데이터 공개를 위한 트랜잭션 데이터의 익명화 알고리즘[1]

김영훈○ 박형민* 심규석□
○□서울대학교, *University of British Columbia
{○yhkim, *hmpark, □shim}@kdd.snu.ac.kr

Anonymizing Transaction Data for Publication

Younghoon Kim○ Hyoungmin Park* Kyuseok Shim□
○□Seoul National University, *University of British Columbia

각종 데이터가 정부이나 의료 기관, 회사에 의해 공개되어 연구목적이나 마케팅을 위해 유용하게 활용되고 있다. 그러나 공개된 데이터를 통해 개인의 사생활이 노출될 가능성이 있기 때문에 이를 막기 위해 익명화(anonymization) 방법이 활발히 연구되고 있다. 최근 미국의 아메리카온라인(AOL)은 최근 사용자들의 ID를 삭제하고 URL은 도메인만 남긴 웹 검색 로그 데이터를 일반에 공개하였다. 그러나 [1]에서 공개된 로그데이터를 통해 사용자 개인을 알아낼 수 있다는 사실이 밝혀졌다. 사용자의 ID와 같이 명백한 개인정보를 지우더라도 나머지 공개된 정보를 통해 어떤 사용자의 데이터인지를 알 수 있기 때문이다. 데이터를 공개할 때 AOL이 취한 방법과 같이 흔히 주민등록번호나 이름 등 명백히 개인을 특정 지을 수 있는 정보(identifier)를 지우는 방법을 사용한다 하더라도 이 방법으로는 불충분하다.

[2]에서는 (h,k,p)-coherence라는 개념을 소개한다. 이는 공격자가 p개의 물품을 구입한 사람을 알고 있고 이 사람의 데이터가 공개된 데이터에 들어 있음을 알고 있을 때 1/k보다 높은 확률로 어떤 데이터인지를 특정할 수 없으며 같은 p개의 물품을 구입한 사람들 중 민감한 물품을 갖고 있는 사람의 비율이 h이상이 되지 않는 다는 것을 말한다. 예를 들어 그림 1의 왼쪽 표는 장바구니 데이터는 데이터에 포함된 모든 가능한 2개 물품 집합에 대해서 그 집합을 포함하는 데이터에 민감한 물품의 비율이 50%를 넘어가지 않으며 그 집합들이 나타나는 트랜잭션 역시 2개 이상이기 때문에 (50%,2,2)-coherence를 만족 한다고 말 할 수 있다. 그러나 이 모델은 공격자가 물품의 부재 정보를 갖고 있는 경우의 개인정보를 특정할 수 있는 문제는 해결하지 못한다. 만약 공격자가 특정인이 특정 물품을 사지 않았다는 정보를 알며 그 사람의 데이터가 공개된 데이터 안에 있다는 것을 안다면 특정인의 개인정보를 알아 낼 수 있을 것이다.

TID	*Public Items*	*Priv. Items*
T1	{Alcohol, Diapers, Preg. Test}	
T2	{Alcohol, Diapers}	{Diamond Ring}
T3	{Alcohol, Pregnancy Test}	
T4	{Alcohol }	{Playboy}
T5	{Water, Diapers, Preg. Test}	
T6	{Water, Diapers}	{Adult Video}
T7	{Water, Pregnancy Test}	

TID	*Public Items*	*Priv. Items*
T1	{Alcohol, Diapers, Preg. Test}	
T2	{Alcohol, Diapers}	{Diamond Ring}
T3	{Alcohol, Pregnancy Test}	
T4	{Alcohol }	{Playboy}
T5	{Water, Diapers, Preg. Test}	
T6	{Water, Diapers}	{Adult Video}
T7	{Water, Pregnancy Test}	
T8	{Water, Diapers}	

그림 1 (50%, 2, 2)-coherence 데이터와 Extended (50%, 2, 1, 1)-coherence 데이터

[1] 이 논문 또는 저서는 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0000793)

○ 발표자, □교신저자

예를 들어 그림 1의 왼쪽 표는 (50%, 2, 2)-coherence를 만족하는 공개 데이터이지만 만일 공격자가 Water를 사고 Pregnancy Test를 사지 않은 사람이 데이터에 속한다는 것을 알면 바로 T6가 그 사람의 데이터이며 Adult Video를 같이 구입했다는 것을 알 수 있을 것이다.

우리는 (h,k,p)-coherence를 확장하여 공격자가 특정인이 p개의 어떤 물품들을 구입하고 n개의 어떤 물품들은 구매하지 않았다는 정보를 알 경우에도 특정인을 1/k 확률이상 특정해내지 못하고 같이 구입한 개인적 물품까지도 h의 확률 이상 알 수 없는 (h,k,p,n)-coherence 모델을 제안하였다. 만약 (h,k,p,n)-coherence 모델에서 n을 0으로 둔다면, 즉, 공격자가 물품의 부재정보를 모른다면 곧 (h,k,p)-coherence와 같은 모델이 되기 때문에 (h,k,p,n)-coherence모델은 (h,k,p)-coherence 보다 더 강력한 정보 보호 모델이다. 예를 들어 표 1의 오른쪽 표는 (50%,2,1,1)-coherence를 만족하는 데이터이다. 사용자가 구입한 물품 하나와 구입하지 않은 물품 하나를 알고 있더라도 그림4에서는 그 사용자의 데이터를 1/2 확률 이상으로 찾을 수 없고 또한 그 사용자가 같이 산 개인적인 물품 또한 50% 이상 정확도로 추측할 수 없다.

본 연구에서는 쇼핑몰의 장바구니 데이터나 검색엔진의 쿼리 로그와 같이 관계형 데이터와 달리 정해진 속성이 없이 집합 형태를 갖는 데이터를 외부 사람들에게 공개하기 위한 익명화를 연구한다. 특히 물건을 구입한 내역 뿐만 아니라 구입하지 않은 정보를 통해서도 개인 정보 노출이 일어날 수 있다는 점을 고려해 (h,k,p,n)-coherence라는 모델을 제시하고 또한 정보 손실량을 최소로 하기 위한 그리디 알고리즘을 제안하였다.

주어진 데이터 $D=\{T_1,\ldots,T_n\}$는 트랜잭션 데이터 이며 각 트랜잭션 T_i는 U에 속하는 물품들의 집합이다. 트랜잭션 데이터에는 예를 들어 장바구니 데이터나 검색엔진의 쿼리 세션 로그가 있다. $U=\{e_1,\ldots,e_m\}$는 트랜잭션 데이터에 포함된 물품 목록이고 물품 목록은 공개 가능한 일반 물품들 U_{pub}와 공개를 기피하는 민감한 물품 U_{priv}로 나뉜다. 민감한 물품은 예를 들어 종교, 정치적 성향과 같은 민감한 내용을 나타내는 것을 말한다. 본 연구에서는 일반 물품과 민감한 물품이 미리 분류되어 있다고 가정한다. 일반 물품의 정보에 대한 사전 지식은 개인정보를 알아내고자 하는 공격자가 얻을 수 있으며 민감한 물품은 쉽게 얻을 수 없다고 가정한다. β를 공격자가 알 수 있는 물품의 목록이며 이 중 사용자가 구입한 물품은 p개, 구입하지 않은 것은 n개가 포함되어 있다고 할 때, D에 나타나는 모든 가능한 β에 대해서 $Sup(\beta) \geq k$이거나 $P_{breach}(\beta) \leq h$이면 데이터는 extended (h,k,p,n)-coherence라고 한다. 여기서 $P_{breach}(\beta)$는 모든 개인적 물품에 대해서 β와 같이 나타나는 비율 중 최대 값이라 정의한다. 또한 익명화로 인해 손실되는 정보의 양을 [3]에서 제안한 NCP를 확장하여 정보 손실량 MCP를 정의하였다.

우리는 [4]의 정의에 따라 D에 나타나는 사용자가 구입한 물품은 p개, 부재물품은 n개인 모든 물품 집합 β중 $Sup(\beta) < k$이거나 $P_{breach}(\beta) > h$인 β를 mole이라고 부르고, 데이터가 (h,k,p,n)-coherence 만족시키도록 데이터를 변형하기 위해 물품의 일반화(item generalization)와 트랜잭션 추가(transaction appending)을 통해 정보 손실량을 최소로 하며 모든 mole을 제거하는 그리디 알고리즘을 고안하였다. 그리고 실생활 데이터를 이용한 실험을 통해 기존의 연구와 비교하고 정보 손실량을 더욱 줄일 수 있음을 검증하였다.

[1] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In ICDE, pages 217–228, 2005.

[2] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. PVLDB, 1(1):115–125, 2008

[3] J.Xu, W.Wang, J.Pei, X.Wang, B.Shi, and A.Fu. Utility-based anonymization using local recording. In SIGKDD, 2006.

[4] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In KDD, pages 767–775, 2008