International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 06, June-2019

ANALYSIS ON DATA ANONYMIZATION USING L-DIVERSITY & T-CLOSENESS

Prof. Shameem Akhter[1], Sabreen Kausar[2]

[1]*Professor, Dept. Of CSE, K.B.N College of Engineering, Kalaburagi, VTU Belagavi,*
[2]*PG Student) Dept. Of CSE, K.B.N College of Engineering, Kalaburagi, VTU Belagavi,*

ABSTRACT

The knowledge analyst processes information by records and analytics. A most common technique of data anonymization developed for privacy protecting statistics publishing. The k-anonymity is the important method which is important method for publishing the microdata k-anonymity should contain k-tuples for a set of records. Numerous authors' process data that cannot save attribute disclosure. Many varieties of researches were carried out for isolating quasi-identifiers from sensitive attribute. For evaluation of records, information analyzer has to offer safety to be getting attacked. The belief of l-range has been proposed to address this; each equivalence class has l-diversity as minimum l properly-represented values for every sensitive characteristic. So we proposed a privateness belief called t-closeness, then we apply the Earth mover distance degree for t-closeness. We explain for the method t-closeness. Finally, we have two approaches which are an anonymization system and a reconstruction process. This method provides overview on popular information protecting techniques. On this examine, concepts behind those strategies are analyzed and explained with illustration.

Index Terms- k-anonymity, data mining, l-diversity, background knowledge, privacy disclosure, t-closeness

I. INTRODUCTION

With the digitalization, data from different domain application such as retail companies, social networks, healthcare departments, public transit have been collecting a huge quantity of data when service is the condition. When such data is collected can help us to interpret which is not easy process. For researchers, analysts and companies privacy is most important that they fear to publish as they are responsible for security breaches. There are three types of attack, similarity, linkage, background knowledge attack. As a result, a few amounts of data sets are released and are available online. Because of this issue of limitation to analyze such data to get the information that could be beneficial to the public. So it is very significant for the developer to develop tools that should release the data privately using some datasets. To keep the information save we use a method of data anonymization.

First we apply the k-anonymity method to the original database that is divided into attributes into three categories, data identifier attribute identifies and distinguishes the individual information and protect the individual data. The individual record can be identified by these attribute values Called Quasi-identifiers which are not so special and however can be combined to get the unique QIDs. QIDs are of two type's numeric and classification attribute. Sensitive attribute is sensitive information of the individual and the information or value of the attribute should be protected from an intruder. For example we store the blood group information into the blood donor database where blood group is a sensitive value and age is an attribute, in the table form. The information anonymity is an effective method to protect. The main motive is to transform the statistics through generalization and compression. Anonymizing the real facts set to meet the requirements of k-anonymity, which is used for publishing the information. Presently, there are numerous algorithms to implement k-anonymity. k-anonymity in method of working that allow the information extracting from data, while preserving data mining of individual privacy which minimizes the data manipulation in the clustering procedure. L-diversity approach satisfies if each of the equivalence class in a need to incorporate well represented values for sensitive characteristic. We have shown that l-diversity has some restrictions and we have proposed a unique privateness perception referred to as t-closeness which permits us to gain of anonymization techniques different than generalization of quasi-identifier and suppression of information. T-closeness is a technique where it advanced to enhance the privacy preserved in the statistics units. The generalization is the anonymization technique, which replaces quasi-identifier with the values which is not so unique. This leads to information that has same set of quasi-identifier values and we get the set of records of anonymized table that has similar values.

II. RELATED WORK

In paper [1] localization service for protecting the person using the smart phone from the attacker of location tracking without making it difficult for someone to do tracking and providing great attention location updates continuously. The technique of temporal vector Map is the efficient technique in protecting data but it doesn't allows a user to make use of the k-Anonymity Bloom filter method and best Neighbors method. The TVM makes easy, keep the attention and make location update in magnitude less energy and number of messages in the process of result in win-lose outcomes, from server-side localization processes.

In paper [2] Searching in the candidate list for the best match is performed. To obtain the data we need to use the layers of safety the robust hash values is given for safety layer one and the second is the server quality should be increased when person wants to remove but in first surface some bits which is used to identify secret information, but it is hard to get the data of the person. The person will get all the hash values from the server. The person and the server privacy are protected. Privacy protection uses these techniques one is on random projections, discrete wavelet transform. By using these techniques the result will be the threat if files are duplicate, documents, if hash values are similar.

In paper [3] Web-based applications are nowadays very common and popular to develop and to use as it's far used much less client-side information and to deliver without difficulty. Recent studies and researches have shown that the high Profile Web applications can sometimes get leaked from encryption due to fact the side-channel attacks is any attack rely on personal data advantage that utilizes the unlike other sizes of packet and timing. Hiding the pattern of traffic and information update is an identical problem. In PPTP problems analyze the complexity and the application scenarios, design efficient algorithm. These algorithms are applied and compared to web applications through some experiments.

In paper [4] Improving the Utility data anonymization opposes the previous k-anonymity to the latest differential privacy method whose result is limited because of the noise that added to the output. K-anonymous of the dataset is reduced if the amount of noise of differential privacy .where micro aggregation is designed and reached from k-anonymity. If there is a decrease in the noise then the dataset increase as output. On a reference data set theoretical benefits are illustrated in a practical setting.

In paper [5] for private histogram proposed the two sanitization techniques, for the private data release the promising privacy models is differential privacy. Schemes have been releasing for the information and personal histogram in this method. Fourier perturbation Algorithm (FPA) procedure is released first and is true for the starting. The other scheme depends on the resource and clustering .clustering is not used into bins. The technique used has the exact distribution and histogram attribute that improves the range queries.

II. METHODOLOGY

- In our method we use the attributes in database that use randomization method and the reconstruction method and the anonymize algorithm inserting the record in the individual database by generating the record. The generalization is the anonymization technique, which replaces quasi-identifier with the values which is not so unique. This leads to information that has same set of quasi-identifier values and we get the set of records of anonymized table that has similar values.
- For generalization we use the novel linking-based anonymity which can protect from the attack by similarity.
- After generalization we use the micro aggregation for t-closeness that divides the data depending upon the homogeneity of the attributes in the most compressed form of data sets and generalize it using the centroid class.
- Quasi-sensitive attribute is not sensitive but after combining they become sensitive
- For private dataset the noise is added can be used for micro aggregation based k-anonymity they first generate the database for differentially private database and then noise is added for each class
- Then we make use of the consideration method to divide the data record into different and then apply technique to each group.
- Then after publishing individually the quasi-identifier and sensitive attribute to decrease the connection between them.
- The mostly used method clustering for anonymity for data needs to be similar in database to make less information loss.
- Then method used simply adds random values to the attribute for l- diversity and for t-closeness and doesn't change original value but changes the probability by adding different values.
- We use the baye's method for adding value for reconstruction method, selecting the required values of sensitive values in real record is categorized. The probability is evaluated by the data analyzer then real record is anonymized.

IV.PROPOSED ALGORITHM

In our proposed method we make use of the randomization that uses the uniformly random values and take random choices when executing in order to get the good possible values for random values. On basis of parameter for each record anonymization algorithm generates the anonymize record.
First we assume the blood donation database table that is not anonymized. For k-anonymity we use two method first is suppression where the values are replaced by asterisk (*) then we generalization method where the values are exchanged by the large class.

A. L-diversity from k-anonymity: The k-anonymity provides protection which is easy. If a database satisfies k-anonymity for k record, then if value is known of quasi-identifier then person cannot identify the record. At the same time k-anonymity protects identity disclosure. The attacks are of two types' similarity and background knowledge attack. The table 1 is the original database and table 1(b) is the anonymized record for the database. Suppose Bill is the blood donor age is 25 and has blood group B+ and the parameter for privacy be k and the utility parameter t .we use the micro aggregation for dividing t-closeness. We develop a graph-based model for all the records of generalization without the record which is greater than k. For each sensitive value A is generalized to A*. The result we get in the aggregated expression for anonymize record for the database records. Applying this anonymization algorithm we can illustrate in the database table.

Table 1: Example for anonymization by proposed method

Table 1(a): Original record

Name	Age	Blood group	Disease
Bill	25	B+	HIV

Table 1(b): Anonymized record

Age	Blood group	Disease
2*	A*	PID
2*	A**	HIV
>=21	B*	Breast cancer
>=22	B**	Kidney caner
2*	B*	HIV

Table1(c): Aggregated expression for the above record

Age	Blood group	Disease
{20,25}	{A+,B+}	{HIV,PID)

B. L-diversity for entropy: for privacy preserving we use entropy l-diversity. K-anonymity cannot protect the background knowledge attacks. We can calculate using the baye's-optimal protection which is used for probability distribution for background information and similarity attack of the attributes and uses the techniques to protect information. We use the entropy l–diversity for log (l) for equivalent class for every Entropy (E) for every r* assure the condition. The entropy used for different sensitive values, log (l) for uniformly distributed different sensitive values. But l-diversity may not be easy and not obligatory to achieve.

C. T-closeness: the private information can be measured by the data of an observer .our method can separate the statistics into parts that the information of the population and the unique people. This approach doesn't affect quasi-identifier and is not required to achieve k-anonymity if the value is removed it decreases diversity and is not required to achieve l-diversity We need to calculate the distance between the two distributions and using the optimized parameter algorithm we calculated the t-closeness. The distance in the equivalence class between the attributes of distribution and sensitive that should not be greater than the threshold t. An interesting query is the way to generalize and suppress to attain limitations. For our t-closeness we use the Earth mover Distance measure has the importance for taking the closeness of attributes and can provide more privacy.

D. Reconstruction algorithm : we expand our work by adding values using Bayesian method we assume by combining the sensitive data in the original database and then anonymized and calculate the conditional probability we use the equation by adding law of total probability cab be represented by

$$P\left(\frac{\alpha_i}{\beta}\right) = \frac{\mathrm{P}(\frac{\beta}{\alpha_j})\mathrm{P}(\alpha_j)}{\sum_{\gamma=1}^{n} p(\frac{\beta}{\alpha_i})P(\alpha_i)}$$

V. IMPLEMENTATION

- **Doctor**

In this module, the doctor performs operations such as View My Profile, View User Request Solution and Reply

- **User**

In this module, the user registers and logs in by his/her name and passcode for identity. When user Login the receiver will perform operations like View My Profile, Add Patient Details, and View All Patient Details

- **Patient**

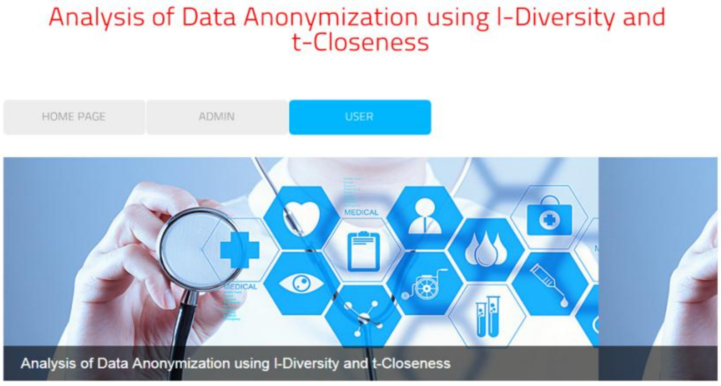
In this module, patient can do following operations such as Viewing patient Profile, View All Disease, and Solution Details.

- **Admin**

The Admin manages all the data stored in server to provide service and can perform the service such as Viewing Data Holder, patient, Doctors and Authorize, View All Patient Details, View Age(l-Diversity) and Disease(t-Closeness), View Disease(l-Diversity) and Age(t-Closeness), View Disease(l-Diversity) and Job(t-Closeness), View Disease(l-Diversity) and Pin code(t-closeness), View k-Anonymity, l-diversity and t-closeness Data, View Disease Based Results, View Age Based Results, View Pin code Based Results, View Blood Group Based Results.

VI. EXPERIMENTAL RESULT

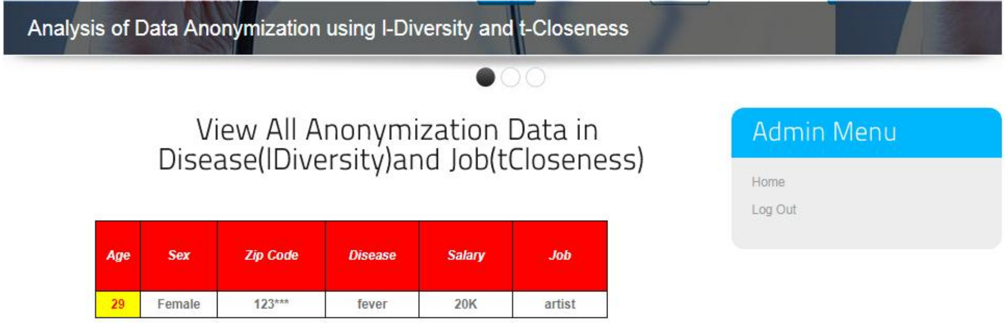
Figure 6.1: Home page



Analysis of Data Anonymization using l-Diversity and t-Closeness

HOME PAGE
ADMIN
USER
MEDICAL
MEDICAL
Analysis of Data Anonymization using l-Diversity and t-Closeness

Fig 6.2: View All Data by Age

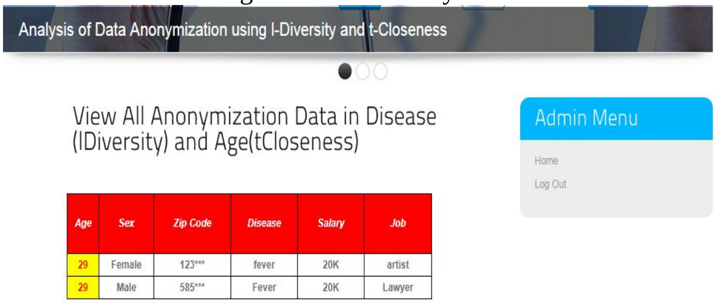


Analysis of Data Anonymization using l-Diversity and t-Closeness
Admin Menu
Home
Log Out

View All Anonymization Data in Disease(lDiversity)and Job(tCloseness)

Age	Sex	Zip Code	Disease	Salary	Job
29	Female	123***	fever	20K	artist

Fig 6.3: View Data by Job

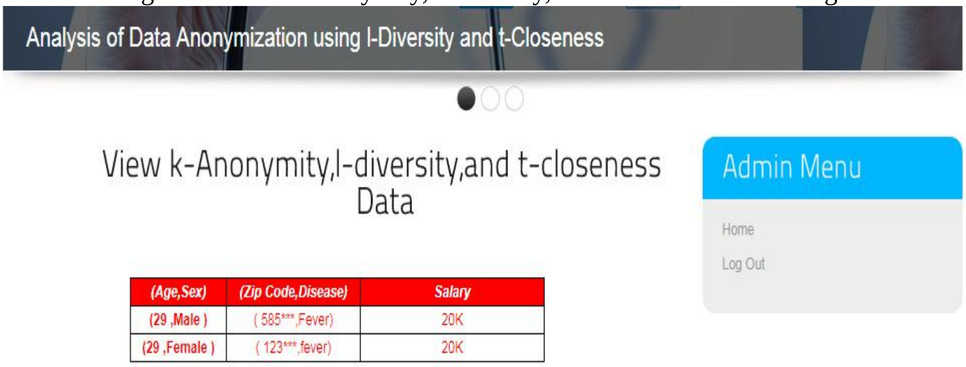


Analysis of Data Anonymization using l-Diversity and t-Closeness
Admin Menu
Home
Log Out

View All Anonymization Data in Disease (lDiversity) and Age(tCloseness)

Age	Sex	Zip Code	Disease	Salary	Job
29	Female	123***	fever	20K	artist
29	Male	585***	Fever	20K	Lawyer

Fig 6.4: View K-Anonymity, l-diversity, and t-closeness Data in age



Analysis of Data Anonymization using l-Diversity and t-Closeness
Admin Menu
Home
Log Out

View k-Anonymity,l-diversity,and t-closeness Data

(Age,Sex)	(Zip Code,Disease)	Salary
(29 ,Male)	(585***,Fever)	20K
(29 ,Female)	(123***,fever)	20K

Fig 6.5: View K-Anonymity, l-diversity, and t-closeness Data

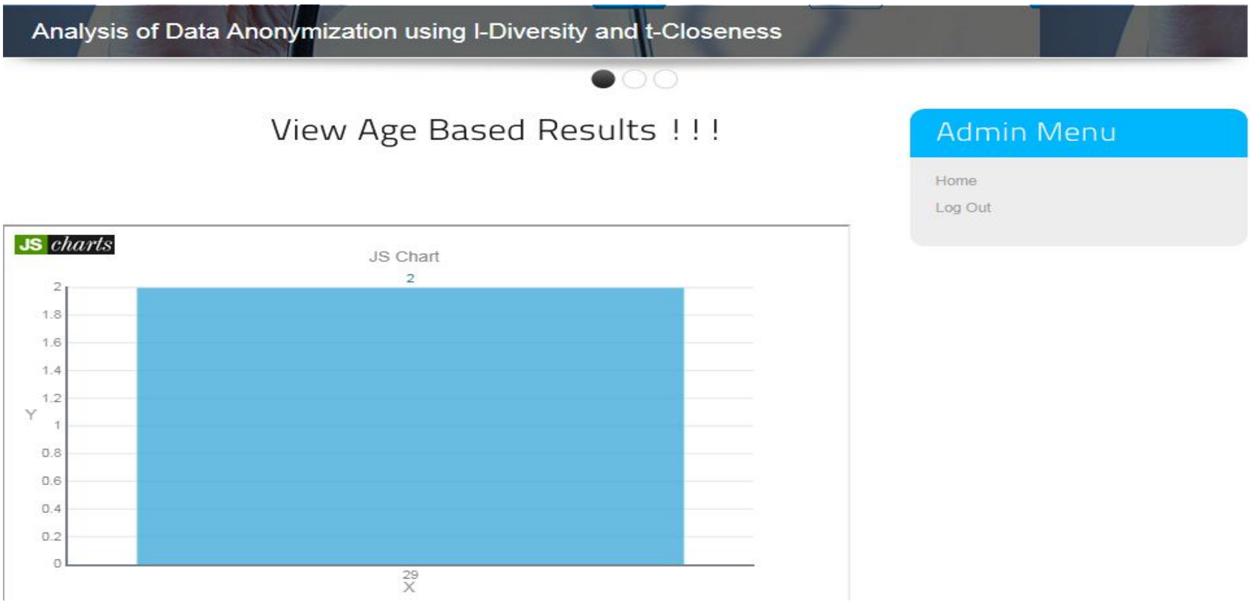

Analysis of Data Anonymization using l-Diversity and t-Closeness
View Age Based Results ! ! !
Admin Menu
Home
Log Out
JS charts
JS Chart
2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Y
29
X

Fig 6.6: View Age Result

VII CONCLUSION

The l-diversity method, t-closeness and algorithms considered and applied carefully to preserve the private data us. By using our proposed method information loss is decreased even if the attributes are increased. The method is reconstructed the true distribution in our proposed method and we try to protect the real data information of the probability distribution. We have shown that there are restrictions in l-diversity and we have proposed a unique perception referred to as t-closeness which allows us to obtain anonymization method different than generalization and suppression of quasi-identifier.the results show that there are similar data that has a chance to get attack by using our method there is decrease in the range and the values of different data. In t-closeness convey the overall distribution and every viable method is to generalize a sensitive attribute as opposed to hiding it completely. We can successfully integrate those strategies with generalization and suppression to obtain the better attributes. For future exertion the method used in proposed system is thread for different methods.

REFERENCES

[1] A. Konstantinidis, G. Chatzimilioudis, D. Zeinalipour-Yazti, P. Mpeis, N. Pelekis, and Y. Theodoridis, “Privacy-Preserving Indoor Localization on Smart phones,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3042–3055, 2015.

[2] Li Weng, Member, IEEE, Laurent Amsaleg, April Morton, and Stéphane Marchand-Maillet” A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval”, *IEEE Transactions on Information Forensics and Security*,vol. 10, no. 1, pp. 152–167, 2015.

[3] KWen Ming Liu, Lingyu Wang, Pengsu Cheng, Kui Ren, Shunzhi Zhu and Mourad Debbabi” PPTP: Privacy-Preserving Traffic Padding in Web-Based Applications” *IEEE Transactions on Dependable and Secure Computing*, 1060 vol. 11, no. 6, pp. 538–552, 2014.

[4] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sa´nchez, and Sergio Mart´ınez” Improving the Utility of Differentially Private Data Releases via k-Anonymity”, in *Proc. IEEE ICDM* , 2012, pp. 1–10, 2013.

[5] Gergely , Claude Castelluccia , Rui Chen” Differentially Private Histogram Publishing through Lossy Compression” *Proc. VLDB*, vol. 22,no. 6, pp. 797–822, 2013.,2012.