 2 shows an example of distorted dataset after sanitization and shows that sensitive item is distorted.

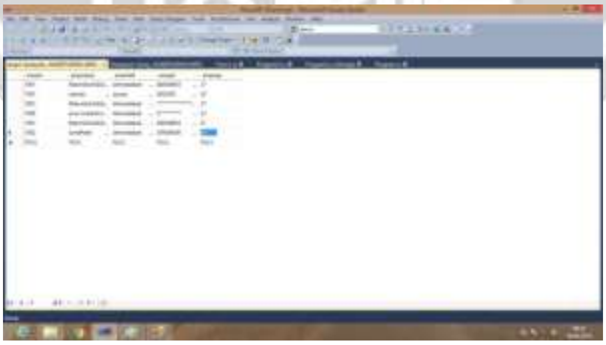


Fig-2: Distorted Database

4. RESULT ANALYSIS

In the experiments, the sensitive item sets and the user- specific minimum support thresholds were defined the same for the three proposed algorithms. The minimum support thresholds were set at 3%, for the database, and for the proposed approach, it can efficiently sanitize the sensitive terms. Table-1 shows the transactions on each node. Table-2 describes the item sets in each transaction and Table-3 describes the execution time for existing and proposed system. Finally, using Table-3 as the input we have achieved experimental results for the existing and proposed system.

Node	Transaction ID
S1 = {cfh}	T1,T3,T4,T8,T10
S2 = {af}	T1,T4,T6,T9,T10
S3 = {c}	T1,T3,T4,T5,T6,T7,T8,T10

Table-1: Transactions on Each Node

Tid	Item
1	a, b, c, d, f, g, h
2	a, b, d, e
3	b, c, d, f, g, h
4	a, b, c, f, h
5	c, d, e, g, i
6	a, c, f, i
7	b, c, d, e, f, g
8	a, d, e, f, i
9	a, c, e, f, h

Table-2: Item sets in Each Transaction

The execution times of the lattice-like approach in two datasets		
	BMS-Position	BMS-Webview
Lattice Approach	3,434.33	20.53
Table Approach	2468.33	NA

Table-3: Execution Time for Existing System and Proposed System

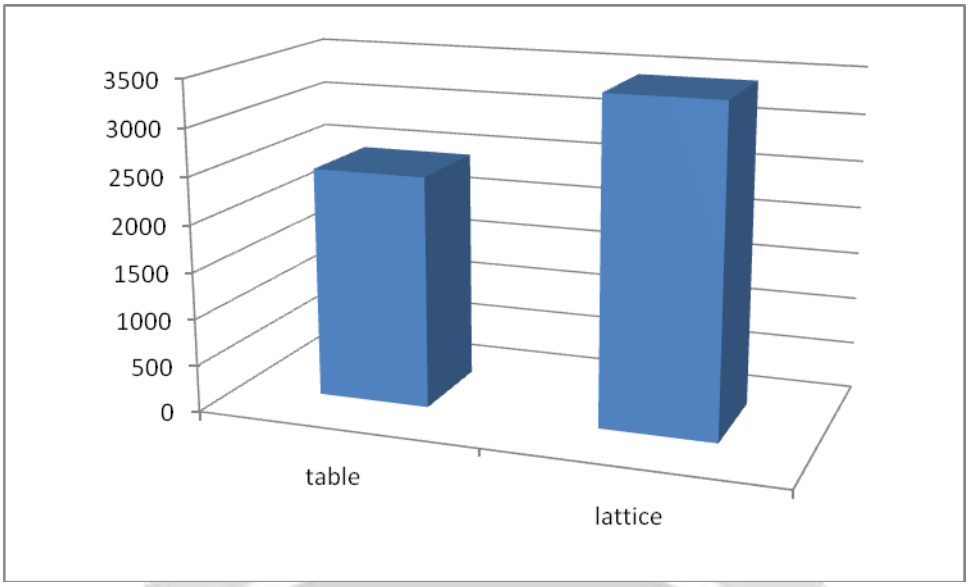


Fig-3: Comparison of Existing and Proposed System

5. CONCLUSION AND FUTURE WORK

Privacy of sensitive data is a critical issue in data mining. There should be some mechanism to overcome this issue. In this study we analyzed several of techniques to efficiently hide the sensitive data and defined a mechanism based on distortion to hide the data.In future experiments could be done for dataset containing images and media files and also same experiments can be done for remaining data like web view.

6. REFERENCES

[1]Montserrat Batet and David Sánchez, “Privacy Protection of Textual Medical Documents”, IEEE Data Mining, vol. 1, no. 1, pp. 4799-0913-, Feb. 2014

[2] SHYUE-LIANG WANG, RAJEEV MASKEY, AYAT JAFARI, TZUNG-PEI HONG,” Efficient Sanitization of Informative Association Rules with Updates”,IEEE Data mining, vol. 1, pp.4244-0555, 2006

[3] Stanley R. M. Oliveira , Osmar R. Zaiane “Protecting Sensitive Knowledge by Data Sanitization” IEEE Data Mining, vol. 4, no. 3, pp. 7695-1978, 2003

[4] Jerry Chun-Wei Lin, Tsu-Yang Wu, Philippe Fournier-Viger, Guo Lin, Tzung-Pei Hong and Jeng-Shyang Pan “A Sanitization approach of Privacy Preserving utility mining ” Springer International Publishing Data Mining, DOI: 10.1007/978-3-319-23207-2_6

[5] R.Hemalatha, M.Elamparithi “Privacy Preserving Data Mining Using Sanitizing Algorithm” International Journal of Computer Science and Information Technologies Data Mining, Vol. 6, no. 3, 4174-4179, 2015

[6] Tzung-Pei Hong , Chun-Wei Lin, Kuo-Tung Yang , and Shyue-Liang Wang, “A Lattice-based Data Sanitization Approach”, IEEE Data Mining, Vol. 1, no. 1, pp. 4577-0653, 2011

[7] Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, “Efficient Techniques for Document Sanitization” ACM 978-1-59593-991-3, 2008