International Journal of Internet, Broadcasting and Communication Vol.11 No.2 59-66 (2019)
http://dx.doi.org/10.7236/IJIBC.2019.11.2.59

IJIBC 19-2-8

Strategy Design to Protect Personal Information on Fake News based on Bigdata and Artificial Intelligence

Jangmook Kang, Sangwon Lee*

Department of Bigdata & Industry Security, Namseoul University
kangjm@nsu.ac.kr
**Department of Computer & Software Engineering, Wonkwang University*
sangwonlee@wku.ac.kr

Abstract

The emergence of new IT technologies and convergence industries, such as artificial intelligence, bigdata and the Internet of Things, is another chance for South Korea, which has established itself as one of the world's top IT powerhouses. On the other hand, however, privacy concerns that may arise in the process of using such technologies raise the task of harmonizing the development of new industries and the protection of personal information at the same time. In response, the government clearly presented the criteria for deidentifiable measures of personal information and the scope of use of deidentifiable information needed to ensure that bigdata can be safely utilized within the framework of the current Personal Information Protection Act. It strives to promote corporate investment and industrial development by removing them and to ensure that the protection of the people's personal information and human rights is not neglected. This study discusses the strategy of deidentifying personal information protection based on the analysis of fake news. Using the strategies derived from this study, it is assumed that deidentification information that is appropriate for deidentification measures is not personal information and can therefore be used for analysis of big data. By doing so, deidentification information can be safely utilized and managed through administrative and technical safeguards to prevent re-identification, considering the possibility of re-identification due to technology development and data growth.

Keywords: *Fake News, Bigdata, Artificial Intelligence, Deidentification.*

1. Introduction

Major advanced countries such as the U.S. and the U.K. are pushing policies to revitalize the data industry as demand for data use is surging due to the development of IT convergence technologies such as bigdata and IoT. In response, specific strategies on the criteria, procedures and methods of deidentification measures needed to utilize bigdata [1-11] are needed. In addition, opening up public information, sharing it to

Manuscript Received: Mar. 24, 2019 / Revised: Apr. 1, 2019 / Accepted: Apr. 9, 2019
Corresponding Author: sangwonlee@wku.ac.kr
Tel: +82-63-850-6566
Department of Computer & Software Engineering, Wonkwang University, Iksan, Republic of Korea

transparent and efficient government operations, and utilizing bigdata is becoming an essential means of implementing scientific policies and providing customized services. In particular, the value of data is increasing in the creation of new services through bigdata analysis and IoT technology and the revitalization of new industries. However, social calls continue to be made to strengthen privacy policies as personal information leakage incidents, both large and small, continue. With the development of new industries and technologies that require diverse data utilization, the risk of personal information infringement is also on the rise. Major advanced countries such as the U.S. and the U.K. are pursuing policies to revitalize the data industry while minimizing the possibility of personal information breaches. Specific strategies are needed to ensure that deidentified information can be used in the industry while safeguards are in place to prevent invasion of privacy. This study aims to design strategies for deidentification to protect personal information by analyzing fake news.

2. Related Works

We look at the concepts of personal information and deidentification that are mainly used in this study.

2.1 Personal Information

If personal information can be identified by combining information that is not identified in itself with other information, it is regarded as personal information and restricted from its use. Such an attitude can cause inconvenience to the bigdata industry and others, but it is difficult to make decisions in a way that would make it easier for individuals to make self-determination for the industry. Personal information is a concept that contains a variety of information about an individual. It can be a name, it can be a food preference, it can be a job. It’s all personal information about which websites you use frequently and which reason you like. Individual biometric information for iris recognition or fingerprint recognition is also personal information. It is distinct from this general public perception to see how far one should view personal information from a legal perspective. In Korea, where laws on the protection of personal information vary, the meaning of personal information prescribed by individual laws can be interpreted differently, but in one legal system, the concept of personal information should be interpreted the same. The Information and Communication Network Act stipulates that ‘personal information’ is information about a surviving individual, including information such as code, text, voice, sound, and video information that can identify the individual by name, resident registration number, etc. (If the information in question is easily combined with other information even if it is not possible to recognize a particular individual, the information is included). Personal information includes information that can be easily combined with other information that can identify a particular person with that information alone. However, not all other information that can be combined with that information should be held by the same person. On the other hand, it refers to the fact that information and other information can be easily combined to identify a particular individual, regardless of whether it is easy to obtain or difficult to obtain, rather than meaning to obtain other information easily. The progress of today’s information services has led to information that previously could not be easily combined to serve as an identifiable individual. Therefore, it is objectively obvious that even mechanical information has been given to a particular individual, and it is reasonable to view it as personal information if it is likely that an individual will be identified through such information.

The Korea Communications Commission said on Jan. 27, 2016, that it will allow anonymous personal information to be freely used to revitalize key IT industry sectors such as big data, the Internet of Things and the cloud. It is said that the Act will stipulate the basis of deidentification and anonymity measures so that individuals can freely use the information, and provide specific criteria to prevent such information from being re-identified and used. It is said that if the government wants to use personal information and information about use anonymously and anonymously, it will introduce an "opt-out" method to enable its useable. Despite such industrial needs, privacy aspects should not be taken lightly in light of the potential infringement of the bigdata industry's self-determination. A privacy model is known as a way of thinking that values privacy over the use of personal information. When designing legislation for personal information, we should not forget this view, either. If information that is not identified in itself can be identified in combination with other information, it is regarded as personal information and restricted from its use. Such an attitude can inconvenience industrialization such as the bigdata industry, but it is difficult to make decisions in a way that makes individuals' self-determination easier to sacrifice for the industry. Balance, that is a task we must constantly pursue, even if it is inconvenient.

2.2 Deidentification

A set of processes or methods that transform part or all of personal information so that a particular individual cannot be identified. Personal identification elements include identifiers that can identify a particular individual directly in themselves, such as name, address, social security number, date of birth, phone number, email address, medical record number, etc., and quasi-identifier that can be identified in combination with other information such as age, gender, residence area, nationality, homepage URL, etc. An alias treatment that replaces the identification elements with other values; total processing that shows only the aggregate value of the data and does not show the individual identification elements; data that erases some identification elements; categorization that hides the exact value of the data and converts it into category values; and data masking that prevents the critical identifier from being seen. Deidentified personal information is subject to strict management as it can be re-identified during the process of collecting and analyzing bigdata such as Web, SNS and medical records. Anonymous is used in terms similar to deidentification, which means that personal information is no longer identifiable. The ISO/IEC 20889 standard defines 'the process of deidentification' as 'the process of removing links between the information subject and the set of identification attributes' and 'the method of transforming a data set with the aim of reducing the degree of information being linked to a personal information subject'.

3. Designing Deidentifiable Action Steps

This step presents the criteria for actions that must be followed by operators or others who intend to use or provide personal information through deidentifiable measures. If personal information is collected and used in accordance with the relevant statutes, such as the Statistics Act, it shall be handled according to the relevant statutes. The following are step-by-step actions: (1) Pre-Review, (2) Deidentification Measures, (3) Adequacy Assessment, and (4) Post Management.

3.1 Pre-Review

After reviewing whether personal information is applicable, use it freely without legal regulations if it is evident that it is not personal information. Businesses that want to process information for analysis of bigdata will make a judgment based on the criteria below regarding whether the information is personal information. If it is clear that such information does not correspond to personal information, it can be used for analysis of bigdata without any action. Separate measures are taken if it is deemed to be in personal information. In principle, identifiers included in the information collection should be deleted.

3.2 Deidentification Measures

To prevent individuals from being recognized by using methods such as deleting or replacing all or part of an individual's identifiable elements in the information collection (data set). An 'identifier' is a value or name that is uniquely assigned to an individual or an object related to an individual. However, identifiers that are essential for data use should be utilized after deidentification measures. In principle, attributes included in the information collection are deleted if they are not relevant to the purpose of data use. A 'property' is an individual-related information that, when easily combined with other information, can recognize a particular individual. If there is an identification element among the attributes related to the purpose of data use, an identification measure shall be taken using a technique such as alias processing and total processing. Attributes such as rare diseases and rare experiences are highly likely to be personally identifiable depending on the specific circumstances, requiring strict deidentification measures. Several techniques are used independently or in combination, such as alias processing, total processing, data deletion, data categorization, and data masking. If only 'assigned-in' technology is used alone, it is difficult to say that it is a sufficient deidentification measure. Each technique has a variety of detailed technologies that can be implemented, and appropriate techniques or detailed technologies are selected and utilized by considering the purpose of data use and the advantages and disadvantages of each technique. Once the deidentification measures are completed, the next step of action is required.

3.3 Adequacy Assessment

The 'Performance Group for Deidentifiable Action Assessments' evaluates whether individuals can be identified easily by combining them with other information. There is a concern that individuals may be identified through combinations of information, such as public information, and various reasoning techniques, if deidentification measures are not sufficient. A 'Assessment Group on the adequacy of deidentification measures' participated by external experts under the responsibility of the personal information protection manager shall be organized to carry out a rigorous assessment of the possibility of personal identification. The k-anonymity among privacy protection models is utilized for adequacy evaluation. k-anonymity is the minimal means of evaluation, and additional evaluation models (l-diversity, t-proximity) are utilized, if necessary.

The adequacy test procedure is carried out as follows. (1) Preparation of basic data: The personal information processor prepares basic data such as the data specifications, status of deidentification measures, and the level of management by the use agency. (2) Company of assessment group: The assessment team consists of three or more persons in charge of personal information protection. (3) (Perform assessment: The assessment team evaluates the adequacy of the level of deidentification measures using the basic data prepared by the personal information processor and the k-anonymous model. (4) Additional deidentification measures: If the assessment results are 'unacceptable', the personal information processor carries out additional deidentification measures by reflecting the opinions of the survey team. (5) Using data: If deidentification measures are assessed as appropriate, they are allowed to be used or provided for analysis of big data.

3.4 Post Management

The medical institution carries out necessary measures to prevent re-identification in the course of the utilization of deidentification information, such as safety measures for deidentification and monitoring of the possibility of re-identification.

Essential safeguards should be implemented, as deidentified information is likely to be identified in combination with other information when leaked. Protection measures are divided into management and technical safeguards. Managed safeguards provide actions such as designation of management personnel for deidentification information files, prohibition of sharing information related to deidentification measures, and destruction of the purpose of use. Technical safeguards provide measures such as control of access to deidentifiable information files, control of access records, and installation and operation of security programs. In case of deidentification information leakage, management and technical protection measures are carried out to analyze the cause of leakage and prevent further leakage. The leaked deidentification information is retrieved and destroyed.

On the other hand, operators who use deidentifiable information or want to provide it to third parties should be regularly monitored for possible re-identification of such information. Additional deidentification measures shall be taken if the monitoring results correspond to any of the following check items. If the person in charge of providing or delegating deidentification information finds the possibility of a re-identification, he shall immediately notify the person in charge of processing the information, demand that the information be discontinued, and retrieve and destroy the information.

In the event of deidentified information being provided to third parties or processed and entrusted, the contract shall include information on risk management for re-identification. (1) Prohibition of re-identification: Any operator who has received or has been entrusted with deidentification information shall be prohibited from attempting to re-identify through combination with other information and other information. (2) Restriction of re-offering or re-commissioning: Any person who provides deidentifiable information or consigns to be processed shall have a range for re-offering or re-commissioning. (3) Notifying in the event of a risk of a re-identification: In the event of a situation in which the re-identification or probability of a re-identification is increased, the obligation to stop processing the data and to notify it shall be specified.

If the deidentified information is reidentified, the necessary measures should be taken to stop processing the information and prevent the leakage of such personal information. The recidivism information shall be destroyed immediately, but the deidentifiable action procedure shall be re-examined in order to utilize the information again.

4. Data Governance for Deidentification

A support system is needed to safely utilize personal information through deidentification measures. It should help small businesses and startups utilize bigdata through consulting and specialized training required for deidentification measures such as support for assessment of the adequacy of deidentification measures carried out by personal information processors. It should actively respond to the risks of re-identification due to the emergence of artificial intelligence [12-15] and new combined technologies.

The specialized institutions for each sector shall be operated by the relevant departments designated and publicly announced. When combining information collections held by different operators for use in bigdata analytics, the identifiers assigned to each individual should be used as matching kits. In this case, using the identifiers themselves to match the identity of the informant may violate the current law. Therefore, for information collection to be combined and analyzed, the matching key function is temporarily performed only in the process of coupling. It is necessary to use ‘temporary substitute’ keys. Even when a temporary replacement key is allowed to be used, a support and management system is needed, such as to require a combination to be made only by a specialized agency (a third-party public institution) to prevent the possibility of personal information infringement through reckless combination.

5. Conclusions

In carrying out data analysis for fake news judgment, we studied strategies for personal information deidentification. It is pushing for policies to revitalize the data industry as demand for data use has surged due to the development of IT convergence technologies such as bigdata and IoT. In response, specific strategies for the criteria, procedures and methods of deidentification necessary to utilize bigdata will seek to establish a safe base for the use of bigdata and strengthen the protection of personal information. Personal information for reading fake news not only applies to text data [16-23] but also to relational databases. For systematic, deidentifiable measures of personal information, research on systematic action laws and the establishment of action institutions is imperative. In addition, specific algorithms for deidentification and re-identification of personal information and research on step-by-step measures are also needed. In the future studies, there should be some empirical evidence to support the analysis of the contents.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03038738 / Algorithm Design & Software Architecture Modeling to Judge Fake News based on Artificial Intelligence).

References

[1] S. Park, J.S. Hwang, and S. Lee, “A Study on the Link Server Development Using B-Tree Structure in the Bigdata Environment”, Journal of Internet Computing and Services, Vol. 16. No. 1. pp. 75-82, 2015.
DOI: https://doi.org/10.7472/jksii.2015.16.1.75.

[2] S.B. Park, S. Lee, S.W. Chae, and H. Zo, “An Empirical Study of the Factors Influencing the Task Performances of SaaS Users”, Asia Pacific Journal of Information Systems, Vol. 25. No. 2. pp. 265-288, 2015.
DOI: https://doi.org/10.14329/apjis.2015.25.2.265.

[3] S. Park, and S. Lee, “Big Data-oriented Analysis on Issues of the Hyper-connected Society”, The E-Business Studies, Vol. 16. No. 5. pp. 3-18, 2015.
DOI: https://doi.org/10.15719/geba.16.5.201510.3.

[4] Jumin Lee, S.B. Park, and S. Lee, “Are Negative Online Consumer Reviews Always Bad? A Two-Sided Message Perspective”, Asia Pacific Journal of Information Systems, Vol. 25. No. 4. pp. 784-804, 2015.
DOI: https://doi.org/10.14329/apjis.2015.25.4.784.

[5] J.K. Kim, S.W. Lee, and D.O. Choi, “Relevance Analysis Online Advertisement and e-Commerce Sales”, Journal of the Korea Entertainment Industry Association, Vol. 10. No. 2. pp. 27-35, 2016.
DOI: https://doi.org/10.21184/jkeia.2016.04.10.2.27.

[6] S.W. Lee, and S.H. Kim, “Finding Industries for Bigdata Usage on the Basis of AHP”, Journal of Digital Convergence, Vol. 14. No. 7. pp. 21-27, 2016.
DOI: https://doi.org/10.14400/JDC.2016.14.7.21.

[7] S. Lee, and S.Y. Shin, "Design of Health Warning Model on the Basis of CRM by use of Health Big Data", Journal of the Korea Institute of Information and Communication Engineering, Vol. 20. No. 4. pp. 1460-1465, 2016.
DOI: https://doi.org/10.6109/jkiice.2016.20.8.1460.

[8] M. Nam, and S. Lee, "Bigdata as a Solution to Shrinking the Shadow Economy", The E-Business Studies, Vol. 17. No. 5. pp. 107-116, 2016.
DOI: https://doi.org/10.20462/TeBS.2016.10.17.5.107.

[9] S.H. Kim, S. Chang, and S.W. Lee, "Consumer Trend Platform Development for Combination Analysis of Structured and Unstructured Big Data", Journal of Digital Convergence, Vol. 15. No. 6. pp. 133-143, 2017.
DOI: https://doi.org/10.14400/JDC.2017.15.6.133.

[10] Y. Kang, S. Kim, J. Kim, and S. Lee, "Examining the Impact of Weather Factors on Yield Industry Vitalization on Bigdata Foundation Technique", Journal of the Korea Entertainment Industry Association, Vol. 11. No. 4. pp. 329-340, 2017.
DOI: https://doi.org/10.21184/jkeia.2017.06.11.4.329.

[11] S. Kim, H. Hwang, J. Lee, J. Choi, J. Kang, and S. Lee, "Design of Prevention Method Against Infectious Diseases based on Mobile Bigdata and Rule to Select Subjects Using Artificial Intelligence Concept", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 174-178, 2018.
DOI: https://doi.org/10.14419/ijet.v7i3.33.18603.

[12] I. Jung, H. Sun, J. Kang, C.H. Lee, and S. Lee, "Bigdata Analysis Model for MRO Business Using Artificial Intelligence System Concept", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 134-138, 2018.
DOI: https://doi.org/10.14419/ijet.v7i3.33.18593.

[13] S. Kim, S. Park, J. Kang, and S. Lee, "The Model of Bigdata Analysis for MICE Using IoT (Beacon) and Artificial Intelligence Service (Recommendation, Interest, and Movement)", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 314-318, 2018.
DOI: https://doi.org/10.14419/ijet.v7i3.33.21192.

[14] S.H. Kim, J.K. Choi, J.S. Kim, A.R. Jang, J.H. Lee, K.J. Cha, and S.W. Lee, "Animal Infectious Diseases Prevention through Bigdata and Deep Learning", Journal of Intelligence and Information Systems, Vol. 24. No. 4. pp. 137-154, 2018.
DOI: https://doi.org/10.13088/jiis.2018.24.4.137.

[15] S. Lee, and I. Jung, "Development of a Platform Using Big Data-Based Artificial Intelligence to Predict New Demand of Shipbuilding", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 19. No. 1. pp. 171-178, 2019.
DOI: https://doi.org/10.7236/JIIBC.2019.19.1.171.

[16] H. Hwang, S. Lee, S. Kim, and S. Lee, "Building an Analytical Platform of Bigdata for Quality Inspection in the Dairy Industry: A Machine Learning Approach", Journal of Intelligence and Information Systems, Vol. 24. No. 1. pp. 125-140, 2018.
DOI: https://doi.org/10.13088/jiis.2018.24.1.125.

[17] Y. Shon, J. Park, J. Kang, and S. Lee, "Design of Link Evaluation Method to Improve Reliability based on Linked Open Bigdata and Natural Language Processing", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 168-173, 2018.
DOI: https://doi.org/10.14419/ijet.v7i3.33.18601.

[18] T. Minami and K. Baba, "A Study on Finding Potential Group of Patrons from Library's Loan Records", International Journal of Advanced Smart Convergence, Vol. 2, No. 2, pp. 23-26, 2013.
DOI: https://doi.org/10.7236/IJASC2013.2.2.6

[19] S.H. Kim, M.S. Kang, and Y.G. Jung, "Big Data Analysis using Python in Agriculture Forestry and Fisheries", International Journal of Advanced Smart Convergence, Vol. 5. No. 1, pp. 47-50, 2016.
DOI: https://doi.org/10.7236/IJASC.2016.5.1.47

[20] W.Y. Kim, “A Practical Study on Data Analysis Framework for Teaching 3D Printing in Elementary School”, International Journal of Internet, Broadcasting and Communication, Vol. 8, No. 1, pp. 73-82, 2016.
DOI: https://www.earticle.net/Article/A263475

[21] H.C. Kang, K.B. Kang, H.K. Ahn, S.H. Lee, T.H. Ahn, and J.W. Jwa, “The Smart EV Charging System based on the big data analysis of the Power Consumption Patterns”, Vol. 9, No. 2, pp. 1-10, 2017.
DOI: https://www.earticle.net/Journal/Issues/821/22509

[22] Y.I. Kim, S.S. Yang, S.S. Lee, S.C. Park, “Design and Implementation of Mobile CRM Utilizing Big Data Analysis Techniques”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 14, No. 6, pp. 289-294, 2014.
DOI: https://doi.org/10.7236/JIIBC.2014.14.6.289

[23] S.J. Oh, “Design of a Smart Application using Big Data”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 15, No. 6, pp. 17-24, 2015.
DOI: https://www.earticle.net/Article/A259710