산업기술연구(강원대학교 산업기술연구소 논문집), 제36권, 2016.
Journal of Industrial Technology, Kangwon Natl. Univ., Korea, No. 36, 2016.

익명 그룹 기반의 효율적인 데이터 익명화 알고리즘

An Efficient Algorithm of Data Anonymity based on Anonymity Groups

권 호 열*
Kwon, Ho Yeol

Abstract

In this paper, we propose an efficient anonymity algorithm for personal information protections in big data systems. Firstly, we briefly introduce fundamental algorithms of *k*-anonymity, *l*-diversity, *t*-closeness. And then we propose an anonymity algorithm using controlling the size of anonymity groups as well as exchanging the data tuple between anonymity groups. Finally, we demonstrate an example on which proposed algorithm applied. The proposed scheme gave an efficient and simple algorithms for the processing of a big amount of data.

키워드 : *익명화, k-*익명성, *l-*다양성, *t-*근접성, *빅 데이터*
Keywords : *anonymity, k-anonymity, l-diversity, t-closeness, big data*

1. 서론

대규모 데이터를 수집 및 분석하여 유용한 정보를 얻는 빅데이터 시스템에서 민감한 개인 정보를 보호하기 위하여 데이터의 개인식별성을 낮추는 익명화 기법이 필요하다.

데이터 익명화 기법은 원시 데이터의 집단화를 통하여 개인정보를 보호하고 프라이버시를 확보하는 방법으로서 *k*-익명성(k-anonymity)[1-3] 개념이 발표된 이래, *l*-다양성(l-diversity)[4], *t*-근접성(t-closeness)[5], *s*-균일성(uniformity)[6] 등으로 발전하였으며, 최근에는 데이터 스트림 환경 등 다양한 환경에서의 익명화[7]~[9], 익명화 척도[10]에 대한 연구가 이루어져 왔다.

본 연구에서는 빅데이터 시스템의 효율적인 개인정보보호를 구현하기 위하여 지금까지 제안된 주요 익명성 개념을 소개하고, 이를 바탕으로 효율적인 익명성 알고리즘를 제안하고 제안한 방법의 유효성을 보이기 위하여 예제를 이용한 실험을 수행하였다. 제안된 방법은 기존 방법에 비하여 단순한 알고리즘으로서 대량의 데이터 처리에 효과적으로 응용될 것으로 기대된다.

* 강원대학교 컴퓨터정보통신공학전공 교수, 공학박사, 교신저자

2. 데이터 익명화 기법

통계적으로 수집된 빅데이터는 이름 및 주민등록번호와 같이 개인의 신원을 직접적으로 나타내는 식별자(IDentifier, ID) 정보, 해당 정보만으로는 대인 식별이 불가능하지만 데이터의 조합을 통해 간접적으로 개인 식별이 가능한 준식별자(Quasi-Identifier, QI) 정보, 개인의 프라이버시에 관계된 병명정보와 같은 민감 속성(Sensitive attribute, S) 정보 등으로 구성된다. 익명화 처리는 제3자에게 제공되는 민감 속성(S)에 대한 프라이버시 보호를 위해서 식별자(ID)를 제거하고 준식별자(QI)를 익명화한다.

먼저 원시데이터의 익명성을 살펴보자. 표 1에 나타낸 2개의 데이터는 독립적으로는 민감한 정보를 노출하지 않으나 결합 공격을 받으면 철수는

감기를, 준호는 몸살을, 우치는 기관지염을 앓고 있음을 알 수 있다. 여기서 이름은 ID 정보, 나이, 성별, 우편번호는 QI 정보, 병명은 S 정보에 해당한다.

표 1. 배포된 데이터 (a) 유권자정보 (b) 의료정보

	ID	QI			QI			S
	이름	나이	성별	우편번호	나이	성별	우편번호	병명
t1	철수	23	M	11324	23	M	11324	감기
t2	준호	24	M	23124	24	M	23124	몸살
t3	수일	25	M	12312	25	M	12312	장염
t4	중배	26	M	12345	26	M	12345	장염
t5	순애	28	F	32141	28	F	32141	감기
t6	길동	29	M	12139	29	M	12139	피부염
t7	몽룡	30	M	22140	30	M	22140	폐렴
t8	우치	35	M	32111	35	M	32111	기관지염

(a) (b)

k-익명성(anonymity) 모델[1]은 원래 데이터 값은 보존하되 집단화를 통하여 데이터 테이블에서 모든 튜플이 서로 구분되지 않는 *k*-1개 이상의 후보 튜플을 갖도록 변환한다. 이와 같은 익명화 과정으로 준식별자 속성값들이 같아져 서로 구별할 수 없는 튜플들의 그룹을 동질 클래스(Equivalent Class)라고 한다.

표 2는 8개의 데이터를 2개씩 4개의 그룹으로 묶은 것으로써 표 1의 데이터를 2-익명성 처리한 것이다.

표 2. *k*-익명성 처리된 데이터

	나이	성별	우편번호	병명
t1	[23,24]	M	11***	감기
t2	[23,24]	M	23***	몸살
t3	[25,26]	M	12***	장염
t4	[25,26]	M	12***	장염
t5	[28,29]	Person	32***	감기
t6	[28,29]	Person	12***	피부염
t7	[30,35]	M	22***	폐렴
t8	[30,35]	M	32***	기관지염

k-익명성 기법에서는 익명화된 데이터 테이블의 t3와 t4처럼 집단화된 튜플들의 민감 속성(S)인 병명의 값이 동일한 경우 나이, 성별, 우편번호 정보(QI) 만 있으면 동질성 공격에 의하여 장염을 앓고 있다는 개인 정보가 추론될 수 있다.

l-다양성(diversity) 모델[2]은 이러한 취약점을 해결하기 위한 것으로써 동질 클래스 내의 서로 구분되지 않는 튜플들 사이에서 민감 속성은 최소한 *l* 개 이상이어야 프라이버시가 보호된다는 점을 이용한다.

표 3은 그룹의 크기를 확장함으로써 각 그룹마다 민감 속성이 2개 이상 존재하도록 만든 것이다. 표 2의 데이터에 대하여 2-다양성 처리가 이루어진 것을 나타낸다.

표 3. *l*-다양성 처리된 데이터

	나이	성별	우편번호	병명
t1	[23,25]	M	11***	감기
t2	[23,25]	M	23***	몸살
t3	[23,25]	M	12***	장염
t4	[26,29]	Person	12***	장염
t5	[26,29]	Person	32***	감기
t6	[26,29]	Person	12***	피부염
t7	[30,35]	M	22***	폐렴
t8	[30,35]	M	32***	기관지염

한편, *l*-다양성 기법에서는 민감 속성의 값이 불균형하게 분포될 수가 있어, 하나의 동질 클래스에 분류 체계 상 가까운 관련 값들이 몰릴 수 있다.

t-근접성(closeness)[3] 기법은 민감 속성의 도메인 분류 체계를 고려하여 고른 분포를 보장한다. *t*-근접성 기법은 먼저 *k*-익명성 처리와 *l*-다양성 처리를 수행한 후, 데이터 분류 체계상 근접도와 동질 클래스내의 근접도를 계산하여 가장 차이가 큰 근접도를 갖는 동질 클래스의 튜플을 다른 동질 클래스의 EC의 튜플과 교환하는 과정을 거쳐 분류 체계상의 근접도를 고려한 고른 분포를 갖게 한다.

표 4는 t7, t8 그룹이 갖는 민감 정보가 폐렴과 기관지염으로 분류체계상 의미적 근접도가 높으므로 t7을 이웃한 동질 클래스의 t6와 교환함으로써 t6, t8으로 이루어진 그룹에서는 민감 정보가 분산되도록 한 것으로서 표 3의 데이터를 t-근접성 처리한 것이다.

표 4. *t*-근접성 처리된 데이터

	나이	성별	우편번호	병명
t1	[23,25]	M	11***	감기
t2	[23,25]	M	23***	몸살
t3	[23,25]	M	12***	장염
t4	[26,30]	Person	12***	장염
t5	[26,30]	Person	32***	감기
t7	[26,30]	Person	22***	폐렴
t6	[29,35]	M	12***	피부염
t8	[29,35]	M	32***	기관지염

s-균일성(uniformity)[4] 기법은 각 민감 정보의 민감도를 평가한 후, t-근접성 기법에서 분류체계상 의미적 근접도 대신 민감 정보의 민감도를 고르게 분포하도록 하는 방법이다. 예를 들어 민감 정보인 병명에서 민감도는 해당 질병의 사망률이 될 수 있다.

3. 제안된 익명그룹 기반 익명화 방법

앞에서 살펴본 k-익명성, l-다양성, t-근접성 등 방법에서의 데이터 처리는 1) 튜플들의 집단화 및 2) 익명 집단 간 튜플 교환 등 주로 두 가지 동작을 통하여 익명성을 개선한다. 따라서 튜플들로 이루어진 익명 그룹의 익명도를 평가하여 익명성이 원하는 수준에 이르도록 반복적으로 익명 집단의 크기를 증가시키는 작업과 익명집단 간의 튜플 교환 작업을 반복적으로 수행함으로써 익명성을 개선할 수 있다.

일반적으로 데이터의 식별성(IDF_{Data})은 테이블 내의 총 데이터 수(N_{total})에 대하여 개인정보 식별이 가능한 데이터의 수(N_{ident})로 나타낼 수 있다. 만일 모든 데이터의 개인정보가 식별 가능하다면 식별성은 IDF = 1 이며, 전혀 식별가능하지 않다면 IDF = 0 이 된다. 데이터의 익명성(ANM_{Data})은 데이터의 식별성과 서로 역의 관계를 갖는다.

표 5. 익명그룹 기반 익명화 알고리즘

```
입력: N개의 데이터 튜플, 목표 익명도 ANM 설정
출력: M개의 익명그룹으로 집단화 분류된 데이터 튜플
k*-익명성 처리 {
    N개의 데이터 튜플을 읽어온다
    i = 0
    do {
       // 익명그룹의 크기 조정
       i = i +1 //익명그룹의 크기 조정
        i 개의 튜플로 이루어진 M개의 익명그룹 구성
        익명그룹들의 익명도 ANM_AG 계산
        전체 익명도 Σ_AG ANM_AG 계산
        if(Σ_AG ANM_AG < ANM) break

       // 최소 익명그룹 튜플의 교환
        최소익명도 ANM_AG,MIN 익명그룹 선정
            선정된 그룹의 튜플을 이웃 그룹과 교환
        익명그룹들의 익명도 ANM_AG 계산
        전체 익명도 Σ_AG ANM_AG 계산
    } while (Σ_AG ANM_AG < ANM)
    익명성 처리 종료
}
```

한편, 익명집단(Anonymity Group, AG)은 동일한 준식별자(QI)에 의하여 정의되는 데이터 튜플들의 집합이다. 익명집단의 익명성(ANM_{AG})은 익명집단(AG) 내 데이터 민감 속성(S_i, S_j)들이 갖는 거리의 합으로 나타낼 수 있다. 이 때 거리 척도는 t-근접성에서 사용하는 분류체계 상의 근접도 뿐만 아니라, s-균일성에서 사용하는 민감도 등이 사용될 수 있다. 전체 데이터의 익명도는 익명그룹익명도의 합이다.

$$ANM_{AG} = \sum_i \sum_j |S_i - S_j| \quad (1)$$

$$ANM = \sum_{AG} ANM_{AG} \quad (2)$$

본 연구에서 제안하는 k^*-익명성 처리 알고리즘은 표 5와 같다.

4. 실험 결과

앞에서 소개한 예제를 본 연구에서 제안한 방법으로 익명화하는 과정을 표 6에서 표 9까지 나타내었다.

표 6. 익명화하기 전 원시 데이터

	이름	나이	성별	우편번호	병명
t1	철수	23	M	11324	감기
t2	준호	24	M	23124	몸살
t3	수일	25	M	12312	장염
t4	중배	26	M	12345	장염
t5	순애	28	F	32141	감기
t6	길동	29	M	12139	피부염
t7	몽룡	30	M	22140	폐렴
t8	우치	35	M	32111	기관지염

표 7. i = 2 로 익명그룹화

	나이	성별	우편번호	병명
t1	[23,24]	M	11***	감기
t2	[23,24]	M	23***	몸살
t3	[25,26]	M	12***	장염
t4	[25,26]	M	12***	장염
t5	[28,29]	Person	32***	감기
t6	[28,29]	Person	12***	피부염
t7	[30,35]	M	22***	폐렴
t8	[30,35]	M	32***	기관지염

표 8. 익명그룹간 교환: t4, t5

	나이	성별	우편번호	병명
t1	[23,24]	M	11***	감기
t2	[23,24]	M	23***	몸살
t3	[25,28]	Person	12***	장염
t5	[25,28]	Person	32***	감기
t4	[26,29]	M	12***	장염
t6	[26,29]	M	12***	피부염
t7	[30,35]	M	22***	폐렴
t8	[30,35]	M	32***	기관지염

표 9. 익명 그룹간 교환: t6, t7

	나이	성별	우편번호	병명
t1	[23,24]	M	11***	감기
t2	[23,24]	M	23***	몸살
t3	[25,28]	Person	12***	장염
t5	[25,28]	Person	32***	감기
t4	[26,30]	M	12***	장염
t7	[26,30]	M	22***	폐렴
t6	[29,35]	M	12***	피부염
t8	[29,35]	M	32***	기관지염

기존의 k-익명성, l-다양성, t-근접성 처리를 차례로 거친 결과인 표 4와 본 연구에서 제안한 방법에 의한 결과인 표 9를 비교하면 단순한 과정을 거쳐 익명화가 얻어지며 최종결과인 익명화그룹의 수도 증가함을 알 수 있다.

5. 결론

본 연구에서는 빅데이터 시스템에 적용할 수 있는 효율적인 개인정보 익명화 알고리즘에 대하여 논하였다. 제안된 알고리즘은 기존의 k-익명성, l-다양성, t-근접성 처리를 차례로 거친 경우에 비하여 단순한 알고리즘을 반복적으로 사용하며 예제에서 보는 바와 같이 최종 결과가 보다 많은 익명그룹을 유지하는 장점이 있다.

본 연구와 관련하여 향후 연구해야할 과제는 제안된 익명성 알고리즘을 실데이터에 적용하여 성능을 평가하는 것과 수시로 데이터의 추가, 삭제, 갱신이 이루어지는 동적인 데이터에 대한 익명화 방안 등이다.

참 고 문 헌

[1] L. Sweeney, “k-Anonymity: A model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based System, vol.10, no.3, pp.557-570, 2002.

[2] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. R., “Incognito - Efficient full-domain k-anonymity”, Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005.

[3] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity”, Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, 2006.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy beyond k-Anonymity”, Proceedings of Int'l Conference on Data Engineering, pp.24-35, 2006.

[5] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-anonymity and l-diversity”, ICDE, pp.106-115, 2007.

[6] 고한영, 정강수, 박석, “s-uniformity: 데이터의 민감도를 고려하여 고른 분포를 보장하는 익명화 기법”, *2009 한국컴퓨터종합학술대회 논문집*, Vol.36, No.1(C), pp.70-75, 2009.

[7] 성민경, 정연돈, “데이터 스트림에서 프라이버시 보호를 위한 익명화 기법”, *정보과학회 논문지 : 데이타베이스*, 41(1), pp.8-20, 2014. 2.

[8] 황치광, 최종원, 홍충선, “데이터 유용성 향상을 위한 서비스 기반의 안전한 익명화 기법 연구”, *정보과학회논문지*, 42(5), pp.681-689, 2015. 5.

[9] 강주성, 강진영, 이옥연, 홍도원, “익명성 관련 측도에 기반한 데이터 프라이버시 확보 알고리즘에 관한 연구”, *정보보호학회논문지*, 21(5), pp.149-160, 2011. 10.

[10] 권호열, “빅데이터 시스템의 효율적인 데이터 익명성 척도에 관한 연구”, *2015 대한전자공학회 하계학술대회 논문집*, 제주, pp.22-24, 2015. 6.