프라이버시 보호와 데이터 유용성 향상을 위한 서비스 기반의 안전한 익명화 기법

황치광○ 홍충선*
경희대학교 컴퓨터공학과
{chikwang16, cshong*}@khu.ac.kr

Service-based Secure Anonymization Technique for Privacy Protection and Data Utility Enhancement

Chi Kwang Hwang○ Choong Seon Hong*
Department of Computer Engineering, Kyung Hee University

요 약

개인정보는 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보를 말한다. 개인정보는 정보주체의 민감한 정보를 포함하고 있기 때문에, 유출시 각종 범죄에 악용될 가능성이 있다. 이를 막기 위한 조치로 데이터를 배포 및 공개하기 전에 개인 식별 요소를 제거하고 있다. 그러나 이름이나 주민등록번호 등의 식별자를 삭제 또는 변경하더라도, 다른 데이터와 연계하여 분석하면 개인정보가 노출될 수 있다. 본 논문에서는 서비스에 활용될 속성은 낮은 강도의 익명화를 수행하여 실제 사용될 정보의 유용성을 높이면서도, 연결 공격에 대한 우려 없이 하나의 원본 데이터 테이블로부터 둘 이상의 익명화 테이블을 동시에 제공할 수 있는 익명화 기법을 제안한다.

1. 서 론

'개인정보'는 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보를 말한다. 이 개인정보에는 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다[1]. 개인정보가 포함된 데이터는 학교, 기업, 정부와 같은 많은 기관에서 사용된다. 이러한 기관들은 개인정보가 포함된 데이터를 역학조사 등의 연구 목적으로 활용하기 위해 수집하고 공유하고 있다. 특히, 기업에서는 광고의 효과를 높이기 위한 개인 맞춤형 광고를 위해 방대한 개인정보를 수집하고 있다. 그러나 이렇게 수집된 개인정보는 정보주체의 민감한 정보를 포함하고 있기 때문에, 유출시 각종 범죄에 악용될 가능성이 있다. 이런 이유로 금융정보와 같은 개인정보를 노린 각종 보안 위협이 늘어나는 추세이다. 이러한 부작용을 막기 위한 조치로 데이터를 배포 및 공개하기 전에 가명처리, 총계처리, 범주화, 데이터 마스킹 등 개인 식별 요소를 제거하여 데이터로부터 개인을 식별하지 못하게 하는 방법을 사용하고 있다. 그러나 데이터에서 개인의 신원을 명백히 파악할 수 있는 이름이나 주민등록번호 등의 식별자를 삭제 또는 변경하더라도, 다른 데이터와 연계하여 분석하면 개인정보가 노출될 수 있다.

투표인명부				↔	진료기록			
이름	나이	성별	우편번호		나이	성별	우편번호	병명
Ahmed	25	남	53711		25	남	53711	독감
Brooke	28	여	55410		25	여	53712	간염
Clair	31	여	90210		27	여	53712	AIDS
…	…	…	…		…	…	…	…

그림 1. 연결 공격의 예

그림 1은 유권자 정보와 의학정보를 연계해서 환자의 신원을 확인할 수 있음을 보여주는 연결 공격의 예이다. 두 데이터 테이블을 이용한 연결 공격을 통해 Ahmed의 성별, 연령, 우편번호는 물론, 질병 정보까지 얻을 수 있다[2].

본 논문의 구성은 2장에서 프라이버시 관련 연구를 분석하고, 3장에서 이에 기반을 둔 새로운 익명화 기법을 제안한다. 4장에서 제안 사항에 대해 평가한 후, 5장에서 결론을 맺는다.

2. 관련 연구

식별자 준 식별자 민감한 속성

이름	나이	성별	우편번호	병명
홍길동	21	남	482010	암
김철수	24	남	482750	암
이영희	47	여	420760	감기
김희영	49	여	420880	당뇨

그림 2. 데이터 테이블의 예

그림 1과 같이 단순히 식별자를 삭제 또는 변경한 데이터는 익명화하지 않은 데이터와 마찬가지로 개인정보를 노출할 가능성이 있다. 보다 안전한 프라이버시 보호를 위해 다른 데이터와의 연관성을 줄이며 익명화하여야 한다. 일반적으로 통계적으로 수집된 데이터는 그림 2와 같이 식별자(identifier), 준 식별자(quasi-identifier), 민감한 속성(sensitive attribute) 으로 구성되며, 흔히 데이터 테이블이라고 부른다. 데이터 테이블의 각 행은 레코드라 하고 각 속성의 값을 나타낸다. 개인의 신원을 명백히 나타내는 이름, 주민등록번호 등을 식별자라고 하며, 생년월일, 성별, 우편번호 등 개인의 특징을 나타내는 속성인 준 식별자는, 직접적으로 대상을 알 수는 없지만 조합을 통해 간접적으로 개인 식별이 가능한 속성이다. 또한, 민감한 속성은 데이터 테이블이 제공하고자 하는 개인의 민감한 정보를 나타낸다. 일반적으로 민감한 속성에 대한 정보를 제공하기 위해, 식별자를 제거하고 준 식별자를 익명화함으로써 프라이버시 보호를 수행한다.

본 논문은 산업통상자원부 산업핵심기술개발사업으로 지원된 연구결과입니다 [10049079, 퍼스널 빅데이터를 활용한 마이닝 마인즈 핵심 기술 개발]

*Dr. CS Hong is the corresponding author

2.1. K-익명성 (K-anonymity)

K-익명성은 다른 데이터와의 연관성을 줄이기 위해 L. Sweeney에 의해 고안된 익명화 기술이다. K-익명성을 만족한다는 의미는 데이터 테이블 내 준 식별자의 모든 기록들이 적어도 K번 데이터 테이블에서 나타난다는 것을 뜻한다[3]. 그림 3에서 t1과 t2를 구분할 수 없으며, 마찬가지로 t3과 t4를 구분할 수 없고, t5와 t6, t7을 구분할 수 없다.

	나이	성별	우편번호	병명
t1	**20-24**	**남**	**482*****	암
t2	**20-24**	**남**	**482*****	암
t3	**45-49**	**여**	**420*****	감기
t4	**45-49**	**여**	**420*****	당뇨
t5	**30-34**	**남**	**410*****	감기
t6	**30-34**	**남**	**410*****	감기
t7	**30-34**	**남**	**410*****	피부염

그림 3. K-익명성의 예 (K=2)

2.2. L-다양성 (L-diversity)

	나이	성별	우편번호	병명
t1	**20-24**	**남**	**482*****	**감기**
t2	**20-24**	**남**	**482*****	**암**
t3	**45-49**	**여**	**420*****	**감기**
t4	**45-49**	**여**	**420*****	**당뇨**
t5	**30-34**	**남**	**410*****	**감기**
t6	**30-34**	**남**	**410*****	**감기**
t7	**30-34**	**남**	**410*****	**피부염**

그림 4. L-다양성의 예 (L=2)

K-익명성을 만족시키면서 데이터 테이블을 익명화하더라도 프라이버시 보호에 실패할 수 있다. 그림 3은 K-익명성을 만족하지만 t1의 병명이 반드시 암이라는 사실을 알 수 있다. 이런 문제로 인해 A. Machanavajjhala는 L-다양성이라는 기술을 고안하였다. L-다양성을 만족한다는 의미는 데이터 테이블 내 준 식별자의 모든 기록들이 같은 하나의 클래스의 민감한 속성에서 적어도 L개의 서로 다른 속성 값이 존재한다는 것을 뜻한다. 여기서 클래스는 모든 준 식별자의 값이 동일한, K-익명성을 이루는 그룹을 의미한다[4]. 그림 4는 L이 2이므로 t1의 병명이 감기일 확률은 1/2이다.

3. 제안 사항

준 식별자는 개인을 식별할 때 사용할 뿐만 아니라, 서비스의 데이터나 연구 자료로 충분히 활용될 수 있기 때문에, 정보의 손실을 줄여 데이터의 유용성을 높이는 연구가 필요하다. 그러나 기존의 익명화 관련 연구들의 대부분은 데이터의 유용성(Utility)보다는 프라이버시 보호에 집중하였다. 데이터의 유용성을 고려한 몇몇 기존 연구들이 있었지만, 이들은 보통 전체 데이터 테이블의 정보 손실을 줄이는 방향으로 연구되었다. 그러나 데이터의 유용성은 서비스 및 애플리케이션에 따라 다르기 때문에 전체 데이터의 정보 손실이 적다고 하더라도, 데이터를 실제로 활용하는 서비스나 애플리케이션의 관점에서 유용성이 높다고 보장할 수는 없다. 반대로 전체 데이터 테이블의 정보량이 적다고 하더라도 서비스나 애플리케이션 관점에서 유용한 정보가 적다고 단정 지을 수도 없다. 따라서 서비스의 관점에서 데이터의 유용성을 높일 수 있는 익명화 기법이 필요하다. Jian Xu는 [5]에서 애플리케이션에 따라 속성들의 유용성에 차이가 있음을 언급하였으며, S. Kiyomoto는 [6]에서 사용자의 요구사항에 맞는 익명화 테이블을 만드는 기법을 제안하였다.

원본 데이터

이름	나이	성별	우편번호	병명
김철수	22	남	482534	A
이희영	23	여	410264	C
박수철	32	남	410342	B
이민수	33	남	482289	D

익명화 테이블 1

나이	성별	우편번호	병명
20-24	*	******	A
20-24	*	******	B
30-34	남	******	C
30-34	남	******	D

익명화 테이블 2

나이	성별	우편번호	병명
*	남	482***	A
*	*	410***	B
*	*	410***	C
*	남	482***	D

복원한 데이터

나이	성별	우편번호	병명
20-24	남	482***	A
20-24	*	410***	B
30-34	남	410***	C
30-34	남	482***	D

그림 5. 동일한 원본 데이터에 대한 연결공격

본 논문에서는 서비스를 기반으로 한 익명화 기법을 제안한다. 준 식별자 중 서비스에 집중적으로 활용될 속성은 낮은 강도의 익명화를 수행하여 실제 사용될 정보의 유용성을 높인다. 반면, 이외의 속성은 익명화 강도를 높여 K-익명성 및 L-다양성을 만족시킬 수 있도록 한다. 그러나 동일한 원본 데이터로부터 여러 개의 익명화 테이블을 만들 경우, 이를 민감한 속성을 기준으로 비교하면 프라이버시 노출의 위험이 있다. 그림 5는 K-익명성과 L-다양성을 만족한 두 개의 익명화 테이블을 비교해 복원한 데이터 테이블이 K-익명성과 L-다양성을 만족하지 못함을 보여준다.

이를 방지하기 위해 본 논문에서는 S(L)-다양성(S(L)-diversity) 모델을 새롭게 제안하고, 이를 활용한 Service-based Secure Anonymization 기법을 제안한다. S(L)-다양성은 민감한 속성에 대해 L-다양성을 만족하면서, 서로 다른 클래스에 동일한 값이 S개 존재하는 새로운 익명화 기법이다. 본 기법을 만족하기 위한 조건은 다음과 같다. 첫째, K-익명성과 L-다양성을 만족한다. 둘째, 동일한 민감한 속성 값이 서로 다른 클래스에 S개 존재한다. 셋째, 최초의 익명화 테이블이 출력된 후 두 번째 익명화부터 이전에 공개되지 않은 준 식별자들의 값은 공개가 가능하다. 예를 들어, 최초의 테이블에서 나이 속성이 '*' 로 익명화 되어 공개되지 않았다면, 두 번째 테이블은 나이 속성을 21-30과 같이 공개할 수 있다. 넷째, 최초의 익명화 테이블이 출력된 후 두 번째 익명화부터 이전에 공개된 준 식별자의 값과 같지 않은 준 식별자 값은 공개하지 않는다. 예를 들어, 최초의 테이블에서 성별 속성의 일부 레코드에서 '남자' 라는 값이 공개되었다면, 두 번째 테이블은 성별 속성에서 '남자'라는 값은 공개할 수 있지만, '여자' 라는 값은 공개할 수 없다. 동일한 원본 데이터에서 S(L)-다양성을 만족하도록 익명화된 테이블들은 1/S 확률로 연결 공격에 대한 내성을 가진다. 이를 통해 완전히 같은 원본데이터로부터 다른 방식으로 익명화된 여러 개의 테이블을 이용한 연결 공격을 막을 수 있다.

Service-based Secure Anonymization는 S(L)-다양성을 적용하여, 서비스에 활용될 속성은 낮은 강도의 익명화를 수행하여 실제 사용될 정보의 유용성을 높이면서도, 연결 공격에 대한 우려 없이 하나의 원본 데이터 테이블로부터 둘 이상의 익명화 테이블을 동시에 제공할 수 있는 익명화 기법이다. 사용자에게 익명화된 데이터가 필요한 서비스의 개수를 입력 받고, 각 서

비스마다 주로 활용할 속성을 입력 받는다. 익명화 모듈은 S(L)-다양성을 만족하는 선에서 요청받은 익명화 테이블을 작성하고, 사용자에게 제공한다.

알고리즘1. Service-based Secure Anonymization

Input : Original data table **OT**; Quasi-identifier attribute which will be used **QI**; **Max K**, **Min K** for K-anonymity; **S**, **L** for S(L)-diversity
Output : Anonymized data table **AT**

```
1  generalize all records to max hierarchy
2  K = Max K
3  IF  K < Min K  THEN
 3.1  AT is unavailable
4  ELSE  de-generalize in other to satisfy K-anonymity
 4.1  IF  the table satisfies K-anonymity  THEN
  4.1.1  check L-diversity of sensitive attribute from the table
  4.1.2  IF  the table satisfies L-diversity  THEN
   4.1.2.1  check S(L)-diversity of sensitive attribute from the table
   4.1.2.2  IF  the table satisfies S(L)-diversity  THEN
    4.1.2.2.1  de-generalize other quasi-identifiers within K-anonymity
    4.1.2.2.2  publish the anonymized table AT
   4.1.2.3  ELSE  K=K-1 and go to the step 3
  4.1.3  ELSE  K=K-1 and go to the step 3
 4.2  ELSE  K=K-1 and go to the step 3
```

4. 평 가

그림 6은 동일한 원본 데이터에서 3(4)-다양성을 만족하도록 익명화된 테이블들이다. 이 두 테이블의 데이터는 1/3 확률의 민감한 속성을 이용한 연결 공격에 대한 내성을 가진다. Table 1의 첫 번째 레코드는 Table 2의 레코드 중 3개와 일치할 확률이 있으므로, 두 테이블의 레코드가 연결될 확률은 1/3이다.

Table 1. 나이 활용

나이	성별	우편번호	병명
20-24	*	******	A
20-24	*	******	B
20-24	*	******	C
20-24	*	******	D
45-49	*	******	A
45-49	*	******	B
45-49	*	******	C
45-49	*	******	D
30-34	남	******	C
30-34	남	******	D
30-34	남	******	A
30-34	남	******	B

Table 2. 우편 번호 활용

나이	성별	우편번호	병명
*	남	482***	A
*	남	482***	B
*	*	410***	C
*	*	410***	D
*	*	420***	A
*	*	420***	B
*	남	482***	C
*	남	482***	D
*	*	420***	C
*	*	420***	D
*	*	410***	A
*	*	410***	B

그림 6. 3(4)-다양성을 만족하여 1/3 확률로 연결 공격을 방지하는 예

익명화된 테이블의 데이터 유용성을 측정하기 위해 Latanya Sweeney가 [7]에서 고안한 측도인 Precision을 변형한 Weighted Precision을 사용한다. 이는 서비스에 활용될 속성의 중요도를 고려한 측도이다. Weighted Precision을 사용하여 그림 6의 익명화 테이블에 대한 데이터 유용성을 측정할 수 있다. Weighted Precision의 최댓값은 1이며, 이는 활용될 속성의 정보손실이 없음을 의미한다. 나이와 질병의 상관관계를 이용한 서비스를 예로 들어보자. 이 서비스는 나이, 성별, 우편번호 중 나이 데이터가 90%, 성별 데이터가 10% 필요하며, 우편번호 데이터는 필요하지 않다고 가정한다. 이 경우, 나이를 활용하기 위해 익명화한 테이블의 Weighted Precision 값은 0.633이다.

$$WeightedPrec(Table) = 1 - \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\frac{h_{ji}w_{A_i}N_A}{|GH_{A_i}|}}{N \bullet N_A}$$

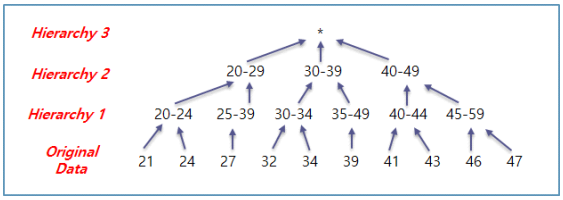

그림 7. 일반화 계층의 예

표1. Weighted Precision의 파라미터

파라미터	설명
W_{A_i}	속성 A_i의 가중치
N	모든 레코드의 개수
N_A	모든 준 식별자 속성의 개수
A_i	준 식별자 속성 i를 나타내는 기호
h_{ij}	A_i에서 j번째 레코드의 일반화 계층
$\|GH_{A_i}\|$	A_i의 전체 일반화 계층

5. 결 론

개인정보는 정보주체의 민감한 정보를 포함하고 있기 때문에, 유출시 각종 범죄에 악용될 가능성이 있지만, 이를 활용한 연구가 필요하기 때문에 개인 식별 요소를 제거한 후, 데이터를 배포 및 공개하고 있다.

본 논문에서는 서비스에 활용될 속성은 낮은 강도의 익명화를 수행하여 실제 사용될 데이터의 유용성을 높이면서도, 연결 공격에 대한 우려 없이 하나의 원본 데이터 테이블로부터 둘 이상의 익명화 테이블을 동시에 제공할 수 있는 익명화 기법을 제안하였다. 향후에는 실제 서비스에 데이터를 제공할 때 필요한 가격 정책에 대한 연구를 수행하고자 한다.

참 고 문 헌

[1] 신신애 외, “빅데이터 활용을 위한 개인정보 비식별화 사례집”, 한국정보화진흥원, 2014년 5월
[2] 김형중, “통계적 익명성을 위한 Privacy 보호 기술”, NIA Privacy Issues, 제 2호, 2012년 6월
[3] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp.557-570, 2012
[4] A. Machanavajjhala et al. “L-Diversity: Privacy Beyond k-Anonymity”, ACM Transactions on Knowledge Discovery from Data, 1(1), March 2007
[5] Jian Xu at al, “Utility-Based Anonymization Using Local Recoding”, The 12th ACM SIGKDD international conference on Knowledge discovery and data mining, March 2007
[6] Shinsaku Kiyomoto at al, “A User-Oriented Anonymization Mechanism for Public Data”, Data Privacy Management and Autonomous Spontaneous Security, vol.6514, pp.22-35, 2011
[7] L. Sweeny, “Achieving k-anonymity privacy protection using generation and suppression”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp.571-588, 2012