

정형데이터와 비정형데이터를 동시에 고려한 가명·익명 처리 연구

성민경, 이진규, 한주연, 유수훈*, 김동례*, 조민용**, 김태훈**

한국정보통신기술협회, (주)이지서티*, (주)딥핑소스**

[mksung, jklee, hanjy]@tta.or.kr, [shyoo, drkim]@easycerti.com, [minyong.cho, pete.kim]@deepingsource.io

A study on pseudonymization/anonymization considering both structured and unstructured data

Sung Min Kyoung, Lee Jin Kyu, Han Ju Yeun, Yoo Soo Hoon*, Kim Dong Rye*, Min Yonh Cho**, Pete Kim**

Telecommunications Technology Association, EASYCERTI Co., Ltd.*, Deeping Source Inc.**

요약

데이터 중심의 사회가 구현됨에 따라 데이터의 형태, 양, 이용 방법이 점차 다양해지고 있으며, 데이터에 포함된 개인정보도 다양한 형태로 나타나고 있다. 이와 같이 다양한 방식으로 존재하는 개인정보에 대한 보호 기술 역시 데이터가 활용되는 환경에 맞게 발전해 나가야 한다. 본 논문에서는 사진에 포함된 이미지 데이터와 메타 데이터(시간/위치정보)를 동시에 고려한 가명·익명처리 방법론을 제시한다.

I. 서론

데이터 3법 개정으로 데이터 활용에 대한 기반이 마련되었으며, 데이터 가명·익명처리에 대한 관심도 높아지고 있다[1,2]. 데이터 가명·익명처리란 데이터에서 개인식별 정보가 포함된 부분을 삭제 또는 변경하여 해당 데이터만으로 개인을 식별할 수 없게 만든 기술 행위를 뜻한다[3]. 데이터 가명·익명 대상은 주로 정형데이터와 비정형데이터로 나뉜다. 정형데이터는 미리 정해 놓은 형식과 구조에 따라 저장되도록 구성된 데이터로, 주로 관계형 데이터베이스(RDB: Relational Database)로 표현한다. 비정형데이터는 정의된 구조가 없이 정형화되지 않은 데이터로 동영상, 오디오, 사진, 문서 등이 있다[4]. 데이터 가명·익명처리는 그동안 주로 정형데이터를 대상으로 하였으나 최근에는 비정형데이터를 대상으로 한 기술 또한 개발되고 있다. 정형데이터 가명·익명처리는 삭제, 마스킹, 총계처리, 라운드 등의 기술이 사용되며, 비정형데이터 가명·익명처리 기술은 이미지 데이터를 기준으로 이미지 필터링(블러링), 이미지 암호화, 얼굴합성, 인페인팅 등이 있다[5]. 이처럼 정형데이터와 비정형데이터를 가명·익명처리하는 기술이 각각 존재하지만, 이를 동시에 고려하여 개인정보 노출을 방지하는 방법론은 아직 널리 알려져 있지 않다. 예를 들어, 스마트폰으로 촬영한 사진은 이미지(비정형)와 촬영 장소, 시간 등 메타 데이터(정형)가 함께 포함되어 있어 두 가지 정보를 동시에 고려하여 가명·익명 처리해야 개인정보 노출을 방지할 수 있다.

본 논문에서는 메타 데이터(위치정보, 시간정보)를 포함한 이미지 데이터를 가명·익명 처리하는 방법론을 제시하고, 해당 방법론을 적용한 결과와 정형/비정형 가명·익명처리 기술을 따로 적용한 결과와의 차이점을 분석한다.

II. 본론

1. 기존 연구

정형데이터 익명처리는 K-익명성, L-다양성, T-근접성, 차분프라이버시 등의 기술이 활용되며[6,7,8,9], 가명처리는 '가명정보 처리 가이드라인'에 따라 삭제기술, 통계도구 활용, 일반화(범주화) 기술, 암호화, 무작위화 기술 등이 활용된다[3]. 익명처리의 경우 K, L, T 등 해당 모델을 만족시킬 때까지 데이터를 삭제/수정하며, 가명처리의 경우 유용성 보장을 위해 전문가 적정성

검토를 통한 데이터 삭제/수정 등의 범위를 정한다.

비정형데이터 중 영상/이미지에 대한 익명처리는 2단계로 구성된다. 첫 번째 단계는 얼굴, 차량번호판 등 비식별 대상 영역 탐지기술로 영상/이미지에서 얼굴 영역을 검출하는 기술이다. 특히, 최근 딥러닝 기술의 발달로 탐지 수준은 점차 높아지고 있다. 두 번째 단계는 탐지된 영역을 변형하여 익명처리 하는 기술로 이미지 필터링(Image Filtering)과 이미지 암호화(Image Encryption)가 널리 사용된다. 이미지 필터링은 영상의 각 프레임에 필터를 적용하여 이미지를 변형하는 기술로, 블러(blur)필터와 픽셀레이트(pixelate)필터를 주로 활용한다. 이미지 암호화는 원본 이미지의 전체 또는 일부를 암호화하는 기술로, 복호화가 가능한 경우 가명처리 기술로 분류된다.

2. 제안 방법론

메타데이터(위치정보, 시간정보)와 이미지 데이터를 함께 가명·익명처리 하는 제안 방법은 다음과 같은 절차로 구성된다.

- ① 이미지 데이터에서 개인정보(본 논문에서는 사람 얼굴, 차량 번호판)를 식별할 수 없게 처리
- ② 메타 데이터(본 논문에서는 위치정보, 시간정보)를 정해진 규칙(본 논문에서는 K-익명성을 따르며, K 값은 5로 설정)에 따라 가명·익명 처리
- ③ 동일한 또는 유사한 위치정보와 시간정보를 가지는 데이터는 추가로 가명·익명 처리
 - 위치정보가 같을 경우 시간정보에 대해 가명·익명처리를 한 단계 더 수행
 - 시간정보가 같을 경우 위치정보에 대해 가명·익명처리를 한 단계 더 수행
 - 위치정보와 시간정보가 같을 경우 위치정보와 시간정보에 대해 각각 가명·익명처리를 한 단계 더 수행

위 과정에서 ①과정은 이미지 데이터의 가명·익명처리, ②과정은 메타 데이터 가명·익명처리이며, ③과정에 해당하는 작업이 본 논문에서 추가로 제안하는 방법론이다.

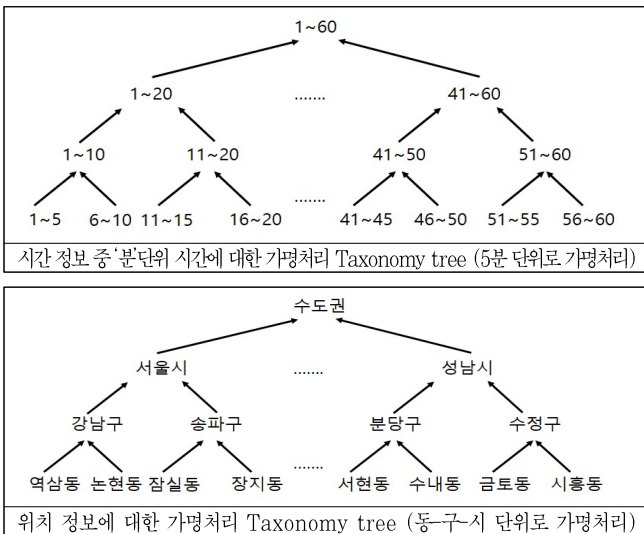
3. 실험수행

3-1 수집 데이터

실험을 위해 1,000장의 사진을 직접 촬영하여 수집하였으며, 사진은 사람의 얼굴 또는 차량번호판을 포함하고 있다. 수집 기간은 21년7월~22년4월이며, 수집 위치는 서울시, 성남시, 수원시, 용인시, 하남시로 한정하였다. 사진에 포함된 사람 얼굴과 차량번호판은 비정형데이터의 개인정보로, 시간정보와 위치정보는 정형데이터의 개인정보로 정의하였다.

3-2 가명·익명처리

비정형데이터인 이미지의 경우 얼굴과 차량번호판을 검출하여 모자이크 또는 블러링 처리 하였으며, 정형데이터는 그림1과 같은 트리구조를 활용하여 ‘문자 데이터 범주화’ 기법을 이용하여 가명 처리했다.



[그림1] 시간과 위치 정보에 대한 Taxonomy Tree

3-3 실험결과

수집한 사진 데이터에 대해 2장에서 제시한 ①, ② 절차만 수행할 경우 정형/비정형 데이터를 각각 처리한 결과가 되며, ③의 절차를 추가로 적용하는 경우 정형/비정형 데이터를 함께 고려한 가명·익명처리 결과물이 된다. ③의 절차를 적용하면 그림1의 Taxonomy 트리에서 추가로 문자 데이터 범주화를 하여야 하므로, 원본 데이터의 값보다 더욱 상위 개념으로 치환된다. 실험수행 결과 1,000장의 데이터에 대해 ③의 과정을 추가했을 때, 420장의 사진의 시간 정보, 378장의 위치정보가 상위 개념으로 치환되었다. 상위 개념으로 치환은 시간/위치정보가 동일 또는 유사한 경우 빈번하게 발생하였다.

III. 결론

본 논문은 사진의 메타 데이터와 이미지 데이터를 이용하여 정형/비정형 데이터를 동시에 고려한 가명·익명처리 방법론을 제안하였다. 데이터의 형태와 분량이 점차 늘어남에 따라 데이터에 포함된 개인정보도 증가하고 있으며, 이를 효과적으로 보호할 방안도 지속적으로 연구해야 할 것이다. 향후 연구로는 위치 연속추적을 통한 트래킹, 이미지로 파악할 수 있는 위치 정보(랜드마크 등), 시간정보(이미지에 찍힌 시계 등) 등을 고려한 개인정보 보호 기술 개발이 있다. 또한, 가명·익명 처리된 데이터의 유용성을 검증하기 위한 연구로 원본 데이터로 학습한 결과와 가명·익명 처리된 데이터로 학습한 결과의 비교 및 학습 결과의 오차 줄이기 등이 있을 것이다.



[그림2] 이미지 데이터를 익명처리한 결과물 (실험에 사용된 모든 이미지는 개인의 동의를 받은 후 사용함)

ACKNOWLEDGMENT

이 논문은 2021년 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00634 ‘대용량 정형 데이터 대상 개인정보 가명·익명처리 자동화 및 안정성 검증 기술개발’, No.2021-0-00393 ‘영상 등 멀티미디어 데이터의 온전한 AI 학습활용을 보장하는 복원 불가능 개인정보식별정보 익명처리 핵심 기술 개발’)

참 고 문 헌

- [1] 개인정보보호위원회, “개인정보 보호·활용기술 R&D 로드맵,” 2021.
- [2] 과학기술정보통신부, “데이터보호 핵심기술 개발 전략,” 제7차 데이터 특별위원회 심의안건 제2호, 2021.
- [3] 개인정보보호위원회, “가명정보 처리 가이드라인,” 2021.
- [4] 한국정보통신기술협회, “정보통신용어사전,” 2022.
- [5] K-ICT 빅데이터센터, “영상 데이터 익명화 기술 및 평가방안,” 2019
- [6] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 557-570, 2002.
- [7] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkatasubramanian, ℓ -diversity: Privacy beyond k-anonymity, in Proceedings of the IEEE International Conference on Data Engineering, 2006.
- [8] N. Li, T. Li and S. Venkatasubramanian, t-Closeness: Privacy beyond k-anonymity and ℓ -diversity, in Proceedings of the IEEE International Conference on Data Engineering, 2007.
- [9] C. Dwork, Differential Privacy. In: van Tilborg H.C.A., Jajodia S. Encyclopedia of Cryptography and Security. Springer, Boston, MA, 2011.