한국빅데이터학회지
제5권 제2호, 2020, pp. 43-52 https://doi.org/10.36498/kbigdt.2020.5.2.43

재식별 시간에 기반한 k-익명성 프라이버시 모델에서의 k값에 대한 연구

Analysis of k Value from k-anonymity Model Based on Re-identification Time

김채운 · 오준형 · 이경호†

고려대학교 정보보호대학원[1]

요 약

빅데이터 활용 기술의 발전으로 데이터의 저장 및 공유가 늘어나면서 그에 따른 프라이버시 침해가 일어나게 되었다. 이 문제를 해결하기 위해 비식별 기술이 도입되었지만 비식별된 데이터에 대해서도 재식별이 가능하다는 것이 여러 차례 증명되었다. 재식별 가능성이 존재하기 때문에 완전히 안전할 수 없지만 그럼에도 불구하고 충분한 비식별처리가 이루어져야 하는데, 현재 법령이나 규제는 어느 정도로 비식별 처리를 해야 하는지 정량적으로 규정하고 있지 않다. 본 논문에서는 재식별 작업을 할 때 소요되는 시간을 고려하여 적절한 비식별 기준을 제시하려고 한다. 다양한 비식별 평가 모델 중에서 k-익명성 모델에 대해 집중적으로 연구하였으며 어느 정도의 k값이 적절한 지 판단하였다. 본 연구의 결과를 일반화시킬 수 있다면 각종 법률 및 규제에서 적절한 비식별 강도를 규정하는 데 사용할 수 있을 것이다.

■ 중심어 : 데이터 비식별화, 데이터 프라이버시, 데이터 보안

Abstract

With the development of data technology, storing and sharing of data has increased, resulting in privacy invasion. Although de-identification technology has been introduced to solve this problem, it has been proved many times that identifying individuals using de-identified data is possible. Even if it cannot be completely safe, sufficient de-identification is necessary. But current laws and regulations do not quantitatively specify the degree of how much de-identification should be performed. In this paper, we propose an appropriate de-identification criterion considering the time required for re-identification. We focused on the case of using the k-anonymity model among various privacy models. We analyzed the time taken to re-identify data according to the change in the k value. We used a re-identification method based on linkability. As a result of the analysis, we determined which k value is appropriate. If the generalized model can be developed by results of this paper, the model can be used to define the appropriate level of de-identification in various laws and regulations.

■ Keyword : Data de-identification, Data Privacy, Data security

2020년 11월 20일 접수; 2020년 12월 10일 게재 확정.

† 교신저자 (kevinlee@korea.ac.kr)

Ⅰ. 서 론

정보의 공유와 활용이 늘어남에 따라 수집 및 저장되는 데이터 또한 증가하고 있으며 수집된 데이터 간의 조합과 연결을 통해 새로운 정보를 도출할 수 있다. 의료 분야를 비롯한 많은 분야에서 빅데이터를 구축하고 유용한 가치를 창출하고 있다. 그러나 대규모로 데이터를 수집하고 사용하면서 데이터에 포함된 개인정보로 인한 프라이버시 침해 문제가 발생하였다. 소규모의 데이터를 다룰 때와는 달리 빅데이터의 시대로 변화하면서, 데이터의 개인정보를 일일이 삭제하는 것은 사실상 불가능하기 때문에 프라이버시 침해가 더욱 중요한 문제로 주목받게 되었다.[4][11][20].

데이터 비식별화는 이러한 문제를 해결하기 위해 사용된다[6]. 그러나 데이터를 비식별 처리한다고 해도 완전히 안전할 수는 없으며 여전히 재식별의 위험을 안고 있다[9]. 개인정보를 식별할 수 없게 비식별화된 데이터라도 많은 양이 모이면 조합을 통해 다시 식별될 수 있다. 기존의 여러 연구에서 재식별이 충분히 가능하다는 것이 증명되었기 때문에 비식별화를 한다고 해서 완전히 안심할 수는 없다. 그렇지만 완전하지 못한 방법이라 하더라도 그 효용성은 분명히 있기 때문에 충분한 비식별처리가 이루어져야 하지만, 어느 정도가 '충분한' 수준인지에 대해 명확하게 기술하고 있는 규정은 현재 없다.

EU GDPR[1]에서는 비식별화와 익명화의 개념에 대해서 설명하고 그 방법에 대해 설명하고 있다. k-익명성 모델 이외에도 다양한 비식별화 기법을 소개하고 있지만 각 모델에 대해 어느 정도로 비식별화해야 충분히 비식별화 되었다고 말할 수 있는지에 대한 수치화 된 기준을 제시하고 있지는 않다. 국내에서는 2016년 '개인정보 비식별 조치 가이드라인'이 만들어졌는데 비식별 조치를 어떤 식으로 해야 하는지 대략적인 가이드만 제시하고 있으며 구체적인 방법은 규정하고 있지 않다. 또한 프라이버시 모델은 k-익명성 모델만을 언급하고 있고 k값의 기준에 대한 명확한 가이드라인은 기술하고 있지 않다.

본 논문의 목적은 충분히 비식별화가 되었다고 말할 수 있는 적정선을 판단하는 방법을 제안하는 것이다. 서로 다른 수준의 비식별 처리가 된 데이터를 재식별하는 데 걸리는 시간을 각각 측정하고, 소모되는 시간적 비용과 재식별 가능성이 적절한 균형을 이루는 지점을 찾는다. 실험을 위해 미국 뉴욕 주에서 공개한 병원 이용 데이터인 Hospital Inpatient Discharges (SPARCS De-Identified): 2017[12] 데이터셋을 사용하였다.

Ⅱ. 관련 연구

기존의 연구들을 통해 데이터가 비식별화 되었음에도 불구하고 다시 식별해내는 것이 충분히 가능하다는 것이 증명되었다.

[17]의 연구에서는 기계학습 알고리즘을 통해 불완전한 데이터셋에서도 개인을 식별해낼 수 있는 모델을 만들었다. 해당 모델은 15개의 속성만 알아내면 익명화된 데이터에서도 99.98%의 정확도로 재식별해내는 것이 가능하다.

[19]의 연구에서는 미국 메사추세츠 주 보험위원회가 수집한 약 13500명의 공무원과 그 가족에 대한 환자 데이터와 캠브리지 시의 유권자 등록 명부 데이터를 구매하여 연결 공격을 사용해 당시 주지사였던 William Weld의 레코드를 식별하는 것에 성공하였다.

[15]의 연구에서는 전자의료기록(Electronic Medical Record, EMR)에 포함된 진단 코드를 통해 전체 2762개의 레코드에서 96% 이상의 개인을 식별하는 데 성공하였다. 이 연구에서는 비식별화를 통해서는 임상적으로 의미 있는 정보를 유지하면서 개인정보를 충분히 보호할 수 없음을 보였다.

[7]에서는 New York City 택시 데이터에서 완전히 가명화된 택시 ID를 통해 거의 모든 택시를 식별하는 데 성공하였다.

[3]에서는 Gaussian and randomized skew를 사용하여 익명화된 지도상의 좌표 데이터를 재식별하였다. 좌표 데이터를 비식별화하는 데 사용되는non-deterministic blurring algorithm의 취약점을 이용하여 원본 데이터에서 여러 개의 익명화된 데이터셋을 생성한 후 결과를 평균화하여 재식별된 지점과 원래 위치의 거리를 줄였다.

[5]에서는 호주 연방 보건부가 공개한 비식별처리 된 의료비 청구 기록을 재식별하였다. 해당 데이터셋은 호주 인구의 10%인 약 290만명의 의료비 청구 기록으로 각 레코드에는 1984~2014년까지의 공적으로 지불된 모든 의료비, 의약품 청구서가 포함되어 있다. 이 연구에서는 조산사 등의 ID를 복호화, 데이터셋의 암호화되지 않은 부분과 개인에 대해 알려진 정보를 연결하여 재식별하였다.

Ⅲ. 비식별화와 프라이버시 모델

3.1 데이터 비식별화

비식별화란 데이터를 공유하기 전에 익명화하는 과정으로 데이터 주체의 프라이버시를 보호하면서 데이터를 공유하기 위해 널리 사용되는 방법이다[17]. 미국 국가표준기술연구소에서 발표한 NIST SP 800-188[10]에서는 비식별화를 “식별 데이터셋과 데이터 주체 간의 연관성을 제거하는 프로세스에 대한 일반적인 용어”로 정의한다. 이와 유사하게 국제표준화기구에서 작성한 ISO/TS 25237-2008[2]에서는 비식별화를 “식별되는 데이터셋과 데이터 주체 간의 연관성을 줄이는 모든 프로세스에 대한 일반적인 용어” 로 정의하고 있다. 또한 EU GDPR[1]의 WP29 권고사항에서는 특정 데이터가 하나의 개인으로 특정될 가능성(Singling out), 특정 데이터가 하나의 개인과 연결될 가능성(linkability), 특정 데이터로부터 특정 개인을 추론할 수 있을 가능성(inference)의 세 가지중 일부 또는 전부가 제거되었다면 데이터가 비식별화 되었다고 한다[13].

비식별화는 데이터의 개인식별정보에 대해 수행된다. 개인식별정보(Personally identifiable information, PII)란 개인을 직접 식별하거나 유추를 통해 식별해 낼 가능성이 있는 모든 정보를 말한다. 개인정보보호법 제 1장 제 2조에서는 개인식별정보를 “해당 정보만으로는 식별이 불가능하더라도 다른 정보와의 결합을 통해 ‘쉽게’ 식별할 수 있는 것을 포함한다”고 정의하고 있다. 개인식별정보에는 크게 직접 식별자, 준식별자, 민감 정보가 있다. 직접 식별자는 그 이름대로 해당 속성 하나만으로도 개인을 직접적으로 식별할 수 있는 정보이다. 준식별자는 그 자체만으로는 식별이 불가능하지만 다른 속성과의 결합을 통해 개인을 식별할 수 있는 정보이다. 민감 정보는 병명, 카드 사용 금액 등 개인의 프라이버시가 공개될 수 있는 정보를 가지고 있는 속성이다.

일반적으로 비식별화에 사용되는 방법으로 크게 마스킹, 일반화 및 억제 기법 등이 있다[8]. 데이터 마스킹은 직접 식별자를 조작하는 것으로 일반적으로 데이터셋에서 직접 식별자를 제거하거나 임의의 값 또는 가명으로 대체한다. 일반화는 데이터 값을 보편적인 범위로 변환하여 특정 개인을 식별하지 못하게 하여 데이터의 정밀도를 감소시킨다. 억제는 데이터에서 값을 제거하여 식별을 어렵게 한다.

3.2 프라이버시 보호 모델

3.2.1 k-익명성 모델

k-익명성 모델은 1998년 Samarati와 Sweeney

Age	Gender	Zip code	Diagnosis
30	M	0284*	Diabetes
30	M	0284*	Diabetes
10	F	0285*	Measles
10	F	0285*	Measles
50	M	0286*	Cancer
50	M	0286*	Cancer
50	M	0286*	Cancer

〈그림 1〉 k값이 2인 k-익명화된 데이터의 예시

Age	Gender	Zip code	Diagnosis
40	M	0284*	Hypertension
40	M	0284*	Diabetes
40	M	0284*	Cancer
40	M	0284*	Cancer
50	F	0286*	COPD
50	F	0286*	pneumonia
50	F	0286*	Emphysema

〈그림 2〉 l값이 3인 l-다양화된 데이터의 예시

가 제안한 개념이다[18]. k-익명성 모델에서는 한 동치류(equivalent class, 동일한 준식별자 값을 가지고 있는 레코드의 개수)의 크기가 k 이상이 되도록 일반화하여 특정 개인의 식별을 불가능하게 한다. 동일한 준식별자 값을 가지는 레코드가 적어도 k개 존재하기 때문에 개인이 식별될 확률이 1/k로 낮아지게 된다. Sweeney는 [19]의 연구에서 정보를 가공하여 가능한 많은 사람들에게 매핑될 수 있게 해 후보자의 수가 많을수록 연결이 더 모호하므로 데이터의 익명화 강도를 올려 연결성을 애매하게 하면 연결 공격을 막을 수 있음을 보였다.

3.2.2 l-다양성 모델

l-다양성 모델은 k-익명성 모델의 문제점을 보완하기 위해 2007년 Machanavajjhala가 제안한 프라이버시 모델이다[16]. k-익명성 모델은 공격자가 공격 대상에 배경지식을 가진 경우 재식별 위험이 증가하며 동질성 공격에 취약하다. 즉 k-익명성을 만족시키더라도 민감 정보의 다양성이 적은 경우 민감 정보의 값을 추론할 수 있다. 예를 들어 만약 그림 1에서 zip code가 0284*인 지역에 사는 30세 남성이 있고 현재 질병을 가지고 있다면 높은 확률로 당뇨를 가지고 있을 것이라고 추론할 수 있다. 이런 문제에 대응하기 위해 l-다양성 모델이 제안되었다.

l-다양성 모델에서는 한 동치류에 속하는 레코드들은 적어도 l개의 서로 다른 민감 정보 값을 가져야 한다. 그림 2는 l=3인 3-다양화 된 데이터의 예시이다. 만약 fig. 2에 포함된 zip code가 0284*인 지역에 사는 40대 남성이 있고 질병이 있다는 것을 알고 있다고 해도 최소 3가지의 가능한 후보가 있기 때문에 어떤 질병을 가지고 있는지 쉽게 특정할 수 없다.

3.2.3 t-근접성 모델

t-근접성 모델은 2007년에 제안된 프라이버시 모델로 l-다양성 모델을 더욱 향상시키기 위해 만들어졌다[14]. l-다양성 모델은 각 동치류에서 민감 정보 값의 '다양성'을 보장하지만, 그 사이의 의미적 유사성을 고려하지 않았다는 문제가 있다. 예를 들어 그림 2에서 zip code가 0286*인 지역에 사는 50대 여성이 있을 때 그 사람이 어떤 질병을 가지고 있는지 특정할 수는 없지만 최소한 폐에 관련된 질병이 있다는 것을 알아낼 수 있다.

t-근접성 모델에서는 한 동치류 안에서의 민감 정보의 분포와 전체 데이터에서의 분포와의 거리가 임계값 t 이하인 경우 해당 동치류는 t-근접성을 가졌다고 한다. 모든 동치류가 t-근접성을 가지는 경우 해당 데이터셋은 t-근접성을 충족시킨다.

Ⅳ. 실험 및 결과

4.1 실험의 개요

본 실험에서는 서로 다른 수준의 비식별 처리가 된 데이터를 재식별하는 데 걸리는 시간을 각각 측정하고, 소모되는 시간적 비용과 재식별 가능성이 적절한 균형을 이루는 지점을 찾는다. 실험을 위해 데이터셋을 공통된 열을 가지는 두 개의 데이터셋으로 분리하고, 분리한 두 데이터셋 중 하나에 비식별화 작업을 수행하고 나머지 한 데이터셋을 이용해 매칭을 시도하였다.

비식별화에는 k-익명성 모델을 사용하였다. k-익명성 모델은 간단하고 이해하기 쉽기 때문에 실생활에서 많이 사용되며, 더욱 효과적일 수 있는 프라이버시 모델 중 비전문가가 이해하기가 어렵기 때문에 많이 사용된다. 특히 정책 입안자가 쉽게 이해할 수 있기 때문에 실생활에서는 k-익명성 모델이 많이 적용된다. 데이터 마스킹 및 범주화를 이용하여 2부터 10까지의 k값을 가지는 k-익명화 된 데이터셋을 생성하였다.

본 실험에서는 비식별화된 데이터에 다른 데이터를 연결하여 추가적인 속성을 얻으려고 한다. 연결하는 데이터셋의 각 레코드를 비식별화된 레코드와 매칭하여 연결 가능한 레코드의 개수 n을 세고, 연결의 정확도 1/n을 계산하였다. 이 과정에서 비식별화의 정도가 강할수록 데이터를 연결하는 데에 더 많은 시간이 걸릴 것이라 가정하고, k=2부터 k=10까지의 각 경우에 대해 정확도가 일정 이상으로 매칭되는 경우를 찾는 데 걸리는 시간을 측정하였다.

4.2 데이터 설명

실험을 위해 미국 뉴욕 주 SPARCS(Statewide Planning and Research Cooperative System)에서 제공하는 병원 이용 데이터인 SPARCS De-Identified: 2017[12] 데이터를 사용하였다. 이 데이터셋에는 환자의 특징, 진단, 치료, 치료비 등의 병원 기록 데이터가 포함되며 HIPAA에 따라 보호되는 의료 정보(Protected health information, PHI)가 포함되어 있지 않다. 데이터는 총 34개의 속성을 가지고 있다.

결측값이 포함된 레코드를 삭제하는 과정에서 'Abortion', 'Born weight'열의 데이터는 필요 없는 것으로 판단하여 삭제하였다. 또한 Description 열의 데이터는 각각 앞의 속성에 대한 설명이기 때문에 삭제하였다. 최종적으로 데이터 처리를 통해 1000개의 레코드와 27개의 열을 가진 데이터를 얻었다.

4.3 실험 수행

3가지 케이스를 가정하고 데이터 비식별화 및 재식별을 각각 수행하였다. 케이스 1에서는 비식별된 데이터셋과 연결할 데이터셋이 'age group', 'zip code', 'gender', 'race', 'ethnicity', 'length of stay' 속성을, 케이스 2에서는 'age group', 'gender', 'length of stay', 'type of admission', 'patient disposition' 속성을, 케이스 3에서는 'age group', 'CCS Diagnosis Code', 'APR Severity of Illness Code', 'Payment Typology 1', 'Emergency Department Indicator' 속성을 공통으로 가진다고 가정하였다.

익명화 과정에서 'Age Group' 및 'Length of Stay'열은 데이터 범주화를 사용하여 일반화하였고, 'Zip Code', 'Gender', 'CCS Diagnosis Code' 속성은 데이터 마스킹을 통해 '*'로 값을 대체하였다. 그 외 속성은 각각 2~3단계의 일반화 계층 구조를 구성하여 일반화하였다.

각 케이스에 대해 일반화 방법의 조합을 통해 k=2에서 k=10까지 k-익명화된 데이터 세트를 생성하고 연결되는 데이터셋을 각각 매칭해 연결 가능한 레코드의 개수 n을 세고, 연결의 정확도 1/n을 계산하였다. k=2부터 k=10까지의 각

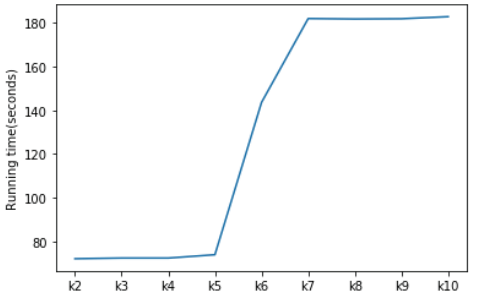
〈표 1〉 데이터셋의 예시 및 속성 별 설명

#	Column name	Example	Data description
1	Hospital Service Area	Hudson Valley	A description of the Health Service Area (HAS) in which the hospital is located.
2	Hospital County	Westchester	A description of the county in which the hospital is located.
3	Operating Certificate Number	5903001	The facility Operating Certificate Number as assigned by NYS Department of Health.
4	Permanent Facility Id	001061	Permanent Facility Identifier
5	Facility Name	Montefiore Mount Vernon Hospital	The name of the facility where services were performed based on the Permanent Facility Identifier (PFI), as maintained by the NYSDOH Division of Health Facility Planning,
6	Age Group	50 to 69	Age in years at time of discharge.
7	Zip Code - 3 digits	105	The first three digits of the patient's zip code.
8	Gender	M	Patient gender.
9	Race	Black//African American	Patient race.
10	Ethnicity	Not Span/Hispanic	The ethnicity of the patient.
11	Length of Stay	3	The total number of patient days at an acute level and/or other than acute care level (excluding leave of absence days)
12	Type of Admission	Emergency	A description of the manner in which the patient was admitted to the health care facility.
13	Patient Disposition	Short-term Hospital	The patient's destination or status upon discharge.
14	Discharge Year	2017	The year (CCYY) of discharge.
15	CCS Diagnosis Code	660	AHRQ Clinical Classification Software (CCS) Diagnosis Category Code.
16	CCS Procedure Code	222	CCS ICD-9 Procedure Category Code.
17	APR DRG Code	280	APR-DRG Classification Code.
18	APR MDC Code	07	All Patient Refined Major Diagnostic Category (APR MDC) Code.
19	APR Severity of Illness Code	Major	APR-DRG Severity of illness Code.
20	APR Risk of Mortality	Major	All Patient Refined Risk of Mortality (APR ROM) Description.
21	APR Medical Surgical Description	Medical	APR-DRG specific classification of Medical, Surgical or Not Applicable.
22	Payment Typology 1	Medicare	A description of the type of payment for this occurrence.
23	Payment Typology 2	Medicare	
24	Payment Typology 3	Medicare	
25	Emergency Department Indicator	N	Emergency Department Indicator is set based on the submitted revenue codes.
26	Total Charges	$49,290.08	Total charges for the discharge.
27	Total Costs	$18,503.26	

경우에 대해 정확도가 일정 이상으로 매칭되는 경우를 찾는 데 걸리는 시간을 측정하였다.

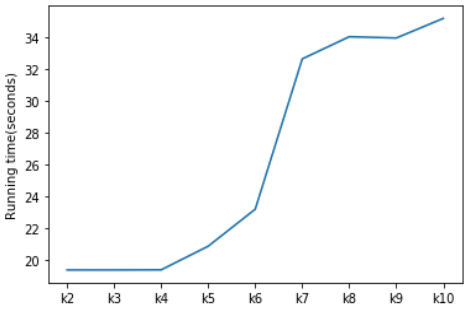
4.4 실험 결과

그림 3은 케이스 1에서 각 경우에 대한 시간 측정 결과를 그래프로 나타낸 것이다. k값이 5일 때와 7일 때 그래프의 기울기가 크게 변한다. k가 5에서 7 사이일 때는 매칭에 걸리는 시간이 증가하는데 7 이후에는 유의미한 차이를 보이지 않는다. 따라서 k=7보다 커지게 되어도 비식별화에 들이는 시간적 비용이나 노력은 늘어나도 연결에 필요한 시간적 비용은 그에 비례하여 늘어나지 않아 큰 의미가 없다고 볼 수 있다. 결론적으로 케이스 1의 경우 k=7일 때 충분히 비식별화 되었다고 볼 수 있다.



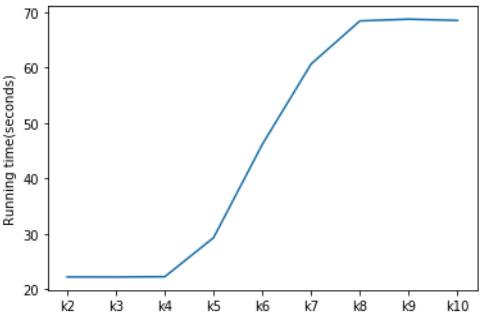
〈그림 3〉 케이스 1의 매칭 시간 측정 그래프

그림 4는 케이스 2에서의 시간 측정 결과를 나타낸 것이다. 시나리오 2의 경우 k값이 4일 때부터 그래프의 기울기가 늘어나다가 다시 k가 7이 되면 감소한다. k값이 6에서 7로 변할 때보다 7에서 8로 변할 때의 시간 측정값 변화 폭이 적다. 케이스 2의 경우 k=7 이후에도 시간 측정값이 계속 증가하기는 하지만 그 이전에 비해 증가폭이 줄어든 모습을 보인다. 따라서 k값이 7보다 커지게 되어도 비식별화에 들이는 노력은 늘어나도 연결에 드는 시간적 비용은 그에 비례하여 늘어나지 않아 큰 의미가 없다고 볼 수 있다. 결론적으로 케이스 2의 경우 k=7일 때 충분히 비식별화 되었다고 볼 수 있다.



〈그림 4〉 케이스 2의 매칭 시간 측정 그래프

그림 5는 케이스 3에서의 시간 측정 결과를 나타낸 것이다. 케이스 3의 경우 k값이 4일 때까지는 변화를 보이지 않다가, 4일 때부터 증가하기 시작해 k값이 8이 된 후 유의미한 증가를 보이지 않는다. 따라서 시나리오 3의 경우 k값이 8보다 커지도록 비식별화를 해도 연결에 드는 노력이 늘어나지 않기 때문에 k=8이면 충분하다고 볼 수 있다.



〈그림 5〉 케이스 3의 매칭 시간 측정 그래프

4.5 토론

본 논문에서는 공격자가 연결 공격을 시도하는 상황을 가정하여 연결하는 데이터셋이 매칭되는 레코드 수의 역수를 재식별 위험으로 정의하였고, 위험이 일정 수준 이상인 매칭을 찾는

데 걸리는 시간을 기반으로 적절한 k값의 판단 기준을 제시하고자 하였다.

[21]의 연구에서는 배경 지식 공격을 하는 공격자가 유사 식별자 이외 다른 속성에 대한 지식을 가지고 있는 경우 이론상 최악의 경우보다 10배 가까이 높은 위험을 가진다는 것을 보였다. 이 연구에서는 radix 트리 구조를 사용해 알려진 속성 집합의 발생 횟수를 계산하여 그 횟수에 통일성이 있다면 재식별 위험이 존재하는 것으로 판단하였고, 공격자가 가지고 있는 배경지식의 정도에 따른 재식별 위험을 분석해 재식별 시간에 기반한 본 연구와는 차이가 있다.

[22]의 연구에서는 기존에 많이 사용되었던 인구 고유성에 기반하여 재식별 위험성을 측정하였다. HIPAA(Health Insurance Portability and Accountability Act)에서는 재식별 위험을 측정하는 방법으로 고유성만을 언급하고 있다[23].

[24]에서는 재식별 위험은 잠재적인 동기와 리소스에 따라 달라지기 때문에 위험을 추정할 순 있지만 허용 가능한 위험의 정해진 임계값은 없다고 언급하고 있다. 법률 및 규제에서 위험의 정도가 낮은 경우 정량적인 임계값을 의도적으로 알리지 않고 있지만 작은 위험이라도 그것을 수용 가능한지 여부는 개별 데이터 주체에 따라 다르다는 점을 지적하고 있다. 본 연구에서는 그러한 경우에도 적절한 합의점을 찾을 수 있는 판단 기준을 제시하고자 하였다.

[25]의 연구는 모바일 멀티미디어 환경에서 비식별화된 데이터셋에 대한 신속한 재식별 위험 평가 모델을 제안하고 있다. 해당 연구에서는 재식별 위험을 데이터셋의 사용자 속성 그룹과 일치하는 레코드 수의 역수로 정의하고 있다. 본 연구에서의 재식별 위험 측정 방식과 맥락을 같이 하며 개인정보 위험 평가의 핵심 지표로 사용하고 있다.

Ⅴ. 결 론

데이터의 수집과 저장이 증가하면서 공유와 활용도 늘어나고 있다. 그러나 이 과정에서 다양한 경로로 프라이버시 침해가 일어나게 된다. 프라이버시 문제를 해결하기 위해 데이터 비식별화 기술이 등장하여 프라이버시를 보호하면서도 데이터를 연구 등의 여러 목적으로 사용할 수 있도록 하고 있다. 그러나 데이터를 비식별화하더라도 다시 식별될 가능성이 충분히 존재하며 완벽하게 안전한 비식별이란 존재하기 어렵다. 본 논문에서는 재식별에 소요되는 시간적 비용을 고려하여 적절한 비식별 수준을 찾는 방법을 제안한다. 다양한 비식별 평가 모델 중에서 k-익명성 모델이 비전문가도 이해하기 쉽고 실생활에서 많이 사용되고 있기 때문에 k-익명성 모델에 대해 집중적으로 연구하였다. 3개의 케이스에서 데이터를 연결하는 데 걸리는 시간을 k값에 변화에 따라 분석하였고 분석 결과 어떤 k값이 적정한 지 판단하였다. 본 논문에서는 특정한 경우를 가정하여 분석했지만 이를 확장하여 향후 일반화된 모델을 만들 수 있다면 각종 법률 및 규제에서 적절한 수준의 비식별 정도를 규정하는 데 사용할 수 있을 것이다.

참 고 문 헌

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) 2016.

[2] ISO DIS 25237 “Health informatics - Pseudonymization,” 2017.

[3] C. A. Cassa, S. C. Wieland, and K. D. Mandl, “Re-identification of home addresses from spatial

locations anonymized by Gaussian skew," International journal of health geographics, vol. 7, no. 1, p. 45, 2008.

[4] A. Cavoukian and D. Castro, "Big data and innovation, setting the record straight: de-identification does work," Information and Privacy Commissioner, vol. 18, 2014.

[5] C. Culnane, B. Rubinstein, and V. Teague, "Health data in an open world: a report on re-identifying patients in the MBS/PBS data set and the implications on future releases of Australian government data," 2017.

[6] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," BMC medical informatics and decision making, vol. 12, no. 1, p. 66, 2012.

[7] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing nyc taxi data: Does it matter?," in 2016 IEEE international conference on data science and advanced analytics (DSAA), pp. 140-148. 2016.

[8] K. El Emam, "Methods for the de-identification of electronic health records for genomic research," Genome Medicine, vol. 3, no. 4, p. 25, 2011.

[9] K. El Emam, E. Jonker, and B. M. Luk Arbuckle, "A systematic review of re-identification attacks on health data," PloS one, vol. 6, no. 12, 2011.

[10] S. Garfinkel of National Institute of Standards and Technology (NIST) "De-Identifying Government Datasets (2nd Draft)," 2016.

[11] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," Journal of biomedical informatics, vol. 50, pp. 4-19, 2014.

[12] N. Y. S. D. o. Health. Hospital Inpatient Discharges (SPARCS De-Identified): 2017.

[13] R. Leenes, R. Van Brakel, S. Gutwirth, and P. De Hert, Data protection and privacy: the age of intelligent machines. Bloomsbury Publishing, 2017.

[14] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in 2007 IEEE 23rd International Conference on Data Engineering, pp. 106-115. 2007.

[15] G. Loukides, J. C. Denny, and B. Malin, "The disclosure of diagnosis codes can breach research participants' privacy," Journal of the American Medical Informatics Association, vol. 17, no. 3, pp. 322-327, 2010.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, pp. 3-es, 2007.

[17] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature communications, vol. 10, no. 1, pp. 1-9, 2019.

[18] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[19] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557-570, 2002.

[20] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: privacy and data mining," Ieee Access, vol. 2, pp. 1149-1176, 2014.

[21] A. Basu, T. Nakamura, S. Hidano and S. Kiyomoto, "k-anonymity: Risks and the Reality,"

IEEE Trustcom/BigDataSE/ISPA, pp. 983-989, 2015.

[22] F. K. Dankar, K. El Emam, A. Neisa and T. Roffey, “Estimating the re-identification risk of clinical data sets,” BMC Medical Informatics and Decision Making, vol. 12, no. 66, 2012.

[23] Office for Civil Rights, HHS. “Standards for privacy of individually identifiable health information. Final rule,” Fed Regist. 2002 Aug 14;67(157): 53181-273, 2002.

[24] G. E. Simon, S. M. Shortreed, R. Y. Coley, R.B. Penfold, R. C. Rossom, B. E. Waitzfelder, K. Sanchez, and F. L. Lynch, “Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records,” EGEMS (Washington, DC), 7(1), 6, 2019.

[25] Z. Yang, R. Wang, D. Luo and Y. Xiong, “Rapid Re-Identification Risk Assessment for Anonymous Data Set in Mobile Multimedia Scene,” IEEE Access, vol. 8, pp.41557-41565, 2020.

사 사

This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2020-2015-0-00403) supervised by the IITP (Institute for Information &communications Technology Planning &Evaluation)

저 자 소 개

김 채 운(Chaewoon Kim)

·2017 고려대학교 컴퓨터학과 학사

·2017~현재 고려대학교 정보보호대학원 석사과정

·관심분야: 정보보호, 데이터 보안, 프라이버시

오 준 형(Junhyoung Oh)

·2017 고려대학교 전기전자전파공학부 (학사)

·2017~현재 고려대학교 정보보호대학원 석·박사 통합과정

·관심분야: 정보보호, 위험관리, 프라이버시

이 경 호(Kyungho Lee)

·1989 서강대학교 수학과 학사

·1997 서강대학교 정보통신대학원 석사

·2009 고려대학교 정보보호대학원 박사

·2017~2019 고려대학교 정보전산처장

·2011~현재 고려대학교 정보보호대학원 교수

·관심분야: 정보보호 정책, 개인정보보호 정책, 위험관리, 머신러닝, 블록체인