Preservation of Privacy using Multidimensional K-Anonymity Method for Non-Relational Data

Abhijit J. Patankar, Kotrappa Sirbi, Kshama V. Kulhalli

Abstract*: Mining of huge data having complexity is a challenging issue also maintaining Privacy of data is also equally important ,sometimes there is a need to release data for use of researchers or for the purpose of gaining knowledge or earn money this release of data includes releas e of all attributes of personal data. when this type of data like Insurance record data, Medical diagnosis data, funding scheme data is release even if we remove sensitive attribute like Name for hiding personal details still data re-identification is possible by linking public data like voters data with these released data and by linking the quasi identifiers we are able to get sensitive information about person like critical disease, financial position etc. by applying k–Anonymization using multiple dimensions of attributes we are able to hide these sensitive attributes by generalising and suppressing the Quasi identifiers so that when linking with public database is done no records are re-identified, also we obtained results for quality measures for anonymisation and observed that the value of k once we start increase after some threshold anonymity starts decreasing so there is a need to choose proper value of k on non-relational data.

Keyword:-re-identification

I. INTRODUCTION

In most of the researches the more personal data we access more accurate outcome we achieve and also obtained results need to publish or release online for benefit of other researchers

While publishing data online one should avoid displaying personal information as when personal information is published anyone can link this information with public databases like voter database and sensitive information about a person gets disclosed. Consider a common example of mobile phone, a person identity may be disclosed with the help of mobile company or using telephone directory an adversary can track address or personal information.

To achieve k-anonymity, attributes such as common identifier are required to process . common-identifier have attributes which matched with common published public data to re-identify other fields. Therefore, k-anonymization provides maximum protection by allowing minimum data loss. Also in future work we can do Private information Protection for Relational Data using Access Control Mechanism and we can achieve K-Anonymization with relational data

Revised Manuscript Received on September 25, 2019

Mr. Abhijit J. Patankar, Research Scholar, Vishvesvaraya Technological University, Belagavi, Karnataka

Dr. Kotrappa Sirbi, Professor, Department of CSE, KLE's Dr.M.S.S.C.E.T, Belgaum

Dr. Kshama V. Kulhalli, Principal, D. Y. Patil C.E.T, Kolhapur Maharashtra Email id: abhijitpatankarmail@gmail.com

II. REVIEW OF LITERATURE

2.1 PRIVACY DEFINITION:-

Privacy related to how data is collected, used and shared by the users. So privacy definition varies from one environment to the other, Privacy is the generalized term which is used for securing confidential or personal information associated with individual or any group or organization

2.2 REVIEW OF LITERATURE

To achieve K-Anonymity we have studied 16 different papers in these papers different methods and techniques were discussed for obtaining animalization on dataset in Survey paper [1] advantages and drawbacks of PPDM discussed but here all probable techniques were not discussed. In Protecting Privacy by Multi-dimensional K-anonymity [8] basic method for Anonymity was discussed and also advantages and drawbacks were discussed Detail comparative Review of Literature as follows including Gaps

2.3 RESEARCH OBJECTIVES:-

The research objectives of this research work are as follows:-

- At the time of displaying of real critical and minute level-data need to avoid losses of data.
- To protect individuals private data and improve confidentiality,
- For Preservation of Privacy of persons or groups identified in released data
- To avoid from Linking, Background Knowledge and Homogeneity Attacks.
- To prevent from various information disclosure.
- To improve anonymity and data security on web 3.0.
- To use Nearest Neighborhood strategy to improve k-Anonymity on multidimensional data.

2.4 PROBLEM DEFINITION:

To obtain K-anonymity, quasi-identifier (Key attributes) attributes which we found from dataset are required to be processed, for reducing the chances that intruder will obtain the sensitive data by linking datasets on Key attributes attributes and prevent from homogeneous type of attack and background knowledge type of attack. to increase the anonymization and information loss should be less, and less time complexity for web 3.0.we can build K-Anonymization as a web based service which will convert normal data to protected data before release

DOI:10.35940/ijrte.B1096.0982S1019

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. METHODOLOGY

3.1 PRAPOSED SYSTEM ARCHITECTURE

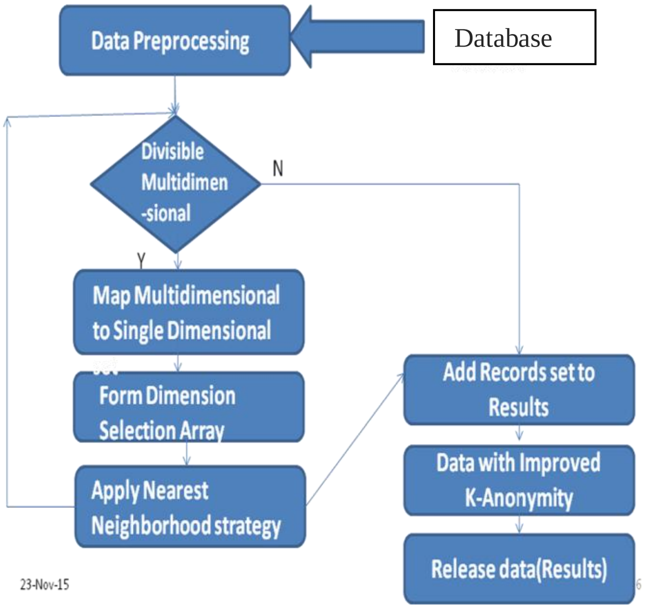

Data Preprocessing
Database
Divisible Multidimen -sional
N
Y
Map Multidimensional to Single Dimensional
Form Dimension Selection Array
Apply Nearest Neighborhood strategy
Add Records set to Results
Data with Improved K-Anonymity
Release data(Results)
23-Nov-15

Fig3.1 Proposed system Architecture

In above Architecture as shown in Fig3.1 after data is preprocessed if divisible dimensions are obtained then we will map Multi to single dimension and form the dimension selection Array and later we apply nearest neighborhood strategy and add obtained resultsets or records to improve Anonimization and later we release the Results this is a continuous process till we release all records

3.2 MODULES Implemented

3.2.1. Preprocessing and cleaning of Data

In data preprocessing all basic data mining processes were carried out like data cleaning,data Transformation to make uniform data.

3.2.2. Multidimensional to single dimensional Mapping.

In this module mapping of different multidimensional datasets with single dimensional dataset takes place

this mapping will convert all multi dimensionally sets to single dimension

3.2.3. To array for dimension selection

If information loss need to be reduced, by not compromising availability, while selecting the dimensions closer

value tuples are grouped together in same partition also there is a need to check interdependency among the relations which will separate selection of dimensions.

3.2.4. Nearest Neighborhood Strategy applied on datasets to calculate distance

As per requirement of Normalized Euclidean distance measures are used for calculating the distance.

$$d=\sqrt{\sum_{i=1}^{v}\left(\frac{(P_{1i}-P_{2i})^{2}}{v}\right)}$$

p = indicates how many attributes used per person,

v= indicates difference at maximum scale

d = indicates total distance between person/attribute.

IV. SCOPE OF THE WORK

4.1 Scope

The proposed system will help to protect privacy and minimize information loss on web 3.0

Following are the identified problems to achieve privacy

- How to protect individuals data on web 3.0 ?
- How to quantify privacy protection?
- How to maximize usefulness of published data?
- How to implement privacy protection in Relational data?

Here our system implement multidimensional K-Anonymity using nearest neighborhood strategy

To achieve security and confidentiality on micro data release. Also in future work we can do data protection for relational data using Access Control Mechanism and we can achieve K-Anonymity with relational data.

This Technique will provide wide scope to maintain Data Integrity and Confidentiality

4.2. Facilities Available at research Centre

- The adequate library facility at the research centre.
- The computing facility in the computer laboratory of RRC and in the personal computer at home.
- Full access to Digital library including IEEE and ACM
- Good Research laboratory with 24 X 7 Internet Facility
- Availability of Tools like Weka
- Availability of Training Dataset like LIC,Election Database,Medical Database.
- State of Art Systems with Printing Facility

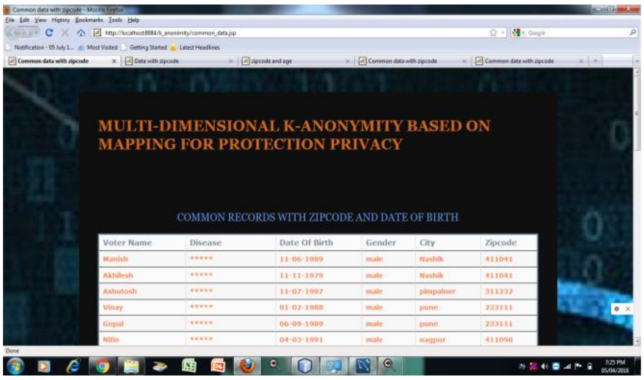
V. OUTCOME OF THE WORK

The outcome of this work is justified by creating different records of Non-Relational dataset such as Adult dataset and LIC dataset

After applying preprocessing activities on the nonrelational dataset we apply K-Anonymity and also prevent from background knowledge attack ,Linking attack so that integrity of data is maintained this is called as protecting data privacy when doing micro-release

Above fig shows step by step execution and outcome

5.1 . Common Records with zipode and Date of Birth this record is having values without anonymity


Common data with zipcode - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8084/k_anonimity/common_data.jsp
Common data with zipcode
Data with zipcode
zipcode and age
Common data with zipcode
Common data with zipcode
Done
7:25 PM
05/04/2018

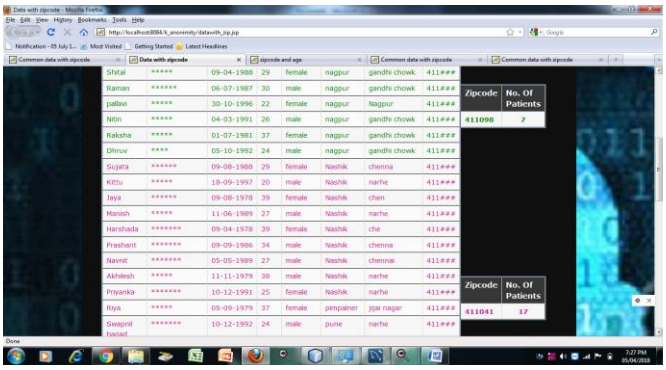
MULTI-DIMENSIONAL K-ANONYMITY BASED ON MAPPING FOR PROTECTION PRIVACY

COMMON RECORDS WITH ZIPCODE AND DATE OF BIRTH

Voter Name	Disease	Date Of Birth	Gender	City	Zipcode
Manish	*****	11-06-1989	male	Nashik	411041
Akhilesh	*****	11-11-1979	male	Nashik	411041
Ashutosh	*****	11-07-1997	male	pimpalner	311232
Vinay	*****	01-02-1988	male	pune	233111
Gopal	*****	06-09-1989	male	pune	233111
Nitin	*****	04-03-1991	male	nagpur	411098

5.2 Data with zipcode and Diseases Anatomized

Here we applied anonymization on datasets such as zipcode and disease to obtain K-anonymity and to protect from attacks


Data with zipcode - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8084/k_anonimity/datawith_zip.jsp
Common data with zipcode
Data with zipcode
zipcode and age
Common data with zipcode
Common data with zipcode
Done
7:27 PM
05/04/2018

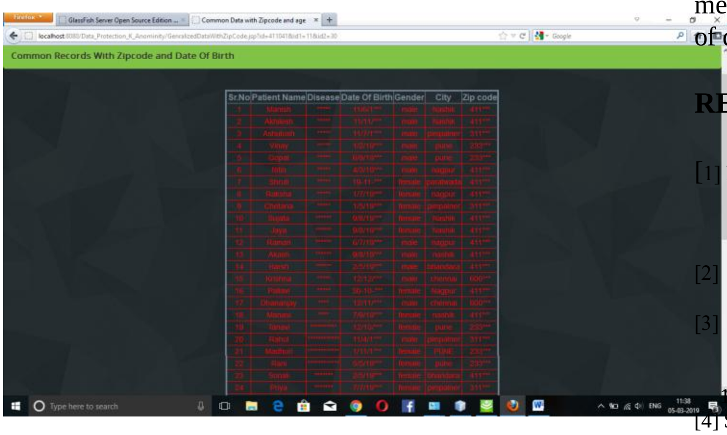
Shital	*****	09-04-1988	29	female	nagpur	gandhi chowk	411###
Raman	******	06-07-1987	30	male	nagpur	gandhi chowk	411###
pallavi	*****	30-10-1996	22	female	nagpur	Nagpur	411###
Nitin	*****	04-03-1991	26	male	nagpur	gandhi chowk	411###
Raksha	*****	01-07-1981	37	female	nagpur	gandhi chowk	411###
Dhruv	****	05-10-1992	24	male	nagpur	gandhi chowk	411###
Sujata	******	09-08-1988	29	female	Nashik	chenna	411###
Kittu	*****	18-09-1997	20	male	Nashik	narhe	411###
Jaya	******	09-08-1978	39	female	Nashik	chen	411###
Manish	*****	11-06-1989	27	male	Nashik	narhe	411###
Harshada	*******	09-04-1978	39	female	Nashik	che	411###
Prashant	*******	09-09-1986	34	male	Nashik	chenna	411###
Navnit	*******	05-05-1989	27	male	Nashik	chennai	411###
Akhilesh	*****	11-11-1979	38	male	Nashik	narhe	411###
Priyanka	*******	10-12-1991	25	female	Nashik	narhe	411###
Riya	*****	05-09-1979	37	female	pimpalner	jijai nagar	411###
Swapnil bagad	*******	10-12-1992	24	male	pune	narhe	411###

Zipcode	No. Of Patients
411098	7

Zipcode	No. Of Patients
411041	17

5.3. Common Records with Zipcode and Date of Birth Anonymized

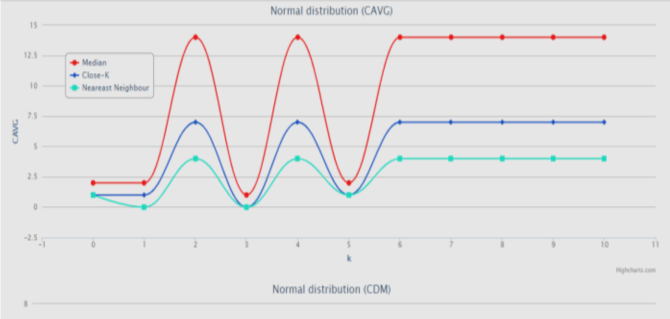
Here if we try to implement any attack then linking and background knowledge attacks are not possible on datasets


Firefox
GlassFish Server Open Source Edition ...
Common Data with Zipcode and age
localhost:8080/Data_Protection_K_Anonimity/GeneralizedDataWithZipCode.jsp?id=411041&id1=11&id2=30
Google
Type here to search
ENG
11:38
05-03-2019

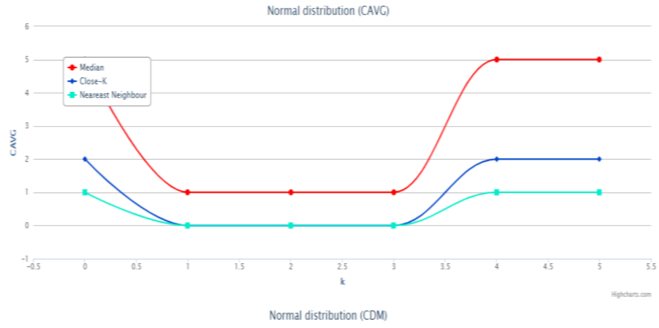
Common Records With Zipcode and Date Of Birth

Sr.No	Patient Name	Disease	Date Of Birth	Gender	City	Zip code
1	Manish	*****	11/6/1***	male	Nashik	411***
2	Akhilesh	*****	11/11/***	male	Nashik	411***
3	Ashutosh	*****	11/7/1***	male	pimpalner	311***
4	Vinay	*****	1/2/19***	male	pune	233***
5	Gopal	*****	6/9/19***	male	pune	233***
6	Nitin	*****	4/3/19***	male	nagpur	411***
7	Shruti	*****	10-11-***	female	[illegible]	411***
8	Raksha	*****	1/7/19***	female	nagpur	411***
9	[illegible]	*****	1/5/19***	female	pimpalner	311***
10	Sujata	******	9/8/19***	female	Nashik	411***
11	Jaya	******	9/8/19***	female	Nashik	411***
12	Raman	******	6/7/19***	male	nagpur	411***
13	Akash	******	9/8/19***	male	nashik	411***
14	Harish	******	2/5/19***	male	bhandara	411***
15	Krishna	*****	12/12/***	male	chennai	600***
16	Pallavi	*****	30-10-***	female	Nagpur	411***
17	Dhananjay	****	12/11/***	male	chennai	600***
18	Manasi	****	7/8/19***	female	nashik	411***
19	Tanvi	*********	12/10/***	female	pune	233***
20	Rahul	***********	11/4/1***	male	pimpalner	311***
21	Madhuri	***********	1/11/1***	female	PUNE	233***
22	Rani	***********	5/5/19***	female	pune	233***
23	Sonali	*******	2/5/19***	female	bhandara	411***
24	Priya	*******	7/7/19***	female	pimpalner	311***

5.4.Graph showing result of Normal distribution of values (Cavg) with the value of k and we obtain better results using Nearest Neighbour than Monderian and close k methods



5.5 Graph shows change of values of (Cavg) when value of k varies it proves that for normal distribution CDM alues of Nearest Neighbour are better than Median and close-k method



VI. CONCLUSION

With the obtained results and best on comparison of different methods for two quality valued variables Cavg and Cdm based on variation of value of k it proves that nearest neighbor with multidimensional k-anonymity is the better method for Anonymity improvement the and reduce the loss of data and different types of attacks

REFERENCES

[1] Pingshui Wang, Jiandong Wang1, Xinfeng Zhu1, Jian Jiang, "Reserch on Privacy Preserving Data Mining", International Conference on Biological and Biomedical Sciences *Advances in Biomedical Engineering, Vol.9*, pp.251-257, 2012.

[2] Charu C. Aggarwal, "*A General survey of Privacy-Preserving Data Mining Models And Algorithms"*, IEEE, pp 11-52,2008.

[3] Sweeney L., "*K-anonymity: A Model for protecting privacy"*, International Journal of Uncertainty, Fuzziness and Knowledge based system, 10(5), pp.557-570, 2002.

[4] Slava kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira,"Efficient Multi-dimensional Suppression for k-anonymity", IEEE transaction, pp1-14,2009.

[5] Kshitij Pathak, Nidhi Maheshwarkar, "*Performance issues of various K-anonymity Strategies"*, IJCTEE ,ISSN:2231-2307, Vol.1,Issue 2, pp18-22,2011.

[6] LeFevre K, DeWitt D J, Ramakrishnan R,"Mondrian multidimensional K-anonymity", IEEE International Conference on Data Engineering(ICDE06), Atlanta, GA, USA, pp1- 11,April 2006.

[7] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. "The WEKA data mining software: an update", ACM SIGKDD Explorations Newsletter, v.11 n.1, pp10- 18june 2009 [doi:10.1145/1656274.1656278].

[8-1] Qian Wang, Cong Xu, Min Sun, " Multi-dimensional K-anonymity based on Mapping for Protecting Privacy", Journal of Software, Vol. 6, No. 10, pp1937-1947,October 2011 .

[8-2] Qian Wang, Cong Xu, Min Sun, " Protecting Privacy by Multi-dimensional K- anonymity", Journal of Software, Vol. 7, No. 8, August 2012,pp1873-1880.

[9] Yongbin Yuan, Jing Yang, Sheng Lan "P-sensitive k-anonymity Based on Nearest Neighborhood Search in Privacy Preserving", Journal of Information and Computational Science 9: 5(May 2012) ,pp1385-1393.

[10] Gionis A, Tassa T., "k-Anonymization with Minimal Loss of Information", Knowledge and Data Engineering, IEEE Transactions, pp. 206-219, 2009.

[11] Madhan Subramaniam, Senthil R., "An Analysis on Preservation of Privacy in DataMining", International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010,pp1696-1699.

[12] Abdullah H. Wahbeh, Qasem A. Al- Radaideh, Mohammed N., "A Comparison Study between Data Mining Tools over some Classification Methods", International Journal of Advanced Computer Science and Applications, December 2011.

[13] C. Blake, C. Merz., "UCI repository of machine learning databases", 1998.
http://www.ics.uci.edu/mlearn/M1Repository.html.

[14] Samarati, Latanya Sweeney, "Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression", IEEE Transactions on Knowledge and Data Engineering, 2001

[15]Accuracy-Constrained Privacy- Preserving Access Control Mechanism for Relational Data Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.26, NO. 4, APRIL 2014.

[16] Liu, Kai-Cheng & Kuo, Chuan-Wei &Liao, Wen-Chiuan & Wang, Pang-Chieh. (2018). Optimized Data de-Identification Using Multidimensional k-Anonymity. 1610-1614. 10.1109/TrustCom/BigDataSE.2018.00235.