논문 14-39B-06-05
The Journal of Korea Information and Communications Society '14-06 Vol.39B No.06
http://dx.doi.org/10.7840/kics.2014.39B.6.370

안전한 다중집합 빈도 계산 기법

김 명 선◆°, 박 재 성*

A Secure Frequency Computation Method over Multisets

Myungsun Kim◆°, Jaesung Park*

요 약

잘 알려진 바와 같이 데이터마이닝 (Data Mining)은 대용량의 데이터를 분석하여 필요한 정보를 추출하는데 있어서 매우 중요한 역할을 수행한다. 그중에서 집합에 포함된 원소들의 빈도수 (Frequency)를 알아내는 것은 데이터마이닝에서 기본적으로 지원되어야 하는 필수기능이다. 동시에 사용자가 소유한 다중집합 (혹은 집합) 자체의 공개를 원하지 않는 경우에 대비하여 다중집합의 원소는 공개하지 않고 빈도수만 계산하는 방법이 필요하다. 본 논문에서는 암호학적 도구를 기반으로 사용하여 이러한 조건을 만족하는 기법을 개발하고, 이것의 안전성을 엄밀하게 증명한다. 본 논문에서 제안된 기법은 기존 기법들과 달리 첫째, 시스템 가정이 일반적이고 둘째, 통신/연산 복잡도가 효율적이고 마지막으로 엄밀한 안전성 증명을 제시한다.

Key Words : Privacy-preserving, Frequency, Multiset

ABSTRACT

It is well known that data mining plays a crucial role in varities of real-world applications, by which extracts knowledge from large volume of datasets. Among functionalties provided by data mining, frequency mining over given multisets is a basic and essential one. However, most of users would like to obtain the frequency over their multisets without revealing their own multisets. In this work, we come up with a novel way to achive this goal and prove its security rigorously. Our scheme has several advantages over existing work as follows: Firstly, our scheme has the most efficient computational complexity in the cardinality of multisets. Further our security proof is rigorously in the simulation paradigm. Lastly our system assumption is general.

Ⅰ. 서 론

데이터마이닝 (Data Mining)은 다양한 데이터로부터 필요한 지식 (Knowledge)을 추출하는 중요한 도구로서 현실에서 아주 중요한 역할을 한다. 그러나 입력으로 주어지는 데이터 자체에 대한 프라이버시 (Privacy)의 훼손과 나아가 데이터를 처리하여 얻은 결과가 프라이버시를 훼손하는 결과로 이어지곤 한다. 이러한 이유로 데이터마이닝 분야에서 프라이버시 문제가 크게 주목받고 있다. 이러한 이유로 프라이버시 문제를 기술적으로 해결하기 위한 다양한 연구 결과들이 제시되고 있다[2,7,14].

데이터마이닝 분야에서 프라이버시 문제를 해결하는 방법은 크게 두 가지로 분류할 수 있다. 첫 번째 부류는 섭동기반 (Perturbation-based) 기법으로 연산이 매우 효율적인 것을 특징으로 한다. 두 번째 부류는 암호기반 (Cryptography-based) 기법으로 효율성보다는 프라이버시와 정확성 (Accuracy) 보장을 주목적으

◆° First Author and Corresponding Author : Suwon University, msunkim@suwon.ac.kr, 정회원
* 수원대학교 정보보호학과, jaesungpark@suwon.ac.kr, 종신회원
논문번호 : KICS2014-04-121, Received December April 7, 2014; Revised May 28, 2014; Accepted May 28, 2014

로 한다. 첫 번째 부류에 속하는 기법의 대표적인 예로는 Agrawal과 Aggarwal 등에 의해서 제안된 기법[1]과 Evmievski 등에 의해서 제안된 기법[5]을 들 수 있다. 전술한 바와 같이 섭동을 기반으로 하는 기법은 매우 효율적이나 프라이버시와 정확성간의 상충관계(Tradeoff)로 인하여 높은 정확성을 얻기 위해서는 프라이버시 수준을 낮추어야 한다.

반면에 암호학적 기법을 기반으로 하는 기법들 (예, [7,17,11])은 프라이버시와 정확성을 모두 달성할 수 있지만, 연산과 통신복잡도가 높다. 그러나 두 가지 부류의 기법은 모두 폭넓은 현실세계의 응용을 가능하게 하고, 동시에 사용자가 소유한 데이터의 프라이버시를 보장하면서 사용자 집합들을 이용하여 연산을 수행한다.

주어진 집합의 빈도수를 계산하는 연산은 데이터마이닝의 기본 연산으로 다른 데이터마이닝 연산들에 폭넓게 이용되기 때문에 다양한 기법들이 제안되었다. Vaidya와 Clifton에 의해서 제안된 기법은 다자간연산 (Multiparty Computation, MPC) 기법을 수직으로 분할된 데이터베이스 (Database)에 적용하여 나이브 베이지언 분류 (Naive Bayes Classification), 연관규칙 (Association Rule), 결정트리 (Decision Tree)를 얻을 수 있도록 한다[12,13]. Kantarcoglu와 Vaidya는 수평으로 분할된 데이터베이스에 동일한 방법을 적용하여 나이브 베이지언 분류가 가능함을 보였다[9]. 데이터베이스의 분할을 가정하지 않는 일반적인 시스템 모델에서 프라이버시를 보장하면서 빈도를 계산하려면 좀 더 복잡한 기법을 필요로 하지만 성능상의 이유로 암호기반 기법대신 섭동기반 기법을 이용한다[3,19]. 더구나 프라이버시에 대한 정의와 안전성 분석이 엄밀하지 않다.

본 논문에서는 기존 기법들과 달리 연산과 통신복잡도가 높은 MPC 기법을 사용하지 않고, 일반적인 시스템 모델에서 다수의 사용자들이 소유한 다중집합으로부터, 프라이버시를 보장하면서 빈도수를 계산하는 방법을 제안하고자 한다. 특히 암호학적으로 프라이버시를 정의하고 제안하는 기법이 안전하다는 것을 증명하고자 한다.

다음과 같은 응용 사례에 제안하는 기법을 적용할 수 있다. 특정 질병과 생활 습관의 관계를 진료 기록과 환자의 설문조사를 바탕으로 연구하려는 상황을 생각하자. 병원과 환자는 모두 진료기록과 자신의 설문조사 결과가 드러나는 것을 원하지 않을 것이다. 특히 연산의 결과가 높은 정확성을 요구하는 동시에 프라이버시가 보장될 필요가 있기 때문에 섭동기반 기법으로는 한계가 있다.

1.1 기여하는 점

본 논문에서는 이러한 상황을 고려하여 암호학적 기법을 활용하여 다수의 사용자의 다중 집합으로부터 다음과 같은 요구조건을 만족하는 안전한 프로토콜을 제안하고자 한다:

(1) 정확성과 프라이버시가 동시에 보장된다.
(2) 프라이버시가 보장된다는 것이 암호학적으로 엄밀하게 증명된다.
(3) 실용적으로 받아들일 수 있는 수준의 연산복잡도를 제공한다.

1.2 관련연구

본 논문에서 제안하는 기법과 유사한 목적을 달성하는 다양한 기법들이 관련 분야에서 수 년간 연구되어 왔다.

앞에서 언급한 바와 섭동기반 기법을 적용하여 데이터마이닝하는 방법으로는 2000년 SIGMOD에 제안된 Agrawal과 Srikant에 의해서 제안된 기법이 처음이다[2]. 그 후, [1,5,3]가 대표적이다. 핵심 아이디어는 사용자의 데이터를 데이터마이닝하기 전에 특정 노이즈 (Noise) 값을 적용하여 난수값처럼 보이게 변형하는데, 이것을 섭동이라 한다. 그 후 데이터마이닝 과정을 통하여 사전에 결정된 에러 범위안에서 원본 데이터를 재구성하고, 이 결과를 바탕으로 원하는 함수에 적용하는 것이다. 데이터마이닝 시점에서 원본 데이터와 유사한 데이터 상에서 연산이 이루어지므로 원하는 함수 연산이 매우 효율적이다. 그러나 원본 데이터와 동일한 데이터가 재구성되는 것이 아니므로 정확성이 입력 데이터에 적용한 노이즈에 의존한다. 즉 노이즈가 클수록 정확성은 떨어지나 프라이버시는 강화되는 것이다.

이러한 정확성과 프라이버시 간의 상충관계 문제를 해결하기 위해 암호학적 기법을 적용하려는 다양한 시도가 이루어졌다. 먼저 [12,13]에서 MPC 기법을 적용하여 데이터마이닝을 할 수 있는 기법이 적용되었으나, 이미 잘 알려진 바와 같이 MPC는 실용적인 수준의 연산복잡도를 허용하지 않기 때문에 이론적인 가능성을 제시한 것으로 보아야 할 것이다. 더구나 데이터베이스가 수직이나 수평으로 분할되어 있다는 가정은 상황에 따라 적용하기 어렵기도 하다. [17,16]에서 프라이버시를 보장하면서 빈도수를 계산할 수 있는 방법이 제안되었으나 2명의 사용자만 참여할 수 있는 환경으로 제한되며 더구나 사용자의 집합을 입

력으로 사용하여 빈도수를 계산하는 별도의 사용자가 필요하다. 본 연구와 가장 유사한 결과는 Luong와 Ho에 의해서 제안되었으나 참여하는 사용자들의 쌍이 서로 데이터베이스의 속성의 부분집합을 공유해야 하는 가정을 필요로 한다[8]. 이러한 가정은 극히 일부의 경우를 제외하고 실용적인 응용이 불가능하며 다중집합을 고려하지 않는다.

섭동기반 기법과 암호기반 기법의 하이브리드 방법으로 k-anonymization을 이용하여 유사한 결과를 얻을 수 있는 기법도 제시되었다[19,18]. 이 기법들은 공개되는 데이터베이스의 k개의 투플 (Tuple)들이 서로 구분되지 않도록 하는 것이 핵심 아이디어이다. 익명화 (Anonymize)되지 않는 속성값이 공개되어 [4]에 제시된 공격에 취약하다.

1.3 논문의 구성

본 논문은 다음과 같이 구성된다: 먼저 Ⅱ장에서는 본 논문에서 사용되는 기본 도구에 대한 배경지식을 설명하고 Ⅲ장에서는 제안하는 기법을 설명한다. 이어서 Ⅳ장에서는 제안하는 기법의 안전성과 복잡도를 분석하고 Ⅴ장의 맺음말로 마치고자 한다.

Ⅱ. 배경지식

본 장에서는 제안하는 기법을 설계하는데 필요한 암호학적 도구들을 간략하게 설명하고자 한다. 추가로 본 논문 전체에서 사용되는 표기법을 제시한다.

2.1 표기법

$X=\{a_1,\cdots,a_n\}$를 다중집합이라 하자. 즉, 어떤 $i\neq j$에 대하여 $a_i=a_j$일 수 있다. 이때 $F(a)$는 A의 원소 a의 빈도수 (Frequency)를 나타내며 $F(A)$는 모든 원소의 빈도수의 집합을 나타낸다. 함수 $\mu:\mathrm{N}\rightarrow\mathrm{R}$이 *negligible*하다는 의미는 모든 다항식 $p()$에 대하여 $\mu(\lambda)<1/p(\lambda)$을 만족하는 $\lambda>L$인 어떤 L이 존재한다는 의미이다.

$X=\{X(a,\lambda)\}_{a\in\{0,1\}^*,\lambda\in N}$와 $Y=\{Y(a,\lambda)\}_{a\in\{0,1\}^*,\lambda\in N}$을 어떤 분포의 앙상블 (Ensemble)이라 할 때, 모든 다항식 시간 알고리즘과 모든 $a\in\{0,1\}^*$ 및 $\lambda\in\mathrm{N}$에 대하여 두 앙상블 X,Y를 구분할 수 있는 가능성이 negligible할 때 두 앙상블은 구분불가능하다 (*Indistinguishable*)고 하며 $X\approx Y$로 표기한다.

끝으로 모든 사용자는 u_i로 표기하고 각 사용자가 소유한 다중집합은 X_i로 표기한다. 이에 더하여 각 $X_i=\{a_{i1},a_{i2},\cdots,a_{ik}\}$로 나타내며 편의를 위해 모든 다중집합의 크기는 k로 동일한 것으로 가정한다. 표기의 편의를 위해 $[n]=\{1,2,\cdots,n\}$로 표기한다.

2.2 암호학적 가정

제안하는 기법의 암호학적 가정은 크게 두 가지이다. 하나는 discrete logarithm (DL) 문제의 어려움에 대한 가정이고 다른 하나는 Decisional Diffie-Hellman (DDH) 문제의 어려움에 대한 가정이다.

정의 1. (**DL 문제** 및 **가정**) G를 군 (Group)의 위수 (Order)가 q이고 생성자가 g라 할 때, $y=g^x$를 만족하는 x를 찾는 문제를 DL 문제라 하며, 이 문제를 효율적으로 풀 수 있는 알고리즘이 존재하지 않는다는 것이 DL 가정이다.

반면에 DDH 가정은 다음과 같다.

정의 2. (**DDH 문제** 및 **가정**) G,q,g는 위와 동일하다고 하자. $a=g^{\alpha},b=g^{\beta}$와 $c=g^{\alpha\beta}$에 대하여 c가 $g^{\alpha\beta}$인지 혹은 어떤 난수 γ에 대하여 g^{γ}인지 판단하는 문제를 DDH 문제라고 하며, 이것을 효율적으로 판단하는 알고리즘이 존재하는 않는다는 가정을 DDH 가정이라 부른다.

2.3 El Gamal 암호기법

본 논문에서 제안하는 기법은 El Gamal 암호기법을 사용한다.

El Gamal 암호기법은 어떤 그룹 G에서 DDH 문제의 어려움에 기반한다. 특히 DDH 문제를 풀기 어려운 그룹을 DDH 그룹이라 하자. 이러한 DDH 그룹에서 El Gamal 암호기법은 선택평문공격 (Chosen Plaintext Attack, CPA)에 안전한 것이 증명되었다. 임의의 큰 소수 $p=2q+1$에 대하여 곱셈순환군 Z_p^*의 부분군을 G라 하면 이것은 DDH 그룹이 되며 생성자 g를 하나 선택할 수 있다.

이러한 El Gamal 암호기법의 개인키는 $x\in Z_q$이며 공개키는 $y=g^x \bmod p$에 의해서 주어진다. 평문 m이 주어지면 난수 r을 하나 선택한 후 $u=g^r$과 $v=m\cdot y^r$을 계산하여 $c=(u,v)$를 암호문으로 설정한다. 복호화는 암호문 c를 수신한 후, $m=v/u^x$에 의하여 얻을 수 있다. 추가로 El Gamal 암호화 기

법은 다수의 사용자가 사용하는 환경으로 쉽게 변형할 수 있다. 암호화 및 복호화 방법은 동일하며 각자 개인키 x_i를 생성한 후, $y_i = g^{x_i}$를 모든 사용자가 공유한 후 $y = \prod_{i=1}^{n} y_i$를 공개키로 사용한다.

2.4 Double Encryption 기법

Double encryption 기법은 Kim 등에 의해 제안된 기법으로 El Gamal 암호화기법과 함께 사용되어 다음 조건들을 만족한다[10].

(KG, Enc, Dec)를 El Gamal 암호화 기법의 키생성, 암호화 및 복호화 알고리즘이라 할 때

(1) 어떤 함수 $E: G \rightarrow G$가 존재하여
$Enc \circ E(m) = E \circ Enc(m)$을 만족하고

(2) 모든 $c = Enc(m)$에 대하여 다음 조건을 만족한다:
$E(m) = Dec \circ E(c)$

(3) 끝으로 함수 $D: G \rightarrow G$가 존재하여
$m = D \circ E(m)$을 만족한다.

위 세가지 조건을 만족하는 (E, D)가 존재하면 *double encryption* (DE) 이라 하며, 좀 더 자세한 것은 [10]를 참조한다.

본 논문에서는 double encryption의 구체적인 예로서 El Gamal암호 기법을 사용하였으나 조건을 만족하는 다른 공개키 암호 기법을 사용하여 구현하는 것이 가능하다.

2.5 안전성 정의

본 장에서는 프라이버시가 보장되는 것을 의미하는 프로토콜의 안전성을 형식적으로 정의한다. 다자간 프로토콜의 안전성에 대한 자세한 내용과 동기 등은 [6]를 참조한다.

인덱스 집합 $I = [n]$에 대하여 $u_{i \in I}$와 $X_{i \in I}$를 각각의 사용자 및 그의 다중집합이라 하자. 추가로 공격자 A가 제어할 수 있는 사용자의 집합을 I라 하자. 그리고 f를 입력된 다중집합의 원소를 드러내지 않고 빈도수만 계산하는 functionality라 하자. λ를 안전성 매개변수라 할 때 $IDL_{f,A,I}(X_1, \cdots, X_n; \lambda)$을 이상적인 제 3자 (Trusted Thrid Party, TTP)가 존재하는 가상의 공간에서 연산을 수행하는 경우의 출력이라 하자. 여기서 이상적인 제 3자가 의미하는 것은 어떠한 공격자에 의해서도 안전성이 훼손되지 않고 모든 사용자가 신뢰하는 제 3의 기관을 의미한다. 그러나 현실세계에 이러한 TTP는 존재하지 않으며 이상적인 공간에만 존재한다. 그에 반하여 f를 현실세계에서 구현한 어떤 프로토콜 π를 동일한 입력에 대하여 수행하여 얻은 출력을 $RAL_{\pi,A,I}(X_1, \cdots, X_n; \lambda)$이라 하자.

직관적으로 설명하면 이상적인 공간에서 TTP가 계산하는 수준의 안전성을 제공한다는 것을 형식적으로 기술하기 위한 것으로, 이상적인 공간은 공격자가 존재하여도 TTP가 계산을 해주므로 이러한 상황과 동치인 결과를 만드는 어떤 프로토콜이 있다면 이것은 이상적인 공간의 TTP를 흉내낸 것이고 안전하다고 말할 수 있다는 것이다. 이제 이러한 개념과 표기법을 사용하여 안전한 빈도수 계산 프로토콜을 정의한다.

정의 3. (안전한 빈도수 계산 프로토콜) f, π가 위와 같이 정의된다. 이때 현실세계의 모든 다항식 시간 공격자 A에 대하여, 다음 조건을 만족하는 이상적인 공간의 공격자 S가 존재하면 semi-honest 모델에서 빈도수 계산 프로토콜 π가 *안전하다* (또는 *프라이버시를 보존한다*)고 한다:

$$IDL_{f,A,I}(\{X\}_{i \in I}; \lambda) \approx RAL_{\pi,A,I}(\{X\}_{i \in I}; \lambda).$$

Ⅲ. 제안하는 기법

이 장에서는 El Gamal 암호화 기법과 이를 이용한 DE 기법을 적용하여, 본 논문에서 제안하는 기법을 설명한다.

이를 위하여 우선 λ는 안전성을 나타내는 매개변수라 하고, El Gamal 암호화 기법과 DE 기법을 위한 $p = 2q + 1$에 대하여 군 G를 위수가 p인 곱셈순환군 Z_p^*의 부분군으로 위수가 q라 하자.

3.1 빈도수 계산을 위한 구체적 DE 기법

El Gamal 기법과 함께 사용할 구체적인 DE 기법은 이미 [10]에서 제시된 바와 같이 군 G에서의 s에 의한 지수승 연산을 이용하는 것이다. 좀 더 구체적으로 설명하면

$$E: G \rightarrow G, \quad m \mapsto m^s$$

로 정의되면 이에 대응하는 복호화 함수 D는 $s \cdot t = 1 \bmod q$를 만족하는 t에 의한 지수승 연산

을 이용하는 것이다. 각각을 $E_s(\cdot), D_t(\cdot)$로 표기한다.

주의할 것은 본 논문에서 제안하는 기법은 DE 기법의 첫 번째 조건만 만족하면 된다. 왜냐하면 빈도수를 계산하는 경우에는 복호화 함수 D가 필요하지 않기 때문이다. 그러므로 실제로는 비밀키로 (s,t)를 사용하는 대신 s만 사용하는 것이 [10]에서 제안된 기법과 다른 점이다.

3.2 제안하는 기법의 구체적 내용

주어진 n명의 사용자가 제시한 다중집합의 빈도수를 안전하게 계산하기 위한 프로토콜은 다음과 같다.

Protocol π: Private Frequency Counting

Setup 단계

각 사용자 $u_{i\in[n]}$는 각자 순서대로 다음 작업을 수행한다.

Z_q에 속하는 난수 x_i를 선택한 후, $y_i = g^x$를 계산한 후 자신의 개인키 $sk = x_i$로 설정한 후 다른 모든 사용자에게 y_i를 전송한다.

다른 모든 사용자들로부터 $y_{j\neq i}$를 수신한 후 다음을 계산하고

$$y = y_i \cdot \prod_{j\in[n]\setminus\{i\}} y_j$$

공개키 $pk = (G,p,q,g,y)$을 지정한다.

추가로 $s_i \in Z_q$를 하나 선택하여 자신의 비밀키로 설정한다.

DE 단계

각 사용자 u_i는 전단계에서 설정한 공개키 y를 이용하여 다음 단계를 수행한다.

자신의 다중집합 X_i의 모든 원소 a_{ij}에 El Gamal 암호화 기법을 적용하여 암호화한다. 즉, 모든 $j\in[k]$에 대해 $\overline{a}_{ij} = Enc_{pk}(a_{ij})$을 계산하여 $\overline{X}_i = \{\overline{a}_{ij}\}_{j\in[k]}$을 얻는다.

u_1을 제외한 모든 사용자 u_i는 $\overline{X}_i$를 u_1에게 전송한다.

모든 사용자들로부터 $\overline{X}_i$를 수신한 u_1은 자신의 비밀키 s_1을 사용하여 $E_{s_1}(\overline{X}_i)$를 모든 $i\in I$에 대해서 수행한다. 여기서 각 $E_{s_1}(\overline{X}_i)$는 다음과 같이 정의된다:

$$E_{s_1}(\overline{X}_i) := \left\{E_{s_1}(\overline{a}_{i1}),\cdots,E_{s_1}(\overline{a}_{ik})\right\}$$

여기서 각 $E_{s_1}(\overline{a}_{ij})$는 다음과 같이 계산된다:

$$\begin{aligned} E_{s_1}(\overline{a}_{ij}) &= (\overline{a}_{ij})^{s_1} = Enc_{pk}(a_{ij})^{s_1} \\ &= Enc_{pk}(a_{ij}^{s_1}) = Enc_{pk}(E_{s_1}(a_{ij})). \end{aligned}$$

Mixing 단계

각 사용자 $u_1,\cdots,u_n$는 차례대로 다음 단계를 수행한다.

u_1은 $[n]$상의 임의의 permutation σ_1을 하나 선택하고 $Y_0 = \left(E_{s_1}(\overline{X}_1),\cdots,E_{s_1}(\overline{X}_n)\right)$에 σ_1을 적용하여 섞은 후, 이 벡터를 Y_1으로 설정하고 u_2에게 전송한다.

$\ell\in\{2,\cdots,n\}$에 대하여, 사용자 u_ℓ는 $Y_{\ell-1}$을 수신한 후, 자신의 비밀키 s_ℓ을 이용하여 $Y_\ell = E_{s_\ell}(Y_{\ell-1})$을 계산한다.

그 다음에 사용자 u_ℓ은 $[n]$상의 임의의 permutation σ_ℓ을 선택하여 벡터 Y_ℓ을 섞은 후, $u_{\ell+1}$에게 전달한다.

마지막으로 u_n은 Y_{n-1}을 받은 후 다른 사용자들과 같이 $Y_n = E_{s_n}(Y_{n-1})$을 계산한 후, σ_n을 이용하여 섞은 후 모든 사용자에게 전송한다.

Reveal 단계

모든 사용자들이 El Gamal 암호화 기법의 복호화에 공동으로 참여하여 Y_n의 복호화를 수행한다. 이를 통하여 각 사용자는

$$\begin{aligned} Dec_{sk}(Y_n) &= \left(E_s(a_{\sigma(1)}),\cdots,E_s(a_{\sigma(nk)})\right) \\ &= \left(a_{\sigma(1)})^s,\cdots,(a_{\sigma(nk)})^s\right) \end{aligned}$$

를 얻는다, 여기서 $s = \prod_{i=1}^{n} s_i$이며, permutation $\sigma = \sigma_n \circ \cdots \circ \sigma_1$이다.

Remark. User의 순서를 결정하는 방법은 순서를 결정하는 그룹관리자를 선출하는 것이 가장 간단한 방법의 하나이다. 이러한 경우 선출 알고리즘으로 다양한 선택이 가능한데, 예를 들면 Wang 등에 의해 제안된 [15]을 사용할 수 있다.

Ⅳ. 제안하는 기법의 분석

이 장에서는 제안하는 기법의 복잡도와 안전성을 분석한다. 군 G상에서의 지수승의 개수를 세어서 전체 연산복잡도를 제시한다. 그리고 안전성은 semi-honest 모델에서 제안하는 기법이 안전하다는 것은 simulation 기법을 이용하여 분석한다.

4.1 연산/통신 복잡도 분석

먼저 각 사용자 별로 계산해야 하는 지수승 연산의 횟수는 단계별로 설명하면 다음과 같다:

Setup 단계: 1번

DE 단계: $4k$번

Mixing 단계: $2nk$번

Reveal 단계; nk번

그러므로 총 연산량은 $n(1+4k+3nk)$이므로 $O(n^2k)$임을 알 수 있다. 여기서 사용자의 제곱에 연산량이 비례하는 것으로 보이나 사용자의 수 n에 대하여 $k \gg n$이며 많은 응용에서 n을 상수로 취급할 수 있으므로, 이러한 경우 k에 선형의 연산복잡도를 갖는다고 할 수 있다.

다음으로 사용자별 통신량을 계산하면 단계별로 다음과 같다:

Setup 단계: $\log p$ 비트

DE 단계: $2k\log p$ 비트

Mixing 단계: $2nk\log p$ 비트

Reveal 단계: $nk\log p$ 비트

그러므로 총 통신량은 $n(1+2k+3nk)\log p$이므로 $O(n^2k\log p)$ 비트이며, 위와 같은 이유로 k에 선형의 통신 복잡도를 갖는다.

마지막으로 라운드 (Round) 복잡도를 살펴본다. 모든 사용자에 대하여 단계별로 다음과 같다:

Setup 단계: 1 회

DE 단계: 1회

Mixing 단계: n회

Reveal 단계: 1회

그러므로 총 라운드의 횟수는 $n+3$이므로 $O(n)$으로 나타낼 수 있고, 사용자의 수에 선형으로 비례한다. 그러나 위에서 언급한 바와 같이 사용자의 수 n은 상수로 취급할 수 있는 경우나, $n \leq 10$인 경우에 효율적으로 수행할 수 있다.

4.2 안전성 분석

먼저 개념적으로 안전한 이유를 살펴보면 DE 단계에서 사용자가 다른 사용자에게 공개하는 값은 El Gamal 암호화 기법으로 암호화한 자신의 집합 $\overline{X_i}$이므로 프라이버시가 보장된다. 그 후 Mixing 단계에서 다른 사용자가 전송한 암호화된 다중집합에 자신의 비밀키를 이용하여 $Y_\ell = E_{s_\ell}(Y_{\ell-1})$을 수행하는데 이때 $Y_{\ell-1}$로부터 어떠한 정보도 얻을 수 없다. 더구나 σ_ℓ을 사용하여 벡터 Y_ℓ의 원소를 섞기 때문에 다음 사용자는 이전 사용자의 σ_ℓ을 모르는 상태에서 어떤 원소가 사용자 u_i에 속하는지 알 수 없다.

여기서 permutation을 사용하는 이유는 u_ℓ이 $u_{\ell-1}$에게 수신한 집합의 원소가 어떤 사용자의 것인지를 숨기기 위한 것이다. 만약 permutation을 하지 않으면 전체 집합의 빈도수뿐만 아니라 개별 다중집합의 빈도수가 드러나는 문제점이 있다.

형식적인 증명을 위하여 먼저 빈도수를 계산하는 이상적 공간의 functionality f를 먼저 정의한다.

정의 4. (이상적 Functionality F_{fred}) 이상적 공간(Ideal-World)에는 n명의 사용자가 존재하여 $\{u_i\}_{i\in[n]}$로 나타내고, TTP는 T로 나타낸다. 이러한 공간의 공격자는 simulator라고 하며 S로 나타낸다. 그러면

1) 각 사용자의 다중집합 X_i를 T에게 전송한다.
2) T는 다음을 계산하여 각 사용자 u_i에게 Φ_i를 전송한다:

$$\Phi_i = \left\{ F(a) | a \in \bigcup_{i=1}^{n} X_i \right\}$$

정리 1. 앞에서 주어진 다중집합의 빈도수를 계산하는 프로토콜 π는 모든 사용자의 다중집합의 빈도수를 옳게 계산한다.

증명. 먼저 전체 집합을 $U = \bigcup_{i=1}^{n} X_i$로 나타내자.

이때 $|U| = nk$이다. 먼저 DE 단계를 수행한 후 각 사용자는 자신들의 다중집합의 암호화된 값들을 갖게 된다. 그 후 사용자 u_1이 모든 사용자들의 다중집합 Y_0에 대하여 $Y_1 = E_{s_1}(Y_0)$를 계산하고 σ_1을 사용하여 섞어서 다음 사용자에게 전달하는 과정을 거치면 최종적으로 u_n은 Y_n을 받게 된다. 이때 벡터 Y_n은 Y_0에 $\sigma = \sigma_n \circ \cdots \circ \sigma$을 수행하여 섞은 후, 지수승 $s = \prod_{i=1}^{n} s_i$을 수행한 것과 같다. 그러므로 임의의 $a_{ij} \in U$에 대하여 $E_s(\bar{a}_{ij}) \in Y_n$이고 복호화를 수행한 후 $E_s(a_{ij}) \in Dec_{sk}(Y_n)$이다. 그런데 함수 E는 군 G에서 단사함수이므로 모든 $a_{ij} \in U$에 대하여 다음이 성립한다:

$$\{F(a_{ij}) | a_{ij} \in U\} = \{F(E_s(a_{ij})) | E_s(a_{ij}) \in \tilde{U}\},$$

여기서 $\tilde{U} = Dec_{sk}(Y_n)$이다. □

정리 2. 앞에서 주어진 다중집합의 빈도수를 계산하는 프로토콜 π는 이상적인 공간에서 TTP에 의해서 수행되는 이상적 functionality F_{freq}를 안전하게 계산한다.

증명. 안전성을 증명하기 위하여 다항식 시간안에 수행되는 simulator S가 존재함을 보이는 것으로 충분하다. 이를 위하여 simulator S는 다음과 같이 동작한다. 사용자의 인덱스 집합은 $I = [n]$로 나타내며, 공격자에 의해서 제어되는 semi-honest 사용자의 집합을 Γ라 할 때 이러한 사용자의 집합에 대한 인덱스 집합을 J라 하자. 다르게 이야기 하면 인덱스 집합 J에 포함된 인덱스를 갖는 사용자는 공격자에 의해서 제어되는 semi-honest 사용자일 것이다. 증명의 편의를 위하여 $u_1 \notin \Gamma$라 하자. 당연히 $u_1 \in \Gamma$ 경우는 더 간단하기 때문에 이러한 경우는 제시된 증명으로부터 쉽게 이해할 수 있다.

Setup 단계: 먼저 S는 정직한 사용자를 흉내내기 위하여 $\alpha \in I \setminus J$에 속하는 인덱스를 사용하는 사용자 u_α의 개인키 x_α를 생성하고 $y_\alpha = g^{x_\alpha}$를 계산하여 모든 $u_{j \in J}$에게 전송한다. 추가로 s_α도 선택하여 저장한다.

DE 단계: Simulator S는 임의의 난수를 k개 생성하여 사용자 u_α의 다중집합 X_α를 생성한다. 그 후 공개키 y를 이용하여 El Gamal 암호화를 수행하고 모든 $u_{j \in J}$의 암호화된 다중집합 $Enc_{pk}(X_j)$의 수신을 기다린다. 이제 u_1이 정직한 사용자이므로 S는 다음 집합을 얻는다:

$$Y_0 = \bigcup_{\alpha \in I \setminus J} Enc_{pk}(X_\alpha) \cup \bigcup_{j \in J} Enc_{pk}(X_j)$$

그 후, $E_{s_1}(Y_0)$을 계산한 후, 오름차순으로 정렬하여 $u_{j'}$에게 전송한다. 이때 $j' = \min\{J\}$이다.

Mixing 단계: $\alpha \in I \setminus J$에 속하는 모든 u_α에 대하여 난수 s_α와 임의의 permutation σ_α를 생성한다. 이때 $j \in J$를 사용하는 semi-honest 사용자 u_j로부터 Y_j를 수신하면 s_α, σ_α를 사용하여 Y_α를 계산한다. 만약 $(\alpha+1) \in I \setminus J$이라면 $j' = \min\{(\alpha+1) < j \wedge j \in J\}$를 만족하는 $u_{j'}$에게 Y_α를 전송한다. 이와 달리 $(\alpha+1) \in J$라면 $u_{\alpha+1}$에게 Y_α를 전송한다. 마지막으로 사용자 u_n이 정직한 사용자라면 S는 앞서 미리 생성한 s_n, α_n을 이용하여 Y_n을 모든 사용자 $u_{j \in J}$에게 전송한다.

Reveal 단계: Simulator S는, 모든 $\alpha \in I \setminus J$에 대하여 x_α를 이용하여 복호화에 참여한다.

이제 각 단계에서 simulator S의 출력을 현실세계에서 프로토콜 π의 출력과 비교한다. 먼저 Setup 단계에서 π의 출력 y_i와 y_α는 DDH 가정에 의해 구분할 수 없다.

두 번째, DE 단계에서 S는 정직한 사용자의 다중집합을 알 수 없기 때문에 동일한 개수의 난수를 생성한 후 이것을 암호화한 $Enc_{pk}(X_\alpha)$ 값을 다른 사용자에게 전송한다. 그러나 El Gamal 암호화 기법의 안전성에 의해

$$Enc_{pk}(X_i) \approx Enc_{pk}(X_\alpha)$$

이다. 마찬가지로 Mixing 단계에서 S는 정직한 사용자의 비밀키를 모르기 때문에 모든 $\alpha \in I \setminus J$에 대하여 난수와 permutation의 쌍 $(s_\alpha, \sigma_\alpha)$을 생성하고 Y_α

를 계산하여 전송한다. 이때 DE 기법의 E의 첫 번째 특성에 의하여

$$E_{s_\alpha}(Y_{\alpha-1}) = \{E_{s_\alpha}(Enc_{pk}(X_i))\}_{i\in I} = \{Enc_{pk}(E_{s_\alpha}(X_i))\}_{i\in I}$$

이고 다시 El Gamal 암호화 기법의 안전성에 의하여

$$\{Enc_{pk}(E_{s_\alpha}(X_i))\}_{i\in I} \approx \{Enc_{pk}(E_{s_i}(X_i))\}_{i\in I}$$

임을 알 수 있다. 끝으로 Reveal 단계에서 S는 각 x_α를 사용하여 복호화에 참여하므로 구분할 수 있는 정보가 드러나지 않는다. 이상의 사실에 의하여

$$IDL_{F_{\text{freq}},A,I}(\{X_i\}_{i\in I};\lambda) \approx RAL_{\pi,A,I}(\{X_i\}_{i\in I};\lambda)$$

임을 알 수 있다. 이것으로 정리 2가 증명됨을 알 수 있다. □

V. 결 론

본 논문에서는 사용자의 다중집합을 입력으로 사용하여 입력 다중집합의 프라이버시를 보존하면서 원소의 빈도수만을 안전하게 계산하는 암호학적 기법을 제안하였다. 프라이버시를 보존한다는 것은 프로토콜을 수행한 후 결과와 결과로부터 자명하게 알 수 있는 것을 제외한 모든 정보를 드러나지 않고 빈도수만을 계산하는 것을 의미한다. 본 논문에서 제안하는 기법은 $O(n^2k)$의 연산 및 통신 복잡도를 갖는다. 특히 기존 연구 결과와 달리 일반적인 시스템 모델에서 엄밀한 안전성 증명을 보였다. 그러나 현재 시스템은 라운드 수가 사용자의 수 n에 비례하는 문제점이 있다. 그래서 n이 상수로 고정되거나 작은 경우에 효율적으로 사용할 수 있다. 이러한 문제를 해결하는 것은 이것 자체로 흥미로운 연구일 것으로 판단하며, 향후 연구 주제로 제시하고자 한다.

References

[1] D. Agrawal and C. Aggawal, “On the design and quantification of privacy preserving data mining algorithms,” in *Proc. PODS 2001*, pp. 247-255, 2001.

[2] R. Agrawal and R. Srikant, “Privacy preserving data mining,” in *Proc. SIGMOD 2000*, pp. 439-450, 2000.

[3] R. Agrawal, R. Srikant, and D. Thomas, “Privacy preserving OLAP,” in *Proc. SIGMOD 2005*, pp. 251-262, 2005.

[4] http://w2.eff.org/Privacy/AOL/

[5] A. Evmievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in *Proc. KDD 2002*, pp. 217-228, 2002.

[6] O. Goldreich, *Foundations of cryptography*: Vol. 2, Cambridge Press, 2004.

[7] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” in *Proc. Crypto 2000*, pp. 36-53, 2000.

[8] T. Luong and T. Ho, “Privacy preserving frequency mining in 2-party fully distributed setting,” *IEICE Trans. Inf. Syst.* Vol. E93-D, No. 10, Oct. 2010.

[9] M. Kantarcoglu and J. Vaidya, “Privacy preserving naive Bayes classifier for horizontally partitioned data,” in *Proc. ICDM 2003*, pp. 3-9, 2003.

[10] M. Kim, A. Mohaisen, J. H. Choen, and Y. Kim, “Private over-threshold aggregation protocols,” in *Proc. ICISC 2012*, pp. 472-486, 2012.

[11] J. Sakuma and R. Wright, “Privacy preserving evaluation of generalization error and its application to model and attribute selection,” in *Proc. ACML 2009*, pp. 338-353, 2009.

[12] J. Vaidya and C. Clifton, “Privacy prreserving association rule mining in vertically partitioned data,” in *Proc. SIGKDD 2002*, pp. 639-644, 2002.

[13] J. Vaidya and C. Clifton, “Privacy preserving naive Bayes classifier for vertically partitioned data,” in *Proc. SDM 2004,* pp. 522-526, 2004.

[14] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving data mining,” in *Proc. SIGMOD Record*, vol. 3, no. 1, pp. 50-57, 2004.

[15] Y. Wang, S. Hur, Y. Park, and J.-H. Choi,

"Efficient user selection algorithms for multiuser MIMO systems with zero-forcing dirty paper coding," *J. Commun. Networks(JCN)*, vol. 13, no. 3, pp. 232-239, 2011.

[16] F. Wu, J. Liu, and S. Zhong, "An efficient protocol for private and accurate mining of support counts," *Pattern Recognit. Lett.*, vol. 20, no. 1, pp. 80-86, 2009.

[17] Z. Yang, S. Zhong, and R. Wright, "Privacy preserving classification of consumer data without loss of accuracy," in *Proc. SDM 2005*, pp. 21-23. 2005.

[18] S. Zhong, Z. Yang, and T. Chen, "k -anonymous data collections," *J. Inf. Sci.*, vol. 179, no. 17, pp. 2948-2963, 2009.

[19] S. Zhong, Z. Yang, and R. Wright, "Privacy-enhancing k-anonymization of consumer data," in *Proc. PODS 2005*, pp. 139-147, 2005.

김 명 선 (Myungsun Kim)

1994년 2월 : 서강대학교 전자계산학과 졸업
2002년 8월 : KAIST 정보통신학과 석사
2012년 8월 : 서울대학교 수리과학과 박사
2012년 9월~현재 : 수원대학교 정보보호학과 조교수
<관심분야> 암호이론, 다자간연산

박 재 성 (Jaesung Park)

1995년 2월 : 연세대학교 전자공학과 졸업
1997년 2월 : 연세대학교 전자공학과 석사
2001년 2월 : 연세대학교 전기, 전자공학과 박사
2001년~2002년 : University of Minnesota (PostDoc.)
2002년~2005년 : LG전자(선임연구원)
2005년 현재 : 수원대학교 정보보호학과 부교수
<관심분야> 네트워크 성능 분석 및 프로토콜 개발