비식별 처리 데이터에 대한 정보손실 측정방법 조사연구

A survey of the information loss measuring methods for de-identification dataset

저자 (Authors)
성민경, 유정승
Min Kyoung Sung, Jeong Seung Yu

출처 (Source)
한국정보과학회 학술발표논문집 , 2018.6, 164-166(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07502941

APA Style
성민경, 유정승 (2018). 비식별 처리 데이터에 대한 정보손실 측정방법 조사연구. 한국정보과학회 학술발표논문집, 164-166

비식별 처리 데이터에 대한 정보손실 측정방법 조사연구

성민경○ 유정승
한국정보통신기술협회
mksung@tta.or.kr, js.yu@tta.or.kr

A survey of the information loss measuring methods for de-identification dataset

Min Kyoung Sung○ Jeong Seung Yu
Telecommunications Technology Association

요 약

빅데이터, 사물인터넷(IOT)의 발달로 정보 활용기술이 나날이 발달하고 있으나 그 과정에서 개인정보 침해에 대한 우려가 커지고 있다. 이에 따라 데이터에 포함된 개인정보의 비식별 처리 데이터를 이용하는 방안이 활성화되고 있으며, 정부 관계부처는 '개인정보 비식별 조치 가이드라인'을 제시하여 개인정보 침해 방지에 힘쓰고 있다. 가이드라인에 따라 비식별 처리된 데이터는 원본 데이터의 개인정보를 수정/삭제하여 만들기 때문에 원본데이터의 정보가 손실될 수 있다. 본 논문에서는 비식별 처리된 데이터의 정보손실 측정 방법에 대한 연구들을 조사하여 소개하고자 한다.

1. 서 론

빅데이터, 사물인터넷(IOT)의 발달로 수많은 데이터가 매일 생성되고 있으며, 이를 이용한 새로운 IT 기술과 융합산업의 발전은 가속되고 있다. 데이터를 공개하여 활용하고자 하는 요구는 점차 커지고 있으나 데이터에 포함된 개인정보 침해 우려 또한 나날이 증대되고 있다. 이에 따라 데이터에 포함된 정보로 개인을 유추할 수 없도록 개인정보를 비식별화 또는 익명화하는 기법이 널리 연구되었으며, 2016년에는 정부 관계부처에서 '개인정보 비식별 조치 가이드라인' [1]을 제시하여 비식별 처리 기준 및 관리체계를 구체화 하였다.

개인정보 비식별 처리를 위해 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 등의 기법을 이용할 수 있으며, 이 기법들을 이용하여 k-익명성 (k-anonymity) 모델[2], ℓ-다양성 (ℓ-diversity) 모델[3], t-근접성 (t-closeness) 모델[4] 등을 만족하게 데이터를 수정한다.

<표 1> 비식별 대상 원본데이터

이름	연령	주소	질병
김 일	21	서울시 강남구 개포동	감기
이 이	31	서울시 강동구 명일동	고혈압
박 삼	40	서울시 강서구 등촌동	고혈압
최 사	22	서울시 강남구 대치동	위암
정 오	32	서울시 강동구 암사동	감기
강 육	43	서울시 강서구 방화동	고혈압
유 칠	23	서울시 강남구 역삼동	위암
장 팔	34	서울시 강동구 천호동	위암
조 구	45	서울시 강서구 화곡동	위암

'개인정보 비식별 조치 가이드라인' 에서는 k-익명성 모델을 최소한의 평가 기준으로 설명하고 있다. k-익명성 모델은 '주어진 데이터 집합에서 같은 값이 적어도 k개 이상 존재하도록 하여 쉽게 다른 정보로 결합할 수 없도록 함' 으로 정의되어 있다. <표2>와 <표3>은 원본데이터 <표1>에서 개인을 식별할 수 있는 '이름' 을 삭제하고, 개인을 추론할 수 있는 '연령', '주소' 를 범주화 하여 k-익명성 모델 (k=3)을 만족하게 만든 비식별 처리된 데이터이다. ('질병' 은 민감한 속성이므로 그대로 남겨 두었으며, {연령, 주소}의 데이터 집합으로 같은 값이 3개씩 있어 3-익명성을 만족한다.)

<표 2> k-익명성 (k=3)으로 비식별 처리된 데이터 1

연령	주소	질병
21-40	서울시	감기
21-40	서울시	고혈압
21-40	서울시	고혈압
22-43	서울시	위암
22-43	서울시	감기
22-43	서울시	고혈압
23-45	서울시	위암
23-45	서울시	위암
23-45	서울시	위암

<표 3> k-익명성 (k=3)으로 비식별 처리된 데이터 2

연령	주소	질병
21-23	서울시 강남구	감기
21-23	서울시 강남구	고혈압
21-23	서울시 강남구	고혈압
31-34	서울시 강동구	위암
31-34	서울시 강동구	감기
31-34	서울시 강동구	고혈압
40-45	서울시 강서구	위암
40-45	서울시 강서구	위암
40-45	서울시 강서구	위암

<표2>와 <표3>은 동일한 기법을 사용하여 동일한 수준으로 비식별 처리된 데이터지만 <표3>이 <표2>에 비해 원본데이터와 차이가 적다. 이 차이의 발생 원인은 비식별 처리 과정에서 원본데이터의 데이터 분포 및 특성의 고려 여부이다. <표2>는 데이터 분포를 고려하지 않은 채 원본데이터의 순서대로 '연령', '주소'를 범주화 하였고, <표3>은 데이터 분포를 고려하여 동일한 작업을 수행한 결과이다.

비식별 처리 과정에서 원본데이터와 비식별 처리된 데이터의 차이를 정보손실(Information Loss)이라 한다. 정보손실 측정 방법은 학계에서 연구하고 있으나 '개인정보 비식별 조치 가이드라인'에서는 언급하고 있지 않다.

본 논문에서는 널리 알려진 정보손실 측정 방법 연구에 대해 소개한다. 분량의 제한으로 본 논문의 조사 범위는 데이터 범주화를 통한 k-익명성 모델을 만족하는 비식별 처리 데이터의 정보손실 측정 방법으로 한정한다.

2. 정보손실 측정기법

비식별 처리된 데이터의 정보손실 측정을 위해서는 두 가지 방법을 사용할 수 있다. 첫 번째는 동일한 분석작업 또는 쿼리를 수행하여 결과의 차이를 비교하는 방법이다. 두 번째 방법은 비식별 처리된 데이터의 정보손실의 양적 측정기법을 개발해서 이용하는 것이다. 본 논문에서는 정보손실의 양적 측정기법을 소개할 것이다.

2.1. 범주화 횟수 측정

정보손실을 측정하기 위해 초기에 제시된 방법은 범주화된 작업의 횟수를 계산하는 것이다 [5], [6].

예제 1. <표2>는 <표1>을 범주화하여 비식별 처리된 데이터이다('이름'은 반드시 삭제해야 하는 식별자 이므로 정보손실에 포함시키지 않는다.). '연령' 속성에서는 총 9개의 튜플이 범주화 되었으며, '주소' 속성에서도 총 9개의 튜플이 범주화 되어 총 18개의 범주화 작업이 발생하였다. □

이 방법은 범주화 작업의 횟수만 측정하기 때문에 간단하고 직관적으로 정보손실을 측정할 수 있다. 그러나 모든 범주화 작업을 같은 손실로 간주하는 단점이 있다.

예제 2. <표3>은 <표1>을 범주화하여 비식별 처리된 데이터이며, 총 18개의 범주화 작업이 발생하였다. □

<표3>이 <표2>에 비해 원본데이터의 정보를 더 많이 보존하고 있지만 같은 횟수의 범주화 작업을 수행했기 때문에 정보손실이 같게 계산된다. 이러한 문제점을 보완하기 위해 범주화 양 계산 방법이 연구되었다.

2.2. 범주화 양 측정

범주화 양 계산 방법은 비식별 처리된 데이터의 각 튜플의 속성마다 손실 값을 측정하여 한 튜플의 정보손실을 계산하고, 전체 데이터의 정보손실은 데이터에 속한 모든 튜플의 정보손실의 합으로 계산한다 [7], [8], [9]. 이 때, 정보손실은 각 속성의 도메인 크기와 비식별 처리된 속성정보의 적용범위 크기 비율로 계산된다. 이 때, 속성 값이 숫자형 데이터인 경우 수의 범위가 크기가 되며, 속성 값이 범주형 데이터인 경우 그림 1과 같은 분류 트리를 이용하여 크기를 계산한다.

수식 1. 정보손실 계산

비식별 처리된 데이터 D의 튜플 t, 속성 A에 대해 $|A|$를 속성 A의 도메인 크기라 하고, t[A]를 튜플 t의 속성 A 값이 포함하는 범위라 할 때, 튜플 t의 속성 A에 대한 정보손실은

$$\frac{t[A]}{|A|}$$

으로 계산된다. 튜플 t의 정보손실은 각 속성들의 정보손실의 합이며, 데이터 D의 정보손실은 모든 튜플의 정보손실의 합과 같다. □

예제 3. <표2>의 첫 번째 튜플 t_1의 정보손실은 '연령' 속성과 '주소' 속성의 정보손실의 합과 같다. '연령' 속성의 도메인 크기는 25이며 (데이터의 연령 범위가 21~45), '주소' 속성의 도메인 크기는 <그림1>의 분류 트리의 리프 노드 수인 9이다. t_1[연령]은 20, t_1[주소]는 9 ('서울시'가 포함하는 리프 노드 범위는 9) 이다. 따라서 t_1의 정보손실은 20/25 + 9/9 = 1.8 이다. t_3, t_3의 정보손실도 t_1과 같은 1.80이다. 같은 방식으로 t_4(t_5, t_6), t_7(t_8, t_9)의 정보손실을 계산하면 각각 1.88, 1.92 이다. <표2> 데이터의 정보손실은 (1.80*3 + 1.88*3 + 1.92*3) = 16.80 이다. □

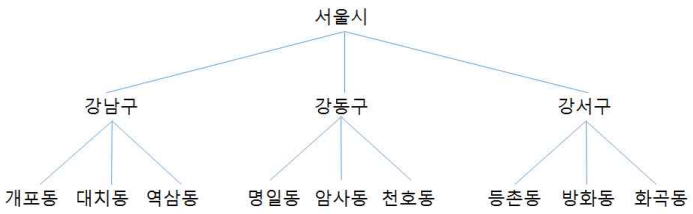

그림 1. 주소에 대한 분류 트리

예제 4. <표3>의 데이터도 같은 방식으로 정보손실을 계산 할 수 있다. 튜플 t_1, t_2, t_3의 정보손실은 각각 3/25 + 3/9 = 0.453 이다 ('강남구'가 포함하는 리프 노드 범위는 3). t_4, t_5, t_6의 정보손실은 각각 4/25 + 3/9 = 0.493 이며, t_7, t_8, t_9의 정보손실은 각각 6/25 + 3/9 = 0.573 이다. <표3> 데이터의 정보손실은 0.453*3 + 0.493*3 + 0.573*3 = 4.557 이다. □

이와 같이 범주화 양 계산하면 원본데이터의 정보를 많이 유지할수록 정보손실이 적게 측정되는 결과를 얻을 수 있다. 그러나 숫자형 데이터를 제외한 데이터에 대한 분류 트리가 반드시 필요하며, 속성의 도메인 크기에 따라 각 속성의 가중치가 왜곡될 수 있는 단점이 있다.

2.3. 모호성 측정

모호성 측정 방법은 Nergiz 와 Clifton에 의해 제안되었다[10]. 이 방법은 비식별 처리된 데이터의 각 튜플 t에 대해 원본데이터에서 t로 매핑(mapping)될 수 있는 튜플의 수를 측정한다. 이 측정수치가 클수록 비식별 처리된 데이터와 원본데이터의 연관성이 사라져 데이터 간의 모호성이 커진다.

예제 5. 비식별 처리된 데이터 <표2>의 첫 번째 튜플 t_1에 대해 t_1 으로 매핑될 수 있는 원본데이터 튜플 집합 P는 { 김일, 이이, 박삼, 최사, 정오, 유칠, 장팔 } 이다. 튜플 t_3, t_3도 t_1과 같은 집합을 가진다. 튜플 t_4, t_5, t_6에 매핑되는 원본데이터 튜플 집합 Q는 { 이이, 박삼, 최사, 정오, 강육, 유칠, 장팔 } 이며, 튜플 t_7, t_8, t_9에 매핑되는 원본데이터 튜플 집합 R은 { 이이, 박삼, 정오, 강육, 유칠, 장팔, 조구 } 이다. 정보손실은 (3*｜P｜ + 3*｜Q｜ + 3*｜R｜ = 21 + 21 + 21 = 63 이다. □

예제 6. <표3>의 데이터도 같은 방식으로 정보손실을 계산 할 수 있다. 튜플 t_1, t_2, t_3에 매핑되는 원본데이터 튜플 집합 P는 { 김일, 최사, 유칠 }, 튜플 t_4, t_5, t_6에 매핑되는 원본데이터 튜플 집합 Q는 { 이이, 정오, 장팔 }, 튜플 t_7, t_8, t_9에 매핑되는 원본데이터 튜플 집합 R은 { 박삼, 강육, 조구 } 이다. 정보손실은 (3*｜P｜+ 3*｜Q｜ + 3*｜R｜ = 9 + 9 + 9 = 27 이다. □

2.4. 엔트로피(Entropy) 측정

정보손실 측정에 엔트로피를 이용한 연구도 있다. 각 속성 값의 엔트로피 합을 이용한 연구 [11], 원본데이터와 비식별 처리된 데이터를 비교하여 속성 값의 엔트로피 차이를 이용한 연구 [12], 엔트로피의 전반적인 변화(Overall Change)를 이용한 연구 [13], 조건부 엔트로피를 사용한 연구 [14] 등이 있다.

3. 결 론

빅데이터, 인공지능 등의 분야가 발전하면서 분석을 위한 데이터 공개의 요구는 점차 커지고 있다. 데이터 공개를 위해서는 필연적으로 데이터에 포함된 개인정보 보호 조치를 취해야한다. 그러나 개인정보 보호에만 치중하여 활용도가 떨어지는 데이터를 공개한다면 본래의 목적과 취지에 부합하지 않는 일이 된다. 본 논문에서는 비식별 처리된 데이터의 정보손실을 측정하는 기법들을 조사하고 소개하여 활용할 수 있게 하였다.

향후 연구로는 비식별 처리된 데이터의 정보손실 측정을 위해 동일한 분석작업 또는 쿼리를 원본데이터와 비식별 처리된 데이터에 수행하여 결과의 차이를 비교하는 방법으로 정보손실을 측정하는 방법에 대한 조사가 있다. 또한, 분량의 제한으로 인해 생략한 여러 비식별 기법 (가명처리, 총계처리, 데이터 삭제, 데이터 마스킹 등) 및 모델 (ℓ-다양성 모델, t-근접성 모델)에 대한 조사연구도 진행할 예정이다.

참 고 문 헌

[1] 개인정보 비식별조치 가이드라인, 2016.

[2] L. Sweeney, *k*-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 557-570, 2002.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramanian, ℓ-diversity: Privacy beyond k-anonymity, in Proceedings of the IEEE International Conference on Data Engineering, 2006.

[4] N. Li, T. Li and S. Venkatasubramanian, *t*-Closeness: Privacy beyond k-anonymity and l-diversity, in Proceedings of the IEEE International Conference on Data Engineering, 2007.

[5] A. Meyerson and R. Williams, On the complexity of optimal k-anonymity, in Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 2004.

[6] P. Samarati, Protecting repondents' identities in microdata release, in Transactions on Knowledge and Data Engineering, pp.1010-1027, 2001.

[7] V. S. Iyengar, Transforming data to satisfy privacy constraints, in ACM International Conference on Knowledge Discovery and Data Mining, 2002.

[8] M. K. Sung, K. Y. Lee, J. B. Shin and Y. D. Chung, A privacy protection method for social network data against content/degree attacks, IEICE Transaction on Information and Systems, Vol.E95-D, pp.152-160, 2012.

[9] 성민경, 이기용, 정연돈, 소셜 네트워크에서 구조정보과 내용정보를 고려한 프라이버시 보호 기법, 한국컴퓨터정보학회논문지, 15(1), pp.119-128, 2010.

[10] M. E. Nergiz and C. Clifton, Thoughts on k-anonymization, Data & Knowledge Engineering, Vol. 63, No. 3, pp. 622-645, 2007.

[11] A. Gionis and T. Tassa, k-anonymization with minimal loss of information, in TKDE, 2008.

[12] S. Xu and X. Ye, Risk & distortion based k-anonymity, Information Security Application, 2008.

[13] S. Gomatam and A. F. Karr, Distortion measures form categorical data swapping, Technical Report, National Institute of Statistical Sciences, Technical Report, No. 131, 2003.

[14] L. Willenborg and T. de Wall, Elements of Statistical Disclosure Control, Springer, 2000.