경로정보 개인 비식별화를 위한 K-익명성의 K값에 따른 데이터 활용과 익명성의 상충관계

Trade-off between data use and anonymity according to K value of K-anonymity for personal trajectory de-identification

저자 (Authors)

이인경, 엄수현, 아로샤, 자파르, 이우기
InKyung Lee, Chris Soo-Hyun Eom, Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee

출처 (Source)

한국정보과학회 학술발표논문집 , 2018.6, 1241-1242(2 pages)

발행처 (Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07503307

APA Style

이인경, 엄수현, 아로샤, 자파르, 이우기 (2018). 경로정보 개인 비식별화를 위한 K-익명성의 K값에 따른 데이터 활용과 익명성의 상충관계. 한국정보과학회 학술발표논문집, 1241-1242

경로정보 개인 비식별화를 위한 K-익명성의 K값에 따른 데이터 활용과 익명성의 상충관계

이인경, 엄수현, 아로샤, 자파르, 이우기
인하대학교
{22171259, sunching, Arousha.haghighian, jafar.afshar, trinity}@inha.edu

Trade-off between data use and anonymity according to K value of K-anonymity for personal trajectory de-identification

InKyung Lee, Chris Soo-Hyun Eom, Arousha Haghighian Roudsari, Jafar Afshar, and Wookey Lee
INHA University

요 약

스마트 기기의 확산에 따라 개인 위치정보 빅데이터는 광범위하게 수집 및 활용될 수 있지만, 프라이버시로 인해 수집이 제한되기도 한다. 데이터베이스의 비식별화 처리를 통한 활용이 필수적이지만 안전의 비중이 큰 경우 데이터의 활용이 심각히 제한될 수 밖에 없다. 본 연구에서는 데이터의 활용과 익명성의 상충관계를 보이고 적절한 K값의 중요성을 입증하기위해 K-익명성 모델에 대해 설명하고 경로 데이터베이스에 적용하여 정량적으로 이를 입증하였다.

1. 서 론

스마트 기기들의 증가와 아울러 급증하고 있는 개인의 위치정보는 광범위하게 수집, 보관, 분석되어야 하지만 프라이버시 문제 때문에 수집과 활용이 법적으로 금지되는 경우가 많다. 이 문제는 빅데이터 시대에 접어들면서 반드시 해결해야 할 과제의 하나이다. 위치정보는 GPS를 통해 얻어진 데이터를 기반으로 가공되고 있으며 최근 관련 연구들이 활발히 이루어지고 있다 [1,2]. 이동경로의 경우, 도로 건설, 교통량 조정, 범죄발생 그래프, 대중 교통 계획, 관광정보, 맛집 분석, 이벤트 참석자 분석, 소셜 네트워크 분석 등에 응용되고 있다 [3]. 기본적으로 이동 객체(object)의 경로 정보는 시간과 공간 데이터를 포함하는 스트리밍 데이터로 Mobile 기능이 더해지면서 계속 확산되고 있다.

2. K-익명성과 비식별화

개인의 위치 정보가 포함된 데이터베이스의 배포는 LBS를 포함한 다양한 연구를 수행하는 데 있어 필요하다. 하지만, 비식별화 처리가 없이 단순하게 해당 데이터베이스를 배포할 경우 악의적인 공격자에 의해 개인이 식별되어 정보가 노출되어 심각한 피해를 볼 수 있기 때문에 비식별화 처리는 데이터베이스의 배포에 있어 매우 필수적이라고 할 수 있다. 비식별화란 일반적으로 데이터베이스에 있는 개인을 직접 식별할 수 있는 식별정보(이름, 주민번호, 전화번호)와 타 정보들과 연계하여 개인화할 수 있는 준식별정보(연령, 주소, 성별, 몸무게) 및 프라이버시 정보(질병, 약품, 성적 등) 등을 모호한 데이터로 일반화하는 기법, 데이터 제거 등의 기법을 사용하여 개인의 식별을 방지하는 방법을 말한다.

비식별화를 위한 여러 모델이 개발되고 연구되었으며 그중 K-익명성은 가장 기본적이면서도 중요한 모델로 연구되어 왔다. K-익명성은 데이터베이스내에 동일한 데이터를 가지는 레코드가 적어도 K개가 존재하도록 하여 특정 개인을 식별할 수 없게 하여 프라이버시를 보호하는 모델이다. 예컨대, 공격대상이 되는 피해자의 단편적인 위치정보를 배경지식으로 보유하고 있는 어떤 악의적인 한 공격자가 피해자가 포함된 경로 데이터베이스를 입수하였다고 가정하자. 만약, 경로 데이터베이스에 공격자가 알고 있는 배경지식에 해당하는 피해자의 단편적인 위치정보가 드러난 레코드가 하나뿐이라면 공격자는 100%의 확률로 해당 레코드를 피해자 것으로 특정화 할 수 있을 것이고 이를 통해 레코드에 포함된 다른 정보를 악용할 위험이 매우 높다. 하지만, K-익명성 모델에 따라 그 위치정보를 포함하는 레코드가 10개 즉, K=10이라면 공격자가 피해자의 레코드를 특정화 할 수 있는 확률은 10분의 1로 낮아진다. 즉, K값이 높을수록 데이터베이스내에 동일한 데이터를 가진 레코드가 많을수록 높은 익명성이 보장된다고 할 수 있다.

3. 데이터 활용과 익명성의 상충관계

높은 익명성을 보장하기 위해 K값을 높이기만 하면 좋은 것인가? 그렇지 않다. 데이터베이스가 동일한 데이터를 가진 레코드를 다수 보유하도록 처리하기 위해서는 앞서 말한 데이터를 일반화 또는 제거, 잡음 추가 등의 기법을 통해 정확하지 않은 데이터로 변환하거나 다수의 데이터를 제거해야 하기 때문에 이는 데이터의 활용성이 저하되는 결과를 초래한다. 즉, 익명성과 데이터 활용성은 상충관계(Trade-off)가 있다.

데이터의 활용과 익명성의 보장을 동시에 달성하기 위해서 적절한 혹은 최적 K값을 찾는 것은 중요한 문제라고 할 수 있다. 상세한 이론전개는 지면 한계상 생략한다. 본 연구에서는 종래의 단순 위치정보에 대한 비식별화에서 나아가 이동경로를 보여주는 경로 데이터베이스에 대한 비식별화를 통해 K-익명성에 따른 데이터의 효율성과 익명성의 상충관계를 보이기 위해 실험을 통해 입증하고 적절한 K값을 찾는 문제를 다루고자 한다. 이를 위해 경로 데이터베이스의 K-익명성을 다룬 K_L-Privacy 방법론[4,5]을 채택하고 K값의 변화에 따른 실험을 통해 이를 입증한다. K_L-Privacy 방법은 경로 데이터베이스에서 K개 미만의 빈도수로 나타나는 데이터 시퀀스를 공격자가 보유한 배경지식에 대상이 될 수 있는 위반 시퀀스(Violating Sequence)로 둔다. 또한, 모든 위반 시퀀스를 지울 경우 너무 많은 데이터를 지워 데이터 활용에 매우 좋지 않은 결과를 초래하므로 이를 고려한 최소 위반 시퀀스(Minimal Violating Sequence)를 정의하고 이들을 삭제하여 K-익명성을 달성하는 억제(Suppression) 기반의 방법이다. 또한, 기존 연구에서는 특정 건물 및 지역만을 위치정보에 반영하였으나 본 연구에서는 여러 교통수단이나 모바일 기기 등 좀 더 다양한 위치 데이터를 반영하여 일반적으로 폭넓게 적용할 수 있는 방법을 적용하였다. 이를 위해 위도 및 경도 좌표 데이터를 사용하여 그리드에 맵핑하는 전처리를 하였다.

4. 실 험

실험에서는 10,000개의 레코드와 50,029개의 데이터로 임의로 생성한 Synthetic 데이터셋과 실제 데이터로 구성된 T-Drive 데이터셋을 실험 데이터로 사용하였다.

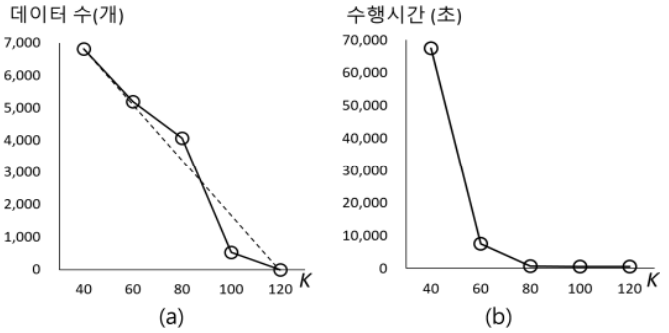

Figure 1. Synthetic 데이터셋 실험결과

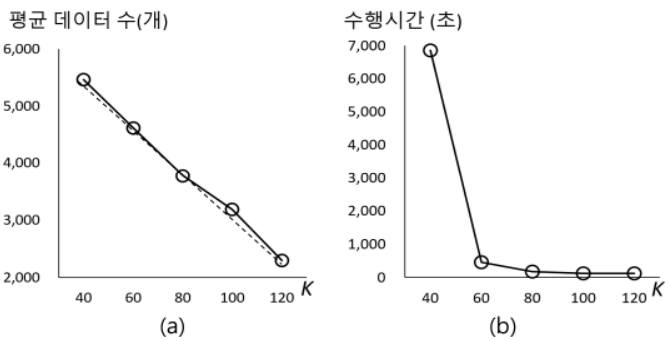

Figure 2. T-Drive 데이터셋 실험결과

T-Drive 데이터셋은 중국 북경의 다수 택시들의 GPS 데이터 기반의 스마트한 운전 방향 서비스로부터 생성되었다. 해당 데이터셋은 택시 10,357 대의 한 주간에 걸쳐 운행한 데이터를 보이며 총거리 900만Km에 약1,500만 개의 데이터로 구성된다. 본 연구는 T-Drive 데이터셋을 사용하였다. 실험에서는 K의 증가에 따라 달라지는 위반 시퀀스를 제거하는 과정을 통해 남은 데이터의 크기와 처리 수행 속도를 비교하였다. 그림 1과 2는 각각 임의의 데이터셋과 T-Drive에 대하여 실험한 결과이다. 그림 1과 2의 (a)에서 K값이 증가함에 따라 많은 데이터를 제거함으로써 활용할 데이터의 수는 줄어드는 것을 확인할 수 있다. 하지만, K가 증가하면서 익명성이 증가할 뿐만 아니라 수행 시간이 대폭 개선되는 것을 확인할 수 있다. 결과적으로 데이터 활용과 익명성 및 수행 시간의 상충관계를 통해 양자를 모두 고려한 K값을 설정하는 것이 중요하다는 사실을 확인할 수 있었다.

5. 결 론

본 연구에서는 위치정보 빅데이터의 활용에 따른 개인 경로정보 비식별화에 대하여 데이터 활용과 익명성의 상충관계를 다루었다. 비식별화를 위해 임의 데이터셋과 더불어 실제 데이터셋에 경로정보의 비식별화 연구를 적용하였으며 이 과정에서 보다 일반적이고 다양한 빅데이터에 적용될 수 있도록 그리드 기반의 셀 데이터로 맵핑하는 전처리 과정을 수행하였다. 본 연구의 실험 결과 높은 익명성과 더불어 수행시간이 좋아질수록 데이터 활용성이 떨어지는 상충관계를 확인하였으며 적절한 K값 설정의 필요성을 명확히 제시하였다. 추후 이에 관한 이론개발 및 검증절차가 필요할 것임을 입증하였다.

6. Acknowledgement

“이 논문은 2016년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(NRF-2016R1A2B4014245)”.

7. 참고문헌

[1] S. Ma, Y. Zheng, O. Wolfson, "T-share: A large-scale dynamic taxi ridesharing service,” *ICDE*. 2013

[2] S. Park, J. Song, J. Lee, W. Lee, S. Ree: How to measure similarity for multiple categorical data sets? *Multimedia Tools Appl.* 74(10): 2015

[3] X. Yang, et al.: “A Novel Representation and Compression for Queries on Trajectories in Road Networks,” *IEEE TKDE*,30(4) 2018

[4] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, K. Wang, “Privacy-preserving trajectory data publishing by local suppression,” *Information Sciences*, Vol. 231, No. 5, pp. 83-97, 2013

[5] Y. Dong, D. Pi: “Novel Privacy-preserving algorithm based on frequent path for trajectory data publishing,” *KBS*. 148:55-65, 2018