Survey on privacy preserving association rule hiding in data mining

Jinal Patel[1], Gurcharan Sahani[2]
[1]*Computer Engineering, Sardar Vallabhbhai Institude Of Technology*
[2]*Assi. Prof., Computer Engineering, Sardar Vallabhbhai Institude Of Technology*
Vasad, India

***Abstract-* Data mining refers to mining information from large volume of data. Association rule mining is a technique in data mining that finds identifies the regularities found in large volume of data . such technique may identify and reveal hidden information that is private for an individual or organization. privacy preserving association rule mining needs to prevent disclose not only the confidential or a personal information from the aggregated data also to prevent data mining techniques from discovering sensitive information. Association rule hiding is one of the techniques of privacy preserving data mining to protect the association rules generated by association rule mining. In this paper we present survey on privacy preserving association rule hiding techniques and their merit and demerits.**

***Index Terms-* Privacy preserving ,association rule hiding, sensitivity**

I. INTRODUCTION

Data Mining [1] refers to extracting or mining knowledge from large amounts of data. Data mining is the process of discovering the intensive knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. It is an interdisciplinary field and it involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, pattern recognition, information retrieval, and spatial or temporal data analysis. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. So, there might be a conflict among data mining and privacy. Privacy [2] refers to extraction of sensitive information using data mining. On the other hand, the excessive processing power of intelligent algorithms puts the sensitive and confidential information that resides in large and distributed data stores at risk. Recent developments in information technology have enabled the collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. Undoubtedly, this information is very useful in many areas, including medical research, law enforcement and national security. However, there is an increasing public concern about the individuals' privacy. Privacy is commonly seen as the right of individuals to control information about themselves. The general privacy issues are secondary use of the personal information, handling misinformation, and granulated access to personal information. The concept of privacy preserving data mining involves in preserving personal information from data mining algorithms. PPDM technique is a research area in data mining and statistical databases where mining algorithms are analyzed for the side effect they acquire in data privacy.

II.PRIVACY PRESERVING DATA MINING (PPDM)

Data mining is the process of gathering information about the user specific data, also called knowledge discovery, on internet. The problem with data mining output is that it also discloses some information, which is considered to be private and personal. Effortless access to such personal data causes a peril to individual privacy. Official statistics, Health information, and E-commerce are some key concern for privacy. Privacy preserving data mining technique gives novel way to solve this problem. The main purpose of privacy preserving data mining is to design competent frameworks and algorithms that can extract relevant knowledge from a large amount of data without revealing of any sensitive information. It protects sensitive information by providing sanitized database of original database on the internet or a process is used in such a way that private data and private knowledge remain private even after the mining process. It is PPDM due to which the benefits of data mining be enjoyed, without compromising the privacy of concerned individuals. PPDM Techniques can be classified over five dimensions .The first dimension is related to distribution of data i.e. Centralized or Distributed. The second dimension refers to the modification of original values of data that are to be released for data mining task. Modification is carried out using perturbation, blocking, aggregation, merging, swapping or sampling or any combination of these. The third dimension is that of data mining algorithms. The data mining algorithm are applied on the transformed data to get useful nuggets of information that were hidden previously. The fourth dimension refers to whether the raw data or aggregated data should be hidden. The fifth and the final dimension refer to the techniques that are used for protecting privacy. Based on these dimensions, different PPDM techniques may be classified into following five categories[6]

1. Anonymization based PPDM
2. Randomized Response based PPDM
3. Condensation approach based PPDM
4. Cryptography based PPDM
5. Perturbation based PPDM

A. Anonymization based PPDM

Anonymization

To protect individuals’ identity when releasing sensitive information, data holders often remove explicit identifiers. These data may accidentally disclose the sensitive data about the individuals. The risk of linking the private information is handled by various privacy preserving techniques. The main concern is sensitive information which should not be disclosed. There are two types of disclosures such as, identity disclosure and attribute disclosure. Identity disclosure happens when an individual is uniquely identified from the published data. Attribute disclosure happens when the information of an individual can be inferred from the published data. The number of privacy models are discussed which are succeeded in solving the problems such as attribute disclosure and identity disclosure by preserving private information. Some of popular techniques such as k-anonymity, l-diversity and tcloseness models. Generalization involves replacing a value with a less specific (generalized) but semantically reliable value. For example, the age of a person could be generalized to a range such as youth, middle age and adult without specifying appropriately, so as to reduce the risk of identification. Suppression involves reduction in exactness of applications and it doesn’t liberate any information .By using this method it reduces the risk of detecting exact information. In kanonymity, it is difficult for an imposter to determine the identity of the individuals in collection of data set containing personal information. Each release of data contains every]combination of values of quasi-identifiers and that is indistinctly matched to at least k-1 respondents [4]

Table 1 original patterns table

S.NO	ZIP CODE	AGE	DISEASE
1	546177	39	Heart disease
2	546102	32	Heart disease
3	546178	37	Heart disease
4	549105	53	Gastritis
5	549209	62	Heart disease
6	546205	57	Cancer
7	546205	40	Heart disease
8	546273	46	Cancer
9	546207	42	Cancer

Table 2. A3 –Anonymous version table 1

S.NO	ZIP CODE	AGE	DISEASE
1	546***	3*	Heart disease
2	546***	3*	Heart disease
3	546***	3*	Heart disease
4	549***	>=50	Gastritis
5	549***	>=50	Heart disease
6	546***	>=50	Cancer
7	546***	4*	Heart disease
8	546***	4*	Cancer
9	546***	4*	Cancer

B. Randomized Response based PPDM

TheRandomization Method The randomization technique uses the data distortion methods in order to create private representations of the records. In most of the cases, the individual records cannot be recovered, but aggregate distributions can be recovered.

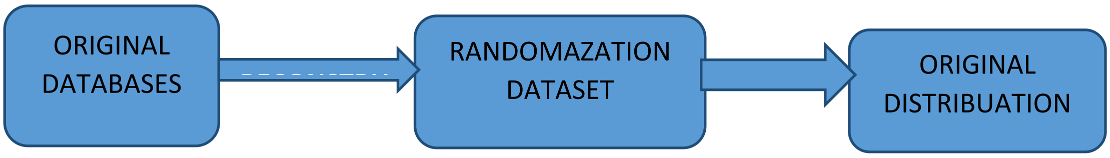

Fig 1. Ramdomazation method

These distributions can be used for data mining purposes. Two kinds of perturbation are possible with the randomization methods:

Additive Perturbation: In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms designed to work with these data distributions.

Multiplicative Perturbation: In this case, either the random projection or random rotation techniques is used in order to perturb the records.

C. Condensation approach based PPDM Condensation approach constructs constrained clusters in dataset and then generates fake data from the statistics of these clusters It is called as condensation because of its approach of using condensed statistics of the clusters to generate fake data. It constructs groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its secrecy level. Subsequently, fake data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data This approach can be effectively used for the problem of classification . The use of fake data provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data. Moreover, the aggregate behavior of the data is preserved, making it useful for a variety of data mining problems This approach helps in better privacy preservation as compared to other techniques as it uses fake data rather than modified data. Moreover, it works even without redesigning data mining algorithms since the fake data has the same format as that of the original data. It is very effective in case of data stream problems where the data is highly dynamic. At the same time, data mining results get affected as large amount of information is lost because of the condensation of a larger number of records into a single statistical group entity

D. Cryptography based PPDM Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information Cryptographic techniques find its utility in such scenarios because of two reasons : First, it offers a well defined model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. Although cryptographic techniques ensure that the transformed data is exact and secure but this approach fails to deliver when more than a few parties are involved

E. Perturbation based PPDM Perturbation has a long back history, being used in statistical disclosure control as it has an inherent property of simplicity, efficiency and ability to preserve statistical information . In perturbation the original values are replaced with some synthetic data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not correspond to real-world record owners, so the attacker cannot perform the sensitive linkages or recover sensitive information from the published data. In perturbation approach, records released is synthetic i.e. it does not correspond to real world entities represented by the original data. Therefore the individual records in the perturbed data are meaningless to the human recipient as only statistical properties of the records are preserved. Perturbation can be done by using additive noise or data swapping or synthetic data generation. Since the perturbation method does not reconstruct the original values but only the distributions, new algorithms are to be developed for mining of the data .

Table 3 Comparative Study of existing approaches

TECHNIQUES	Merit	Demerit
1. Anonymization based PPDM	This method is protects identity disclosure when it is releasing sensitive information	It is prone to homogeneity attack and the background knowledge attack. Does not protect attribute disclosure to sufficient extent It has the limitation of k-anonymity model which fails in real scenario when the attackers try other methods
2.Randomized Response based PPDM	It is easily implemented in data collection phase and all the other knowledge is not required and there is no need of server. It is useful for hiding individual sensitive data.	It is not required for multiple attribute databases It results in high information loss .
3.Condensation approach based PPDM	This approach works with pseudo-data rather than with modifications of original data	The pseudo-data have the same format as the original data. So, it is no longer necessitates the redesign of data mining algorithms
4. Cryptography based PPDM	Cryptography offers a well-defined model for privacy for proving and quantifying it. There exit a vast range of cryptographic algorithms	It is difficult to scale when more than a few parties are involved It does not guarantee that the disclosure of the final data mining result may not violate the privacy of individual records.
5. Perturbation based PPDM	It is very simple technique. Different attributes are treated independently	Does not reconstruct the original vale rather than only distortion The perturbation approach does not provide a clear understanding of the level of indistinguishability of different records

III. ASSOCIATION RULE MINING

Let I = {i1,…., in} be a set of items. Let D be a database which contains set of transactions. Each transaction t _ D is an item set such that t is a proper subset of I. As transaction t supports X, a set of items in I, if X is a propersubset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X_Y, where X and Y are subsets of I and X_Y= Ø. The support of rule X_Y can be calculated by the following equation: Support(X_Y) = |X_Y| / |D|, where |X_Y| denotes the number of transactions containing the itemset XY in the database, |D| denotes the number of the transactions in the database D. The confidence of rule is computed by Confidence(X_Y) = |X_Y|/|X|, where |X| is number of transactions in database D that contains itemset X. A rule X_Y is strong if support(X_Y) _ min_support and confidence(X_Y) _ min_confidence, where min_support and min_confidence are two given minimum thresholds.

Association rule mining algorithms calculate the support and confidence of the rules. The rules having support and confidence higher than the user specified minimum support and confidence are retrieved. Association rule hiding algorithms prevents the sensitive rules from being revealed out. The problem can be declared as follows "Database D, minimum confidence, minimum support are given and a set R of rules are mined from database D. A subset SR of R is denoted as set of sensitive association rules.SR is to be hidden. The objective is to modify D into a database D' from which no association rule in SR will be mined and all non sensitive rules in R could still be mined from D[1]

IV.PRIVACY PRESERVING ASSOCIATION RULE MINING

Privacy preserving association rule mining needs to prevent disclosure not only of confidential personal information from original or aggregated data, but also to prevent data mining techniques from discovering sensitive knowledge. It is known that each strong rule extracts from frequent itemsets. To prevent sensitive rules (determined by the experts) being mined in the process of association rule mining, many methods are developed all of which are based on reducing the support and confidence of rules that specify how significant they are. In order to achieve this goal, transactions are modified by removing some items, or inserting new items depending on the hiding strategy. In the following, we will discuss some general methods for hiding sensitive rules. The conceptual framework for association rule hiding is shown in the fig2

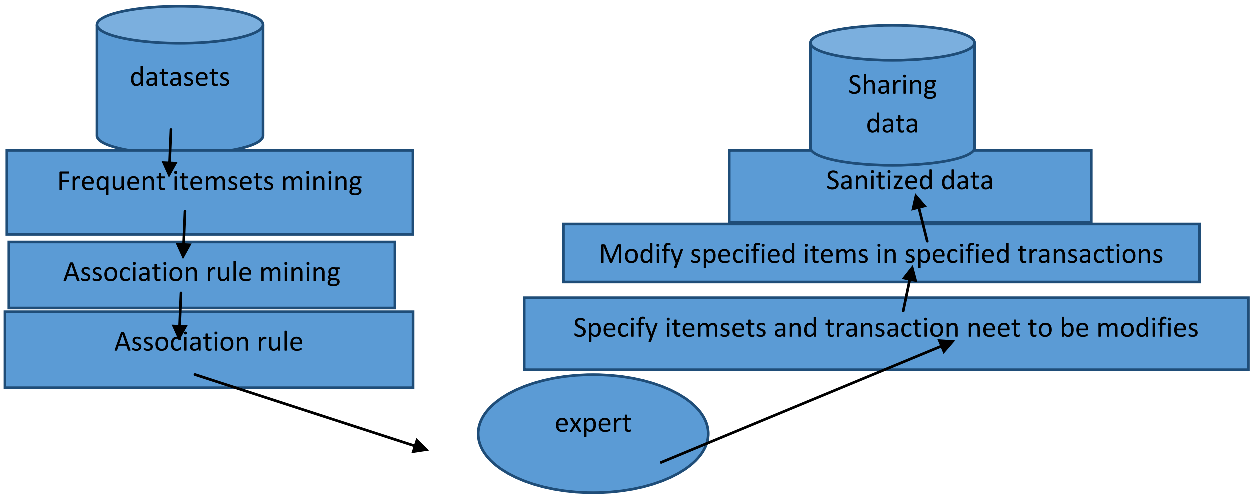

Fig 2.Conceptual framework for association rule hiding

This approach hides rules having sensitive items either in the right side or in the left side. The common approaches used in association rule hiding algorithms are [11]

1) Heuristic approaches
2) Border-based approaches
3) Exact approaches.

(1).Heuristic Approach

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

(2) Border Based Approaches:

Border based approach is hide the sensitive association rules by modifying the border in the lattice of the frequent and the infrequent item sets in original database[3]. this approach is make the border between the frequent and infrequent items. that way this border is divided the frequent and in frequent item sets. The first frequent item set hiding methodology that is based on the notion of the border[l]. It maintains the quality of database by greedily selecting the modifications with minimal side effect.

(3) Exact Approaches:

Another name of the exact approach is the non-heuristic algorithm which is formulated to constrain satisfaction problem(CSP) and solve by using the binary integer programming(BIP). It Provide the optimal solution to all constrain. In [4] is first used the exact approach for hide the rules. and it provides an optimal solution of rule hiding problem. In [7] to hide sensitive rules by formulating constraint satisfaction problem without any side effects with the concepts of positive and negative border sets. By using adopting divide and conquer technique on constraints.

Table 4. Comparative Study of existing approaches rule hiding

ALGORITHM	Merit	Demerit
1) Heuristic approaches	It is a more efficient, fast and scalable.	This approach is difficult to Handel the changes in Database
2) Border-based approaches	It maintain the database quality by selecting the modification with the minimal side effect.	The border is not easily identify. then it is difficult to understand based on the heuristic approaches
3)exact approaches	it gives guarantees to provides the optimal solution without any side effect	the approach is require the high complexity due to the binary integer programming.

V. METRICS AND PERFORMANCE ANALYSIS

The performance of two hiding algorithms can be compared on the basis of following Metrics. These Metrics determine the most efficient algorithm between the two.[11]

A. Hiding Failure (HF) It is the percentage of the sensitive data that remain exposed in the sanitized dataset. It is defined as the fraction of the restrictive association rules that appear in the sanitized database divided by the ones that appeared in the original dataset.

$$HF = \frac{\#\mathrm{Sp}(\mathrm{D}')}{\#\mathrm{Sp}(\mathrm{D})}$$

Formally, Where |SR (D')| is number of the sensitive rules discovered in the sanitized dataset D', |RR (D)| is the number of sensitive rules appearing in the original dataset D. Ideally, the hiding failure should be 0%.

B. LOST RULES :

when some non-sensitive patterns falsely hidden during the hiding process, we call this Lost Rules, is the percentage of non-sensitive patterns that cannot be discovered from sanitize dataset D` and can be measured by formula as shown in equation

$$LRs = \frac{\#\mathrm{Sp}(\mathrm{D}) - \#\mathrm{Sp}(\mathrm{D}')}{\#\mathrm{Sp}(\mathrm{D})}$$

Where # ~ Sp (D), denotes the number of non-sensitive association rules or pattern discovered from database D. Moreover, the lost rules and hiding failure is directly proportional. Similarly, the more sensitive pattern we hide, the more non-sensitive pattern we loss.

C. GHOST RULE:

when some unwanted patterns discover during hiding process, we call this Ghost Rules, is the percentage of artificial pattern that is discovered from sanitize dataset D` but not discover from original dataset D. It is measured by formula as shown in equation

$$GRs = \frac{|\mathrm{p}| - |\mathrm{p} \sqcap \mathrm{p}'|}{|\mathrm{p}'|}$$

Where | P |, denotes the number sensitive patterns discovered from D and | P |, denotes the number of artificial patterns discovered from D`.

D. Recovery Factor (RF) :

This measure expresses the possibility of an adversary to recover a sensitive rule based on the non-sensitive ones. The recovery factor of a pattern takes into account the existence of its subsets. If all the subsets of a sensitive rule can be recovered from the sanitized dataset, then the recovery of the rule itself is possible, thus it is assigned an RF value of 1; otherwise RF = 0. However, this measure is not certain since, for instance, an adversary may not learn an itemset despite knowing its subsets..These "process performance" measures are clustered into four categories, as follows:

1) Efficiency: This category consists of measures that quantify the ability of a privacy preserving algorithm to efficiently use the available resources and execute with good performance. Efficiency is measured in terms of CPU-time, space requirements (related to the memory usage and the required storage capacity) and communication requirements.

2) Scalability: This category consists of measures that evaluate how effectively the privacy preserving technique handles increasing sizes of the data from which information needs to be mined and privacy needs to be ensured. Scalability is measured based on the decrease in the performance of the algorithm or the increase of the storage

requirements along with the communications cost (if in a distributed setting), when the algorithm is provided with larger datasets.

3) Data Quality: The data quality of a privacy preservation algorithm depends on two parameters. There are the qualities of the dataset after the sanitization process, and the quality of the data mining results when applied to this dataset, compared to the ones attained when using the original dataset. Among the various possible measures for the quantification of the data quality, the most preferable are: (i) accuracy, which measures the proximity of a sanitized value to the original one and is closely related to the information loss resulting from the hiding strategy, (ii) completeness, which is used to evaluate the degree of missed data in the sanitized database and (iii) consistency, which is related to the relationships that must continue to hold among the different fields of a data item or among data items in a sanitized database.

4) Privacy Level: This category consists of measures that estimate the degree of uncertainty according to which, the protected information can still be predicted. Measures, such as the information entropy, the level of privacy and the measure , are some among the possible metrics that one can apply to quantify the privacy level attained by a hiding scheme

VII OVERVIEW OF HEURISTIC APPROACH FOR HIDING THE SENSITIVE RULE

Heuristic approaches hide sensitive association rules by directly modifying, or we say, sanitizing the original database DB, and get the released database DB' directly from DB. Heuristic Based Approaches can be divided into two groups based on data modification techniques: data distortion techniques and data blocking techniques.[5]

.

(i) Data distortion: This type is done by alteration of old attribute value to new value. It changes 1 's to D's or vice versa in selected transactions to increase or decrease support or confidence of sensitive rule[I ,2]. In heuristic approach give optimum solution because of some side effect to non sensitive rule. two basic technique for data is reduced the confidence of rule and reduce the support of the rule. example explain........ the problem is finding an optimum solution using NP-hard. Proposes algorithm using the heuristic based data distortion technique.

Advantages: It is a more efficient, fast and scalable.

Disadvantages: This approach is difficult to Handel the changes in Database.

(ii) Data Blocking: This technique is using the maximum confidence or not reduce the sensitive rule. In database there are D's and I's must be hidden during blocking, because D's or 1 's replace with "?". In some applications where publishing wrong data is not acceptable, then unknown values may be inserted to blur the rules. so, that support of certain items goes down to certain level and rule mining algorithm nit able to mine the sensitive rules[5].

Advantages: It Maintain database , instead of inserting false value to block the original value.

Disadvantages: Difficult to reproduce the original database. and it is the various side effect like lost rule, ghost rule, false rule etc.

VIII.RELATED WORKS

chirag N. Modi, Udai Pratap Rao and Dhiren R Patel, They proposed heuristic algorithm its name was DSRRC which is hide rule only the certain level. the heuristic algorithm better used then the other hiding algorithm. DSRRC algorithm is hide rule that rule contain single item on R.H.S of the rule. [3]

Komal shah, Amit Thakkar, Amit Ganatra, they proposed two algorithm ADSRRC and RRLR to hiding sensitive association rule. this two algorithm are overcome the limitation of DSRRC algorithm. ADSRRC overcomes limitation of multiple sorting in database as well as it selects transaction to be modified based on different criteria than DSRRC algorithm. Algorithm RRLR overcomes limitation of hiding rules having multiple R.H.S. items. they also reduce the side effect and complexity.[2]

Dharmendra Thakur, Prof. Hitesh Gupta , they present various association rule hiding algorithm. And they work on the heuristic based algorithm because this approach is efficient, fast algorithm to hide the sensitive knowledge. This paper are using the support based algorithm and confidence based algorithm.[11]

R.Natarajan, Dr.R.Sugumar, M.Mahendran, K.Anbazhagan, They using the privacy preserving data mining to provide the confidentiality and improve the performance of at the time when database stores and retrieves huge amount of data. they present ISL(Increase support of L.H.S) and DSR(Decrease support of R.H.S).In which DSR algorithm to hide useful association rule from transactions data with binary attributes. and ISL algorithm confidence of a rule is decreased by increasing the support value of LHS of the rule. they conclude that using this two algorithm only the 19% efficiency increase[13]

Vikram Garg, Anju Singh, Divakar Singh , Data mining is extract the hidden predictive information from the data warehouse without revealing their sensitive information. they show all PPDM(privacy-preserving data mining) technique and also show the recently research area to deal with the association rule hiding. Show the comparative analysis of the all association rule hiding approaches.[12]

Table 5.Comparatitive study algorithm

parameter	DSRRC	MDSRRC	ADSRRC	RRLR
Hf	**0**	**0**	**0**	**0**
Mc	**36**	**26.66**	**36.36**	**22.73**
ap	**0**	**0**	**0**	**0**
Diss(D,D’)	**6.5**	**5.4**	**5.40**	**0**
sef	**36.5**	**26.66**	**11.00**	**20**

IX.CONCLUSION

In this survey, we have discussed the basic of PPDM and its different approaches. Subsequently, association rule hiding approaches and metrics for performance comparison of those approaches are discussed. Before we conclude our study we have provided an overview of heuristic approaches. These approaches involve efficient, fast algorithms to hide the sensitive knowledge. Due to their efficiency and scalability, the heuristic approaches have been the focus of attention for the vast majority of researchers in the knowledge hiding field. Different algorithms can be designed and developed by taking ideas from existing algorithms and compare their efficiency using metrics

REFERENCES

[1]J. Han and M. Kamber , “Data Mining: Concepts and Techniques”, 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[2] Komal shah, Amit Thakkar, Amit Ganatra," Association Rule Hiding by Heuristic Approach to Reduce Side Effects & Hide Multiple R.H.S. Items " International Journal of Computer Applications (0975 - 8887) Volume 45- No.1, May 2012.

[3] Chirag N. Modi, Udai Pratap Rao, Dhiren R. Patel, "Maintaining Privacy and Data Quality in Privacy Preserving AssociationRule Mining" Second International conference on Computing, Communication and Networking Technologies,2010.

[4] L.Sweeney, “Achieving k-Anonymity Privacy Protection Using Generalization and Suppression”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5.

[5]bhumika mistry, amit desai “Privacy preserving heuristic approach For Association Rule Mining in Distributed Database”IEEE 2015

[6] K.Saranya ,K.Premalatha S.S.Rajasekar “A Survey on Privacy Preserving Data Mining”IEEE 2015

[7] S. Kasthuri ,T. Meyyappa “Detection of Sensitive Items in Market Basket Database using Association Rule Mining for Privacy Preserving” IEEE 2013

[8] “A Conceptual Framework for Privacy Preserving of Association Rule Mining in E-Commerce “ Hai Quoc Le IEEE 2013

[9]” Methods and Techniques to Protect the Privacy Information in Privacy Preservation Data Mining “ N.Punitha, R.Amsaveni Ijcit 2014

[10] “A Review on “Privacy Preservation Data Mining (PPDM) “Dwipen Laskar ,Geetachri Lachit IJCAT 2014

[11] Dharmendra Thakur, prof. Hitesh Gupta, " An Exemplary Study of Privacy Preserving Association Rule Mining Techniques" International Journal of Computer Science and software engineering Volume 3, Issue 11, November 2013

[12]Vikram Garg, Anju Singh, Divakar Singh, " ASurvey of Association Rule Hiding Algorithms", Fourth International Conference on Communication Systems and Network Technologies,IEEE 2014.

[13] R.Natarajan, Dr.R.Sugumar, M.Mahendran, K.Anbazhagan, " Design and Implement an Association Rule hiding Algorithm for Privacy Preserving Data Mining ", international Journal of Advanced Research in Computer and Communication Engineering Vol. I, Issue 7, September 2012