

Anonymity Personal Information Secure Method in Big Data environment

Sunghyuck Hong*, Sang-Hee Park

Division of Information & Communication, Baekseok University

빅데이터 환경에서 개인정보 익명화를 통한 보호 방안

홍성혁*, 박상희
백석대학교 정보통신학부

Abstract Big Data is strictly positioning one of method to deal with problems faced with mankind, not an icon of revolution in future anymore. Application of Big Data and protection of personal information have contradictoriness. When we weight more to usage of Big Data, someone's privacy is necessarily invaded. otherwise, we care more about keeping safe of individual information, only low-level of research using Big Data can be used to accomplish public purpose. In this study, we propose a method to anonymize Big Data collected in order to investigate the problems of personal information infringement and utilize Big Data and protect personal. This will solve the problem of personal information infringement as well as utilizing Big Data.

Key Words : Big Data, Protection System of Personal Information, Non-identified, Network, Protection

요 약 빅데이터는 이제 더 이상 미래 혁신의 아이콘이 아니라 인류가 당면한 과제를 해결하기 위한 하나의 수단으로써 공고히 자리매김해 가고 있다. 빅데이터의 활용과 개인정보 보호는 분명 양면성을 갖고 있다. 데이터의 활용을 강조할 경우 개인이 공개를 원하지 않는 사생활은 필연적으로 침해 될 것이고, 개인정보 보호를 강조할 경우 어설픈 수준의 빅데이터 연구만 가능해 공공의 목적을 달성 하는데 어려움을 겪을 수 있다. 본 연구에서는 개인정보 침해의 문제점을 알아보고 빅데이터의 활용과 개인정보의 보호를 하기 위해서 취합하는 빅데이터를 익명화하는 방안을 제시하였다. 이를 통해 빅데이터 활용 뿐만 아니라 개인정보 침해의 문제점을 해결할 수 있을 것으로 보인다.

주제어 : 빅데이터, 개인정보보호관리체계, 비식별화, 네트워크, 보호

1. Introduction

Big Data is no longer an icon of future innovation, it is firmly positioned as a means of solving the challenges facing mankind. It seems that the current trend is changing, not the discussion of 'need big data', but 'how do you use it to create high value?'. The utilize of Big Data and the protection of personal information have double-sidedness. If you emphasize

the utilize of data, privacy that an individual does not want to disclose will inevitably be violated, and if you emphasize personal information protection, you can only have difficulty in achieving public purposes because you can only study fragile Big Data[1]. The nation should therefore move forward in a policy manner, talking this into consideration.

The composition of this study describes the definition and trends of Big Data in Chapter 2, the

*This paper was supported by 2017 Baekseok University research fund.

*Corresponding Author : Sunghyuck Hong(sunghyuck.hong@gmail.com)

Received January 31, 2018

Accepted February 20, 2018

Revised February 09, 2018

Published February 28, 2018

problems of privacy infringement in Chapter 3, the protection measures of personal information in Chapter 4. Finally, Chapter 5 concludes the paper.

2. Definition and Trend of Big Data

2.1 Big Data Definition

Big Data is a series of course for storing, collection, extracting, analyzing, and commercializing large amounts of data. The historical data was to simply store or collect. Recent areas of data have been easily conveyed by finding valuable information in the data collected various digital devices and presenting it in an information graphic, and it covers all the business processes that sell information to the person or institution that wants it[2]. At the heart of Big Data is the business of data. It means to exchange data that flows like air or water into money. To create new business, you should pay attention to data as software, not hardware. Data that is scattered on the Web is a business that is created as new secondary data according to its own analysis rules and sells the processed secondary data to the persons concerned.

2.2 Big Data Trends

2.2.1 Overseas Trend

The UK think tank Economic Management Research Center outlook that the economic effect of the analytics business on Big Data over the next five years will amount to £216billion and create more than 58,000 new jobs. In Germany, the Big Data market is expected to grow by about 59% over the previous year's €3.9 billion, reaching a total of €6.1 billion in 2014, and about €13.6 billion in 2016, more than twice that of 2014.

2.2.2 Domestic Trend

I think Korea is the best country in the world about interest in Big Data. However, despite this interest, the domestic Big Data market seems not to be well formed. The current business is in the form of ordering projects

from large companies with large amounts of data, and the market is being formed by creating business strategies and delivering technology and solutions[3]. Also, In domestic, the number of specialized companies with specialized Big Data solutions is limited. Furthermore, the number of companies that have worked on the actual Big Data and have done business has very small. Except for businesses that target large companies with data, it is a pity that there are no successful cases or not shared enough.

3. Infringement problem of Personal Information

3.1 Type of Infringement in Big Data Environment

As the information is easy to find with Big Data, the risk of exposure to privacy is increasing. Among them, problems of personal data leakage of Big Data are getting more and more. Big Data is currently becoming the most important underlying technology. Among them, customer relationship management activities that perform marketing activities utilize customer data can analyze the data of customers as well as their behaviors so that they can reach the purpose through integrated information of consumers[4]. And, other affiliates also include affiliate marketing, which also utilizes marketing services and cost forecast data. As an individual, privacy is exposed in the blind spots of huge data such as consumer purchase history information, log analysis, and location based service. Especially, social network service is used as a repository for freely expressing individual daily information, and many people provide personal information such as their contact, residence, school, and dating status to individual SNS[5]. For example, when an assault by a schoolgirl is issued, people can easily search the portal or SNS search for personal information such as the student's name, photo, residence, school, etc. This is because unofficial organizations, called netizen, find out the personal

information described in the perpetrator's social network and distribute them. Personal information is exposed to everyone like this. People with malicious intent can easily retrieve data information records of others that they want only by searching. Therefore, the type of personal information infringement that can occur in the Big Data environment is divided into the Table 1, Table 2, Table 3, Table 4 as follows.

Table 1. Step of Collection

Type of Infringement	Threat and Cause of Infringement of Personal Data
Inappropriate way to collect information	-Collecting personal data from unrecognized software and etc. -Crawling contents including personal information.
Private monitoring without consent	-Persistingly collect and analysis of access data from private information donee. -Gathering personal location data without consent.
Collecting data unnecessarily	-Gathering individual data for reason of commercial purpose or managing convenience. -taking personal delicate information without right purpose.

Table 2. Step of Storage

Type of Infringement	Threat and Cause of Infringement of Personal Data
protected database/ System	-management oversight of database/system saving personal data.
exposure of private data caused by carelessness	-leaking personal information by hacking system or mistake. -leak and exposure of private information by authorization error.

Table 3. Step of Using and Supply

Type of Infringement	Threat and Cause of Infringement of Personal Data
unsuitable analysis	-analysing personal product purchase list for customized service without consent. -analysing one's moving path way for malicious purpose.
SPAM without agreement	-sending commercial advertisement or SPAM without informed consent. -receiving commercial SPAM mail, SMS text, telemarketing by providing information to third party without user agreement.

Table 4. Step of Destruction

Type of Infringement	Threat and Cause of Infringement of Personal Data
saving personal data after period of possession	-not destructing location data and individual data after period of possession. -negligence of hard disks including private data without formatting
illegally destructing private data	-arbitrarily delete personal information without authorization. -erasing data by mistake of administrator.

4. Privacy protection measures

4.1 Non-identification of personal information

As personal information is exposed, protection must be. Big Data contains a lot of information about individuals used by public companies such as banks. Therefore, there is a need for a way to protect who owns personal information for each individual. That is 'non-identification'. The non-identification combines multiple pieces of personal information into a new data set to make it impossible to identify the data. Types of data sets are divided into five categories as alias processing, total processing, data value deletion, categorization, and data masking[6].

However, non-identification techniques are not always safe. Because they can re-identify and find information. Non-identification methods always have the risk that there is a possibility of re-identification. The information that is non-identification to be re-identified is likely to lose its value as Big Data analysis and utilization information because of the large data loss[7]. Therefore, to protect personal information in Big Data, strong regulation and management of Non-identification strength is needed.

4.2 Non-identifying process processing structure

In the processing structure of the non-identification process, data is collected from the data subject to which the data refers. The collected personal data is combined into a data set containing personal information. And non-identification creates a new data

set so that it can not identify the data. This dataset can be used internally by the institution instead of using the original dataset to reduce the risk of personal privacy[8].

As shown in Fig. 1 below, the non-identification can be performed when the data collection step(flow(b)) or the identified data has been collected but the identification information is not actually needed(flow(a)). That is, it is not necessary to collect identifiers that are not needed for data management. Instead, non-discrimination can be applied after data conversion and before data storage to avoid obtaining identification information(flow(c)). If fully identified data is required by the institution, the identification information shall be deleted before the data is released to the data set for data use(flow(d)). This data set can then be provided to trusted data recipients. That is, to the recipient associated with additional, administrative controls such as data usage agreements. Alternatively, the data may be made publicly available to a number of unknown data recipients, such as releasing non-identified data to the Internet[9].

Applying a non-identified process based on the data lifecycle model can reduce the risk of personal privacy and facilitate the public disclosure process.

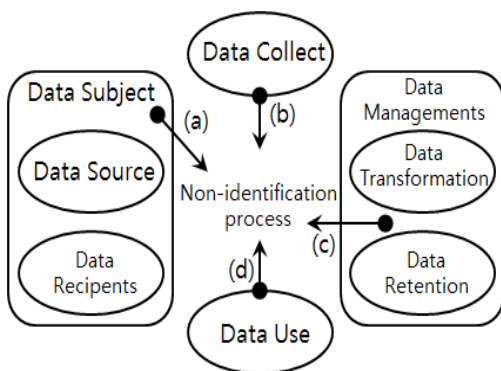


Fig. 1. Data Life Cycle in Data Circumstance

However, due to the interrelationships among people involved in the data flow, it is affected by when the non-identification process is to be performed. That is,

it can be performed before data collection(flow(a)), or after data collection(flow(b)), or before data storage(flow(c)), or before sharing data with the next participant(flow(d)).

4.3 Information Security Management System

Create an information protection policy for a specific organization. And, the certification body assesses and certifies whether the information protection management system that establishes, manages and operates the information management system continuously and systematically in order to protect the important information assets[10]. In other words, it is a comprehensive management system including management, technical, and physical protection measures to secure the stability and reliability of information and communication network. In the IT era, it aims to cope with the paradigm shift, the cyber infringement risk at all times, and organically manage various security measures.

4.4 Personal Information Protection Management System

To assess the degree of risk of all processes of collecting, using, providing, and destroying personal information for the purpose of achieving technological, managerial, physical protection measures and conformity of personal information, documenting the risk management procedures, And thus a comprehensive system that continuously operation and manages necessary countermeasures[11]. In the event of personal information leakage, the benefits of PIMS certification can be reduced to less than 50% of the damage caused by the victims, and the same benefits can be granted such as when acquiring the information protection management system certification.

4.5 Personal Information Protection Act

4.5.1 8 principles of OECD personal information protection

First, The principle of collection limitation : Personal information must be collected in a legitimate and fair manner and collected after information should be collected after consent is requested or communicated to the responsible parties.

Second, Principle of information accuracy : Personal information should be accurate, complete, and updated to the extent necessary for its intended use.

Third, Purpose Specification Principle : Personal information must be used in the collection process to specify the purpose of collection and to suit the stated purpose.

Fourth, The principle of use restriction : It can not be used or disclosed for any purpose other than the consent of the information entity or the provision of legal regulations.

Fifth, The principle of ensuring safety : Physical, organisational and technical safeguards should be secured to prevent the invasion, destruction, disclosure, or theft of personal information[12].

Sixth, Principle of disclosure : Under the management policy for the processing and protection of personal information, the identity and address information of the manager shall be disclosed.

Seventh, The principle of individual participation : The right to request access / correction / deletion of the personal information of the information subject should be guaranteed.

Eighth, Principle of accountability : The manager of personal information should be bound and accountable for adhering to principles.

4.5.2 US Information Protection Act

The privacy laws of the United States have individual laws for each field, but the privacy laws enacted as "watergate cases" are regarded as general laws of the public sector[13]. The private sector can be identified with the legal system of Korea prior to the enactment of the Personal Information Protection Act in March 2011 as a separate individual legislation.

4.5.3 Japan's Information Protection Act

Since the 1970s, Japan has been pushing forward the introduction of the legal system for the protection of personal information in order to meet the global trend. In response to the EU's request for strengthening the protection of personal information, the revision of the personal information protection legislation was enacted, and in May 2003, five laws relating to personal information protection were enacted and revised[14]. Here, there is the Privacy Act, which has the status of the basic law that applies to both the public sector and the private sector, and the status of the general law of the private sector. Under the Personal Information Protection Act, the public sector has an administrative agency privacy protection law.

4.5.4 Personal information regulation in domestic law

First, the Personal Information Act "Personal Information" is information about a living individual and information that can identify an individual through a name, resident registration number, and images.

Second, the "Personal Information" of Information and Communication Network Act refers to information such as codes, texts, voices, sounds, and images that can identify a specific individual by information such as name, resident registration number, etc.

Third, the location information law "personal location information" refers to location information of a specific individual.

Fourth, the "Credit Information" of the Credit Information Act refers to the information set forth in the Presidential Decree as information on the following items necessary for judging the credit of counterparties in transactions such as financial transactions and commercial [15].

5. Conclusion

Currently, there are no legal provisions in domestic that can provide clear protection. Therefore, individuals

should write personal information on the web or be cautious about their consent. In addition, similar personal information protection management system and information protection management system should be established, and actual companies should utilize to gain credibility of information. In order to establish such a personal information protection system, it is necessary to abolish the vague current standards for judging whether the government can cause defects, and to provide the company with benefits such as exemption from investigation of the personal information management. In addition, the agency for information human rights protection should actively engage in information security issues and carry out victim relief. Although there is a problem that the use of Big Data should be prevented from being overly focused on the protection of information in the Big Data environment, efforts should be made to continuously utilize the Big Data. Therefore, in order to reduce the risk of privacy due to leakage of personal information, it is necessary to anonymity data when it is necessary.

REFERENCES

- [1] M. Dave & J. Kamal. (2017). Identifying big data dimensions and structure. *2017 4th International Conference on Signal Processing, Computing and Control*. (pp. 163-168). Solan : IEEE.
DOI : 10.1109/ispcc.2017.8269669
- [2] D. Sik, K. Csorba & P. Ekler. (2017). Implementation of a geographic information system with big data environment on common data model. *2017 8th IEEE International Conference on Cognitive Infocommunications*. (pp. 181-184). Debrecen : IEEE.
DOI : 10.1109/coginfocom.2017.8268238
- [3] L. Mertz. (2018). Machine Learning Takes on Health Care: Leonard D'Avolios Cyft Employs Big Data to Benefit Patients and Providers. *IEEE Pulse*, 9(1), 10-11.
DOI : 10.1109/impul.2017.2772686
- [4] K. Dounya, K. Okba, S. Hamza, S. Safa, H. Iman, & B. Omar. (2017). A new approach based mobile agent system for ensuring secure big data transmission and storage. *2017 International Conference on Mathematics and Information Technology*. (pp. 196-200). Adrar : IEEE.
DOI : 10.1109/mathit.2017.8259716
- [5] A. Bagheri, M. H. Bollen, & I. Y. Gu. (2017). Big data from smart grids. *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe*. (pp. 1-5). Torino : IEEE.
DOI : 10.1109/isgteurope.2017.8260155
- [6] J. Ahmad, K. Muhammad, J. Lloret & S. W. Baik. (2018). Efficient Conversion of Deep Features to Compact Binary Codes using Fourier Decomposition for Multimedia Big Data. *IEEE Transactions on Industrial Informatics*, PP(99), 1-1.
DOI : 10.1109/tii.2018.2800163
- [7] J. Wu, M. Dong, K. Ota, J. Li & Z. Guan. (2018). Big Data Analysis-based Secure Cluster Management for Optimized Control Plane in Software-Defined Networks. *IEEE Transactions on Network and Service Management*, PP(99), 1-1.
DOI : 10.1109/tnsm.2018.2799000
- [8] N. Kumar, S. Antwal, G. Samarthiyam, & S. Jain. (2017). Genetic optimized data deduplication for distributed big data storage systems. *2017 4th International Conference on Signal Processing, Computing and Control*. (pp. 7-15). Solan : IEEE.
DOI : 10.1109/ispcc.2017.8269581
- [9] A. Prysyzhnyuk, R. Baevsky, A. Berseneva, A. Chernikova, E. Luchitskaya, V. Rusanov & C. McGregor. (2017). Big data analytics for enhanced clinical decision support systems during spaceflight. *2017 IEEE Life Sciences Conference*. (pp. 296-299). Sydney : IEEE.
DOI : 10.1109/lsc.2017.8268201
- [10] S. Long. (2017). Information Service Research and Development of Digital Library in the Era of Big Data. *2017 13th International Conference on Semantics, Knowledge and Grids*. (pp. 150-153). Beijing : IEEE.
DOI : 10.1109/skg.2017.00032
- [11] Z. Zhou, H. Yu, C. Xu, Y. Zhang, S. Mumtaz & J. Rodriguez. (2018). Dependable Content Distribution in D2D-Based Cooperative Vehicular Networks: A Big Data-Integrated Coalition Game Approach. *IEEE Transactions on Intelligent Transportation Systems*, PP(99), 1-12.
DOI : 10.1109/tits.2017.2771519
- [12] I. Notarnicola, Y. Sun, G. Scutari & G. Notarstefano. (2017). Distributed big-data optimization via block-iterative convexification and averaging. *2017 IEEE 56th Annual Conference on Decision and Control*.

- (pp. 2281-2288). Melbourne : IEEE.
DOI : 10.1109/cdc.2017.8263982
- [13] J. J. Harwood. (2016). Spectral ageing in the era of big data: integrated versus resolved models. *Monthly Notices of the Royal Astronomical Society*, 466(3), 2888-2894.
DOI : 10.1093/mnras/stw3318
- [14] K. B. Sindoori, L. Karthikeyan, S. Sivakumar, G. Abirami & R. B. Durai. (2017). Multiservice product comparison system with improved reliability in big data broadcasting. *2017 Third International Conference on Science Technology Engineering & Management*. (pp. 48-53). Chennai : IEEE.
DOI : 10.1109/iconstem.2017.8261256
- [15] P. Ezatpoor, J. Zhan, J. M. Wu & C. Chiu. (2018). Finding Top-k Dominance on Incomplete Big Data Using MapReduce Framework. *IEEE Access*, PP(99), 1-1.
DOI : 10.1109/access.2018.2797048

홍 성 혁(Hong, Sung hyuck)

[정회원]



- 2007년 8월 : Texas Tech University, Computer Science (공학박사)
- 2007년 9월 ~ 2012년 2월 : Texas Tech University, Office of International Affairs, Senior Programmer
- 2012년 3월 ~ 현재 : 백석대학교 정보통신학부 부교수
- 관심분야 : Network Security
- E-Mail : sunghyuck.hong@gmail.com

박 상 희(Park, Sang hee)

[학생회원]



- 2013년 3월 ~ 현재 : 백석대학교 정보통신학부 재학
- 관심분야 : Network Security, Hacking, Secure Sensor Networks
- E-Mail : nmas1994@gmail.com