빅데이터 환경에서의 개인정보 비식별 처리 방법 분석

임 형 진*

* 금융보안원 보안연구부 보안기술연구팀(e-mail : hjlim@fsec.or.kr)

요 약

빅데이터는 기존 데이터에 비해 양(Volume), 다양성(Variety) 및 처리속도(Velocity) 측면에서 더 광범위한 데이터 특성을 가진다. 또한 빅데이터 분석을 통해 새로운 부가 가치(Value)를 창출할 수 있는 정보를 추출할 수 있으며 빅데이터 처리 기술을 통해 데이터 접근 가능성, 비즈니스 프로세스의 생산성 향상 및 비용 절감 등의 이점을 제공받을 수 있다.

빅데이터의 중요한 특징은 다양한 출처로부터 데이터를 결합할 경우 더 많은 새로운 정보를 도출할 수 있다는 것이다. 일반적으로 기업 간에는 대규모 데이터로부터 새로운 가치를 창출하기 위해 상호 간 데이터 교환을 필요로 한다. 하지만 이러한 데이터 교환을 통해 데이터를 결합할 경우 특정 개인과 관련된 패턴이 나타날 수 있는 위험이 야기된다. 이러한 문제점을 해결하기 위해 기업들은 데이터 교환 과정에서 비식별 처리를 수행하게 되는데, 특정 개인의 식별이 가능한 데이터를 수정하여 특정 개인을 식별할 수 없도록 하는 프로세스를 비식별 처리라 한다.

본 고에서는 금융보안원에서 2016년부터 추진 중인 비식별 처리에 대한 국제표준 개발 내용을 기반으로 최근의 비식별 처리에 대한 국제 표준 및 관련 가이드라인을 분석하여 비식별 처리를 위한 고려사항과 비식별 처리를 위해 요구되는 방법을 구조적으로 분류하여 제시한다.

우선, 최근 발간된 비식별 처리 관련 해외의 주요 가이드라인 및 표준인 ISO의 비식별 기술 표준(2015년), 미국 NIST의 비식별 처리 관련 가이드라인(2016년), 영국의 익명화 프레임워크(2016년), EU 보안연구기관인 ENISA의 빅데이터 프라이버시 설계(2015년)에 대하여 소개하고 비교 분석 결과를 제시한다.

또한 국가별·기관별 가이드라인 내용을 기반으로 빅데이터 환경에서 비식별 처리 과정, 처리자의 역할, 공유 방법 및 비식별 수준에 따른 특징과 고려사항을 제시하고, 비식별 처리 시 데이터의 유용성 혹은 프라이버시를 고려하는 데이터 중심의 처리 방법과 비식별 처리에 참여하는 참여자들 간의 역할 중심의 비식별 처리 방식을 제시한다. 마지막으로 비식별 처리에 대한 국제표준 추진의 향후과제와 전망을 제시한다.

01 빅데이터 환경에서의 개인정보 비식별 처리 방법 분석

I 서 론

정보 통신 기술의 급속한 발전과 폭발적인 성장으로 데이터의 양이 늘어나고 복잡해짐에 따라 기존의 데이터 처리 방법과 도구를 사용하여 주어진 시간 내에 데이터를 효율적으로 분석하는 것이 매우 어려워진다. 이러한 문제를 해결하기 위해 개발된 패러다임을 빅데이터라고 하며, 기존 데이터에 비해 양, 다양성 및 처리속도 측면에서 더 광범위한 데이터 특성을 가진다. 다양한 출처로부터 데이터를 결합하여 더 많은 새로운 정보를 도출할 수 있다는 것은 빅데이터의 대표적 활용 방법이다. 일반적으로 기업 간에는 대규모 데이터로부터 새로운 가치를 창출하기 위해 상호 간 데이터 교환을 필요로 한다. 하지만 이러한 데이터 교환을 통해 데이터를 결합할 경우 특정 개인과 관련된 패턴이 나타날 수 있는 위험이 야기된다.

대용량 데이터 기술은 데이터 보존 정책을 유지함과 동시에 전체 데이터 처리 과정에서 개인 식별 정보 및 민감한 데이터의 보호를 요구한다. 결과적으로, 기업들은 데이터 교환 과정에서 데이터에 대한 비식별 처리가 필요하게 된다. 비식별 처리란 식별 가능한 데이터를 수정하여 특정 개인을 식별 할 수 없도록 처리하는 프로세스를 의미한다. 데이터의 유용성을 손상시키지 않으면서 높은 수준의 비식별 처리는 불가능하다. 일반적으로 단순히 식별 정보만 삭제하는 낮은 수준의 비식별 처리는 재식별 가능성을 차단하기에 충분하지 않다. 반면에 매우 높은 수준의 비식별 처리는 다른 데이터에서 동일한 혹은 유사한 개인의 데이터를 연결하지 못하게 하여

빅데이터의 많은 잠재적 이점을 저해할 수 있다. 또 다른 쟁점으로, 데이터 처리자가 수집한 데이터를 분석한 후 어떤 목적으로 사용하는지 알 수 있어야 한다는 것이다. 즉 개인 정보의 원래 수집 목적에 한해 사용되는 것이 보장되어야 한다.

본 고에서는 금융보안원에서 2016년부터 추진 중인 비식별 처리에 대한 국제표준[1] 개발 내용을 기반으로 최근의 비식별 처리에 대한 국제 표준 및 관련 가이드라인을 분석한 후 빅데이터 환경에서 비식별 처리를 위한 고려사항과 비식별 처리를 위해 요구되는 방법을 구조적으로 분류하여 제시하고자 한다.

표 1 용어정의

용어	정의
익명화 (anonymization)	식별 데이터 집합과 개인 데이터 주체 간의 연관성을 제거하는 프로세스
비식별 (de-identification)	일련의 식별 데이터와 개인 데이터 주체 간의 연관성을 제거하는 모든 프로세스에 대한 일반적인 용어 참고 – 일부 가이드라인에서는 익명화라는 용어가 비식별의 동의어로 자주 사용되고 있으며 문맥에 따라 약간 다른 정의를 가질 수 있다. 본고에서는 수집, 사용, 보관 및 공유되는 데이터의 개인 정보를 제거, 수정 또는 난독화하는 것을 설명하기 위해 비식별이라는 용어를 사용하여 정보 위험을 예방하거나 제한하는 기술적 및 관리적 목표를 수행하는 활동으로 정의한다.
비식별 기술	데이터 주체의 신원을 모호하게 하는 목적으로 데이터 의미를 조작하는 기법
재식별 (re-identification)	식별 데이터와 데이터 주체 간의 관계를 재확립하는 모든 프로세스에 대한 일반적인 용어
가명화	식별 정보를 별명으로 대체하는 개인 식별 정보에 적용되는 프로세스
데이터 관리자 (data controller)	타인과 단독으로 또는 공동으로 개인 데이터에 의한 조직의 처리 목적과 수단을 결정하는 자연인, 법인, 공공 기관, 기관 또는 기타 단체
데이터 처리자 (data processor)	데이터 관리자를 대신하여 개인 정보를 처리하는 사람, 공공 기관, 기관 또는 기타 기관, 즉 서비스 제공자
개인 데이터 주체	이름, 식별 번호, 위치 데이터 또는 온라인 식별자, 즉 개인과 같은 식별자를 참조하여 직접 또는 간접적으로 식별 가능한 자연인

자료 : ITU-T draft Recommendation X.fdip (2017.3)

1) 금융보안원은 2016년 ITU-T SG17(정보보호) 정기회의에서 비식별 처리 프레임워크에 대한 표준 개발 필요성을 제안하였고, 회의 참여 국가들과 논의를 거쳐 2019년까지 해당 표준을 개발하는 과제가 채택됨

Ⅱ 비식별 처리 방법에 대한 가이드라인 분석

본 절에서는 비식별 처리에 대한 기준 및 개념을 제시하는 해외의 주요 가이드라인 및 표준인 ISO의 비식별 기술 표준(2015년), 미국 NIST의 비식별 처리 관련 가이드라인(2016년), 영국의 익명화 프레임워크(2016년), EU 보안연구기관인 ENISA의 빅데이터 프라이버시 설계(2015년)에 대하여 살펴보고자 한다.

1. 최근의 주요 비식별 처리 가이드라인 주요내용

가. (ISO/IEC WD 20889) 비식별 기술 표준

ISO/IEC WD 20889 표준은 현재 ISO/IEC JTC1에서 개발을 진행 중이며[2] ISO/IEC 29100(privacy framework)에서 프라이버시를 강화하기 위한 방법으로 비식별 기술을 제시하고 있다. 또한 비식별화와 관련된 용어, 비식별화 기술에 대한 분류, 재식별 위협을 경감하기 위한 방안을 제시하고 있다.

표 2 재식별 판별을 위한 4가지 기준

기준	내용
개별화 (single out)	전체 데이터 집합에서 특정 개인에 해당하는 집합을 식별 가능한 정도
연결 가능성 (linkability)	한 정보가 특정 개인을 알 수 없게 개별화 하였더라도 다른 정보와 동일 값 연결을 통하여 특정 개인의 정보임을 식별할 수 있는 정도
추론 가능성 (inference)	개별화로 특정 개인을 구별해 낼 수 없더라도 특정 정보의 속성과 값(흔하지 않은 값 등)을 통해서 특정 개인임을 유추해 낼 수 있는 정도
구별 불가능성 (indistinguihishability)	특정 정보의 값이 특정 그룹이나 소속에 포함됨을 확인할 수 있어 특정 개인을 구분해낼 수 있는 정도

자료 : ISO/IEC WD 20889, 2016.

2) ISO/IEC 2nd WD 20889 – Information technology – Security techniques – Privacy enhancing data de-identification techniques, 2016-05-30.

동 표준에서는 마스킹, 가명화, k-익명성[3], l-다양성[4], t-유사성[5], 샘플링, 총계 등 다양한 방식의 비식별 기술을 제시하고 있다. 또한, [표 2]와 같이 각 비식별 기술을 적용하였을 경우 재식별 가능성을 판별하는 4가지 기준을 제시하고 있다.

동 국제표준은 비식별 처리에 이용 가능한 기술을 전반적으로 제시하고 있으나, 비식별 처리 절차와 비식별 처리된 데이터에 대한 재식별 위험 관리 방법에 대한 내용은 포함하지 않고 있다.

나. (미국) NIST 비식별 처리 가이드라인

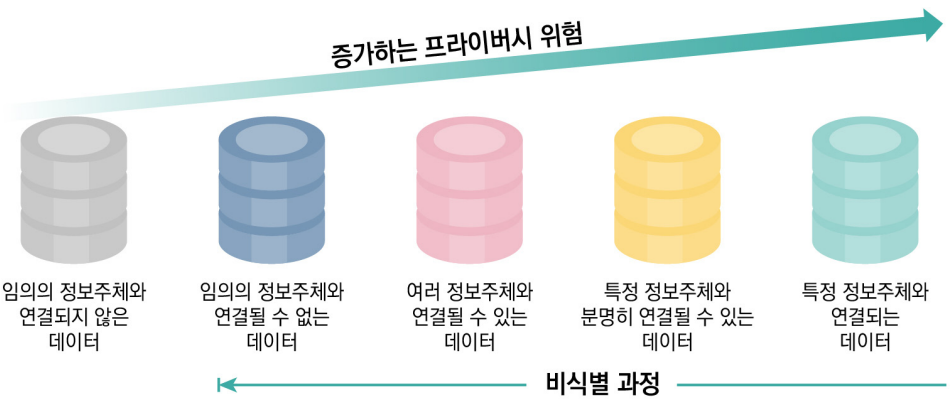
미국 상무부 산하의 표준화 기구인 NIST에서는 20여 년 간의 비식별화에 대한 논의를 정리하는 "개인 식별 정보의 비식별 처리" 가이드[6]를 2015년 10월 발간하였으며, 이전 보고서의 확장으로 "공공데이터에 대한 비식별 처리" 가이드[7]를 2016년 12월 추가로 발간하였다.

"개인 식별 정보의 비식별 처리" 가이드는 모든 데이터 비식별화 기술들에는 재식별 위험성이 존재한다는 것을 전제로 하고 있다. 특히, [그림 1]과 같이 특정인과 정보 간 연결가능성 등 비식별화 데이터 유형에 따라 프라이버시 침해 위험성의 정도를 표현하는 개념을 제시하였다. 즉, 데이터가 특정인과 연결되었는지, 특정인과 연결될 잠재적 가능성이 있는지, 특정인은 아니지만 어느 정도의 사람들과 연결될 가능성이 있는지 등에 따라 해당 정보가 식별되어 프라이버시를 침해할 위험성이 달라짐을 표현하고 있다. 또한 비식별화, 재식별화의 개념들을 개괄적으로 정의하고, 이미지나 유전자 정보와 같은 복잡한 정보의 비식별화에 대한 요구사항을 제시하였다. 특히, 데이터 비식별화를 위한 방법으로 삭제, 마스킹 등의 다양한 방법들을 설명하였고, 비식별화는 단일한 기법만 존재하는 것이 아니라 개별적인 활용 목적이나 상황에 따라 다양한 비식별 처리 기법이 수행될 수 있음을 제시하였다.

---

3) 특정인임을 추론할 수 있는지 여부를 검토하여 일정 확률수준 이상 비식별화가 되도록 하는 기법
4) 특정인인 추론이 안된다 하더라도 민감한 정보의 다양성을 높여 추론 가능성을 낮추는 기법
5) L-다양성 뿐만 아니라, 민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮추는 기법
6) Simon L. Garfinkel, "De-Identification of Personally Identifiable Information", NISTR 8053
7) Simon L. Garfinkel, "De-identifiying Government Datasets", NIST Special Publication 800-188(2nd DRAFT)

그림 1 비식별화 데이터 유형



(출처 : NIST, NISTR 8053, 2015.)

"공공데이터에 대한 비식별 처리" 가이드에서는 이전 가이드에서 다루지 않은 빅데이터 처리 과정에서의 데이터 생명주기 사례를 제시하고 생명주기 별 비식별 처리 고려사항을 설명하였다. 또한 비식별 처리 결과를 평가하기 위해 전문가가 참여하는 위원회에 대한 필요성을 별도의 절로 포함하였고, 비식별 처리 방법에 대한 교육과 홍보·연구의 중요성 등을 언급하였다.

다. (영국) 익명화 프레임워크

비식별화에 대한 최신의 사례를 구현하는 수단으로 영국정보국(ICO)에 의해 2012년 설립된 민간조직 UKAN에서는 "익명화 프레임워크"라는 가이드를 2016년에 발간하였다. 동 가이드는 2012년부터 약 5년간 20여 명의 보안 전문가들이 참여하여 비식별 처리를 위한 방법에 대해 논의한 결과를 다루고 있다.

동 가이드에서는 비식별 처리를 위해서는 각 국가에서 제정된 법률 및 규정이 우선 고려되어야 함을 언급하고 있다. 기존 비식별 처리 방법을 다루는 가이드와는 다르게 [표 3]에서 보여주는 바와 같이 비식별 방법을 정형 비식별화, 보장형 비식별화, 통계적 비식별화, 기능적 비식별화의 4가지로 분류하였다. 또한 비식별 처리 방법을 재식별 위험과 정보 유용성 관점에서 분류를 제시하였다.

표 3 비식별 방법

비식별 처리 타입	내용
정형 비식별화	특정 개인을 구분할 수 있는 식별자를 제거하거나 마스킹(masking) 하여 처리하는 방법
보장형 비식별화	특정 개인이 식별될 수 있는 위험을 제로(zero)에 가깝도록 데이터를 처리하는 방법으로 차분 프라이버시 방법[8] 등의 수학적 모델링에 의한 알고리즘을 적용하여 처리하는 방법
통계적 비식별화	재식별 가능성을 0의 확률로 축소하는 것이 어렵기 때문에 특정 속성 값에 대한 노출 빈도를 통계적으로 균일하게 처리하는 비식별 처리로써 데이터의 속성 값에 적용한다는 부분에서 정형 비식별화 대상과 차이가 있음
기능적 비식별화	데이터의 활용 목적에 따라 비식별 처리 기법을 결정하는 방식으로 재식별이 미치는 영향, 노출이 가능한 데이터, 비식별 데이터의 관리 방법, 재식별 가능성을 분석하여 적합한 방식을 선정하는 방법

자료 : UKAN, "The anonymisation decision-making framework", 2016.9.

마지막으로 동 가이드에서는 비식별 처리를 위해 3가지 관점에서 고려해야하는 10가지 요소를 제시하였다.

- **보유한 데이터의 활용 특성 이해** - ① 데이터 처리 요구사항 이해, ② 데이터 처리를 위한 법적 요구사항 이해, ③ 데이터에 포함된 속성 이해, ④ 데이터 활용 목적, 방법 이해, ⑤ 데이터 활용과 제한 사항 이해
- **프라이버시 노출 위협 평가와 대응** - ⑥ 프라이버시 노출 위협 평가 수행을 통해 필요한 비식별 처리 방안 수립, ⑦ 데이터 유용성을 고려한 노출 위협 경감 방안 수립
- **프라이버시 노출 위험 관리** - ⑧ 데이터를 이용하는 주체를 식별하고 위험 발생 시 처리 방안 수립, ⑨ 데이터를 제공 및 공유한 이후에 관리 방안 수립, ⑩ 위험 발생 시 대응 방안 수립

라. (ENISA) 빅데이터 프라이버시 설계

EU의 유럽 연합 대응기구인 ENISA에서는 빅데이터 환경에서 프라이버시 보호를 위한 가이드를 2015년 12월에 발간하였으며, 비식별 기술을 포함하여 암호화,

8) 익명화된 데이터 자체를 배포하는 방식이 아니라 데이터 요청 질의에 대해 해당하는 데이터를 응답하는 환경에서 응답 데이터에 수학적 모델링에 기반하여 ε(노이즈)를 포함하여 실제 데이터가 어떤 내용인지 알 수 없도록 하는 기법

접근제어 등의 보안 요구사항을 제시하였다. 특히 본 가이드에서는 빅데이터 프라이버시 보호를 위한 주요 기술 중 하나로 비식별 처리를 제시하고 있으며, 비식별 처리를 위해 프라이버시를 우선할 것인지 혹은 데이터 유용성을 우선할 것인지에 따라 비식별 처리 결과가 달라짐을 제시하였다. 또한, 빅데이터 활용을 위한 데이터 처리 과정을 '데이터 수집'→'데이터 분석'→'데이터 저장'→'데이터 사용'의 4가지 단계로 정의하고 단계별 차별적 요구사항을 제시하였다.

- **데이터 수집** - 데이터 분석 과정에서 프라이버시 노출을 방지하고자 데이터 수집 시점에 식별자를 제거할 수 있다.
- **데이터 분석** - 데이터 분석 시 다양한 추론에 의한 재식별이 발생하지 않도록 k-익명화 및 확장 방식, 차분 프라이버시 모델 등을 적용할 수 있다.
- **데이터 저장** - 데이터 유출 시 프라이버시 보호를 위하여 식별자 이외의 데이터에 대하여 암호화를 적용할 수 있다.
- **데이터 사용** - 데이터 이용 시 특정 개인을 추론하는 것을 방지하고자 비식별 기술을 적용할 수 있다.

2. 비식별 처리 가이드라인의 비교 분석

[표 4]에서는 Ⅱ.1절에서 소개한 가이드라인들을 항목별로 비교하고 있다. 우선 비식별 용어를 다르게 정의하고 있는데, 미국과 ISO의 경우 '비식별'이라는 용어를 사용하고 있고, 영국 및 EU에서는 '익명화'라는 용어를 사용하고 있다. 비록 용어는 다르지만 ISO, 영국 및 EU 모두 비식별이 다양한 기술을 이용하여 데이터의 재식별 위험에 대응할 수 있는 처리 수단을 의미한다. 둘째로, 각 가이드라인에는 빅데이터 분석 과정에서의 생명주기 정의 필요성과 이에 따른 비식별 처리 요구사항이 공통적으로 포함되어 있다. 가장 최근의 가이드라인인 NIST 800-188에서는 현재 표준화된 데이터 모델이 존재하지 않음을 언급하며 기존 데이터 모델의 사례들을 제시하고 있다. 셋째로, 최근의 가이드라인들은 비식별 처리 결과의 공유 방법에 따라 비식별 처리 방식이 달라질 수 있음을 언급하고 있으나 아직 표준화된 용어와

정형화된 형태로 제시하고 있지 않다. 넷째로, ISO/IEC WD 20889 등 초기 가이드라인에서는 비식별 처리 기술을 중심으로 가이드라인의 내용이 구성되어 있으나, 최근의 가이드라인에서는 타 연구 사례를 기반으로 비식별 처리 결과의 수준을 분류하는 개념까지 제시하고 있다. 또한 단순한 비식별 처리 기술에 대한 분류 뿐만이 아니라 비식별 처리 방법과 절차, 평가를 위한 방법을 포함하고 있다. 마지막으로, NIST 800-188에서는 비식별 처리를 수행하는 기업·기관에서 활용할 수 있는 비식별 처리를 위한 소프트웨어의 기능적 요구사항을 포함하고 있다.

표 4 가이드라인 비교

항목/가이드 (출간년도)	ISO WD 20889 (2015.12)	ENISA guideline (2015.12)	NISTTR 8053 (2015.10)	UKAN Guideline (2016.9)	NIST 800-188 (2016.12)
비식별 용어	De-identification	Anonymization	De-identification	Anonymization	De-identification
데이터 모델 (생명주기)	Data flow scenario	Data value chain	Data flow model	×	Data lifecycle models
데이터 공유모델	×	×	○	○	○
비식별 처리 수준의 개념	×	×	○	×	○
비식별 처리 기술 분류	○	○	○	○	×
비식별 처리 절차 및 방법	×	×	○	○	○
비식별 처리 결과 평가 방법	×	×	○	○	○
비식별 처리 SW 기능요구사항	×	×	×	×	○

초기 가이드들이 비식별 처리 기술 자체에 대한 정의와 소개 중심이었다면, 최근 가이드들은 효과적인 비식별 처리 방법과 비식별 처리 결과의 수준 평가 방안들을 포함하고 있다. 또한 비식별 용어는 다르지만 비식별 데이터는 재식별 가능성이 존재하기 때문에 기술적 선택과 관리적 방안들이 요구됨을 공통적으로 제시하고 있다.

Ⅲ 빅데이터 분석 환경에서 비식별 처리 시 고려사항

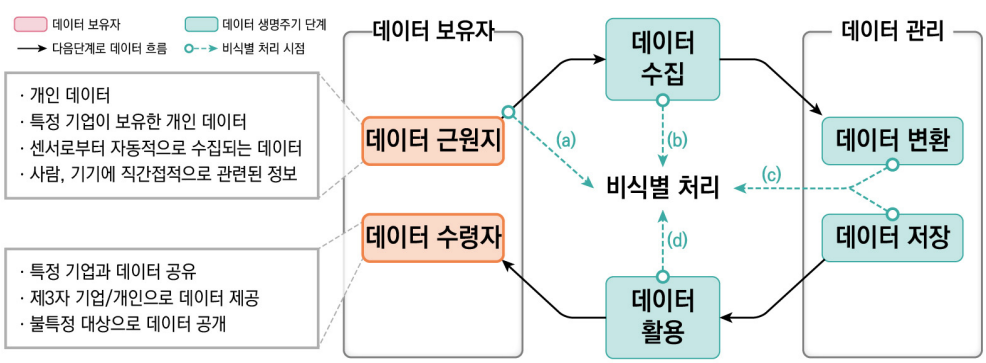
본 절에서는 Ⅱ절에서 검토하였던 국가별·기관별 가이드라인의 내용을 기반으로 빅데이터 환경에서 비식별 처리 시 처리 과정, 처리자의 역할, 공유 방법 및 비식별 수준에 따른 비식별 처리 특징과 고려사항을 제시한다.

1. 데이터 처리 과정

가. 빅데이터 분석 과정에서 데이터 생명주기

일반적으로 기업은 비즈니스 요구 사항, 관련 법률 및 규정에 따라 데이터 보호를 비롯하여 개인정보보호 및 보안 조치를 위한 목표를 수립하고 보안 정책을 운영한다. 기업이 빅데이터 처리과정에서 비식별 처리를 적용한다는 것은 개인의 사적인 영역에 영향을 주지 않으면서 개인 데이터를 사용하기 위한 목적이라고 볼 수 있다.

그림 2 빅데이터 환경에서 비식별 처리를 위한 생명주기



본 고에서 데이터 수명주기는 [그림 2]와 같이 데이터 보유자(데이터 근원지) → 데이터 수집 → 데이터 관리(데이터 변환 및 데이터 저장) → 데이터 활용 → 데이터 보유자(데이터 수령자)의 사이클을 가지는 것으로 정의하며, 각 단계에서 비식별 처리 고려사항을 제시한다. 빅데이터 처리 과정에서 생명주기는 비식별화된

데이터에 발생 가능한 프라이버시 위협과 취약성을 예측하여 비식별 처리가 의도된 목적과 용도에 따라 사용될 수 있도록 지원한다. 따라서 데이터 생명주기 개념은 활용 목적에 따라 재식별 가능성에 대한 합리적인 분석을 기반으로 적절한 대응 방안을 선택하는 데 활용될 수 있다.

나. 빅데이터 분석 과정에서 비식별 처리

빅데이터 분석을 위한 데이터는 개인 데이터의 주체, 데이터 관리자·처리자 및 각종 기기 등 다양한 데이터 보유자로부터 수집될 수 있다. 개인 정보가 포함된 수집 데이터 집합을 비식별 처리하면 그 자체로 특정 개인을 식별할 수 없는 새로운 데이터 집합이 생성되며, 따라서 비식별 처리된 데이터 집합은 개인 정보 노출 위험을 줄이기 위해 원래 데이터 집합 대신 기업에서 내부적으로 사용될 수 있다.

비식별 처리는 [그림 2]의 생명주기 모델에서 데이터를 수집하는 동안((a) 또는 (b)) 수행될 수 있다. (a)과정에서의 비식별 처리는 데이터가 수집되지만 식별 정보가 실제로 필요하지 않은 경우 수행될 수 있다. 즉, 개인정보보호 목적을 위해 데이터를 관리하는 단계인 데이터 변환 및 데이터 저장 단계에서 필요하지 않은 식별자를 수집 시점에서 미리 제거할 수 있다. 또한 식별 정보의 관리를 피하기 위해 데이터 변환 후 및 데이터 저장 전((c)과정)에 비식별 처리를 수행할 수 있다. 마지막으로, 기업 내에서 완전히 식별된 데이터가 필요한 경우 식별 정보는 분석 용도로 사용되고 데이터 공유 전((d)과정)에 비식별 처리 할 수 있다[9].

데이터 생명 주기 전반에 걸쳐 목적에 맞는 비식별 처리를 수행할 경우 개인정보 노출 위험을 최소화하고 데이터 공유를 상당히 쉽게 수행할 수 있는 장점이 있다. 빅데이터 분석을 위한 데이터 활용 목적과 성격은 비식별 처리를 어떤 시점에 수행해야하는지에 대한 정책 결정에 영향을 미칠 수 있다. 또한 이 결정은 각 빅데이터 분석 및 활용 목적에 따라 보다 효율적이고 데이터 유용성을 증대시킬 수 있는 특정 비식별 기술을 선택하는데도 도움을 줄 수 있다.

9) 빅데이터 분석 시 특정 개인을 식별하는 것이 법·규정에 저촉되느냐의 문제는 수집 데이터의 활용이 대해 이용 목적에 부합하는 것인지에 따라 다르게 판단할 수 있으며, 본 보고서의 데이터 생명주기에서는 기술적으로 적용 가능한 시나리오를 제시

1) 데이터 수집 단계

빅데이터 수집 단계에서 비식별 처리 시 불필요한 데이터를 최소화하는 데이터 최소화 원칙을 고려해야한다. 데이터를 수집하고자 하는 빅데이터 처리자는 관련 데이터 보존 기간을 포함하여 데이터 이용 목적에 부합하도록 실제로 필요한 개인 데이터를 사전에 정의하고 수집해야 한다. 따라서 전체 데이터 항목 중 수집 및 전송에 불필요한 데이터를 줄이기 위한 방법과 절차가 마련되어 있어야 한다.

대표적인 예로는 센서 장치, 휴대폰 등 다양한 정보 근원지로부터 수집하여 집계된 정보를 통계적인 분석 목적으로 사용하는 경우이다. 이 경우 분석 과정에서 특정인의 식별이 불필요하기 때문에 데이터 수집 전에 개인정보에 대한 비식별 처리를 적용하거나 식별자가 제거된 정보를 수집하여 데이터 분석 목적에 맞게 활용할 수 있다.

2) 데이터 관리 단계 (데이터 변환 및 저장)

데이터 관리 단계는 데이터 변환 단계와 데이터 저장 단계로 구성될 수 있다. 데이터 변환 단계는 정보의 유용성 및 개인정보보호 수준 개선을 위해 데이터를 비식별 처리하거나, 통계적 집계, 통계화 또는 암호화 하는 등 다양한 데이터 변환 처리가 포함된다. 데이터 변환은 데이터를 수집한 직후, 저장 전 또는 데이터 공유 전 등 언제든지 수행할 수 있다. 데이터를 수집한 후 즉시 개인정보유출 방지를 위해 데이터 변환을 수행할 경우 데이터 유출 사고 시 예상되는 피해를 줄일 수 있다. 그러나 수집 직후 비식별 처리를 일괄적으로 수행할 경우 데이터 간의 연관성이 저하되어 빅데이터 분석이 어려워질 수 있다.

데이터 변환 방법은 요구되는 개인정보 노출 위협을 방지하기 위한 적절한 기법이 무엇인지 고려하고 데이터의 활용 목적에 부합하는 분석이 가능하도록 선택되어져야 한다. 예를 들어 개인정보 노출을 우려한 나머지 데이터 원형을 심각하게 변환하는 경우 데이터 활용성이 떨어질 수 있다.

데이터 관리 활동 중 데이터 저장은 개인 정보를 포함한 데이터를 데이터 처리자

또는 데이터 이용자가 저장 장치에 저장하는 것을 의미한다. 일반적으로 데이터 저장·처리 과정에서 정보 보안 및 개인정보보호 통제 요구사항은 기존 국제표준 등에서 다루고 있으므로 본 고에서는 세부적인 사항은 기술하지 않고 항목만을 요약한다. 주요 통제 요구사항으로는 접근 제어, 유지 보수, 보안 평가, 인증 절차, 사고 모니터링 및 대응, 보안 감사와 같은 정보 보안 및 개인 정보보호 제어 등이 있다. 일반적으로 기업은 관련 법, 규정 혹은 계약 조항에 따라 데이터가 필요 이상으로 유지되지 않고 데이터 백업이 일정 기간 후에 파괴되도록 데이터 보존, 폐기 등 관리 정책을 운영한다.

3) 데이터 활용 단계

비식별 처리의 주요 목적은 개인의 사생활을 보호하는 것이다. 비식별 처리된 데이터는 데이터 유용성에 따라 다양한 용도와 목적으로 수집, 저장 또는 공개될 수 있다. 비식별 처리된 데이터는 다른 사람들에게 제공되며 원시 데이터가 가지고 있는 가치와 속성을 이용하여 분석된다. 따라서 비식별 처리는 개인정보의 노출을 방지하면서도 가능한 한 정보의 유용성을 유지해야 한다.

데이터 활용 모델은 다양한 상황에서 빅데이터 분석을 효과적으로 수행하며 비식별 처리의 목적을 달성 할 수 있도록 지원하기 위한 모델로써 공개 범위에 따라 공개, 반공개 및 비공개 모델로 분류될 수 있으며 모델에 따라서 다양한 수준의 정보 유용성 및 개인정보보호 특성을 가질 수 있다. 데이터 활용 모델은 법률적 요구사항 혹은 공유 목적에 따라 각 모델의 사용 적합성이 판별되는데, 어떤 활용 모델을 선택하는가에 따라 비식별 처리 대상 데이터와 방법이 달라질 수 있기 때문에 데이터 활용 모델의 선택은 비식별 처리 방법 선택을 위한 주요 기준으로 영향을 미칠 수 있다.

데이터 처리 기관은 비식별 데이터를 활용할 때 해당 데이터와 관련된 다양한 이해관계자를 포함한 전문가가 참여하는 위원회를 통하여 개인정보보호 수준과 데이터 유용성을 검토해야 한다. 또한 비식별 처리된 데이터의 위험 평가와 이를

위한 점검리스트는 프라이버시와 데이터 유용성의 균형을 맞추기 위한 적절한 비식별 처리 방법과 활용 모델 선택에 도움을 줄 수 있다.

결론적으로 비식별 처리 방법과 수준 선택은 빅데이터 활용 목적과 방법에 따라 달라질 수 있다. 따라서 빅데이터가 의도하는 목적에 활용될 수 있는 수준의 충분한 데이터 유용성을 보장하기 위해서는 비식별 처리 기법 선택 전에 데이터의 유용성이 먼저 분석되어야 한다.

2. 데이터 처리에 대한 역할과 책임

비식별 처리를 수행하는 참여자의 역할은 빅데이터 분석 환경 및 처리 환경에 따라 달라질 수 있으며, 비식별 처리 수행에 대한 책임에 따라 데이터 관리자, 데이터 처리자, 개인 데이터 주체로 분류될 수 있다. 하나의 참여자가 하나의 책임만을 가지는 것이 아니고 동시에 역할을 수행하는 경우도 발생할 수 있다.

빅데이터 분석 시 특정 데이터에 대한 법률적 처리 권한에 대한 이해는 비식별 처리 과정에서 연관되는 모든 데이터 처리자의 책임을 명확하게 수립하는데 중요한 요소가 된다. 따라서 빅데이터 처리 환경과 참여자의 형태에 따라 비식별 처리의 필요성, 비식별 처리 시점, 비식별 처리 주체 등은 참여자들의 비식별 처리에 대한 역할을 정의하는 기준이 될 수 있다.

가. 데이터 관리자

데이터 관리자는 개인정보가 포함된 데이터에 대해 처리 목적과 수단을 결정함으로써 데이터를 임의[10]로 활용하는 것을 방지한다. 데이터 관리자는 자신이 보유한 특정 개인의 데이터에 대한 처리 권한이 자신에게 유지되는 경우에만 다른 참여자가 데이터를 활용할 수 있도록 제공할 수 있다. 즉, 데이터 관리자는 빅데이터 처리 과정에서 다음의 대표적인 사례들을 포함한 전반적인 책임을 가지게 된다.

10) 한 예로 정보 수집 시 개인 데이터 주체로부터 정보수집에 대한 이용 동의가 선행되었거나 혹은 법률에서 정하는 요건에 따른 데이터의 합법적인 사용이 아닌 경우가 해당 될 수 있다.

○ 개인정보를 수집하고 활용하기 위한 법적 근거 제시
○ 수집할 개인 데이터 항목 및 데이터를 사용할 목적 제시
○ 데이터 공유 가능 여부 및 데이터 공유·제공 대상 제시
○ 데이터 상황을 고려한 비식별 처리 필요성 제시 등

다양한 참여자가 빅데이터 처리과정에 참여하는 환경에서 하나 이상의 데이터 관리자가 존재할 수 있으며 개인정보를 처리하고 결정하는 방식에 따라 “협력” 혹은 “공동”의 두 가지 형태로 분류할 수 있다. “협력형 데이터 관리자”는 다른 데이터 관리자와 동등하게 참여하여 개인 정보 처리 방식을 결정한다. “공동형 데이터 관리자”는 개인정보를 일종의 풀(pool)형태로 공유하면서 분석하는 경우로 각 관리자는 개별적으로 데이터를 활용한다.

나. 데이터 처리자

데이터 관리자와 대조적으로 데이터 처리자는 데이터 관리자가 제시하는 기준과 방법 하에 개인정보를 처리한다. 데이터 처리자는 법적 근거에 따라 특정 개인의 프라이버시를 보호할 수 있는 기술적·관리적 조치의 수행 능력을 보장할 수 있어야 한다. 따라서 데이터 처리자는 제공 받은 데이터를 데이터 관리자의 명시적 허가 없이 제 3자에게 다시 공유할 수 없다. 데이터 관리자와 데이터 처리자 간에 데이터가 공유된 이후에는 이들 상호 간에 빅데이터 활용 행위에 대한 근거가 기본적인 법적 의무와 개인정보보호의 원칙에 기반하여 결정되어야한다.

다. 개인 데이터 주체

개인 데이터 주체는 직접 또는 간접적으로 개인을 식별 가능한 데이터에 대한 주체를 의미한다. 이러한 개인 데이터는 자신을 식별할 수 있는 식별자 또는 자신의 신체적, 생리적, 정신적, 경제적, 사회적 또는 문화적 정체성과 관련된 하나 이상의 속성을 포함한다. 데이터 관리자나 데이터 처리자가 아닌 개인 데이터 주체가

다른 특정인의 데이터를 보유할 경우 이러한 데이터는 비식별 처리되어 다른 특정 개인을 식별할 수 없어야한다.

3. 데이터 활용 방법

데이터 활용 방법은 빅데이터 처리 과정 중 재식별 가능성을 줄이기 위한 데이터 이용·제공 방법으로 각 방법에 따라 서로 다른 재식별 위험을 가지게 된다. 따라서 데이터의 활용 목적(예, 여러 종류의 데이터 공유 혹은 주기적인 데이터 공유 등)과 이에 따른 프라이버시 위험을 고려하여 계층적 접근 제어를 이용한 다양한 활용 모델을 결합하여 사용할 수 있다. 데이터 활용을 위한 모델은 접근 제한이 없는 공개적인 모델부터 엄격한 제한을 요구하는 모델로 분류될 수 있으며, [표 5]는 세 가지 모델의 특징을 비교하고 있다.

표 5 데이터 활용 모델 비교

항목	공개적 데이터 활용 모델	반공개적 데이터 활용 모델	비공개적 데이터 활용 모델
접근 용이성	모든 사람이 자유롭게 데이터 접근 가능	허가된 사람 혹은 기관은 누구나 데이터에 접근 가능	특정 개인 혹은 기관만이 데이터에 접근 가능
활용 사례	웹 포털을 통한 데이터 공개	• 제한된 공간에서의 데이터 접근 • 요청에 의해 정보를 전달 • 원격 접근을 통한 데이터 열람 • 분석 결과만을 전달	기업·기관 간에 데이터 제공
재식별 위험 수준	매우 높은 위험	매우 높은 위험	엄격한 제한 하에 보통 위험
이용 권한	무제한 사용과 재사용	접근 권한이 있는 개인과 기업에게만 제공	데이터의 재활용 금지
가능한 재식별 공격	가능한 모든 공격	• 재식별을 위한 내부자의 의도적인 공격 • 부주의한 데이터 설정에 의한 개인정보 노출 • 데이터 유출	

자료 : ITU-T draft Recommendation X.fdip(2017.3)

가. 공개적 데이터 활용 모델

공개적 데이터 활용 모델은 데이터 사용에 대한 사전 등록이나 조건 없이 누구나 데이터에 접근할 수 있는 전형적인 공유 방법이다. 이러한 활용 모델의 예로 웹 포털을 통해 기업이 자유롭게 공개하는 정보로써 누구나 자료를 열람할 수 있는 형태가 있다. 이때 기업은 공개를 위한 데이터를 미리 준비하고 누구든지 해당 데이터를 재사용하거나 활용할 수 있도록 게시한다. 공개적 데이터 활용 방식은 데이터 접근에 대한 최대한의 가용성을 제공할 수 있지만 데이터에 포함된 개인정보보호를 위해서는 높은 수준의 비식별 처리가 필요하다. 공개적 데이터 활용 모델은 해당 정보에 접근 할 수 있는 사람과 방법 등에 제한을 두지 않는 것이 일반적이다. 따라서 공개적 데이터 활용 모델은 정보를 다운로드해서 사용하는 사람이 누구인지 식별할 수 없다.

뒤에서 언급할 "Ⅲ.다.절 반공개적 데이터 활용 모델"의 사례 중 정보 요청에 의한 데이터 접근 방식이라고 하더라도 정보를 요청한 사람에게 개인정보보호 의무나 데이터 사용에 대한 어떤 조건이나 동의를 요구하지 않는다면 공개적 데이터 활용 모델로 볼 수 있다.

나. 비공개적 데이터 활용 모델

일반적으로 개인정보가 포함된 데이터는 법·규정에 따라 공개가 허용되는 경우에만 기업·기관의 내·외부로 공유될 수 있다. 따라서 공개가 허용되지 않는 경우 기업·기관은 모든 개인정보가 삭제되어야 해당 정보를 타인에게 제공할 수 있다. 비공개적 데이터 활용 모델은 정보 접근에 대한 가용성은 매우 낮지만 다양한 기술적·관리적 개인정보보호 정책을 적용할 수 있어 상대적으로 낮은 수준의 비식별 처리를 요구한다.

비공개적 데이터 활용 모델은 기업 혹은 기관 간에 정보를 공유할 때 데이터에 대한 접근이 특정 조직으로 제한되기 때문에 정보 이용·제공 시 계약서를 통해 개인정보보호 및 보안에 대한 요구사항을 명시할 수 있다.

따라서 비공개적 데이터 활용 방식을 위해서는 상호 정보를 이용·제공하는 당사자 간 데이터 활용을 위한 계약이 요구된다. 데이터 활용을 위한 계약서에는 다음과 같은 항목이 포함되어 데이터 활용 시 개인정보 침해 위협을 경감하기 위한 요구사항으로 제시될 수 있다.

- ○ 정보에 접근할 수 있는 사람에 대한 명시 (정보의 이용자 제한)
- ○ 데이터 보안을 위한 보안 요구사항 명시 (정보처리 환경에 대한 제한)
- ○ 사용 제한, 특히 다른 파일과의 연결 및 의도적인 재식별 행위에 대한 금지 명시 (데이터 통제 방안)
- ○ 데이터의 사용 목적과 기한이 만료되면 데이터를 파괴해야한다는 요구사항 명시 (데이터 통제 방안)

비공개적 데이터 활용 시 계약서에 정보보호를 위한 사항을 명시하는 하는 목적은 다음의 세 가지 측면이 있다.

- ○ 데이터 관리자가 신뢰하는 개인이나 조직과 그렇지 않은 개인이나 조직을 명확하게 구별
- ○ 데이터 활용을 위한 접근이 발생할 수 있는 조건을 지정하기 위한 물리적·관리적·인적 요건의 명시
- ○ 개인 혹은 조직이 계약서에 명시된 조건을 변경하거나 따르지 않을 경우 제재 또는 벌칙을 지정

개인 혹은 기업은 비공개 데이터를 활용할 시에 개인정보보호를 위한 법률 혹은 계약 등을 명시해야한다.

다. 반공개적 데이터 활용 모델

공개적 데이터 활용에 비하여 반공개적 데이터 활용 모델은 정보 이용자가 데이터 이용 권한을 얻기 위해 공식적인 요청 및 승인 프로세스를 거치도록 요구하기 때문에 정보 접근에 더 제한적인 모델이다.

반공개적 데이터 활용 모델의 한 사례로 어떤 사용자가 인터넷으로 데이터를 다운로드 받을 때 이용약관 혹은 동의사항을 클릭하여 서명한 이후에 데이터를 열람하는 경우가 해당될 수 있다. 데이터 이용자는 제공받는 데이터를 통해 수행할 수 있는 작업 범위와 데이터 처리 방법의 제한 등을 확인할 수 있지만 누구든지 이용약관에 서명하면 데이터를 열람하여 활용할 수 있다.

반공개적 데이터 활용 모델은 정보를 요청하는 자에게 정보 접근의 용이성을 제공하면서도 특정 개인을 식별할 수 없도록 비식별 처리가 이루어져야한다. 이 방식에서 정보를 제공하고자 하는 조직은 다양한 접근 제어 방식을 적용할 수 있다.

- ○ 데이터에 접근하여 열람하기를 원하는 모든 데이터 사용자의 신원 정보를 사전에 검증·등록하도록 요구
- ○ 개인의 신원을 확인하기 위해 인증 프로토콜을 사용
- ○ 데이터별 계층적인 접근 권한을 부여하여 적합한 이용 권한을 가진 경우에만 해당 데이터 열람이 가능토록 제어

이러한 정보시스템을 활용할 경우 연구목적을 위한 특정 사용자를 대상으로 정보를 제공하게 하는 등 승인된 소수의 사람만이 데이터에 접근하도록 할 수 있다. 또 다른 반공개적 데이터 활용 모델 중 하나는 데이터 관리자가 빅데이터 분석 후 해당 결과를 사용자에게 전달하는 형태로 사용자가 직접적인 데이터가 필요하지 않은 경우 이러한 방식을 사용할 수 있다.

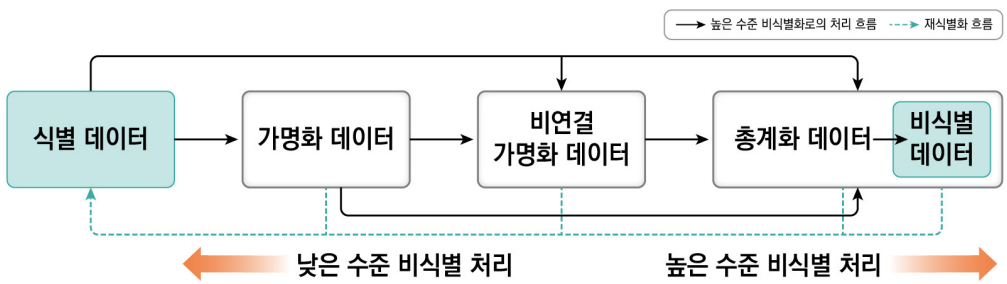
4. 데이터 형태와 비식별 수준

본 절에서는 데이터의 형태에 따라 특정 개인을 직접적으로 식별할 수 있는 정도와 방법을 설명한다. 데이터의 사용 목적 및 처리 방법 등에 따라 빅데이터 분석에 요구되는 데이터의 형태가 정의될 수 있으며 여기에는 데이터의 구성, 개인을 식별할 수 있는 정보, 개인을 식별하는데 활용될 수 있는 속성이 포함되어야 한다.

[그림 3]은 비식별 처리 과정에서 나타날 수 있는 데이터 형태를 식별 데이터

부터 비식별 데이터까지 표현하고 있다. 각 데이터 형태는 모두 재식별 위험이 존재하며 각기 다른 수준을 가짐을 나타내고 있다. [그림 3]에서 보여주는 바와 같이 “식별 데이터 형태”는 비식별 처리가 수행됨에 따라 점차적으로 가장 높은 수준인 “비식별 데이터 형태”로 변환되는 것을 보여주고 있다.

그림 3 비식별 처리과정의 데이터 형태와 비식별 수준



(출처 : ITU-T draft Recommendation X.fdip(2017.3))

[그림 3]의 오른쪽 끝은 특정 개인을 식별할 수 없는 데이터 형태로 재식별 가능성이 매우 낮은 형태를 나타내며, 왼쪽 끝은 특정 개인과 직접 연결된 식별 가능한 데이터 형태를 나타낸다. 이 두 데이터 형태 사이에는 특정 개인을 식별하기 위한 시도를 통해 특정 개인이나 개인이 포함된 그룹을 찾아낼 수 있는 데이터 형태가 존재한다. 또한 특정 개인 데이터를 기반으로 하지만 원래의 데이터로 복구할 수 없는 데이터 형태도 존재한다.

빅데이터의 효율적 활용을 위해서는 높은 수준의 비식별 처리가 진행되더라도 요구되는 데이터의 유용성을 유지하면서 [그림 3]의 오른쪽 방향의 비식별 데이터 형태로 변환이 되어야 한다. 다음은 각 데이터 형태에 따른 비식별 처리 특징 및 재식별 방법을 기술한다.

가. 식별 데이터

식별 데이터 형태에서는 데이터에 포함된 정보가 개인의 것이라는 것을 관찰

가능하기 때문에 데이터가 특정 개인과 명확하게 연관될 수 있다. 한 예로 주민번호, 여권번호 등 고유 식별자를 포함한 데이터는 식별 데이터 형태라 말할 수 있다.

나. 가명화 데이터

가명화된 데이터 형태에서는 모든 식별자가 다른 값으로 대체되기 때문에 대체 처리를 수행한 당사자가 아닌 사람은 특정인과 연결될 수 있는 원래의 데이터를 알 수 없다.

다. 비연결 가명화 데이터

비연결 가명화 데이터 형태에서는 모든 식별자를 지우거나 혹은 가명화를 위한 대체 방법도 유지하지 않기 때문에 비식별 처리를 수행한 당사자도 비식별 처리 이전의 원래 데이터로 복구가 불가능하다.

라. 총계화 데이터

총계화 데이터 형태에서는 특정 개인을 식별할 수 있는 값들을 포함하지 않도록 서로 다른 사람에 대한 정보를 구성한다. 총계화 방법을 통해 형성된 데이터는 특정 값을 통해서 식별할 수 있는 사람들의 수(예, k-익명화의 k값 등)를 설정하고 그 수 미만으로 데이터를 형성하여 특정 사람을 식별할 수 없도록 한다.

마. 비식별 데이터

비식별 데이터는 특정인에 해당하는 데이터 값을 변경하여 직·간접적으로 다른 데이터와 결합이 불가능한 형태로써 데이터 자체 혹은 다른 데이터와 결합을 통해서도 재식별이 어려운 형태이다. [표 1] 용어정의에서 기술하였던 바와 같이 비식별(de-identification)은 비식별을 위한 기술적인 조치뿐만이 아니라 비식별 데이터의 재식별 가능성을 차단하기 위한 관리적 행위도 포함한다.

Ⅳ 비식별 처리 방법의 분류와 특성

본 절에서는 비식별 처리 시 데이터의 유용성 혹은 프라이버시를 고려하는 데이터 중심의 처리 방법과 비식별 처리에 참여하는 참여자들 간의 역할에 따른 역할 중심의 비식별 처리 방식을 제시한다.

1. 데이터 중심의 처리 방법

비식별 처리가 개인 정보 노출을 막기 위해 원본 데이터를 수정하는 것이라고 할 때, 정보의 유용성과 프라이버시 보호 간에 상대적인 역학 관계가 존재하게 된다. 따라서 빅데이터 환경에서 비식별 처리 시 데이터 유용성 손실을 최소화 할 수 있는 비식별 처리 방법을 선택하는 것이 중요하다. 즉 데이터 사용자는 원본 데이터를 이용하여 빅데이터 분석을 수행하는 것처럼 데이터의 정확성을 잃지 않는 수준의 비식별 데이터를 확보하기를 기대한다.

데이터가 다양하고 양이 방대한 빅데이터 환경에서 데이터의 유용성을 손상시키지 않으면서 완벽하게 비식별 처리를 수행하는 것은 매우 어려운 작업이다. 식별자만을 제거하는 정도의 낮은 수준의 비식별 처리는 재식별을 방지하기에 충분하지 않다. 또한 매우 높은 수준의 비식별 처리는 다른 데이터에 포함된 동일한 개인 또는 유사한 개인의 데이터를 결합하여 분석하는 것을 불가능하게 하여 빅데이터 활용의 궁극적인 장점을 훼손할 수 있다.

본 절에서는 데이터 유용성 확보와 프라이버시 보호 수준 유지 간에 균형을 맞추기 위한 두 가지 비식별 처리 접근법을 제시한다.

가. 데이터 유용성 우선 비식별 처리

빅데이터 환경에서 개인에 대한 정보는 다양한 출처로부터 수집될 수 있기 때문에 동일한 개인 또는 유사한 개인에 속하는 데이터의 연결은 빅데이터 분석의 핵심적인 기능 중 하나이다.

데이터 유용성 우선 비식별 처리 방법에서는 전체 데이터 집합 대상이 아닌 개별 데이터(레코드)에 대한 유용성이 보장될 수 있는 비식별 기술을 적용해보면서 적정 방식을 선택하게 된다. 따라서 데이터 유용성 우선 비식별 처리 방식은 적절한 비식별 처리 방법을 찾는데 시간이 걸리면서도 프라이버시 보장 수준이 낮아지게 된다. 예를 들어 원본 데이터와 비식별 데이터 간의 연결을 시도해 봄으로써 경험적으로 재식별을 추정할 수 있는 위험이 존재한다. 이 방식에서 재식별 위험이 매우 높다고 판단되면 이러한 위험을 제거하기 위해 데이터 유용성 손실을 감수하면서라도 적정 비식별 수준을 제공할 수 있는 다른 비식별 처리 방법을 찾아야한다.

즉, 비식별 데이터가 다른 데이터와 결합될 수 있는 특징은 데이터 유용성 관점에서 바람직하지만 이러한 재식별 가능성은 프라이버시 보호를 저해하는 결과를 가져오게 된다. 따라서 다른 데이터와의 결합 정확성은 원본 데이터보다 비식별 데이터 형태에서 훨씬 적어야한다. 전문가는 비식별 처리된 데이터의 적절성 분석을 통해 데이터 결합 정도와 재식별 위험을 수용할 수 있는 비식별 처리 방법을 결정할 수 있다.

나. 프라이버시 우선 비식별 처리

프라이버시 우선 비식별 처리는 특정 개인에 대한 식별자와 특정 속성 값의 노출 위험의 상한을 보장할 수 있도록 적절한 변수들을 적용하는 방법이다. 이와 같은 프라이버시 보장 기법은 노출 위험을 최소화할 수 있는 특정 비식별 처리 방법에 적합한 변수들을 적절하게 설정하게 된다. 잘 알려진 프라이버시 보장 기법으로는 k-익명성, l-다양성, t-유사성 및 ε-차등 프라이버시 기법[11] 등이 있다.

프라이버시 우선 비식별 처리에서 비식별 데이터의 정보 유용성이 매우 낮다면

11) 익명화 된 데이터 자체를 배포하는 방식이 아니라 데이터 요청 질의에 대해 해당하는 데이터를 응답하는 환경에서 응답 데이터에 수학적 모델링에 기반하여 ε(노이즈)를 포함하여 실제 데이터가 어떤 내용인지 알 수 없도록 하는 기법

정보 유용성을 높일 수 있는 비식별 처리 방법으로 변경하거나 다른 프라이버시 보장 기법을 선택해야한다.

2. 역할 중심의 처리 방법

본 절에서는 비식별 처리를 수행하는 과정에서 참여자의 역할과 책임에 따라 중앙 집중형, 지역형 및 협력형의 세 가지 비식별 접근법을 제시한다. 역할 중심의 처리 방법은 '누가', '무엇을', '어디에서', '어떻게' 비식별 처리를 수행하는가에 따라 다른 특징을 가지게 된다.

- ○ 데이터를 이용하는 사람 혹은 기관은 누구인가?
- ○ 어떤 종류의 데이터 분석이 수행되는가? 혹은 수행되지 않는 분석은 무엇인가?
- ○ 데이터의 접근 및 분석은 어디에서 수행되며 어떻게 데이터에 접근할 수 있는가?

가. 중앙 집중형 비식별 처리

중앙 집중형 비식별 처리는 일반적으로 원본 데이터 전체에 대해 접근이 가능한 환경에서 활용될 수 있다. 즉, 중앙 집중형 비식별 처리 방식에서는 전체 데이터를 수집하거나 보유하고 있는 데이터 관리자가 통계적인 노출을 방지할 수 있는 총계화 기법을 적용하는 것이 가능하다. 이러한 중앙 집중형 비식별 처리는 다음과 같은 장단점을 가진다.

1) 장점

중앙 집중형 비식별 처리의 경우 데이터 원본을 제공한 개인 혹은 조직이 직접 비식별 처리를 수행하지 않아도 된다. 데이터 관리자는 비식별 처리를 위하여 충분한 전산처리 능력을 보유할 수 있고 원본을 제공하는 개인 혹은 조직에 비하여 비식별 처리에 대한 전문지식을 가진다. 따라서 전체 수집된 데이터에 대한 적정한 비식별 처리 능력을 보유하고 있다.

데이터 관리자는 모든 데이터를 확보하고 있기 때문에 원본 데이터를 비식별 처리함에 있어서 데이터 유용성과 노출의 위험을 최적화하기 위한 평가 능력을 확보하고 있다.

2) 단점

데이터 관리자는 원시 데이터를 제공한 모든 당사자가 신뢰할 수 있어야 한다. 데이터 관리자가 여러 출처로부터의 데이터를 통해 공식 통계를 제공하는 국가 통계 연구소인 경우는 문제가 되지 않지만 데이터 관리자가 단순한 사기업(예: 데이터 브로커)일 경우 해당 기업에 대한 신뢰 여부가 빅데이터 활용의 주요 장애가 될 수 있다.

특히 빅데이터의 경우 단일 관리자가 비식별 처리 시 정보처리 능력에 부담이 커질 수 있다. 또한 여러 데이터 관리자가 존재하여 특정한 빅데이터 처리를 개별적으로 수행 할 경우 중앙 집중형의 장점이 훼손될 수 있다.

위와 같은 장단점을 통해 지역형 비식별 처리와 협력형 비식별 처리 모델이 고려될 수 있다.

나. 지역형 비식별 처리

지역형 비식별 처리는 데이터 보유자가 데이터를 수집하는 데이터 관리자를 신뢰하지 않는 경우에 적합한 방법이다. 지역형 비식별 처리는 각 주체가 자신의 데이터를 데이터 관리자에게 전달하기 전에 직접 비식별 처리를 수행하고 제공하는 형태이다.

하지만 데이터 제공 전에 비식별 처리를 함에 따라 정보의 손실로 인해 중앙 집중형에 비해 데이터 유용성이 떨어질 수 있다. 즉, 개인 데이터 주체는 전체적인 데이터 집합에서 데이터 유용성 확보 방안에 대한 고려 없이 비식별 처리를 하기 때문에 정보 노출 위험을 줄이면서 정보의 유용성을 살릴 수 있는 절충점을 찾기가 어렵게 된다.

다. 협력형 비식별 처리

협력형 비식별 처리는 중앙 집중형의 장점인 유용성 확보와 지역형의 장점인 높은 프라이버시 보호 능력을 결합한 방식이다.

중앙 집중형의 문제는 데이터 관리자의 적절한 데이터 사용 및 비식별 처리 능력을 신뢰하지 않으면 개인 데이터 주체가 허위 정보를 제공하거나 정보 자체를 전혀 제공하지 않을 수 있다는 것이다. 따라서 데이터 보유자는 협력 작업을 통해 다음과 같이 정보의 유용성을 검토하고 적절한 수준의 비식별 처리를 적용하여 자신의 데이터와 관련된 노출 위험 수준을 결정할 수 있다.

- ○ 데이터 보유자가 제공하는 데이터에 동일한 프라이버시 수준을 적용함으로서 정보 손실을 최소화하여 지역형 비식별 처리의 방식을 보완
- ○ 데이터 보유자나 데이터 관리자 모두 최종 비식별화된 데이터에 포함된 정보 이상의 정보를 확보할 수 없기 때문에 정보가 집중되는 중앙 집중형 데이터 관리자의 신뢰 문제를 해소

이러한 협력형 비식별 처리는 외부의 강제 메커니즘 없이 참여자들 간 상호 협력을 위한 프로토콜[12]로 구현될 수 있다. 협력형 비식별 처리 방식에서 정보의 유용성을 높이기 위해서는 안전한 다중참여자 데이터 결합 방법이 필요하다. 즉, 참여자들이 다른 참여자의 데이터를 관리하지 않으면서 다른 참여자의 데이터가 모두 포함되어 활용될 수 있도록 하는 협력 프로토콜이 요구된다.

안전한 다중참여자 결합 방법은 중앙 집중형의 저장 공간의 필요 없이 협력 프로토콜에 의해 데이터 결합을 수행할 수 있기 때문에 데이터 유출에 의한 침해 가능성을 줄일 수 있고 각기 다른 참여자들을 완전히 신뢰하지 않더라도 참여자들 간 데이터 결합을 수행할 수 있다.

---

12) 협력 프로토콜은 전자적으로 자동화 되어 구현될 수 있지만, 빅데이터 활용을 위한 데이터 결합 전에 상호 협의에 의하여 결합 요구사항을 상호 제공할 수 있음

V 향후과제와 전망

본 고는 앞서 언급하였던 바와 같이 금융보안원이 주도적으로 추진하고 있는 ITU-T 국제표준에 제안한 비식별 처리 표준에 대한 내용을 기반으로 하였다. 본 고에서는 기술 중심의 비식별화를 언급하던 기존의 국내·외 가이드라인 및 표준에는 포함되지 않은 빅데이터 활용을 위한 효율적인 비식별 처리 방안을 위한 요소들을 제시하고 있다.

본 고의 제안과 국내의 비식별 처리 가이드라인을 비교한다면 우리나라는 산업 영역별 6개 비식별 처리 전문기관에 의해 비식별 처리(산업 영역별 분리된 중앙집중형 방식)를 수행하면서도 데이터 결합을 요청하는 기업에서도 비식별 처리(지역 비식별 처리)를 동시에 수행하기 때문에 기관 간 데이터 결합과정에서 비식별 처리 시 정보의 유용성을 떨어뜨리게 되고 데이터 결합을 위해 전문기관으로 모든 데이터를 전송하게 된다. 또한 빅데이터 분석을 위한 기관 간 데이터 결합 시 두 기관 간에 데이터를 공유(비공개 데이터 공유 모델)하지만, 데이터 활용 제약에 대한 근거가 당사자 간 계약보다는 개인정보보호 법에 근거하고 있다.

향후 정보의 유용성을 살리면서도 비식별 수준을 적정히 유지할 수 있는 협력형 비식별 처리 방안의 개발이 요구된다. 또한 빅데이터를 처리하는 과정에서 여러 참여자들 간에 다양한 형태의 정보 교환 상황이 발생할 수 있으며, 이 경우 개인정보보호에 대한 책임 소재 이슈가 발생할 수 있다. 빅데이터 활용 과정에서 요구되는 책임 소재를 명확히 규정하고 각 참여자는 부여된 역할 범위 내에서 데이터 처리를 수행하는 것이 필요할 것이다.

ITU-T 국제표준회의에서는 이러한 고려사항에 기반하여 우리나라 비식별 처리 가이드라인의 내용을 구조화하고, 타 국가 및 기관들의 기준들을 수용하여 표준안에 반영할 예정이다. 또한 향후 리스크 평가와 관리에 관한 부분과 비식별 처리 절차 등도 포함하여 2019년까지 비식별 국제표준을 개발 완료할 예정이며, 비식별 기술을 중점적으로 다루고 있는 ISO/IEC에서 개발 중인 비식별 처리 기술 전문가들과도 표준개발상황을 공유하여 협력할 예정이다.

참고문헌

[1] HyungJin Lim, A new work item proposal for framework of de-identification processing service, ITU-T SG17 WP4/Q7, 2016.

[2] HyungJin Lim, Proposal for the 1st revised text for draft Recommendation X.fdip : Framework of de-identification processing service for telecommunication service providers, ITU-T SG17 WP4/Q7, 2017.

[3] NISTIR 8053, De-Identification of Personal Information, 2015.

[4] NIST Special Publication 800-188(2ndDRAFT),De-identifying government datasets, 2016.

[5] ISO/IEC JTC 1/SC 27, Information technology - Security techniques - Privacy enhancing data de-identification techniques, 2016.

[6] ISO/IEC JTC 1/WG 9, Big data - Reference architecture - Part 4: Security and privacy fabric (draft), 2015.

[7] ISO/IEC, Information technology - Security technique - Privacy framework, 2011.

[8] ISO/IEC JTC1/SC 38, Information technology - Cloud computing - Cloud services and devices : data flow, data categories and data use, 2016.

[9] Recommunication ITU-T Y.3600, Big data - Cloud computing based requirements and capabilities, 2015.

[10] Health Information Trust Alliance, De-Identification Framework, 2015.

[11] ENISA, Privacy by design in big data - An overview of privacy enhancing technologies in the era of big data analytics, 2015.

[12] The Scottish Government, Joined-up data for better decisions: Guiding principles for data linkage, 2012.

[13] UK Anonimyzation Network, The anonymisation decision-making framework, 2016.

[14] Berkeley Technology Law Journal, Towards a modern approach to privacy-aware government data release, 2015.