International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com
Volume 7, Issue 2 (May 2013), PP. 21-32

Survey: Privacy Preservation Association Rule Mining

Lalita Sharma[1], Vinit Kumar[2],Pushpakraval[3]
[1,2,]Departmentofcomputerengineering, Hasmukhgoswami College Of Engineering, Vehlal, Gujarat.
[3]Department Of Computer Engineering, DAIICT, Gandhinagar, Gujarat.

ABSTRACT*–Privacy is an important issue when one wants to make use of data that involves individuals' sensitive information. Research on protecting the privacy of individuals and the confidentiality of data has received contributions from many fields, including computer science, statistics, economics, and social science. In this paper, we survey research work in privacy-preserving Data Publishing and Association Rule Mining. Association rule mining is the most important technique in the field of data mining. It aims at extracting interesting correlation, frequent pattern, association or casual structure among set of item in the transaction database or other data repositories. Association rule mining is used in various areas for example Banking, department stores etc. In this paper, we provide the preliminaries of basic concepts about association rule mining and survey the list of existing association rule mining techniques. Of course, a single article cannot be a complete review of all the algorithms, yet we hope that the references cited will cover the major theoretical issues, guiding the researcher in interesting research directions that have yet to be explored.*

Keywords **–**K-anonymity, Perturbation, Cryptography, Randomization, Association Rule Mining, Support, and Confidence.

I. INTRODUCTION

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [1]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two subproblems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … ,Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1}⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second subproblems is quite straight forward, most of the researches focus on the first subproblems.

The first sub-problem can be further divided into two sub-problems: candidate large itemsets generation process and frequent itemsets generation process. We call those itemsets whose support exceed the support threshold as large or frequent item-sets, those itemsets that are expected or have the hope to be large or frequent are called candidate itemsets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "nonredundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength. Hegland [16] reviews the most well-known algorithm for producing association rules - Apriori and discuss variants for distributed data, inclusion of constraints and data taxonomies. The review ends with an outlook on tools which have the potential to deal with long itemsets and considerably reduce the amount of (uninteresting) returned. In this paper, we surveyed the most recent existing association rule mining techniques. The organization of the rest of the paper is as follows. Section 4 provides the preliminaries of basic concepts and their notations to facilitate the discussion and describes the well-known algorithms. Section 2 describes different methods of privacy preserving section 3 describes merit and demerit of privacy preserving methods section 4 basic Concepts & basic Association Rules Algorithms 5 describes the methods that have been proposed for increasing the efficiency of association rules algorithms. Section 6 refers to the categories of databases in which

association rule can be applied. Section 7 presents the recent advances in association rule discovery. Finally, Section 9 concludes the paper.

II. THEORETICAL ANALYSIS AND LITERATURE SURVEY

2.1 Different Approaches to achieve Privacy

2.1.1 K-anonymous

When releasing micro data for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Samarati et al. [42]; Sweeney [43] introduced the *k*-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least *k*-other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization. Unlike traditional privacy protection techniques such as data swapping and adding noise, information in a *k*-anonymous table through generalization and suppression remains truthful. In particular, a table is *k*- anonymous if the QI values of each tuple are identical, to those of at least *k* other tuples. Table3 shows an example of 2-anonymous generalization for Table. Even with the voter registration list, an adversary can only infer that Ram may be the person involved in the first 2 tuples of Table1, or equivalently, the real disease of Ram is discovered only with probability 50%. In general, *k* anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/*k*.

ID	**Attributes**			
	Age	Sex	Zip Code	Disease
1	36	Male	93461	Headache
2	34	Male	93434	Headache
3	41	Male	93867	Fever
4	49	Female	93849	Cough

TABLE-1MICRODATA

ID	**Attributes**			
	Name	Age	Sex	Zip code
1	Ram	36	Male	93461
2	Manu	34	Male	93434
3	Ranu	41	Male	93867
4	Sonu	49	Female	93849

TABLE-2 VOTER REGISTRATION LIST

ID	**Attributes**			
	Age	Sex	Zip Code	Disease
1	3*	Male	934**	Headache
2	3*	Male	934**	Headache
3	4*	*	938**	Fever
4	4*	*	938**	Cough

TABLE-3 2-AN0NYM0US TABLETABLE-4 ORIGINAL PATIENTS TABLE

ID	**Attributes**		
	Zip Code	Age	Disease
1	93461	36	Headache
2	93434	34	Headache
3	93867	41	Fever
4	93849	49	Cough

TABLE-5 ANONYMOUS VERSIONS OF TABLE1

While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the *k*-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods. Example1. Table4 is the Original data table, and Table5 is an

anonymous version of it satisfying 2-anonymity.The Disease attribute is sensitive. Suppose Manu knows that Ranu is a 34 years old woman living in ZIP 93434 and Ranu's record is in the table. From Table5, Manu can conclude that Ranu corresponds to the first equivalence class, and thus must have fever. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Sonu's age and zip code, Manu can conclude that Sonu's corresponds to a record in the last equivalence class in Table5. Furthermore, suppose that Manu knows that Sonu has very low risk for cough. This background knowledge enables Manu to conclude that Sonu most likely has fever.

2.1.2 Perturbation approach

The perturbation approach works under the need that the data service is not allowed to learn or recover precise records. This restriction naturally leads to some challenges. Since the method does not reconstruct the original data values but only distributions, new algorithms need to be developed which use these reconstructed distributions in order to perform mining of the underlying data. This means that for each individual data problem such as classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. For example, Agrawal [44] develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton and Rizvi and Haritsa[46] develop methods for privacy-preserving association rule mining. While some clever approaches have been developed for distribution-based mining of data for particular problems such as association rules and classification, it is clear that using distributions instead of original records restricts the range of algorithmic techniques that can be used on the data [45].

In the perturbation approach, the distribution of each data dimension reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations. For example, the classification technique uses a distribution-based analogue of single-attribute split algorithm. However, other techniques such as multivariate decision tree algorithms cannot be accordingly modified to work with the perturbation approach. This is because of the independent treatment of the different attributes by the perturbation approach.
This means that distribution based data mining algorithms have an inherent disadvantage of loss of implicit information available in multidimensional records. Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast tool set of cryptographic algorithms and constructs to implement privacy -preserving data mining algorithms. However, recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

2.1.3 Cryptographic technique

Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular [47] for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work [48] has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

2.1.4 Randomized response techniques

The method of randomization can be described [50] as follows. Consider a set of data records denoted by $X = \{x_1. ..x_N\}$. For record xi X, we add a noise component which is drawn from the probability distribution$f_Y(y)$.Thesenoise components are drawn independently, and are denoted $y_1. . . y_N$ Thus, the new set of distorted records are denoted by $x_1 + y_1. . .x_N + y_{N.}$ We denote this new set of records by $z_1. . . z_{N.}$In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered. Thus, if X be the random variable denoting the data distribution for the original record, Y be the random variable describing the noise distribution, and Z be the random variable denoting the final record, we have:

$Z = X + Y$

$$X = Z - Y$$

Now, we note that N instantiations of the probability distribution Z are known, whereas the distribution Y is known publicly. For a large enough number of values of N, the distribution Z can be approximated closely by using a variety of methods such as kernel density estimation. By subtracting Y from the approximated distribution of Z, it is possible to approximate the original probability distribution X. In practice, one can combine the process of approximation of Z with subtraction of the distribution Y from Z by using a variety of iterative methods Such iterative methods typically have a higher accuracy than the sequential solution of first approximating Z and then subtracting Y from it.

The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. Although information from each individual user is scrambled, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for decision-tree classification since decision-tree classification is based on aggregate values of a data set, rather than individual data items. Randomized Response technique was first introduced by Warner as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute A, queries are sent to a group of people Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models: Related-Question Model and Unrelated-Question Model have been proposed to solve this survey problem. In the Related-Question Model, instead of asking each respondent whether he/she has attribute A the interviewer asks each respondent two related questions, the answers to which are opposite to each other .When the randomization method is carried out, the data collection process consists of two steps .The first step is for the data providers to randomize their data and transmit the randomized data to the data receiver. In the second step, the data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm. The model of randomization is shown in Figure 1.

Figure 1: The Model of Randomization

One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. Therefore, the randomization method can be implemented at datacollection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process.

2.1.5 Data-Blocking Techniques

Data-Blocking is another data modification approach for association rule hiding. Instead of making data distorted (part of data is altered to false), blocking approach is implemented by replacing certain data items with a question mark "?". The introduction of this special unknown value brings uncertainty to the data, making the support and confidence of an association rule become two uncertain intervals respectively. At the beginning, the lower bounds of the intervals equal to the upper bounds. As the number of "?" in the data increases, the lower and upper bounds begin to separate gradually and the uncertainty of the rules grows accordingly. When either of the lower bounds of a rule's support interval and confidence interval gets below the security threshold, the rule is deemed to be concealed.

2.1.6 Data Reconstruction Approaches

Data reconstruction methods put the original data aside and start from sanitizing the so-called "knowledge base". The new released data is then reconstructed from the sanitized knowledge base. This idea is first depicted in [49]. They give a coarse Constraint-based Inverse Itemset Lattice mining procedure (CIILM) for hiding sensitive frequent itemsets. Our work is inspired by it. The main differences are: 1) their method aims at hiding frequent itemsets, while ours addresses hiding association rules; 2) their data reconstruction is based on itemset lattice, while ours is based on FP-tree. In phase 2 I shall work on privacy preserving association rule mining based on FP-tree.

3 MERITS AND DEMERITS OF DIFFERENT TECHNIQUES OF PRIVACY IN DATA MINING

Techniques of PPDM	**Merits**	**Demerits**
ANONYMIZATION	This method is used to protect respondents' identities while releasing truthful information. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.	There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the *k*-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods.
PERTURBATION	Independent treatment of the different attributes by the perturbation approach.	The method does not reconstruct the original data values, but only distribution, new algorithms have been developed which uses these reconstructed distributions to carry out mining of the data available.
RANDOMIZED RESPONSE	The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in privacy preserving data mining. The randomization method is more efficient. However, it results in high information loss.	Randomized Response technique is not for multiple attribute databases.
DATA BLOCKING	Replacing certain data items with a question mark "?". The introduction of this special unknown value brings uncertainty to the data, making the support and confidence of an association rule become two uncertain intervals respectively	Rule Eliminated Ghost Rule Created
CRYPTOGRAPHIC	Cryptography offers a well-defined Model. For privacy, which includes methodologies for proving and quantifying it. There exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms.	This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

TABLE–6 PROS AND CONS OF DIFFERENT PRIVACY PRESERVING TECHNICS

III. BASIC CONCEPTS & BASIC ASSOCIATION RULES ALGORITHMS

Let I=I1, I2, … , Imbe a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where X, $Y \subset I$ are sets of items called itemsets, and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support(s) and confidence(c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. Support(s) of an association rule is defined as the percentage/fraction of records that contain $X \cup Y$ to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule $X \Rightarrow Y$ is 80%, it means that 80% of the transactions that contain X also contain Y together. In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k-itemsets. Generally, an association rules mining algorithm contains the following steps:

- The set of candidate k-itemsets is generated by 1-extensions of the large (k -1) -itemsets generated in the previous iteration.
- Supports for the candidate k-itemsets are generated by a pass over the database.
- Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

This process is repeated until no more large itemsets are found. The AIS algorithm was the first algorithm proposed for mining association rule [1]. In this algorithm only one item consequent association rules are generated, which means that the consequent of those rules only contain one item, for example we only generate rules like $X \cap Y \Rightarrow Z$ but not those rules as $X \Rightarrow Y \cap Z$. The main drawback of the AIS algorithm is too many candidate itemsets that finally turned out to be small are generated, which requires more space and wastes much effort that turned out to be useless. At the same time this algorithm requires too many passes over the whole database. Apriori is more efficient during the candidate generation process [2]. Apriori uses pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness. These are the itemsets that the algorithm can prove will not turn out to be large. However there are two bottlenecks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements.

IV. INCREASING THE EFFICIENCY OF ASSOCIATION RULES ALGORITHMS

The computational cost of association rules mining can be reduced in four ways:

- by reducing the number of passes over the database
- by sampling the database
- by adding extra constraints on the structure of patterns
- through parallelization.

In recent years much progress has been made in all these directions.

4.1 Reducing the number of passes over the database

FP-Tree [15], frequent pattern mining, is another milestone in the development of association rule mining, which breaks the main bottlenecks of the Apriori. The frequent itemsets are generated with only two passes over the database and without anycandidate generation process. FP-tree is an extended prefix-tree structure storingcrucial, quantitative information about frequent patterns. Only frequent length-1 itemswill have nodes in the tree, and the tree nodes are arranged in such a way that morefrequently occurring nodes will have better chances of sharing nodes than less frequentlyoccurring ones.FP-Tree scales much better than Apriori because as the support threshold goesdown, the number as well as the length of frequent itemsetsincrease dramatically.The candidate sets that Apriori must handle become extremely large, and the patternmatching with a lot of candidates by searching through the transactions becomes veryexpensive. The frequent patterns generation process includes two sub processes:constructing the FP-Tree, and generating frequent patterns from

the FP-Tree. Themining result is the same with Apriori series algorithms.To sum up, the efficiency of FP-Tree algorithm account for three reasons. First theFP-Tree is a compressed representation of the original database because only thosefrequent items are used to construct the tree, other irrelevant information are pruned.Secondly this algorithm only scans the database twice. Thirdly, FP-Tree uses a divideand conquer method that considerably reduced the size of the subsequent conditionalFP-Tree. In [15] there are examples to illustrate all the detail of this mining process.Every algorithm has his limitations, for FP-Tree it is difficult to be used in an interactivemining system. During the interactive mining process, users may change hethreshold of support according to the rules. However for FP-Tree the changing ofsupport may lead to repetition of the whole mining process. Another limitation is thatFP-Tree is that it is not suitable for incremental mining. Since as time goes on databaseskeep changing, new datasets may be inserted into the database, those insertionsmay also lead to a repetition of the whole process if we employ FP-Tree algorithm.TreeProjection is another efficient algorithm recently proposed in [3]. The generalidea of TreeProjection is that it constructs a lexicographical tree and projects a largedatabase into a set of reduced, item-based sub-databases based on the frequent patternsmined so far. The number ofnodes in its lexicographic tree is exactly that of thefrequent itemsets. The efficiency of TreeProjection can be explained by two mainfactors: (1) the transaction projection limits the support counting in a relatively smallspace; and (2) the lexicographical tree facilitates the management and counting ofcandidates and provides the flexibility of picking efficient strategy during the treegeneration and transaction projection phrases.Wang and Tjortjis [38] presented PRICES, an efficient algorithm for mining associationrules. Their approach reduces large itemset generation time, known to be themost time-consuming step, by scanning the database only once and using logicaloperations in the process. Another algorithm for efficient generating large frequentcandidate sets is proposed by [36], which is called Matrix Algorithm. The algorithmgenerates a matrix which entries 1 or 0 by passing over the cruel database only once,and then the frequent candidate sets are obtained from the resulting matrix. Finallyassociation rules are mined from the frequent candidate sets. Experiments resultsconfirm that the proposed algorithm is more effective than Apriori Algorithm.

4.2 Sampling

Toivonen [33] presented an association rule mining algorithm using sampling. Theapproach can be divided into two phases. During phase 1 a sample of the database isobtained and all associations in the sample are found. These results are then validatedagainst the entire database. To maximize the effectiveness of the overall approach, theauthor makes use of lowered minimum support on the sample. Since the approach isprobabilistic (i.e. dependent on the sample containing all the relevant associations)not all the rules may be found in this first pass. Those associations that were deemednot frequent in the sample but were actually frequent in the entire dataset are used toconstruct the complete set of associations in phase 2.Parthasarathy [24] presented an efficient method to progressively sample for associationrules. His approach relies on a novel measure of model accuracy (self-similarityof associations across progressive samples), the identification of a representativeclass of frequent itemsets that mimic (extremely accurately) the self-similarityvalues across the entire set of associations, and an efficient sampling methodologythat hides the overhead of obtaining progressive samples by overlapping it with usefulcomputation.Chuang et al. [11] explore another progressive sampling algorithm, called SamplingError Estimation (SEE), which aims to identify an appropriate sample size formining association rules. SEE has two advantages. First, SEE is highly efficient becausean appropriate sample size can be determined without the need of executingassociation rules. Second, the identified sample size of SEE is very accurate, meaningthat association rules can be highly efficiently executed on a sample of this size toobtain a sufficiently accurate result.Especially, if data comes as a stream flowing at a faster rate than can be processed,sampling seems to be the only choice. How to sample the data and how big the samplesize should be for a given error bound and confidence level are key issues forparticular data mining tasks. Li and Gopalan [19] derive the sufficient sample sizebased on central limit theorem for sampling large datasets with replacement.

4.3 Parallelization

Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of the higher speed and greater storage capacity thatthey offer [41]. The transition to a distributed memory system requires the partitioningof the database among the processors, a procedure that is generally carried outindiscriminatelyCheung et al. [9] presented an algorithm called FDM. FDM is a parallelization ofApriori to (shared nothing machines, each with its own partition of the database. Atevery level and on each machine, the database scan is performed independently onthe local partition. Then a distributed pruning technique is employed. Schuster andWolff [29] described another Apriori based D-ARM algorithm - DDM. As in FDM,candidates in DDM are generated level wise and are then counted by each node in itslocal database. The nodes then perform a distributed decision protocol in order to findout which of the candidates are frequent and which are not.Another efficient parallel algorithm FPM (Fast Parallel Mining) for mining associationrules on a shared-nothing parallel system has beenproposed by [10]. It adoptsthe count distribution approach and has

incorporated two powerful candidate pruningtechniques, i.e., distributed pruning and global pruning. It has a simple communicationscheme which performs only one round of message exchange in each iteration. Anew algorithm, Data Allocation Algorithm (DAA), is presented in [21] that uses PrincipalComponent Analysis to improve the data distribution prior to FPM.Parthasarathy et al. [23] have written an excellent recent survey on parallel associationrule mining with shared memory architecture covering most trends, challengesand approaches adopted for parallel data mining. All approaches spelled out andcompared in this extensive survey are apriori-based. More recently, Tang and Turkia[25] proposed a parallelization scheme which can be used to parallelize the efficientand fast frequent itemset mining algorithms based on FP-trees.

4.4 Constraints based association rule mining

Many data mining techniques consist in discovering patterns frequently occurring inthe source dataset. Typically, the goal is to discover all the patterns whose frequencyin the dataset exceeds a user-specified threshold. However, very often users want torestrict the set of patterns to be discovered by adding extra constraints on the structureof patterns. Data mining systems should be able to exploit such constraints to speed upthe mining process. Techniques applicable to constraint-driven pattern discoverycan be classified into the following groups:

- post-processing (filtering out patterns that do not satisfy user-specified patternconstraints after the actual discovery process);
- pattern filtering (integration of pattern constraints into the actual mining processin order to generate only patterns satisfying the constraints);
- dataset filtering (restricting the source dataset to objects that can possibly containpatterns that satisfy pattern constraints).

Wojciechowski and Zakrzewicz [39] focus on improving the efficiency of constraint-based frequent pattern mining by using dataset filtering techniques. Datasetfiltering conceptually transforms a given data mining task into an equivalent oneoperating on a smaller dataset. Tien Dung Do et al [14] proposed a specific type ofconstraints called category-based as well as the associated algorithm for constrainedrule mining based on Apriori. The Category-based Apriori algorithm reduces thecomputational complexity of the mining process by bypassing most of the subsets ofthe final itemsets. An experiment has been conducted to show the efficiency of theproposed technique.Rapid Association Rule Mining (RARM) [13] is an association rule miningmethod that uses the tree structure to represent the original database and avoids candidategeneration process. In order to improve the efficiency of existing mining algorithms,constraints were applied during the mining process to generate only thoseassociation rules that are interesting to users instead of all the association rules.

V. CATEGORIES OF DATABASES IN WHICH ASSOCIATION RULES ARE APPLIED

Transactional database refers to the collection of transaction records, in most cases they are sales records. With the popularity of computer and e-commerce, massivetransactional databases are available now. Data mining on transactional databasefocuses on the mining of association rules, finding the correlation between items inthe transaction records.One of data mining techniques, generalized association rule mining with taxonomy,is potential to discover more useful knowledge than ordinary flat association rulemining by taking application specific information into account [27]. In particular inretail one might consider as items particular brands of items or whole groups likemilk, drinks or food. The more general the items chosen the higher one can expect thesupport to be. Thus one might be interested in discovering frequent itemsets composedof items which themselves form a taxonomy. Earlier work on mining generalizedassociation rules ignore the fact that the taxonomies of items cannot be keptstatic while new transactions are continuously added into the original database. Howto effectively update the discovered generalized association rules to reflect the databasechange with taxonomy evolution and transaction update is a crucial task. Tsenget al [34] examine this problem and propose a novel algorithm, called IDTE, whichcan incrementally update the discovered generalized association rules when the taxonomyof items is evolved with new transactions insertion to the database. Empiricalevaluations show that this algorithm can maintain its performance even in largeamounts of incremental transactions and high degree of taxonomy evolution, and ismore than an order of magnitude faster than applying the best generalized associationsmining algorithms to the whole updated database.Spatial databases usually contain not only traditional data but also the location orgeographic information about the corresponding data. Spatial association rules describethe relationship between one set of features and another set of features in aspatial database, for example (Most business centres in Greece are around City Hall), and the spatial operations that used to describe the correlation can be within, near, next to,etc. The form of spatial association rules is also $X \Rightarrow Y$, where X, Y are sets of predicatesand of which some are spatial predicates, and at

least one must be a spatialpredicate [30].Temporal association rules can be more useful and informative than basic associationrules. For example ratherthan the basic association rule {diapers}⇒{beer},mining from the temporal data we can get a more insight rule that the support of{diapers}⇒{beer} jumps to 50% during 6pm to 9pm every day, obviously retailerscan make more efficient promotion strategy by using temporal association rule. In[35] an algorithm for mining periodical patterns and episode sequential patterns wasintroduced.

VI. RECENT ADVANCES IN ASSOCIATION RULE DISCOVERY

A serious problem in association rule discovery is that the set of association rules can grow to be unwieldy as the number of transactions increases, especially if the supportand confidence thresholds are small. As the number of frequent itemsets increases,the number of rules presented to the user typically increases proportionately. Many ofthese rules may be redundant.

6.1 Redundant Association Rules

To address the problem of rule redundancy, four types of research on mining associationrules have been performed. First, rules have been extracted based on user-definedtemplates or item constraints [6]. Secondly, researchers have developed interestingness measures to select only interesting rules [17]. Thirdly, researchers have proposedinference rules or inference systems to prune redundant rules and thus present smaller,and usually more understandable sets of association rules to the user [12]. Finally,new frameworks for mining association rule have been proposed that find associationrules with different formats or properties [8].Ashrafi et al [4] presented several methods to eliminate redundant rules and toproduce small number of rules from any given frequent or frequent closed itemsetsgenerated. Ashrafi et al [5] present additional redundant rule elimination methods thatfirst identify the rules that have similar meaning and then eliminate those rules. Furthermore,their methods eliminate redundant rules in such a way that they never dropany higher confidence or interesting rules from the resultant rule set.Jaroszewicz and Simovici [18] presented another solution to the problem using theMaximum Entropy approach. The problem of efficiency of Maximum Entropy computationsis addressed by using closed form solutions for the most frequent cases.Analytical and experimental evaluation of their proposed technique indicates that itefficiently produces small sets of interesting association rules.Moreover, there is a need for human intervention in mining interesting associationrules. Such intervention is most effective if the human analyst has a robust visualizationtool for mining and visualizing association rules. Techapichetvanich and Datta[31] presented a three-step visualization method for mining market basket associationrules. These steps include discovering frequent itemsets, mining association rules andfinally visualizing the mined association rules.

62 Other measures as interestingness of an association

Omiecinski [22] concentrates on finding associations, but with a different slant. That is, he takes a different view of significance. Instead of support, he considers othermeasures, which he calls all-confidence, and bond. All these measures are indicatorsof the degree to which items in an association are related to each other. With allconfidence,an association is deemed interesting if all rules that can be produced fromthat association have a confidence greater than or equal to a minimum all-confidencevalue. Bond is another measure of the interestingness of an association. With regardto data mining, it is similar to support but with respect to a subset of the data ratherthan the entire data set. This has similarities to the work in [26] except in their workthey define data subsets based on the data satisfying certain time constraints. The ideais to find all itemsets that are frequent in a set of user-defined time intervals. In thiscase, the characteristics of the data define the subsets not the end-user. Omiecinski[22] proved that if associations have a minimum all-confidence or minimum bond,then those associations will have a given lower bound on their minimum support andthe rules produced from those associations will have a given lower bound on theirminimum confidence as well. The performance results showed that the algorithm canfind large itemsets efficiently.In [8], the authors mine association rules that identify correlations and considerboth the absence and presence of items as a basis for generating the rules. The measureof significance of associations that is used is the chi-squared test for correlationfrom classical statistics. In [7], the authors still use support as part of their measure ofinterest of an association. However, when rules are generated, instead of using confidence,the authors use a metric they call conviction, which is a measure of implicationand not just co-occurrence.In [20], the authors present an approach to the rare item problem. The dilemmathat arises in the rare item problem is that searching for rules that involve infrequent(i.e., rare) items requires a low support but using a low support will typically generatemany rules that are of no interest. Using a high support typically reduces the numberof rules mined but will eliminate the rules with rare items. The authors attack thisproblem by allowing users to specify different minimum supports for the variousitems in their mining algorithm.

6.3 Negative Association Rules

Typical association rules consider only items enumerated in transactions. Such rules are referred to as positive association rules. Negative association rules also considerthe same items, but in addition consider negated items (i.e. absent from transactions).Negative association rules are useful in market-basket analysis to identify productsthat conflict with each other or products that complement each other. Mining negativeassociation rules is a difficult task, due to the fact that there are essential differencesbetween positive and negative association rule mining. The researchers attack twokey problems in negative association rule mining: (i) how to effectively search forinteresting itemsets, and (ii) how to effectively identify negative association rules ofinterest.Brinet. al [8] mentioned for the first time in the literature the notion of negative relationships.Their model is chi-square based. They use the statistical test to verify theindependence between two variables. To determine the nature (positive or negative)of the relationship, a correlation metric was used. In [28] the authors present a newidea to mine strong negative rules. They combine positive frequent itemsets withdomain knowledge in the form of taxonomy to mine negative associations. However,their algorithm is hard to generalize since it is domain dependant and requires a predefinedtaxonomy. A similar approach is described in [37]. Wu et al [40] derived anew algorithm for generating both positive and negative association rules. They addon top of the support-confidence framework another measure called mininterestfor abetter pruning of the frequent itemsets generated. In [32] the authors use only negativeassociations of the type $X \Rightarrow \neg Y$ to substitute items in market basket analysis.

ACKNOWLEDGMENT

I would be thankful to my guide assistant professor VinitKumar Gupta here for help when I have some troubles in paper writing. I will also thanksto Mr. PuspakRaval (Assistant Professor, DAIICT,Gandhinagar) and my other faculty members and class mates for their concern and support both in study and life.

CONCLUSION

Association rule mining has a wide range of applicability such market basket analysis,medical diagnosis/ research, Website navigation analysis, homeland security andso on. In this paper, we surveyed the list of existing association rule mining techniques.The conventional algorithm of association rules discovery proceeds in twosteps. All frequent itemsets are found in the first step. The frequent itemset is theitemset that is included in at least **minsup**transactions. The association rules with theconfidence at least **minconf**are generated in the second step.End users of association rule mining tools encounter several well-known problemsin practice. First, the algorithms do not always return the results in a reasonable time.It is widely recognized that the set of association rules can rapidly grow to be unwieldy,especially as we lower the frequency requirements. The larger the set of frequentitemsets the more the number of rules presented to the user, many of which areredundant. This is true even for sparse datasets, but for dense datasets it is simply notfeasible to mine all possible frequent itemsets, let alone to generate rules, since theytypically produce an exponential number of frequent itemsets; finding long itemsetsof length 20 or 30 is not uncommon. Although several different strategies have beenproposed to tackle efficiency issues, they are not always successful.

REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules betweensets of items in large databases. *In Proceedings ofthe* 1993 *ACM SIGMODInternational Conference on Management of Data*, 207-216.

[2] Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. *In Proc.20th Int. Conf. Very Large Data Bases*, 487-499.

[3] Agarwal, R. Aggarwal, C. and Prasad V., A tree projection algorithm for generation of frequentitemsets. *In J. Parallel and Distributed Computing*, 2000.

[4] Ashrafi, M., Taniar, D. Smith, K. A New Approach of Eliminating Redundant AssociationRules, *Lecture Notes in Computer Science,* Volume 3180, 2004, Pages 465 – 474.

[5] Ashrafi, M., Taniar, D., Smith, K., Redundant Association Rules Reduction Techniques,*Lecture Notes in Computer Science*, Volume 3809, 2005, pp. 254 – 263.

[6] Baralis, E., Psaila, G., Designing templates for mining association rules. *Journal of IntelligentInformation Systems*, 9(1):7-32, July 1997.

[7] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. 1997. Dynamic itemset counting andimplication rules for market basket data. In SIGMOD 1997, *Proceedings ACM SIGMODInternational Conference on Management of Data*, May 13-15, 1997, 255-264.

[8] Brin, S., Motwani, R. and Silverstein, C.,"Beyond Market Baskets: Generalizing AssociationRules to Correlations," *Proc. ACM SIGMOD Conf., pp*. 265-276, May 1997.

[9] Cheung, D., Han, J., Ng, V., Fu, A. and Fu, Y. (1996), A fast distributed algorithm formining association rules. *In `Proc. of* 1996 *Int'l. Conf. on Parallel and Distributed InformationSystems'*, Miami Beach, Florida, pp. 31 - 44.

[10] Cheung, D., Xiao, Y., Effect of data skewness in parallel mining of association rules,*Lecture Notes in Computer Science*, Volume 1394, Aug 1998, Pages 48 – 60.

[11] Chuang, K., Chen, M., Yang, W., Progressive Sampling for Association Rules Based onSampling Error Estimation, *Lecture Notes in Computer Science*, Volume 3518, Jun 2005,Pages 505 – 515.

[12] Cristofor, L., Simovici, D., Generating an informative cover for association rules. *In Proc.of the IEEE International Conference on Data Mining*, 2002.

[13] Das, A., Ng, W.-K., and Woon, Y.-K. 2001. Rapid association rule mining. *In Proceedings*of *the tenth international conference on Information and knowledge management*.ACM Press, 474-481.

[14] Tien Dung Do, Siu Cheung Hui, Alvis Fong, Mining Frequent Itemsets with Category-Based Constraints, *Lecture Notes in Computer Science*, Volume 2843, 2003, pp. 76 – 86.

[15] Han, J. and Pei, J. 2000. Mining frequent patterns by pattern-growth: methodology andimplications. *ACM SIGKDDExplorations Newsletter* 2, 2, 14-20.

[16] Hegland, M., Algorithms for Association Rules, *Lecture Notes in Computer Science*,Volume 2600, Jan 2003, Pages 226 – 234.

[17] Hilderman R. J., Hamilton H. J., Knowledge Discovery and Interest Measures, *KluwerAcademic, Boston*, 2002.

[18] Jaroszewicz, S., Simovici, D., Pruning Redundant Association Rules Using MaximumEntropy Principle, *Lecture Notes in Computer Science*, Volume 2336, Jan 2002, pp 135-142.

[19] Li, Y., Gopalan, R., Effective Sampling for Mining Association Rules, *Lecture Notes inComputer Science*, Volume 3339, Jan 2004, Pages 391 – 401.

[20] Liu, B. Hsu, W., Ma, Y., "Mining Association Rules with Multiple Minimum Supports,"*Proc. Knowledge Discovery and Data Mining Conf., pp*. 337-341, Aug. 1999.

[21] Manning, A., Keane, J., Data Allocation Algorithm for Parallel Association Rule Discovery,*Lecture Notes in Computer Science*, Volume 2035, Page 413-420.

[22] Omiecinski, E. (2003), Alternative Interest Measures for Mining Associations in Databases,*IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 1, pp. 57-69.

[23] Parthasarathy, S., Zaki, M. J., Ogihara, M., Parallel data mining for association rules onshared-memory systems. *Knowledge and Information Systems: An International Journal*,3(1):1–29, February 2001.

[24] Parthasarathy, S., Efficient Progressive Sampling for Association Rules. *ICDM* 2002:354-361.

[25] Tang, P., Turkia, M., Parallelizing frequent itemset mining with FP-trees. *Technical Reporttitus.compsci.ualr.edu/~ptang/papers/par-fi.pdf*, Department of Computer Science,University of Arkansas at Little Rock, 2005.

[26] Ramaswamy, S., Mahajan, S., Silbershatz, A.,"On the Discovery of Interesting Patterns inAssociation Rules*," Proc. Very Large Databases Conf., pp*. 368-379, Sept. 1998.

[27] Sarawagi, S., Thomas, S., "Mining Generalized Association Rules and Sequential PatternsUsing SQL Queries*". In Proc. of KDD Conference*, 1998.

[28] Savasere, A., Omiecinski, E., Navathe, S.: Mining for strong negative associations in alarge database of customer transactions. *In: Proc. of ICDE*. (1998) 494–502.

[29] Schuster, A. and Wolff, R. (2001), Communication-efficient distributed mining of associationrules, *In `Proc. of the* 2001 *ACM SIGMODInt'l. Conference on Management of Data'*, Santa Barbara, California, pp. 473-484.

[30] Sharma, L.K., Vyas, O.P., Tiwary, U.S., Vyas, R. A Novel Approach of Multilevel Positiveand Negative Association Rule Mining for Spatial Databases, *Lecture Notes in ComputerScience*, Volume 3587, Jul 2005, Pages 620 – 629.

[31] Techapichetvanich, K., Datta, A., Visual Mining of Market Basket Association Rules,*Lecture Notes in Computer Science*, Volume 3046, Jan 2004, Pages 479 – 488.

[32] Teng, W., Hsieh, M., Chen, M.: On the mining of substitution rules for statistically dependentitems. *In: Proc. of ICDM*. (2002) 442–449.

[33] Toivonen, H. (1996), Sampling large databases for association rules, in *`The VLDB Journal',pp*. 134-145.

[34] seng, M., Lin, W., Jeng, R., Maintenance of Generalized Association Rules UnderTransaction Update and Taxonomy Evolution, *Lecture Notes in Computer Science*, Volume 3589, Sep 2005, Pages 336 – 345.

[35] Verma, K., Vyas, O.P., Vyas, R., Temporal Approach to Association Rule Mining Using F P-Tree, *Lecture Notes in Computer Science*, Volume 3587, Jul 2005, Pages651 – 659..

[36] Yuan, Y., Huang, T., A Matrix Algorithm for Mining Association Rules, *Lecture Notes inComputer Science*, Volume 3644, Sep 2005, Pages 370 – 379.

[37] Yuan, X., Buckles, B., Yuan, Z., Zhang, J.: Mining negative association rules. *In: Proc. Of ISCC.* (2002) 623–629.

[38] Wang, C., Tjortjis, C., PRICES: An Efficient Algorithm for Mining Association Rules,*Lecture Notes in Computer Science*, Volume 3177, Jan 2004, Pages 352 – 358.

[39] Wojciechowski, M., Zakrzewicz, M., Dataset Filtering Techniques in Constraint-BasedFrequent Pattern Mining, *Lecture Notes in Computer Science*, Volume 2447, 2002, pp.77-83.

[40] Wu, X., Zhang, C., Zhang, S.: Efficient Mining of Both Positive and Negative AssociationRules, *ACM Transactions on Information Systems*, Vol. 22, No. 3, July 2004, Pages 381–405.

[41] Zaki, M. J., Parallel and distributed association mining: *A survey. IEEE Concurrency*,*Special Issue on Parallel Mechanisms for Data Mining*, 7(4):14--25, December 1999.

[42] P.Samarati,(2001). Protecting respondent's privacy in micro data release. *In IEEE Transaction on knowledge and Data Engineering, pp*.010-027.

[43] L. Sweeney, (2002)."K-anonymity: a model for protecting privacy ", *International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp*. 557-570.

[44] Agrawal, R. and Srikant, R, (2000)."Privacy-preserving data mining. "*In Proc. SIGMOD00*, pp. 439-450.

[45] Hong, J.I. and J.A. Landay2004).Architecture for Privacy Sensitive Ubiquitous Computing", *In Mobisys04, pp*. 177- 189.

[46] Evfimievski, A.Srikant, R.Agrawal, and Gehrke J(2002),"Privacy preserving mining of association rules*". In Proc.KDD02,*
pp. 217-228.

[47] Laur, H. Lipmaa, and T. Mieli' ainen,(2006)."Cryptographically private support vector machines*". In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp*. 618-624.

[48] Ke Wang, Benjamin C. M. Fung1 and Philip S. Yu, (2005) "Template based privacy preservation in classification problems", *In ICDM, pp*. 466-473.

[49] Chen, X. and Orlowska, M. A further study on inverse frequent set mining. *In: Proc. of the1st Int'l Conf. on Advanced Data Mining and Applications (ADMA'05). LNCS 3584, Springer-Verlag*. 2005. 753-760.

[50] MohammadRezaKeyvanpour, SomayyehSeifiMoradi. Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework.*International Journal on Computer Science and Engineering (IJCSE).* Feb 2011, ISSN 0975-3397.