Balancing Utility with Anonymity of Data

Latest Trends and Challenges in Anonymization Technology

That's **Collaboration** 2

Anonymization technology is essential for achieving protection of privacy when using personal data. How deeply has anonymization technology evolved? And what challenges have emerged for actual operation? Professor Ichiro Satoh at NII (the chief administrator of the Technology Review Working Group [hereinafter "Technology WG"] of the Study Group on Personal Data of the government's IT Strategic Headquarters) interviewed Katsumi Takahashi of NTT Secure Platform Laboratories, a member of the Technology WG, to hear about the latest trends in anonymization technology and the outlook and challenges for its practical use.

What is anonymization in the big data age?

Satoh First, could you tell me about "anonymization," which is the key for use and application of big data in the future?

Takahashi The basic definition of anonymization is the process of preventing individuals from being identified by deleting and/or changing information such as names, birth dates or addresses included in personal data. In other words, we call technologies and methods to increase the anonymity of data and the combination thereof anonymization. However, since we are in the age of big data and a great deal of information has been accumulated in the world, there are issues wherein individuals are identified by matching with other data even if only the name and address are deleted.

Anonymization in big data is requested under the current interpretation of the legal system, behind which are a variety of needs for free use of personal data by making them anonymous. In many cases, however, it is not easy to process personal data into non-personal data. Some may say that all we have to do is use anonymization technology, but they lack understanding of this technology. Naive anonymization such as deleting names in data is relatively easy and in the world this is what the term commonly refers to, but it is impossible to eliminate the risk of individuals being identified only with naive anonymization.

Satoh So you are saying that anonymization cannot be accomplished simply and uniformly?

Takahashi That's right. On the other hand, it is possible to process personal data into a state that has almost entirely eliminated individuality included in data. Statistical data is a good example. In this case, however, since the level of abstraction of data is too high, the data will often become unsuited for analysis. This is a dilemma between the anonymity and utility of personal data.

Technology that ensures privacy without destroying the utility of data

Satoh What should we aim for to overcome the dilemma?

Takahashi A theoretical goal is to leave only the necessary minimum amount of information according to the analytical purpose. In the Technology WG, we expressed as a basic concept of anonymization: "There is no such thing as versatile anonymization method. It is on a case-by-case basis according to the types, features and utilization purposes."

Under such circumstances, it will not necessarily be a fundamental issue whether or not information is personal. Rather, it will be important to adopt the best data processing method according to the nature of personal data and the purpose of the analysis. It also needs to be ensured that data will be processed safely, including the provision of institutional security on the user side by accurately expressing risk regarding privacy to the information providers.

Ichiro Satoh
Professor, Information Systems Architecture Science Research Division, NII
Professor, Department of Informatics, School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies

Katsumi Takahashi
Executive Research Scientist
General Manager of Information Security Project
NTT Secure Platform Laboratories,
Nippon Telegraph and Telephone Corporation

Figure 1. Example of k-Anonymization

Membership number	Birth date	Address	Age	Item purchased
1001	April 1, 1979	A-cho, Chuo-ku, Tokyo	34	Bread
1002	December 10, 1986	A-cho, Yokohama-shi, Kanagawa-ken	26	Comic book
1003	October 10, 1974	B-cho, Shibuya-ku, Tokyo	38	Ice
1004	May 5, 1991	B-cho, Kamakura-shi, Kanagawa-ken	22	Paperback book
1005	November 10, 2006	A-cho, Kawagoe-shi, Saitama-ken	17	Cola
1006	February 6, 1990	C-cho, Atsugi-shi, Kanagawa-ken	23	Timetable
1007	August 15, 2003	B-cho, Urawa-shi, Saitama-ken	19	Milk
1008	September 30, 2000	C-cho, Omiya-shi, Saitama-ken	9	Tea
1009	January 1, 1983	C-cho, Nerima-ku, Tokyo	30	Boxed lunch
1010	July 7, 1994	D-cho, Yono-shi, Saitama-ken	18	Water

State Fulfilling K-Anonymity (k=3)

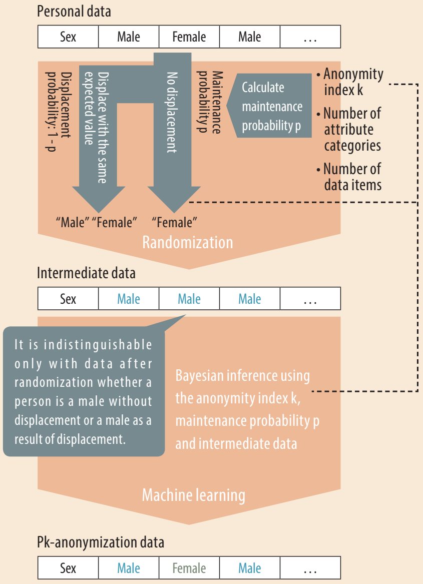
Membership number	Birth date	Address	Age	Item purchased	
1001	April 1, 1979	Tokyo	30s	Food	3
1003	October 10, 1974	Tokyo	30s	Food	
1009	January 1, 1983	Tokyo	30s	Food	
1002	December 10, 1986	Kanagawa	20s	Book	3
1004	Deletion 5, 1991	Kanagawa	20s	Book	
1006	February 6, 1990	Kanagawa	20s	Book	
1005	November 10, 2006	Saitama	Minor	Beverage	4
1007	August 15, 2003	Saitama	Minor	Beverage	
1008	September 30, 2000	Saitama	Minor	Beverage	
1010	July 7, 1994	Saitama	Minor	Beverage	

Figure 2. Conceptual Diagram of Pk-Anonymization

Pk-anonymization is the world's first randomization method with safety equivalent to k-anonymization.

Randomization: k-anonymity is achieved (mathematically guaranteed) only with stochastic displacement of data.

Machine learning: Correct data by estimating data suitable for analysis, using the parameter of randomization


Personal data
Sex
Male
Female
Male
…
Displacement probability: 1 - p
Displace with the same expected value
No displacement
Maintenance probability p
Calculate maintenance probability p
• Anonymity index k
• Number of attribute categories
• Number of data items
"Male" "Female"
"Female"
Randomization
Intermediate data
Sex
Male
Male
Male
…
It is indistinguishable only with data after randomization whether a person is a male without displacement or a male as a result of displacement.
Bayesian inference using the anonymity index k, maintenance probability p and intermediate data
Machine learning
Pk-anonymization data
Sex
Male
Female
Male
…

Satoh What methods do we have as technology?

Takahashi An example is that if we want to handle personal data as is for an analytical purpose, ciphers are useful. We have been studying and developing secret sharing and secure computation technologies to use for processing data while they are encrypted in order to achieve protection of privacy and safety management measures. This group of technologies is a tool to pursue the confidentiality of data to the utmost limit without impairing the data's accuracy.

On the other hand, if we place emphasis on anonymization, we also have a method called "k-anonymity," which is a parameter to express the risk regarding privacy. This is an indicator for evaluating data's anonymity, and means the state in which there are always k or more people with a similar attribute. For example, if there are at least 10 people included in the target attribute, whether they are in their 20s or 30s, the anonymity of the data is expressed as "k = 10." In other words, the larger the k value, the smaller the privacy risk. A technology that processes personal data to achieve k-anonymity is k-anonymization, which can be achieved through generalization that makes the value of an attribute rougher or there is deletion of data for a low number of attributes, so that there will be at least k records with the same attribute combination (Figure 1).

Satoh You are making more advanced efforts at NTT Secure Platform Laboratories.

Takahashi Our team has developed a method called Pk-anonymization. This makes data incomprehensible in terms of who they belong to through randomization[*1], which is processing to change individual data probabilistically. In randomization in this method, records are processed to be identified with a probability of 1/k or less. We call this nature Pk-anonymity (probabilistic k-anonymity). After that, we execute processing to estimate the original state of the data by using a machine learning method called Bayesian inference[*2] (Figure 2). By doing so, practical anonymous data for analysis will be constructed. We might say that this is pseudo-personal data based on actual personal data. We think that Pk-anonymization is effective for anonymization of big data while retaining a similar nature to k-anonymization.

Satoh Can we use Pk-anonymization for any data?

Takahashi It is particularly effective for anonymization of long personal data with a number of items. For long data, I believe that order-made anonymization, which conducts anonymization by selecting necessary items from data, is repeated in many cases, but the possibility of k-anonymity being damaged by checking with multiple anonymized data has been pointed out. To the contrary, since Pk-anonymization is resistant to the loss of k-anonymity, it is possible to conduct order-made anonymization repeatedly and process personal data with many items into highly valuable data for analysis while protecting privacy. These are the features of Pk-anonymization.

Protecting privacy both technically and institutionally

Satoh I think that k-anonymity will be greatly involved in a variety of scenarios in the future, associated with the revision of the Act on the Protection of Personal Information, etc. Do you think that k-anonymity will be one of the indicators for deciding to license the provision of data to a third party? And in that case, is it possible to say that provision of data is safe if the k-value is a certain number?

Takahashi That's a difficult question. There is still an issue of whether safety can really be judged only with the k value, and I think that there are some cases in which there is little risk even if k is one. I think that the adequacy of the k value will be determined in its social operation, including the development of the legal system in the future.

Satoh So you are saying that for the appropriate use and application of anonymized data in the future, it is necessary to promote it, with the technology and the institution working as a pair of wheels?

Takahashi For personal data, I think it will be enough to deal with it obeying the legal system. Also, we will be able to freely use data that are clearly separated from individuals. The future challenge is how to deal with "anonymized data" other than these two types. Protection of data privacy will be possible by establishing social rules for the use and provision of anonymized data. Data with a risk regarding privacy can be provided only to a person who can be trusted. In other words, I imagine that problems concerning anonymized data will be gradually solved by establishing and operating the legal system, such as permitting provision of data to a person who observes the rules, but declining permission or imposing punishment if the rules are not observed.

(Written by Hideki Itoh)

[*1] Randomization: An anonymization method to add noise data to the original data to the extent to which the addition will not have an impact on data analysis

[*2] Bayesian inference: A stochastic method to estimate the probability of an event that causes an observed event from the observed event