재식별화 동향 연구

김재원, 이선영*
순천향대학교

{jwk0016, *sunlee}@sch.ac.kr

Re-identification Cases and Trends

Jae Won Kim, Sun-Young Lee*
Dept. of Information Security Soonchunhyang Univ.

요 약

공공 데이터와 개인정보의 활용도는 시간이 지날수록 더 높아지고 있다. 특히, 데이터 학습을 통한 산업과 소비가 발전함에 따라 국가에서는 보유한 공공 데이터를 적극적으로 개방하여 국민들과 공유하기 위해 공공데이터 정책을 추진하게 되었다. 그러나 개인정보를 포함한 데이터가 악용될 소지가 있기 때문에, 개인정보를 식별되지 않도록 가명/익명처리 등 비식별 조치로 안전성을 확보해야 한다. 비식별화 처리된 데이터에서 개인정보를 식별하기란 어려운 일이다. 다만, 동일한 데이터를 다른 비식별 조치를 취해 여러 곳에서 공개할 경우, 비식별화 처리된 데이터들을 종합해 유의미한 정보의 획득과 개인 식별이 가능해진다. 이를 재식별화라고 한다. 본 논문에서는 재식별화 기술의 동향을 분석하였다.

Ⅰ. 서 론

공공 데이터 및 개인정보의 활용도가 높아짐에 따라 데이터 비식별화의 중요성이 높아지고 있다. 데이터 비식별화는 개인정보를 보호하기 위해 정보를 식별할 수 없게 하는 기법이다. 그러나 비식별화된 데이터를 다른 정보와 조합하여 개인을 식별하는 기법인 재식별화로 인해 완벽한 비식별화가 불가능하다.

여러 연구에 따르면, 비식별 처리된 대량의 데이터와 개인정보가 다른 데이터베이스를 사용하여 식별된 사례가 있다. 이 연구들은 비식별화된 데이터가 여러 데이터베이스를 통하여 충분히 재식별화가 가능하고 개인 식별이 가능함을 보였다. 따라서, 본 논문에서는 재식별화 연구 사례와 재식별화 기술의 동향에 대해 분석하였다.

Ⅱ. 본 론

2.1 비식별화(De-identification)

비식별화는 개인정보를 식별할 수 없는 형태로 변환하는 과정이다. 비식별화의 주요 기법으로는 가명화, 일반화, 잡음 추가, 총계처리, 암호화 등이 있다[1].

비식별화를 진행할 때, 데이터의 특성과 보안 요구사항을 고려하여 적절한 비식별화 기법을 선택하고 적용해야 한다. 비식별화는 그 자체로 개인을 식별할 수 있는 정보 및 다른 정보와 결합하여 개인을 알아볼 수 있는 정보를 대상으로 한다. 비식별화를 진행할 때, 다른 정보와 결합에 따른 재식별 위험을 줄여야 한다. 그러나 비식별화 기술은 재식별에 대한 일정한 한계를 가진다. 따라서, 명확한 비식별화 목적으로 재식별에 대한 위험요소를 최소화해야 한다. 비식별화 과정에서 불필요한 개인정보가 생성되거나 비식별화 처리된 정보가 재식별화 된 경우에는 지체없이 삭제해야 한다[2]. 이러한 과정을 통해, 비식별화는 개인정보 유출로부터 개인 정보를 보호할 수 있는 중요한 기법이 되었다.

2.2 재식별화(Re-identification)

재식별화는 비식별화된 데이터를 활용하여 개인을 식별하는 과정이다. 비식별화를 하는 주된 목적은 승인되지 않은 재식별을 방지하는 것이기 때문에 재식별화는 종종 재식별 공격이라고도 부른다.

개인이나 조직이 재식별을 하는 데에는 여러 이유가 있다. 일반적으로 비식별처리의 질을 시험하고 재식별 수행에 관한 대중의 관심 또는 전문가적 관점을 얻기 위해 사용된다. 그러나, 악의적으로 비식별처리를 수행한 기관 및 사람들에게 피해를 야기하기 위해 사용되기도 한다[3]. 따라서 재식별화는 의료 또는 금융 데이터와 같은 민감한 개인정보 맥락에서 논의된다.

재식별화로 인하여 발생하는 피해를 줄이기 위해 재식별 위험 분석을 진행한다. 재식별 위험 분석이란, 민감한 정보를 분석하여 대상이 식별되거나 개인의 민감한 정보가 주출되는 위험을 높일 수 있는 속성을 찾는 프로세스이다. 이 방법은 재식별 전에 사용하여 효과적인 익명화 전략을 결정하거나 익명화 이후에 사용하여 변경 또는 이상점을 모니터링 할 수 있다[4].

2.3 재식별화 동향

현재 지속적으로 데이터를 재식별하려는 시도와 연구가 진행되고 있다. 기존에 존재하는 일반적인 비식별화 적용기술들만으로는 한계점이 존재하고 취약좀 또한 존재함이 밝혀져 재식별 시도가 늘어났다.

2008 년 발표된 ‘넷플릭스 상 익명성을 깨는 방법’이라는 논문에서는 2007 년 넷플릭스에서 공개한 고객들의 개인정보가 제거된 고객들이 이용한 영화들의 순위가 포함된 데이터베이스에서 고객의 정보를 재식별화할 수 있었음을 보였다. [5]. 2013 년에 게시된 ‘군중 속에서의 독특함: 인간 이동성의 개인정보 보호 범위’라는 기사에서 사람과 차량을 ‘이동 궤적’을 이용하여 식별할 수 있다는 점을 보여주었다 [6]. 2014 년 뉴욕 시 택시 리무진 위원회는 2013 년의 모든 뉴욕 시 택시 승차 기록이 담긴 정보 집합을 공개하였다. 이 정보에는 택시 기사나 탑승자의 성명이 포함되어 있지 않았지만, 택시의 번호판 번호로 쉽게 변환할 수

있는 32 자리로 된 알파벳 숫자 코드가 포함되어 있었다. 한 정보 과학자는 택시 번호판이 선명하게 보이는 곳에 택시에서 타고 내리는 유명인사가 담긴 시간이 찍혀 있는 사진을 인터넷에서 찾아낼 수 있었다[8].

이러한 사례를 보아 재식별화 시도가 끊임없이 증가하고 있는 추세임을 알 수 있다. 이에 따라 재식별을 방지하고 데이터 유출을 막기 위한 연구가 정부와 기업의 지원 아래 이루어지고 있다. 최근에는 머신러닝 알고리즘 훈련을 통한 차등 프라이버시 기술과 같은 머신러닝을 함께 활용하여 재식별화를 막기 위한 기술의 연구가 활발히 이루어지고 있는 추세이다[9].

Ⅲ. 결론

본논문에서는 비식별화된 데이터들의 재식별화 동향을 분석하였다. 다른 데이터들을 종합하여 유의미한 정보 및 개인을 식별하는 것이 가능하기 때문에 비식별화 기술을 지속적인 연구와 머신러닝과 같은 기술을 함께 활용하는 연구가 필요하다.

ACKNOWLEDGMENT

본 연구는 2021 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1A4A2001810)

참 고 문 헌

[1] Kim Hai Won, “A study on institutional problems and improvement directions of personal information de-identification to utilize big data,” Aug, 2022

[2] 양현철, 김자영, 김진철, 김배현, 신신애, “A guide to the use of personal information de-identification technology for the use of big data,” CISP, 2015.

[3] Simon L. Garfinkel, “De-Identification of Personal Information,” NIST, 2015. (http://dx.doi.org/10.6028/NIST.IR.8053)

[4] “Re-identification risk analysis (2023),” Retrieved Jan. 1. 2024, from https://cloud.google.com/dlp/docs/concepts-risk-analysis

[5] Arvind Narayanan, Vitaly Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” IEEE Symposium on Security and Privacy, pp.111-125, 2008 (https://doi.org/10.1109/SP.2008.33)

[6] Ynes-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen and Vincent D. Blondel, “Unique in the Crowd: The privacy bounds of human mobilty,” Scientific Reports 3, Article 1376, 2013 (https://www.nature.com/articles/srep01376;)

[7] Ma, C.Y.T.; Yau, D.K.Y.; Yip, N.K.; Rao, N.S.V., "Privacy Vulnerability of Published Anonymous Mobility Traces," Networking, IEEE/ACM Transactions on, vol.21, no.3, pp.720-733, June. 2013 (https://doi.org/10.1145/1859995.1860017)

[8] Anthony Tockar, “Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset (2014),” Retrieved Dec, 30, 2023, from https://agkn.wordpress.com/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

[9] Yuichi Sei, Hiroshi Okumura, Akihiko Ohsuga, “Re-Identification in Differentially Private Incomplete Datasets,” IEEE Open Journal of the Computer Society, vol.3, pp.62-72, 2022. (https://doi.org/10.1109/OJCS.2022.3175999)