An Incremental Sanitization Approach in Dynamic Databases

Cynthia Selvi P
Associate Professor, Dept. of Computer Science, KNGA College(W), Thanjavur 613007
Affiliated to Bharathidasan University
Tiruchirapalli, TamilNadu, India.
Email: pselvi1501@gmail.com

Mohamed Shanavas A.R
Associate Professor, Dept. of Comuter Science, Jamal Mohamed College, Tiruchirapalli 620 020
Affiliated to Bharathidasan University
Tiruchirapalli, TamilNadu, India.
Email: vas0699@yahoo.co.in

Abstract— In most organizations the historical dataset is dynamic which means it is periodically updated with fresh data. In such an environment, when the counterpart organizations need to share their information for a decision making strategy in a periodical way, the protection of sensitive knowledge need to be done as a recurring activity, especially if the dataset has been significantly updated. In this scenario, it would not be effective to perform the process of preserving sensitive knowledge in the updated database as a whole repeatedly. Hence this work proposes an incremental approach for knowledge protection in dynamic databases.

Keywords-Data mining, Incremental Sanitization, Restricted Patterns

I. INTRODUCTION

Advanced improvements in information science have resulted in the rapid accumulation of large amount of data and this requires efficient management of very large databases and quick retrieval of useful information. Data mining or knowledge discovery techniques are normally used to discover useful information or patterns from data warehouses. These techniques represent an important field of research and have been applied extensively to several areas [1], including financial analysis, market research, industrial retail and decision support. Pattern discovery is highly an exploratory activity which operates as a feedback process where new discoveries are searched with the guidance of previously discovered patterns.

In this context, it is worthwhile to consider the use of previously discovered patterns for discovering new ones instead of processing the updated dataset from scratch. Eventually, it may be necessary to mine patterns from the entire updated dataset or just from the increment or from both. In this context, it is worthwhile to consider the possibility of incremental mining which attempts to use the previously mined patterns in order to perform the new mining operation more efficiently in the original dataset which is very larger than the fresh data in many orders of magnitude. However, the incremental mining algorithms are usually able to process the original data very quickly. These techniques can result in substantial benefits, especially when the dataset grows exponentially.

In the real world where enormous amount of data grow steadily, some old association rules can become useless, and new databases may give rise to some implicitly valid patterns or rules. Hence, updating rules or patterns is also important. When a database has been updated, some interesting new rules may be introduced and some existing rules may become obsolete. Thus, the maintenance of the association rules in the updated database is a significant problem in order to keep track of the generation of new rules and the invalidation of some existing rules. Conceptually when this updated dataset is to be shared in a collaborative environment, the data mining techniques may pose security problems and lead to privacy concerns.

Privacy Preserving Data Mining(PPDM) techniques have thus become a critical research issue for hiding confidential or secure information. The goal in PPDM is to hide the sensitive patterns with the minimal side effects and this problem has been addressed by many researchers [2-10]. Atallah et al.[2] referred this task as sanitization. This article addresses the task of incremental sanitization of the work presented in [9], which made an attempt to perform sanitization in the static transactional databases.

II. Incremental Sanitization

Many researchers have proposed several algorithms to sanitize the sensitive patterns but most of them are static in nature. In real world, data is growing frequently and it is being dynamic in the sense that the users may periodically or occasionally insert data to or remove data from the database. The static algorithms may not work efficiently whenever any change happens to the original database. Moreover existing support threshold for interestingness measure as well as association rule may be invalidated. If support is static(once it is set to very high or low), important information loss may occur. Hence support should be flexible, based on real time event to avoid loss of important information.

When the original database is inserted with new transactions in the case of dynamic database, not only some existing association rules may be invalidated but also some new association rules may be discovered. This is the case because frequent itemsets can be changed after inserting new transactions into a dynamic database. Therefore, an association rule discovery algorithm for a dynamic database has to maintain frequent itemsets when new transactions are inserted into the dynamic database. However, we also endeavor to maintain intemsets that are false positive itemset, ie infrequent itemsets that have the potential to become frequent itemsets when new transactions are added to original database.

Moreover, static methods cannot be applied to mine a *publication-like* database(a transaction database where each item involves an individual exhibition period) efficiently. In many situations, new information is more important than old information like publication database, stock transactions, grocery markets, or web-log records. Consequently, a frequent itemset in the incremental database is also important even if it is infrequent in the updated database. Traditional static sanitization techniques ignore the exhibition period of each item and use a fixed and unique minimum support threshold. This measure may not be suitable for new items. The field of information science has achieved an impressive ability to store data. Furthermore, our skills and interest to manipulate them are also remarkable. There has been a wide variety of research going on in the field of PPDM and most of the techniques are implemented for static data. As the world is filled with dynamic data which grows rapidly than what we expect, this article focus on privacy criteria in dynamic dataset and present algorithm that sanitize data to make it secured for release while preserving useful information.

III. Proposed Algorithm

A. *Notations:*

D^o – Old Dataset

D^n – New Dataset (to be appended with D^o)

Rp – Restricted Patterns

~Rp – UnRestricted Patterns

B. *Methodology :*

Table-I. Various Strategies of Sanitization in Dynamic Databases

D^n D^o	Rp	~Rp
Rp	Strategy-1	Strategy-3
~Rp	Strategy-2	Strategy-4

Strategy-1 : Condition-a: The rules considered for sanitization in D^n is the same as that of the old Dataset D^o.

Condition-b: Some new rules are considered for sanitization in the new dataset D^n, in addition to the ones considered in the old dataset D^o.

Strategy-2: Completely new rules are to be considered for sanitization.

Strategy-3: The rules considered for sanitization in the old dataset D^o are not restricted when it is updated with the new dataset D^n .

Strategy-4: The rules are not restricted in both old and new datasets.

In the above situations, strategy-1 and strategy-2 needs sanitization to be done in either one of the datasets or both, whereas strategy-3 and strategy-4 needs no revisions in the sanitization process.

C. Algorithm:

```
Input : Old dataset-$D^o$, New dataset $D^n$, Restricted Patterns.
Output : Sanitized & Merged dataset of $D^o$ & $D^n$.
Begin
{
  Select Rules;
  Check with $D^o$;
  Switch
  {
    Case-i: / Strategy-1a /
        Sanitize $D^n$ ;
    Case-ii: / Strategy-1b /
        Sanitize $D^o$ & $D^n$ in parallel;
    Case-iii: / Strategy-2 /
        Sanitize $D^o$ & $D^n$ in parallel;
  }
  $D' \leftarrow D^o \cup D^n$;
}
```

IV. Implementation

The algorithm was tested for real dataset T10I4D100K[11] and the Patterns to be hidden are given in the table-I. The test run was made on Intel core i5 processor with 2.3 GHz speed and 4GB RAM operating on 32 bit OS; The implementation of the proposed algorithm was done with windows 7-Netbeans 6.9.1-SQL 2005. The frequent patterns were obtained using Matrix Apriori[12] and the algorithm presented in [9] is used for sanitization.

Table-II. Details of Restricted Patterns

	Str-1a (Same)	Str-1b (Combined)	Str-2 (New)
R1	39,48	41,36	41,36
R2	39,48,41	39,38	38,36
R3	39,41	39,170	48,65
R4	39,48,38,41	39,41,32	38,170
R5	39,38	39,41	39,170

The effectiveness of this approach is studied based on the following measure and the execution time observed for the sanitization process for the different strategies are tabulated(Table-II) and presented using graph(Fig.1)

A. Effectiveness Measure :

The effectiveness of this approach is measured in terms of time gain which is defined as given below: The readings observed for various strategies are given in Table-III and the corresponding graph is shown Fig.1.

Time Gain = $((T_1 – T_2)/T1)$ X 100,

where T_1 = Sanitization time on D^o & D^n combined together,

T_2 = Sanitization time (D^n) for strategy-1a &

Sanitization time [Max (D^o, D^n)] for strategy-1b & strategy-2

Table-III. Time Gain with Incremental Sanitization

No.of Transactions in D^o & D^n		5K + 1K	5K + 3K	5K + 5K
Str-1a (Same)	D^o U D^n	24.95	19.24	39.42
	D^n Incremental	5.62	7.54	10.49
	Time Gain (%)	**77.47**	**60.81**	**73.39**
Str-1b (Combined)	D^o U D^n	7.78	14.99	21.98
	D^o & D^n parallel(Max)	4.00	10.14	20.78
	Time Gain (%)	**48.59**	**32.35**	**5.46**
Str-2 (New)	D^o U D^n	9.54	10.20	12.08
	D^o & D^n parallel(Max)	8.36	7.71	8.61
	Time Gain (%)	**12.37**	**24.41**	**28.93**

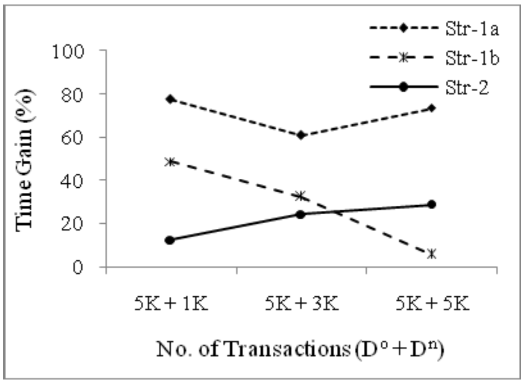

Fig.1. Time Gain

V. CONCLUSION

The proposed incremental approach for preserving sensitive patterns has highly reduced the execution time for all possible strategies; especially when the same set of patterns are to be protected in the existing and the new dataset to be updated the time gain is more than 70%. Hence the incremental sanitization would improve the performance issues like time, storage and effort.

REFERENCES

[1] Yubo, J., Yuntao, D., Yongli, W.: An Incremental Updating Algorithm for Online Mining Association Rules. In: 2009 International Conference on Web Information Systems and Mining (2009)

[2] Atallah M, Bertino E, Elmagarmid A, Ibrahim M and Verykios V " Disclosure Limitation of Sensitive Rules", In Proc. of IEEE Knowledge and Data Engineering Workshop, pages 45–52, Chicago, Illinois, November 1999.

[3] Dasseni E, Verykios V.S, Elmagarmid A.K & Bertino E, "Hiding Association Rules by Using Confidence and Support", In Proc. of the 4th Information Hiding Workshop, pages 369– 383, Pittsburg, PA, April 2001.

[4] Saygin Y, Verykios V.S, and Clifton C, "Using Unknowns to Prevent Discovery of Association Rules", SIGMOD Record, 30(4):45–54, December 2001.

[5] OliveiraS.R.M, and Zaiane O.R, "Privacy preserving Frequent Itemset Mining", in the Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining, Pages 43-54, Maebashi City, Japan, December 2002.

[6] OliveiraS.R.M, and Zaiane O.R, "An Efficient One-Scan Sanitization for Improving the Balance between Privacy and Knowledge Discovery", Technical Report TR 03-15, June 2003.

[7] Yildz B, and Ergenc B, "Hiding Sensitive Predictive Frequent Itemsets", Proceedings of the International MultiConference of Engineers and Computer Scientists 2011, Vol-I.

[8] Cynthia Selvi P., Mohamed Shanavas A.R.,"An Improved Item-based Maxcover Algorithm to protect Sensitive Patterns in Large Databases", International Organization of Scientific Research - Journal of Computer Engineering(IOSRJCE) ISSN : 2278-0661 (Impact Factor 1.686)– Vol.14, Issue 4, Oct 2013, Pages 1-5.

[9] Cynthia Selvi P., Mohamed Shanavas A.R., "Output Privacy Protection with Pattern-Based Heuristic Algorithm", International Journal of Computer Science & Information Technology(IJCSIT) Vol 6, No 2, April 2014, DOI:10.5121/ijcsit.2014. 6210, Pages 141 – 152.

[10] Cynthia Selvi P., Mohamed Shanavas A.R., “Towards Information Privacy Using Transaction-Based Maxcover Algorithm”, World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques): 06-11, 2014, ISSN 1818-4952, © IDOSI Publications, 2014,DOI:10.5829/idosi.wasj.2014.29. dmsct.2, Pages. 06-11

[11] The Dataset used in this work for experimental analysis was generated using the generator from IBM Almaden Quest research group and is publicly available from http://fimi.ua.ac.be/data/.

[12] Pavon J, Viana S, Gomez S, “Matrix Apriori: speeding up the search for frequent patterns,” Proc. 24th IASTED International Conference on Databases and Applications, 2006, pp. 75-82.