빅데이타 비식별화 기술과 이슈

우성희

한국교통대학교

De-identification Techniques for Big Data and Issues

SungHee Woo

Korea National University of Transportation

E-mail : shwoo@ut.ac.kr

요 약

최근 스마트폰, SNS, 사물인터넷이 확산되면서 생겨나는 빅데이타의 처리와 활용이 ICT 분야의 새로운 성장 동력으로 부상하고 있다. 하지만 이러한 빅데이터의 활용을 위해서는 개인정보 비식별화가 이루어져야한다. 비식별화는 개인의 데이터가 특정인과 연결되지 않도록 데이터 셋으로부터 식별정보를 제거하는 것으로 정보를 수집, 처리, 보관 혹은 배포하는데 있어 발생할 수 있는 개인정보노출의 위험을 줄이며 그 정보를 활용하고 공유하는데 그 목적을 두고 있다. 비식별화된 정보는 또한 재식별화되어 개인정보보호의 논란이 되고 있지만 빅데이터등의 개인정보가 비식별 처리되어 활용되는 사례는 점차 증가하고 있다. 또한 많은 비식별화 가이드라인의 등장과 함께 개인정보 비식별화 방법이 제시되고 있다. 따라서 본 연구에서는 빅데이타 비식별화 과정과 사후관리를 서술, 비식별화 방법을 비교분석하고 비식별화와 개인정보보호 관련 이슈와 해결과제를 제시한다.

ABSTRACT

Recently, the processing and utilization of big data, which is generated by the spread of smartphone, SNS, and the internet of things, is emerging as a new growth engine of ICT field. However, in order to utilize such big data, De-identification of personal information should be done. De-identification removes identifying information from a data set so that individual data cannot be linked with specific individuals. De-identification can reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing information, thus it attempts to balance the contradictory goals of using and sharing personal information while protecting privacy. De-identified information has also been re-identified and has been controversial for the protection of personal information, but the number of instances where personal information such as big data is de-identified and processed is increasing. In addition, many de-identification guidelines have been introduced and a method for de-identification of personal information has been proposed. Therefore, in this study, we describe the big data de-identification process and follow-up management, and then compare and analyze de-identification methods. Finally we provide personal information protection issues and solutions.

키워드

비식별화, 정보보호, 재식별화, 빅데이터

Ⅰ. 서 론

빅데이터는 IT 환경의 새로운 화두로 부상하고 있으며 새로운 마케팅 시장을 형성하는 원동력이 되고 있다. 하지만 개인 정보 해킹등으로 개인의 주민등록번호, 카드번호, 프라이버시등 민감정보가 유출될 경우 국가와 기업의 신뢰가 낮아질 뿐 아니라 더 심각한 위험에 노출 될 수 있어 개인정보를 활용 시에는 개인의 정보는 감추고 나머지 기타 정보를 사용해야 한다. 즉 개인 정보 비식별화가 이루어진 후에야 데이터의 활용이 가능하다는 것이다[1].

전 세계 기업의 30%가 이미 교통문제 해결을 위한 미래예측, 재난 및 안전관리, 신규세원 발굴, 세금징수 및 복지서비스를 위해 빅데이터를 구현하고 활용하는 추세이며 2013년 20%였던 빅데이터의 도입은 2015년 3월 기준 30%로 증가하고 있다. 국내에서도 고객관리, 재난공공분야, 제조분야, 보건의료분야, e-Business에 이르기 까지 외국과 비교하여 다양성과 사업 단계 측면에서 유사한 수준으로 활성화되고 있으며 국내 시장 경우 2017년 4,260억원(연평균 26.9%)으로 성장될 예정이다. 따라서 비식별된 개인정보는 제약사항이 적어 처리가 용이하기 때문에 활용되는 사례가 증가하고 있으며 이와 더불어 각종 가이드라인과 개인정보를 비식별 방안을 제시하고 있다. 하지만 이러한 가이드라인의 실효성과 가이드라인을 준수하더라도 개인정보보호 관련법령에 부합하게 되는 것인지 확신을 주지 못하고 있다. 미국, 영국[2]등 선도국에서도 개인정보를 보호하면서 빅데이터 활용을 위하여 개인정보 비식별화 활용을 권고하고 있다. 이에 우리는 빅데이터 활용 노하우가 부족한 상황으로 비식별화를 기반으로 한 빅데이터 활용과 도입 확산이 필요한 시점이라 할 수 있다. 따라서 본 연구에서는 빅데이타 비식별화 과정과 사후관리를 서술, 비식별화 방법을 비교분석하고 비식별화와 개인정보보호 관련 이슈와 해결과제를 제시한다.

II. 빅데이터 비식별화 프로세스와 사후관리

비식별화는 그림 1과 같이 수집 또는 사용, 저장, 공유되는 데이터로부터 개인을 식별하지 못하게 조치하는 일련의 방법으로 데이터를 어떤 개인과도 연결시킬 수 없도록 만드는 것으로 비식별화 정도가 높을수록 데이터의 유용성은 떨어지지만 프라이버시 침해 위험도는 낮아진다[3].

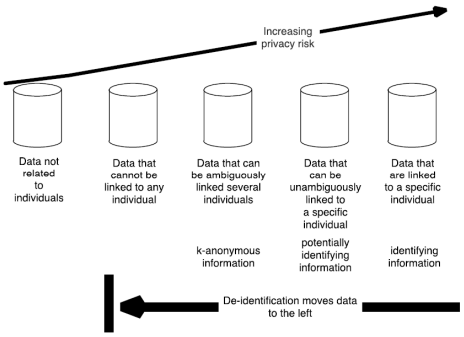

그림 1. 비식별화와 개인정보보호

2014년 12월 방송 통신 위원회는 빅데이타의 현실을 반영하여 수집되는 정보로부터 개인을 식별할수 없도록 조치를 취할 경우 기업들이 활용을 할수 있도록 빅데이타 가이드라인을 발표하였다.

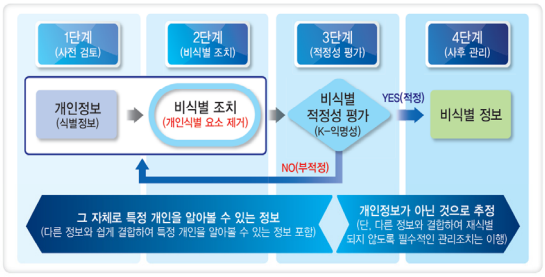

그림 2. 비식별화과 과정과 사후처리

단계별 조치 사항은 그림2와 같다. 개인정보에 해당하는지 여부를 검토 후, 개인정보가 아닌 것이 명백한 경우 법적 규제 없이 자유롭게 활용하도록 사전검토 후 비식별 조치로 데이터 셋에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 활용, 개인을 알아볼 수 없도록 하고 다음단계로 다른 정보와 쉽게 결합하여 개인을 식별할 수 있는지를 「비식별 조치 적정성 평가단」을 통해 적정성 평가 후 비식별 정보 안전조치, 재식별 가능성 모니터링 등 비식별 정보 활용 과정에서 재식별 방지를 위해 필요한 조치를 수행하는 사후 관리단계를 거친다[4].

III. 비식별화 기술과 비식별화 정도 평가 방법

비식별화는 빅데이터의 수집·활용의 모든 단계에서 개인정보가 식별되는 경우 또는 이후 정보의 추가 가공 등을 통하여 개인이 식별되는 경우에 적용한다. 가령 개인정보의 수집 및 저장 시, 개인정보가 포함되어 있을 수 있는 데이터의 활용 시, 다른 기관과의 정보 공유 시, 기관내의 서로 다른 부서간의 정보 공유 시 등의 상황에서 비식별화가 필요할 수 있다[7]. 식별화의 주요 기술로는 다음 표 1과 같이 개인정보 중 주요 식별요소를 다른 값으로 대체하여 개인식별을 곤란하게 하는 방법인 가명처리, 데이터의 총합 값을 보임으로써 개별 데이터의 값을 보이지 않도록 하는 총계처리, 데이터 공유·개방 목적에 따라 데이터세트에 구성된 값 중에 필요 없는 값 또는 개인식별에 중요한 값을 삭제하는 데이터 값 삭제, 데이터의 값을 범주의 값으로 변환하여 명확한 값을 감추는 범주화, 공개된 정보 등과 결합하여 개인을 식별하는데 기여할 확률이 높은 주요 개인 식별자가 보이지 않도록 처리하여 개인을 식별하지 못하게 하는 데이터 마스킹등의 방법이 있다. 각 처리 대상과 내용, 식별기법 및 장단점은 표 2,3과 같다. 그리고 18개의 처리기법별 세부기술을 보면 표 3과[5][7] 같다.

표 1. 비식별화 기법의 개념과 대상

처리 기법	개념	처리대상 식별정보
가명처리	개인 식별이 가능한 데이터를 다른 값으로 대체	성명, 기타 고유특징
총계처리	개인정보에 통계 값 적용	개인과 직접 관련된 날짜 정보, 기타 고유특징 등
데이터 값 삭제	개인정보 식별이 가능한 특정 데이터 값 삭제 처리	쉽게 개인을 식별할 수 있는 정보(이름, 전화번호, 주소, 생년월일, 사진 등), 고유식별정보(주민등록번호, 운전면허번호 등), 생체정보(지문, 홍채, DNA 정보 등), 기관·단체 등의 이용자 계정(등록번호, 계좌번호, 이메일 주소 등)
범주화	단일 식별 정보를 해당 그룹의 대표값으로 범주화, 구간 값으로 범위화	
데이터 마스킹	개인 식별 정보를 전체 또는 부분적으로 대체	

표 2. 비식별화 기법의 장단점

처리 기법	장점	단점
가명처리	자체로 완전 비식별화 가능, 데이터의 변형, 변질 수준이 적음	일반화된 대체값으로 가명 처리, 성명을 기준으로 한 분석에 한계
총계처리	민감한 정보의 비식별화 가능, 다양한 통계분석용 데이터 셋 작성에 유리	집계 처리된 데이터 기준으로 정밀한 분석 어려움, 집계 수량이 적을 경우 데이터 결합 과정에서 개인정보 추출 또는 예측가능
데이터 값 삭제	민감한 개인 식별 정보의 완전한 삭제 처리 가능하여 예측, 추론 등이 어려움	데이터 삭제로 인한 분석의 다양성, 분석 결과의 유효성, 분석정보의 신뢰성 저하
범주화	통계형 데이터 형식으로 다양한 분석 및 가공이 가능	범주로 표현되어 정확한 수치 값에 따른 분석, 특정한 분석 결과 도출 어려움, 데이터 범위 구간이 좁아질 경우 추적, 예측 가능
데이터 마스킹	완전 비식별화 가능, 원시 데이터의 구조에 대한 변형이 적음	과도한 마스킹 적용 시 필요한 정보로 활용하기 어려움, 마스킹 수준이 낮을 경우 특정한 값의 추적 예측 가능

표 3. 비식별화 세부 기술

기법	세부기술	
가명처리	휴리스틱 익명화	식별자에 해당하는 값들에 몇 가지 정해진 규칙 적용, 혹은 사람의 판단에 따라 가공하여 개인정보 숨김
	K-익명화	하나의 데이터세트 안에 동일한 속성값을 가진 데이터를 k개 이상 유지
	암호화	일정 규칙의 알고리즘을 적용, 암호화
	교환 방법	데이터베이스의 레코드를 미리 정해놓은 변수/항목들의 집합과 연계하여 교환
총계처리	총계처리	데이터 집단 또는 부분으로 집계 처리를 하여 민감성 낮춤
	부분집계	분석 목적에 따라 부분 그룹만 비식별처리
	라운딩	집계하여 처리된 값에 대하여 올림, 내림, 반올림 같은 라운딩 기준 적용
	데이터 재배열	기존 정보의 값은 유지하면서 개인정보와 연관되지 않도록 데이터 재배열
데이터값삭제	속성값 삭제	원시 데이터에서 민감한 속성값 등 개인식별항목 제거
	속성값 부분 삭제	민감한 속성값의 일부값을 삭제하여 대표성을 가진값으로 보이도록 함
	데이터 행 삭제	다른 정보와 비교하여 값이나 속성이 구별되는 식별정보 전체 삭제
	준식별자 제거	위에서 나열된 형태들을 제외한 모든 제거 형태의 익명화 기법
범주화	범주화	명확한 값을 숨기기 위해 데이터의 평균 또는 범주 값으로 변환
	랜덤 올림 방식	개인식별 정보의 수치데이터에 대하여 임의의 수를 기준으로 올림(round up) 또는 절사(round down)
	범위 방법	개인식별정보의 수치데이터를 임의의 수를 기준으로 범위 설정
	제어 올림	랜덤 올림 방법에서, 어떤 특정 속성값을 변경시 행과 열의 합이 맞지 않는 것을 제어하여 일치시킴
데이터마스킹	임의 잡음 추가	민감한 개인식별항목에 대해 임의의 숫자 등의 잡음을 추가, 식별정보의 노출 방지
	공백과 대체	비식별 대상 데이터를 공백과 대체의 방식

Ⅳ. 비식별화와 개인정보보호 관련 이슈

오늘날 광고시장은 TV, 신문등 전통적인 오프라인 매체에서 PC, 그리고 다시 휴대전화, 태블릿 등 모바일 기기로 확대되어가고 있다. 특히, 사용자의 행위를 분석하여 관심항목이 무엇인지 파악하고 관련된 광고를 제공하는 온라인 행동기반 맞춤형 광고로 진화하고 있다. 이것은 소비자의 구매행동을 예측하여 사용자에게 정확하고 유효한 광고를 제공하는 기술을 의미한다. 하지만 이러한 맞춤형 광고를 위해 사용자의 행위 정보를 수집 및 활용할 때는 개인정보 침해 가능성이 존재한다. 즉 개인정보보호법이나 정보통신망법에서 요구하는 동의절차 등을 고려해야 한다. 이것은 광고 분야뿐만 아니라 빅데이터나 IoT 관련된 분야에서도 논란이 되고 있다. 개인정보의 개념은 범위도 너무 넓고, 불필요한 형사처벌 규정도 많고 행정기관의 규제도 강해 글로벌 회사와의 경쟁이나 스타트업들이 사업 시작에 장애물이 될 수가 있다. 따라서 빅데이터와 같은 정보를 활용하면서도 정보주체의 개인정보자기결정권과 같은 기본권을 침해하지 않는 방법이 필요하다. 그것은 바로 비식별화 방법이다. 하지만 여기에는 부가적인 이슈와 해결과제가 존재한다.

1) 재식별화 문제

비식별화에 관한 많은 기술적·법적 문제가 있지만 가장 문제가 되는 것은 재식별화 문제이다. 정보의 주체가 누구의 것인지 모르는 정보도 그 양과 종류가 늘어나면 식별될수 있기 때문이다. 즉 비식별화는 재식별화 가능을 배제할 수 없다는 것이다. 비식별화된 정보가 재식별화 되어 개인정보로 취급되면, 결국 법의 제제를 받아야 한다. 이렇듯 비식별화된 개인정보를 포함한 빅데이터의 활용은 논란이 될 수밖에 없다. 빅데이터, 사물인터넷의 우선 과제는 개인정보보호 문제로 종결된다.

2) 사회적 합의 필요

개인정보보호 수준은 사회적 합의가 필요한 문제이다. 현재 정부가 창조경제 활성화와 서비스 경제 육성을 위해 명목으로 '개인정보 비식별조치 가이드라인'을 배포하였지만 미래창조과학부 하의 소프트웨어정책연구소는 기업들이 정부의 가이드라인을 준수하다보면 가이드라인이 법적 효력을 갖지 못하기 때문에 개인정보 불법 유출이 되면 개인정보보호법 위반으로 법적 책임을 질수 있다고 발표하였다[6]. 그리고 비식별화 기술과 적정성 평가기준을 잘 활용하면 개인의 정보도 침해하지 않고 개인정보보호법을 준수하면서 개인정보를 산업에 활용 및 기술 발전이 가능하다고 하였다. 특히 정부의 비식별조치 가이드라인은 시민단체들과 법조계와 학계에서도 비판이 제기되고 있다. 비식별화된 개인정보는 다른 정보와 결합되면 재식별이 가능하다는 한계점을 가지고 있다. 따라서 개인정보 보호와 빅데이터 활용의 상반관계의 적정선을 찾기 위해서는 법 개정 노력이 이루어져야 하고 비식별 정보의 유통에 대한 정부의 관리체계를 정비해야 할 필요가 있다.

Ⅴ. 결 론

최근 빅데이터 분석 시장이 확대되고 데이터 분석의 경제적 가치가 높아지고 있다. 또한 데이터의 활용성에 대한 사회적 수요가 증가하는 상황에서 무분별한 개인정보의 수집과 분석은 제어되어야 할 것이다. 그러나 비식별화의 기준을 까다롭게 하여 정보의 유용한 활용 가능성을 차단하는 것은 시대의 흐름에 역행하는 것과 같다. 따라서 비식별화의 기준을 합리적으로 설정하고 적절한 규범적 통제를 하는 것은 개인정보를 보호하는 동시에 유용한 정보의 활용가능성을 열어주는 유익한 기능을 할 것이다.

참고문헌

[1] 김동국, 이혁, "빅데이터 기반의 개인정보 비식별화 동향", 한국인터넷 정보학회, 제16권 제2호, 2015. 12.
[2] "빅데이터비식별화_기술활용안내서_ver_1.0", 미래창조부.
[3] Simson L. Garfinkel, "De-Identification of Personal Information", NISTIR 8053, 2015. 10.
[4] "개인정보 비식별 조치 가이드라인", 미래창조과학부 보건복지부, 2016. 6.
[5] "개인정보 비식별화기술 활용 안내서", 미래창조부, NIA, 2015. 6.
[6] 이현승, 송지환, "개인정보 비식별화기술의 쟁점 연구", 소프트웨어 정책연구소, 2016. 8.
[7] 고학수, 최경진, "개인정보의 비식별화 처리가 개인정보보호에 미치는 영향에 관한 연구" 개인정보위원회, 2015.12.