A Discussion for Recognition about Big-data and Shared Personal Information

- Focusing on the twenties as a main user of SNS -

Junyong Jeon*, Jinho Yoo**

***Abstract*—**Big data can analyze not only structured data but also unstructured data to detect a pattern and find new value. These unstructured data is typically extracted from open data on the SNS, such as text, picture, etc. But there is some conflict between utilization of big data and protection of personal information in the use of these open data. In this study, we practiced targeted surveys to the twenties as a main user of SNS and tried to find relationship between recognition about utility of big data and frequency of SNS use and agreement about big data analytics based on open SNS data.

***Keywords*—**Big Data, Privacy, Shared Personal Information, SNS

I. Introduction

DUE to advances in computing and information technology, capacity of processing data with the computer and network has grown explosively than ever in the past.

Through these changes, The Big Data which beyond the dimensions of the traditional data analysis has emerged.

According to KISA Internet & Security Focus(2013), big data has features called '3V', which is Volume, Velocity, Variety and in recent years, it is called '4V' because 'Value' has added. Variety, one of the features of big data, means extended data type including structured and unstructured data and extended data collect pool including inside and outside. The result derived from these 3V features of big data can be referred as 'Value.

As we can see in those features, big data exceed the scope of conventional data analysis. In addition, the development of information and communication technologies were promoting information exchange among individual users, so internet users are now spontaneously producing and distributing massive structured and unstructured data.

Published and distributed data from individual users are not much valuable by itself as information. But if it is combined and analyzed through a big data technology, it could be a valuable information.

Social Networking Service (SNS) is typical data source for big data analysis, because SNS data is not limited to standardized and structured as same as for most of the existing data. Instead, SNS data has variety forms of data such as unstructured text, audio, picture, video, location information, logs.

However, unstructured data created from the SNS is easy to combine with the other information that can identify the specific individual and it is easy to collect. In other words, unstructured data created from the SNS has a feature as 'shared(disclosed) personal information.

Recently, the Korea Communications Commission announced a plan about 'Big Data Guidelines' which established for the purpose of preventing misuse of personal information in big data analysis and activating big data industry.

However, there are pros and cons for this 'Big Data Guidelines' and this makes some conflict. One of the biggest issues about this guideline is [article 3: collection of shared personal information]. It said that if telecommunication service provider wants to collect shared personal information, they don't have to get separate agreement from the subject of information.

Although the guideline obliged duty for notice and de-identification when telecommunication service provider collect and use shared personal information to assure Personal Information Control Right, Some of the NGOs are concerned that human rights abuse when shared personal information indiscriminately collected and analyzed because various personal data are uploaded on SNS, electronic bulletin board, blog, etc.

In this situation, these questions could be emerged : Is it possible that utilizing big-data without invasion of privacy? Is it possible to achieve preventing personal privacy invasion and activating big-data industry at the same time? Do the people will disagree to collect and analyze their shared personal information as a big-data when the legal and technical measures to protect personal information and privacy are assured?

In this situation, this paper focused on some points as follows : recognition rate in twenties as a SNS user about the fact that their SNS open data can be used as a source of big-data on the Internet and their opinion about accommodating these kind of collect and analysis. Especially, we have researched about expected big-data utility and frequency of using SNS.

Also, we tried to find about expectation and concerns from the twenties as SNS user in provide or refuse to provide shared personal information. In addition, we examined the necessity about technical and systematic approach to protect personal information and privacy in collecting and analyzing big-data through literature research. After that, we considered validity about using shared personal information in big-data era. Finally, this paper aims to derive implications for making policies related to vitalization of big data industry and protection of

*Junyong Jeon is in the master's course at Department of Information Security Management in Sangmyung University, Seoul, Korea(e-mail : goseonglove@gmail.com)

**Jinho Yoo is a professor at Department of Information Security Management in Sangmyung University, Seoul, Korea(e-mail : jhyoo@smu.ac.kr)

This research was supported by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the "Employment contract to support Master's Degree in Information Security" supervised by the KISA(Korea Internet Security Agency)

privacy.

To achieve these purposes of this study, we will analyze existing research in Chapter 2 and will describe about hypothesis and research methods in Chapter 3. In Chapter 4, we provide result of the survey and analyze this result to verify a hypothesis. Finally, we suggest implication of this study in Chapter 5.

II. Existing Research

In this paper, firstly we organize the existing studies about risk of unintended privacy invasion in big-data analysis especially on SNS. Then we find some implication which can be applied in this study.

Dae-Sun Choi (2013) proposed a technique for risk analysis in disclosed personal information on the online. He said that various personal information such as name; personal schedules, hobbies, etc. are disclosed on the online space like SNS. So this makes privacy risks to persons because it could be used for the source of data-analysis without de-identification.

Especially he emphasized the essentials that personal information and identity could be disclosed through the combination and inference of unstructured data while analyzing big-data.

Hwan-Su Lee(2013) researched about influences of overflow personal information on SNS to privacy risks and privacy concerns. He said SNS data which is voluntarily uploaded have a influence for awareness on privacy risks for the persons and this increase awareness level of the user's privacy concerns.

The above studies shows that using shared personal information on SNS and other online spaces for big-data analysis could make unintended personal information leaks and using shared personal information on SNS for big-data analysis could be a burden for users because of privacy problem.

The importance of legal measures to protect personal privacy when big-data management is studied by Jeong-Min Ahn(2013), in「A study about online personalized advertising in the privacy aspects」.

In this study, He insisted that online personalized advertising has a risk for infringement of personal information because it observes online usage patterns of persons for a long time and users don't know about this. But he also said if these kind of information like 'online usage patterns' are strictly interpreted as personal information, it is hard to make some innovation through big-data analysis.

These 'dilemma' is closely related to the debate around Big Data Guidelines. So it is needed to find the recognition of twenties about big-data and privacy because they are actively using the SNS.

III. Research Methods And Hypothesis

A. Research Methods

To achieve the goal of this study, we performed a survey to Internet users of twenties in early May 2014. The total number of samples is 137, research summary is as follows:

TABLE I
SURVEY OUTLINE

Survey Target : Internet users aged 19 - 29
Respondents : 133 people
Survey Method : offline survey using questionnaire
Survey Periods : 2014. 5. 14 - 15 (2 days)

The questionnaire was composed to nine questions, the main question is as follows:

1) Using SNS or not and frequency of use
2) Awareness of Big-Data
3) Expected Big-Data utility
4) Awareness of the fact that their SNS data could be used for big-data analysis as shared personal information
5) Agreement levels for using their SNS data as shared personal information to analyze big-data

TABLE II
CLASSIFICATION OF EXPECTED BIG-DATA UTILITY

Data Analytics Development of Statistics, Bio-informatics, Weather Research, Language Translation Technology
Healthcare Contribution for Medical Law, Improving therapeutic efficacy through data analysis(ex. epidemic path analysis)
Business Management Providing future prediction technique, business analytics including statistical and probability analysis
Marketing Providing personalized service through Internet usage pattern and location information analysis

B. Hypotheses

In this paper, we propose several hypotheses to analyze public awareness of sheared personal information and big-data.

[Hypothesis 1] Frequency of using SNS and expected utility of services that is provided through big-data analysis will have a positive correlation.

We assume that people who actively using SNS are much more exposed to the new technology or trend like big data because they can easily acquire and receive newest information through PC and Internet, or smart phones. In this respect, people who using SNS frequently could have more expectation about new services through new technology like big-data.

[Hypothesis 2] Expected utility for big-data analysis and intention to provide their SNS data for big-data analysis will have a positive correlation.

Having high level of intention to provide their SNS data for big-data analysis means that they already aware about SNS data as a 'shared personal information.' So they may expect high utility of big-data as a result of providing their open personal information.

[Hypothesis 3] If they already recognized the fact that their SNS data could be used for a data source for big-data analysis, it has a negative correlation between frequency of using SNS and intention to provide their SNS data for big-data analysis.

To the SNS user, the fact that their SNS data could be used

for a data source for big-data analysis is a burden because it has a risk for unintended personal information leaks. Therefore, It is expected that if someone use the SNS actively and has large amount of data in their SNS, they may have low-intention to provide their SNS data for big-data analysis because they have psychological burden for privacy risks. On the other hand, if someone use the SNS occasionally and has small amount of data in their SNS, they may have more intention to provide their SNS data for big-data analysis because they have relatively small burden for unintended personal information leaks.

IV. RESULTS ANALYSIS AND HYPOTHESIS TESTING

A. Summary of Survey

Summarizing the survey results are as follows.

TABLE III
DEMOGRAPHIC FACTOR(GENDER)

Male	Female	N/R	Total
82	50	1	133

TABLE IV
DEMOGRAPHIC FACTOR(AGE)

Distribution of Age	Respondents
Aged 19 ~ 20	29
Aged 21 ~ 22	44
Aged 23 ~ 24	34
Aged 25 ~ 26	19
Aged 27 ~ 28	4
Aged over 29	1
N/R	2
Total	133

TABLE V
RESULT OF SURVEY

Questionnaires	Results
Use or not to use SNS	82% of Respondents are using SNS
Frequency of uploading data to SNS	More than three times the monthly, Less than once a week(on average)
Recognition for Big-Data	45% of Respondents are recognized about big-data
The expected utility of Big Data	Average 7.6 (out of 10)
Recognition for the fact that shared personal information could be used for big-data analysis	45% of Respondents are recognized
Intention to provide shared personal information for big-data analysis	Average 6.3 (out of 10)

The results that analyzing the correlation for the hypothesis testing are as follows

TABLE VI
RESULT OF CORRELATION ANALYSIS

Hypothesis	Correlation
Frequency of SNS usage and Expected utility of big data	0.178
Expected utility of big data and Intention to provide shared personal information for big-data analysis	0.319
(If respondents are already aware about shared personal information on SNS as a big-data source) Frequency of SNS usage and Intention to provide shared personal information for big-data analysis	0.057

B. Hypothesis Testing

[Verification of Hypothesis 1] Frequency of SNS usage and Expected utility of services based on big-data have a very weak positive correlation.

The correlation between Frequency of SNS usage and Expected utility of services based on big-data is showed a very weak positive correlation, 0.178. Expected utility of services based on big-data was measured constantly high-level regardless of using SNS.

[Verification of Hypothesis 2] Expected utility of big data and Intention to provide shared personal information for big-data analysis have a positive correlation.

The correlation between Expected utility of big data and Intention to provide shared personal information for big-data analysis is showed a positive correlation, 0.319. This correlation is little weak, but significant. This suggests that there is a possibility that enhanced recognition of big-data utility could lead agreement to provide shared personal information for big-data analysis.

[Verification of Hypothesis 3] In hypothesis 3, correlation is measured 0.057. 0It seems to have no relation.

We assumed that if respondents are recognized about the fact that shared personal information could be a data source for big-data analysis, frequency of SNS usage and intention to provide shared personal information for big-data analysis has a negative correlation. But the result showed that there is barely any relation.

V. CONCLUSION AND IMPLICATIONS

The value of unstructured data has been revalued unlike the old days because application and analysis field based on unstructured data like SNS data are gradually widened in big-data era. SNS data, which is typical unstructured data source of big-data, is generally interpreted as personal information under existing legislation because it can identify a specific person through combination of private information. This situation can lead to high-level of regulation and this can be a limitation on activating big-data industry.

This study imply that emphasizing utility of big-data to the SNS users as unstructured data-producer can lead to agreement to provide their data for big-data analysis and this can be

activating big-data industry. This can be determined by the survey results. The group that has high-level of expected utility for big-data are generally answered positively about providing their SNS data in the survey.

Also, the group that answered a negative response to big-data analysis generally has distrust for a privacy protection in big-data analysis because of unintended privacy risks and they also concerned about identifying specific person through combination and analysis of SNS data.

Finally, this study suggests that enhanced awareness for the usefulness of big-data and reinforcement of trust for privacy protection in big-data analysis are needed to activate big-data industry.

This study includes three limitations. Firstly, survey target was limited to SNS user of twenties without other age group and people who produce data and information through other channels. Secondly, number of samples is absolutely insufficient. Thirdly, the analysis method that used in this study is limited to correlation analysis. Due to those limitations, this study has a meaning as an exploratory study.

In the future study, we have a plan to increase a number of samples and to elaborate a research paper and methods with regression analysis.

If these studies are progressed successfully, it will contribute to policy development related with activating big-data industry.

REFERENCES

[1] Jae-Sik Lee, "Technology for Personal Information Protection in Big-Data Environments" in Internet & Security Focus, 2013
[2] Yoon-Ki Kim, "Big-data usage, The Data Alchemy of Smart Era", unpublished, 2011
[3] Korea Internet and Security Agency, "2013 Internet Use Survey", 2013
[4] Jung-Min Ahn, "Customized online advertising and Privacy", in Cyber Communications, Volume 30, No. 4, 2013.12, p43-86
[5] Hwan-Su Lee, "Personal Information Overload and User Resistance in the Big Data Age", Intelligence and Information Systems, Chapter 19 No. 1, 2013.3, p125-139
[6] Dae-Sun Choi, "Privacy Risk Analysis Techniques in Big Data", in Information Security Journal No. 23, 2013.6, p56-60