데이터 결합 시 재식별에 취약한 칼럼 패턴 분석

이강원, 성민경, 한주연
한국정보통신기술협회
[blong116, mksung, hanjy]@tta.or.kr

An analysis of column pattern vulnerable to re-identification due to combining datasets

Lee Kangwon, Sung Min Kyoung, Han Ju Yeun
Telecommunications Technology Association

요 약

데이터 3법 개정을 통해 정보주체의 동의없이 가명정보를 활용할 수 있는 기반이 마련되었으며, 안전한 가명정보 결합 및 활용 활성화를 위해 전문가 집단을 활용한 적정성 검토를 수행한다. 그러나 데이터 결합 시 발생할 수 있는 재식별 위험을 정성적으로 판단할 경우 예상치 못한 정보 조합이나 데이터의 특수성으로 인해 검토가 불완전할 수 있다. 본 논문에서는 재식별 위험을 최소화하기 위한 정량적 방법론 연구 중 하나로 데이터 결합 시 재식별에 취약할 수 있는 칼럼 조합 패턴에 대해 분석한다.

Ⅰ. 서 론

데이터 3법이 개정됨에 따라 가명처리된 개인정보 즉, 가명정보를 정보주체의 동의없이 활용할 수 있는 기반이 마련되었다. 가명정보의 안전한 제공 및 활용을 위해 개인정보보호위원회에서는 가명정보 처리 가이드라인[1]을 제공하고 있으며, 가명정보 결합 및 반출 절차 구성을 통해 가명정보 결합 및 활용을 활성화하여 이종 분야 간 데이터에 내재된 잠재력을 활용하기 위한 발판을 마련하고 있다. 또한, 가명정보의 안전한 결합 및 활용을 위해 결합전문기관에서 결합된 가명정보의 적정성을 검토하고, 필요한 경우 추가처리를 수행하도록 하고 있다. 그러나 전문가의 경험이나 지식을 활용하여 정성적으로 적정성을 판단하고 있어 데이터 결합으로 인한 예상치 못한 재식별 사례[2-4]가 발생하거나, {대전 거주, 어업 종사자, ...}와 같이 도서산간으로 구분되는 지리적 특수성과 연관성이 낮은 직업군과의 결합으로 인해 특정 개인이 식별 위험성에 노출될 수 있는 상황에 대한 판단이 어려울 수 있다.

이러한 문제를 해소하기 위해 정량적인 방법론[5-10]의 적용을 고려해볼 수 있지만, 주로 최대값/최소값 또는 빈도수를 활용한 이상치 즉, 특이정보 판단을 통한 재식별 위험 감소에 대한 방법론으로 데이터 결합 시 데이터의 특수성으로 인해 발생하는 상황에 대한 식별 위험성 검토에 그대로 적용하기에는 적합하지 않다. 본 연구는 데이터 결합 시 발생할 수 있는 식별 위험성 검토를 위해 데이터 결합 시 재식별에 취약할 수 있는 칼럼 조합에 대한 패턴을 분석하고자 한다.

Ⅱ. 개인정보의 종류 및 유형

개인정보보호위원회에서는 개인정보 포털을 통해 표 1과 같이 개인정보의 종류 및 유형을 분류하고 있다[11]. 개인정보의 종류는 유형에 따라 인적사항, 신체적 정보, 정신적 정보, 사회적 정보, 재산적 정보, 기타 정보로 분류되어 있다.

구분	유형	내용
인적사항	일반정보	성명, 주민등록번호, 주소, 연락처, 성별 등
	가족정보	가족관계 및 가족구성원 정보 등
신체적 정보	신체정보	얼굴, 홍채, 음성, 지문, 키, 몸무게 등
	의료·건강 정보	건강상태, 진료기록, 신체장애, 병력, 검사정보 등
정신적 정보	기호·성향 정보	물품구매내역, 웹사이트 검색내역 등
	내면의 비밀 정보	사상, 신조, 종교, 가치관, 정당 및 노조 활동 등
사회적 정보	교육정보	학력, 성적, 생활기록부, 건강기록부 등
	병역정보	병역여부, 제대유형, 근무부대 등
	근로정보	직장, 근무처, 근로경력, 직무평가기록 등
	법적정보	전과·범죄 기록, 재판 기록, 과태료 납부내역 등
재산적 정보	소득정보	봉급액, 보너스 및 수수료, 이자소득, 사업소득 등
	신용정보	대출 및 담보설정 내역, 신용평가 정보 등
	부동산 정보	소유주택, 토지, 자동차, 상점 및 건물 등
	기타 수익 정보	보험, 가입현황, 휴가, 병가 등
기타 정보	통신정보	E-mail 주소, 전화통화내역, 로그, 쿠키 등
	위치정보	GPS 및 휴대폰에 의한 개인의 위치정보
	습관 및 취미정보	흡연여부, 음주량, 여가활동, 도박성성향 등

<표 1. 개인정보의 종류 및 유형>

인적사항과 같이 정보 자체만으로도 특정 개인을 식별할 수 있는 민감한 정보가 존재하며, 신체적 정보, 정신적 정보와 같이 해당 정보만으로는 식별 위험성이 낮지만 인적사항과 같이 민감한 정보와 결합되면 정보주체를 식별할 가능성이 매우 높아지는 경우가 존재할 수 있다. 또한, 사회적 정보, 재산적 정보, 기타 정보는 다른 정보와 결합된 경우 적정한 수준의 가명처리가 적용되었다면 개인에 대한 식별 위험에 비교적 안전할 수 있지만, {대전 거주, 어업 종사자, ...}와 같이 각 정보 자체로는 특수한 성격의 정보가 아니지만 데이터의 특수성으로 인해 두 정보가 결합된다면 특정 개인의 식별 위험성이 매우 높아질 수 있다. 이처럼 어떤 정보들이 결합되는지에 따라 식별 위험성 수준이 다르며, 특수한 성격을 가진 정보로 인해 전문가 집단을 통한 정성적인 적정성 검토 시 일부 위험 요소의 검토가 누락될 가능성이 존재할 수 있다. 이와 같이 다양한 정보 즉, 칼럼 조합에서 발생할 수 있는 위험 요소에 대해 고려한 적정성 검토가 이루어져야 한다.

Ⅲ. 재식별에 취약한 칼럼 패턴 분석

본 논문에서는 개인정보보호위원회에서 분류한 개인정보의 종류 및 유형을 참고하여 데이터 즉, 칼럼 결합 시 특정 개인의 재식별 위험이 존재할 수 있는 칼럼 조합 패턴을 표 2과 같이 분석하였다.

칼럼 A	칼럼 B	식별 위험 요소
근로정보	주소정보	주소정보와 연관성이 매우 낮은 직업군이 있을 수 있음
	부동산정보	근로소득으로 취득할 수 없는 수준의 부동산 소유 정보가 있을 수 있음
	의료정보	의료정보를 통해 해당 직업군의 업무수행 가능여부를 판단할 수 있음
	신체정보	신체정보를 통해 해당 직업군의 업무수행 가능여부를 판단할 수 있음
	출생지정보	특정한 직업군을 가질 수 없는 출생지정보가 있을 수 있음
소득정보	주소정보	근로소득으로 유지하기 어려운 주소정보가 있을 수 있음
	부동산정보	근로소득으로 취득할 수 없는 수준의 부동산 소유 정보가 있을 수 있음
신체정보	병역정보	신체정보가 군 복무 수행에 지장이 있으나 병역 이력이 있을 수 있음
	취미정보	취미활동에 제약이 있음을 알 수 있는 신체정보가 있을 수 있음
	보험관련정보	특정 신체정보로 인해 가입하기 어려운 유형의 보험 가입정보가 있을 수 있음
의료정보	병역정보	의료정보를 통해 군 복무 수행에 어려움이 있음을 알 수 있음
	취미정보	취미활동에 제약이 있음을 알 수 있는 의료정보가 있을 수 있음
학력정보	법적정보	전과 및 범죄기록이 있는 경우 학력정보를 통한 미성년자 여부 판단으로 인해 식별 위험에 노출될 수 있음
	흡연 및 음주량 정보	흡연여부 및 음주량 정보가 있는 경우 학력정보를 통한 미성년자 여부 판단으로 인해 식별 위험에 노출될 수 있음
	위치정보	학력정보를 통한 미성년자 여부 판단이 가능한 경우 위치정보를 통해 생활패턴이 유추될 수 있으며 범죄 위험에 노출될 수 있음
위치정보	병역정보	군 복무 수행 중인 경우 위치정보를 통해 지도 상에 표시되지 않은 군사지역이 노출될 수 있음
	물품구매내역	물품구매내역과 위치정보를 통해 동거인여부와 주거지 정보가 유추될 수 있음
	E-mail 정보	사내메일과 같은 E-mail 정보로 특정된 직업군과 위치정보를 통해 생활패턴이 유추될 수 있음

<표 2. 데이터 결합으로 인해 식별 위험 요소가 존재하는 칼럼 조합 패턴>

결합 전 데이터셋 A, B에 포함될 수 있는 칼럼은 각각 칼럼 A, 칼럼 B에 나타내었으며, 칼럼 A, B 조합 패턴에서 발생할 수 있는 상황은 식별 위험 요소에 나타내었다. 칼럼 조합 패턴 분석을 통해 근로정보, 소득정보, 신체정보, 의료정보, 학력정보, 위치정보가 다양한 유형의 정보와 결합되었을 때 식별 위험에 노출될 수 있음을 알 수 있었다. 주소정보 및 부동산정보는 근로정보와 결합되었을 때 상식적인 수준을 벗어난 경우 식별 위험성에 노출될 수 있으며, 소득정보와도 이와 유사한 특징으로 인해 식별 위험성에 노출될 수 있는 칼럼 패턴이다.

병역정보 및 취미정보는 신체정보와 결합되었을 때 신체적 활동의 제약 여부에 따라 식별 위험성에 노출될 수 있는 칼럼 패턴이다. 또한, 의료정보를 통해 신체적 활동에 제약이 있음을 유추할 수 있다면 병역정보 및 취미정보 값에 따라 식별 위험에 노출될 가능성이 존재한다.

법적정보, 흡연 및 음주량 정보와 같은 기타 정보, 위치정보는 다른 정보와 결합되더라도 식별 위험에 노출될 가능성이 높은 정보는 아니지만 학력정보와 결합되는 경우 재식별에 취약할 수 있는 칼럼 조합 패턴이다. 학력정보를 통해 미성년자임을 알 수 있는 경우 전과 및 범죄기록 등과 같은 법적정보 또는 흡연이나 음주량 정보를 확인할 수 있다면 이를 통해 정보주체의 식별 위험성이 상당히 높아질 수 있다. 또한, 미성년자임을 알 수 있는 상태에서 위치정보가 제공된다면 특정 개인에 대한 식별 위험은 다소 낮을 순 있으나 범죄 대상이 될 가능성에 노출될 수 있다.

병역정보, 물품구매내역, E-mail 정보는 위치정보와 결합되는 경우 식별 위험에 노출될 수 있다. 위치가 노출되지 않아야 하는 군사지역은 일반적으로 지도 상에 표시되지 않지만, 병역정보와 위치정보의 결합으로 인해 군사지역이 노출될 가능성이 매우 높다. 물품구매내역은 정보값에 따라 동거인여부, 성별, 연령대 등을 유추할 수 있으며 위치정보와 결합되는 경우 범죄의 대상이 될 수 있다. E-mail 정보와 위치정보의 결합은 식별 위험성이 높은 칼럼 조합 패턴은 아니지만, 사내메일 정보를 통해 특정 직업군으로 인식할 수 있다면 생활패턴 유추를 통한 식별 위험에 노출될 가능성이 존재한다.

Ⅳ. 결 론

본 논문에서는 데이터 결합 시 재식별을 방지하고 안전성을 확보하기 위한 정량적인 방법론을 연구하기 전 선행 연구로 데이터 결합 시 식별 위험에 노출될 수 있는 칼럼 조합 패턴에 대한 분석을 진행하였다. 분석된 칼럼 조합 패턴을 활용한다면 전문가의 경험이나 지식 등을 통한 적정성 검토 시 누락될 수 있는 경우를 최소화할 수 있을 것이다. 향후 연구를 통해 분석한 칼럼 조합 패턴을 수치화하고 이를 통해 재식별 위험성을 최소화하기 위한 정량적인 방법론 연구를 진행할 예정이다.

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)

참 고 문 헌

[1] 개인정보보호위원회, "가명정보 처리 가이드라인," 2024.

[2] Sweeney, L. “k-anonymity: A model for protecting privacy,” International journal of uncertainty, fuzziness and knowledge -based systems, Vol. 10, No. 1, pp. 557-570, 2002.

[3] Naranyanan, A., & Shmatikov, V. “Robust de-anonymization of large datasets,” Proceedings of the 2008 IEEE Symposium on Security and Privacy, May. 2008.

[4] Cohen, A. “Attacks on Deidentification's Defenses,” 31st USENIX Security Symposium, pp 1469-1486, 2022.

[5] Grubbs, F. E. “Sample criteria for testing outlying observations,” The Annals of Mathematical Statistics, Vol. 21, No. 1, pp. 27-58, 1950.

[6] S. Walfish, "A review of statistical outlier methods", Pharmaceutical Technol., Vol. 30, No. 11, pp. 1-5, 2006.

[7] A. Taha, and O. M. Hegazy, “A Proposed Outliers Identification Algorithm for Categorical Data Sets,” 2010 the 7th international conference on informatics and systems (INFOS). IEEE, pp. 1-5, 2010

[8] X. Zhao, and J. Liang, and F. Cao, “A simple and effective outlier detection algorithm for categorical data,” International Journal of Machine Learning and Cybernetics, 5, pp. 469-477, 2014

[9] 이강원, 성민경, 한주연, “데이터 재식별 가능성 감소를 위한 행 단위 방법론 연구,” 한국통신학회 동계학술발표회, 2023.02.

[10] 이강원, 성민경, 한주연, “개인정보 식별위험성 감소를 위한 특이정보 판단 방법론 연구,” 한국통신학회 추계학술발표회, 2023.11.

[11] 개인정보 포털, “개인정보의 종류”, https://www.privacy.go.kr/front/contents/cntntsView.do?contsNo=35