Intelligent Information Management, 2011, 3, 219-229
doi:10.4236/iim.2011.36027 Published Online November 2011 (http://www.SciRP.org/journal/iim)

Hiding Sensitive XML Association Rules with Supervised Learning Technique

Khalid Iqbal[1], Sohail Asghar[2], Abdulrehman Mirza[3]
[1]*Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science & Technology, Islamabad, Pakistan*
[2]*Associate Professor, Center of Research in Data Engineering, Mohammad Ali Jinnah University, Islamabad, Pakistan*
[3]*Associate Professor Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia*
E-mail: *mykhalidiqbal@yahoo.com, sohail.asghar@jinnah.edu.pk, amirza@ksu.edu.sa*
Received July 7, 2011; *revised August* 17, 2011; *accepted August* 30, 2011

Abstract

In the privacy preservation of association rules, sensitivity analysis should be reported after the quantification of items in terms of their occurrence. The traditional methodologies, used for preserving confidentiality of association rules, are based on the assumptions while safeguarding susceptible information rather than recognition of insightful items. Therefore, it is time to go one step ahead in order to remove such assumptions in the protection of responsive information especially in XML association rule mining. Thus, we focus on this central and highly researched area in terms of generating XML association rule mining without arguing on the disclosure risks involvement in such mining process. Hence, we described the identification of susceptible items in order to hide the confidential information through a supervised learning technique. These susceptible items show the high dependency on other items that are measured in terms of statistical significance with Bayesian Network. Thus, we proposed two methodologies based on items probabilistic occurrence and mode of items. Additionally, all this information is modeled and named PPDM (Privacy Preservation in Data Mining) model for XARs. Furthermore, the PPDM model is helpful for sharing markets information among competitors with a lower chance of generating monopoly. Finally, PPDM model introduces great accuracy in computing sensitivity of items and opens new dimensions to the academia for the standardization of such NP-hard problems.

Keywords: XML Document, Association Rules, Bayesian Network, PPDM Model, NP-Hard, K2 Algorithm

1. Introduction

The combination of statistics, artificial intelligence, and database makes data mining and knowledge discovery a hot research area. The purpose of such development is to find the formerly unknown, potentially useful knowledge, rules, or models [1] from a large set of data. This useful knowledge can be comprehended and interpreted for making decisions [2]. For that reason, it can be observed in insurance agencies [2,3], web mining [2,4], financial institutes [2,5] and marketing contexts [2,6,7]. Moreover, this application is obtained by the open use of data with presupposition of data mining and knowledge discovery. The problem in such use of data causes complications in the real world. Thus, data mining techniques disclose the private information which should be hidden. Consequently, privacy preserving data mining comes forward with its great importance in such perspective. Furthermore, the reasons behind the privacy preserving data mining are the existence of typical problems in data mining and knowledge discovery including clustering, classification, sequential patterning and association rule mining [8]. Therefore, privacy preserving data mining and knowledge discovery must be developed with an aim to address these problems. The problem of privacy preserving data mining can be illustrated with an example as given in [9].

"*Suppose there are two Supplier-X and Supplier-Y who supply milk to the supermarket. If the transactional database of the supermarket is released, both the suppliers can mine the association rules. The objective of such mining of association rules is to promote sales as well as the supply of goods. Ultimately, supplier-X can go for the lower price of goods which is in the favor of supermarket.*

In either case, supplier-X finds association rules concerning to the supplier-Y's milk. Thus, supplier-X can introduce a scheme by giving discount of d% for the promotion of his/her milk selling together with tea. Slowly but surely, the amount of sales of supplier-Y's milk will go down without lowering the price to the supermarket. Consequently, supplier-X will build monopoly without lowering the price. In such aspect, the released database is dreadful for supermarket."

Currently, standardization issues [10] are being focused in privacy preserving data mining but solution to the right privacy is NP-hard [11]. Therefore, our primary focus is to identify the sensitive item(s) with the analysis of side effects such as new rules [12-14], lost rules [12, 13] and hidden rules [12-16] incurred in XML association rules privacy preservation. Hence, we presented a model for privacy preserving data mining which quantifies sensitivity through Bayesian Network over vertically partitioned data. After quantification, XML association rules are generated over horizontally partitioned data through apriori algorithm.

2. Literature Review

Dafa-Alla *et al.* [18] employed PRBAC in order to protect sensitive information while different users are working in sessions according to their roles with variable privileges. Such control of sensitivity is dependent upon relational format. This technique does not identify Sensitive Objects (SOBS) but adding it to the model for preservation before publishing of data. In contrast to this technique, Dasseni *et al.* [15] investigated confidentiality issues related to association rules. In this approach, a rule is characterized as sensitive which has the confidence measure above the certain privacy threshold. Thus, a simple selection condition which can be applied to protect the sensitivity of objects is randomly distributed in the transactions. Thus, such sensitive objects cannot be selected as in [15] because obvious rules can be hidden using a condition. Moreover, FHSAR (Fast Hiding Sensitive Association Rules) algorithm is presented by Weng *et al.* [19]. This algorithm scans database only once to reduce the execution time. This technique hides already known sensitive association rules. Furthermore, it assigns weights using Equation (3.1) in order to estimate side effects.

$$w_i = \frac{MIC_i}{2^{(|t_i|-1)}} \rightarrow (3.1)$$

where $MIC_i = \max(|R_k|)$

Thus, prior weights associated with transactions are used to modify the transactions until all sensitive rules are hidden. This technique does not provide clear picture of selecting criteria of sensitive item to modify the transactions while our proposed technique is very accurate in the identification of such items as well as recommends transactions to modify with the use of sensitive items. Besides this, Guo [20] presented FP-tree-based method for inverse frequent set mining called as reconstruction-based framework. This technique assumes the sensitive rules for hiding but loses or generates non-informative rules as lost or ghost rules. In this scenario, we do not lose any such information. In the same context, Rajalaxmi *et al.* [23] introduced effective data sanitization algorithm with minimum side effects in the original database. In this data sanitization, the presented HCR (Hybrid Conflict Ratio) approach picks victim transactions and victim items. These victimized items are used to modify the data source in order to minimize the legitimate loss during sanitization. The limitation of this technique is that it does not follow any criterion for the identification and the way of selection of victimized items as well as victimized transactions. In contrast to this, we come up with clear criterion for selecting and identifying sensitive item(s) as well as for transactions keeping support as constant. In addition to support measure in association rule mining, Wang *et al.* [12] focused to increase and decrease the support of association rules based on their antecedent and consequent respectively. For this purpose, ISL (Increase Support of Left Hand Side) and DSR (Decrease Support of Right Hand Side) algorithms are presented including their side effects. These algorithms hide sensitive rules based on predicted items existence in the database while we automatically quantify items to measure their sensitivity without prior knowledge of these items. The side effects results (such as new rules, lost rules, hidden rules and transaction modification) are generated through ISL and DSR are based on predicted items. In addition to quantification, Krishna *et al.* [24] preserves association rules with quantitative data using mean and standard deviation named as Statistical Association Rules (SARs) and Fuzzy Association Rules (FARs) are expressed in linguistic terms. FARs was quantified with the use of the member function [40] as shown in Equation 3.2

$$m_{f_x}(x) : D \rightarrow [0,1] \rightarrow (3.2)$$

This function is used to booleanize the original data source to generate FARs. Similarly, Gupta *et al.* [13] hides association rules discovered form quantitative database. This approach integrates fuzzy rules and apriori [25] concepts to find useful association rules by decreasing the support of rules. Also, an attribute has three fuzzy values as shown in table of section which are produced by the membership function. In this method, a rule is selected as sensitive which has support or confidence beyond the defined threshold value. Thus, the criterion for making decision about sensitivity is poor but allows room to add in especially in fuzzy association rule mining as discussed by Krishna *et al*. [24] and Gupta *et al*. [13].

Saygin *et al*. [14] introduced metric based approach. The purpose of such metric introduction is to demonstrate the security issues in general framework. In this framework, rules are preserved by reducing the support and confidence. Such preservation of association rules can be carried out by introducing "?" in place of "1" in the original data source. Based on this transformed database, safety margin is introduced in order to measure the uncertainty of rules. In case of uncertainty, our approach converts uncertainty into certainty heuristically using BN.

Additionally, association rules can be generated from relational format [31], transactional databases [19,23] as well as from XML documents [26,27,29-31]. From the critically reviewed literature on PPDM, a question can be raised "*Which domain is not yet focused for privacy preservation of association rule mining*?" To dig out this answer, the literature is reviewed in Section 2.2 which highlights the importance of XML documents. This area is not sufficiently focused by researchers in the context of privacy preservation of XML association rules. Thus, we present the critical evaluation of such an important and ignored area for preserving the disclosure risk involved in it.

Gonzing [26] presented *FreqTree* and *DFreqtTree* algorithm based on DOM (Document Object Model) tree. These algorithms mine association rules form XML file in an efficient way. The performance of these algorithms is not compared and the security risks involved in the generation of association rules are ignored. Similarly, Abazeed *et al*. [27] comes up with the modified version of FLEX (MFLEX) algorithm. This implementation is carried out by using DOM and SAX (simple API for XML). In this case, a result of mining algorithm is displayed in XML format. In this mining process, algorithm is not jotted down without ensuring the disclosure risks involved in XML association rules. Besides this, Combi *et al*. [28] turns up with the query based approach to XML file. This query can either be structural or content based for extracting information. The problem with flexibility of XML structure increases the complexity in the process of privacy preservation in terms of association rules. Therefore, XML will get importance in future regarding security risks. In addition to querying XML documents, Bei *et al*. [29] suggested query recommended technique. This technique has five components such as rule miner, result recommender, query miner, query recommender and querier [29]. The ultimate goal of this technique is to provide the best query for the user amongst the frequently asked ones but remains silent about security risks involvement as suggested by Dafa-Alla *et al*. [18]. Furthermore, Wang *et al*. [30] suggested SDST (Standard Data Source Template) for mining XML file. The purpose of such template is to address the complexity and irregularity of XML structure. Therefore, XSL and XSLT are used for standardization. XSLT can be generated for specific XML documents which limit the adaptability related to standardization. Moreover, the performance of XQuery and XSLT is not compared and left the disclosure risks involvement as for future.

Besides querying XML documents, Li *et al*. [31] proposed a high adaptive data mining technique for XML. In this technique, an index table is built from XML documents without user involvement while we transform XML document into transactional itemset and binary table. The conversion of XML documents into index table creates relational dependency but our proposed model measures the dependency amongst items rather than entities. In addition to adaptability, HiLOP (Hierarchical Layered Structure of Pairset) is presented by Shin *et al*. [32]. This methodology contributes the data structure named as Pairset for XML association rule mining. In this structure, join operation is used to save time by reducing the number of rounds for *candidate-tree-item-pruning*. The technique becomes expensive while traversing the depth of the tree. This ignores the likeliness of the structure.

After critical evaluation of XML association rules and PPDM techniques regarding association rules, the question arises that "*How can privacy of XML association rules be preserved*?" To answer this question, more literature is reviewed in Section 2.3. The purpose of this reviewed literature is to provide a solid base for quantification of the occurrence of vertically partitioned items with reliability. Thus, the critical evaluation of Section 2.3 is presented as below.

Doguc *et al*. [33] presented a generic approach to estimate the behavior of the system in a reliable way. In this estimation process, K2 algorithm [33, 34] is used for quantification of associations amongst the vertically partitioned data. This algorithm uses a scoring function heuristically for reducing search space. Moreover, the K2 algorithm [34] uses a defined order of items which helps in the identification of sensitive items while quantifying them. Furthermore, system operation effectiveness is investigated by Doguc *et al*. [36]. This investigation assesses the related dependencies amongst items/attributes. Thus, a problematic item is identified for review to improve the system operation. Therefore, such problematic item identification can be hooked up with our proposed model for the identification of sensitive items in order to modify the original data source. Similarly, Richiardi *et al*. [35] measured modality reliability information by combining of acoustic environment and classifier behavior. The overall modality reliability average accuracy and variability results show the effectiveness of reliability measurement using BN. Also, Vaidya *et al*. [37] proposed a naïve bayes classifier which is used for vertically partitioned data with an objective of preserving privacy. This preservation of privacy is carried out with the same number of entities but variable number of columns while sharing data. Similar to this, we also used the vertically partitioned data while preserving privacy of XML association rules with 26 attributes. In addition to preserving privacy, Wright *et al*. [38] presented preserving privacy protocol for distributed heterogeneous data. Such privacy is carried out by producing the joint data after summing the intermediate values cryptographically. In our proposed model, associated items are quantified without cryptography.

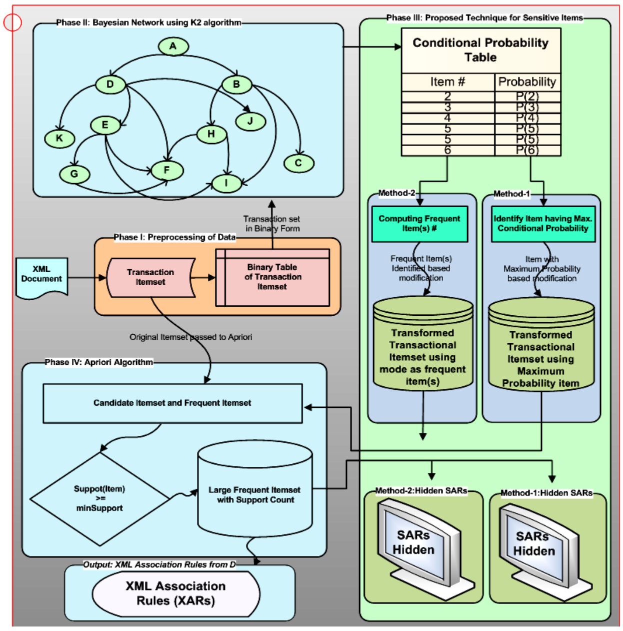
Phase II: Bayesian Network using K2 algorithm
A
B
C
D
E
F
G
H
I
J
K
Phase III: Proposed Technique for Sensitive Items
Conditional Probability Table
Item #
Probability
2
P(2)
3
P(3)
4
P(4)
5
P(5)
5
P(5)
6
P(6)
Transaction set in Binary Form
Phase I: Preprocessing of Data
XML Document
Transaction Itemset
Binary Table of Transaction Itemset
Original Itemset passed to Apriori
Phase IV: Apriori Algorithm
Candidate Itemset and Frequent Itemset
Suppot(Item) >= minSupport
Large Frequent Itemset with Support Count
Output: XML Association Rules from D
XML Association Rules (XARs)
Method-2
Computing Frequent Item(s) #
Frequent Item(s) Identified based modification
Transformed Transactional Itemset using mode as frequent item(s)
Method-1
Identify Item having Max. Conditional Probability
Item with Maximum Probability based modification
Transformed Transactional Itemset using Maximum Probability item
Method-2:Hidden SARs
SARs Hidden
Method-1:Hidden SARs
SARs Hidden

Figure 1. Proposed model for XARs privacy preservation in date mining.

3. Proposed Model for Privacy Preservation of XARs in Data Mining

In order to understand the flow of our proposed model, it has been divided into four phases. Initially, XML document is prepared on a given dataset. This document is automatically read in phase-1 to build transactional symbolic items and binary table. Binary table shows the presence or absence of an item rather than the symbolic item name. Hence, phase-1 is the preprocessing of the data based on XML document. These transactional symbolic item names are passed to apriori algorithm [17] to generate the XARs on the original data source as shown in phase-4 of **Figure 1**. Also in phase-2, K2 algorithm [34] uses binary table and generates BN. In BN generation process, Conditional Probability Table (CPT) is generated which is used to measure the dependency of items on each other. Thus, this table is recorded during the execution of K2 algorithm [34] and stored in MS-Excel file for later use. In phase-3, the stored MS-Excel probabilistic file is read which has two columns named as "*Item#*" and "*Probability*" in the same phase of **Figure 1**. Using these two columns, we presented two methods such as method-1 and method-2. In method-1 of phase-3, maximum conditional probability is picked from the "*probability*" column which shows the maximum dependency of an item on the other items. Based on this probability, an item number is identified from column "*Item #*". Additionally, this number has a unique item in the transactions. In this way, an item is identified as sensitive. After identification of sensitive item symbol, the largest size transactions are modified by deleting this item. Thus, largest size transactions are modified and passed to the apriori algorithm [17] in phase-4 for the generation XML association rules. Similarly in method-2, "*Item #*" is used of CPT and mode (frequent item number) is found out. In this case, the data in "*Item #*" column may be uni-modal, bi-modal or multi-modal. Therefore, many modes can be computed in terms of their occurrence frequency. Thus, these items are declared as sensitive because of their frequent occurrence in BN. After computation of mode, the largest size transactions are modified based on the computed mode(s). These modified transactions are passed to apriori algorithm [17] and XML association rules are generated. The reason of selecting the largest transaction is to keep the effect minimum to transformation of the original data source in order to save the generation of new rules and lost rules as a side effect. Finally, the method-1 and method-2 XML association rules are compared with the output of the original transaction symbolic items XML association rules. In this way, model validation is ensured related to XML association rules in sensitivity preservation in data mining. This comparison shows the difference in results with no new rules as well as lost rules with the use of same support for the original and modified transactional symbolic items. The reason behind keeping the support constant is the truthful identification of sensitive item.

After the detailed understanding of the overall flow of the model, we present the important components of the model in this section.

- ***XML Document***: *A document that contains the transactions in XML.*

- ***Transactional Symbolized Items***: *A group of symbolized items that forms a transaction based on XML document items.*
- ***Binary Table of Transaction***: *A table containing* 1'*s and* 0'*s to represent the presence or absence of an item in a transaction.*
- ***Apriori Algorithm***: *An algorithm suggested in* [17] *for generating XML association rules after preprocessing.*
- ***Bayesian Network and K*****2** ***Algorithm***: *A graphical model that trains interesting relationship among nodes in a probabilistic manner is known as Bayesian Network* (*BN*) *using K*2 *algorithm* [34].
- ***Conditional Probability Table*** (***CPT***): *This table contains items and their conditional probabilities according to their dependency in Bayesian Network.*
- ***Frequent Item Identification*** (***Mode***): *From CPT, the most frequent item*(*s*) *are identified for the modification of transactions.*
- ***Frequent Item Identification Based on Probability***: *Also from CPT, an item that has the maximum dependency on other items in terms of probability is used to modify transactions.*

4. Architectural Implementation

The model for PPDM has been proposed in the previous section which hides the sensitive information disclosed by the association rules. This model demonstrates a very flexible, usable, and reliable performance. In this section, we discuss the architectural implementation of the PPDM model in detail. Thus in such implementation of such architectural implementation, we selected a very high performance language such as Matlab 7.6 among the variety of available tools which have been used for technical computation, visualization and programming to express solutions to problems in an easy to use and friendly environment. The main steps of our proposed model implementation can be expressed in four phases.

Main Steps of the PPDM Model

***STEP*-1:** *Initially, xml_read*() *function is used to read an XML document.*

***STEP*-2:** *This XML document is then converted into symbolized transactional itemset and binary table.*

***STEP*-3:** *Transactional Itemset are passed to apriori algorithm* [17] *as an original data source to generate the association rules without considering the disclosing aspect.*

***STEP*-4:** *Binary table based on transactional itemset is passed to K*2 *algorithm* [34] *to generate Bayesian Network.*

***STEP*-5:** *In this step, Conditional Probability Table of items is stored in MS-Excel file from STEP*-4.

***STEP*-6:** *In STEP*-5, *two columns such as Item# and Item Occurrence Probability is read using xlsread*() *function to calculate.*

***Sub-Step* 6.1:** *Identify the maximum probability based Item# with its symbolized item.*

***Sub-Step* 6.2:** *Compute the Mode* (*Frequently Occurred Item*(*s*) *in Bayesian Network*) *using Item# Column.*

***STEP*-7:** *Modify the transactional Itemset based on Sub-Step* 6.1 *and Sub-Step* 6.2.

***STEP*-8:** *Pass Modified Transactional Itemsets in STEP*-7 *to apriori algorithm* [17].

***STEP*-9:** *Output the Result based on STEP*-8 (*Hidden Sensitive Association Rules*).

In phase-1, step-1 and step-2 are performed. apriori algorithm [17] is used in phase-4 which includes the step-3 and step-4 while K2 algorithm [43] is presented in phase-2 which follows step-4 only. Phase-3 (step-5, step-6 with the sub-step 6.1 and sub-step 6.2, step-7 and step-9) is formulated based on phase-2 to prepare the conditional probability table of items. Also, Method-1 (sub-step 6.1) and method-2 (sub-step-6.2) are presented in the same phase which transforms the original transaction symbolized items. These transformed transactions are passed to phase-4 in order to generate the non-obvious XML association rules. Finally, the original transactional symbolized items output and the modified transactional symbolized items output of Method-1 and Method-2 is compared. The purpose of this comparison is to measure the effectiveness of the proposed technique in this paper.

The important and main suggested processes algorithms are presented while the rest of the functions such as apriori [17], conversion of transactional itemset into binary table, *xml_read*() and *xlsread*() are ignored.

Therefore, we have used K2 algorithm [34] with slight modifications. These modifications can be observed from *line* 2 *to line* 7, *line* 19 *to line* 21, *line* 28 and *line* 29 from the following presented K2 algorithm [34]. The purpose of these modifications is to record and store the conditional probability table as a file of MS-Excel for later use in the implementation.

After preparing MS-Excel file with the use of BN based on the original transactional symbolized items, the proposed model is started as shown in phase-3 of our model in the previous section. From phase-2, we obtained the conditional probabilities of symbolized items. Method-1 finds the maximum probability item from the conditional probability table. This identified item with the highest probability in BN is declared as sensitive because of its conditional occurrence with other items. After identifying the "*item #*" in the BN, we have to search for the item symbol because the transactional symbolized items need to be transformed with the deletion of sensitive item. This transformation is applied only to the assumption based largest size transactions in order to keep the minimal effect to the original transactional symbolized itemset. These modified transactional symbolized items are passed to aprioi algorithm [17] for the generation of non-obvious XML association rules. Therefore, the algorithm developed for such probability based item modification of transaction is shown in **Figure 2**.

Similarly, in phase-3, Method-2 is presented based on mode. Mode means to calculate the most frequent item(s) occurring in the BN. In this case, "*item #*" is used to compute the most frequent items which may be single called as uni-modal, having two modes called as bi-modal, tri-modal and multi-modal as frequent items are more than three. For a huge set of data, many frequent items equal in their occurrence frequency can be obtained from BN.

```
Input: XML document, SupportThreshold
Output: Non-obvious XML association rules
$ConProbTable \leftarrow Call\ MaxCPT()$ // Prune Conditional
    // Probability Table for each Item with maximum probability
$MaxProbItem \leftarrow \max([ConProbTable\{:,1\}])$
$[cptRow, cptCol] \leftarrow size(ConProbTable)$
$FOR\ i = 1 : cptRow$
    $IF\ ConProbTable\{i,1\} = MaxProbItem\ THEN$
        $ItemNumber \leftarrow ConProbTable\{i,2\}$
    $END\ IF$
$Next\ i$
$ItemSymbol \leftarrow itemLookUpTable\{itemNumber, 2\}$
$t\_A\_L \leftarrow [\ ]$; // transaction and Lengths
$tC \leftarrow 0$; // transaction counter
$FOR\ k = 1 : len(lot)$ //lot : list Of Transactions
    $IF\ any(lot\{k\} = itemSymbol)\ THEN$
            $tC \leftarrow tC + 1$
        $t\_A\_L(tC,1) \leftarrow k$
        $t\_A\_L(tC,2) \leftarrow len(lot\{k\})$
    $END\ IF$
$NEXT\ k$
$mL \leftarrow \max(t\_A\_L(:,2))$; // maximum transaction length
$t\_T\_D\_F \leftarrow t\_A\_L(find(t\_A\_L(:,2) = mL),1)$
    //transaction To Delete From
$FOR\ m = 1 : len(t\_T\_D\_F)$
    $transaction \leftarrow lot\{t\_T\_D\_F(m)\}$
    $transaction(transaction == ItemSymbol) \leftarrow [\ ]$;
    $lot\{t\_T\_D\_F(m)\} \leftarrow transaction$
$Next\ m$
$Call\ Apriori(lot)$ // use apriori algorithm to
    hide SARs on modified lot
$END$
```

Figure 2. Method-1: Maximum probability based sensitive item identification & transaction modification.

These mode based items are called as sensitive and used for the transformation of assumption based largest size transactional symbolized items. Finally, transformed transactional symbolized items are passed to aprioialgorithm [17] as shown in phase-3 and phase-4 of our model in the previous section. In this way, the sensitive XML association rules disclosure risk is minimized. Hence, the algorithmic steps are presented and can be viewed in **Figure 3** while.

```
Input: XML document, SupportThreshold
Output: Non-obvious XML association rules
$ModeTable = xlsread('filename.xls')$
$UniqueNodes = unique(ModeTable(:,2))$
$[unRow, unCol] = size(UniqueNodes)$
$[mtRow, mtCol] = size(ModeTable)$
$FOR\ i = 1 : unRow$
    $C = 0$ //Counter is set as 0
    $FOR\ j = 1 : mtRow$
        $IF\ UniqueNodes(i,1) == ModeTable(j,2)\ THEN$
            $C \leftarrow C + 1$
        $END\ IF$
    $NEXT\ j$
    $UniqueNodes(i,2) = C$
$NEXT\ i$
$BFN = mode(UniqueNodes(:,2))$ //Bayesian Frequent Node
$MM = [\ ]$; // Multiple Modes
$FNIS = \{\ \}$; // Frequent Node Item Symbol(s) Initialized
$MC = 1$; // Mode Counter is set as 1
$FOR\ i = 1 : unRow$
    $IF\ UniqueNodes(i,2) == BFN\ THEN$
        $MM(MC,1) = UniqueNodes(i,1)$
        $MM(MC,2) = UniqueNodes(i,2)$
        $FNIS(MC,1) = itemLookUpTable\{MM(MC,1), 2\}$
        $MC \leftarrow MC + 1$
    $END\ IF$
$NEXT\ i$
$t\_A\_L \leftarrow [\ ]$; // transaction and Lengths
$tC \leftarrow 0$; // transaction counter
$[fnisRow, fnisCol] \leftarrow Size(FNIS)$;
$FOR\ j = 1 : fnisRow$
    $FMIS \leftarrow FNIS\{j,1\}$ //Frequent Item Mode Symbol
    $FOR\ k = 1 : len(lot)$ //list Of Transaction
        $IF\ any(lot\{k\} == FMIS)\ THEN$
            $tC \leftarrow tC + 1$
            $t\_A\_L(tC,1) \leftarrow k$
            $t\_A\_L(tC,2) \leftarrow len(lot\{k\})$
        $END\ IF$
    $NEXT\ k$
$NEXT\ j$
$mL \leftarrow \max(t\_A\_L(:,2))$;
$t\_A\_D\_F \leftarrow t\_A\_L(find(t\_A\_L(:,2) == mL),1)$;
$FOR\ j = 1 : fnisRow$
    $FMIS \leftarrow FNIS\{j,1\}$;
    $FOR\ m = 1 : len(t\_T\_D\_F)$
        $transaction \leftarrow lot\{t\_T\_D\_F(m)\}$;
        $transaction(transaction == FMIS) \leftarrow [\ ]$;
        $lot\{t\_T\_D\_F(m)\} \leftarrow transaction$
    $NEXT\ m$
$NEXT\ j$
$Call\ Apriori(lot)$
$END$
```

Figure 3. Method-2: Mode based sensitive item(s) identification & transactions modification.

5. Experimental Results

The assessment of the proposed model is pursued on ZOO dataset [39] with the consideration of 15 attributes out of 17 after implementing it in Matlab 7.6 which is suitable for technical computations and visualization of output. Our proposed model suggests two methodologies such as Method-1 and Method-2. Initially, the implemented program reads XML document. In phase-I, the preprocessing of XML items am carried out by assigning symbol to transactional items as well as form a binary table of these items in the XML document. The following sample **Table 1** shows the presence or absence of an item (animal) in a transaction of XML document.

Thus in phase-II, transactions based binary table of XML document items is passed to K2 algorithm [34] to generate Bayesian Network (as shown in **Figure 4**. by recording the occurrence of items and their conditional probabilities in MS-Excel file as shown in **Table 2** and **Figure 5**.

Table 1. Sample binary table of ZOO dataset.

ID	J	H	G	L	A	B	M	O	C	D	Q	I	N	F	E
1	1	0	0	1	0	0	1	1	1	1	0	0	0	0	1
2	1	0	0	1	0	0	0	1	1	1	0	0	1	0	1
3	0	0	1	0	0	1	1	1	1	0	0	1	1	0	0
4	1	0	0	1	0	0	1	1	1	1	0	0	0	0	1

Table 2. Conditional probability table.

Item #	2	3	4	4	5	5	5	6	7
P(Item)	0.968	0.723	0.495	0.474	0.933	0.884	0.854	0.577	0.835
Item #	7	7	8	8	8	9	9	9	9
P(Item)	0.766	0.692	0.503	0.43	0.366	0.438	0.422	0.354	0.307
Item #	10	10	10	11	11	12	12	12	13
P(Item)	0.499	0.455	0.411	0.267	0.207	0.936	0.911	0.904	0.525
Item #	13	14	14	14	14	14	15	15	15
P(Item)	0.511	0.674	0.474	0.416	0.291	0.288	0.422	0.387	0.38

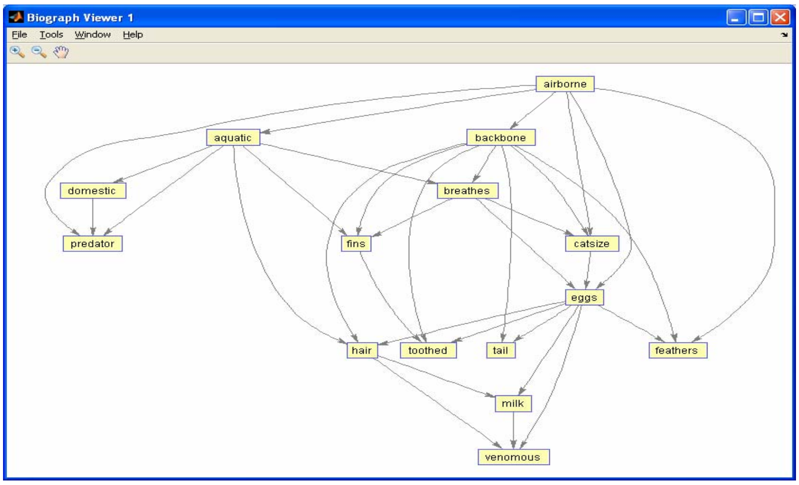

Figure 4. Bayesian network by K2 algorithm on ZOO da- taset.

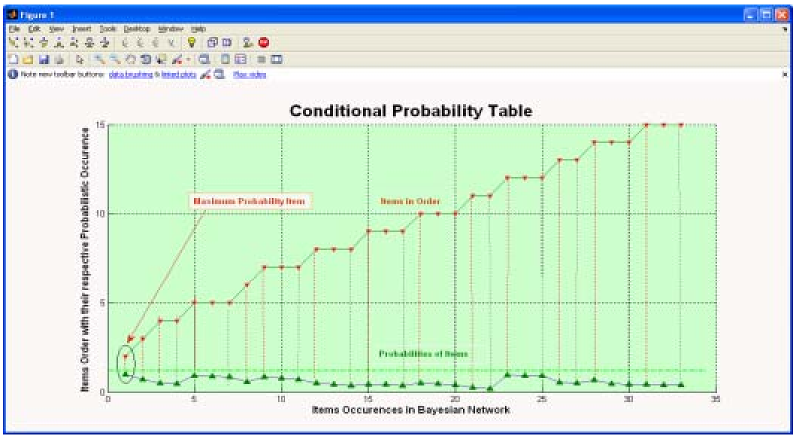

Figure 5. Graphical view of Conditional Probabilities of Items (CPT).

Besides this in phase-IV, the original transactional itemset is passed to apriori algorithm for the production of XML association rules. In phase-III, the MS-Excel file is read for Method-1 and Method-2. In method-1, sensitive item is recognized with the help of maximum conditional probability quantified through Bayesian Network. Also in method-2, sensitive item(s) are recognized with the computation of frequently occurring items in Bayesian Network. Both the suggested techniques are presented below. Before highlighting the further details of K2 algorithm [34], symbolized transactions are passed to apriori algorithm [17] in order to generate association rules. These association rules are generated with the use of *minimum support* 26. Based on the ZOO dataset [39], the XML Association Rules are presented in the following **Table 3**. The total XML association rules generated is 132 with their respective support, confidence and lift ratio. Among these generated rules, there are five B➔C, B➔G, B➔L, B➔M and B, C➔M are the sensitive XARs based on the identified sensitive item.

The results generated in this method are the same regarding the K2 algorithm [34] and apriori algorithm [17]. The difference of computing frequent item(s) called Mode is being identified based on the Item# column in CPT rather than maximum probability. From **Table 2**, Mode is computed. There may be more than one mode which will be known as multi mode computation. Thus, mode represents the frequently occurrence of the items in Bayesian Network. Therefore, the following **Table 4** shows the frequency and multiple items as sensitive items.

The graphical rpresentation of **Table 5** is shown in **Figure 6** with all items according to their occurrence in Bayesian Network.

Based on these items, the original data source of transaction itemset is modified by deleting these sensitive items from the largest itemset consisting transaction in the Zoo dataset. In this dataset, there is one dominating transaction in its length which is the only transaction to be modified based on our assumption. Therefore, this aspect restricts modifying transaction with a minimal effect but increases the side effect.

In this method, Sensitive XARs are picked in which these items are present. So these XARs need not to be disclosed while sharing information in collaboration. The XARs generated using the apriori [17] on the original data source are shown as in **Table 5**.

There are 103 XML Sensitive Association Rules which need not be disclosed. In this method hidden rules are shown in **Table 6** which is only 6 because of the domi- nating nature of the largest transaction as well as multi- ple modes generated by the BN according to their occur- rences. Thus, we can conclude that increase in Mode computed through BN will generate more sensitivity with an introduction of more complexity in the hiding process.

As the PPDM is effective and reliable, so the implemented algorithms hide the sensitive XML Association

Table 3. XML Association rules using apriori on *D*.

Rule #	XARs	s%	c%	LR
1	B==>C	29.7	83.33	1
2	B==>G	29.7	83.33	1.41
3	B==>L	28.71	80.56	1.44
4	B==>M	25.74	72.22	0.96
5	C==>D	68.32	83.13	1.04
6	C==>E	42.57	51.81	1.18
	⋮			
34	M==>N	51.49	69.33	1.14
35	B,C==>M	25.74	86.67	1.16
36	C,D==>E	38.61	56.52	1.28
	⋮			
130	C,D,E,J,K==>N	28.71	96.67	1.58
131	C,D,E,K,M==>N	25.74	96.3	1.58
132	C,D,J,K,M==>N	31.68	96.97	1.59

Table 4. Summary of PPDM model: Method-1.

Sensitive XARs	Hidden XARs	Side Effects	New XARs	Modified Transactions	Total Transaction
5	2	3	0	1	101

Table 5. Sensitive items as mode in BN.

Item Name	Item Symbol	Item #	Frequency
“catsize”	E	5	3
“eggs”	G	7	3
“feathers”	H	8	3
“hair”	J	10	3
'predator'	L	12	3
'venomous'	O	15	3

Table 6. Hidden rules using mode in bayesian network.

Rule #	XARs	s%	c%	LR
107	D,E,M==>N	25.74	76.47	1.25
113	E,K,M==>N	25.74	96.3	1.58
119	C,D,E,M==>N	25.74	76.47	1.25
125	C,E,K,M==>N	25.74	96.3	1.58
128	D,E,K,M==>N	25.74	96.3	1.58
131	C,D,E,K,M==>N	25.74	96.3	1.58

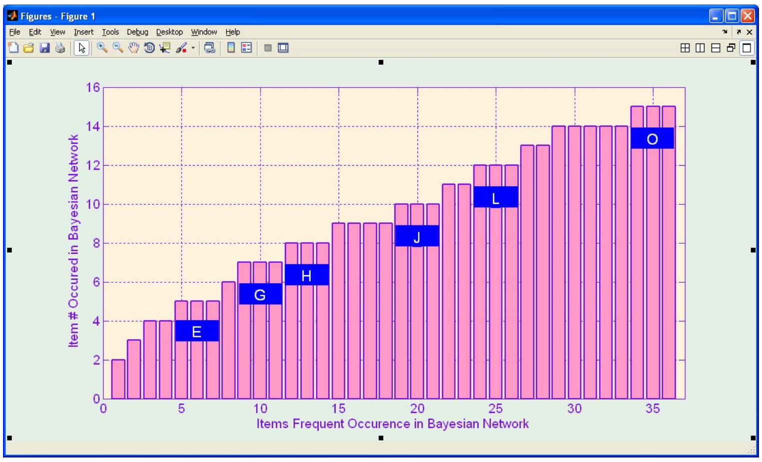

Figure 6. Multiple modes in BN.

Rules using both the methods. Thus, the comparison of the PPDM model method-1 and method-2 are given with the ISL algorithm and DSR algorithm [14] and FHSAR [21].

In this Figure, we use the same support as used by ISL and DSR [12]. The purpose of using the same support is to standardize the results of existing technique as well as our own technique. The reason of this standardization is to measure the effectiveness of the proposed technique because such NP-hard [11] problem has a variety of parameters. Thus, from **Figure 8**, we can observe the 100 sensitive XML association rules are hidden with support = 3% and 148 sensitive XML association rules are preserved with the use of same support through Method-1 and Method-2 respectively. Moreover, the disclosure risk of association rules minimizes and we can observe from the **Figure 8** that less rules were hidden through Method-1 and Method-2 respectively with the use support = 16%. In contrast to other techniques as in [12,19], hidden rules are below 10. The difference is quite clear that these techniques could not identify the sensitive item(s) in the roots of the original database with certain accuracy as in the proposed solution. At this stage, the question arises, "*In how many rules, did we achieve success in preserving the privacy of XML association rules*?" To answer this question, we have presented the total number of sensitive rules and the number of attributes considered by the proposed technique as well as the existing techniques [12,19]. For this purpose in zoo dataset, there is one largest size transaction and it consists of 15 attributes and 101 transactions while the existing techniques are evaluated on different dataset with fifty attributes and 10,000 [21] and 15,000 [14] transactions. Therefore, such a huge difference cannot show a clear comparison. In spite of such valid reason, we presented the total rules generated on the variable number of attributes in order to understand the NP-hard [11] problem as shown in **Figure 7**.

With such variable background knowledge of various techniques, they produce variable results according to their own parameters. Despite this fact, we have converted these results into percentage for the given parameters. The reason of this conversion is to know the facts about NP-hardness [11] in order to catch a useful and common methodology. Hence, we have shown the proposed model results as well as the existing techniques' results at one place for better understanding of a problem with no proper solution. These results are shown in the following **Figure 8** which provides an easy and understandable view to the generated results by various techniques [12,19] including the proposed model.

6. Conclusions and Future Work

Data mining technology is benefitted by privacy pre- serving data mining due to the potential increase in the electronic data. However, we should provide justification for the perpetuation of confidentiality. For this purpose, privacy needs to be defined clearly. Currently, the whole academia is presenting its own meaning of privacy. Therefore, such blend of definitions leads to misunderstanding.

Consequently in this paper, we present a PPDM model for hiding the XML association rules which discloses perceptive information to the external parties. Thus, we show how the supervised learning technique is effective in the confidentiality of XML association rules. Moreover, we also presented two methodologies based on Bayesian Network such as the maximum conditional probability of an item and mode of items according to their occurrences in a supervised manner. Additionally, the proposed model automatically identifies the sensitive items to modify largest and sized transactions in the database without changing support measure. The purpose of such supervision approach is to keep the complete

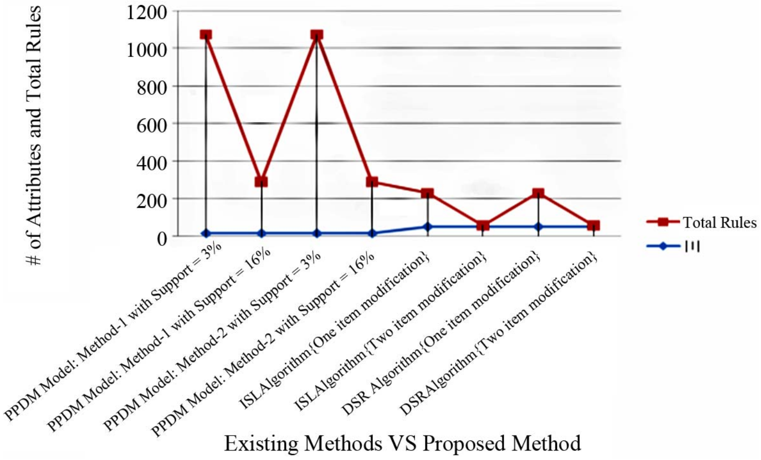

Figure 7. Comparison of rules generation from the specified # of attributes.

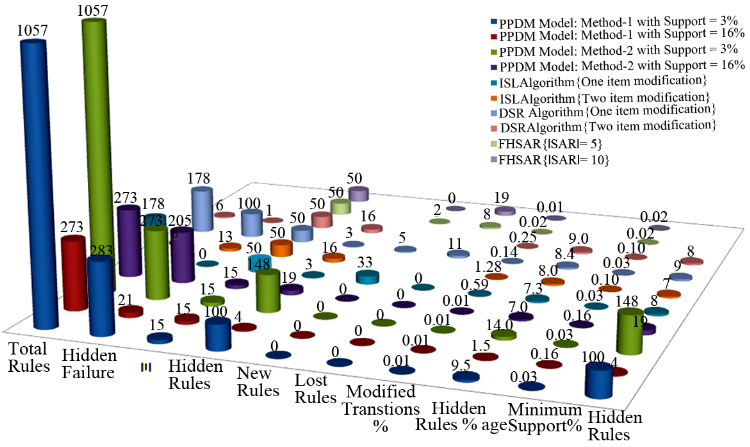

Figure 8. Overall comparison of proposed methodology.

control on the manipulation of information before sharing. This model does not generate any new or ghost rules as well as complete elimination of lost rules. However, the model keeps sensitive rules as insecure as side effect. In future, we shall introduce a supplementary approach in order to remove the largest transaction modification assumption by quantifying the items dependency as well as transactional dependency for more exact solution while addressing confidentiality issues in association rule mining.

7. References

[1] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, Vol. 17, No. 3, 1996, pp. 37-54.

[2] T. Porter and B. Green, "Identifying Diabetic Patients: A Data Mining Approach, Americas Conference on Information Systems," *Proceedings of the Fifteenth Americas Conference on Information Systems*, San Francisco, 2009.

[3] C. Apte, B. Liu, E. P. D. Pednault and P. Smyth, "Business Applications of Data Mining," *Communications of the ACM*, Vol. 45, No. 8, 2002, pp. 49-53. doi:10.1145/545151.545178

[4] L. Chen, T. Sakaguchi and M. N. Frolick, "Data Mining Methods," *Applications and Tools, Information Systems Management*, Vol. 17, No. 1, 2002, pp. 65-70.

[5] A. Berson, S. Smith and K. Thearling, "Building Data Mining Applications for CRM," McGraw-Hill Companies, New York, p. 510.

[6] T. H. Davenport, J. G. Harris and A. K. Kohli, "How Do They Know Their Customers So Well?" *MIT Sloan Management Review*, Vol. 42, No. 2, pp. 63-73.

[7] A. Scime, "Web Mining: Applications and Techniques," Idea Group Publishing, 2004, pp. 1-442. doi: 10.4018/978-1-59140-414-9

[8] W.-M. Ouyang and Q.-H. Huang, "Privacy Preserving Association Rules Mining Based on Secure Two-Party Computation," In: D.-S. Huang, K. Li and G. W. Irwin, Eds., Springer-Verlag, Berlin Heidelberg, 2006, pp. 969-975.

[9] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining," *Proceedings of ACM Workshop Research Issues in Data Mining and Knowledge Discovery*, Montreal, 1996.

[10] S. R. M. Oliveira and O. Zaïane, "Toward Standardization in Privacy-Preserving Data Mining," *Proceedings of the* 3nd *Workshop on Data Mining Standards*, 2004.

[11] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, "Disclosure Limitation of Sensitive Rules," *Proceedings of the* 1999 *Workshop on Knowledge and Data Engineering Exchange*, IEEE Computer Society Washington, DC.

[12] S.-L. Wang, B. Parikh and A. Jafari, "Hiding Informative Association Rule Sets," *Expert Systems with Applications*: *An International Journal*, Vol. 33, No. 2, 2007, pp. 316- 323.

[13] M. Gupta and R. C. Joshi, "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data," *International Journal of Computer Theory and Engineering*, Vol. 1, No. 4, 2009, pp. 1793-1820.

[14] Y. Saygın, V. S. Verykios and A. K. Elmagarmid, "Privacy Preserving Association Rule Mining," *IEEE Computer Society*, Washington, DC, 2002, p. 151.

[15] E. Dasseni, V. S. Verykios, A. K. Elmagarmid and E. Bertino, "Hiding Association Rules by Using Confidence and Support," IBM, Almaden Research Center, San Jose, 2000.

[16] K. Duraiswamy, D. Manjula and N. Maheswari, "A New Approach to Sensitive Rule Hiding," *Computer and Information Journal*, Vol. 1, No. 3, 2008, pp. 107-111.

[17] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proceedings of ACM SIGMOD International Conference on Management of Data*, Washington DC, 1993.

[18] A. F. A. Dafa-Alla, G. Sohn and K. H. Ryu, "Employing PRBAC for Privacy Preserving Data Publishing," Seoul, 2009.

[19] C.-C. Weng, S.-T. Chen and H.-C. Lo, "A Novel Algorithm for Completely Hiding Sensitive Association Rules," *Eighth International Conference on Intelligent Systems Design and Applications*, Vol. 3, 2008, pp. 202-208. doi:10.1109/ISDA

[20] Y.-H. Guo, "Reconstruction-Based Association Rule Hiding," *Proceedings of SIGMOD*2007 *Ph.D. Workshop on Innovative Database Research*, Beijing, 2007.

[21] Y. H. Guo, Y. H. Tong, S. W. Tang and D. Q. Yang, "A FPtree-Based Method for Inverse Frequent Set Mining," *Proceedings of the* 23*Prd P British National Conference on Databases*, Springer-Verlag 2006, pp. 152-163.

[22] J. W. Han, J. Pei and Y. W. Yin, "Mining Frequent Patterns without Candidate Generation," In: C. Weidong and F. Jeffrey, Eds., *Proceedings of the ACM SIGMOD Conference on Management of Data*, ACM Press, Dallas, 2000, pp. 1-12. doi:10.1145/342009.335372

[23] R. R. Rajalaxmi and A. M. Natarajan, "Hybrid Conflict Ratio for Hiding Sensitive Patterns with Minimum Information Loss," *International Journal of Computer Theory and Engineering*, Vol. 1, No. 4, 2009, pp. 1793-8201.

[24] G. V. Krishna and P. R. Krishna, "A Novel Approach for Statistical and Fuzzy Association Rule Mining on Quantitative Data," *Journal of Scientific and Industrial Research*, Vol. 67, 2008, pp. 512-517.

[25] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables," *Proceedings of ACM SIGMOD*, New York, 1996, pp. 1-12.

[26] G.-X. Wu, "A Study on the Mining Algorithm of Fast Association Rules for the XML Data," *International Conference on Computer Science and Information Technology*, 2008. doi:10.1109/ICCSIT.2008.89

[27] A. Abazeed, A. Mamat and M. Nasir, "Hamidah Ibrahim, "Mining Association Rules from Structured XML Data," 2009 *International Conference on Electrical Engineering and Informatics*, Selangor, 2009. doi:10.1109/ICEEI.2009.5254708

[28] C. Combi, B. Oliboni and R. Rossato, "Querying XML Documents by Using Association Rules," *Proceedings of the* 16*th International Workshop on Database and Expert Systems Applications*, 2005.

[29] Y.-J. Bei, G. Chen, L.-H. Yu, F. Shao and J.-X. Dong, "XML Query Recommendation Based On Association Rules," *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, 2007. doi:10.1109/SNPD.2007.378

[30] X.-W. Wang and C.-J. Cao, "Mining Association Rules from Complex and Irregular XML Documents Using XSLT and Xquery," *International Conference on Advanced Language Processing and Web Information Technology*, 2008. doi:10.1109/ALPIT.2008.48

[31] X.-Y. Li, J.-S. Yuan, Y.-H. Kong, "Mining Association Rules from XML Data with Index Table," *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 2007.

[32] J. Shin, J. Paik and U. Kim, "Mining Association Rules from a Collection of XML Documents Using Cross Filtering Algorithm," *International Conference on Hybrid Information Technology*, 2006.

[33] O. Doguc and J. E. Ramirez-Marquez, "A Generic Method for Estimating System Reliability Using Bayesian Networks, Reliability Engineering & System Safety," Vol. 94, No. 2, 2009, pp. 542-550. doi:10.1016/j.ress.2008.06.009

[34] G. F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning*, Kluwer Academic Publishers, Hingham, Vol. 9 , No. 4, 1992, pp. 309-347,

[35] J. Richiardi, P. Prodanov and A. Drygajlo, "A Probabilistic Measure of Modality Reliability in Speaker Verification," 2005.

[36] O. Doguc and W. Jiang, "A Bayesian Network (BN) Model for System Operational Effectiveness Assessment & Diagnosis," 26*th ASEM National Conference Proceedings*, 2005.

[37] J. Vaidya and C. Clifton, "Privacy Preserving Naive Bayes Classifier for Vertically Partitioned Data," 2003.

[38] R. Wright and Z.-Q. Yang, "Privacy Preserving Bayesian Network Structure Computation on Distributed Heterogeneous Data," Seattle, 2004.

[39] Zoo Dataset. http://mlearn.ics.uci.edu/databases /zoo/