International Journal of Grid and Distributed Computing
Vol. 10, No. 11 (2017), pp.33-42
http://dx.doi.org/10.14257/ijgdc.2017.10.11.04

An Information Loss Reduction Scheme in Big Data *k*-Anonymization

Sung-Bong Jang[1] and Young-Woong Ko[2*]

[1]Department of Industry-Academy Cooperation, Kumoh National Institute of Technology, 61 Daehak-ro, Gumi, Kyoung-Buk, 39177, Republic of Korea
[2]Department of Computer Engineering, Hallym University, Chuncheon, Gangwon, 200-702, Republic of Korea
sungbong.jang@kumoh.ac.kr, yuko@hallym.ac.kr

Abstract

Information loss is one of the critical issues to be resolved when applying k-anonymization to the publishing data. To solve the problem, several solutions has been proposed by many researchers. However, the existing approaches cannot be used almost for BigData because it takes too long time. For Big Data, the computational time required to reduce the information loss is regarded as a NP-hard problem. To deal with the limitation, this paper presents a scheme that is based on heuristic threshold definition. To evaluate the proposed approach, we have implemented a pro-type evaluation system. The experimental results shows that our approach improve the execution time a little for Big Data when compared with the existing approaches.

Keywords*: Privacy Protection, Big Data Anonymization, k-anonymity, Information Loss Reduction*

1. Introduction

Recent advances in Big Data analysis technology boosts up a new era of fourth industry evolution. In this situation, to efficiently use the huge Big Data, it must be delivered to a third party without worrying about the revelation of private information. However, the more the Big Data, the more danger of private information attacks. Hence, the private information securities becomes more and more important in Big Data technology. The *k*-anonymization schemes proposed by Samariti and Sweeny [1] can be applied for normalized Big Data as illustrated Figure 1. The recent advances in Big Data anonymization rooted from their technologies. The basic philosophy of *k*-anonymization is to always keep the number of the same records in the original data as more than the number of *k*. to drop down the probability of the linkage attack to 1/*k* [2][3]. To do this, data publishers have to convert or generalize the original data into another ones [4][5]. Due to this conversion, some information are lost or distorted, and this lead to the wrong results of the data analysis [6][7]. If the information is not important, it is not problem. However, if the data users are a data mining expert who have to deeply the data using like machine learning. Then, these distortion cannot be ignored. In general, privacy protection has trade-off relationships with the number of k in *k*-anonymization scheme [8-10]. Therefore, before starting the anonymization, we need to find an optimal number of *k* while satisfying the privacy constraints. To solve this problem, this paper present an enhanced method that is based on a threshold definition of the information loss.

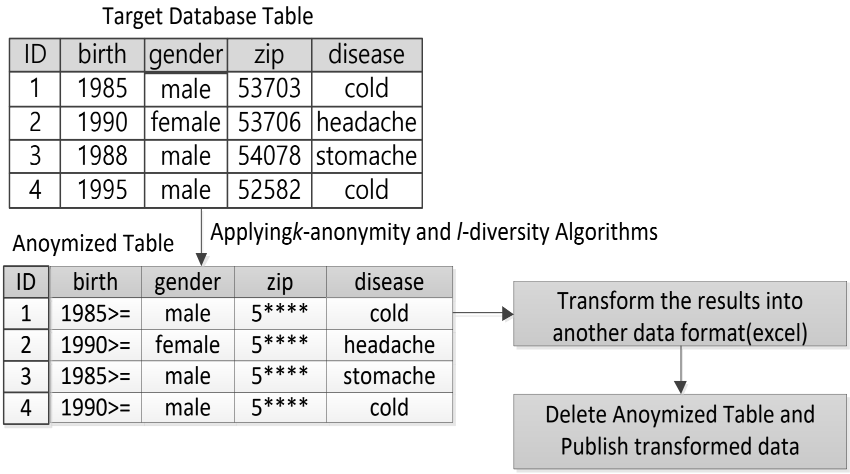
The remainder of this work is organized as follows: Section 2 describes the existing research works. In Section 3, an idea and approach are presented to reduce an information

Received (April 15, 2017), Review Result (September 13, 2017), Accepted (October 10, 2017)

[*] Corresponding Author

ISSN: 2005-4262 IJGDC

loss. Section 4 discusses about the performance evaluation and the experimental results. Finally, Section 5 are concluding remarks.


Target Database Table
Applyingk-anonymity and l-diversity Algorithms
Anoymized Table
Transform the results into another data format(excel)
Delete Anoymized Table and Publish transformed data

ID	birth	gender	zip	disease
1	1985	male	53703	cold
2	1990	female	53706	headache
3	1988	male	54078	stomache
4	1995	male	52582	cold

ID	birth	gender	zip	disease
1	1985>=	male	5****	cold
2	1990>=	female	5****	headache
3	1985>=	male	5****	stomache
4	1990>=	male	5****	cold

Figure 1. *k*-anonymization and l-diversity

2. Related Works

The state-of-the-art of the information loss reduction in *k*-anonymization can be found as follows. X. Xu *et al.*, [11] present an enhanced anonymization algorithm that is based on clustering technique. In the method, they define a special function which is used to computer the distance between optimal anonymization and the outputs. Also, they categorized the identifiers types into numeric, categorical, and structure. Using this categorization, the anonymization work is processed in two steps. In first step, all records in a database table are categorized into clusters. If there are records not contained in a cluster, then, push them one by one to a cluster. When they push it, the cluster selection is done in a way that the information loss can be minimized. They have conducted an experiment in which they compared their scheme with other existing schemes. The results shows that the presented schemes shows a better performance than the traditional ones. Gabriel et al. [12] presented a method that is based on a dimensional shrink to efficiently reduce the computational rime of the information loss reduction in *k*-anonymity. The method map the multi-dimensional attributes space to one dimensional attribute space, and detect a special properties of the narrowed space. To shrink the space, they have used Hilbert curve and iDistance transformations. By doing this, they could solve a NP-hard problem in information loss reduction. To evaluate their approaches, they have implemented an evaluation system. The experimental results shows that the methods reveal better performance than the existing solutions from the aspect of processing time and data distortion. Also, they could attain linear increase with the large datasets. Terrovitis *et al.*, [13] presented a method that is based on the cut detection in a taxonomy tree to protect private information while minimizing the information distortion and loss. The method use a modified Incognito algorithm to find an optimal point. In the previous Incognito algorithm, there was a performance problem that it reveals very slow execution time for huge datasets which is mainly caused by the dimensionality of a taxonomy tree in *k*-anonymity algorithm. To overcome the problem, they presented a partial searching scheme that is based on heuristics. By doing this, they detect the threshold of the privacy constraint. The output from the technique is a threshold level at the taxonomy tree to satisfy the k^m-anonymity while guaranteeing minimal data distortion. Furthermore, to remove a full time search to refer to a taxonomy tree and to predict how the generalization may distort the original data, they have used a special purpose of data structure, so called count-tree. The tree is a kind of modified FP-trees. Aderonke *et al.*, [14] discuss about a privacy protection when collecting and analyzing mobile crime data in various countries. They point out that privacy protection is very important because the crime data are very critical data. When applying *k*-anonymity algorithm to the mobile data, real-time collection and analysis are required because the source is composed of streaming data. The mobile environment increase the data distortion rate due to the fluctuant data stream. To deal with this problem, they present a method that is based on a buffer size adjustment according to packet arrival time. Moreover, sometimes, it fails to anonymize the arrived database records in a mobile environment. To solve this problem, they delay the records or integrate the records into a group of successful records by force. The contributions of their approaches are that the solution can stably guarantee the privacy constraint and minimize the data distortion by increasing the data accuracy in a mobile environment. Experimental results shows that information loss ratio can be effectively reduced. Jerry Chun-Wei Lin *et al.,* [15] proposed a method for anonymizing the transactional database. The approach is composed of three software blocks. The first block is responsible for preliminarily encoding record transaction into bitmaps. Then, the records are sorted in the order of Gray to minimize the data loss. After that, the sorted records are clustered into several groups to decrease down the processing time. The second block is to prevent a cyclic loop among transactions by applying a approach of solving the Traveling Salesman Problem. By doing this, they can minimize the data loss for each group. The last one is the anonymization block. The block anonymize each group of records by the following scheme. First, they partition all transactions into clusters by checking their similarities among them. Then, they compute a central spot for every cluster through the use of map and vote approach. The experiments have been conducted to evaluate their approach in terms of processing time and data loss.

The existing researches of information loss reduction are not proper to BigData anonymization because of long execution time. To overcome the limitation, this paper presents an enhanced scheme that is based on the heuristic threshold definition.

3. Proposed Approach

The proposed approach is illustrated in Figure 2. As seen from the figure, the system ask if administrators want to control information loss or not before anonymization. If they don’t want, the system perform anonymization without information loss control. On the other hand, if they want to minimize the information loss, then, the system starts anonymization with data loss control. The first step for loss control is to define the data distortion threshold. To reduce the processing time of the information loss reduction, our solution define a maximum allowed threshold of information loss base on the basic parameters which are determined by data publishers in advance before starting BigData anonymization. The calculation is done using the equation (1) as follows.

$$BigDataAnoLoss_{metric} = AnoDomain \times w_{bd} + AnoPurpose \times w_{ap} + AnoCriticality \times w_{ac} \quad \text{- -- (1)}$$

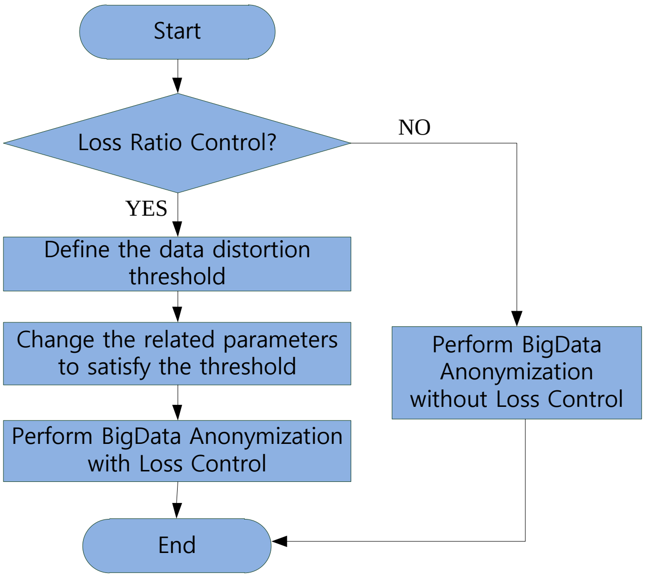

Start
Loss Ratio Control?
NO
YES
Define the data distortion threshold
Change the related parameters to satisfy the threshold
Perform BigData Anonymization with Loss Control
Perform BigData Anonymization without Loss Control
End

Figure 2. Proposed Approach: Data Distortion Minimization based on Threshold Definition

In equation (1), *BigDataAnoLoss*$_{th}$ represents the predefined threshold that is computed according to three parameters as represented by *AnoDomain*, *AnoPurpose*, and *AnoCriticality*, respectively. The *AnoDomain* represents a value of allowed information loss depending on the data characteristics. Here, data characteristics refers to what type of data are included in the source data. For example, if the source data contain the medical records, there may exists much private information like identification number, disease information, cures, and home address. In this case, we give higher priority to privacy preservation than to data loss. On the other hand, suppose that the anonymization source contain the winning and losing information of the football players in the game, then, we give lower priority to the privacy preservation than to data loss because those data are regarded as much less critical. Based on these heuristics, the value of *AnoDomain* is set to be one of the default weight values as shown in Table 1.

Table 1. Assignment of AnoDomain Values

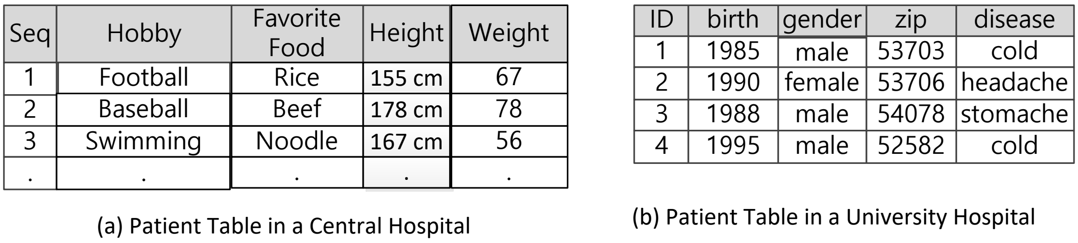
Industry Domain	Weight of *AnoDomain*	Assigned Value
Medical	High	75
Financial	Very High	100
Sports Records	Low	25
Public Organization's Open Data	Low	25
Military	High	75
Product Sales Data	Middle	50

Next, the *AnoPurpose* denotes a score of loss information according to a data acquiring purpose. This value is assigned by data administrators by investigating the purpose of the data acquisition through simple interview or e-mail. For example, if data receivers are researchers who have to deeply analyze the acquired data for data mining, we see that information loss should be minimal when distributing the BigData to provide more accurate data with them. In this case, to minimize the distortion, lower priority must be given to privacy preservation. On the other hand, if deep analysis is not required, higher priority is given to privacy preservation. For instance, suppose that some teachers need to get middle school scores to get trivial information such as average or standard deviation every month of the students. Based on this observation, appropriate weight values are assigned to *AnoPurpose* as shown in Table 2.

Table 2. Values Assignment Depending on Data Acquisition Purposes

Data Acquisition Purpose	Data Distortion Constraint	Assigned Values
Simply Review	Low	25
Get Simple Statistical Information	Middle	50
Complex Statistical Processing	High	75
Data Mining	Very High	100

Basically, we have divided weight into three: very high, high, middle, and low. Of course, sometimes, there are some cases where the data characteristics and data acquisition are not enough to define the loss threshold because our heuristics are not accurate. For example, suppose that there are no critical information in medical records to be anonymized. Then, we don't have to give high priority to privacy preservation. To manage such heuristics miss, our solution considers one more parameter, data criticality, as represented "*AnoCriticality*" in equation (1). The value can be determined by analyzing the quasi and sensitive attributes. For example, suppose two type of medical database tables as shown in Figure 3.



Seq	Hobby	Favorite Food	Height	Weight
1	Football	Rice	155 cm	67
2	Baseball	Beef	178 cm	78
3	Swimming	Noodle	167 cm	56
.	.	.	.	.

(a) Patient Table in a Central Hospital

ID	birth	gender	zip	disease
1	1985	male	53703	cold
2	1990	female	53706	headache
3	1988	male	54078	stomache
4	1995	male	52582	cold

(b) Patient Table in a University Hospital

Figure 3. Example of Medical Database Tables

The first table contains personal's hobby, favorite food, height, weight, and the second table contains disease, address, social security number. When compared two table, we can see that the first table are less sensitive and critical than the second table. In this case, for the first table, our work gives less priority to privacy protection, on the other hand, for the second one, we give more priority to privacy protection.

In equation (1), there are three basic constants for each parameters as represented by the w_{bd}, w_{bd}, w_{bd}. Those constant represent weight values for each parameter. The summation of three constants is one. These values are heuristically determined by data publishers. Heuristically, the order of important parameters are a data acquisition purpose, data criticality, and data domain. For example, the default values for weight of each parameter are 0.5, 0.3, and 0.2, respectively. These weights can be changed according to the situation.

After finishing the computation of the loss threshold σ_{th}, the system change the related parameters. The most important parameter to be changed is k value in k-anonymity algorithm. In general, the data loss ratio almost linearly increases with the value of k. Therefore, if we want to satisfy the σ_{th}, we have to reduce the k. At the same time, we have to consider the basic privacy constraints. Hence, our work search for the largest value of k that satisfy the defined threshold, and then, the value is set to be the k value for anonymization. The next parameter to be changed is a level of generalization in a taxonomy tree. For example, consider the following taxonomy as shown in Figure 4.

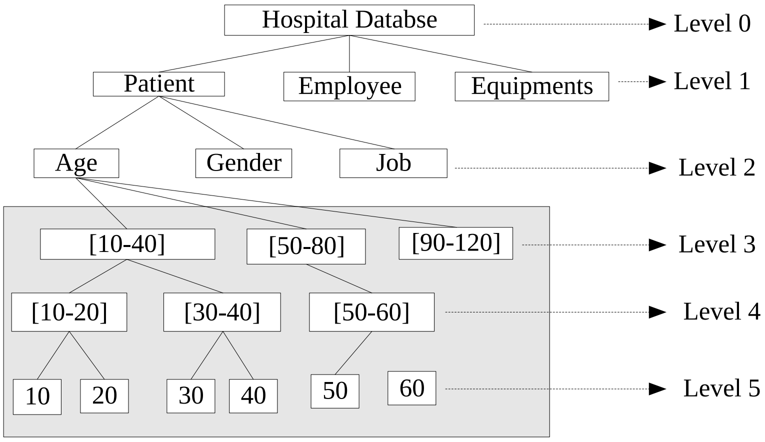


Figure 4. Example of Generalization in a Taxonomy Tree

Suppose that a data mining expert would like to get statistics of the patients whose age ranges from 10 to 20 for a disease cold. Then, a data publisher is planning to anonymize the age through generalization using the taxonomy tree shown in Figure 4. At this situation, if the number of 20 at level 5 of age is generalized into <10-40> at level 3, then, the expert could not get what he want because changed data would contain a patient of 40. In other words, it does not satisfy the data distortion constraint. Therefore, the generalization level must be restricted to satisfy a level threshold. To do this, our work changes *k* value and the level while satisfying the following equation.

$$max\_k\_value * 0.7 + max\_generalization\_level*0.3 <= \sigma_{th} \text{ - - - - - } (2)$$

Here, the maximum *max_k_*value represents the default maximum value of the *k* and *max_generalization_level* represents the highest level of the taxonomy tree that can be changed when apllying *k*-anonymity without the consideration of information loss.

4. Performance Evaluation

To evaluate the propose approach, we have implemented the prototype system. The basic system architecture is shown in Figure 5. The architecture has a three-tier database system. In the system, the MySQL DBMS [16][17] is used. The administrators can start anonymization task through Web interface [18][19]. The database server manages creation, modification, and deletion of database records, and process users' SQL query [20][21]. In anonymization server, *k*-anonymity software block and web server are running. K-anonymity block is responsible for anonymization based on *k*-anonymity algorithm, and it also convert the anonymized data into another format of data such as Excel. By using the pro-type system, we have measured two kinds of metrics.

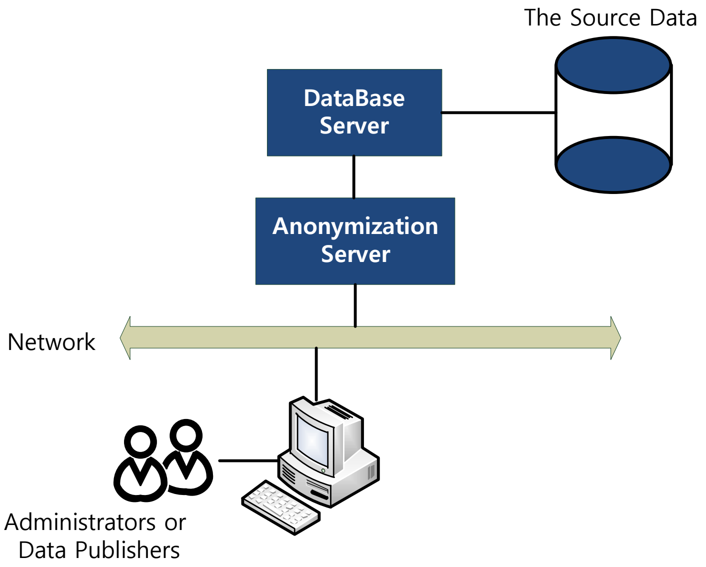


Figure 5. Basic Architecture of Evaluation System

First, we have measured the processing time for the variable *k* in the proposed approach and conventional scheme. In the experiment, the conventional method for information loss reduction was the one proposed by Terrovitis [10]. In the scheme, they have used the full time search to satisfy the complicated metric which is based on the following equation.

$$NCP(t) = \begin{cases} 0, & |U_t| = 1 \\ |U_t| \big/ |\rho|, & otherwise \end{cases} \quad \text{- -} \qquad \text{- - - (3)}$$

Here, U$_t$ represents the target item of *t* in the taxonomy tree. The $|U_t|$ represents the number of lead nodes located below U$_t$ and $|\rho|$ represents the number of nodes of the entire taxonomy. In reality, we have tried to implement the full algorithm, however, it is too difficult to finish the job because the algorithm was NP-Hard problem. So, to conduct an experiment, we have implemented the algorithm in an alternative way where they do not perform the complicate computation as shown in equation (3). Instead, we have used the computational time that was generated based on random number. In other words, when we starting the computing, we generate the long random number. And if it is not finished within that time, we have stopped the computation, and we have used the random number time as final computational time. So, the experimental results are illustrated in Figure 6.

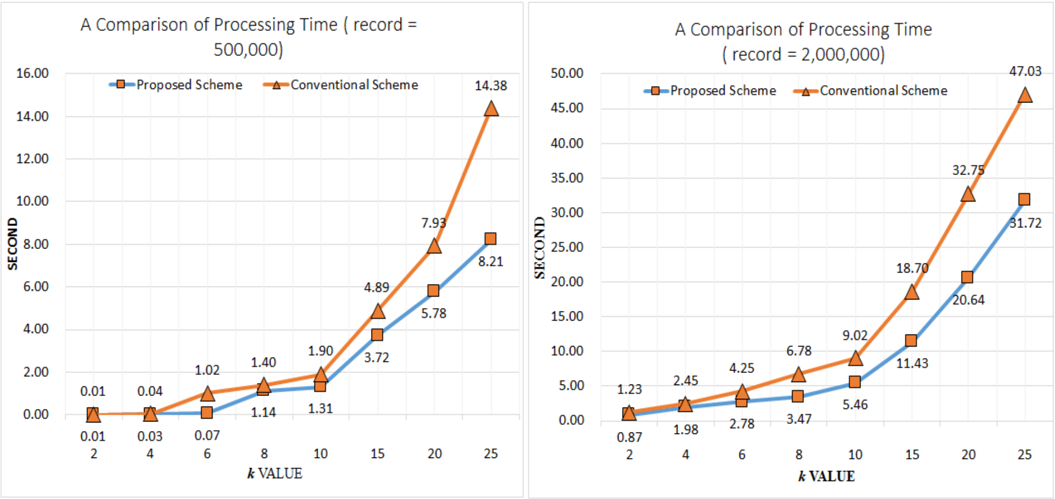


Figure 6. Comparison of Processing Time

In the test, we have considered only generalization scheme only in *k*-anonymization. In the experiment, the initial value of *k* was two. The figure- (a) shows the measurement results of processing time when the number of records is 500,000 and Figure-(b) shows that of processing time when the number of record is 2,000,000. From the results, we see that the proposed scheme can reduce the processing time a little when compared with the conventional one. The reason that the result has a difference is because the conventional scheme use two complicated measure to compute the information loss when they anonymize the BigData.

In the next experiment, we have measured the information loss for a specific *k* value. The metric to measure the information loss is shown in equation (3). The experimental results are shown in Figure 7.

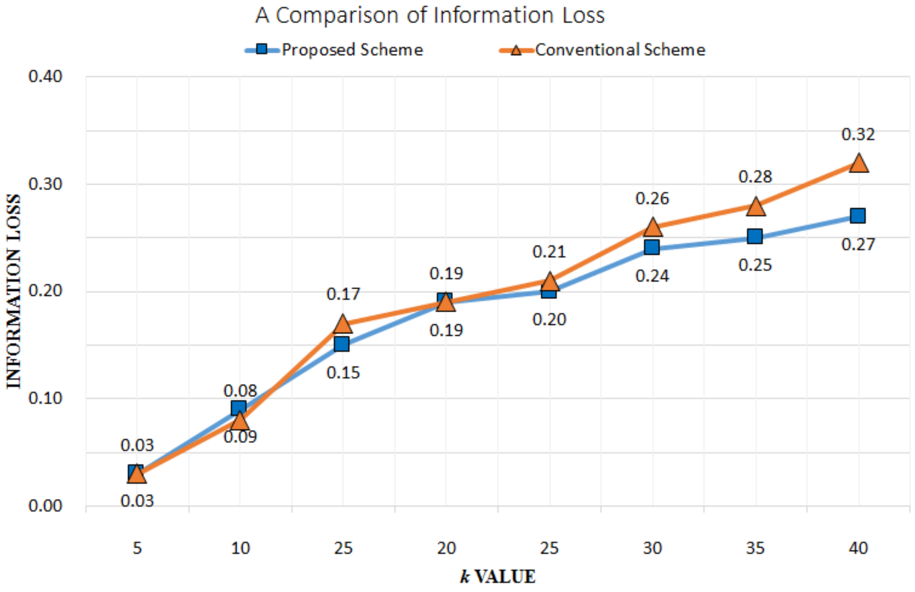


Figure 7. Comparison of Information Loss Ratio

From the results, we see that the information loss is almost similar to that of the conventional scheme for the k value smaller than 25. However, if the value is larger than 25, the loss of our solution is definitely less than that of the conventional scheme. In reality, the *k* value in our work changes dynamically as smaller one according to the according to the defined threshold.

5. Conclusions

This paper discusses about how to find an optimal point to reduce the information loss while guranteeing privacy when applying k-anonymity to Big Data anonymization. To solve the limitation, a method that is based on a predefined threshold definition is presented. The experimental results shows that the proposed solutions achieve a better performance than the existing approaches in terms of data loss ratio and processing time.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2016R1C1B1014346).

References

[1] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information:k-anonymity and Its Enforcement through Generalization and Suppression", IEEE, **(1998)**.

[2] C. Dandan, L. Yidong, W. Tao, Z. Lei Zhang and S. Hong, "Practical Anonymization for Protecting Privacy in Combinatorial Maps", Proceedings of the 15th International Conference on Parallel and Distributed Computing, Applications and Technologies, **(2014)**.

[3] M. Neelam, L. Grigorios and S. Jianhua, "A Parallel Method for Scalable Anonymization of Transaction Data", Proceedings of the 14th International Symposium on Parallel and Distributed Computing, **(2015)**.

[4] S.-C. Jordi, D.-F. Josep, S. David and M. Sergio, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation", IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, **(2015)**.

[5] W. Sai, W. Xiaoli, W. Sheng, Z. Zhenjie, K. H. T. Anthony, "K-Anonymity for Crowdsourcing Database", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, **(2014)**.

[6] J. P. Jisha, S. P. Anitha and V. N. Krishnachandran, "Disclosure Risk of Individuals: A k-Anonymity Study on Health Care Data Related to Indian Population", Proceedings of the International Conference on Data Science & Engineering (ICDSE), **(2015)**.

[7] [7] L. Yuechuan and L. Yidong, "On Preserving Private Geosocial Networks against Practical Attacks", Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), **(2015)**.

[8] G. Longkun and S. Hong, "Privacy-Preserving Internet Traffic Publication", IEEE Trustcom/BigDataSE/ISPA, **(2016)**, pp.884-891.

[9] A. M. Mona, H. N. Magdy and M. G. Sahar, "A Clustering Approach for Anonymizing Distributed Data Streams", Proceedings of the 11th International Conference on Computer Engineering & Systems (ICCES), **(2016)**.

[10] O. A. Otgonbayar, P. Zeeshan and D. Keshav, "Toward Anonymizing IoT Data Streams via Partitioning", Proceedings of the IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), **(2016)**.

[11] X. Xu and M. Numao, "An Efficient Generalized Clustering Method for Achieving K-Anonymization", Proceedings of the Third International Symposium on Computing and Networking (CANDAR), **(2015)**.

[12] G. Gabriel, K. Panagiotis, K. Panos and M. Nikos, "Fast Data Anonymization with Low Information Loss", Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), **(2007)**.

[13] M. Terrovitis, N. Mamoulis and P. Kalnis, "Privacy-preserving Anonymization of set-valued Data", Proceedings of the VLDB Endowment, **(2008)**.

[14] B. S. Aderonke and V. D. M. K. Anne, "Adaptive Buffer Resizing for Efficient Anonymization of Streaming Data with Minimal Information Loss", Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP), **(2015)**.

[15] J. Chun-Wei Lin, Q. Liu, P. Fournier-Viger and T.-P. Hong, "PTA: An Efficient System for Transaction Database Anonymization", IEEE Access, vol. 4, **(2016)**, pp. 6467-6479.

[16] Q. Liu, H. Shen and Y. Sang, "Privacy-preserving Data Publishing for Multiple Numerical Sensitive Attributes", Journal of Tsinghua Science and Technology, vol. 20, no. 3, **(2015)**, pp. 246-254.

[17] L. Grigorios and G.-D. Aris, "Utility-preserving Transaction Data Anonymization with Low Information Loss", Expert Systems with Applications, vol. 39, no. 10, **(2012)**, pp. 9764-9777.

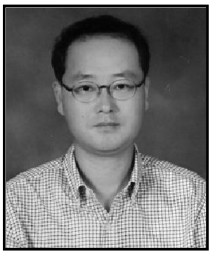
[18] R. Mehri and S. H. Mostafa, "An Improved Ambiguity+ anonymization Technique with Enhanced Data Utility", Proceedings of the 7th Conference on Information and Knowledge Technology (IKT), **(2015)**.

[19] Z. Jianpei, Z. Ying Zhao, Y. Yue and Y. Jing, "A K-anonymity Clustering Algorithm Based on the Information Entropy", Proceedings of the IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD), **(2014)**.

[20] L. Na Li, Z. Nan and K. D. Sajal, "Relationship Privacy Preservation in Publishing Online Social Networks", Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing, **(2011)**.

[21] G. Gabriel, K. Panos and T. Yufei, "Anonymous Publication of Sensitive Transactional Data", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 2, **(2011)**, pp. 161-174.

Authors



Sung-Bong Jang, he received his a M.S. and Ph.D. degrees from Korea University, Seoul, Korea in 1999 and 2010, respectively. He worked at the Mobile Handset R&D Center, LG Electronics from 1999 to 2012. Currently, he is an associate professor in the Department of Industry-Academy Cooperation, Kumoh National Institute of Technology in Korea. His interests include BigData *k*-anonymization, Mobile Video Communication, and Privacy Protection.

Young-Woong Ko, he received both a M.S. and Ph.D. in computer science from Korea University, Seoul, Korea, in 1999 and 2003, respectively. He is now a professor in Department of Computer engineering, Hallym University in Korea. His research interests include operating systems, embedded systems and multimedia systems.