情報保護學會論文誌
第 21 卷 第 5 號, 2011. 10

익명성 관련 측도에 기반한 데이터 프라이버시 확보 알고리즘에 관한 연구*

강 주 성,[1†] 강 진 영,[1] 이 옥 연,[1] 홍 도 원[2]
[1]국민대학교, [2]한국전자통신연구원

A study on the algorithms to achieve the data privacy based on some anonymity measures*

Ju-Sung Kang,[1†] Jin-young Kang,[1] Okyeon Yi,[1] Dowon Hong[2]
[1]Kookmin University, [2]ETRI

요 약

익명화 기법은 마이크로 데이터에서 프라이버시를 보호하기 위해 제안된 방법 중의 하나이다. 원본 데이터로부터 그룹화를 기반으로 프라이버시를 확보하고자 하는 익명화 기법은 k-익명성(k-anonymity) 개념을 효시로 하여 ℓ-다양성(ℓ-diversity), t-밀접성(t-closeness) 등의 개념이 차례로 제안되면서 발전된 모습을 보여주었다. 프라이버시 측도 관점에서 각각의 익명성 관련 개념들이 상호 보완적인 관계에 놓여 있으나, 데이터의 유용성과 익명성 개념들을 복합적으로 고려한 실질적인 익명화 알고리즘 개발에 관한 연구는 아직까지 미진한 상태이다. 본 논문에서는 먼저 기존에 발표된 익명성 개념들에 기반한 익명성 측도들과 정확성 관련 측도들에 대하여 비교 분석한다. 또한, k-익명성을 만족하는 데이터로부터 블록 합병 방법에 의하여 ℓ-다양성을 확보하는 알고리즘을 새롭게 제안한다.

ABSTRACT

Technique based on the notions of anonymity is one of several ways to achieve the goal of privacy and it transforms the original data into the micro data by some group based methods. The first notion of group based method is k-anonymity, and it is enhanced by the notions of ℓ-diversity and t-closeness. Since there is the natural tradeoff between privacy and data utility, the development of practical anonymization algorithms is not a simple work and there is still no noticeable algorithm which achieves some combined anonymity conditions. In this paper, we provides a comparative analysis of previous anonymity and accuracy measures. Moreover we propose an algorithm to achieve ℓ-diversity by the block merging method from a micro-data achieving k-anonymity.

Keywords: Privacy, Accuracy, k-anonymity, l-diversity, t-closeness.

접수일(2011년 7월 13일), 게재확정일(2011년 10월 7일)
* 본 논문은 2011년도 국민대학교 교내연구비와 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단-차세대정보컴퓨팅기술개발사업(No. 2011-0029925)의 지원을 받아 수행된 연구임
† 주저자, jskang@kookmin.ac.kr

I. 서 론

컴퓨터의 계산 능력과 네트워크 시스템 기술의 비약적인 발달과 함께 대용량 데이터에 대한 공유가 편리해진 이면에는 개인의 프라이버시나 조직의 기밀에 대한 위협이 상대적으로 증가하고 있다는 부작용이 존재한다. 오늘날의 사회는 다양한 목적을 위해 개인 정보가 포함된 데이터의 수집과 공유를 필요로 한다. 국가 및 공공기관이나 기업에서는 다양한 목적을 위하여 개인 정보가 내재되어 있는 데이터의 수집과 공유를 필요로 한다. 이러한 데이터 내에 명백한 식별자 정보가 포함되어 있지 않을 경우조차도 연결공격(linkage attack)[1] 등을 통하여 프라이버시 관련 정보가 노출될 가능성이 존재한다. 프라이버시 관련 정보가 포함된 데이터의 경우 이름, 주민등록번호 등과 같은 명백한 식별자 기록만을 삭제하여 데이터를 공유하는 단순한 방법만으로는 우리가 원하는 수준의 프라이버시를 달성하기가 어렵다. 그러므로 개인의 프라이버시를 보호하기 위한 좀 더 면밀하고 기술적인 방법이 필요하다.

1.1 익명성 관련 개념의 진보

프라이버시가 보호된 상태에서 유용한 정보를 공유하기 위한 실용적인 방법 중의 하나로 주목받고 있는 개념이 k-익명성(k-anonymity)[2]이다. k-익명성 개념은 데이터 내에 특정 자료와 식별이 불가능한 데이터 목록이 적어도 $(k-1)$개 이상 존재하도록 데이터를 변형하여 프라이버시를 확보하고, 데이터의 유용성을 보장하는 정확도(accuracy)를 일정 수준 이상으로 유지시키는 방법을 일컫는다.

프라이버시 보호 관점에서 k-익명성을 만족하는 데이터일지라도 구분되지 않는 목록 내에서 민감한(sensitive) 속성값(attribute value)들이 모두 같거나 다양하게 구분되지 않는 경우에는 민감한 속성값을 추론해낼 수 있다는 문제점이 발생한다. 이러한 문제점을 지적하면서 Machanavajjhala 등[3]은 k-익명성 개념의 프라이버시 강화 기법으로 ℓ-다양성(ℓ-diversity)이란 개념을 제안하였다. ℓ-다양성 개념은 익명성을 만족하는 블록들 내에서 서로 다른 민감한 속성값들이 ℓ 개 이상 존재하도록 하여 보다 높은 수준의 프라이버시 보호 요구 조건이 만족되게 하기 위한 것이다.

한편, Li 등[4]은 k-익명성과 ℓ-다양성을 만족하는 마이크로 데이터라 할지라도 민감한 속성에 대한 프라이버시 노출이 가능할 수 있음을 지적하면서 새로운 프라이버시 개념으로 t-밀접성(t-closeness)이란 것을 제안하였다. k-익명성을 달성하기 위해서는 마이크로 데이터에서 식별 불능인 기록(record)들의 집합인 동치클래스(equivalence class) 내에 적어도 k개의 기록들이 존재해야 하고, ℓ-다양성 조건을 만족하기 위해서는 각각의 동치클래스 내에 적어도 ℓ 개의 서로 다른 민감한 속성값들이 존재해야 한다. 하지만 이러한 조건을 만족하는 마이크로 데이터인 경우에도 어떤 동치클래스 내의 민감한 속성값의 분포가 전체 속성값의 분포와 차이가 많이 생길 경우에는 프라이버시 정보의 노출이 가능하다는 것이 t-밀접성 개념이 제안된 배경이다. 두 분포들 사이의 거리를 측정하여 그 값이 t 이하가 되도록 하면 프라이버시 정보 노출을 제어할 수 있다는 개념이 바로 t-밀접성 개념이다. 여기에서 사용된 분포들 사이의 거리 측도는 EMD(Earth Mover's Distance)이다.

비교적 최근에 발표된 t-밀접성 개념은 k-익명성과 ℓ-다양성 개념만으로는 해결할 수 없는 프라이버시 문제를 다룬 측도를 제공해주었다는 점에서는 의미가 있다고 할 수 있다. 하지만 주어진 데이터로부터 t-밀접성 조건까지 만족된 마이크로 데이터로 변환시키는 실질적인 방법은 그리 단순하지 않다. k-익명성과 ℓ-다양성의 조건이 만족된 상태에서 t-밀접성 요구 조건이 달성되어야할 뿐만 아니라 데이터 유용성을 결정짓는 정확도 수준도 고려되어야 하기 때문이다. 본 논문에서는 아직까지 크게 진전된 연구 결과가 발표되지 않고 있으나 실용적인 관점에서는 매우 중요한 실질적인 익명화 적용 방법에 초점을 맞추고자 한다. 이를 위해서 우선 k-익명성이 달성된 마이크로 데이터로부터 ℓ-다양성 요구 조건을 만족시키는 알고리즘에 대해서 연구한다. t-밀접성 요구 조건까지 만족시키는 알고리즘에 관한 연구는 향후의 과제로 남기고 여기에서는 t-밀접성 관련 논의를 더 이상 언급하지 않기로 한다.

1.2 본 논문의 연구 결과

본 논문에서는 먼저 k-익명성을 달성하기 위한 알고리즘들 중에서 대표적인 일반화(generalization) 기법과 특수화(specialization) 기법에 대한 심층적인 분석을 실시한다. 이들 기법에 사용된 프라이버시 및 정확도 관련 측도(measure)들을 조사 분석하고, 상호 관계를 탐구하며, 대표적 알고리즘인 MinGen[5]

과 TDS[6]에 대한 특성과 장단점을 비교분석한다. 기존의 정확성 관련 측도들은 제안한 저자들 나름대로 상이한 적용 대상을 생각하고, 각각 특색 있는 기호들을 사용하였으므로 통찰력 있게 전체를 비교하여 그 의미를 분석한다는 것이 그리 간단한 작업만은 아니라 할 수 있다.

ℓ-다양성 개념의 원저자들은 k-익명성만이 달성된 데이터의 프라이버시 문제점을 지적하고 이를 해결하기 위한 대책으로 ℓ-다양성 개념을 제시하였을 뿐, ℓ-다양성을 만족시키기 위한 구체적 방법론을 제안하지는 못했다. 본 논문에서는 k-익명성이 만족되는 데이터로부터 ℓ-다양성 조건까지 만족되도록 하는 실질적인 방법을 제안한다. 즉, k-익명성과 ℓ-다양성을 모두 만족시킬 수 있는 구체적인 알고리즘을 제안하고 그 효율성을 논한다.

II. k-익명성 관련 측도

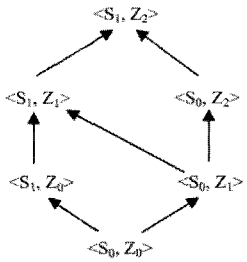
일반적으로 개인정보를 포함한 데이터는 행과 열로 이루어진 표(table)로 구성된다. 한 행의 데이터를 기록(record)이라 하며, 하나의 열은 속성(attribute)이라 부르고, n개의 속성을 갖는 데이터 표를 $T[A_1, \cdots, A_n]$과 같이 나타낸다. 데이터 표에 포함된 속성은 두 가지로 나뉜다. 식별자(identifier) $I_1, I_2, \cdots, I_l$은 이름이나 주민등록번호와 같이 개인을 식별할 수 있는 속성을 의미한다. 성별이나 우편번호와 같이 직접적으로 개인이 노출되지 않는 속성들은 준식별자(quasi-identifier)라 하여 $K_1, K_2, \cdots, K_m$으로 표현한다.

정의 1. k-익명성(k-anonymity)

데이터 표 $T[A_1, \cdots, A_n]$와 준식별자 K_T가 주어졌을 때, T가 k-익명성을 만족한다는 의미는 $T[K_T]$의 모든 기록들이 k개 이상 $T[K_T]$ 내에 존재한다는 것이다.

k-익명성을 만족하는 데이터 표가 배포될 경우, 이를 소유한 사람은 특정인이 배포된 데이터 내에 포함되어 있다는 사실은 알 수 있을지 모르지만, 정확히 어떤 기록이 특정인의 것인지 알 수 있는 확률은 $1/k$ 이하가 된다.

주어진 데이터 표를 단순히 k-익명성을 만족하도록 변환시키는 것은 그리 어려운 작업이 아니다. 하지만 k-익명성이 만족되도록 변형된 데이터 표가 원본 데이터와 너무 많은 차이를 보인다면 그 데이터의 유용성



〔그림 1〕 일반화 격자(GL)

(data utility)은 낮을 수밖에 없다. 그러므로 k-익명성을 만족한 상태에서 데이터 유용성을 유지시키기 위한 변형 방법이 의미를 갖게 될 것이고, 이 데이터 유용성을 측정하기 위한 정확성 측도(accuracy measure)가 필요할 것이다. 여기에서는 지금까지 발표된 정확성 측도들 중 주목할 만한 것들을 비교 분석해본다.

2.1. 높이(height)

높이(height) 개념은 Truta-Bindu[7]에 의하여 소개된 것으로 두 개 이상의 속성을 일반화시킬 때, 데이터 소유자가 〔그림 1〕과 같이 일반화된 도메인의 가능한 모든 조합을 나타내는 일반화 격자(GL)를 고려하는 것에서 출발한다. 그림에서 S는 성별 속성을 나타내고, $S_0 = \{남, 여\}$, $S_1 = \{*\}$를 의미한다. 즉, 성별은 $S_0 \rightarrow S_1$으로 일반화 된다. Z는 우편번호 속성을 의미하고, $Z_0 \rightarrow Z_1 \rightarrow Z_2$의 단계로 일반화 된다. 여기에서

$Z_0 = \{48201, 48275, 41075, 41076, 41088, 41099\}$,

$Z_1 = \{482**, 410**\}$, $Z_2 = \{*****\}$

이다.

정의 2. GL에서의 최소원에서 노드 X까지의 최단거리를 $height(X, GL)$로 정의한다.

〔그림 1〕에서 각 노드들의 높이는 다음과 같이 계산된다.

$$height(\langle S_0, Z_0\rangle, GL) = 0,$$
$$height(\langle S_1, Z_0\rangle, GL) = 1,$$
$$height(\langle S_0, Z_1\rangle, GL) = 1,$$
$$height(\langle S_1, Z_1\rangle, GL) = 2,$$
$$height(\langle S_0, Z_2\rangle, GL) = 2,$$
$$height(\langle S_1, Z_2\rangle, GL) = 3.$$

2.2. 정확도(precision)

정확도(precision)는 일반화 기법을 소개하면서 데이터 유용성을 측정하기 위하여 Sweeny[5]가 제안한 측도이다. [표 1]은 [그림 1]의 일반화 격자에 의한 속성 단위로 시행되는 PT의 일반화를 보여준다. $GT_{[1,0]}$는 첫 번째 속성인 성별을 한 단계 일반화한 것이고, $GT_{[0,1]}$은 두 번째 속성인 우편번호를 한 단계 일반화한 것이다. 전 소절에서 소개한 측도에 의하면 두 개의 데이터 표는 1이라는 같은 높이를 갖는다. 그러나 두 개의 데이터 표를 보면 $GT_{[0,1]}$이 $GT_{[1,0]}$에 비해 더 많은 정보를 포함하고 있다는 것을 직관적으로 알 수 있다. 이것은 각 속성에 대해서 똑같이 한 단계의 일반화가 진행된 데이터 표라 할지라도 $GT_{[1,0]}$의 경우 가장 일반화된 값으로 일반화하였다는 것을 감안하지 않았기 때문이다. 전체 일반화 단계에서 몇 단계의 일반화가 진행되었는지를 고려한 측도가 정확도 개념이다.

정의 3. $PT(A_1, \cdots, A_{N_A})$와 $GT(A_1, \cdots, A_{N_A})$가 주어졌을 때, GT의 정확도 $Prec(GT)$는 다음과 같이 정의된다.

$$Prec(GT) = 1 - \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\frac{h_{ji}}{|DGH_{A_i}|}}{N \bullet N_A}.$$

여기에서 h_{ji}는 j번째 기록에서 속성 A_i가 일반화된 계층의 수를 의미하고, $|DGH_{A_i}|$는 속성 A_i의 전체 계층의 수를 의미한다. 그리고 N은 기록의 전체 개수, N_A는 속성의 전체 개수를 각각 나타낸다.

[표 1]의 예를 바탕으로 일반화된 각 데이터 표에 대한 정확도를 구해보면 다음과 같다.

[표 1] PT를 일반화한 테이블 GT의 예

S_0	Z_0
남	48201
남	48275
남	41076
남	41088
남	41099
여	48201
여	48275
여	41076
여	41088
여	41099

PT

S_1	Z_0
*	48201
*	48275
*	41076
*	41088
*	41099
*	48201
*	48275
*	41076
*	41088
*	41099

$GT_{[1,0]}$

S_0	Z_1
남	482**
남	482**
남	410**
남	410**
남	410**
여	482**
여	482**
여	410**
여	410**
여	410**

$GT_{[0,1]}$

$$Prec(GT_{[1,0]}) = 0.5,\ Prec(GT_{[0,1]}) = 0.75,$$
$$Prec(GT_{[1,1]}) = 0.25,\ Prec(GT_{[0,2]}) = 0.5.$$

네 가지 일반화된 데이터 표 모두 2-익명성 조건을 만족하므로, 정확도 측도 관점에서는 이 중에서 가장 정확도가 높은 $GT_{[0,1]}$을 선택하는 것이 현명하다고 볼 수 있다.

2.3. 비용(cost)

앞에서 소개한 두 가지 측도는 상향식인 일반화 기법에서 사용되는 반면, 앞으로 소개될 두 가지 측도는 하향식인 특수화 기법에 사용된다. 비용(cost) 개념은 Roberto 등[8]에 의해 제안된 측도이다. 비용을 계산하기 위해서는 각 속성들이 가지는 모든 값들에 전체적으로 순서를 정한 후, 그 값들을 원소로 갖는 모든 가능한 부분집합을 고려한다. 속성들의 모든 값들이 n개라면, 2^n개의 익명화된 데이터 표를 만들 수 있다. 이 데이터 표 중에서 최적의 k-익명화를 찾기 위해서 비용을 계산한다. 비용은 익명화된 데이터 표에서 하나의 기록이 다른 기록들과 얼마나 구별되지 않느냐에 기초한 측도이다. 비용의 정의는 다음과 같다.

정의 4. k-익명성 관점에서 일반화된 데이터 GT의 비용 $C(GT, k)$는 다음과 같이 정의된다.

$$C(GT, k) = \sum_{|E| \geq k} |E|^2 + \sum_{|E| < k} |GT||E|.$$

여기에서 $|E|$는 동치클래스 E에 속하는 기록들의 개수를 나타내고, $|GT|$는 전체 기록들의 개수를 나타낸다.

앞에서의 측도들은 원래의 데이터 표에서 익명화된 데이터 표로 변환되었을 때의 일반화 정도를 가지고 높이, 정확도 등을 나타낸다. 반면 비용은 원래의 데이터가 얼마나 변환되었는지는 상관없이 단순히 익명화된 데이터 표가 얼마만큼의 익명성을 갖고 있는지 나타낸다는 특징이 있다. [표 1]의 예에서 일반화된 각 데이터 표에 대한 비용을 구해보면 다음과 같다.

$$C(GT_{[1,0]}, 2) = 20,\ C(GT_{[0,1]}, 2) = 26,$$
$$C(GT_{[1,1]}, 2) = 50,\ C(GT_{[0,2]}, 2) = 50.$$

2.4. 점수(score)

점수(score)는 Fung 등[6]에 의해서 제안된 측도로 일반화에 비해 매우 효율적인 것으로 알려진 Top-Down 알고리즘에 사용된 것이다. 저자들은 [그림 2]의 분류나무에서 각 노드들을 연결한 경로들 중의 하나를 컷(cut)이라 부르고, 가장 일반화된 상태의 노드를 지나는 컷을 초기 컷으로 놓아 점수를 계산하여 속성별로 한 단계씩 특수화를 진행하는 과정이 Top-Down 알고리즘이다. 점수의 정의는 다음과 같다.

정의 5. 분류나무의 한 노드 v의 점수 $Score(v)$는 다음과 같이 정의된다.

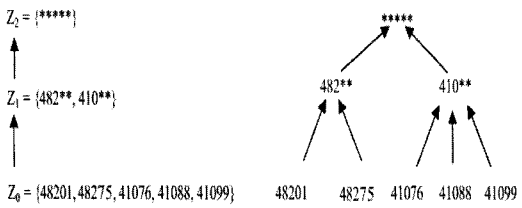
$$Score(v)=\begin{cases}\dfrac{InfoGain(v)}{AnonyLoss(v)}, & AnonyLoss(v)\neq 0\\ InfoGain(v), & AnonyLoss(v)=0.\end{cases}$$

여기에서 $InfoGain$은 정보이론의 엔트로피를 사용하여 계산되는 양이고, $AnonyLoss$는 특수화 과정 중에 발생하는 익명성 손실 정도를 나타내는 양이다. 자세한 수학적 정의는 참고문헌 [6]에 나타나 있다.

점수를 계산할 때 사용되는 양은 정보이론의 엔트로피와 연관되어 있다. [표 1]의 데이터에서 각 노드별 점수를 계산한 예는 다음과 같다.

$$Score(\text{ANY\_성별})=0.529,$$
$$Score(\text{ANY\_우편번호})=0.026,$$
$$Score(482^{**})=0.156,$$
$$Score(410^{**})=0.155.$$

한편, 점수(score)는 어떤 속성을 특수화시킬 때 얻어지는 정보량과 손실되는 익명성을 내포하고 있다. 본 논문은 다양한 측도들을 가능한 공정한 시각으로 비교하고자 하므로 점수(score)를 분류나무의 노드별 측도가 아닌 데이터 표의 정보량과 익명성을 나타내는 측도로 사용하고자 한다. 따라서 먼저 데이터 표의 일반화 단계를 나타내는 분류나무의 컷을 찾아 그 컷을 지나는 노드에서 부분나무(subtree)를 만들어 부분나무 안에 속하는 모든 노드에서의 점수(score)를 합산한 값을 정의한다. 이러한 테이블 점수는 데이터 표가 가장 특수화된 데이터가 되기까지 얻어야 할 정보량을 의미한다.



[그림 2] 도메인 일반화 및 값 일반화 계급

정의 6. 데이터 표 GT의 점수 $Score(GT)$는 다음과 같이 정의된다.

$$Score(GT)=\sum_{v}\sum_{c\in S_v} Score(c).$$

여기에서 S_v는 노드 v에서 뻗어 내린 부분나무이고, c는 S_v의 한 노드이다.

[표 1]의 예에서 각 데이터 표 GT의 점수를 계산하면 다음과 같다.

$$Score(GT_{[1,0]})=0.529,\ Score(GT_{[0,1]})=0.311,$$
$$Score(GT_{[1,1]})=0.84,\ Score(GT_{[0,2]})=0.337.$$

III. 정확성 측도들 사이의 상호 관계

3.1. 데이터 표에 대한 측도 값

데이터 표가 주어졌을 때 계산되는 네 가지 측도들 사이의 일관성과 개별성을 조사해보자. [그림 1]의 일반화 격자에 나타난 여섯 가지 데이터 표에 대한 측도 값들을 비교분석한 결과가 [표 2]에 나타나 있다. [표 2]에서 대소 관계는 데이터 표의 측도들에 대한 순서를 나타내며, 최소 일반화를 우선순위로 한 것이다.

높이는 단순히 일반화 단계 수를 의미하므로 다른

[표 2] 측도에 따른 데이터 표의 정확성

측 도	데이터 표의 정확성(accuracy)
height	[0, 0] > [1, 0]=[0, 1] > [1, 1]=[0, 2] > [1, 2]
Prec	[0, 0] > [0, 1] > [1, 0]=[0, 2] > [1, 1] > [1, 2]
Cost	[0, 0] > [1, 0] > [0, 1] > [1, 1]=[0, 2] > [1, 2]
Score	[0, 0] > [0, 1] > [0, 2] > [1, 0] > [1, 1] > [1, 2]

〔표 3〕 측도들의 특성 비교

측도	k-익명성 알고리즘 방식	측도 대상	측도의 기준
height	상향식 (bottom-up)	속성	속성의 일반화 수
Prec	상향식 (bottom-up)	속성	속성의 일반화 비율
Cost	하향식 (top-down)	기록	데이터의 익명성
Score	하향식 (top-down)	기록	정보량 및 익명성

측도들에 비해 데이터 표의 측도 값이 다양하지 못함을 볼 수 있다. 정확도는 하나의 열에 속하는 모든 값들이 같은 손실을 갖는다. 비용은 정확도와는 상이한 결과를 갖는데, 이는 원본 데이터에서의 손실을 계산하는 것이 아니라 일반화된 데이터 표의 익명성만을 고려하기 때문이다. 점수는 정보 이론적 관점에서 엔트로피를 사용하고 익명성 관련 수치로 나누어 줌으로써 데이터 왜곡 정도와 익명성 손실 정도를 동시에 고려한 측도이므로 좀 더 종합적인 의미를 담고 있다고 볼 수 있다.

3.2. 측도들에 대한 특성 분석

아래에 나타나 있는 〔표 3〕은 다양한 측도들에 대한 특성을 종합적으로 비교 분석한 결과이다. 〔표 3〕에서 보는 바와 같이 각 측도들은 그것들의 다양한 특성 때문에 〔표 2〕에서와 같이 일관성을 갖지 않는다는 점을 발견할 수 있다. 이는 각각 값 일반화 계급과 일반화된 데이터 표에서의 기록들의 분포에 따라 측도 값이 다르게 계산되기 때문이다.

IV. 일반화 및 특수화 기법

일반화와 특수화 기법을 이용하여 k-익명성을 달성하는 메커니즘을 설명하기 위해서는 준식별자에 대한 엄밀한 정의가 먼저 필요하다.

정의 7. 모집단 $U=\{p_i\}$와 개인정보를 포함하는 데이터 표 $T[A_1,\cdots,A_n]$가 주어지고, 함수

$$f_c: U\rightarrow T\ ,\quad f_g: T\rightarrow U'\ (U\subset U')$$

가 주어졌을 때, 준식별자(quasi-identifier) $QI_T\subset\{A_1,\cdots,A_n\}$는 다음 조건을 만족하는 속성들의 부분집합이다.

$$\exists p_i\in U\ ,\ f_g(f_c(p_i)[QI_T])=p_i\ .$$

4.1. 일반화(generalization) 기법

일반화 기법은 속성의 값을 보다 일반화된 값으로 대체하는 기법이다. 각 속성의 값을 보다 일반화된 도메인 상의 값으로 변환함으로써 k-익명성 조건을 만족시키고자 하는 것이다.

일반화 정도가 심할수록 k-익명성은 더욱 쉽게 만족되겠지만, 데이터 표가 너무 심하게 변형된다면 유용한 정보를 얻을 수 없을 것이다. 그러므로 k-익명성 조건을 만족하는 최소의 일반화 방법을 모색하는 것은 매우 의미 있는 작업이다.

정의 8. (k-최소일반화) 데이터 표 $T[A_1,\cdots,A_n]$와 준식별자 QI_T가 주어졌을 때, 일반화 과정 중의 데이터 표를 T_l과 T_m이 $T_l\leq T_m$을 만족한다고 하자. 이 때, T_m이 다음 두 가지 조건을 만족하면, T_m을 준식별자 QI_T 상에서 k-익명성에 대한 T_l의 최소일반화(minimal generalization)라 한다.

1. T_m은 QI_T 상에서 k-익명성 조건을 만족한다.
2. $T_l\leq T_z\leq T_m$을 만족하는 임의의 데이터 표 T_z가 k-익명성 조건을 만족한다면, $T_z=T_m$이다.

정의 8에서 $T_l\leq T_m$가 의미하는 바는 T_l을 일반화하면 T_m이 얻어진다는 것이다. 주의할 점은 여기에서 사용된 기호 "$\leq$"은 부분적으로만 순서를 결정할 수 있는 부분순서(partial ordering)라는 점이다. 서로 다른 k-최소일반화가 존재할 때 더 좋은 일반화된 데이터 표를 선택하기 위한 기준으로는 3절에서 논한 정확성 관련 측도들 중에서 정확도(precision)가 사용된다.

4.2. 특수화(specification) 기법

특수화 기법은 일반화 기법과는 반대로 속성의 값을 보다 더 적은 값으로 대체하는 기법이다. 데이터 표 $T(A_1,\ldots,A_m,Class)$가 주어졌다고 하자. 기록은 $\langle v_1,\ldots,v_m,ds\rangle$의 형태를 갖는다. 여기에서 ds는 $Class$에 대한 속성값을 의미한다. 특수화 기법에서는 클래스를 다른 속성과 구별하기 때문에 준식별자 QI_T를 *VID*(virtual identifier)로 표현한다. 특수화 기법에서 사용된 익명성 조건은 다음 정의와 같다.

정의 9. 데이터 표 T에서 준식별자 VID를 고려해 보자. T에서 $a(vid_i)$는 VID의 속성값이 vid_i인 기록들의 수를 의미하고, $A(VID)$는 $a(vid_i)$ 중에서 가장 작은 수를 의미한다. 이 때, $A(VID) \geq k$일 때, T는 준식별자 VID에 대하여 k-익명성 요구조건을 만족한다고 말한다.

일반화 기법이 상향식(bottom-up) 방법이라면, 특수화 기법은 속성의 최상위 계층에서 시작하여 정보손실이 작은 속성들부터 차례로 아래 계층으로 특수화시키는 하향식(top-down) 방법이다. 특수화시킬 속성을 결정할 때, 점수(score)라는 측도를 사용하고, k-익명성 조건이 위배되는 순간 특수화 과정은 멈추게 된다. 이 과정의 핵심은 어떤 속성을 특수화 할 것인지 결정하는 것인데 이때 사용하는 측도가 바로 점수(score)이다.

4.3. MinGen 및 TDS 알고리즘 비교 분석

일반화와 은폐 기법을 사용하여 k-익명성 조건을 만족시키고자 하는 알고리즘으로 알려져 있는 대표적인 것으로는 Datafly[9]와 μ-Argus[10]를 들 수 있다. Sweeney[5]는 Datafly 알고리즘은 너무 많은 왜곡을 요하고, μ-Argus 알고리즘은 적절한 프라이버시 보호가 이루어지지 않고 있음을 지적하면서 이들에 대한 개선책으로 MinGen 알고리즘을 제안하였다. 한편, Fung 등[6]은 상향식 알고리즘인 일반화 기법과 반대되는 개념의 하향식 알고리즘인 TDS를 제안하여 범주적 자료뿐만 아니라 연속형 수치 자료도 효율적으로 익명성 조건을 만족시킬 수 있다고 주장하였다. 우리는 여기에서 일반화와 특수화 기법을 대표하는 MinGen 알고리즘과 TDS 알고리즘의 효율성과 장단점을 비교분석한다. 이들 알고리즘은 각각 [표 4]와 [표 5]에 나타나 있다.

MinGen과 TDS 알고리즘의 특성과 장단점을 비교 분석해보자. 우선 가장 눈에 띄는 차이는 일반화 기법을 대표하는 MinGen 알고리즘이 상향식(bottom-up) 방법인 반면, 특수화 기법을 대표하는 TDS 알고리즘은 하향식(top-down) 방법이라는 점이다. 두 기법에서 달성하고자 하는 프라이버시 관련 요구조건은 k-익명성으로 동일하다. 하지만 MinGen에서는 k-익명성 조건을 만족하지 않는 데이터 표로부터 출발하여 k-익명성 조건이 만족되는 최소의 일반화 과정을 찾게 되지만, TDS 알고리즘에서는 k-익명성 조건이 충족되는 분류나무의 컷에서 출발하여 k-익명성 조건이 위배되기 직전까지 특수화 과정을 진행시키게 된다. 즉, k-익명성 조건은 두 알고리즘에서 동일하지만 이를 충족시키기 위한 절차적 과정은 서로 반대 방향으로 이루어진다는 차이점이 있다.

한편, 프라이버시 측도와 대척점에 존재하는 정확성(accuracy) 측도의 관점에서 MinGen과 TDS 알고리즘은 약간의 차이를 보인다. MinGen 알고리즘에서 최소 일반화를 결정하는 측도로 정확도를 의미하는 $Prec(\cdot)$이 사용되었다. $Prec(\cdot)$은 직관적으로 원래의 데이터 표에서 일반화 과정의 횟수가 각 속성의 분류나무 단계에서 어느 정도의 비율로 이루어졌는지를 측정하는 양이다. 그리고 TDS 알고리즘에서 사용된 $Score(\cdot)$는 특수화 과정이 진행되면서 얻어지는

[표 4] MinGen 알고리즘

입력: 데이터 표 PT, 준식별자 QI, 상수 k, DGH_{Ai}, $preferred()$명세서.
출력: MGT(상수 k의 선택에 따른 $PT[QI]$의 최소 왜곡 데이터 표)
가정 : $|PT| \geq k$

Method :
1. **if** $PT[QI]$가 k-익명성을 만족
   1.1. MGT ← $\{PT\}$
2. **else do**
   2.1. $allgen \leftarrow \{T_i$: T_i는 QI에 대한 PT의 일반화된 데이터 표이다.$\}$
   2.2. $protected \leftarrow \{T_i$: $T_i \in protected \wedge$ T_i는 k-익명성을 만족한다.$\}$
   2.3. MGT$\leftarrow \{T_i : T_i \in protected \wedge Prec(T_z) > Prec(T_i)$를 만족하는 $T_z \in protected$는 존재하지 않는다.$\}$
   2.4. MGT← **preferred**(MGT)
3. **return** MGT

[표 5] TDS 알고리즘

1. 데이터 표 T에 있는 모든 속성값을 최상위 값으로 초기화한다.
2. 최상위 속성값을 포함하기 위해 $\bigcup cut_i$를 초기화 한다.
3. **While** $x \in \bigcup cut_i$가 타당하면 실행한다.
   3.1 $\bigcup cut_i$로부터 $Best$ 특수화를 찾는다.
   3.2 T에 $Best$를 수행하고, $\bigcup cut_i$을 업데이트 한다.
   3.3 $x \in \bigcup cut_i$에 대한 타당성과 $Score(x)$를 업데이트 한다.
   end while
4. 일반화된 T와 $\bigcup cut_i$를 출력한다.

〔표 6〕 일반화와 특수화의 비교

알고리즘	MinGen	TDS
분류나무 진행방식	상향식 (bottom-up)	하향식 (top-down)
사용 기법	일반화 (generalization)	특수화 (specification)
프라이버시 측도	k-익명성	k-익명성
정확성 측도	$Prec(\cdot)$	$Score(\cdot)$
특징	- 이론적으로 최소 왜곡의 데이터를 찾을 수 있음 - 계산복잡도가 높아 비실용적	- 익명성과 정확성을 동시에 고려한 진행과정 - 상대적으로 높은 계산효율성

정보량을 엔트로피 개념을 사용한 정보 이론적 관점으로 측정하고, 이를 특수화 과정에서 손실되는 익명성 관련 수치로 나누어줌으로써 정의되는 개념이다. 그러므로 $Prec(\cdot)$가 단순히 데이터 표의 왜곡 정도만을 탐지하는 반면, $Score(\cdot)$는 데이터 왜곡 정도와 익명성 손실 정도를 동시에 고려한 측도이므로 좀 더 종합적인 의미를 담고 있는 측도라 할 수 있다. 위와 같은 비교 분석을 간단히 표로 정리해 놓은 것이 〔표 6〕에 나타나 있다.

V. ℓ-다양성 개념

5.1. k-익명성에 대한 공격

k-익명성을 만족하는 마이크로 데이터는 민감한 속성이 연결 공격에 의해서 유일하게 식별 될 수 없다는 것을 보장한다. k-익명성은 개념적으로 간단하기 때문에, 실용적인 프라이버시의 정의로써 광범위하게 논의 되고 있다. 그러나 k-익명성을 만족하는 데이터가 항상 프라이버시를 보장한다고 단언할 수는 없다. 다음 두 가지 공격을 살펴보면 k-익명성의 약점을 알 수 있다.

5.1.1. 동질성 공격

갑과 을이 사이가 좋지 않은 이웃이라고 하자. 어느 날 을은 건강이 나빠져서 구급차에 실려 갔다. 구급차를 보자마자, 갑은 을이 어떤 병에 걸렸는지 궁금해졌다. 갑은 〔표 7〕에 나타나 있는 것과 같이 병원에 의해 출판된 최근 환자들의 데이터를 발견했다. 이 데이터는 $k=4$인 k-익명성을 만족한다. 갑은 우편번호와 나이 정보로부터 을이 이 데이터 표의 기록 중에 포함되어 있다는 것을 알았다. 갑은 을의 이웃이기 때문에 을이 우편번호 13053에 살고 있는 30대라는 것을 알고 있는 것이다. 그러므로 갑은 을의 기록 번호가 9, 10, 11, 12 중의 하나라는 사실을 알았으며, 이 네 기록의 민감한 속성이 모두 같기 때문에 갑은 을의 병명이 암이라는 을의 프라이버시 정보를 알아낼 수 있었다. 이 예는 k-익명성이 만족된 데이터라 할지라도 블록 내에서 민감한 속성의 다양성이 부족할 경우에는 프라이버시 관련 정보가 노출될 수 있음을 보여준다.

〔표 7〕 k-익명성($k=4$)을 만족하는 마이크로 데이터

순번	준식별자			민감한 속성
	우편번호	나이	국적	상태
1	130**	<30	*	심장병
2	130**	<30	*	심장병
3	130**	<30	*	바이러스
4	130**	<30	*	바이러스
5	148**	≥40	*	암
6	148**	≥40	*	심장병
7	148**	≥40	*	바이러스
8	148**	≥40	*	바이러스
9	130**	3*	*	암
10	130**	3*	*	암
11	130**	3*	*	암
12	130**	3*	*	암

5.1.2. 배경지식 공격

갑은 을과 같은 병원에 있는 아사코라는 펜팔 친구를 만들었다. 그리고 〔표 7〕의 환자 기록을 가지고 있다. 갑은 아사코가 현재 우편번호 13068에 살고 있는 21살의 일본 여자임을 알고 있다. 이 정보를 바탕으로 갑은 아사코의 정보가 첫 번째 블록인 기록 1, 2, 3, 4 중에 포함되어 있다는 것을 알아냈다. 추가적인 정보 없이, 갑은 아사코가 바이러스에 걸렸는지 심장병에 걸렸는지 확신할 수는 없다. 그러나 일본인은 심장병에 걸릴 확률이 극단적으로 낮다는 사실을 알고 있다. 이로부터 갑은 아사코가 바이러스에 걸렸다는 것을 거의 확신할 수 있다. 이 예 역시 k-익명성을 만족한 데이터가 배경지식을 이용한 공격에 대해서 프라이버시를 보호하지 못할 수 있음을 보여준다.

5.2. ℓ-다양성

우리는 앞에서 k-익명성을 만족하는 데이터 표가 민감한 정보를 노출시키는 두 가지 공격에 대해서 살펴보았다. 실제로 이 공격들은 가능성이 충분하기 때문에 민감한 속성의 단순성과 배경지식을 이용한 공격에 대비할 수 있는 좀 더 강한 의미의 프라이버시 관련 개념이 필요하다. Machanavajjhala 등[3]은 이러한 필요성을 인식하여 k-익명성의 보완적 개념으로 ℓ-다양성(ℓ-diversity)이란 개념을 제안하였다.

정의 10. (ℓ-다양성) AQ-블록을 익명성 조건을 만족하면서 동일한 준식별자 값을 갖는 기록들의 집합이라고 할 때, AQ-블록 내에 있는 민감한 속성 값 집합의 원소 개수가 적어도 ℓ개 이상인 경우, 이 AQ-블록은 ℓ-다양성을 만족한다고 말한다. 그리고 데이터 표에 있는 모든 AQ-블록이 ℓ-다양성을 만족 한다면 그 데이터 표는 ℓ-다양성을 만족한다고 말한다.

결국 ℓ-다양성 요구사항은 AQ-블록 내에서 $1/\ell$ 보다 작은 확률로 민감한 속성의 노출 위험성을 낮추고자 하는 것이다. [표 8]은 [표 7]의 4-익명성을 만족하는 데이터 표로부터 3-다양성이 만족되도록 표현한 데이터 표를 나타낸 것이다. 4-익명성을 만족하는 데이터 표에 대한 동질성 및 배경지식, 두 가지 공격은 3-다양성을 만족하는 데이터 표에 의해서 막을 수 있다는 것을 알 수 있다. 예를 들어 갑은 을(우편번호:13053, 31세)이 암에 걸렸다는 사실을 3-다양성을 만족하는 데이터 표로부터 알아 낼 수는 없다. 그리고 아사코가 심장병에 걸리지 않았다고 확신하더라도, 바이러스에 걸렸는지 암에 걸렸는지 확신 할 수는 없게 될 것이다.

[표 8] ℓ-다양성($\ell=3$)을 만족하는 마이크로 데이터

순번	준식별자			민감한 속성
	우편번호	나이	국적	상태
1	1305*	≤40	*	심장병
2	1305*	≤40	*	바이러스
3	1305*	≤40	*	암
4	1305*	≤40	*	암
5	1485*	>40	*	암
6	1485*	>40	*	심장병
7	1485*	>40	*	바이러스
8	1485*	>40	*	바이러스
9	1306*	≤40	*	심장병
10	1306*	≤40	*	바이러스
11	1306*	≤40	*	암
12	1306*	≤40	*	암

ℓ-다양성의 원저자들은 위에서 설명한 개념을 명시적 ℓ-다양성(distinct ℓ-diversity)이라 부르고, 이 명시적 ℓ-다양성 외에 엔트로피(entropy) ℓ-다양성 개념과 재귀적(recursive) (c,l)-다양성 개념을 제시하였으나 본 논문에서는 명시적 ℓ-다양성만을 다루기로 한다.

VI. 블록 합병 방법을 이용한 ℓ-다양성 확보 알고리즘

우리는 본 절에서 블록 합병 방법을 이용한 ℓ-다양성 확보 알고리즘을 제안한다. $AQ[i]$는 k-익명성을 만족하는 블록을 의미하고, $DQ[i]$는 ℓ-다양성을 만족하는 블록을 의미한다고 하자. 그리고 $ADQ[i]$는 k-익명성과 ℓ-다양성을 모두 만족하는 블록을 의미하며, $\widehat{DQ}$는 ℓ-다양성을 만족하지 않는 블록을 뜻한다고 하자. 기존의 효율적인 알고리즘을 이용하여 k-익명성 조건은 만족한다고 가정한 상태에서 우리는 ℓ-다양성 조건을 달성하기 위한 구체적 방법을 제안하고자 한다. [표 9]는 블록 합병 방법에 의하여 ℓ-다양성을 만족시키는 알고리즘을 나타내고 있다. 알고리즘 내에서 DGH_{A_i}는 i번째 속성 A_i의 도메인 일반화 단계를 의미하는 것으로 k-익명성을 달성하기 위한 알고리즘 내에서 정확성 관련 측도 값을 계산하기 위해 필요한 것이며, k-익명성과 ℓ-다양성을 모두 만족하는 출력 데이터 표를 ADT라 명명한다. 또한, $|AQ[i]|_s$는 k-익명성과 ℓ-다양성을 모두 만족하는 블록의 민감한 속성 개수를 의미하고, $ADQ(i\|j)$는 i번째 블록과 j번째 블록을 합병한 결과 k-익명성과 ℓ-다양성이 모두 만족되는 블록을 의미한다.

제안하는 알고리즘의 첫 번째 단계는 기존의 효율적인 알고리즘을 이용하여 k-익명성이 만족되는 데이터 표를 생성하는 과정이다. 이 단계의 결과 값은 $AQ[1],\ldots,AQ[n_k]$ 블록들로 재편성된 데이터 표가 된다. 두 번째 단계는 ℓ-다양성을 만족하는 데이터 표가 생성되는 과정을 세부적으로 나타낸 것이다. 첫 번째 단계의 결과 값이 k-익명성을 만족하는 동시에 ℓ-다양성도 만족한다면 그대로 결과 값으로 출력해 낸다. AQ-블록 중에 어느 하나라도 ℓ-다양성을 만족하지 않는다면 다음 세부 절차를 실행한다.

ℓ-다양성을 만족하지 않는 블록 $\widehat{DQ}$이 한 개만 존재한다면 ℓ-다양성을 만족하는 블록 $DQ[i]$들과 차례로 합병시킨 다음, ℓ-다양성을 만족하는 데이터 표들의 정확도($Prec.$) 값들을 계산한 후 그 값이 최대인 데이터 표를 출력 한다. ℓ-다양성을 만족하지 않는 블록 $\widehat{DQ}$들이 두 개 이상이라면 $\widehat{DQ}$들을 각각 합병시킨 다음, ℓ-다양성을 만족하는 데이터 표의 정확도($Prec$) 값들을 계산하여 이 들 중 최대값을 갖는 데이터 표(*Max Prec Table*)를 출력하게 된다.

〔표 9〕에 나타나 있는 알고리즘은 블록 합병 방법을 이용하여 ℓ-다양성을 확보하고자 하는 것으로, 그간 발표된 바가 없었던 구체적인 ℓ-다양성 달성 방법을 처음으로 제안했다는 데에 큰 의미를 둘 수 있다. ℓ-다양성을 만족하지 않은 블록의 개수가 많은 경우에 본 알고리즘의 효율성은 상대적으로 떨어질 수 있으나, k-익명성을 위한 일반화 또는 특수화 과정을 반복함으로써 ℓ-다양성도 확보하고자 했던 원저자들의 생각보다는 효율적인 것이다. 전체 데이터 중에서 일부분인 ℓ-다양성을 만족하지 않은 블록들에 먼저 집중함으로써 k-익명성만을 위해 설계된 알고리즘으로는 우연에 의존했던 결과 데이터의 ℓ-다양성을 처음부터 고려하여 알고리즘의 세부 과정이 진행되기 때문이다.

〔표 9〕 ℓ-다양성을 위한 블록 합병 알고리즘

입력 : 데이터 표 PT, 준식별자 QI,
상수 k, ℓ, DGH_{A_i}

출력 : ADT
(k-익명성과 ℓ-다양성을 만족하는 데이터 표)

Method :
1. k-익명성을 만족하는 데이터 표 생성
($AQ[1], ..., AQ[n_k]$ 블록들로 구성된 데이터 표)
2. ℓ-다양성을 만족하는 데이터 표 생성.
2.1. $\forall 1 \le i \le n_k$, if $|AQ[i]|_s \ge \ell$
2.1.1. $ADQ[i] \leftarrow AQ[i]$
2.1.2. $ADT \leftarrow ADQ$
2.2. **else do**
2.2.1. if $\#(\widehat{DQ}) = 1$, $AQ[j] = \widehat{DQ}$
then, for $1 \le i \le n_k$, $i \ne j$,
다음을 계산
$Prec(T[ADQ(1), ..., ADQ(i \cup j), ..., ADQ(n_k)])$
$ADT \leftarrow$ *Max Prec Table*
2.2.2. else $\#(\widehat{DQ}) = m \ge 2$
$\widehat{DQ}$ 블록들을 모든 블록들이
ADQ 블록으로 바뀔 때까지
블록합병을 반복한다.
2.2.3. 2.2.2를 만족하는 각 블록합병 방법에
대하여 대응하는 데이터 표의 $Prec$을
계산한다
2.2.4. $ADT \leftarrow$ *Max Prec Table*
3. **return** ADT

VII. 결 론

우리는 프라이버시가 보호된 상태에서 유용한 정보를 공유하기 위한 실용적인 방법 중의 하나로 주목받고 있는 개념인 k-익명성 개념과 이를 바탕으로 전개되어 온 ℓ-다양성 및 t-밀접성 개념에 대하여 개괄적으로 살펴보았다. k-익명성을 달성하기 위한 알고리즘들 중에서 대표적인 일반화 기법과 특수화 기법에 대한 심층적인 분석을 실시하여 이들 기법에 사용된 프라이버시 및 정확도 관련 측도들을 조사 분석하고, 상호 관계를 탐구하였으며, 대표적 알고리즘인 MinGen과 TDS에 대한 특성과 장단점을 비교분석하였다. 본 논문에서는 기존 측도들 중 높이, 정확도, 비용, 점수 등을 고려 대상으로 하였다. 이들을 통합적으로 비교 분석하기 위하여 가능한 균형 있고 공정한 시각으로 합리적인 비교 분석을 실시하고자 노력하였다. 원본 데이터와 변형된 데이터가 주어졌을 때, 각 측도들 사이의 마이크로 데이터의 정확성에 대한 일관성과 개별성을 조사하고, 그 측도들의 특성에 따른 의미와 효율성을 비교분석하였다. 향후 좀 더 면밀한 연구를 통하여 각 측도들의 장점을 통합하고 단점을 보완할 수 있는 복합적인 측도를 개발하는 것이 남아있는 과제라고 생각한다.

ℓ-다양성 개념의 원저자들은 k-익명성만이 달성된 데이터의 프라이버시 문제점을 지적하고 이를 해결하기 위한 대책으로 ℓ-다양성 개념을 제시하였을 뿐, ℓ-다양성을 만족시키기 위한 구체적 방법론을 제안하지는 못했었다. 본 논문에서는 k-익명성이 만족되는 데이터로부터 ℓ-다양성 조건까지 만족되도록 하는 실질적인 알고리즘을 제안하였다. 블록 합병 방법을 이용함으로써 k-익명성으로부터 ℓ-다양성 개념이 자연스럽게 전환될 수 있는 세부적 절차를 수립한 것이다. 이 연구 결과는 아직까지 발표된 바 없는 ℓ-다양성 달성 방법을 처음으로 구체적인 알고리즘의 형태로 제안했다는 데에 큰 의미를 둘 수 있다.

제안된 알고리즘은 k-익명성을 만족하면서 동시에 ℓ-다양성을 만족하며, 정확도 측도를 이용하여 가장 정보 손실이 적은 데이터 표를 생성해 낸다. 하지만 알

고리즘의 실용성 관점에서는 ℓ-다양성을 위한 블록 합병 알고리즘이 모든 가능한 경우의 합병 데이터 표를 생성한 후, 정확도가 최대인 데이터 표를 생성해내기 때문에 계산 효율성이 매우 높다고 할 수는 없다. 향후 좀 더 면밀한 분석과 연구를 통하여 효율성 높은 알고리즘이 되도록 개선해야 하는 과제가 남아 있다.

참고문헌

[1] G. Duncan, "State of the Art: An Overview of Policy and Practice on Release of Microdata," www.heinz.cmu.edu/research/32full.pdf, 2002.

[2] L. Sweeney "k−anonymity : A model for protection privacy," International Journal of Uncertainty, Fuzziness and Knowledge-based systems, Vol. 10, No. 5, pp. 557-570, October 2002.

[3] A. Machanavajjhala, J. Gehrke, and D. Kifer, "ℓ-diversity: Privacy beyond k-anonymity," Proceedings of the International Conference on Data Engineering (ICDE2006), pp. 24-35, April 2006.

[4] T. Li, N. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," Proceedings of the International Conference on Data Engineering(ICDE2006), pp. 106-115, April 2006.

[5] L. Sweeny, "Achieving k-anonymity privacy protection using Generalization and suppression," International Journal of Uncertainty, Fuzziness and Knowledge-based systems, pp.571-588, October 2002.

[6] B. Fung, K. Wang, and P. Yu, "Top-Down Specialization for information and privacy preservation," Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, April 2005.

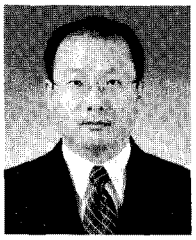
[7] T. Truta and V. Bindu, "Privacy Protection: P-Sensitive K-Anonymity Property," Proceedings of the Workshop on Privacy Data Management, In Conjunction with 22th IEEE International Conference of Data Engineering (ICDE), Atlanta, Georgia, April 2006.

[8] R. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," 21st International Conference on Data Engineering (ICDE'05), pp. 217-228, April 2005.

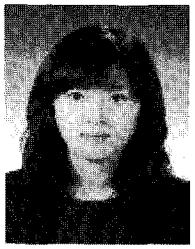
[9] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the Datafly system," Proceedings of the American Medical Informatics Association Annu. Fall Symp., http://dataprivacylab.org/datafly/paper4.pdf, February 1997.

[10] A. Hundepool and L. Willenborg, "μ- and τ-argus: software for statistical disclosure control," Third International Seminar on Statistical Confidentiality, pp. 142-149, November 1996.

〈著者紹介〉

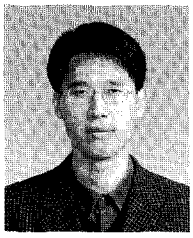


강 주 성 (Ju-Sung Kang) 종신회원
1989년: 고려대학교 수학과(학사)
1991년: 고려대학교 일반대학원 수학과 (이학석사)
1996년: 고려대학교 일반대학원 수학과 (이학박사)
1996년~1997년: 과학재단 박사후연구원
1997년~2004년: 한국전자통신연구원 선임연구원, 팀장
2001년~2002년: 벨기에 루벤대학 COSIC 방문연구원
2004년~현재: 국민대학교 수학과 부교수
〈관심분야〉 암호 알고리즘, 정보보호 프로토콜



강 진 영 (Jin-young Kang)
2006년: 순천향대학교 정보물리학과 졸업
2008년: 국민대학교 일반대학원 수학과 (이학석사)
2008년~2009년: (주)교학사
2009~현재: (주)해법에듀
〈관심분야〉 정보보호, 암호론

이 옥 연 (Okyeon Yi) 정회원
1988년: 고려대학교 수학과 졸업
1990년: 고려대학교 일반대학원 수학과 (이학석사)
1996년: University of Kentucky 수학과 (이학박사)
1999년~2001년: 한국전자통신연구원 선임연구원, 팀장
2001년~현재: 국민대학교 수학과 부교수
〈관심분야〉 정보보호, 이동통신, 암호론 등



홍 도 원 (Dowon Hong) 정회원
1994년: 고려대학교 수학과(학사)
1996년: 고려대학교 일반대학원 수학과 (이학석사)
2000년: 고려대학교 일반대학원 수학과 (이학박사)
2000년~현재: 한국전자통신연구원 책임연구원, 팀장
〈관심분야〉 암호 이론, 정보보호 이론, 이동통신 정보보호