효과적인 k-RDFAnonymity 를 위한 알고리즘 구현

전민혁*, Odsuren Temuujin*, 서광원*, 안진현**, 임동혁*
*호서대학교 컴퓨터공학과
*e-mail : jeoncoder, temuujintemka, rhkddnjs153@gmail.com, dhim@hoseo.edu
**제주대학교 경영정보학과
**e-mail : jha@jejunu.ac.kr

Implementation of algorithm for effective k-RDFAnonymity

Min-Hyuk Jeon*, Odsuren Temuujin* Seo Kwangwon*, Jinhyun Ahn**, Dong-Hyuk Im*
*Dept. of Computer Engineering, Ho-Seo University
**Dept. of Management Information Systems, Jeju National University

요 약

최근 정부 및 기업단체에서 배포하는 데이터의 규모가 점점 방대해지고 있다. 민간에서는 이러한 공개데이터를 자유롭게 사용할 수 있으나, 공개 데이터에는 개인의 프라이버시를 침해할 수 있는 개인정보도 포함되어 있다. 그에 따라 대두된 문제가 공개데이터 중 개개인의 정보를 식별해낼 수 없도록 하는 데이터의 비식별화이며 그로 인해서 비식별화에 관한 많은 익명화 기법과 프라이버시 모델이 발표되었다. 그중 본 논문에서 사용하는 Mondrian algorithm 은 k-익명화 모델을 사용하여 효과적으로 데이터를 비식별화할 수 있다. 또한 방대한 웹 데이터 자원 간의 관계를 표현해놓은 RDF 모델을 DB 로 변환시켜 k-익명화 방법인 kRDF 에 Mondrian algorithm 의 Multi-dimensional 방식을 따라 익명화하여 범용적이고 효과적인 개인정보 데이터의 프라이버시 보호를 구현하고자 한다.

1. 서론

빅데이터를 사용하기 위해서는 데이터를 개인이나 기업 및 정부를 통해서 개인정보 수집 및 동의를 얻은 후에 수집하거나 소셜네트워크서비스(SNS) 또는 정부 및 단체에서 자발적으로 공개하는 공개데이터를 사용할 수도 있으며 웹상의 자원 정보를 표현한 RDF(Resource Description Framework) 데이터를 사용하기도 한다. 이렇게 수집된 개인정보는 데이터의 통계적 유용성을 그대로 유지한 채, 개개인을 식별할 수 없는 비식별화 상태의 데이터로 수집되어야 하는데, 그에 따라 많은 익명화 방법이 제시되었고 k-익명성[1], l-다양성[2], t-근접성[3] 같은 데이터의 익명화 정도를 수치화시킬 수 있는 모델들이 발표되었다. 하지만, 데이터의 익명화는 NP-Hard 한 방법으로써[4] 데이터의 특성마다 가장 최적의 모델이 다르고 그중 Mondrian algorithm[5]은 DB 데이터의 튜플 값에 따라 같은 준식별자를 가진 여러 개의 Partition 으로 나누어 배경지식을 통해 유추될 확률을 낮춰주는 k-익명화 모델의 알고리즘이다. 본 논문에서는 웹 환경에서 사용하는 RDF 데이터를 DB 데이터로 변환하여 Mondrian Multi-dimensional k-익명화 방법으로 익명화를 시도하고, k 값의 변환을 통해 나누어지는 동질클래스의 수를 비교해보고자 한다.

2. 관련연구

공개데이터에서의 데이터 익명화 연구는 한국뿐만 아니라 앞서 외국에서부터 연구가 진행되었다. 익명화 연구의 시초는 본 논문에서 사용하는 Mondrian algorithm 에서 사용하는 k-익명화 모델[1]이라고 할 수 있는데, k-익명화를 포함한 많은 알고리즘은 RDBMS 데이터를 사용한다.

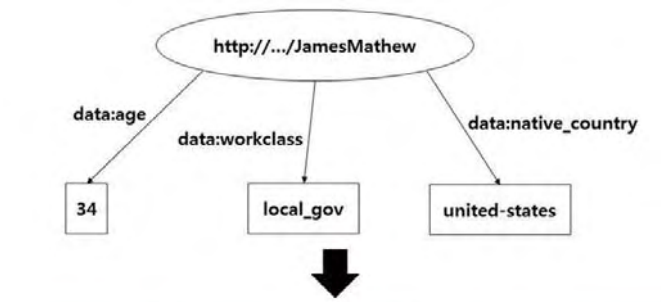
RDBMS(Relational Database Management system) 데이터는 키와 값들의 관계를 테이블 화 시킨 데이터베이스를 의미하며, 여기에는 각각의 속성값을 의미하는 속성(Attribute)과 속성의 집합인 튜플(Tuple)들로 구성되어 있다.

k-익명화 모델에서는 테이블의 속성값을 다음과 같은 여러 유형으로 나눌 수 있다. 속성자체만으로도 특정 개인을 식별할 수 있는 정보를 식별자(Identifier), 단일 속성만으로는 개인의 식별이 불가능하지만, 외부 데이터와 결합할 경우 특정 개인을 식별할 수 있는 속성을 준식별자(QI: Quasi-Identifier), 데이터 분석 혹은 공격자의 공격 대상이 되는 사생활과 밀접한 속성을 민감속성(SA: Sensitive Attribute) 등의 유형이다. 모델 적용 순서는 다음과 같다. 데이터에서 식별자 속성의 값을 제거한 뒤에 테이블의 각 튜플이 최소 k-1 개의 다른 튜플과 구분할 수 없도록 같은 준식별자 값들을 가지는 동질클래스(EC: Equivalence Class)를 생성한다면 k-익명화를 만족한다고 할 수 있다. k-익명화 모델을 만족하면 공격자는 배경지식으로 준식별자를 알고 있더라도 민감속성을 1/k 의 확률로밖에 추정할 수 없다.

3. K-RDFAnonymity 구현

3.1. RDF 데이터 튜플 변환

본 논문에서는 RDF 모델의 데이터를 DB 테이블로 변환하여 익명화시킨다. RDF 모델에서 테이블로의 변환은 Java Programming Language 의 Library 인 Jena framework 를 사용하여 Model[6]로 Parsing 한 후, 데이터베이스에 속성값에 따른 튜플로 저장하였다. 변환 과정은 그림 1 과 같다. JamesMathew 라는 Subject 는 data:age, data:workclass, data:native_country 등의 Predicate 를 가지고 있고 이는 각각의 속성이 된다. 그리고 Predicate 에 해당하는 34, local_gov, united-state 등의 Object 는 하나의 튜플로 합쳐진다. 이렇게 변환된 테이블에 kRDF 익명화[7]를 적용하게 된다.



name	age	workclass	native_country
JamesMathew	34	local_gov	united-states

(그림 1) Adults RDF 테이블 변환

3.2. DB 데이터의 Mondrian algorithm 적용

Mondrian algorithm 의 동작 방법은 동질클래스의 크기를 정함에 있어서 준식별자를 한 개만 쓰는 Single-Dimensional 방법과 2 개 이상의 준식별자를 사용하는 Multi-Dimensional 방법이 있다. 이 중에서 Single-Dimensional 방법을 사용할 경우 동질클래스의 크기는 선형적으로 매우 커질 수가 있어 데이터의 유용성을 해칠 가능성이 크지만, Multi-Dimensional 방법을 사용했을 동질클래스의 최대 크기는 2k 로, 최악의 경우에도 동질클래스는 2k 개의 튜플로 구성된다. 또한 Multi-Dimensional 방법을 사용할 경우 모든 동질클래스의 크기를 정확히 균등 하는 방법인 Stric Multi-Dimensional 과 EC 의 크기가 정확히 같지는 않아도 되는 Relaxed Multi-Dimensional 이 있다. 그리고 실험에서는 Relaxed Multi-Dimensional 의 방법을 사용하였다.

Relaxed Multi-Dimensional 을 두 개의 축을 갖는 2-Dimensional 로 예를 든다면 속성값 중에서 기준이 될 준식별자 2 개를 정하고, 값의 범위가 더 큰 준식별자를 기준으로 모든 튜플에 대한 준식별자의 종류별 개수를 순차 정렬된 Fs 테이블로 생성한다. 그리고 Fs 테이블에서 가장 작은 준식별자부터 차례로 개수를 더한 값이 Fs 테이블에 포함된 준식별자의 전체 개수의 절반 이상이 될 때의 준식별자값을 기준으로 DB 데이터의 모든 튜플을 두 개로 나눈다. 이후 나누어진 튜플은 기준이 된 준식별자보다 작거나 같은 값을 가진 튜플들은 Lhs 테이블로, 큰 값의 튜플들은 Rhs 테이블로 저장한 뒤 Rhs, Lhs 테이블의 크기가 모두 k 개 이하가 될 때까지 재귀적 방법을 통해 나눈 뒤에 분할된 각각의 테이블을 익명화한다.

4. 성능평가

4.1. 실험환경

본 실험은 웹 데이터인 RDF 데이터를 DB 테이블 데이터로 변환하여 Mondrian algorithm 을 적용하여 2-dimensional k-익명화를 실행한 후 k 값의 변화에 따른 나누어진 동질클래스의 값을 비교했다. 여기에는 총 15 개의 속성값과 10 만 개의 튜플을 가진 Adults 데이터[8]가 사용됐다.

4.2. 성능분석

본 논문에서는 Adults 데이터의 age 와 fnlwgt 를 준식별자로 지정한 뒤 native_country 를 민감속성으로 정하였다. 그에 대한 데이터는 아래의 표 1 와 같고, Adults 데이터를 2-dimensional 2-익명화 방법으로 익명화하여 표 2 로 나타내었다.

<표 1> Adults 원본 데이터

name	*Age*	*fnlwgt*	*native_country*
Hills	27	10000	united-state
Runy	28	10000	laos
Emma	31	20000	canada
Simpson	27	20000	cuba
Jone	30	20000	united-state
Clark	32	10000	Jamaica
Park	32	30000	united-state
Cook	27	40000	portugal
Steam	28	30000	united-state
Bula	29	20000	India

<표 2> Adults 2-dimensional 2-익명화 데이터

Age	*fnlwgt*	*native_country*
29-30	20000	india
29-30	20000	united-state
27-32	30000-40000	united-state
27-32	30000-40000	united-state
27-32	30000-40000	prtugal
27-28	10000-20000	united-state
27-28	10000-20000	cuba
27-28	10000-20000	laos
31-32	10000-20000	canada
31-32	10000-20000	jamaica

이후 각각 다시 2, 5, 10, 50 으로의 k 값을 정한 뒤 데이터를 k-익명화 실험을 진행했고, 그 결과는 다음의 표 3 과 같다. 표 3 에서의 k 값이 5 인 상태의 동질클래스 수는 16382 로 Strict Multi-dimensional 의 방법으로의 기댓값인 10000(튜플 수/2k)보다 더 크다. 왜냐하면 최악의 경우에 동질클래스의 최대 개수가

2k 인 경우는 Strict Multi-dimensional 한 방법을 사용했을 때이지만 이는 이론적인 방법이고, 실제로는 Relaxed Multi-dimensional 한 방법을 사용해야 하므로 결과에 약간의 차이가 있다. 이러한 차이는 k 값이 클수록 줄어들며, 준식별자로 사용된 속성값의 분포가 작아도 차이가 줄어든다. 결론적으로 k-익명화는 이러한 차이가 작을수록 익명성이 증가하게 된다.

<표 3> k 값에 따른 동질클래스의 개수 비교

k 값	2	5	10	50
EC 개수	36183	16382	8192	1027

5. 결 론

본 논문에서는 웹 자원 데이터 형식인 RDF 데이터를 Java 언어의 Jena framework 를 통해 테이블 형태의 DB 데이터로 변환한 뒤에 Mondrian algorithm 을 Multi-dimensional 방식으로 k-RDF 익명화 처리를 하는 실험을 하였다. 이 방법을 통해 다양한 웹 환경에서도 쉽게 데이터를 얻어서 최적의 k-익명화 모델로 구현할 수 있었고, 이 데이터를 배경지식을 가지고 있는 공격자에게서부터 공격자가 찾고자 하는 레코드를 정확히 찾을 수 없도록 하여 프라이버시를 안전하게 보호할 수 있었다. 하지만, Mondrian algorithm 은 숫자 데이터에 대해서만 익명화가 가능하고, 또한 k-익명화 모델만으로는 모든 유형의 공격으로부터 보호하기에는 부족하다. 그 예로, k-익명화 모델로 익명화한 데이터 중 특정 동질클래스의 민감속성이 같은 특정 값으로만 있거나 민감속성의 분포가 작을 때는 공격에 취약해진다. 이에 따라 k-익명성을 만족하는 상황에서 l-다양성, t-근접성 등의 다양한 모델들이 더 존재하고, 이러한 모델은 Mondrian algorithm 을 적용하기가 힘들다. 그리고 적용하고자 하는 데이터의 크기가 클 경우 익명화하는 기기의 성능에 따라 처리 속도가 많은 영향을 받는데, 이것과 관련하여 향후 연구 주제는 오픈소스 분산처리 플랫폼인 Apache Spark 를 이용한 여러 컴퓨팅 환경에서의 k-익명성과 l-다양성을 만족하는 익명화 실험을 RDF 데이터에 적용하여 익명화 처리속도를 개선하는 실험을 진행할 것이다.

Acknowledgement

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No.NRF-2017R1C1B1003600)이며 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.R0113-15-0005, 대규모 트랜잭션 처리와 실시간 복합 분석을 통합한 일체형 데이터 엔지니어링 기술 개발). 또한, 이 논문은 2018 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2018R1D1A1B07048380).

참고문헌

[1] L. Sweeney. K-anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness, and Knowledgebased Systems, 10(5):557–570, 2002.

[2] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-Diversity: Privacy Beyond k-Anonymity. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006), p. 24 (2006)

[3] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anon. and l-diversity. In ICDE, 2007

[4] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In Proc.of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 223-228, 2004.

[5] LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Data Engineering, 2006. ICDE'06. Proceedings of the 22nd Inter-national Conference on. pp. 25{25. IEEE (2006)

[6] Wilkinson, Kevin. "Jena property table design." Proceedings of the Jena Users Conference, Bristol, England. 2006

[7] Radulovic, Filip and García Castro, Raul and Gómez-Pérez, A. Towards the anonymisation of RDF data. In: "27th International Conference on Software Engineering and Knowledge Engineering", 2015.

[8] "Uci machine learning repository." [Online]. Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/