International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 7, July-2018

IMPROVED PRIVACY OF PUBLISHING SENSITIVE DATA USING ED-KA TECHNIQUE IN DATA MINING

[1]Ashutosh Sharma, Prof. [2]Ashish Gupta (Professor)

[1]MTech. Computer Science & Engineering
[2]Department Computer Science & Engineering
Nagaji Institute Of Technology & Management,
Gwalior, India

***Abstract*—Data Mining (DM) is the procedure of examine data from specific summarizing and perspectives the final outcome as helpful information.Preserving privacy is becoming a key apprehension as personal data is publicly available in recent years. The proper protection of personal information is increasingly becoming an important issue in an age where many countries have laws on these issues such as the misuse Individual Information Protection and Electronic Documents of individual data and fraud are across the board. Publishing data about individuals without revealing sensitive information about them is an important problem. Publishing individual-related data for big data analysis such as scientific research and merchant analysis has become frequent in this decade. Anonymized information distribution has gotten significant consideration from the examination group as of late. For arithmetical responsive attribute, mainly of the accessible privacy-preserving data publish technique contemplate on top of micro-data with various definite responsive attribute otherwise merely one arithmetical responsive characteristic. During this paper, Entropy l-diversity + k-Anonymity (ED-kA) technique used to increase the retreat of the data.**

Keywords*— *Data Anonymity, Publishing Data, Privacy Preserving, Sensitive data, Entropy l-diversity, Enhanced β-likeness.

I. INTRODUCTION

Data mining strategies are being actualized in a few looks into and applications. Be that as it may, DM strategies raise the issue of protection as well. Privacy refers as crucial factor in the information system. Because of, various works are being devoted to incorporate privacy preserving methods with data mining algorithms with the view to stop sensitive data at the time of informationdetectionprocedure. When a person transfers their database to their server, few sensitive patterns are concealing from its database regarding particular privacy policies. Recently, researchers focus on seriousness of the errors regarding privacy. Privacy preserving refers as the crucial issue in the DM techniques. The major challenge of present DM algorithm is drawing the data at the time of maintenance of privacy datasets. Because of this concern, the privacy preserving data mining (PPDM) systems are being received. The major concern in the privacy preserving data mining is referring as sensitive pattern mining. PPDM techniques implemented for modification of database with the help of insertion wrong information for concealing sensitive information. These issues at display are enhanced in DM systems however they additionally wind up costly, calculation and overhead will occur [1].The PPDM shows privacy preserving data publishing model and multi analyze few associate technologies, k-anonymity, relational kanonymity, 1-diversity and perturbation of privacy preserving.

Usually, any anonymization process passes through one stage only, which mainly addresses the people outside the organization while totally neglecting the data repository from being accessed by any employee or attacker that can reach this data and violate individuals' privacies. According to most of the statistics and surveys1, 70% of every security event come from insiders, whereof misleading the internal users is as important as the external ones and the anonymization at this Level becomes an urgent need toward edge the danger of internalAthreat [2]. Many technique used for preserving privacy within DM contain be a developed more than the last decade such like cryptographic, randomization methods, k-anonymity, I-diversity etc. But these method be able to influence the accurateness of results with can result into the loss of information [3]. Mainly of the present PPDP methods could not process multiple, heterogeneous sensitive attributes with different levels of sensitivity requirements. This motivates us to suggest a novel methodology that can handle multiple heterogeneous (both numerical and categorical) sensitive attributes [4].

II. RELATED WORK

RashadSaeed, AzharRauf [2018] presents a hybrid anonymization approach using bucketization of (l, e) diversity technique and simplification with repression of k-anonymity technique. The proposed method protects data from semantic similarity, membership, and individuality discovery attack. Our results have proved that from the view of data privacy, the proposed method have increased the diversity degree. On the new hand from the view of information loss, the proposed method have reduced the Discernibility Penalty (DP). In addition, the proposed process hasimproved the Normalized Certainty Penalty (NCP). Hence the proposed method preserves data privacy more effectively when compare toward klredInfo and (l, e) diversity techniques while maintaining the value of data[5].

Ricardo Mendes et al. (2017) The accumulation and examination of information are constantly becoming because of the inescapability of processing gadgets. The examination of such data is cultivating organizations and contributing gainfully to the general public in a wide range of fields. Be that as it may, this stockpiling and stream of conceivably touchy information postures genuine protection concerns. Strategies that permit the learning extraction from information, while safeguarding security, are known as privacy-preserving data mining (PPDM) systems. This paper reviews the most pertinent PPDM strategies from the writing and the measurements used to assess such systems and presents average uses of PPDM techniques in significant fields. Besides, the present difficulties and open issues in PPDM are talked about. [6]

Putri et al. (2016) This investigation proposes a mixture change in PPDM, which is a merger of the two existing methods on past examinations, the entropy-based parcel system and consolidated twisting strategies. To quantify the proposed strategy, assessment of the utility and security parameter assessment are utilized. Utility evaluation is used to assess the accuracy of the information and privacy parameter evaluation to assess how close the original value will be obtained from the transformation and how much they are distorted. The experimental outcomes demonstrate that the proposed technique gives preferable outcomes over past strategies in utility and protection, so the information will be saved and can be used for analyzing such as DM. [7]

KhoirunnisaAfifah et al.(2016) In this thesis, we proposed additional steps to perform pre-process on disk image before entering protection mechanism. Besides, we create implementation of protection mechanism as a standalone Java application and extraction mechanism as a module in Autopsy. [8]

Ahmed Ali Mubarket. Al [2016] proposed a way to deal with all out information conservation in light of Domain-based of semantic tenets to defeat the likeness attacks. The investigational results of the proposal approach focused to definite data presented. The results show to the semantic anonymization increases the privacy level with effect data utility [9].

Vanessa Ayala-Riveraet al., [2016] proposed a dynamic VGH evaluation approach which exploit the occurrence allocation of the input data-sets. Our approach yields a score (d-GSL) which acts as predictor of the effectiveness of VGHs to perform the anonymization of data. The d-GSL score enables users to effectively compare multiple VGHs for a given domain and select the one that will better maintain the value of the unique data. We also proposed a ranking scale that will help out user to categorize the VGHs base on top of their value (in terms of d-GSL) into categories. All category include an time with a qualitative descriptor to offer practitioners an intuitive interpretation of the d-GSL [10].

Francisco Dias et al. [2016] present an execution of a robotized anonymization framework, worked in a secluded structure, for archives written in Portuguese . Four diverse techniques for anonymization are assessed and looked at. Two techniques supplant the delicate data by artificial labels: repression with category. The other two techniques supplant the data by literary articulations: arbitrary substitution and speculation. Assessment demonstrated that the utilization of the labeling furthermore, the speculation strategies encourages the perusing of an anonymized content while keeping some semantic floats caused by the remotion of the first data [11].

Shu-Ming Hsieh, Mao-Hsu Yen, Li-Jen Kao [2016] propose a

Semantic-based data anonymization method which employs

Entity ontology to anonymize the graph data for publication.

The anonymized graph can answer Complex queries with assured privacy[12].

III. PROPOSED METHODOLOGY

Micro data: Attribute which must not be disclosed in the released micro data. Let T = {t1, t2, . . . ,tn} be alive a table by attribute A1, . . . , Am. We assume that T is a sub-set of several bigger population Ω wherever every tupple represent an person as of the population. Let A indicate the set of every attribute {A1, A2, . . . , Am} with t[Ai] indicate the value of attribute Ai used for tuple t. If C = {C1, C2, . . . , Cp} ⊆ A after that we utilize the information t[C] to indicate the tuple (t[C1], . . . , t[Cp]), which be the projection of t on the attribute within C.

Sensitive attribute: Into PPDP, there live several important subsets of A. A receptive characteristic is an element whose value used for any exacting individual have to be reserved covert as of people who contain no direct entrance toward the unique data. The data to be released after applying anonymization methods on it is called the receptive attribute.

Quasi-Identifier: The table T with attributes (A1… An), a quasi identifier be a minimum locate of attribute (Ai1… Ail) (1<= i1<il<= n) in T that can be joined with external information to re-identify individual records. single example of a quasi-identifier be a primary key similar to public security digit. Another illustration is the set {Gender, Age, Zip Code} in the GIC dataset that was utilized to recognize the legislative head of Massachusetts as depicted in the presentation.

Equivalence Class: Every set of tables which can't be recognized from each other as for Quasi-Identifier is known as Equivalence class.

L-diversity: An equivalence class be thought toward include ℓ-diversity if there be at a smallest amount ℓ well represented values for the receptive attribute. A table be supposed toward include ℓ-diversity if each equivalence class of the board has ℓ-diversity.

Entropy l-diversity: The entropy of a equivalence group E is defined to be

$$E = -p(E, s) \log p(E, s)$$

Where S be the domain of the sensitive attribute, and p (E, s) be the portion of report within E so as to include sensitive value s. A table be said toward include entropy ℓ -diversity if used for every equivalence set E, Entropy (E) ≥ log ℓ. Entropy ℓ - diversity is tough than separate ℓ -diversity. In order toward contain entropy ℓ -diversity used for every equivalence class, the entropy of the complete table must be at least log (ℓ). Sometimes this might too preventive, since the entropy of the complete table could be low down but a little value is extremely common.

Data preprocessing is a DM method that involve transform raw data keen on an comprehensible set-up. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

In the proposed work, hybrid technique Entropy l-diversity + k-Anonymity (ED-kA) used to improve the privacy of the data. In a k-anonymized dataset, each record is unclear from at any rate k −1 different records as for certain "distinguishing" characteristics. A table T fulfills k-namelessness if for each tuple t ∈ T there exist k − 1 different tuples ti1 , ti2 , . . . , tik−1 ∈ T with the end goal that t[C] = ti1 [C] = ti2 [C] = • = tik−1 [C] for all C ∈ QI. The Anonymized Table T ⋆ . Since the semi identifiers may exceptionally distinguish tuples in T , the table T isn't distributed; it is subjected to an anonymization technique and the subsequent table T ⋆ is distributed. There has been a ton of research on strategies for anonymization. These procedures can be extensively grouped into speculation methods. The speculation with tuple concealment methods, and information swapping and randomization strategies. In this paper we restrict our discourse just to speculation methods.

Attacks on k-Anonymity

In this segment we introduce two assaults, the homogeneity assault and the foundation learning assault, and we indicate how they can be utilized to trade off a k-mysterious dataset.

Homogeneity Attack:

Alice and Bob are opposing neighbors. One day Bob falls sick and is taken by emergency vehicle to the doctor's facility. Having seen the emergency vehicle, Alice embarks to find what sickness Bob is experiencing. Alice finds the 4-unknown table of current inpatient records distributed by the doctor's facility, thus she realizes that one of the records in this table contains Bob's information.

Since Alice is Bob's neighbor, she realizes that Bob is a 31-year-old American male who lives in the postal division 13053. In this manner, Alice realizes that Bob's record number is 9, 10, 11 or 12. Presently, those patients have a similar therapeutic condition (tumor), thus Alice presumes that Bob has malignancy. Note that such a circumstance isn't remarkable. As a back of-the-envelope count, assume we have a dataset containing 60,000 particular tuples where the delicate characteristic can take 3 unmistakable qualities and isn't related with the non sensitive traits. A 5-anonymization of this table will have around 12,000 groups2 and, by and large, 1 out of each 81 gatherings will have no decent variety (the qualities for the delicate trait will all be the same). Along these lines we ought to expect around 148 gatherings with no assorted variety. In this manner, data around 740 individuals would be endangered by a homogeneity assault. This proposes notwithstanding k-secrecy, the disinfected table ought to likewise guarantee "assorted variety" – all tuples that offer similar estimations of their semi identifiers ought to have different qualities for their delicate traits. Our next perception is that an enemy could utilize "foundation" learning to find touchy data.

Foundation Knowledge Attack:

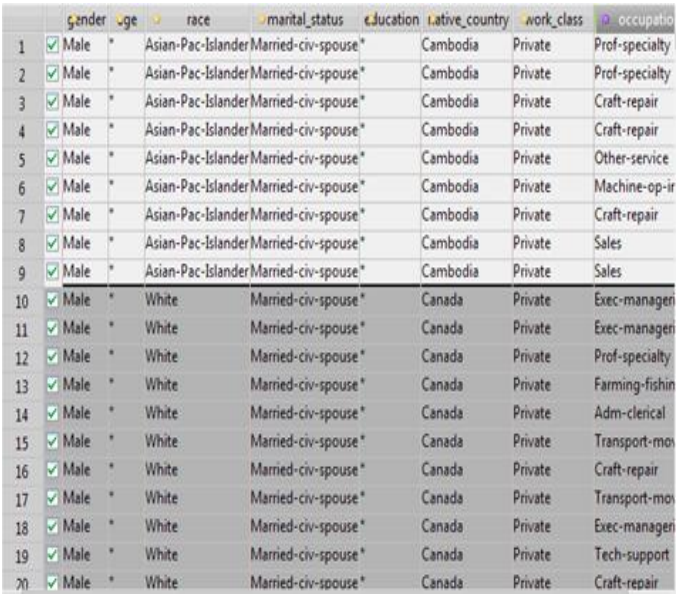
Alice has a penfriend named Umeko who is admitted to an indistinguishable healing facility from Bob, and whose patient records additionally show up. Alice realizes that Umeko is a 21 yearold Japanese female who as of now lives in postal district 13068. In view of this data, Alice discovers that Umeko's data is contained in record number 1,2,3, or 4. Without extra data, Alice isn't sure whether Umeko come down with an infection or has coronary illness. In any case, it is notable that Japanese have a to a great degree low rate of coronary illness. In this manner Alice closes with close assurance that Umeko has a viral disease.

In the proposed work, ED-kA be use to improve the presentation of the privacy protection of the dataset. In the figure below, it show the data which are preprocessed by using MATLAB and then here we performed the anonymization using ED-kA.

	gender	age	race	marital_status	education	native_count
1	Male	27	Asian-Pac-Islander	Married-civ-spouse	Some-college	Cambodia
2	Male	36	Asian-Pac-Islander	Married-civ-spouse	HS-grad	Cambodia
3	Male	37	Asian-Pac-Islander	Married-civ-spouse	1st-4th	Cambodia
4	Male	37	Asian-Pac-Islander	Married-civ-spouse	Bachelors	Cambodia
5	Male	40	Asian-Pac-Islander	Married-civ-spouse	7th-8th	Cambodia
6	Male	46	Asian-Pac-Islander	Married-civ-spouse	HS-grad	Cambodia
7	Male	48	Asian-Pac-Islander	Married-civ-spouse	Some-college	Cambodia
8	Male	50	Asian-Pac-Islander	Married-civ-spouse	12th	Cambodia
9	Male	51	Asian-Pac-Islander	Married-civ-spouse	HS-grad	Cambodia
10	Male	45	White	Married-civ-spouse	Bachelors	Canada
11	Male	47	White	Married-civ-spouse	Bachelors	Canada
12	Male	48	White	Married-civ-spouse	Bachelors	Canada
13	Male	51	White	Married-civ-spouse	Bachelors	Canada
14	Male	56	White	Married-civ-spouse	Bachelors	Canada
15	Male	41	White	Married-civ-spouse	Bachelors	Canada
16	Male	49	White	Married-civ-spouse	Some-college	Canada
17	Male	52	White	Married-civ-spouse	Some-college	Canada
18	Male	52	White	Married-civ-spouse	Some-college	Canada
19	Male	55	White	Married-civ-spouse	Some-college	Canada
20	Male	29	White	Married-civ-spouse	Bachelors	Canada

Fig.1. Input data

In the figure below, it show the anonymous data which the information of age and education. The clustering of the tuple performed to create equivalence class (ECs). It should contain unique quasi identifiers (qid) and distinct sensitive attribute values. Different adversaries can have different background knowledge leading to different inferences. It simultaneously protects against all of them without the need for checking which inferences be able to be completed with which level of backdrop information.



	gender	age	race	marital_status	education	native_country	work_class	occupatio
1	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Prof-specialty
2	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Prof-specialty
3	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Craft-repair
4	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Craft-repair
5	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Other-service
6	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Machine-op-in
7	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Craft-repair
8	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Sales
9	Male	*	Asian-Pac-Islander	Married-civ-spouse	*	Cambodia	Private	Sales
10	Male	*	White	Married-civ-spouse	*	Canada	Private	Exec-managen
11	Male	*	White	Married-civ-spouse	*	Canada	Private	Exec-managen
12	Male	*	White	Married-civ-spouse	*	Canada	Private	Prof-specialty
13	Male	*	White	Married-civ-spouse	*	Canada	Private	Farming-fishin
14	Male	*	White	Married-civ-spouse	*	Canada	Private	Adm-clerical
15	Male	*	White	Married-civ-spouse	*	Canada	Private	Transport-mo
16	Male	*	White	Married-civ-spouse	*	Canada	Private	Craft-repair
17	Male	*	White	Married-civ-spouse	*	Canada	Private	Transport-mo
18	Male	*	White	Married-civ-spouse	*	Canada	Private	Exec-managen
19	Male	*	White	Married-civ-spouse	*	Canada	Private	Tech-support
20	Male	*	White	Married-civ-spouse	*	Canada	Private	Craft-repair

Fig.2. Output Anonymized data

IV. RESULT ANALYSIS

In the result analysis, two different tools such like a MATLAB and Anonymization Tool are used to perform the execution of the proposed work. MATLAB has twisted away to be such a basic instrument are using units of MATLAB programs designed to support a selected challenge. These sets of programs are called toolboxes, and the particular toolbox of interest to us is image processing toolbox. Somewhat give an explanation of whole MATLAB'S capabilities, we will limit ourselves to only those features apprehensive with handling of pictures. We will provoke features, commands and procedures as required.

ARX is isolated into four points of view, which display diverse parts of the anonymization procedure. It supports configuring retreat model, value measures, with renovation method, explore the resolution space, analyzing data usefulness with analyzing privacy risks. The information import wizard likewise underpins the renaming, expelling and reordering of segments. Amid information import, information writes are naturally recognized and information purging might be performed. This implies that qualities that don't adjust to the predetermined information compose will be supplanted with particular invalid qualities, which are taken care of effectively by all strategies executed in ARX. Every single unthinkable datum showed by ARX can be sent out into CSV records through setting menus. ARX utilizes esteem speculation progressive systems to execute a wide assortment of information change strategies. These hierarchies be able to either be formed inside the software (via specific wizards) or import as of CSV files. Hierarchies twisted by ARX be able to also be export towards CSV files.

Description of Dataset:

The dataset be use to performed the execution is Adult dataset which be occupied from UCI machine-learning website. It contains 301061 tuples and 7 characteristic such like age, education, gender, marital status, native country, work class and occupation. Initial 6 attributes are used as quasi-identifiers and final individual be used as sensitive attribute. Now the generation type and distinct values of each attributes be explain below:

Table 1: Dataset Description

S.No.	Attributes name	Distinct Value	Type of Generalization
1.	Age	74	Ranges
2.	Education	16	Hierarchy (3)
3.	Gender	2	Suppression (1)
4.	Marital status	7	Hierarchy (2)
5.	Native country	41	Hierarchy (2)
6.	Work class	7	Hierarchy (2)
7.	Occupation	14	Sensitive Attribute

There are 14 sensitive attribute be separated into three set i.e. in group 1, Tech-support, machine-op-inspct, craft-repair and prof-specialty. In group 2, sales, handlers-cleaners and exec-managerial. In group 3, Farming-fishing, other-service, armed-forces, transport-moving, adm-clerical, protective-serv and priv-house-serv. The effectiveness of proposed work have been publicized in terms of Information thrashing with Discernibility.

A. Information loss:

Information loss is minimize via giving sensitive level used for receptive attribute values. Individuals tuples which feel right towards the elevated sensitive level be merely comprehensive with break of the tuples be available like it is. Information loss improves the solitude of the data and it generally decrease the likelihood of similarity attack. This calculate summarize the amount towards which distorted attribute value cover up the unique sphere of an characteristic.

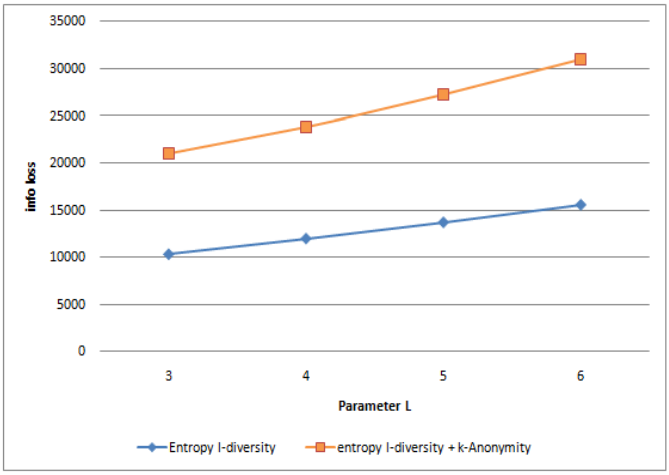


Fig.3. Information Loss for different L values

ARX equipment complicated method used for quantify this reporting base on efficient representation of simplification rules. Additionally, the model be able to be parameterized towards manipulate the quantity of simplification with repression that will be functional towards a dataset. The proposed work has more information loss which means it improves the privacy level at dissimilar integer of L parameter. It is premeditated use the formula below where r is the digit of records and T is generalized table:

$$\text{Info\_loss} = \sum_{r \in T} Info\_loss$$

B. Discernibility:

Discernibility means to calculate the digit of tuples to distinguish the data from each other. This model also estimates data quality base resting on the size of the equivalence program within the output data-set. Records which be concealed be penalize. It do not obtain interested in explanation the authentic attribute value within the output data-set. It defines the simplification with repression information used for the data loss. It is calculated by using the formula below where qid is the record of the quasi identifier group:

$$\text{Disc} = \sum_{qid \in T} |qid|^2$$

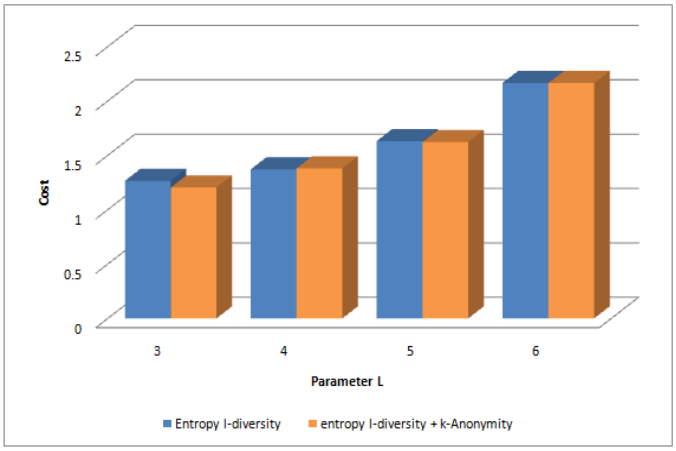


Fig.4. Discernibility for different L values

V. CONCLUSION

Information on people and substances are being gathered generally. These information can contain data that unequivocally distinguishes the individual (e.g., government managed savings number). Information can likewise contain different sorts of individual data (e.g., date of birth, postal district, sex) that are possibly distinguishing when connected with other accessible informational collections. Information are regularly shared for business or lawful reasons code, gender) that are conceivably distinguishing when connected with other accessible informational indexes. Information are frequently shared for business or lawful reasons. This paper addresses the significant subject of preserve the secrecy of the persons otherwise entity through the data distribution process. We discover preserve the secrecy via the exploit of generalization with suppressions resting on the potentially identify segment of the data. We expand previous workings within this district the length of different dimensions. First, rewarding privacy constraint be measured within combination by the procedure use for the data creature distributed. This allow us toward optimize the method of preserve privacy used for the specific procedure Our work improve resting on preceding advance via permit extra elastic simplification used for the data. These extension permit us to convert the data so that they be further useful used for their proposed reason while agreeable the privacy constraints.

In the future work, the security can be improved by applying other cryptographic techniques which also reduce the execution time. Clustering can also be improved by using the better techniques for the cluster formation of the data efficiently.

References

[1] Bhawani Singh Rathore, Anju Singh,Divakar Singh, “A Survey of Cryptographic and Non-cryptographic Techniques for Privacy Preservation”.International Journal of Computer Applications (0975 – 8887) Volume 130 – No.13, November 2015

[2] Hussein Hellani, Rima Kilany, “Towards internal privacy and flexible K-anonymity”, International Conference onApplied Research in Computer Science and Engineering (ICAR),IEEE, 2015.

[3] Rupinder Kaur and Meenakshi Bansal, “Transformation Approach for Boolean Attributes In Privacy Preserving Data Mining”,1st International Conference on Next Generation Computing Technologies (NGCT-2015), Dehradun, India, 4-5 September 2015.

[4] Ashoka K and Dr. Poornima B, “Enhanced Utility in Preserving Privacy for Multiple Heterogeneous Sensitive Attributes using Correlation and Personal Sensitivity flags”, International Conference onAdvances in Computing, Communications and Informatics (ICACCI), IEEE, 2017.

[5] Rashad Saeed, Azhar Rauf, “Anatomization through Generalization (AG): A Hybrid Privacy- Preserving Approach to Prevent Membership, Identity and Semantic Similarity Disclosure Attacks”, International Conference on Computing, Mathematics and Engineering Technologies – iCoMET, IEEE, 2018.

[6] RICARDO MENDES AND JOAO P. VILELA, “Privacy-Preserving Data Mining: Methods, Metrics, and Applications”. Received April 26, 2017, accepted May 15, 2017, date of current version June 27, 2017. Digital Object Identifier 10.1109/ACCESS.2017.2706947

[7] Putri, A walia W., Laksmiwati Hira "Hybrid Transformation in Privacy-Preserving Data Mining" 978-1-5090-5671-2/16/$31.00 ©2016 IEEE

[8] Khoirunnisa Afifah, Riza Satria Perdana" Development of Search on Encrypted Data Tools for Privacy Preserving in Digital Forensic" 978-1-5090-5671-2/16/$31.00 ©2016 IEEE.

[9] Ahmed Ali Mubark et. Al, "Semantic Anonymityin Publishing Categorical Sensitive Attributes", 8thInternational Conference onKnowledge and Smart Technology (KST), IEEE, 2016.

[10] Vanessa Ayala-Riveraet al., "Improving the Utility of Anonymized Datasets through Dynamic Evaluation of Generalization Hierarchies", 17th International Conference on Information Reuse and Integration, IEEE, 2016.

[11] Francisco Dias, "Automated Anonymization of Text Documents",Congress on Evolutionary Computation (CEC), IEEE, 2016.

[12] Shu-Ming Hsieh, Mao-Hsu Yen, Li-Jen Kao, " Semantic-Based Graph Data Anonymization For Big Data Analysis", Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, Jeju, South Korea, IEEE, 10-13 July, 2016.