Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.349 – 357

RESEARCH ARTICLE

A Research on Privacy Preserving Data Mining Using Heuristic Approach

Maulesh Chhatrapati[1], Shilpa Serasiya[2]

[1]M.E in C.E, Kalol Institute of Technology & Research Centre, India

[2]Asst. Prof. in Kalol Institute of Technology & Research Centre, India

[1] cmaulesh@gmail.com; [2] shilpaserasiya@gmail.com

Abstract— Data mining is the process of identifying patterns from large amount of data. Association rule mining aims to discover dependency relationships across attributes. It may also disclose sensitive information. With extensive application of data mining techniques to various domains, privacy preservation becomes mandatory. It has become a very important area of concern but still this branch of research is in its infancy .People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information. Association rule hiding is one of the techniques of privacy preserving data mining to protect the sensitive association rules generated by association rule mining. There are many approaches to hide association rule. In this paper Efficient Heuristic approach method is proposed which is more effective to hide association rule. This paper adopts heuristic approach for hiding sensitive association rules. The proposed technique makes the representative rules and hides the sensitive rules. The objective of this algorithm is to extract relevant knowledge from large amount of data, while protecting at the time sensitive information. In this paper we also focused to hide multiple sensitive items without affecting other sensitive items.

Keywords— heuristic approach, Minimum Confidence, Minimum Support, Item set, Association rules.

I. INTRODUCTION

Data Privacy preserving data mining (PPDM) is a novel research direction in Data Mining (DM), where DM algorithms are analysed for the side-effects they incur in data privacy. The main objective of PPDM is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [1]. In DM, the users are provided with the data and not the association rules and are free to use their own tools; So, the restriction for privacy has to be applied on the data itself before the mining phase. For this reason, we need to develop mechanisms that can lead to new privacy control systems to convert a given database into a new one in such a way to preserve the general rules mined from the original database. The procedure of transforming the source database into a new database that hides some sensitive patterns or rules is called the sanitization process[2]. To do so, a small number of transactions have to be modified by deleting one or more items from them or even adding noise to the data by turning some items from 0 to 1 in some transactions. The released database is called the sanitized database. On

one hand, this approach slightly modifies some data, but this is perfectly acceptable in some real applications [3, 4].

This study mainly focus on the task of minimizing the impact on the source database by reducing the number of removed items from the source database with only one scan of the database. Section-2 briefly summarizes the previous work done by various researchers; In Section-3 preliminaries are given. Section-4 states some basic definitions and of which definition 5 is framed by us which is used in the proposed heuristic based algorithm. In Section-5 the proposed algorithm is presented with illustration and example. As the detailed analysis of the experimental results on large databases is under process, only the basic measures of effectiveness is presented in this paper, after testing the algorithm for a sample generated database.

II. INTRODUCTION TO PRIVACY PRESERVING DATA MINING

In Privacy preserving has originated as an important concern with reference to the success of the data mining. Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very reluctant to share their sensitive information. This may lead to the inadvertent results of the data mining. Within the constraints of privacy, several methods have been proposed but still this branch of research is in its infancy.

In figure 1, framework for privacy preserving Data Mining is shown. Data from different data sources operational systems are collected and are pre-processed using ETL tools. This transformed and clean data from Level 1 stored in the data warehouse. Data in data warehouse is use for mining. In level 2, data mining algorithms are used to fin patterns and discover knowledge from the historical data after mining privacy preservation techniques are used to protect data from unauthorized access. Sensitive data of an
Individual can be prevented from being misused.

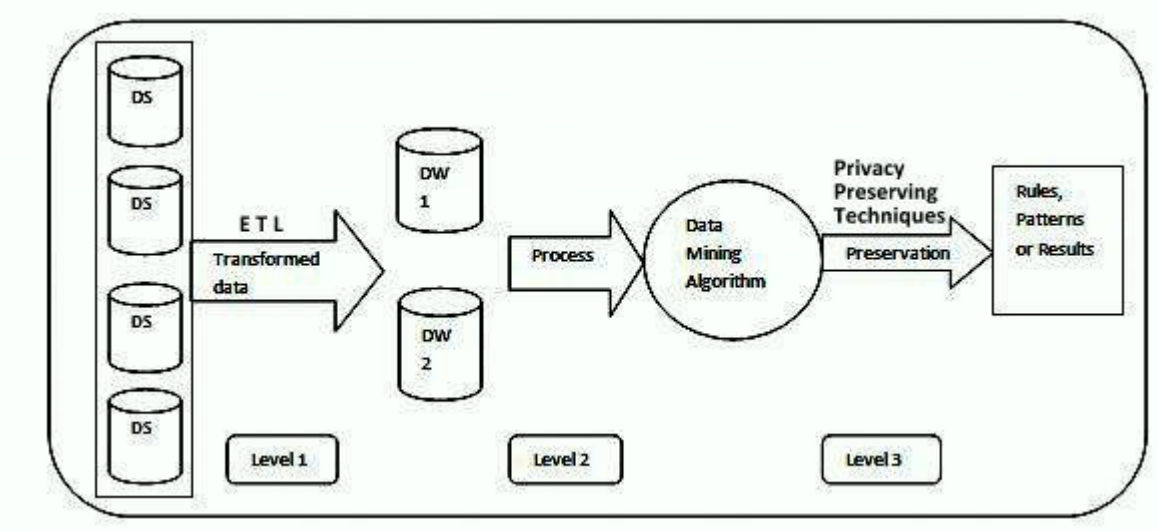

Figure 1: Framework of privacy preserving data mining [3]

III. ASSOCIATION RULE MINING

In a Let I = {i1,…., in} be a set of items. Let D be a database which contains set of transactions. Each transaction t _ D is an item set such that t is a proper subset of I. As

transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X_Y, where X and Y are subsets of I and X_Y= Ø. The support of rule X_Y can be calculated by the following equation: Support(X_Y) = |X_Y| / |D|, where |X_Y| denotes the number of transactions containing the item set XY in the database, |D| denotes the number of the transactions in the database D. The confidence of rule is computed by Confidence (X_Y) = |X_Y|/|X|, where |X| is number of transactions in database D that contains item set X. A rule X_Y is strong if support (X_Y) _ min_support and confidence (X_Y) _ min_confidence, where min_support and min_confidence are two given minimum thresholds.

Association rule mining algorithms calculate the support and confidence of the rules. The rules having support and confidence higher than the user specified minimum support and confidence are retrieved. Association rule hiding algorithms prevents the sensitive rules from being revealed out. The problem can be declared as follows "Database D, minimum confidence, minimum support are given and a set R of rules are mined from database D. A subset SR of R is denoted as set of sensitive association rules.SR is to be hidden. The objective is to modify D into a database D' from which no association rule in SR will be mined and all non sensitive rules in R could still be mined from D'.

IV. APPROACHES FOR ASSOCIATION RULE HIDING

Association rule hiding algorithms can be divided into three distinct approaches. They are heuristic approaches, border-revision approaches and exact approaches.

A. Heuristic Approach

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from 1's to 0's and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

B. Border Revision Approach

Border revision approach modifies borders in the lattice of the frequent and infrequent item sets to hide sensitive association rules. This approach tracks the border of the non sensitive frequent item sets and greedily applies data modification that may have minimal impact on the quality to accommodate the hiding sensitive rules. Researchers proposed many border revision approach algorithms such as BBA (Border Based Approach), Max– Min1 and Max-Min2 to hide sensitive association rules. The algorithms uses different techniques such as deleting specific sensitive items and also attempt to minimize the number of non sensitive item sets that may be lost while sanitization is performed over the original database in order to protect sensitive rules.

C. Exact Approach

Third class of approach is non heuristic algorithm called exact, which conceive hiding process as constraint satisfaction problem. These problems are solved by integer programming. This approach can be concerned as descendant of border based methodology.

V. PROBLEM STATEMENT

Data mining represents a wide range of tools and techniques to extract useful information which can contain sensitive information from a large collection of data. Data should be manipulated or distorted in such a way that sensitive information cannot be discovered through data mining techniques. Sensitive information has to be protected against unauthorized access. The major challenge faced is better balancing the confidentiality of the disclosed with the legitimate needs of the data user. The proposed approach is based on modification of database transactions to hide multiple sensitive item without affecting other sensitive items.

VI. ANALYSIS OF EXISTING TECHNIQUES

1. In Distortion Based Technique (Proposed By Veryki- os Et Al, Etc.) authors propose strategies and a suite of algorithms for hiding sensitive knowledge. In order to achieve this, transactions are modified by removing few items, or inserting new items depending on the hiding strategy.
2. The distortion based Technique (Proposed by shyue-liang wang et al.) hides certain specific items that are sensitive. In this technique, two algorithms are proposed to modify data in the Dataset. If the sensitive item is on the LHS of the rule then the first algorithm increases its support. If the sensitive item is on the right of the rule then the second algorithm decreases its support.

In (1) author tries to hide every rule without checking if rules can be pruned after some transactions have been changed. In (2) the author hides all the rules which contain sensitive items either in the left or in the right. Two different algorithms are applied over the data. The first algorithm hides association rules with sensitive items on the LHS and the second one for sensitive items on the RHS. It takes more number of passes to prune all the rules containing sensitive items.

VII. PROPOSED WORK

In this work, In order to hide an association rule, $X \rightarrow Y$, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). To decrease the confidence of a rule, we can either (1) increase the support o of X, the left hand side of the rule, but not support of $X \rightarrow Y$, or (2) decrease the support of the item set $X \rightarrow Y$ .For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of $X \rightarrow Y$. To decrease support of an item, we will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction.

Based on these two concepts, we propose a new association rule hiding algorithm for hiding sensitive items in association rules. In our algorithm, a rule $X \rightarrow Y$ is hidden by decreasing the support value of $X \rightarrow Y$ and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly. This algorithm first tries to hide the rules in which item to be hidden i.e., X is in right hand side and then tries to hide the rules in which X is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

A. **FLOW OF WORK:**

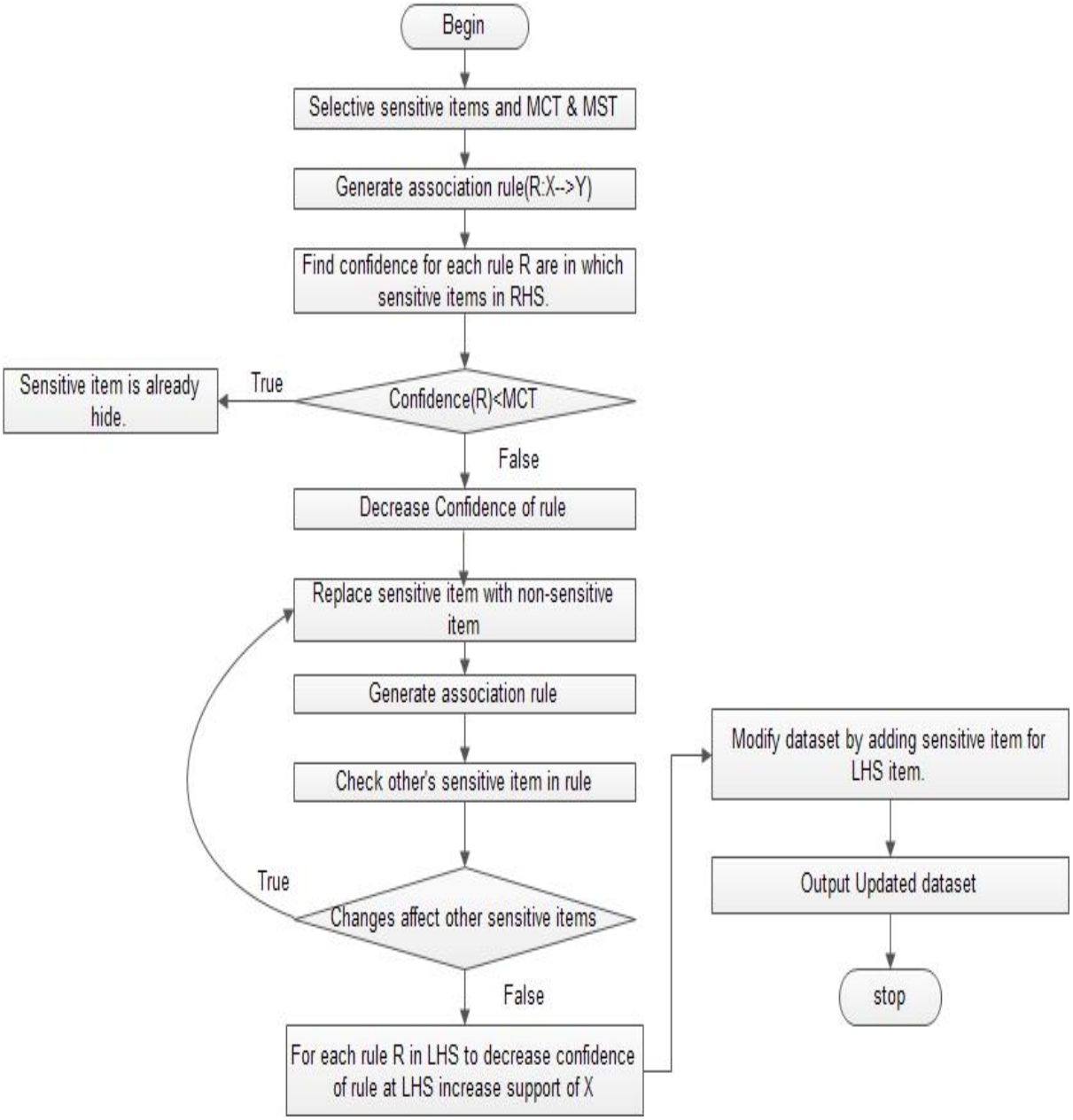

Figure 2 Flow of Proposed Work .

B. **THE PROPOSED ALGORITHM:**

1. Begin
2. Generate all possible rule from given items X;
3. Compute confidence of all the rules for each hidden item H, compute confidence of rule R.
4. For each rule R in which H is in RHS
   4.1 If confidence (R) < MCT, then
   Go to next 2-itemset;
   Else go to step 5
5. Decrease Support of RHS item H.
   5.1 Find T=t in D fully supports R;

5.2 While (T is not empty)
5.3 Check every instance in data for sensitive items
5.4 Get non-sensitive item for replacement
5.5 Create new instance with updated value
5.6 Remove odd instance from data set
5.7 Add new instance to data set
End While
•check other item's rule those are included in sensitive list
•if changes affect other Item repeat process of change with other replaces.
Else
6. Compute confidence of R;
7. For each rule R in which is in LHS
8. Increase Support of LHS;
9. While (T is not empty)
10. Modify t by adding sensitive item;
11. Remove and save the first transaction t from T; End While
12. Compute confidence of R;
End For;
End Else;
End For;
13. Output updates D, as the transformed D;

The framework of the proposed approach is shown in figure

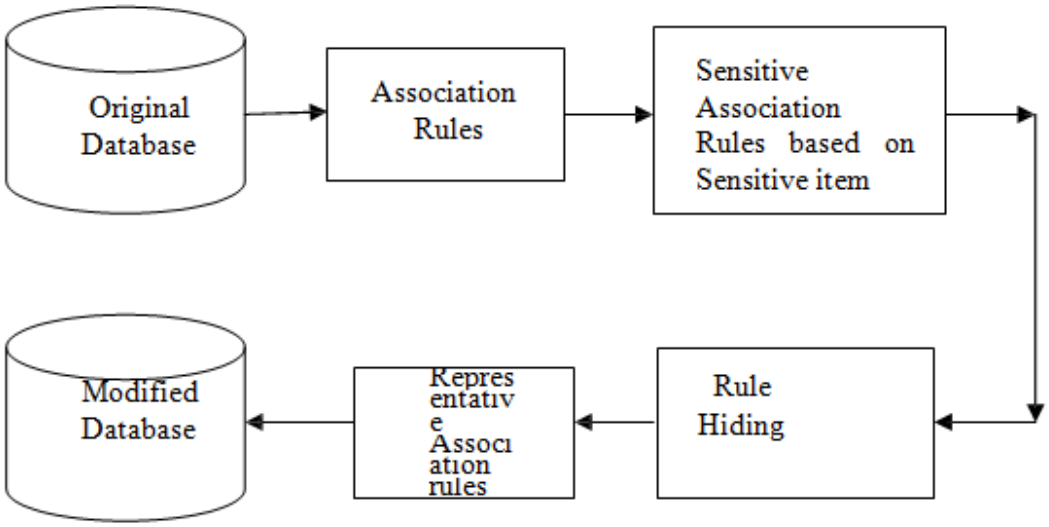

FIGURE 3: ASSOCIATION RULE HIDING FRAMEWORK

VIII. EXPERIMENTAL WORK

This section describes the current status of implementation along with appropriate screen shots.
First, the enter the min_sup and min_conf. Then enter the sensitive items which we want hide fron the users.

Second, Generate the association rules according to min_sup and min_conf.
Third, Replace the sensitive items another items.
And last Update the association rules in which sensitive items are hidden fron the user.

A. **Enter Min_SUP and Min_CONF:**

1. ENTER MIN_SUP & MIN_CONF FOR SUPER MARKET DATA SET:

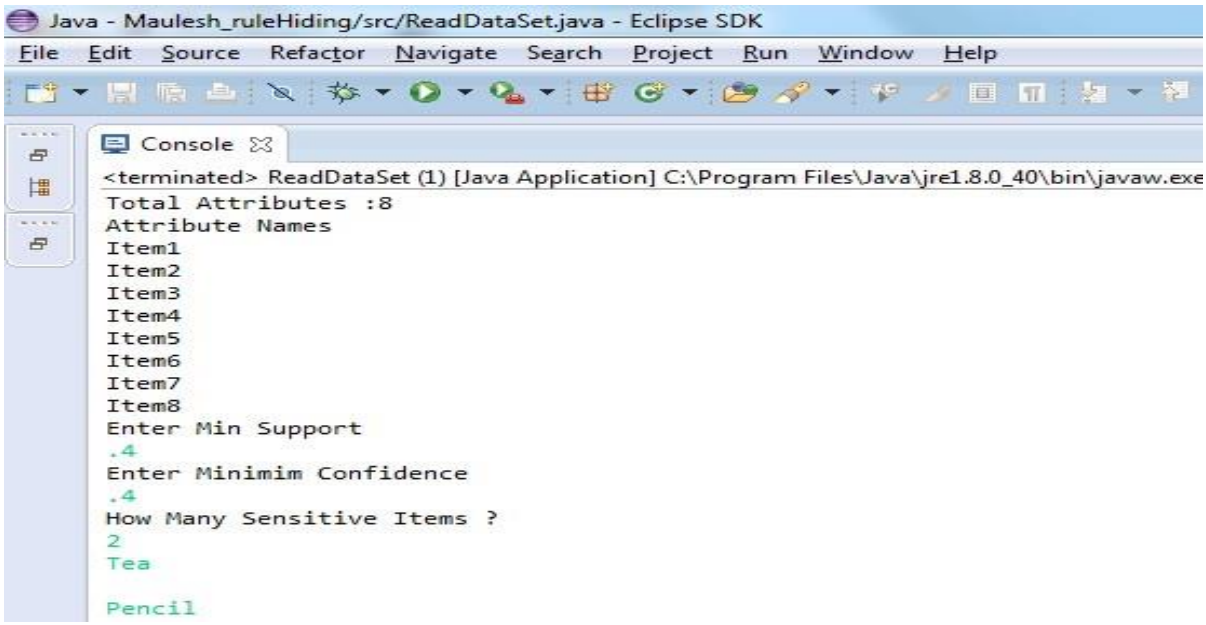

Figure 4 min_sup & min_conf

B. **Generated association rule according to sensitive item:**

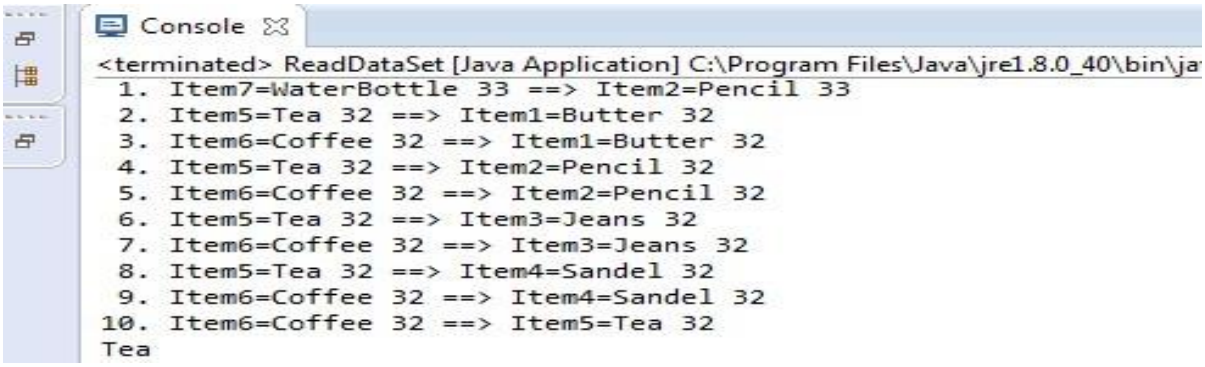

Figure 5 Generate association rule

C. **Replace sensitive item with non-sensitive item**

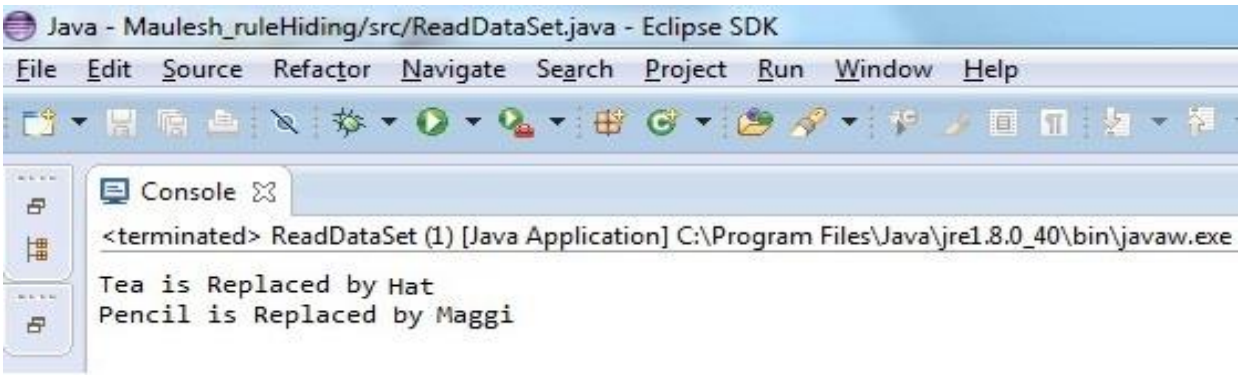

Figure 6 Replace sensitive items

D. **Updated Association Rules**

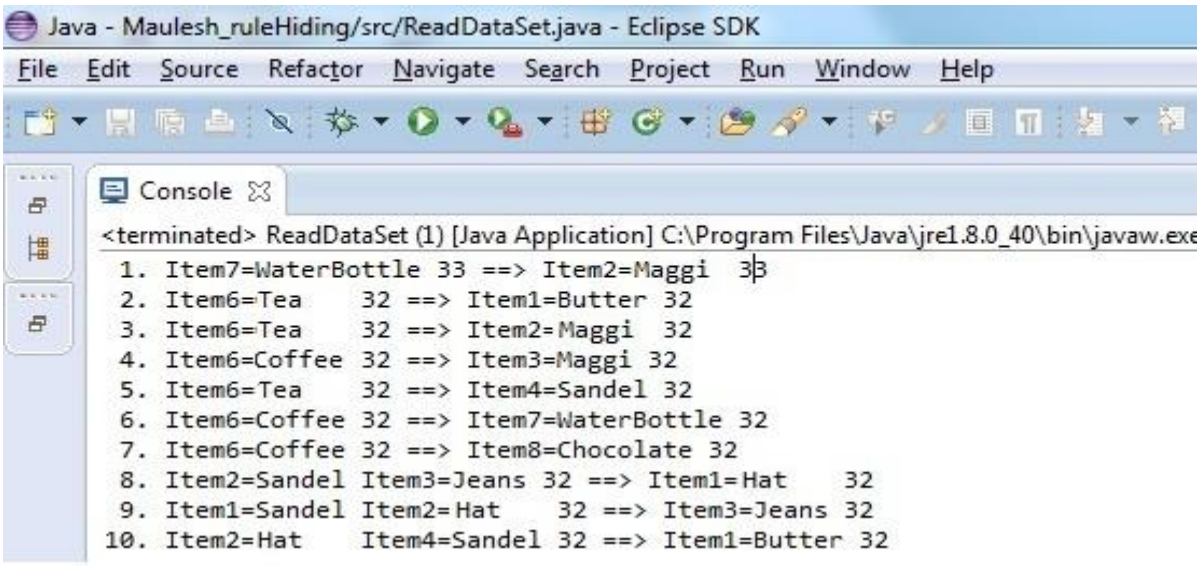

Figure 7 Updated association rules

IX. CONCLUSION AND FUTURE WORK

In this paper, we hide the sensitive item from the association rule with the use of heuristic approach. we conclude from this approach that this approach is best for hide multiple sensitive items in the association rules and dataset. Further the efficiency of the algorithm will be analyzed and improved by reducing the side effects. Further research is in progress to evolve a method which can avoid the computational overhead associated with confidence of the rules. ” Develops the best algorithm to hide multiple sensitive data in timely manner and accuracy is my main goal.”

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their precious comments and suggestions that contributed to the expansion of this work.

REFERENCES

[1] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in *proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.*

[2] Shyue-Liang Wang, Yu-Huei Lee, Billis S., Jafari, A. "Hiding sensitive items in privacy preserving association rule mining", IEEE International Conference on Systems, Man and Cybernetics, Volume 4, 10-13 Oct. 2004 Page(s): 3239 - 3244 .

[3] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012.

[4] A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in proceedings of International Symposium on Computer Science and Society, IEEE 2011.

[5] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.

[6] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita"A Review on Privacy Preserving Data Mining: Techniques and Research Challenges" Shweta Taneja et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014.

[7] Kasthuri S1 and Meyyappan T2"Hiding Sensitive Association Rule Using Heuristic Approach"International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.1, January 2013

[8] Dharmendra Thakur, Prof. Hitesg gupta. Techniques,"An Exemplary Study of Privacy Preserving Association Rule Mining Techniques "International Journal Of Advanced Research In Computer Science And Software Engineering,Volume 3, Issue 11, November 2013.

[9] Supriya S. Borhade1 , Bipin B.Shinde2 "Privacy Preserving Data Mining Using Association Rule With Condensation Approach" Supriya S. Borhade et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014.

[10] 1M.Mahendran, 2Dr.R.Sugumar, 3K.Anbazhagan, 4R.Natarajan "An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach" international journal of advanced research in computer and communication engineering vol. 1, issue 9, November 2012.

[11] "Privacy Preserving Data Mining: Models and Algorithms" BOOK-Edited by: Charu C. Aggarwal and Philip S. yu.

[12] Archana Tomar1, Vineet Richhariya2, Mahendra Ku. Mishra3 "A Improved Privacy Preserving Algorithm Using Association Rule Mining In Centralized Database" International Journal of Advanced Technology & Engineering Research (IJATER) .

