ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 9, November 2012

Emphasizing the Data Quality with Privacy Preservation

K.PRIYANKA [1], S. ASHA VARMA[2],T.RAJESH[3]

[1]Asst.Prof,Department of computer science ,Andhra Loyola Institute Of Engineering And Technology,Vijayawada,Andhra Pradesh,India.

[2]Asst.Prof,Department of computer science ,Andhra Loyola Institute Of Engineering And Technology,Vijayawada,Andhra Pradesh,India.

[3]Asst.Prof,Department of computer science ,Andhra Loyola Institute Of Engineering And Technology,Vijayawada,Andhra Pradesh,India.

Abstract-**In recent years a wide variety of information is available for the research, business applications and other development organizations but before taken this information into consideration we must ensure that the available information is having better quality as well as maintenance of privacy .This paper mainly deals with the issues like quality of the data and the privacy preservation techniques employed on the available data,intern this data will be useful for further enhancements in the field of research.**

***IndexTerms*—accuracy, data quality ,privacy preservation.**

I. INTRODUCTION

Now a days it has become essential for every organization to maintain the quality data in order to face the upcoming competitors in the market. Currently searching this kind of quality data has become a challenge for the researchers.Data quality includes accuracy, completeness, uniqueness, timelines and consistent Quality also includes the less redundant information and more durable. Can we protect our data from the competitors is the challenge .This concept data quality is used various business applications especially in informatica tool. First of all we have to gather the information from the web before placing that the web developers must make sure that quality is maintained or not. Government organizations and other agencies publish the micro data like census data and medical data. The data is in the form of tables. Each table consists of records of individuals .Every record has number of attributes ,there attributes are categorized in three types.1.Explicit Identifiers :Attributes that can clearly identify the individual Ex: Social Security Number 2.Quasi Identifiers: Attributes whose values taken together can potentially identify an individual. Ex: Zip-Code. Here we want to take the sensitive attribute called zip code and the zip code last digits are replaced by '*' so that we can import some of the privacy constraints. we can do for other sensitive attributes also can be replaced by '*'.This is shown in the fig 1.

II. RELATED WORK

Data Mining is the process of retrieving the information from the large databases. We have various data preprocessing techniques to retrieve the relevant information.[3]data cleaning, data integration, data transformation and data Reduction while entering the data into the database we must take care that it should free from flaws which include the missing values and null values and redundant values should not be include into our database.This is also one of the quality issue. Now we discuss the concept of data quality. Data quality not a new concept it is widely using from the past decades but it is changing from its content

III. DATA QUALITY

The definition of data quality is Accuracy here the data was recorded correctly,next Completeness in this measure all relevant data was recorded.In the concept of Uniqueness Entities are recorded once afterwards Timeliness also checked for the data the data is kept up to date. special problems in federated data: time consistency. Consistency here the data agrees with itself. We need a definition of data quality which

- ✓ Reflects the use of the data
- ✓ Leads to improvements in processes
- ✓ Is measurable (we can define metrics)

A Sources of data

- ✓ Relational databases (transactional data, mostly normalised into many tables, with keys between them, continuous and frequent updates)

- ✓ Data warehouses (decision support data, processed and cleaned, historical data, aggregated, updated at certain intervals – daily, weekly, monthly) Internet (click-stream data, log files, HTML, XML, e-mails Simple files like portable text.

For example comma separated values or non-portable, proprietary binary files.

The problems in data sources are manual entry of data may lead to errors uniform standards and formats are not followed while entering the data. Some times it may lead to data replication there by quality of the data is getting deteriorated. The data quality continuum

- Data and information is not static, it flows in a data collection and usage process
- Data gathering
- Data delivery
- Data storage
- Data integration
- Data retrieval
- Data mining/analysis

We need a multi-disciplinary approach to attack data quality, no one approach solves all problem so we have to follow below approaches to solve our problem.

- Process Management Ensure proper procedures**.** Process management is the application knowledge, skills, tools, techniques and systems to define, visualize, measure, control, report and improve processes with the goal to meet customer requirements profitably. It can be differentiate from program management in that program management is concerned with managing a group of inter-dependent projects. Business processes which encourage data quality.

- Assign dollars to quality problems
- Standardization of content and formats
- Enter data once, enter it correctly
- Automation
- Assign responsibility : data stewards
- End-to-end data audits and reviews
- Transitions between organizations.
- Data Monitoring
- Data Publishing
- Statics
- Focus on analysis: find and repair anomalies in data.
- Database Focus on relationships ensure consistency. Metadata / domain expertise

IV. PRIVACY PRESERVATION

Privacy preservation is widely used now a days we have to preserve the sensitive information present in the data Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "*" indicating that any value can be placed instead [7].For example in the employee table we can include the details like age, salary, and zipcode and department number,to preserve the privacy of the individual information like Zipcode {06148, 06149} can be replaced with 0614*, thereby stripping the rightmost digit and semantically indicating a larger geographical area The domains in databases are used to describe the set of values that attributes assume. For example, there might be a ZIP domain, a number domain and a string domain. In the original database, where every value is as specific as possible, every attribute is considered to be in a ground domain. For example, 06148 and 06149 are in the ground ZIP domain, Z0. In order to achieve anonymity the ZIP codes should be less informative. This can be done by making the domain of them at higher level Z1 in which the last digit has been replaced by '*'.

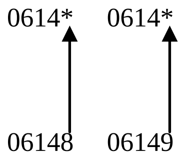

Fig 1 Zip code

Algorithm 1: Multidimensional Suppression

```
If the dataset D does not satisfy Anonymity property then
Identify quasi attribute set Q
for all records in Q do
suppress (Q, D);
end for;
End if
suppress (Q, D)
```

A. *K-Anonymity*

K-Anonymity is a technique that prevents joining attacks by generalizing and/or suppressing portions of the released micro data so that no individual can be uniquely distinguished from a group of size k. The *k*-anonymity requirement is quite simple. Intuitively, it stipulates that no individual record should be uniquely identifiable from a group of *k* on the basis of its quasi-identifier values. We will refer to each group of tuples in *T* with identical quasi identifier values as an *equivalence class*. The concept of *k*-anonymity has been proposed by Samarati and Sweeney [5] to anonymize microdata such that the correctness of the released (anonymized) data can be preserved. In order for microdata to meet the requirement of *k*-anonymity, every record in the micro data must be related to at least *k* other records or individuals. However, *k*-anonymity cannot always guarantee to protect privacy that each record is indistinguishable with at least *k-1* of the records with respect to the quasi-identifier.While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. Machanavajjhala et al. [6].k-anonymity suffers with two types of attacks one is *homogeneous.*

Table A
Sensitive attributes Before Anonymization

S.No	Age	Sal	Zip code
1	25	15k	06147
2	28	19k	06148
3	29	27k	06149

Table B
Sensitive attributes Before Anonymization

S.No	Age	Sal	Zip code
1	2*	15k	0614*
2	2*	19k	0614*
3	2*	27k	0614*

B t-Closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have [4] t-closeness. The closeness property is better approach compare to other techniques like k-anonymity and l-diversity. Here the prior believe of the observer has the great important than others. Information gain= prior probability-posterior probability. We have to limit the difference between *B*1 and *B*2. In other words, we have to assume that Q, the distribution of the sensitive attribute in the overall population in the table, is public information. We should not limit the observer's information gain about the population as a whole, but limit the extent to which the observer can learn additional information about specific individuals.

C (n,t) Closeness

This is more flexible version for the privacy preservation .Here n is the number of attributes and t is the threshold we are setting for the attributes. In this method we have use the distance measure between the two attributes named as count .Here we used the concept of EMD(Earth Movers Distance) for identifying the distance between the attributes which are sensitive in nature like Zip code, Salary and Age. We are measuring how closely the attributes are related to each other. This can be more flexible compare to other anonymization techniques but this alone is not sufficient to measure but we need multidimensional technique that can be used along with this approach can be very much helpful to keep the individual information confidentially. Now we will see this approach of (n,t) closeness which can be distributed entire population of data .This concept quality and privacy measure closeness can be applied on the data before utilized by the researchers and organizations.

V. CONCLUSION

This paper mainly deals with the quality of the data as well as privacy preservation issues so that this data is helpful for the organizations or researchers to use this data.

VI. REFERENCES

[1] "A Metadata Resource to Promote Data Integration", L.Seligman, A. Rosenthal, IEEE Metadata Workshop, 1996
[2] www.informatica.com
[3] Data Mining concepts and techniques 2nd edition By Jiawei Han Micheline Kamber
[4] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," Proc. Int'l Conf. Data Engineering (ICDE),pp. 106115, 2007
[5] Samarati, P. (2001). Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, vol. 13 (6) pp. 1010-1027.
[6] Machanavajjhala, A.; Gehrke, J.; Kifer D. (2006). L-Diversity: Privacy Beyond k-Anonymity. n:*Proceedings of the 2006 International Conference on Data Engineering (Icde'06)*, Atlanta, USA,2006.
[7] N. Li, T. Li, and S. Venkatasubramanian, *t-Closeness: Privacy Beyond k-anonymity and l-Diversity*, In Proc. Of ICDE, 2007, pp. 106-115

VII. BIOGRAPHIES

K.Priyanka received M.Tech CSE from KL University in 2012.Currently She is working as Asst.Prof In Andhra Loyola Institute Of Engineering and Technology Vijayawada. Her Area of interest are data mining and security issues in the data bases.

S.Asha Varma received her M.Tech CSE From VRSE.Acharya Nagarjuna University .Currently working as Asst.Prof In Andhra Loyola Institute of Engineeringand Technology.Her Area of interests are data Mining and Network security.

T. Rajesh received M. Tech CSE From JNTU in 2008.Currently He is working as Asst.Prof in Andhra Loyola Institute of Engineering and Technology.Area of interests are mobile computing,security and data mining.