제 2 호
2012. 6. 20

NIA PRIVACY ISSUES

한국정보화진흥원
개인정보보호정책단

통계적 익명성을 위한 Privacy 보호 기술

고려대학교 정보보호대학원 교수 김 형 중

1. 서론

통계적 데이터베이스에서 프라이버시 보호와 자료의 정확성은 서로 상충한다. 프라이버시를 중시해서 자료를 모호하게 공개할수록 당연히 자료의 정확성은 낮아지고, 프라이버시를 약화시켜 모호성을 누그러뜨리면 정확성은 높아진다. 데이터를 모은 후 공익목적으로 공개할 때 익명화 처리를 하면 그것으로 충분하다고 믿는 것은 순진한 생각이다. 연결공격(linkage attack)으로 인해 실명을 유추할 수 있는 방법이 많기 때문이다.

그래서 통계적 데이터베이스의 정확성은 높이면서 개인의 식별 가능성을 낮추는 것을 목표로 삼는 게 소위 차분프라이버시(differential privacy) 개념이다. 이 개념은 마이크로소프트의 Dwork 등이 2006년 처음 제안했다. 간단히 말하자면 고객이 데이터세트에 참여하거나 탈퇴하는 것과 무관하게 확률분포가 분석결과의 확률분포가 거의 달라지지 않도록 해주자는 것이 주된 개념이다.

통계적 데이터베이스에서는 참여자 수가 많을수록 더 정확한 분석이 가능해진다. 그러자면 참여자가 데이터베이스에 참여하기 전에 프라이버시가 충분히 보장된다는 확신을 심어주어 한다.

간혹 비밀자료를 수집하는 정부기관, 병원, 검색서비스제공 업체 등은 원래의 데이터 수집목적에 어긋나게 다른 용도로 쓸 수 있도록 데이터를 공개하라는 압력을 받을 수 있다. 이런 상황에서 데이터수집 관련된 내부 상태가 악의적 공격자에게 공개되더라도 프라이버시를 유지할 수 있게 하자는 것이 범프라이버시(pan-privacy) 개념이다. 이 개념은 2010년 역시 Dwork 등에 의해 제시되었다.

본 기고에서는 프라이버시 보호기술의 최신동향을 살펴봄으로써 향후 한국에서 어떤 대책을 취해야 하는지 알아보고자 한다.

2. AOL 4417749 사례

2006년 8월 4일 AOL Research는 문제의 압축된 텍스트 파일을 연구목적으로 공개했다. 이 파일에는 3개월간 65만명의 사용자가 2천만 건의 키워드를 검색한 기록을 담고 있었다. 당연히 AOL은 사용자를 식별할 수 없게 했다. 그러나 익명 처리한 검색자 번호 4417749가 일주일도 지나지 않아 Thelma Arnold로 밝혀졌다. 정보가 넘쳐나는 사이버공간에서 익명성 처리가 얼마나 어려운지 알려주는 사례라 할 수 있다.

AOL 데이터에는 다음과 같은 다섯 개 항목만이 포함되어 있다.

1. 익명 아이디: 익명의 사용자 아이디 번호 (실명 등은 포함되지 않음)
2. 질문: 사용자가 올린 질문
3. 질문시각: 검색하기 위해 질의한 시각
4. 아이템 순위: 검색 결과를 사용자가 클릭했을 때 검색된 아이템의 순위에 대한 기록
5. 클릭 URL: 검색 결과를 사용자가 클릭했을 때 검색된 도메인 주소에 대한 기록

이 자료를 바탕으로 뉴욕타임즈 기자는 실명을 추적해 보기로 했다. 이 자료만으로는 추적이 쉽지 않지만 검색한 질문에 포함된 사람 이름이나 지명을 전화번호부 등의 기록과 조합해 보니 실명 확인이 가능해 마침내 같은 달 9일 뉴욕타임즈 1면에 릴번에 사는 62세의 여성인 Thelma Arnold가 개를 안고 있는 사진과 함께 등장했다.

AOL은 사태의 심각성을 깨닫고 3일만인 7일에 데이터를 내렸다. 그러나 삭제 시기가 이미 너무 늦어 데이터가 여러 사이트로 퍼져나갔고 지금도 일부 미러 사이트에서 자료를 구할 수 있다. 이로 인해 담당 임원이 사임했고 AOL은 50억달러에 달하는

집단소송에 휩싸였다. 물론 이 데이터베이스 공개로 인해 실명이 공개된 것은 4417749이 처음으로 65만명이라는 수에 비하면 아주 미미하기는 하다.

3. Netflix 사례

2006년 10월, 넷플릭스(Netflix)는 자사의 영화추천시스템을 개선하기 위해 백만 달러에 달하는 넷플릭스 상을 공표했다. 동시에 연구목적으로 6년간 50만명의 고객들이 작성한 1억 개에 달하는 영화 관련 데이터세트를 공개했다. 이 자료는 고객의 이름은 삭제되고 대신 숫자가 배정되었으며, 영화 이름, 영화 평점 기록을 담고 있다. 이름은 익명으로 처리되었지만 이 자료를 다른 데이터베이스와 비교하면 의외로 실명 추적이 가능하다는 것이 밝혀졌다. 인터넷 무비 데이터베이스라는 영화 평점 자료와 연계시킬 경우 6개의 영화 평점만을 비교해도 84%의 고객 이름을 식별할 수 있다고 한다. 그림 1이 이런 관계를 보여준다.

넷플릭스 연구목적으로 이 자료를 공개했으나 2007년 이처럼 실명확인이 가능하다는 사실이 알려졌고 2009년에는 소송에도 휩싸여 자료를 내렸다. 이 경우에도 두 데이터베이스에서의 클릭 성향이 비슷하므로 둘을 연관 지을 수 있다는 의미이지 정확히 실명이 공개된 것은 아니다. 그러나 유추할 수 있는 가능성은 충분히 높다는데 동의한다.

Netflix Dataset		
User id	Movie id	Rating
12345	Batman	3/5
12345	The Wizard	4/5
12345	Rocky II	3/5
...	...	...
98765	Pulp Fiction	3/5

Internet Movie Database		
User id	Movie id	Rating
daniel	Rocky II	6/10
daniel	The Wizard	8/10
daniel	Batman	6/10
...	...	...
wang75	Rambo	5/10

그림 1. 고객 12345가 Daniel임을 알 수 있는 예

4. Latanya Sweeney

Latanya Sweeney 교수는 다섯 자리의 우편번호, 생년월일, 성별 등 세 가지 정보만 알면 미국에 거주하는 주민 87%를 식별할 수 있다고 했다.

시 이름, 생년월일, 성별 정보만으로는 53%를, 카운티 이름, 생년월일, 성별만 알아도 18%를 알 수 있다고 한다. 그래서 간접적으로 개인을 식별할 수 있는 정보도 엄격히 관리해야 한다.

투표인명부			
Name	Age	Sex	Zip code
Ahmed	25	M	53711
Brooke	28	F	55410
Clair	31	F	90210
Dave	19	M	02174
Evelyn	40	F	02237
Soomin	34	F	32133

진료기록			
Age	Sex	Zip code	Disease
25	M	53711	Flu
25	F	53712	Hepatitis
26	M	53711	Bronchitis
27	M	53710	Broken arm
27	F	53712	AIDS
28	M	53711	Hang nail

그림 2. 투표인명부와 진료기록을 연계한 개인식별 방법의 예

GIC라 불리는 Group Insurance Commission은 공무원들의 병원진료기록을 연구목적 차원에서 공개했다. 데이터베이스에서 개인의 이름은 당연히 익명으로 처리됐다. 그래서 메사츠세츠 주지사 William Weld는 이 기록 공개가 프라이버시를 침해하지 않을 거라 확신했다. 당시 대학원생이던 Sweeney는 이 데이터베이스를 이용해 개인을 식별할 수 있는지 알아보기로 했다. 그래서 그녀는 20달러를 주고 케임브리지 시의 선거인명부를 구입했다. 이 자료에는 투표자 이름, 우편번호, 생년월일, 성별이 담겨있었다. 그녀는 주지사의 이름과 일치하는 사람 6명 찾아냈고, 그 가운데서 남성 3명으로, 다시 같은 우편번호에 사는 사람 1명으로 압축해 나갔다. 그리고 주지사 사무실로 진단내역과 처장기록을 담은 의료기록을 우편으로 보냈다.

그림 2는 유권자 정보와 의학정보를 연계해서 환자의 신원을 확인할 수 있음을 보여준다. 그림 1에서 유권자 Ahmed의 성별, 연령, 우편번호를 얻을 수 있다. 환자정보 데이터베이스에 우연히도 25세의 남성이 우편번호 53711인 지역에 살며 독감에 걸렸던 정보가 공개될 경우 아마도 그 환자는 유권자 Ahmed일 가능성이 매우 높다. 투표자 명단에 최근 이사를 와서 누락된 사람이 있거나 이사간 사람이 있을 경우에는 이 Ahmed가 그 Ahmed가 아닐 수도 있다.

5. K 익명성

K-익명성(K-anonymity)이라는 기술은 이런한 유형의 연관성을 줄이기 위해 2002년 고안되었다.

그림 2에서 개인 이름은 그 자체로 민감한 식별자(sensitive identifier)이고, 생년월일, 성별, 우편번호도 어느 정도 간접적으로 개인을 식별할 능력이 있어 준식별자(quasi-identifier)라 부른다.

민감한 정보인 이름이나 주민등록번호와 다르므로 준식별자만 공개하는 것은 안전하다고 William Weld처럼 쉽게 생각할 수 있으나 투표인명부 등과 연계되면 환자의 이름이 식별될 수도 있어 주의가 요구된다. 당연히 개인을 직접 식별할 수 있는 이름은 삭제하고 데이터베이스를 공개해야 하지만 그렇다고 그것으로 충분하지 않다.

그림 3은 그림 2의 진료기록을 다시 정리한 예로 위의 셋과 아래의 셋은 나이, 성별, 그리고 우편번호 관점에서 볼 때 차이가 없게 만들어져 있다. 각각 3개의 항목이 서로 구별되지 않기 때문에 3-익명성 처리되었다고 한다. 이렇게 될 경우 그림 2의 선거인명부에서 본 Ahmed는 처음 그룹에 드는지 두 번째 그룹에 드는지도 분명하지 않게 된다. 어느 그룹에 들든 정확히 Ahmed를 식별할 확률은 1/3로 낮아진다. 물론 확률적으로는 처음 3명의 기록에 포함될 가능성이 더 높다. 유권자 명부에서 Ahmed는 나이 25세, 성별 남, 거주지역 우편번호 53711로 되어있다. 그러므로 남성이며 우편번호 53711인 곳에 살고 나이는 25세에서 28세 사이인 사람들이 처음 그룹에 들어있기 때문이다.

Age	Sex	Zip code	Disease
25-28	**M**	**53711**	**Flu**
25-28	**M**	**53711**	**Hang nail**
25-28	**M**	**53711**	**Bronchitis**
25-27	*	5371*	Broken arm
25-27	*	5371*	AIDS
25-27	*	5371*	Hepatitis

그림 3. 3-익명성을 구현한 데이터베이스의 예

그러나 나이는 25세에서 27세, 성별은 남녀 모두 포함, 우편번호는 53710부터 53719까지 망라하는 아래 그룹에 포함될 수도 있다. 어느 경우든 비록 Ahmed의 정보를 안다고 해서 병명을 정확히 짚어내기는 상대적으로 어려워졌다.

그러므로 K-익명성의 주된 목적은 데이터와 관련된 개인의 프라이버시를 보호하는 것이다.

K의 값이 커질수록 익명화는 더 쉬워진다. 그러나 K가 커짐에 따라 공개되는 데이터베이스의 효용성은 낮아진다. 예를 들어 극단적으로 K의 값이 무한대라면 이 경우 데이터베이스는 공개하나 마나 하게 된다. 레코드 값이 모두 "*"로 채워질 것이기 때문이다. 그렇다고 K를 1로 하면 익명화가 부실해서 데이터베이스를 원래대로 공개하는 것과 다를 바 없다.

게다가 이런 익명화 방법은 한 두 가지가 아니다. 그림 4에도 3-익명성이 구현되어 있다. 이 경우 적색 그룹은 나이가 25세에서 26세까지, 흑색 그룹은 27세에서 28세까지로 두 살 단위로 묶였다. 나이 관점에서 보면 그림 4의 묶음이 그림 3의 묶음보다 좋다. 그러나 성별에서 볼 때는 그림 4의 경우 모든 사람의 성을 구별할 수 없어 그림 3에 비해 못하다. 우편번호 관점에서도 그림 4의 경우 모든 우편번호가 5371*로 표시되어 정확도 측면에서 그림 3만 못하다. 이와 같이 익명화 데이터베이스를 만드는 경우 구현방법은 무수히 많은데 어떤 방법이 더 효율적인지에 대해서는 고려해 봐야 한다.

K-익명성 개념을 이용한 익명화의 문제는 이미 그림 3과 4에서 본 바와 같이 데이터베이스 분할(clustering) 문제가 됨을 알 수 있다. 그런데 분할할 수 있는 방법은 많기 때문에 그 가운데 어떤 분할이 가장 좋은 방법인지에 대한 문제로 귀결된다. 가장 좋은 것을 찾는다는 것은 최적화 문제라는 의미가 된다. 즉 K-익명성 문제는 최적의 클러스터링 문제가 됨을 알 수 있다.

Age	Sex	Zip code	Disease
25-26	*****	**5371***	**Flu**
25-26	*****	**5371***	**Hepatitis**
25-26	*****	**5371***	**Bronchitis**
27-28	*	5371*	Broken arm
27-28	*	5371*	AIDS
27-28	*	5371*	Hang nail

그림 4. 3-익명성의 또 다른 예

그렇지만 최적의 다차원 익명화는 전산학적으로 NP-hard 문제이다. 그래서 이 문제에 적용할 수 있는 최적의 솔루션을 찾는 것은 레코드 수 N이 커지고 K도 크면 현실적으로 불가능하다. 그림 2의 경우 3-익명성 문제를 풀 경우 모두 ${}_6C_3$개의 경우의 수가 생긴다. 즉, 20 가지의 경우의 수가 생긴다. 우선 6개의 레코드에서 3개를 골라내 그룹을 만드는 방법이 20가지가 있다. 그런데 데이터베이스에는

최소 수백 만개의 개인 기록이 포함되는 예가 비일비재하다. 다만 K가 2인 경우에는 최적의 해를 찾을 수 있다는 것이 알려져 있다. 그래서 일반적으로 최적의 해를 구하는 것은 포기하고 휴리스틱 방법을 적용하게 된다.

굳이 이 방법의 흠을 잡자면 그림 5와 같이 같은 집단의 사람들이 동일한 질병을 앓아서 생기는 동질성(homogeneity)의 문제, 남녀 구분이 불가능하게 했지만 난소염(ovaritis)은 여성에 국한된 병이므로 사실상 배경지식에 의해 성별이 구별되는 문제 등을 안고 있다.

Age	Sex	Zip code	Disease
25-28	**M**	**53711**	**Prostate Cancer**
25-28	**M**	**53711**	**Prostate Cancer**
25-28	**M**	**53711**	**Prostate Cancer**
25-27	*	5371*	Ovaritis
25-27	*	5371*	Ovaritis
25-27	*	5371*	Ovaritis

그림 5. 3-익명성의 또 다른 예

6. L-다양성

비록 K-익명성을 구현했다 해도 프라이버시 보호에 큰 도움이 되지 않을 때가 있다. 그림 5는 3-익명성을 잘 구현한 예에 속하지만 이 데이터베이스를 자세히 보면 Ahmed는 전립선암(prostate cancer)을 앓고 있음을 쉽게 추측할 수 있다. 왜냐하면 병명이 모두 동일하기 때문이다. 이처럼 병명이 모두 같을 경우 K-익명성의 효과가 크게 감소한다.

이런 문제를 해결하기 위해 사용하는 기술이 L-다양성(L-diversity) 기술이다. K-익명성이 확률을 1/K로 낮춘다면 L-다양성도 마찬가지로 1/L로 떨어뜨린다. 그래서 K-익명성이 지켜지게 한 후 병명에서 L-다양성을 유지하는 방법을 적용한다. 그림 6은 2-다양성을 함께 구현한 예라 할 수 있다. 자세히 보면 이 경우는 3-익명성과 2-다양성을 구현한 예라고도 할 수 있다. 병명이 전립선암과 난소염으로 구성되어 있어 병명 예측 확률은 1/2로 감소시킬 수 있다.

그런데 그림 6의 이 예는 그 효과가 매우 떨어지는 경우에 속한다. 병명이 두 가지라 하지만 전립선암은 남성에게 발생하고 난소염은 여성에게 발생하므로 둘을 합쳐 확률을 낮추었다 한들 실제로는 효과가 없게 된다. 이처럼 좋은 기술을 적용하려 해도 현실적으로 많은 제약이 따름을 알 수 있다.

Age	Sex	Zip code	Disease
25-28	*	**5371***	**Prostate Cancer**
25-28	*	**5371***	**Prostate Cancer**
25-28	*	**5371***	**Ovaritis**
25-28	*	5371*	Prostate Cancer
25-28	*	5371*	Ovaritis
25-28	*	5371*	Ovaritis

그림 6. 3-익명성과 2-다양성을 구현한 예

7. 차분프라이버시

프라이버시에 대한 수학적 모델링을 처음 시도한 것이 차분프라이버시라 할 수 있다. 그림 7은 인구통계조사 과정을 묘사하고 있다. 개인들이 답변한 내용을 모아 가공해서 그 자료를 공표하면 정부나 연구자들이 활용하게 된다. 그런데 만일 개인들이 정부를 신뢰하지 않으면 쉽게 질문에 답하지 않게 된다. 특히 성생활이나 범죄와 관련된 민감한 질문에는 더더욱 그렇다. 시스템을 일단 신뢰해야 많은 사람들이 참여하게 되고 그래야 그 결과가 가치를 지니게 된다. 이 데이터베이스에 담긴 자료는 개인을 바로 식별할 수 있는 정보를 담고 있다. 여기에 비해 선호도나 추천 정보에서 개인정보 유출 가능성은 낮지만 가능성이 전혀 없는 것은 아니다.

그런데 프라이버시라는 개념은 매우 모호하기 이를 데 없다. 그래서 수학적인 모델링이 필요하게 되었다. 1977년 통계학자 Tor Dalenius는 "통계적 데이터베이스에서 응답자에 관해 알 수 있는 모든 것은 데이터베이스에 접근하지 않고도 알 수 있다"는 보증을 할 수 있어야 한다고 제안했다. 달리 말하면 응답자가 자발적으로 웹에 공개한 자료 이외에는 데이터베이스에서 어떤 정보도 얻을 수 없다는 것을 보여야 한다는 것이다. 수학적으로는 매우 좋은 아이디어이기는 하나 현실성이 적다

그림 7. 인구통계조사 과정

그림 8은 데이터베이스 D_1과 D_2의 차이를 보여준다. 두 데이터베이스의 차이는 단 하나. D_1에 Jane이 있는데 D_2에는 없다는 점. 이 두 데이터베이스는

가공(sanitization) 과정을 거쳐 $M(D_1)$과 $M(D_2)$로 각각 공개된다. S는 함수 M의 영역에 속하는 멤버라고 하자. 이때 다음과 같은 조건을 만족하면

$$\frac{P[M(D_1) \in S]}{P[M(D_2) \in S]} \leq e^{\epsilon}$$

이 데이터베이스는 ∈-차분프라이버시를 유지한다고 정의한다. 단 여기서 데이터베이스 D_1과 D_2은 Jane이 빠진 것처럼 한 항목만 차이가 나야 한다. 이런 조건을 만족하는 두 데이터베이스 D_1과 D_2을 근접데이터베이스라고 부른다. 여기서의 값은 작을수록 좋은데 만일 그 값이 0이면 두 확률의 비는 1이 된다.

차분프라이버시는 수학적 모델이므로 이 모델에 적합한 구현이 필요하다. 차분프라이버시 모델도 약점이 있는데 아웃라이어(outlier)를 제거하고 결과를 보여주는 경향이 있기 때문에 간혹 아웃라이어를 연구할 경우 오히려 제약을 받을 수 있다.

D_1		
User id	Age	Sex
John	45	M
Jane	24	F
Cain	48	M
...	...	...
Abe	81	M

D_2		
User id	Age	Sex
John	45	M
Cain	48	M
Lee	47	M
...	...	...
Abe	81	M

그림 8. 근접한 두 데이터베이스의 사례

데이터를 공개해야 할 일은 많은데 어떻게 정보를 공개해야 할 지는 여전히 연구해야 할 일이 많다. 소셜 네트워크만 해도 그렇다. 개인과 개인의 링크 정보를 익명화해서 공개한다 해도 다른 소셜 네트워크 링크 정보를 활용하면 실명 확인이 가능하다는 것도 알려졌다. 이런 것을 보면 여전히 연결공격의 가능성은 널리 열려있다. 그래서 충분한 연구를 통해 데이터베이스 공개가 최대 효용, 최소 프라이버시 유출로 귀결되도록 노력해야 할 것이다.

[참고문헌]

C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd TheoryofCryptographyConference*, pp. 265-284, 2006.

C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin, "Pan-private streaming algorithms," in *Proc.* , pp. 66-80, 2010.

1. 본문에 언급된 내용은 한국정보화진흥원의 공식적인 견해와 다를 수 있습니다.
2. 본문의 내용에 대해 무단전재를 금하며, 가공·인용할 때는 반드시 출처를 밝혀주시기 바랍니다.
3. 문의 : (02)2131-0129 / wonjin@nia.or.kr (선원진 선임)

Copyright©한국정보화진흥원 all rights reserved.