k-anonymity와 ℓ-diversity를 이용한 동적 데이터 보호 기법 설계

정은희* · 이병관**

A Design of DDPT(Dynamic Data Protection Technique) using k-anonymity and ℓ-diversity

Eun-Hee Jeong*, Byung-Kwan Lee**

요 약

본 논문에서는 동적 데이터베이스 환경에서 발생할 수 있는 개인 정보 노출 문제를 해결할 수 있는 동적 데이터 보호 기법(Dynamic Data Protection Technique)을 제안하였다. 본 논문에서 제안한 DDPT은 다중 속성 일반화 알고리즘을 이용해 MAG(Multi-Attribute Generalization) 규칙을 생성하고, 그 MAG 규칙에 따라 k-anonymity를 만족하는 EC(Equivalence Class)를 생성한다. 그리고 데이터 변경 시 MAG 규칙에 따라 EC를 재구성 하도록 하여, EC의 변경으로 인한 식별 노출을 방지할 수 있다. 또한, ℓ-diversity를 만족하는 EC의 정보손실 정도를 측정하고, 임계치 이하의 EC를 선정해서 데이터의 정확성을 유지함으로써 개인 정보 보호를 향상시켰다.

ABSTRACT

This paper proposes DDPT(Dynamic Data Protection Technique) which solves the problem of private information exposure occurring in a dynamic database environment. The DDPT in this paper generates the MAG(Multi-Attribute Generalization) rules using multi-attributes generalization algorithm, and the EC(equivalence class) satisfying the k-anonymity according to the MAG rules. Whenever data is changed, it reconstructs the EC according to the MAC rules, and protects the identification exposure which is caused by the EC change. Also, it measures the information loss rates of the EC which satisfies the ℓ-diversity. It keeps data accuracy by selecting the EC which is less than critical value and enhances private information protection.

Key Words : MAG rules, k-anonymity, ℓ-diversity, Private Information, Protection

Ⅰ. 서 론

많은 기업, 병원, 공공기관에서 수집하는 개인정보의 양은 지속적으로 증가하고, 이들 정보는 마케팅 조사, 의학연구, 인구통계 등 다양한 목적으로 다양한 연구 분야에서 사용된다[1]. 그리고 이 정

* 강원대학교 지역경제학과 교수 (jeongeh@kangwon.ac.kr)
** 교신저자 : 관동대학교 컴퓨터학과 교수 (bklee@kwandong.ac.kr)
접수일자 : 2011년 07월 21일, 수정일자 : 2011년 08월 06일, 심사완료일자 : 2011년 08월 22일

보의 활용도를 높이기 위해 기관이나 조직은 서로의 정보를 공유하거나 공공의 목적으로 배포하기도 한다. 이렇게 개인 정보가 배포되었을 때는 더 이상 그것을 수립했던 조직의 통제 하에 있지 않기 때문에 개인의 민감한 정보가 노출될 수 있다[2].

예를 들어, 미국 인구의 약 87%는 성별, 생년월일, 5자리 우편번호(ZIP) 코드의 단 세 가지 정보로 개인이 유일하게 판별될 수 있으며, 개인이 유일하게 판별될 경우 다른 정보와 결합을 통해 개인의 민감한 정보가 드러날 수 있다[3]. 즉, 정보들은 외부로 배포되기 전에 주민등록번호 또는 이름과 같은 개인 신원 정보들은 미리 암호화되거나 삭제되고 나머지 정보들만 배포되지만, 다른 테이블과의 조인을 통해 민감 정보의 식별이 가능하다. 이러한 형태의 개인 정보 노출 공격을 결합공격(linkage attack)이라 한다.

다른 데이터와의 결합을 통한 결합 공격으로부터 정보 노출을 방지하기 위해 k-anonymity 모델[1,3,4]이나 L-diversity 모델[5]과 같은 데이터 익명화 모델들이 제안되었다. 하지만 k-anonymity와 L-diversity는 레코드의 추가나 삭제가 없는 정적인 환경을 가정하기 때문에 레코드가 추가되거나 삭제될 경우 민감한 속성의 값이 유출되는 문제가 발생한다.

본 논문에서는 개인정보 노출 문제를 해결하기 위해, k-anonymity와 L-diversity를 이용해 각 개인에 대한 정보를 가지고 있는 데이터에 대해 결합공격과 배경지식 공격에 의해 개인정보 노출을 방지하는 동적 데이터 보호 기법(Dynamic Data Protection Technique)을 제안한다.

또한, 본 논문에서 제안된 DDPT를 활용해 데이터베이스에 새로운 데이터가 삽입, 수정, 그리고 삭제될 때에도 민감한 속성 값 유출의 가능성을 줄여 개인 정보의 보안을 강화시키고자 한다.

논문의 구성은 다음과 같다. 2장에서 관련연구를 설명하고, 본 논문에서 제안하는 DDPT를 3장에서 설명하고, 4장에서 실험 결과 및 평가에 대해 설명한다. 그리고 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

1. k-anonymity

k-anonymity 모델과 일반화를 이용한 개인정보 보호 기법은 Sweeny와 Samarati에 의해 소개되었다[1,3,4].

k-anonymity의 목적은 Quasi-Identifier(준 식별자)를 이용하여 인스턴스간의 대응관계와 테이블 안의 레코드 사이에 높은 확률을 가지는 관련성을 제거한 테이블을 생성하는 것이다. 이때, 준 식별자에 의해 k-anonymous를 하기 위해서는 테이블 안의 모든 레코드들은 준 식별자에 대해 구별이 불가능한 레코드가 적어도 k-1개 존재해야 한다.

그림 1의 (b)는 (a)의 원본 테이블에서 나이, 성별, 지역을 준 식별자로 가정해 3-anonymous 테이블이다.

나이	성별	지역	병명
23	F	서울	당뇨
25	M	대구	당뇨
27	F	대구	당뇨
32	M	대전	감기
36	M	서울	당뇨
38	M	강릉	암
55	F	부산	에이즈
50	F	대구	백혈병

(a) 원본 테이블

나이	성별	지역	병명
20-30	P	*	당뇨
20-30	P	*	당뇨
20-30	P	*	당뇨
30-40	M	*	감기
30-40	M	*	당뇨
30-40	M	*	암
50-60	F	*	에이즈
50-60	F	*	백혈병

(b) 3-anonymous 테이블

그림 1. k-anonymity의 예(k=3)
Fig. 1 The Example of k-anonymity(k=3)

k-anonymity 모델은 익명성을 제공하지만 그림 1의 (b)처럼 (20-40, P, *)인 경우 병명이 "당뇨"로 동일하므로, 만약 A가 25세의 대구에 사는 누군가를 알고 있다면, 그 사람의 병명이 당뇨인 것을 추측할 수 있다. 즉, k-anonymity 모델은 익명성은 제공하지만 동등 클래스 내의 민감한 속성 값의 분포는 고려하지 않는 문제가 있다.

2. ℓ-diversity

k-anonymity 모델은 개인정보를 완벽하게 보호하지 못함을 Machanavajjhala 등이 밝히고, 이 문제점을 해결하는 방안으로 L-diversity 모델을 제시하였다[5]. L-diversity 모델은 서로 구분 가지 않은 레코드들 사이에서 민감한 정보는 최소한 L

개 이상이어야 개인정보가 잘 보호된다는 조건을 가진다.

따라서 L-diversity 모델은 결합공격과 공격자가 배경지식을 이용하여 특정 레코드가 특정 민감한 값을 가질 수 없음을 추측하여 해당 특정 레코드 및 다른 레코드의 민감한 값을 추론하는 공격인 배경지식 공격에 대한 방어책을 제시하였다[6].

하지만, L-diversity 모델은 동등 클래스 내의 민감한 속성 분포를 고려하기는 하지만 민감한 속성 값이 분포가 한쪽으로 치우치거나 의미가 유사한 경우에는 충분히 개인정보를 보호하지 못한다.

그림 2의 (a) 테이블을 2-diverse를 가지도록 일반화한 그림 2의 (b) 테이블에서 A씨가 "25세의 대구에 사는 여성"이라는 사실을 안다고 하더라도 이 사람이 어떤 병에 걸렸는지는 유추하지 못한다.

나이	성별	지역	병명
23	F	서울	당뇨
25	M	대구	당뇨
27	F	대구	당뇨
32	M	대전	감기
36	M	서울	당뇨
38	M	강릉	암
50	F	부산	에이즈
55	F	대구	백혈병

(a) 원본 테이블

나이	성별	지역	병명
20-35	P	*	당뇨
20-35	P	*	당뇨
20-35	P	*	당뇨
20-35	P	*	감기
30-50	M	*	당뇨
30-50	M	*	암
40-60	F	*	에이즈
40-60	F	*	백혈병

(b) 2-diverse 테이블

그림 2. ℓ-diversity의 예
Fig. 2 The example of ℓ-diversity

이와 같은 기법은 정적인 테이블에 관한 개인정보 보호 기업으로는 적합하지만, 데이터의 수정 및 삭제 등으로 정보가 동적으로 변화할 때, 동일하게 적용하기에는 문제가 있다.

Ⅲ. 동적 데이터 보호 기법 설계

동적 데이터베이스 환경에서의 기존 익명화 기법 대부분은 L-diversity 기반이므로 데이터 삽입 또는 삭제로 인한 식별노출이 가능하다는 문제점과 익명화로 인한 데이터의 정확성이 감소한다는 문제점을 가지고 있다. 본 논문에서는 데이터 변경시마다 EC를 재조정하여 식별 노출을 방지하고, 익명화된 데이터의 정보손실을 측정하여 정보손실이 적은 EC를 선택함으로써 데이터 정확성에 따른 문제점을 해결하고자 한다.

1. 정의

본 논문에서 사용되는 용어들을 다음과 같이 정의한다.

[정의 1] QI(Quasi-Identifier) 속성

해당 정보만으로는 개인의 식별 정보를 포함하고 있지는 않지만, 다른 정보들과의 조인을 통해 개인의 식별 정보를 알 수 있는 수단으로 사용할 수 있는 속성들을 말한다. 예를 들어, 그림 1, 2의 나이, 성별, 지역을 QI라 할 수 있다.

[정의 2] S(Sensitive) 속성

개인의 프라이버시와 관련된 민감한 정보를 포함할 수 있으며, 프라이버시 보호의 대상이 되는 속성을 말한다. 예를 들어, 그림 1, 2의 병명이 S 속성에 해당된다.

[정의 3] EC(Equivalent Class) 속성

k-anonymity 등의 익명화 기법이 적용되어 QI 속성 값만으로는 구별할 수 없는 레코드들의 그룹을 EC 또는 QI 그룹이라 말한다. k-anonymity 기법을 사용할 경우 k개 이상의 EC를 가지게 된다. 예를 들어, 그림 2는 k=3으로, 3-anonymity 2-diversity를 만족하는 테이블이다.

[정의 4] 일반화(Generalization) 기법

레코드의 속성 값을 일반화된 값으로 대체하는 기법으로 수(연속성) 도메인을 갖는 연속성은 좀 더 큰 범위를 갖는 값으로 변경하고, 비연속성은 도메인에 맞는 계층도를 참고하여 일반적인 값으로 변경한다. 예를 들어, 그림 2처럼 수 도메인을 갖는 나이는 [20-35]으로 큰 범위로 대치하고, M, F인 비연속성 도메인을 갖는 성별은 P로 대치하여 성별을 일반화시켜 데이터 속성 값을 익명화시킨다. 이때, 일반화의 레벨이 커질수록 데이터의 정확성은 감소할 수 있다.

[정의 5] 정보 손실(Information Loss)

일반화 계층 구조를 이용해 일반화할 때, 정보의 손실정도를 측정하는 수식은 다음과 같다[7].

$$IL(e) = |e| \times \sum_{j=1,\ldots,m} \frac{|G_j|}{|D_j|} \quad (1)$$

$|e|$: EC내의 레코드의 수

$|D_j|$: 속성 j값의 크기

$|G_j|$: 일반화에 참여한 속성 j의 일반화 정도

2. 익명화 기법 설계

익명화 기법은 기본적으로 개인 정보 노출을 방지하고, 통계분석을 위해 정확한 통계 정보를 제공해야 한다. 그리하여 개인 정보 노출 방지를 위해 개인 정보를 익명화를 시키는데 일반화 기법을 사용한다. 그런데, 이 일반화 기법은 데이터 정확성을 감소시킬 수 있다.

본 논문에서는 데이터의 정확성을 유지하면서 개인 정보 노출을 방지할 수 있는 DDPT인 익명화 기법을 제안하며, 전체적인 흐름은 그림 3과 같으며, 기본적인 절차는 다음과 같다.

[Step 1] 데이터의 QI 속성과 S 속성을 선정한다.

[Step 2] 선정된 QI 속성을 이용해 단일 속성과 다중 속성으로 구성해 일반화 작업을 수행하여 다중 속성 일반화 규칙인 MAG(Multi-Attribute Generalization) 규칙을 생성한다. 이때, 일반화 작업 시 QI 속성 값에 따라 데이터 분류 트리가 생성된다.

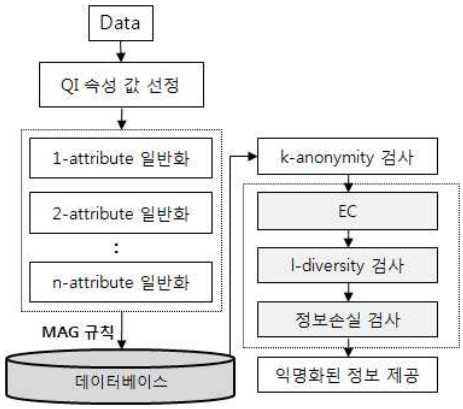


그림 3. 익명화 기법 흐름도

Fig. 3 Flowchart Anonymous technique

[Step 3] step 2의 MAG 규칙들을 이용해 k-anonymity가 만족하는 EC를 생성한다.

[Step 4] EC가 L-diversity를 만족하는지 검사한다. L-diversity가 만족하지 않으면, Step 3으로 이동해 MAG 규칙에 따라 k-anonymity를 만족하도록 새로운 EC를 생성하도록 한다.

[Step 5] k-anonymity와 L-diversity를 만족하는 EC의 정보손실 정도를 계산하여, 관리자가 지정한 임계치와 비교한다.

[Step 6] 임계치보다 작으면 익명화 기법이 적용되어 데이터의 정확성을 유지하면서 개인 정보 노출을 방지할 수 있는 데이터가 생성된 것이다. 그렇지 않으면, Step 3으로 이동해 MAG 규칙에 따라 k-anonymity를 만족하는 새로운 EC를 생성한다.

본 논문에서는 k-anonymity를 만족하는 다중 속성 일반화 알고리즘 설계에 Incognito 알고리즘[8]을 사용하여 MAG 규칙들인 rule-set을 생성하고, 이 rule-set의 MAG 규칙들을 이용해 k-anonymity를 만족하는 테이블인 T를 생성하였다.

그림 4는 MAG(Multi-Attribute Generalization) 규칙을 생성하는 알고리즘인 다중 속성 일반화 알고리즘을 설명한 것이다.

```
MAG(T, Q){
 T : k-anonymity table
 Q : quasi-identifier 속성 값 집합
 for(i=0 ; i<=n ; i++){
  insert_queue(Qi);
  while(NotEmpty(queue)){
   if(IsNotMark(node)){
    if(node(root))
      rule_set = cal_rule_set(Qi);
    else
      rule_set = cal_rule_set(Parent);
    check_k_anonymity(rule_set);
   if(k_anonymity(T)){
    check_direct(node);
   }else{
    delete_queue(node);
   }
  }
 Graph(rule_set);
 }
return(rule_set);
}
```

그림 4. 다중 속성 일반화 알고리즘
Fig. 4 Multi-Attribute Generalization algorithm

그림 5는 k-anonymity 알고리즘을 설명한 것으로 다중 속성 일반화 알고리즘이 생성한 rule-set의 MAG 규칙을 이용해 T의 레코드들을 일반화시키고, k-anonymity를 만족시키는 T의 EC들을 생성한다.

```
k-anonymity(T, rule_set){
 T : 데이터 table
 rule_set : MAG rules
 check = TRUE;
 // 테이블의 각 레코드를 rule_set에 의해 일반화
 while(check){
   for(i=0; i<=n ; i++){
    generalization(T[ri], rule_set)
   }
   if (anonymity(T, k)){
     check = FALSE;
     return(T);
   }
 }
}
```

그림 5. k-anonymity 알고리즘
Fig. 5 k-anonymity algorithm

```
ℓ-diversity(T){
  D : 도메인 크기
  G : 일반화 정도
  EC : equivalence class
  check = TRUE;
 // ℓ-diverse 검사
  if(diversity(EC) < ℓ)
    // EC의 다양성이 ℓ보다 작으면,
   call(k-anonymity(T, Q));
 // 정보 손실 점검
  IL_e = cal_IL(n, D, G);
  if(IL_e > critical_IL) // 정보손실이 크면
    call(k-anonymity(T, Q));
}
```

그림 6. ℓ-diversity 알고리즘
Fig. 6 ℓ-diversity algorithm

그림 6은 L-diversity 알고리즘은 설명한 것이다. k-anonymity 알고리즘에 의해 k-anonymity를 만족하는 테이블 T가 반환 된다. 이 테이블을 이용해 EC들을 생성하고, 이 EC들이 L-diversity를 만족하는지 검사하고, 만약에 L-diversity를 만족하지 않으면, k-anonymity 알고리즘을 재실행하여 MAC 규칙에 맞는 T의 EC를 재조정하여 L-diversity를 만족하는 T를 생성한다.

그리고 데이터의 삽입과 삭제가 발생하면 EC들의 k-anonymity나 L-diversity가 위배될 수 있다. 이때, k-anonymity나L-diversity를 만족시키기 위해, 본 논문에서는 MAG 규칙을 이용해 k-anonymity를 만족하도록 EC를 수정한 후, L-diversity를 만족하도록 수정한다.

Ⅳ. 실험 및 평가

본 논문에서 제안한 DDPT의 실험은 UCI repository of machine learning databases의 Adults 데이터를 이용하였고[9], 실험환경은 운영체제 Windows XP, Intel Core Duo CPU 2.20 GHz, RAM 2.0GB의 노트북으로 하였으며, 데이터베이스 MySQL를 이용하였으며, 알고리즘 구현은 웹 프로그래밍 언어인 PHP를 이용하였다.

실험 데이터의 Adult 데이터베이스에는 32,561개의 레코드와 15개의 속성이 있는데, 본 논문에서

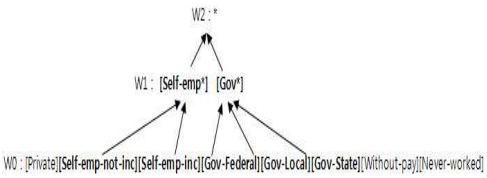
는 이 레코드 중에서 불확실한 레코드를 삭제하여 30,163개의 레코드의 8개의 속성만을 추출해서 7개는 QI 속성, 나머지 한 개인 Education은 S속성으로 실험하였다. 그리고 데이터의 속성 값은 실험을 위해 일부분 수정하였다.

표 1은 Adult 데이터베이스에 추출되어 실험에 사용된 QI 속성과 S 속성을 설명하고, QI 속성들의 일반화 방법을 설명한 것이다.

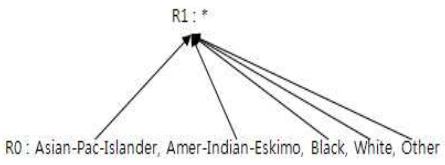
표 1. Adult 데이터베이스.
Table 1. Adult Database

순번	속성	값	일반화
1	age	74	범위[15-25, 26-35, ..]
2	work class	7	데이터 분류 트리
3	marital status	7	데이터 분류 트리
4	occupation	14	데이터 분류 트리
5	race	5	은폐(*)
6	sex	2	은폐(Person)
7	native country	41	데이터 분류 트리
8	education	16	민감한 속성값

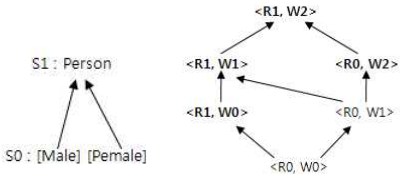
그림 7은 실험에서 사용된 다중 속성 일반화 알고리즘의 결과인 MAG 규칙 중에서 1-attribute 규칙과 2-attribute 규칙 중의 일부를 설명한 것이다. 그림 7의 (d)는 race와 work class 두 개의 속성에 대한 MAG로 이 중에서 <R1, W0>, <R1, W1>, <R0, W2>, <R1, W2>가 2-anonymity를 만족한다.



(a) work class



(b) race



(c) sex (d) race와 work class
그림 7. 일반화의 예
Fig. 7 Generalization samples

실험은 새로운 데이터가 삽입되거나 삭제될 때, 변경된 데이터 셋의 전체를 새롭게 익명화시키는 고정 데이터 익명화 기법과 본 논문에서 제안한 DDPT를 비교하였다.

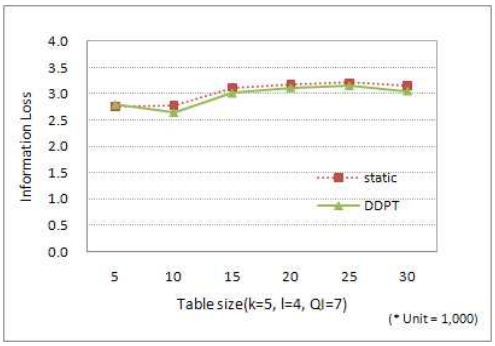


그림 8 데이터 정확성 비교
Fig. 8 The Comparison of data accuracy

그림 8은 고정 데이터와 동적 데이터의 정보 손실정도를 테이블 크기별로 비교한 것이다. 본 논문에서는 L-diversity 알고리즘에서 정보손실을 점검하여 임계치보다 작은 EC를 선정하므로 제안한 DDPT의 정보손실이 기존의 고정 데이터 익명화 기법보다 적은 것을 알 수 있다.

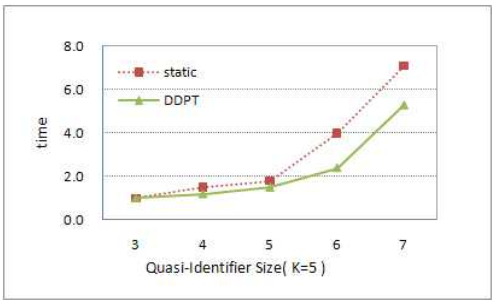


그림 9. QI 크기에 따른 익명화 처리 시간 비교(k=5)
Fig. 9. The comparison of anonymous process time according to QI size

그림 9는 고정 데이터와 동적 데이터의 익명화 처리 시간을 QI 크기에 따라 비교한 것이다. QI 크기가 커질수록 익명화 처리 시간은 증가하고, 고정 데이터 익명화 기법과 본 논문에서 제안한 DDPT의 처리시간이 고정 데이터 익명화기법의 처리시간보다 빠른 것을 알 수 있다.

Ⅴ. 결 론

본 논문에서는 삽입 및 삭제가 일어나는 동적 데이터베이스 환경에서 발생할 수 있는 개인의 프라이버스 침해 문제를 해결할 수 있는 동적 데이터 보호 기법인 DDPT를 제안하였으며, 제안하는 DDPT의 특징은 다음과 같다.

첫째, 본 논문에서 제안한 동적 데이터 보호 기법은 다중 속성 일반화 알고리즘을 이용해 MAG 규칙을 생성하고, 그 MAG 규칙에 따라 k-anonymity를 만족하는 EC를 생성함으로써, EC생성 시간을 단축할 수 있다.

둘째, 데이터 수정, 삽입 그리고 삭제 시 MAG 규칙에 따라 데이터를 일반화하고, k-anonymity를 만족하도록 EC를 재구성함으로써, EC의 변경으로 인한 식별 노출을 방지하여 개인 정보 보호를 강화시키는 효과를 얻을 수 있다.

셋째, L-diversity를 만족하는 EC의 정보손실 정도를 측정하여 관리자가 지정한 임계치 이하의 EC를 선정하므로 데이터의 정확성은 유지하면서 개인 정보를 강화시켰다고 볼 수 있다.

참 고 문 헌

[1] P.Samarati, and L. Sweeney, "Generalizing data to provide anonymity when disclosing information(Abstract)", In Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the principles of Database Systems(PODS'01), pp.188, Seattle, WA, USA, 2001.

[2] 변창우, 김재환, 이향진, 강연정, 박석, "안전한 데이터베이스 환경에서 삭제 시 효과적인 데이터 익명화 유지기법", 정보보호학회논문지, 제17권 제3호, pp.69-80, 2007.

[3] L. Sweeney, "Uniqueness of Simple Demographics in the U.S. Population", Garnegie Mellon University, Laboratory for International Data Privacy, 2000.

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

[5] A. Machanavajjhala, J.Gehrke, and D. Kifer, "ℓ-diversity : Privacy beyond k-anonymity", In Proc. of the International Conference on Data Engineering(ICDE'06), pp.24-35, Atlanta, GA, USA, 2006

[6] 성민경, 정연돈, "소셜 네트워크 데이터의 프라이버시 보호 배포를 위한 모델", 정보과학회 논문지, 데이터베이스 제37권 제4호, pp. 209-219, 2010

[7] J. W. Byun, U. Sohn, E. Bertino, and N. Li, "Secure Anonymization for IIncremental Datasets", 3rd VLDB Workshop, Secure Data Management 2006, pp.48-63, Seoul, Korea, 2006

[8] Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan, "Incognito : Efficient Full-Domain K-Anonymity", Paper presented at the ACM SIGMOD Conference on Management of Data, 2005

[9] C. Blake and C. Merz. UCI repository of machine learning databases, 1998, http://archive.ics.uci.edu/ml/datasets/Adult

저자약력

정 은 희(Eun-Hee Jeong) **정회원**

1991년 2월 강릉대학교 통계학과 이학사
1998년 2월 관동대학교 전자계산공학과 공학석사
2003년 2월 관동대학교 전자계산공학과 공학박사
2003년 9월 ~ 현재 강원대학교 지역경제학과 교수

<관심분야> 네트워크 보안, 전자상거래, 웹 프로그래밍

이 병 관(Byung-Kwan Lee) **정회원**

1979년 2월 부산대학교 기계설계학과 학사
1986년 2월 중앙대학교 전자계산공학과 공학석사
1990년 2월 중앙대학교 전자계산공학과 공학박사
1988년 ~ 현재 관동대학교 컴퓨터학과 교수

<관심분야> 네트워크 보안, 전자상거래, 컴퓨터 네트워크