JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 6, pp. 11-21, June 2022
https://doi.org/10.9708/jksci.2022.27.06.011

A study on the method of measuring the usefulness of De-Identified Information using Personal Information

Dong-Hyun Kim*

*General Researcher, Korea Internet & Security Agency, Naju, Korea

[Abstract]

Although interest in de-identification measures for the safe use of personal information is growing at home and abroad, cases where de-identified information is re-identified through insufficient de-identification measures and inferences are occurring. In order to compensate for these problems and discover new technologies for de-identification measures, competitions to compete on the safety and usefulness of de-identified information are being held in Korea and Japan. This paper analyzes the safety and usefulness indicators used in these competitions, and proposes and verifies new indicators that can measure usefulness more efficiently. Although it was not possible to verify through a large population due to a significant shortage of experts in the fields of mathematics and statistics in the field of de-identification processing, very positive results could be derived for the necessity and validity of new indicators. In order to safely utilize the vast amount of public data in Korea as de-identified information, research on these usefulness metrics should be continuously conducted, and it is expected that more active research will proceed starting with this thesis.

▸ **Key words**: Big data, Personal Information, De-Identification, De-Identified Information usefulness measurement, De-Identified Information usefulness indicators

[요 약]

국내외에서 개인정보의 안전한 활용을 위한 비식별 조치에 대한 관심이 높아지고 있으나 불충분한 비식별 조치 및 추론 등을 통해 비식별 정보가 재식별되는 사례가 발생하고 있다. 이러한 문제점을 보완하고 비식별 조치 신기술을 발굴하기 위해 비식별 정보의 안전성과 유용성을 경진하는 대회를 국내와 일본에서 개최하고 있다. 본 논문은 이러한 경진대회에서 사용되고 있는 안전성과 유용성 지표를 분석하고 보다 효율적으로 유용성을 측정할 수 있는 새로운 지표를 제안하고 검증하고자 한다. 비식별 처리 분야에 수학 및 통계 분야의 전문가가 현저히 부족하여 많은 모집단을 통한 검증은 할 수는 없었지만 신규 지표에 대한 필요성과 타당성에 대해 매우 긍정적인 결과를 도출할 수 있었다. 우리나라의 방대한 공공데이터를 비식별 정보로 안전하게 활용하기 위해서는 이러한 유용성 측정 지표에 대한 연구가 꾸준히 진행되어야 하며, 본 논문을 시작으로 보다 활발한 연구가 진행되길 기대한다.

▸ **주제어**: 빅데이터, 개인정보, 비식별 조치, 비식별 정보 유용성 측정, 비식별 정보 유용성 지표

• First Author: Dong-Hyun Kim, Corresponding Author: Dong-Hyun Kim
*Dong-Hyun Kim (kdonghyun@kisa.or.kr), Korea Internet & Security Agency
• Received: 2022. 04. 26, Revised: 2022. 05. 27, Accepted: 2022. 06. 09.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

데이터 3법 시행 및 데이터경제활성화 정책에 따라 국내에서 개인정보를 비식별 조치하여 활용하는 다양한 신산업이 발굴되고 있다. 비식별 정보는 가명정보와 익명정보로 구분이 되는데 가명정보는 개인정보보호법 제2조에 따라 개인을 식별할 수 있는 요소를 삭제하거나 대체하는 방법 등을 통해 제한된 목적 범위 안에서 활용을 할 수 있도록 하고 있으며, 익명정보는 개인정보가 아닌 정보로써 가명처리 외에 추가적으로 특이치[1] 또는 추론[2] 등을 통한 재식별 가능성에 대해 프라이버시 보호모델 등을 적용하여 보다 강한 비식별 조치를 적용하여 목적 제한 없이 활용을 가능하게 하고 있다. 이러한 익명정보 처리에 대해서는 정부에서 제시하고 있는 명확한 절차는 없지만 `16년 마련된 '개인정보 비식별 조치 가이드라인[3]'에서는 해당 가이드라인에서 제시하는 절차를 적절히 적용하여 처리한 경우 EU GDPR에서 제시하는 익명정보[4]와 동일한 수준으로 법령해설을 제시하고 있다. 이렇게 비식별 정보가 다양하게 활용되고 있는 시점에서 비식별 정보에 대한 안전성과 유용성을 측정하여 검토하기 위한 절차는 아직까지 공식적인 가이드나 지침을 통해 안내하고 있는 사항은 없다. 다만, 일본의 경우 일본정보처리학회 컴퓨터보안연구그룹(Computer Security Research Group)이 주최하여 `15년부터 매년 비식별 정보의 유용성과 안전성을 경진하는 PWSCUP[5]을 개최해오고 있다. 해당 경진대회는 예선전을 통해 비식별 정보의 안전성과 유용성을 평가하고, 본선전에서는 예선전에서 참가자들이 제출한 비식별 정보에 대해 재식별 공격을 시도하여 재식별 가능성을 측정하고 순위를 결정하게 된다. 국내에서도 `18년부터 한국인터넷진흥원이 주최하여 '개인정보 비식별 기술 경진대회[6]'를 개최해오고 있다. 두 가지 대회는 안전성과 유용성을 평가하는 운영 방식은 유사하나 일본의 경우 평가를 측정하는 방식을 모두 정량화하여 지표로 계산하는 반면, 국내의 경우 안전성은 외부전문가들을 통해 정성적으로 평가하고, 유용성 부분만 정량적으로 측정하는 차이점이 있다. 또한 본선전으로 개최되는 재식별 시도 가능성의 경우 일본은 대회 참가자들이 제출한 정보를 이용하여 재식별을 시도하는 반면, 국내에서는 주최 측(한국인터넷진흥원)에서 생성하고 가공한 비식별 정보를 활용하여 재식별을 시도하는 차이점이 있다. 이러한 차이점이 발생하는 이유로는 국내의 경우 `17년 시민사회가 개인정보에 대한 정보주체의 동의 없는 활용을 이유로 `16년에 마련된 '개인정보 비식별 조치 가이드라인[3]'에 따라 빅데이터 서비스를 제공한 20개 기업과 비식별 조치 전문기관 4곳을 개인정보 보호법 위반으로 검찰에 고발[7]하면서 비식별 정보에 대한 안전성이 문제가 되었기 때문이다. 이와 같이 비식별 정보의 활용에 대한 안전성과 사회적 합의가 부족한 시점에서 대회 참가자들이 생성한 비식별 정보가 쉽게 재식별이 된다면 사회적 이슈가 더욱 커질 수밖에 없는 상황으로 대회 주최 측에서는 가능한 한 재식별이 안 되는 비식별 정보를 생성하여 제공을 하는 것이 바람직하다고 판단한 것으로 분석된다.

한편 개인정보의 비식별 조치에 있어 가장 중요한 목표는 안전한 활용이다. 즉 포함된 데이터가 개인정보인 만큼 안전하면서도 분석가들에 의해 쓸모 있게 활용할 수 있도록 유용한 데이터를 만드는 것이다. 사실 이 두 마리 토끼를 잡는 일은 쉬운 일이 아니다. 또한 이 과정에서 중요한 고려사항은 '데이터의 활용 목적이 무엇이냐?'이다. 활용 목적에 따라 비식별 조치의 대상과 처리를 적용하는 기준이 달라지기 때문이다[8]. 이에 본 논문에서는 비식별 정보에 대한 안전성과 유용성을 측정하는 방법을 국내외 비식별 경진대회를 통해 사례를 분석하고, 새로운 유용성 평가 지표를 제안하여 외부전문가를 통해 검증하고자 한다. 이를 위해 본 논문의 2장에서는 일본과 국내의 경진대회에 대해 살펴보고, 3장에서는 경진대회에서 활용한 지표에 대한 분석을, 4장에서는 새롭게 측정할 수 있는 유용성 지표를 제안 및 검증한 다음 5장을 끝으로 결론을 맺고자 한다.

II. De-Identification competitions

비식별 정보에 대한 안전성과 유용성을 측정하는 국내외 경진대회 사례를 살펴보면 일본과 일본의 사례를 벤치마킹하여 개최한 국내의 사례가 유일하다. 미국의 경우 `18년 민감한 개인정보를 활용할 수 있도록 기존의 비식별 조치 방법을 개선하고 제안하기 위해 NIST의 PSCR (Public Safety Communications Research Division)에서 Open Innovation Prize Challenge의 일환으로 비식별 정보를 활용한 공모전을 개최[9]하였으나 Unlinkable Challenge와 Differential Privacy Data Challenge 2개만 진행되고 이후 추진된 사례는 없다. 본 장에서는 일본과 국내 비식별 경진대회의 운영방식을 살펴보고 비식별 정보의 안전성과 유용성을 경진하는 방법에 대해 살펴보도록 한다.

1. PWSCUP in Japan

일본 PWS(Personal WorkShop) CUP[5]은 컴퓨터보안심포지엄(Computer Security Symposium)과 함께 연 1회 개최되며, 개인정보 보호기술의 연구개발을 추진하는 학술기관과 데이터분석 전문가 간의 교류를 적극적으로 지원하기 위해 추진하고 있다. `15년부터 현재까지 총 7회의 대회가 개최되었으며, 온라인상으로 비식별 대회 참가의 현황을 파악할 수 있는 현황판(Reader Board)과 정보의 Format Checker를 Docker 형태로 제공하여 전 세계 어디서나 참여할 수 있도록 구성하고 있다. 경진대회용으로 활용하는 원본 데이터셋은 `15년 일본 가정의 수입과 지출에 관한 국가통계센터(National Statistics Center)의 교육용 마이크로 데이터로, 이 데이터셋은 총 59,500개의 레코드와 197개(14개 준식별자와 183개의 민감 속성)의 컬럼으로 이뤄져 있다.

경진대회는 비식별 정보에 대한 안전성과 유용성을 평가하기 위한 경기로 예선전은 참가자들에게 원본 개인정보를 제공하여 비식별 조치(익명수준)를 수행하고 주최 측에 제공을 한다. 주최 측은 사전에 만들어 놓은 검증S/W를 이용하여 정량적인 측정을 하게 된다. 본선전의 경우 각 참가자가 생성한 비식별 정보를 다른 참가자에게 제공을 하여 재식별을 수행하는 절차로 예선전에서 유용성이 높은 비식별 정보를 생성하기 위해 안전성을 낮춘 경우, 즉 유용성 점수를 높게 받기 위해 비식별 조치 수준을 낮게 처리한 경우 원본정보를 보유하고 있는 상대팀에서 재식별이 쉽게 가능할 수 있다. 일본의 경우 본인 팀에서 제출한 비식별 정보가 상대팀을 통해 재식별될 경우 총점에서 감점을 받는 점수체계를 가지고 있기 때문에 예선전에서 유용성과 안전성을 잘 고려하여 비식별 조치를 하여야 한다.

2. De-Identification Competitions in Korea

국내의 비식별 경진대회[6]는 한국인터넷진흥원 주최로 `18년부터 매년 열리고 있으며 제1회 대회에서는 총 9개 팀 26명이 참가하여 그 중 6개 팀이 본선에 올라 모두 수상의 영예를 안았다. `19년 제2회 대회부터는 참가 접수를 한 팀이 많아져서 온라인 사전 테스트를 통해 20개 팀을 선발하였으며 매년 6~7개 팀이 수상을 받고 있다. 예선 진출자를 20개 팀으로 선발하는 이유로 국내는 일본과 달리 오프라인으로 대회를 진행하고 있어 장소와 시간적 제약이 있다는 점과 아직 평가체계가 온라인을 통해 정량적으로 측정하는 것이 아닌 외부전문가들을 통해 정성적으로 판단하는 점으로 분석되었다.

대회는 일본과 마찬가지로 예선과 본선으로 나뉘어 진행되며 예선에서는 각 참가자들이 주최 측으로부터 대회 문제인 원본 데이터셋 A를 받아 비식별 처리를 수행한 후 주최 측에 비식별 정보와 발표 자료를 USB에 저장하여 제출한다. 본선에서는 예선과 동일한 원본 데이터셋 A와 주최 측에서 사전에 마련한 A에 대한 비식별 정보 B`을 받아 B`정보 중 원본 데이터셋으로 예상되는 재식별 추정 행 번호 BE(일종의 인덱스로 B`의 각 레코드별 원본으로 추정되는 원본의 행 번호 리스트를 말함)와 발표 자료를 제출한다(Fig. 1 참조). 참고로 대회를 위한 재식별 환경은 우리가 실무에서 일어나는 상황과는 달리, 원본 데이터셋을 가지고 재식별을 시도하는 매우 강력한 공격의 유형에 해당한다. 실제 환경에서 대부분의 공격자들은 이러한 원본정보를 대개 보유하고 있지 못하기 때문에 현실적으로 재식별할 수 있는 가능성은 매우 적다고 볼 수 있다.

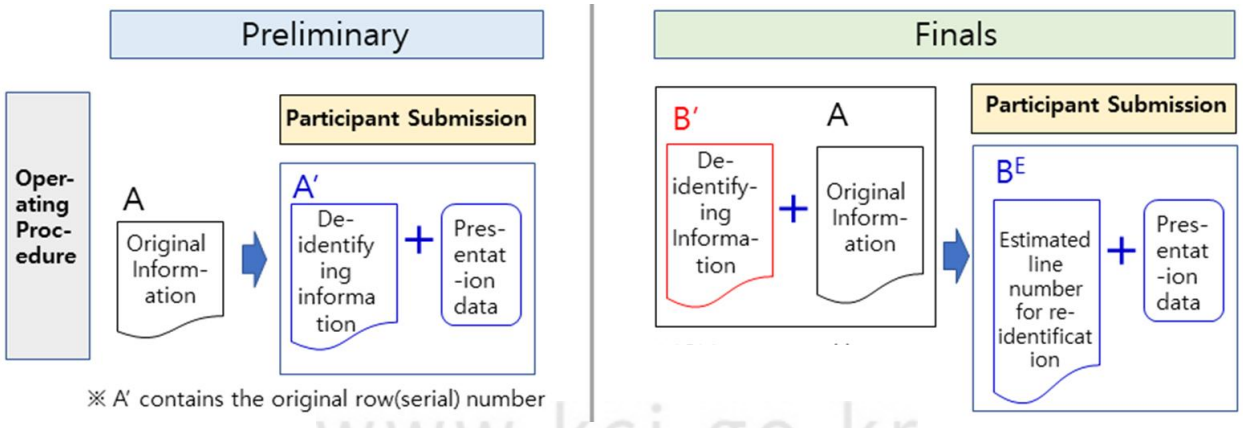

Fig. 1. Operation Procedure of De-identification Competition

3. Implications for domestic and overseas De-Identification competitions

일본과 국내에서 개최하는 비식별 기술 경진대회의 운영방식은 매우 유사하지만 다음과 같은 차이가 있다. 첫 번째로 일본은 정량적인 평가 측정방식을 통해 온라인으로 장소와 시간에 제약 없이 자유롭게 참가가 가능한 반면, 국내에서는 오프라인으로 진행되는 만큼 장소와 시간에 제약을 받게 되어 활성화가 어렵다는 문제가 있다. 두 번째로 두 대회에서 활용하고 있는 본선전의 데이터셋에 대한 차이가 있다. 일본의 경우 예선전을 통해 각 참가자들이 제출한 비식별 정보를 본인 팀 외의 다른 팀에게 제공하여 재식별을 시도함으로써 대회의 흥미를 높이고 경쟁을 유도하여 비식별 정보의 안전성과 유용성을 검토하고 있다. 그러나 국내의 경우 주최 측에서 생성한 비식별 정보를 활용하여 재식별을 시도하기 때문에 정답이 정해진 상태에서 참가자들이 얼마나 많이 찾아내느냐의 정도가 순위로 결정 되고 있다. 그리고 일본처럼 재식별 성공 시 상대팀에 감점요소가 적용되는 경쟁방식이 아니기 때문에 대회 자체의 흥미도도 떨어질 수 있다. 이와 같은 차이점을 정리하면 Table. 1과 같다.

Table 1. Differences between Japanese and Korean competitions

	Japan	Korea
Participation and operation method	On-line	Off-line
Data used in the finals	Information generated by each participant in the qualifier	provided by the organizer
Point deductions for re-identification	Yes	No

두 대회 모두 안전한 수준으로 비식별 정보를 생성하고, 생성한 정보의 유용성을 고려하는 경진대회를 개최하는 것만으로도 전문가가 매우 부족한 비식별 분야의 저변확대 및 기술개발에 큰 도움이 된다는 점에서 매우 우수한 대회라고 판단한다. 경진대회에서 활용한 지표에 대해서는 제 3 장을 통해 구체적으로 살펴보도록 한다.

III. Measurement Indicators of Safety and Usefulness of De-Identified Information

일본 PWSCUP[5]의 조직위원장인 일본 메이지대학 Kikuchi 교수 등이 발표한 논문[10][11]에 따르면 6가지의 유용성 측도들을 일본 경진대회 예선전에서 적용한 것으로 밝히고 있다. 국내에서는 경진대회에서 사용한 지표들에 대해 별도로 논문이나 연구결과로 공개한 사례는 없으나, 지표 개발 시 전문가로 구성된 자문회의를 통하여 확정하는 방식을 사용해 오고 있다. 본 장에서는 각국에서 활용한 안전성과 유용성 지표를 비교 분석하여 보완점을 도출하고 제 4 장을 통해 신규지표를 제시한 다음 이를 검증하고자 한다.

1. Safety indicators

일본과 국내 모두 안전성에 대한 지표는 `02년 미국 하버드대의 Sweeny 교수가 제안한 프라이버시 보호 모델인 k-익명성[12]의 최소와 평균 위험도로 각각 계산하고 있다. 국제표준 ISO/IEC 20889[13]에 따르면 프라이버시 보호 모델은 재식별(Re-Identification) 위험성의 계산을 가능하게 만들고, 경우에 따라 재식별 위험에 대해 수학적 보장을 제공하는 '데이터 비식별 조치 기술의 적용에 관한 접근법'으로 정의하고 있다. 그 중 k-익명성은 대표적인 모델로 다른 정보[1]와 연결 등을 통해 특정 개인을 식별할 수 있는 준식별자(Quasi-Identifier)에 적용할 수 있으며, 단일 또는 다중의 준식별자를 대상으로 동일한 값을 k값만큼 생성하도록 조치하는 것이다. 이러한 프라이버시 보호 모델은 비식별 조치를 수행하는 사람의 입장에서는 안전성에 대해 수치화 할 수 있다는 큰 장점이 있지만 반대로 비식별 정보를 분석하는 사람에 입장에서는 이러한 조치로 데이터의 품질이 훼손되기 때문에 큰 단점으로 작용한다. 즉, 프라이버시 보호 모델은 말 그대로 개인의 프라이버시를 보호하기 위한 안전성에서 바라보는 입장이다. 결과적으로 개인의 정보보호를 위해서는 어느 정도의 데이터 손실은 피할 수 없다는 것이다. 따라서 비식별 조치를 수행하는 사람의 입장에서는 데이터의 활용 목적을 고려하여 비식별 조치 시 데이터의 안전성과 유용성에 대한 균형을 맞추는 시각이 중요하며, 데이터의 손실을 최소화하면서 안전성을 유지하는 최선의 방법을 찾는 것이 중요하다.

1) 비식별 정보와의 결합 등을 통해 특정 개인을 알아볼 수 있도록 하는데 이용되는 정보

다음으로 활용되는 안전성에 대한 지표는 미국 HIPAA 프라이버시 규칙의 '전문가 결정 방식(expert determination)[14]'과 유사하다고 볼 수 있다. 현실적으로 비식별 조치된 정보의 안전성은 활용 목적과 상황에 따라 강도가 달라질 수 있으며 분야별 다양한 케이스별로 결과가 다르게 도출될 수 있다. 대개 미국, 캐나다 등 여러 나라들에서 전문가들이 결정하는 방식을 사용하고 있으며 국내 경진대회에서는 정량적 수치로 나타낼 수 있는 k-익명성과 더불어 외부전문가 평가도 안전성 지표로 함께 검토하고 있다. 외부전문가 평가는 국제표준 ISO/IEC 20889[13]에서 제시하고 있는 3가지(Single-out, 연결 가능성, 추론 가능성)의 재식별 공격 위험을 고려하여 안전성에 대한 평가를 수행한다.

2. Usefulness indicator

비식별 처리된 데이터셋의 유용성(Utility, or Usefulness)에 대한 평가는 일반적으로 데이터가 최초 의도한 용도를 기반으로 수행된다. 즉, 원본 데이터셋과 비식별 처리된 정보 간의 값의 차이가 적을수록 보다 많은 유용성이 보존될 수 있다. 그러나 동일 원본 데이터라 하더라도 목적이나 데이터 상황에 따라 용도가 다양할 수 있으며, 데이터를 공개 시마다 버전이 달라질 수 있다. 이러한 이유로 유용성을 대신하여 정보 손실(Information Loss)이라 부르기도 한다.

일본은 PWSCUP[5]에서는 Table. 2와 같은 6개의 지표를 유용성 지표로 제시하고 있다. U1은 원본 데이터셋(X)과 비식별 정보(Y)간의 특정 컬럼에 대한 속성 값 비교를 통해 비식별 정보의 유용성을 검증한다. U2는 U1과 유사하지만 전체 레코드가 아닌 무작위로 지정한 레코드를 대상으로 유사성을 비교한다. U3은 원본 데이터셋(X)과 비식별 정보(Y)간 같은 준식별자(QI)를 갖는 레코드 간의 개수 차이를 평균한 값이다. 예를 들어 원본 데이터셋에서 같은 준식별자를 갖는 2개의 레코드가 비식별 정보에서 1개의 레코드로 비식별 처리 되었을 경우 그 차이가 클수록 비식별 정보의 유용성은 떨어지게 된다. U4는 원본 데이터셋(X)과 비식별 정보(Y) 사이의 민감정보를 대상으로 피어슨의 상관계수를 적용하여 평균절대오차(MAE, Mean Absolute Error)를 판단한다. 이러한 상관관계를 통해 X와 Y간에 어느 정도 차이가 있는지를 검증한다. U5는 원본 데이터셋과 비식별 정보의 전체 레코드를 비교하여 어느 정도 데이터가 손실이 일어났는지 검증하며, U6은 Domingo- Ferrer[15]가 제안한 재식별 측정 도구를 그대로 사용한 것으로 원본 데이터셋(X)과 비식별 정보(Y) 간의 정형화된 차이 값들을 평균을 계산한다.

국내의 경우 `18년 첫 대회에서는 일본의 지표인 U1~U4와 신규지표인 Mean Distribution(Table. 3 UK4 참조)을 신규로 개발하여 시작하였으며, `19년부터는 국내에 적용하기 힘든 지표를 제외하고 신규지표를 추가로 개발하여 유용성 검증에 활용하였다. 일본 지표에서 삭제된 지표는 U2와 U5, U6로서 U2의 경우 U1지표와 측정방법에는 큰 차이가 없을뿐더러 레코드를 전체로 선택하는 것과 무작위로 선택하는 것에 대해 유용성에서 큰 차이를 발견하지 못했다. U5의 경우 원본 데이터셋과 비식별 정보의 레코드의 단순 차이보다는 k-익명성을 적용하는 경우 유용성 기준이 되는 동질 집합군[2]을 대상으로 그 차이를 비교하는 것으로 보완을 하였으며, U6의 경우 `19년 대회까지는 활용하였으나 U1과 U2의 지표와 유사하여 `20년부터는 삭제를 하였다. 위와 같은 사항을 반영하여 현재까지 유용성 평가에서 활용하고 있는 지표는 Table. 3과 같다.

Table 2. Six usefulness indicators used in the Japanese PWSCUP competition

#	name	meaning	object
U1	meanMAE	Difference in the mean absolute value of all sensitive attribute values in the original dataset and the de-identified dataset	sensitive attribute
U2	crossMean	Difference in the mean absolute value of the sensitive attribute values corresponding to several(designated) quasi-identifiers in the original dataset and the de-identified dataset	quasi-identifier
U3	crossCnt	Difference in the number of records for several(designated) quasi-identifiers in the original and de-identified dataset	quasi-identifier
U4	corMAE	Difference in the average absolute value of the Pearson's correlation coefficient for all sensitive pairs	sensitive attribute
U5	IL	Average of formalized difference values between the original dataset and the de-identified dataset, which are part of the re-identification measurement tool proposed by Domingo-Ferrer[15]	sensitive attribute and quasi-identifier
U6	nrow	Difference in the total number of records in the original and de-identified dataset	NA

2) EC(Equivalent Class)로 표현하며, k-익명성을 적용한 경우 준식별자가 동일하게 k개 있는 항목의 묶음[16]

Table 3. Seven usefulness indicators used in the Korea De-Identification competition

#	name	meaning	object	ranking	year
UK1	MA (Mean Attribute)	Difference in the mean absolute value of all sensitive attribute values in the original dataset and the de-identified dataset * Japan PWSCUP, U1: same as meanMAE	sensitive attributes	Ascending	2018, 2019
UK2	MC (Mean Correlation)	Difference in the average absolute value of the Pearson's correlation coefficient for all sensitive pairs * Japan PWSCUP, U4: same as corMAE	sensitive attributes	Descending	2018, 2019
UK3	CS(Cosine Similarity)	Cosine similarity[17] for the vector values Xi and Yi belonging to the specified numerical attribute i of the original dataset X and the de-identified dataset Y	sensitive attributes	Descending	2019
UK4	MD_ECM (Mean Distribution)	Divide each quasi-identifier(QI) for de-identified dataset by equivalent class(EC), calculate the variance of the specified sensitive attributes(SA) within each EC, and calculate their average and apply it to all QIs, calculate the overall average	sensitive attributes, quasi-identifier	Ascending	2018, 2019
UK5	IL (Information Loss)	Average of formalized difference values between the original dataset and the de-identified dataset, which are part of the re-identification measurement tool proposed by Domingo-Ferrer[15] * Japan PWSCUP, U5: same as IL	sensitive attributes, quasi-identifier	Ascending	2018, 2019
UK6	NA_ECSM (Normalized Average EC Size Metric)	The total number of records in an de-identified dataset divided by the total number of equivalent class(EC) sets, divided by the representative k value of k-anonymity[18]	sensitive attributes, quasi-identifier	Ascending	2020, 2021
UK7	NU_EM (Non-uniform Entropy Metric)	As a measure of information loss for k-anonymity, a representative privacy model, usefulness is measured and calculated using the Non-uniform Entropy method[19]	sensitive attributes, quasi-identifier	Ascending	2020, 2021

국내 지표의 경우 UK1, UK2지표는 일본의 지표를 그대로 인용하였다. UK3은 코사인유사도[17]로 원본 데이터셋과 비식별 정보의 동일한 속성집합 간의 벡터의 스칼라(Scalar)곱[3]과의 크기로 두 속성집합을 계산한다. 유사도는 -1에서 1까지의 값을 가지며, -1은 서로 완전히 반대되는 경우, 0은 서로 독립적인 경우, 1은 서로 완전히 같은 경우를 의미하며 최종적으로 1에 가까운 것을 유용성이 높은 것으로 판단 한다. UK4는 비식별 정보에 대한 동질 집합군에 대한 평균 분포도(분산)로 참가자들 간에 비교를 하여 측정을 하며, 값이 작을수록 유용도가 높다. UK6은 LeFevre 등[18]이 분별력 측도 방식의 대안으로 제안한 것으로 비식별 정보의 전체 레코드 수를 전체 동질 집합(EC)의 수로 나눈 다음, 이 값을 k-익명성의 대표 k값으로 나눈 값을 말한다. 결과 값이 작을수록 분모인 전체 동질 집합(EC)의 수가 많음을 의미하고 동일 레코드에 대비하여 동질 집합(EC)의 수가 많다는 것은 일반화가 상대적으로 적게 되어 손실이 적다는 것을 의미한다. 따라서 결과 값이 작을수록 손실이 적어 유용도가 높다는 것을 의미한다. UK7은 k-익명성의 정보손실 측도(Information Loss metrics)인 비균일 엔트로피(Non-uniform Entropy)[19] 방법을 이용하여 계산을 하며, 계산식은 Fig. 2와 같다.

$$-\sum_{i=1}^{n}\sum_{j=1}^{J}\log_2\left(\Pr\left(R_{ij}|R'_{ij}\right)\right)$$

Fig. 2. Non-uniform Entropy formula

예를 들어 a) 50/950, 남성/여성 분포의 데이터셋은 286의 엔트로피를 가지는 반면, b) 500/500, 남성/여성 분포의 데이터셋은 1,000의 엔트로피를 가진다. 이러한 계산법은 a) -(((log2(50/1000)*50) +(log2(950/1000)*950)) = 286.3, b) -(((log2(500/1000) *500)+(log2(500/1000)* 500)) = 1000으로 나타낼 수 있다. 따라서 첫째 데이터 세트의 정보 손실이 훨씬 낮으며, 이는 직관적으로 더 타당성을 가진다.

3. Implications for Safety and Usefulness Indicators

비식별 정보의 안전성의 경우 국내외에서 비식별 조치 수준을 계량적으로 측정할 수 있는 k-익명성이 일반적으로 사용된다고 볼 수 있다. 프라이버시 보호 모델은 k-익명성 외의 l-다양성과 t-근접성 등이 있지만 이러한 모델들을 적용할수록 비식별 정보에 대한 유용성은 떨어지게

3) 두 벡터의 크기와 이들 벡터들이 이루는 각의 코사인 값의 곱

된다. 그 외 차분프라이버시 모델인 DP(Differential Privacy)가 있지만 DP를 적용할 경우 적용한 수준에 대해 계량적으로 판단할 수 있는 기술이 아직 없어 명확한 지표로 활용되기 어려운 실정이다. 데이터는 데이터가 활용 되는 환경, 즉 데이터 분석가가 요구하는 목적을 잘 파악해서 비식별 조치를 수행해야 하기 때문에 이러한 유동적인 사항을 정량으로만 판단하기 어려우며 외부전문가들의 평가를 통해 판단하는 것도 적절한 방법 중 하나이다[20].

유용성의 경우 국내는 일본의 지표를 인용한 부분이 일부 있었지만 지표가 중복되는 부분과 유용성 판단이 명확하지 않은 지표를 제거 또는 보완을 하였으며, 안전성을 중점으로 보는 국내 환경에 맞춰 k-익명성을 적용한 새로운 지표를 개발하였다. 다만, 일본과 국내 모두 정형 데이터만을 대상으로 평가를 하고 있어 비정형 데이터(문자열 등)에 대한 유용성 평가방안도 필요한 상황이다. 또한 현재 지표들은 대부분 원본 데이터셋과의 차이점을 유용성의 지표로 제시하고 있는데 실무에서는 원본 데이터셋과의 비교가 불가능한 경우가 많기 때문에 비식별 정보 자체에 대한 유용성을 측정할 수 있는 지표가 필요하다. 이러한 사항을 보완하기 위해 제 4 장에서는 문자열에 대한 유용성을 평가할 수 있는 지표와 비식별 정보 자체의 거리 계산을 통해 유용성을 평가할 수 있는 새로운 지표를 제안하고자 한다.

IV. Proposal and validation of new usefulness indicators

1. Proposal of new usefulness indicators

먼저 문자열의 유사도를 판단하는 방법으로 러시아의 과학자 블라디미르 레벤슈테인이 고안한 알고리즘을 많이 사용한다. 레벤슈타인 알고리즘(Levenshtein algorithm)[21]은 두 가지 문자열을 삽입, 삭제, 변경을 하여 몇 번이나 해서 바꿀 수 있는지를 계산하여 그 최솟값을 구해 유사도 판단의 척도로 사용한다. 예를 들어 '1) ELEPHANT, 2) RELEVANT'를 비교를 하면 1회의 삽입이 일어났고, 1회의 변경, 1회의 삭제가 이뤄졌다. 이렇게 변경된 횟수를 계산하면 3만큼 서로 다르다는 결과를 Fig. 3처럼 나타낼 수 있다.

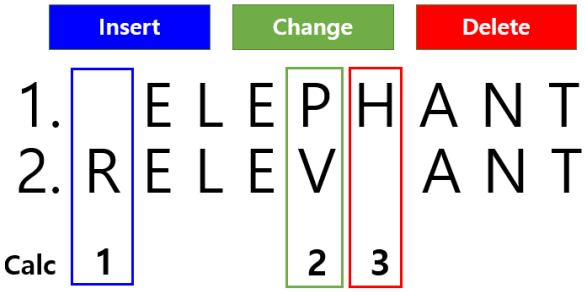

Fig. 3. Levenshtein algorithm measurement method

이러한 방식은 지하철역이나 병명, 주소 등과 같이 사용되는 문자열이 정의되어 있는 경우 유사도를 측정하여 원본 데이터셋과의 유용성을 평가 할 수 있다. 앞서 k-익명성에 대해 살펴보았지만 k-익명성은 데이터의 유용성을 매우 저하시키는 기술로 불특정 다수에게 공개하는 수준의 정보에 필수적으로 적용을 하게 된다. 그러나 실무에서 이러한 데이터는 분석에 효율성이 너무 떨어져서 불필요한 데이터(Garbage Data)로 인식이 되고 있다. 우리나라 정부의 '데이터 경제 활성화 정책[22]'에 따라 품질 높은 공공 데이터를 공개하기 위해서는 k-익명성이 아닌 다른 속성 값들을 원본과 유사하게 비식별 조치하여 처리할 수 있는 방안이 필요하다. 따라서 이러한 지표는 문자열 공개 시 안전성을 강화, 즉 재식별 가능성을 방지하고 유용성이 높은 비식별 정보를 검증할 수 있는 지표로 활용될 수 있다(Table. 4 NU1 참조).

다음으로는 군집화(Clustering) 알고리즘을 활용한 유사도 측정 방법이다. 군집화는 수치들의 군집을 판별하는데 동일한 군집이 소속된 수치가 서로 유사할수록, 즉 집단 내 동질성이 높을수록 선호하게 되고, 상이한 군집에 소속된 수치들은 서로 다를수록, 즉 집단 간 이질성이 높을수록 선호도가 높아진다. 이러한 내용은 Fig. 4와 같이 나타낼 수 있다.

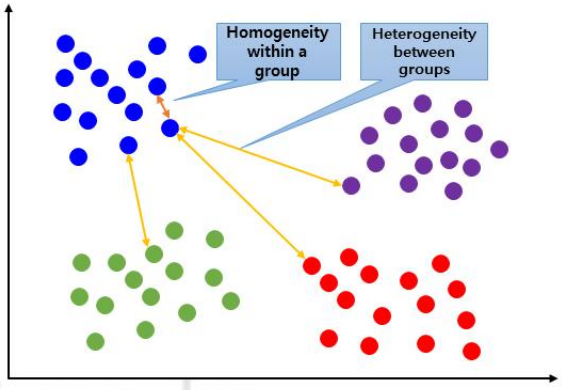

Fig. 4. Clustering algorithm measurement method

Table 4. Three usefulness indicators suggested

#	name	meaning	object	ranking
NU1	LD (Levenshtein Distance)	Use the Levenshtein algorithm for two strings to quantify the number of insertions, deletions, changes, etc. to determine their usefulness	character attributes	Descending
NU2	ED_SSE (Euclidian Distance_Sum of Squared Errors)	Using Euclidean distance, the difference in attribute values between the original dataset and de-identified dataset is calculated and determined as the standard deviation. * Comparison with de-identified dataset between participants	sensitive attributes	Closer to 0
NU3	AR(Anonymisation Ratio)	Determined by calculating the total number of records between the original dataset and de-identified dataset as a ratio * Comparison with de-identified dataset between participants	all record	Ascending

이와 같은 클러스터링 기법은 AI의 머신러닝 기법에서도 기계학습의 정확성을 향상시키기 위한 방법으로 많이 활용되며 이러한 유사도 측정을 위해 유클리디안 거리(Euclidean Distance)[4]를 이용하여 원본 데이터셋과 비식별 정보 간의 속성 값의 차를 표준편차로 계산하여 나타낼 수 있다. 또한, David Sanchez 등[23]은 이러한 계산법을 응용하여 원본 데이터셋과 비식별 정보 사이의 정보 손실을 각 레코드 단위에서 측정하는 새로운 측도를 제안하였다. 해당 유용성 측정 지표의 계산 방법으로는 원본 데이터셋의 속성 값과 이에 대응하는 비식별 정보의 속성 값들 간의 정규화 된 유클리디안 거리의 제곱에 대한 총합을 전체 속성 쌍의 수로 나눈 값이며, 값이 0에 가까울수록 원본 데이터셋과의 차이가 적어 유용도가 높아지고, 값의 차이가 클수록 원본과 비식별 데이터셋 간의 오차가 커서 유용도가 낮아진다는 것을 판단할 수 있다(Table. 4 NU2 참조). 마지막으로 데이터 분석을 위한 비식별 정보를 다양한 사람이 생성한 경우 처리수준의 비교를 통해 유용성을 측정하는 지표로써 원본 데이터셋 전체 레코드 수 대비 비식별 정보의 전체 레코드의 수를 비율로 계산하는 방법이다. 이 계산은 가장 높은 수치(100)에 가까울수록 유용성이 높은 비식별 정보를 생성한 것으로 판단할 수 있다(Table. 4 NU3 참조). 지금까지 새롭게 제안한 지표를 정리하면 Table. 4와 같다.

2. Validation of new usefulness indicators

앞 절에서 제안하였던 신규 유용성 지표에 대한 검증을 위해 `22. 2. 24.~`22. 3. 15.까지 개인정보 비식별 조치 관련 분야 경력 10년 이상의 학계 및 산업계 전문가를 대상으로 조사를 실시하였으며, 이 중 15인(데이터 통계분야 5인, 비식별 조치 전문가 8인, 그리고 개인정보 보호 전문가 2인)이 응답하였다. 설문조사의 모집단이 적은 이유로는 개인정보 비식별 조치 분야가 정보보호 시장에서 아직 연구를 시작하는 단계라 전문가가 많지 않으며, 본 연구 주제인 유용성 지표는 비식별 조치 분야의 지식 외에 수학과 통계 지식이 있어야 판단이 가능한 분야로 많은 전문가들로부터 응답이 어렵다는 의견을 청취하였다. 설문조사는 리커트 5점 척도를 사용하여 질의에 대한 응답점수가 높을수록 타당성이 높은 것으로 판단하였으며, 응답자의 모수가 현저히 적어 통계프로그램(SPSS)의 신뢰도 분석은 실시하지 않았다. 또한 3점 이하의 점수를 받은 경우 추가 인터뷰를 통해 낮은 점수를 배점한 이유와 보완점을 찾아내었다. 구체적인 조사 항목 및 결과는 Table. 5와 같다.

Table 5. Results of feasibility study of the proposed new indicator

#	Question	Ave Score
NU1	Is it reasonable to use a Levenshtein algorithm to measure the similarity of strings	3.5
NU2	Is it reasonable to use a clustering algorithm to measure the usefulness of de-identified dataset by calculating its own distance?	4.5
NU3	Is it reasonable to measure usefulness by utilizing the record deletion rate between the original dataset and the de-identification dataset?	5.0

전체적인 조사 결과는 평균 4.2점으로 이러한 지표를 신규로 개발하는 시도에 긍정적인 의견이 있었으며, 비식별 정보에 대해 계량적으로 유용성을 측정하는 방법에 대한 신규 지표에 대해서도 모두 만족을 하였다. 다만, NU1의 지표의 경우 유용성을 측정하는 지표라기보다는 비식별 정보의 재식별을 방지하기 위한 지표로 판단해야 할 것이라는 의견이 있어 낮은 점수를 받은 것으로 분석되었으며, NU2의 경우 클러스터링 기법을 사용하여 유사한 비식별 정보의 집합군을 만듦으로서 분석 시 보다 유용한 정보

4) 특정 값들의 거리를 구하는 알고리즘으로 유클리디안 유사도(Euclidean Similarity)라고도 함

Table 6. Complements final indicators based on validation results

#	name	meaning	object	ranking
NU1	LD	* This indicator is used to measure the possibility of re-identification of de-identified dataset	String attributes	Ascending
NU2	ED_SSE	same as Table. 4 NU2	sensitive attributes	Closer to 0
NU3	AR	same as Table. 4 NU3	all record	Ascending
NU4	NU4 (CAIL)	Based on k-anonymity, an algorithm called UBDSA (Utility Based Approach for Data Stream Anonymization) is used, and the information loss rate is measured using a new distance measurement method called CAIL (Cardinality Aware Information Loss)[24]	all attributes	Descending

를 생성할 수 있을 것이라는 의견이 있었다. 또한, NU1의 경우 레벤슈타인 알고리즘[21] 외 문자열이나 숫자 데이터를 범주화를 하여 계층화 한 경우 거리 계산을 통해 측정이 가능한 다른 방법에 대한 의견을 받을 수 있었다. 그에 따라 NU1은 유용성보다는 재식별 가능성을 측정하는 지표로 변경을 하고 Table. 6의 NU4와 같은 새로운 지표를 추가하였다. NU4 지표는 Ugur & Osman[24]이 제안한 CAIL(Cardinality Aware Information Loss, Cluster Assignment Distance Metric)이라는 새로운 클러스터 할당 거리 매트릭을 이용하여 거리를 계산하는데, 클러스터링 기법을 사용한다는 점에서 NU2 지표와 유사하지만 데이터를 범주화 후 계층화하여 거리를 계산하는 방식으로 숫자 데이터만이 아닌 문자열 데이터에도 적용이 가능한 차이가 있다. 설문조사에 따른 보완 결과 및 최종 제안 지표는 Table. 6과 같다.

이와 같은 신규 지표를 국내에서 개최하는 비식별 경진대회에 적용하기 위해서는 먼저 문자열 데이터를 비식별 처리 대상으로 포함하는 문제(시나리오)가 제출되어야 하며, 경진대회 사전 설명회 등을 통해 비정형 데이터를 처리할 수 있는 알고리즘에 대한 사전 교육도 필요하다. 경진대회 개최 결과물에 대한 유용성 평가 지표 반영 비율은 Table. 7과 같다.

Table 7. Contest score reflection ratio reflecting new indicators

#	name	ranking	Rate
UK1	MA	Ascending	5
UK2	MC	Descending	5
UK3	CS	Descending	10
UK4	MD_ECM	Ascending	10
UK5	IL	Ascending	5
UK6	NA_ECSM	Ascending	15
UK7	NU_EM	Ascending	15
NU1	LD	Ascending	10
NU2	ED_SSE	Closer to 0	10
NU3	AR	Ascending	5
NU4	NU4	Descending	10
Total			100

반영 비율이 낮은 지표들은 비식별 조치 가이드라인[3]에서 제시하는 기술 및 k-익명성[12]을 단순히 적용한 경우이며, 10점 이상의 점수는 클러스터링 기법 및 거리 계산, 그 외의 수학적 계산법이 어려운 지표를 대상으로 적용할 수 있다. 경진대회의 채점기준은 안전성 50%, 유용성 50% 중 외부 전문가의 발표 평가를 25%, 정량으로 측정할 수 있는 유용성 지표를 25%만 반영하고 있기 때문에 만약 기존 및 신규 유용성 지표로 계량적으로 측정할 수 없는 신기술을 적용한 경우 외부 전문가들을 통한 발표 평가를 통해 정량으로 측정해서 낮게 나온 점수를 높은 점수로 반영할 수도 있다.

V. Conclusion and future research

다양한 개인정보가 빅데이터로 활용되며 비식별 처리에 대한 필요성이 부각되고 있지만 충분하지 않은 비식별 조치로 인한 재식별로 특정 개인이 식별되어 사생활 침해로 이어지는 사례가 발생하고 있다. 빅데이터로서의 분석 효율성을 높이기 위해 비식별 정보의 유용성을 보존하더라도 재식별이 발생하지 않도록 충분한 안전성 확보 조치를 적용하는 것이 필요하며, 비식별 정보가 재식별되지 않기 위해 고려하여야 하는 사항은 다음과 같다. 첫 번째, 비식별 정보의 분석목적을 최소화한다. 비식별 정보의 분석 목적이 다양할 경우 원본 데이터셋에서 많은 항목을 추출해서 비식별 조치를 해야 하며, 분석에 필요한 항목은 유용성을 보존하기 위해 비식별 조치를 약하게 적용을 하여야 한다. 이러한 경우 다른 정보와 결합 등을 통해 재식별이 될 수 있기 때문에 주의하여야 한다. 두 번째, 암호화 등과 같은 가명처리를 적용할 경우 추가정보에 대한 엄격한 관리가 필요하며 낮은 보안강도의 알고리즘을 사용하지 말아야 한다. 한국인터넷진흥원에서는 보안강도별 권고 암호 알고리즘을 안내서를 통해 지침을 제공[25]하고 있다. 세 번째로 특이치(Outlier)에 대한 추가 검토이다. 공인 또는

특정 집단 내에 희귀(질환) 정보, 초고소득, 낙인성 정보는 그 정보 하나 만으로도 특정 개인을 식별해 낼 수 있다. 특이치는 인적 또는 자연적으로 발생 할 수 있는 만큼 비식별 조치 시 추가적으로 검토를 수행하여야 한다. 마지막으로 본 연구의 주제인 비식별 정보의 안전성과 유용성을 계량적으로 측정할 수 있는 지표를 활용하여 기관, 기업 내의 허용 가능한 기준을 마련하고, 비식별 정보의 안전성과 유용성을 검증하여야 한다. `20년 IDC[29]에서는 `25년까지 우리가 접하는 데이터의 80%가 텍스트, 이미지 등의 비정형 데이터가 될 것으로 전망하였다. 따라서 본 논문에서 제안하는 텍스트 및 클러스터링 기법의 거리 계산이 비식별 정보의 유용성을 측정하는데 활용이 된다면 큰 도움이 될 것이라 생각한다. 또한, 비식별 정보에 대한 유용성을 판단하는 지표의 경우 수학 또는 통계분야의 전문지식이 필요하기 때문에 향후에 관련 분야의 전문가들이 더욱 전문인재로 양성되어야 할 것이다. 그나마 국내에서도 비식별 정보의 유용성 강화를 위해 동형암호[26], 차분프라이버시[27] 및 합성데이터[28] 등에 대한 연구가 활발히 진행되고 있지만 유용성에 대한 검증방법은 전무한 실정이다. 다만, 새로운 비식별 처리 기법들을 통해 비식별 정보의 안전성은 강화할 수 있을 것으로 판단한다.

`20년 8월, 개인정보 보호법 개정과 함께 가명정보를 활용하는 다양한 산업이 등장하고 있지만 가명정보는 개인정보로써 활용 목적 및 공개할 수 있는 범위가 제한되어 있다. 방대한 공공데이터 등을 빅데이터 활용하기 위해서는 이러한 유용성 측정 지표의 꾸준한 연구를 통해 목적 제한이 없고, 유용성이 높은 익명정보를 생성할 수 있도록 추가 연구가 진행되어야 할 것이다. 본 논문을 통해 비식별 정보의 안전성과 유용성 지표에 대한 연구가 지속적으로 이뤄지길 바라며, 후속연구로는 본 논문에서 제안한 지표를 활용하여 실증을 통해 구체적으로 검증할 수 있는 방안과 활용 사례에 대한 연구를 수행할 예정이다.

REFERENCES

[1] Sunil Ray, "A Comprehensive Guide to Data Exploration," Analytics Vidhya, pp. 9-13, 2016.

[2] Dschoi, Shkim et al, "Big Data Privacy Risk Analysis Technology," Journal of The Korea Institute of Information Security and Cryptology, Vol. 23, No. 3, pp. 56-60, Jun. 2013.

[3] Joint Government Departments in Korea, "Guide lines for de-identification of personal information," 2016.

[4] EU General Data Protection Regulation, "Recital(26)," 2018.

[5] O. Hidenobu, M. Kunio, "A Study for the practical implementation of the evaluation of utility and security, through the data anonymization and re-identification competition," Information Processing Society of Japan Technical Report, 2016.

[6] KISA, http://datachallenge.kr/challenge/anon-con/

[7] Sgkim, "The Mediating Effect and Moderating Effect of Pseudonymized Information Combination in the Relationship Between Regulation Factors of Personal Information and Big Data Utilization," Informatization Policy, Vol. 27, No. 3, pp. 082-111, Aug. 2020.

[8] Dhkim, Sskim, "A New Scheme for Risk Assessment Based on Data Context for De- Identification of Personal Information", Journal of The Korea Institute of Information Security and Cryptology, Vol. 30, No. 4, pp. 719-734, Jun. 2020.

[9] R. Diane, F. Mary and W. Terese, "Challenge Design and Lessons Learned from the 2018 Differential Privacy Challenges," NIST Technical Note 2151, 2018.

[10] K. Hiroaki, Y. Takayasu et al, "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization," IEEE 30th International Conference on Advanced Information Networking and Applications, pp. 1035-1042, Montana, Switzerland, Mar. 2016.

[11] K. Hiroaki, K. Hamada et al, "Study on Record Linkage of Anonymizied Data," IEICE Trans. FUNDAMENTALS, Vol. E101-A, No. 1, pp. 19-28, Jan. 2018.

[12] Sweeney L, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Vol. 10, No. 3, pp. 557-570, July. 2002.

[13] ISO/IEC 20889, "Privacy enhancing data deident ification terminology and classification of techniques, Annex A," 2018.

[14] U.S. Department of Health & Human Services, "The HIPAA Privacy Rule," 1996.

[15] J. Domingo-Ferrer, V. Torra, "A Quant itative comparison of disclosure control methods for microdata," Confidentiality, Disclosure and Data Access, pp. 111-133, 2001.

[16] N. Guo, M. Yang et al, "Data Anonymization Based on Natural Equivalent Class," IEEE 23rd International Conference on Computer Supported Cooperative Work in Design, Porto, Portugal, May. 2019.

[17] I. Leontiadis, M. Onen et al, "Privacy preserving similarity detection for data analysis," 2013 International Conference on Cloud and Green Computing, pp. 547-552, 2013.

[18] K. LeFevre, D. DeWitt et al, "Mondrian multidimensional k-anonymity," 22nd International Conference on Data Engineering, Atlanta, USA, April. 2006.

[19] K. Emam, F. Dankar et al, "A Globally Optimal k-Anonymity Method for the De-Identification of Health Data," Journal of the American Medical Informatics Association, Vol. 16, No. 5, pp. 670-682, Jun. 2009.

[20] Personal Information Protection Commission, "Guidelines for processing pseudonym information," 2021.
[21] A. Narayanan, V. Shmatikov, "Robust de-anonymiszation of large sparse datasets," 2008 IEEE Symposium on Security and Privacy, Oakland, USA, May. 2008.
[22] Joint Government Departments in Korea, "Plans to revitalize the data and AI economy," 2019.
[23] S. Martinez, J. Domingo-Ferrer et al, "Supplementary materials for How to avoid reidentification with proper anonymization," Science, Vol. 351, No. 6279, pp. 1274, Nov. 2015.
[24] S. Ugur, A. Osman, "A utility based approach for data stream anonymization," Journal of Intelligent Information Systems, Vol. 54, pp. 605-631, Oct. 2019.
[25] KISA, "Cryptographic Algorithms and Key Length User Guide," 2018.
[26] JhCheon, YhEuh et al, "Privacy-Preserving Finance Data Analysis Based on Homomorphic Encryption," Financial Information Society of Korea, Vol. 7, No. 1, pp. 33-60, Feb. 2018.
[27] Hikim, ChPark et al, "A Study on a Differentially Private Model for Financial Data," Journal of The Korea Institute of Information Security & Cryptology, Vol. 27, No. 6, pp. 1519-1534, Dec. 2017.
[28] Ksjung, Spark, "Differentially Private Synthetic Data Generation Technique with k-anonymity," Journal of Computing Science and Engineering Congress 2018, Jeju, Korea, Jun. 2018.
[29] IBM, "IDC stacks up top object storage vendors," 2020.

Authors

Dong-Hyun Kim received Ph.D degree in convergence security from Chung-Ang University. He has been conducting personal information surveys and policy improvement for 6 years from 2010, and since 2016, the Data Utilization Support Team has been working to utilize safe personal information as big data. He is interested in Personal Information Security, De-Identification & De-Identified Information Risk Measure.