情報保護學會論文誌
第19卷 第5號, 2009. 10

프라이버시를 보호하는 접근제어가 가능한 키워드 검색 기법*

노 건 태,† 천 지 영, 정 익 래, 이 동 훈‡
고려대학교 정보경영공학전문대학원

Privacy Preserving Keyword Search with Access Control*

Geon Tae Noh,† Ji Young Chun, Ik Rae Jeong, Dong Hoon Lee‡
Graduate School of Information Management and Security, Korea University

요 약

민감한 개인정보를 보호하기 위해 데이터를 암호화하는 것은 필수적이다. 하지만 복호화 과정 없이 암호화된 데이터를 검색하기 위해서는 암호화된 데이터에서 검색이 가능한 효율적인 기법이 필요하다. 지금까지 수많은 검색 가능한 암호화 기법이 제안되었지만, 아직까지 이러한 기법들은 암호화된 데이터를 공유하기 위해 접근 권한을 갖고 있는 동적인 사용자(dynamic user)에 대해서 적합하지 않다. 기존의 검색 가능한 암호화 기법들에서는 특정 사용자(대칭키 환경에서의 데이터 제공자, 공개키 환경에서 데이터를 암호화한 공개키에 대응되는 비밀키를 갖고 있는 사용자)에 대해서만 암호화된 데이터에 접근이 가능하였다. 이러한 문제를 해결하기 위해 Stephen S. Yau 등은 데이터 공급자의 접근 정책에 따라서 사용자의 검색 능력을 제어할 수 있는 기법을 처음으로 제안하였다. 그러나 이 기법은 데이터 검색자의 프라이버시가 노출되는 문제점을 가진다. 따라서 본 논문에서는 이 기법의 문제점을 분석하고, 이러한 문제를 해결한 프라이버시를 보호하는 접근제어가 가능한 키워드 검색 기법을 제안한다.

ABSTRACT

To protect sensitive personal information, data will be stored in encrypted form. However in order to retrieve these encrypted data without decryption, there need efficient search methods to enable the retrieval of the encrypted data. Until now, a number of searchable encryption schemes have been proposed but these schemes are not suitable when dynamic users who have the permission to access the data share the encrypted data. Since, in previous searchable encryption schemes, only specific user who is the data owner in symmetric key settings or has the secret key corresponding to the public key for the encrypted data in asymmetric key settings can access to the encrypted data. To solve this problem, Stephen S. Yau et al. firstly proposed the controlled privacy preserving keyword search scheme which can control the search capabilities of users according to access policies of the data provider. However, this scheme has the problem that the privacy of the data retrievers can be breached. In this paper, we firstly analyze the weakness of Stephen S. Yau et al.'s scheme and propose privacy preserving keyword search with access control. Our proposed scheme preserves the privacy of data retrievers.

Keywords: Keyword search, Privacy, Encrypted data, Access control, PIR(Privacy information retrieval)

접수일(2009년 7월 2일), 게재확정일(2009년 8월 10일)
* 본 연구는 지식경제부 및 정보통신연구진흥원의 IT산업원천기술개발사업(2009-S-001-02, Car-헬스케어 보안 기술개발)과 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음. (KRF-2008-331-D00581)
† 주저자, oldos@korea.ac.kr
‡ 교신저자, donghlee@korea.ac.kr

I. 서 론

최근 인터넷 서비스 업체들의 데이터베이스(database) 유출사건을 계기로 데이터베이스 암호화 기술에 대한 관심이 급증하고 있다. 데이터베이스 유출은 데이터베이스에 대한 암호화 없이 단순히 접근이 허가된 사람들만 데이터베이스 접근이 가능하도록 한 접근제어방식을 사용하는 환경에서 치명적이다. 이는 해킹이나 내부 사용자의 의도하지 않은 실수 등으로 인해 데이터베이스가 유출되었을 때, 데이터베이스의 정보가 그대로 노출되는 문제를 내포하고 있기 때문이다. 이러한 문제를 해결하기 위해 데이터베이스를 직접 암호화하는 방식을 사용할 수 있다. 하지만 데이터베이스 암호화 방식은 데이터베이스 유출에 대한 안전성을 보장하는 반면, 검색의 효율성을 급격히 저하시킨다.

따라서 최근 암호화된 데이터에서 검색이 가능한 효율적인 기법에 대한 연구가 활발히 진행되고 있다[1-9]. Dawn Song 등에 의해 최초로 제안된 대칭키 기반의 검색 가능한 암호 기법[2-5]은 자신의 데이터를 암호화하여 서버에 저장한 후 필요한 데이터를 검색하는, 데이터 공급자와 검색자가 일치하는 환경으로 이는 대칭키 암호화를 사용한 대칭키 기반의 검색 가능한 암호 기법이다. 또한 Dan Boneh 등에 의해 최초로 제안된 공개키 기반의 검색 가능한 암호 기법[6-9]은 데이터 공급자와 검색자가 일치하지 않는 환경으로 데이터 공급자는 특정 검색자를 위해 공개키 기반의 암호화를 사용하여 데이터를 암호화한 뒤 서버에 저장하여 검색자가 필요한 데이터를 검색할 수 있게 한다.

하지만 이러한 기법들은 모두 특정 사용자만이 복호화가 가능하다. 대칭키 기반의 검색 가능한 암호 기법은 데이터 공급자만이 복호화가 가능하고, 공개키 기반의 검색 가능한 암호 기법은 암호화에 사용되는 공개키에 대응되는 비밀키를 가진 사용자만이 복호화가 가능하다. 따라서 데이터베이스를 공유하는 환경에서 데이터베이스 접근이 허가된 여러 사용자가 암호화된 데이터를 검색하도록 하려면 기존의 기법들을 사용하기 어렵다. 최근 암호화된 데이터에 접근할 수 있는 사용자들을 지정하여 접근 권한을 가진 사용자들만이 접근할 수 있도록, 암호화된 데이터베이스에서 접근제어방식을 사용하기 위한 기법[10]이 Stephen S. Yau 등에 의해 최초로 제안되었다. 이 기법은 암호화된 데이터베이스에 대해 접근 가능한 사용자들만이 접근 가능하도록 데이터 공급자가 접근 권한을 지정할 수 있다는 장점을 가지고 있다. 하지만 이 기법은 신뢰할 수 없는 서버에게 검색자의 인덱스가 노출되는 문제점을 가진다. 따라서 본 논문에서는 서버에 검색자의 프라이버시가 노출되는 문제를 해결한 새로운 기법을 제안한다.

II. 배경지식

2.1 겹선형 군 (bilinear groups)[11]

우리의 기법은 겹선형 맵을 제공하는 합성 위수의 어떤 유한군을 사용한다. 여기서는 다음과 같은 일반적인 표기를 사용한다.

1) G와 G_1은 같은 유한 위수 n을 가지는 두 개의 (곱셈의) 순환군이다.
2) g는 군 G의 생성자이다.
3) e는 겹선형 맵 $e: G\times G\rightarrow G_1$이다. 다시 말하면, 모든 $u,v\in G$와 $a,b\in Z$에 대해서 $e(u^a,v^b)=e(u,v)^{ab}$를 만족한다. 또한 $e(g,g)$는 G_1의 생성자이다.

주어진 위수 n의 겹선형 군을 구성하기 위해, 주어진 제곱 인수가 없는(square-free) 수가 $n>3$이라고 하자. 여기서 우리는 위수 n의 겹선형 군을 다음과 같이 구성한다.

1) $p=ln-1$이 소수이고 $p=2 \bmod 3$을 만족하는 가장 작은 양의 정수 $l\in Z$를 찾는다.
2) F_p에 정의된 타원 곡선 $y^2=x^3+1$ 상의 점들의 군을 고려하자. $p=2 \bmod 3$이기 때문에 곡선은 F_p에서 $p+1=ln$개의 점들을 가진다. 그러므로 곡선상의 점들의 군은 위수 n을 가지는 군 G의 부분군을 가진다.
3) 위수 n를 가지는 $F_{p^2}^*$의 부분군을 G_1이라고 하자. 곡선상에서 수정된 웨일 페어링(weil pairing)은 필요한 속성과 함께 겹선형 맵 $e: G\times G\rightarrow G_1$을 제공한다.

2.2 부분군 결정 문제 (subgroup decision problem)[11]

보호 변수(secret parameter) $\tau\in Z^+$가 주어졌을 경우 튜플 (q_1,q_2,G,G_1,e)를 출력하는 알고리즘 ζ를 정의한다. 여기에서 G, G_1은 위수 $n=q_1q_2$를 가지는 그룹이고, $e: G\times G\rightarrow G_1$는 겹선형 맵이다. 입력 τ에 대

해 알고리즘 ζ는 다음을 수행한다.

1) 두 개의 랜덤한 τ 비트 소수 q_1, q_2를 생성하고, $n=q_1q_2\in Z$라고 하자.
2) 위수 n을 가지는 겹선형 군 G를 생성한다. 그리고 g를 G의 생성자라고 하고, $e: G\times G\rightarrow G_1$를 겹선형 맵이라고 하자.
3) (q_1, q_2, G, G_1, e)를 출력한다.

여기서 겹선형 맵뿐만 아니라 G, G_1에서의 그룹 연산도 다항식 시간 내에 계산될 수 있다.

$\tau\in Z^+$라고 하고 $n=q_1q_2$인 경우에 (q_1, q_2, G, G_1, e)를 $\zeta(\tau)$에 의해 생성된 튜플이라고 할 때, 부분군 결정 문제를 다음과 같이 정의한다.

> (n, G, G_1, e)과 원소 $x\in G$가 주어졌을 때, 만약 x의 위수가 q_1이면 '1'을 출력하고 아니면 '0'을 출력한다.

즉, 군의 위수 n의 인수분해 정보 없이 원소 x가 G의 부분군에 있는지를 결정하는 문제이다. 알고리즘 A에 대하여, 부분군 결정 문제를 풀기위한 A의 이점 $SD-Adv_A(\tau)$는 다음과 같이 정의된다:

$$\begin{aligned} SD-Adv_A(\tau) = &|pr[A(n, G, G_1, e, x)=1 : \\ &(q_1, q_2, G, G_1, e)\leftarrow\zeta(\tau), n=q_1q_2, x\leftarrow G] \\ &- pr[A(n, G, G_1, e, x^{q_2})=1 : \\ &(q_1, q_2, G, G_1, e)\leftarrow\zeta(\tau), n=q_1q_2, x\leftarrow G]| \end{aligned} \quad (1)$$

> 정의 2.1. 만약 어떤 다항식 시간 알고리즘 A에 대하여 $SD-Adv_A(\tau)$가 τ에서 무시해도 좋은(negligible) 함수라면, ζ는 부분군 결정 가정을 만족한다고 한다.

간단히 말해서, 이 가정은 G에서의 균일한 분포가 G의 부분군에서의 균일한 분포와 구별 불가능하다는 것을 설명한다. G의 위수의 인수분해는 다항식 시간 공격자에게 G의 부분군의 위수가 드러나지 않게 하기 위해서 감춰진다.

2.3 준 동형 공개키 시스템 (homomorphic public-key system)[11]

이 시스템은 키생성(KeyGen), 암호화(Encrypt), 복호화(Decrypt)의 세 가지 알고리즘으로 구성된다.

1) 키생성 알고리즘 (KeyGen(τ))
   ① 보안 변수 $\tau\in Z^+$가 주어졌을 경우, 튜플 (q_1, q_2, G, G_1, e)를 얻기 위해 $\zeta(\tau)$를 실행한다.
   ② 위수 $n=q_1q_2$인 군 G에서 두 개의 생성자 $g, u\in G$를 랜덤하게 선택한 후, $h=u^{q_2}$라고 한다.(h는 위수가 q_1인 군 G_1의 생성자가 된다.)
   ③ 공개키 $PK=(n, G, G_1, e, g, h)$와 비밀키 $SK=q_1$를 설정한다.

2) 암호화 알고리즘 (Encrypt(PK, M))
   ① 메시지 m의 메시지 공간을 $\{0, ..., T\}$라고 한다.($T<q_2$)
   ② 랜덤값 $r\in Z_n$을 선택하여 메시지 m의 암호문을 다음과 같이 계산하고, 암호문 C를 출력한다.

$$C=g^mh^r\in G \quad (2)$$

3) 복호화 알고리즘 (Decrypt(SK, C))
   ① 비밀키 $SK=q_1$로 암호문 C를 다음과 같이 복호화 한다.

$$C^{q_1}=(g^mh^r)^{q_1}=(g^{q_1})^m \quad (3)$$

   ② $\hat{g}=g^{q_1}$이라고 할 때, $m=\log_{\hat{g}}C^{q_1}$이고 $0\le m\le T$이므로, 메시지 m을 되찾기 위해서 Pollard's lambda 방법을 이용하면 $O(\sqrt{T})$의 시간이 걸린다.

이 시스템의 안전성은 부분군 결정 문제의 어려움에 기반을 둔다. 이 기법은 다음의 두 가지 성질을 만족한다.

1) 검증 가능성(verifiable) : 메시지 m의 암호문 g^mh^r과 메시지 m', 그리고 검증키 $vk=q_1$이 주어졌을 때, $(g^mh^r)^{q_1}=g^{q_1m'}$을 체크하여 $m=m'$인지를 확인할 수 있다. 여기서 메시지 공간이 작은 경우 q_1을 비밀키(복호화키)로 사용할 수 있지만, 메시지 공간이 큰 경우에는 q_1을 오직 검증키로만 사용할 수 있다.

2) 준 동형 성질(homomorphic) : 메시지 m_1과 m_2의 암호문 $g^{m_1}h^{r_1}$, $g^{m_2}h^{r_2}$이 주어졌을 때, 복호화하지 않고 m_1+m_2와 $m_1\cdot m_2$의 암호문을 $g^{m_1}h^{r_1}\cdot g^{m_2}h^{r_2}$와 $e(g^{m_1}h^{r_1}, g^{m_2}h^{r_2})$을 이용하여 구할 수 있다. 이 때, 덧셈은 여러 번 연산이 가능하지만 곱셈은 한 번의 연산만 가능하다.

2.4 암호화된 다항식에서의 연산

n차 다항식 $f(x)=a_n x^n+a_{n-1}x^{n-1}+\cdots+a_1x^1+a_0x^0$을 고려하자. 그러면 이것에 대한 암호문을 다음과 같이 각각의 계수들의 준 동형 암호화를 수행한 것으로 정의한다.

$$HE_{pk}(f(x))=\{\mathrm{HE_{pk}}(a_n),\mathrm{HE_{pk}}(a_{n-1}),\cdots,\\ HE_{pk}(a_1),HE_{pk}(a_0)\} \quad (4)$$

1) 점 b를 암호화된 다항식 $HE_{pk}(f(x))$에 대입한 값 $HE_{pk}(f(b))$를 구하는 방법

① 점 b에 대한 암호문을 다음과 같이 계산한다.

$$HE_{pk}(b)=\mathrm{HE_{pk}}(b^n),\mathrm{HE_{pk}}(b^{n-1}),\cdots,\mathrm{HE_{pk}}(b^0)\} \quad (5)$$

② 2.3절에서 살펴본 준 동형 성질을 사용하여 다음과 같이 계산한다.

$$HE_{pk}(f(b))=\{\mathrm{HE_{pk}}(a_n b^n),\\ HE_{pk}(a_{n-1}b^{n-1}),\cdots,HE_{pk}(a_0b^0)\} \quad (6)$$

2) 점 b, c를 암호화된 다항식 $HE_{pk}(f(x))$에 대입한 값 $HE_{pk}(f(b)+f(c))$를 구하는 방법

① 점 b, c에 대한 암호문을 다음과 같이 계산한다.

$$HE_{pk}(b+c)=\{\mathrm{HE_{pk}}(b^n+c^n),\\ HE_{pk}(b^{n-1}+c^{n-1}),\cdots,HE_{pk}(b^0+c^0)\} \quad (7)$$

② 2.3절에서 살펴본 준 동형 성질을 사용하여 다음과 같이 계산한다.

$$HE_{pk}(f(b)+f(c))=\{\mathrm{HE_{pk}}(a_n(b^n+c^n)),\\ HE_{pk}(a_{n-1}(b^{n-1}+c^{n-1})),\cdots,\\ HE_{pk}(a_0(b^0+c^0))\} \quad (8)$$

2.5 PIR 기법 (private information retrieval scheme)[11-15]

PIR 기법은 데이터 검색자가 자신의 어떠한 정보도 서버에게 드러내지 않고 원하는 데이터를 검색하는 방법이다. 이 기법에서 서버는 n개의 데이터를 가지고 있고, 데이터 검색자는 준 동형 시스템에 대한 키 쌍 (pk, sk)을 가지고 있다고 가정한다. 데이터 검색자는 자신이 검색하고자 하는 데이터의 인덱스 i를 다항식 암호화를 통해 서버에 전송하고, 서버는 데이터 검색자가 원하는 값이 i번째인지 모르면서도 i번째 데이터베이스에 저장되어 있는 값 $D(i)$를 암호화된 상태로 계산할 수 있다. 결과적으로 서버는 데이터 검색자에게 $HE_{pk}(D(i))$를 전송하게 되고, 데이터 검색자는 복호화를 통해 $D(i)$를 알게 된다. 예로 문헌 [11]에 나온 PIR 기법을 살펴보면 다음과 같다.

1) 데이터 검색자는 자신이 원하는 값의 인덱스를 서버에 드러내지 않기 위해 다음의 조건을 만족하는 $n-1$차 다항식 $p(x)$를 구성한다.

$$p(x)=\begin{cases}0, & x\neq i\\ 1, & x=i\end{cases}(0\le x<n) \quad (9)$$

2) 데이터 검색자는 다음과 같이 다항식 $p(x)$를 암호화하고, 이것을 서버에 전송한다.

$$HE_{pk}(p(x))=\{\mathrm{HE_{pk}}(p_{n-1}),\cdots,\mathrm{HE_{pk}}(p_1),\\ HE_{pk}(p_0)\} \quad (10)$$

3) 서버는 암호화된 다항식과 데이터베이스의 값들을 사용하여 다음을 계산한다. 여기서 $D(i)$는 i번째 데이터베이스에 저장되어 있는 값이다. 이 과정을 통해 데이터 검색자는 서버에게 자신이 원하는 데이터에 대한 정보를 드러내지 않고 값을 얻을 수 있다.

$$HE_{pk}(\sum_{i=0}^{n-1}p(i)D(i))=HE_{pk}(D(i)) \quad (11)$$

4) 서버는 $HE_{pk}(D(i))$를 데이터 검색자에게 되돌려주고, 데이터 검색자는 이 값을 자신이 가지고 있는 sk를 사용해서 복호화하여 $D(i)$의 값을 얻을 수 있다.

III. 관련연구 : Yau 등의 기법[10]

이 기법은 암호화된 DB에서 접근제어방식을 사용하기 위한 것으로 Stephen S. Yau 등에 의해 최초로 제안되었다. 이 기법은 암호화된 DB에 대해 허가된 사용자들만 접근이 가능하도록 데이터 공급자가 접근 권한을 지정할 수 있다는 장점을 가지고 있다. 기존의 대칭키 기반 또는 공개키 기반의 검색 가능한 암호 기법들은 특정 사용자만이 암호화된 데이터에 접근이 가능하였다. 따라서 기존 기법들에서 암호화된 데이터에 대해 여러 사용자가 접근 가능하게 하기 위해서는 같은 데이터를 여러 사용자를 위해 여러 번 암호화할 수밖에 없다. 이러한 번거로움을 피하기 위해 이 기법에서는 우선 데이터를 대칭키 k로 암호화 한 후

이 대칭키 k를 각 사용자들의 접근키(access key, ak)로 감추어 암호화된 데이터에 접근할 수 있는 사용자들만이 자신의 ak를 이용하여 k를 얻을 수 있게 설계하였다. Yau 등이 제안한 기법을 간략하게 알아보면 다음과 같다.

데이터 공급자는 대칭키 k를 사용하여 메시지 m을 암호화한다. 그리고 데이터 공급자는 접근키 ak_i와 키워드 kw_j를 선택하여 다항식 $f(x)$를 생성한다. 만약 테이블 각각의 행별로 ak 안에 ak_i와 kw 안에 kw_j가 포함되어 있다면, $f(ak_i)+f(kw_j)$ 값은 ak_i의 위치를 리턴한다. 데이터 공급자는 ak_i를 사용하여 k를 암호화하고, 결과적으로 데이터 공급자는 암호화된 메시지 $E_k(m)$, 접근키 ak, 키워드 kw, 암호화된 다항식 $HE(f(x))$, 암호화된 대칭키 $E_{ak}(k)$로 이루어진 테이블을 얻을 수 있다. 마지막으로 데이터 공급자는 $E_k(m)$, $E_{ak}(k)$, $HE(f(x))$를 서버에 저장한다. 이후, 검색을 원하는 데이터 검색자는 자신의 $ak_{i'}$과 $kw_{j'}$을 사용하여 $EP_{i'}(ak_{i'},kw_{j'})$를 계산하여 서버에 전송한다. 그러면 서버는 $HE(f(x))$와 $EP_{i'}(ak_{i'},kw_{j'})$를 매치하는 알고리즘을 사용하여 테이블 각각의 행별로 $E_k(m)$가 저장된 위치를 알아낼 수 있고, 그 위치에 저장된 $E_k(m)$의 값을 데이터 검색자에게 전송한다.

먼저, 이 기법에 사용되는 제어된 검색(controlled search)과 검증(verify) 알고리즘에 대해서 살펴보자.

3.1 제어된 검색 (Controlled Search)

$S_1=\{a_1,...,a_s\}$, $S_2=\{b_1,...,b_t\}$일 때 S_1, S_2에 대해서 다음의 조건을 만족하는 제어된 검색 다항식 $f(x)$를 정의한다. 이 때, $f(x)$는 $s+t$차 다항식이고 r은 랜덤 값이다.

$$f(x)=\begin{cases} r+i, & x=a_i\in S_1 \\ -r, & x\in S_2 \end{cases} \tag{12}$$

$a\in S_1$, $b\in S_2$일 때 $f(a)+f(b)$는 S_1의 원소 a의 인덱스를 나타내고, 그렇지 않으면 $f(a)+f(b)$는 랜덤 값이 된다.

이 기법에서는 제어된 검색을 위해 접근키 ak의 집합을 S_1, 키워드 kw의 집합을 S_2라 하고, 데이터 공급자는 S_1, S_2에 대한 제어된 검색 다항식 $f(x)$를 서버에 저장한다. 따라서 S_1의 접근키를 하나라도 알고 있는 검색자는 이 값을 이용하여 S_2에 있는 키워드를 검색할 수 있다. 제어된 검색 알고리즘은 다음과 같다.

> Controlled Search$(f(x),ak,kw)$
> 1. $f(ak)+f(kw)$를 계산한다.
> 2. $1\le f(ak)+f(kw)\le |S_1|$이면 $f(ak)+f(kw)$를 출력하고, 그렇지 않으면 0을 출력한다.

$ak\in S_1$이고 $kw\in S_2$이면, 이 알고리즘은 S_1에서 ak의 인덱스를 출력하고, 그렇지 않으면 0을 출력한다.

두 개의 임의의 원소 a와 b에 대해서, $1\le f(a)+f(b)\le s$의 확률은 $s(s+t)^2|R|/|D|^2$를 넘지 않는다. 여기서 s는 S_1의 크기이고, t는 S_2의 크기이며, D는 다항식 $f(x)$의 변역이고, R은 다항식 $f(x)$의 치역이다.

3.2 검증 알고리즘 (Verify)

신뢰할 수 없는 서버로부터 다항식을 보호하기 위해서는 $f(x)$를 암호화하여 저장하여야 하고, 또한 ak와 kw도 노출되어서는 안 된다. 하지만 서버는 제어된 검색을 하기 위해 ak와 kw, 그리고 $f(x)$에 대한 정보 없이 $f(ak)+f(kw)$를 계산할 수 있어야 한다. 따라서 신뢰할 수 없는 서버가 입력값에 대한 정보 없이 제어된 검색이 가능하게 하기 위해 다음과 같은 검증 알고리즘을 사용한다. 검증 알고리즘은 암호화된 다항식 $f(x)$와 암호화된 점 b, 그리고 상수 c가 주어졌을 때, $f(b)=c$이면 1을, 그렇지 않으면 0을 출력한다.

> Verify(EC,EP,c,pk,vk)
> 1. $C=\prod_{i=0}^{n} e(HE_{pk}(a_i),HE_{pk}(b^i))$를 계산한다.
> 2. 만약 $C^{vk}=e(g,g)^{vk\cdot c}$이면 1을 출력하고, 그렇지 않으면 0을 출력한다.

다항식 $f(x)=a_nx^n+a_{n-1}x^{n-1}+...+a_1x+a_0$ 일 때, EC는 $f(x)$의 각각의 계수를 준 동형 공개키 시스템(2.3절 참조)을 이용하여 암호화한 값으로 $EC=\{HE_{pk}(a_n),...,HE_{pk}(a_0)\}$이다. EP는 점 b를 암호화한 값으로 $EP=\{HE_{pk}(b^n),...,HE_{pk}(b^0)\}$이다. 이 때, pk와 $vk(=sk)$는 준 동형 공개키 시스템에서 공개키와 비밀키(=검증키)이고, $HE_{pk}(\cdot)$은 준 동형 공개키 시스템을 이용하여 암호화한 값을 나타낸다. 서버는 vk로부터 암호문을 복호화할 수 있지만 이는 메시지 공간이 작은 평문이 암호화되었을 때 가능하고, 메시지 공간이 큰 평문이 암호화되었다면 서버는 단지 검증만

할 수 있을 뿐 복호화는 할 수 없다.(2.3절에서 지적한 바와 같이 복호화에 걸리는 시간이 메시지 공간 T의 크기에 대해 다항식 시간만큼 걸리기 때문이다.)

3.3 기법 (main construction)

이 기법의 주요 구성은 다음의 다섯 가지 단계로 이루어져 있다.

3.3.1 초기화

서버는 공개 파라미터 $\{n,e,G,G_1\}$와 암호화키 $pk=\{g,h\}$, 그리고 검증키 $vk=q_1$를 설정한다. 그리고 서버는 안전한 대칭키 암호화 기법 E와, 해시함수 H도 설정한다.

3.3.2 공유데이터(m,ak,kw)

m을 공유하기 원하는 데이터 공급자는 $ak=\{ak_1,...,ak_s\}$와 $kw=\{k_1,...,k_t\}$에 대해서 제어된 검색 다항식 $f(x)=a_{s+t}x^{s+t}+...+a_1x+a_0$를 계산한 후, m과 $f(x)$의 계수를 암호화한다.

① 데이터 공급자는 다음과 같이 $f(x)$를 암호화한다.

$$EC=\{HE_{pk}(a_{s+t}),...,HE_{pk}(a_0)\} \quad (13)$$

② 데이터 공급자는 r_m을 임의로 선택하여 대칭키 $rk=H(g^{r_m})$을 계산한 후, 데이터 m에 대한 대칭키 암호문 $E_{rk}(m)$을 얻는다.

③ 데이터 공급자는 암호화키의 리스트 $C=\{g^{r_m/ak_1},...,g^{r_m/ak_s}\}$를 계산한다.

④ 데이터 공급자는 서버에 암호화된 데이터 $D=(E_{rk}(m),EC,C)$를 저장한다.

3.3.3 질의생성(ak,kw)

접근키 ak를 가지고 키워드 kw로 검색을 원하는 데이터 검색자는 암호화된 질의 $Q=\{HE_{pk}(ak^{s+t}+kw^{s+t}),...,HE_{pk}(ak^0+kw^0)\}$를 생성한다.

3.3.4 매치질의(D,Q)

서버는 데이터 검색자로부터 받은 암호화된 질의 Q와 암호화된 데이터 D를 검증 알고리즘을 이용하여 다음과 같이 매치(match)한다.

① 서버는 $D=(E_{rk}(m),EC,C)$를 스캔한다.

② $1\le i\le s$에 대해서, 서버는 검증 알고리즘 $\mathrm{Verify}(EC,Q,i,pk,vk)$을 실행한다.

③ 만약 검증 알고리즘 $\mathrm{Verify}(EC,Q,i,pk,vk)$이 어떤 i에 대해 1을 출력한다면, 데이터 검색자에게 $E_{rk}(m)$과 C의 i번째 값 g^{r_m/ak_i}을 전송한다.

3.3.5 검색데이터$(E_{rk}(m),g^{r_m/ak_i})$

데이터 검색자는 자신의 접근키 ak_i를 이용하여 대칭키 $rk=H((g^{r_m/ak_i})^{ak_i})$를 계산한 후, $E_{rk}(m)$를 복호화하여 m을 얻는다.

IV. 프라이버시를 보호하는 접근제어가 가능한 키워드 검색 기법

3장에서 살펴본 Yau 등의 기법은 검색자의 인덱스 위치가 서버에 노출되는 문제점을 가진다. 이 기법에서 서버는 매치질의에서 검증 알고리즘을 통해 검색자의 인덱스를 찾고, 이 인덱스를 이용하여 검색자가 대칭키를 구할 수 있도록 암호화키를 전송한다. 따라서 이번 장에서는 서버에 검색자의 프라이버시가 노출되는 문제를 해결한 프라이버시를 보호하는 접근제어가 가능한 키워드 검색 기법을 제안한다. 이 기법에서는 검색자의 인덱스 위치가 서버에 노출되는 문제를 PIR 기법을 사용하여 해결하였다. 우리가 제안한 기법을 간략하게 살펴보면 다음과 같다.

데이터 공급자는 대칭키 k를 사용하여 메시지 m을 암호화한다. 그리고 데이터 공급자는 접근키 ak_i와 키워드 kw_j를 선택하여 다항식 $f(x)$를 생성한다. 만약 테이블 각각의 행별로 ak 안에 ak_i와 kw 안에 kw_j가 포함되어 있다면, $f(ak_i)+f(kw_j)$ 값은 1을 리턴한다. 데이터 공급자는 ak_i를 사용하여 k를 암호화하고, 결과적으로 암호화된 메시지 $E_k(m)$, 접근키 ak, 키워드 kw, 암호화된 다항식 $HE(f(x))$, 암호화된 대칭키 $E_{ak}(k)$로 이루어진 테이블을 얻을 수 있다. 마지막으로 데이터 공급자는 $E_k(m)$, $E_{ak}(k)$, $HE(f(x))$를 서버에 저장한다. 이후, 검색을 원하는 데이터 검색자는 자신의 $ak_{i'}$과 $kw_{j'}$을 사용하여 $Q(ak_{i'},kw_{j'})$를 계산하고, PIR 기법에 사용되는 다항식 $p(x)$을 선택하여 암호화한 P를 Q와 함께 서버에 전송한다. 그러면 서버는 검증 알고리즘을 통해 확인하고, $p(x)$를 암호화한 것과 k를 사용

하여 Ip를 계산한다. 서버는 최종적으로 $E_k(m)$와 Ip를 데이터 검색자에게 되돌려준다.

먼저, 이 기법에 사용되는 제어된 검색과 검증 알고리즘에 대해서 살펴보자.

4.1 제어된 검색 (Controlled Search)

$S_1=\{a_1,...,a_s\}$, $S_2=\{b_1,...,b_t\}$일 때 S_1, S_2에 대한 제어된 검색 다항식 $f(x)$를 다음과 같이 정의한다. 이 때, $f(x)$는 $s+t$차 다항식이고 r은 랜덤 값이다.

$$f(x)=\begin{cases} r+1, & x=a_i\in S_1 \\ -r, & x\in S_2 \end{cases} \tag{14}$$

$a\in S_1$, $b\in S_2$일 때, $f(a)+f(b)$는 1이고, 그렇지 않으면 $f(a)+f(b)$은 랜덤 값이 된다.

제어된 검색을 위해 접근키 ak의 집합을 S_1, 키워드 kw의 집합을 S_2라고 하고, S_1, S_2에 대한 제어된 검색 다항식을 $f(x)$라고 할 경우에, 제어된 검색 알고리즘은 다음과 같다.

> Controlled Search($f(x)$,ak,kw)
> 1. $f(ak)+f(kw)$를 계산한다.
> 2. $f(ak)+f(kw)=1$이면 1을 출력하고, 그렇지 않으면 0을 출력한다.

두 개의 임의의 원소 a와 b에 대해서, $f(a)+f(b)=1$의 확률은 $(t+1)^2|R|/|D|^2$를 넘지 않는다. 여기서 t는 S_2의 크기이며, D는 다항식 $f(x)$의 번역이고, R은 다항식 $f(x)$의 치역이다.

4.2 검증 알고리즘 (Verify)

검증 알고리즘은 암호화된 다항식 $f(x)$와 암호화된 점 b, 그리고 1이 주어졌을 때, $f(b)=1$이면 1을, 그렇지 않으면 0을 출력한다.

> Verify(EC,EP,1,pk,vk)
> 1. $C=\prod_{i=0}^{n} e(HE_{pk}(a_i),HE_{pk}(b^i))$를 계산한다.
> 2. 만약 $C^{vk}=e(g,g)^{vk}$이면 1을 출력하고, 그렇지 않으면 0을 출력한다.

다항식 $f(x)=a_nx^n+a_{n-1}x^{n-1}+...+a_1x+a_0$ 일 때, $EC=\{HE_{pk}(a_n),...,HE_{pk}(a_0)\}$이고, 점 b를 암호화한 값 $EP=\{\mathrm{HE}_{pk}(b^n),...,HE_{pk}(b^0)\}$이다.

4.3 제안하는 기법

제안하는 기법은 다음의 다섯 단계로 구성된다.

4.3.1 초기화

서버는 공개 파라미터 $\{n,e,G,G_1\}$와 암호화키 $pk=\{g,h\}$, 그리고 검증키 $vk=g_1$를 설정한다. 그리고 서버는 안전한 대칭키 암호화 기법 E와, 해시함수 H도 설정한다.

4.3.2 공유데이터(m,ak,kw)

m을 공유하기 원하는 데이터 공급자는 $ak=\{ak_1,...,ak_s\}$와 $kw=\{k_1,...,k_t\}$에 대해서 제어된 검색 다항식 $f(x)=a_{s+t}x^{s+t}+...+a_1x+a_0$를 계산한 후, m과 $f(x)$의 계수를 암호화한다.

① 데이터 공급자는 다음과 같이 $f(x)$를 암호화한다.

$$EC=\{HE_{pk}(a_{s+t}),...,HE_{pk}(a_0)\} \tag{15}$$

② 데이터 공급자는 r_m을 임의로 선택하여 대칭키 $rk=H(g^{r_m})$을 계산한 후, 데이터 m에 대한 대칭키 암호문 $E_{r_k}(m)$을 얻는다.

③ 데이터 공급자는 암호화키의 리스트 $C=\{g^{r_m/ak_1},...,g^{r_m/ak_s}\}$를 계산한다.

④ 데이터 공급자는 서버에 암호화된 데이터 $D=(E_{r_k}(m),EC,C)$를 저장한다.

4.3.3 질의생성(ak,kw,i)

접근키 ak를 가지고 키워드 kw로 검색을 원하는 데이터 검색자는 암호화된 값 $Q=\{HE_{pk}(ak^{s+t}+kw^{s+t}),...,HE_{pk}(ak^0+kw^0)\}$를 생성한다. 그리고 데이터 검색자는 다음과 같은 조건을 만족하는 $s-1$차 다항식 $p(x)$를 선택한다.

$$p(x)=\begin{cases} 1, & x=i \\ 0, & otherwise \end{cases} \tag{16}$$

여기서 i는 데이터 검색자 자신의 인덱스 값이다. 즉, $p(x)=p_{s-1}x^{s-1}+...+p_0$는 자신의 인덱스 값 i에

대해서는 1을 출력하고, 그 이외에는 0을 출력하는 다항식이다. 데이터 검색자는 $p(x)$의 계수들을 자신의 공개키 pk_i로 암호화하여 $P=\{HE_{pk_i}(p_{s-1}),...,HE_{pk_i}(p_0)\}$를 얻는다. 데이터 검색자가 생성한 암호화된 질의는 Q와 P이다.

4.3.4 매치질의(D,Q,P)

서버는 데이터 검색자로부터 받은 암호화된 질의 Q와 암호화된 데이터 D를 검증 알고리즘을 이용하여 다음과 같이 매치(match)한다.

① 서버는 검증 알고리즘 Verify($EC,Q,1,pk,vk$)을 실행한다.

② 만약 검증 알고리즘이 1을 출력하면 다음을 계산하고, 그렇지 않으면 종료한다.

$$Ip=HE_{pk_i}(\sum_{i=1}^{s} g^{r_m/ak_i} \cdot p(i)) \quad (17)$$

③ 서버는 $E_{rk}(m)$과 Ip를 데이터 검색자에게 전송한다.

서버는 암호화키의 리스트 $C=g^{r_m/ak_1},...,g^{r_m/ak_s}$와 $P=\{HE_{pk_i}(p_{s-1}),...,HE_{pk_i}(p_0)\}$로부터 2.3절의 준 동형 성질을 이용하여 $Ip=HE_{pk_i}(\sum_{i=1}^{s} g^{r_m/ak_i} \cdot p(i))$를 계산할 수 있다.

본 단계에서 검색자는 서버에게 자신의 프라이버시를 보호받기 위해 PIR 기법을 사용한다. 제안하는 기법에서는 기법의 통일성을 위해 준 동형 공개키 시스템[11]을 이용한 PIR 기법을 사용하였으나 기존에 제안된 효율적인 PIR 기법[12-15]을 사용하여 효율성을 높일 수 있다.

4.3.5 검색데이터($E_{rk}(m),Ip$)

데이터 검색자는 Ip를 복호화하여 자신의 숨겨진 암호화키 g^{r_m/ak_i}를 찾을 수 있고, 자신의 접근키 ak_i를 이용하여 대칭키 $rk=H((g^{r_m/ak_i})^{ak_i})$를 계산한 후 $E_{rk}(m)$를 복호화하여 m을 얻는다.

4.4 분석

기존에 제안된 Yau 등의 기법에서는 데이터 공급자가 사용자들의 검색 능력을 제어할 수 있는 장점을 가지고 있는 반면, 신뢰할 수 없는 서버에게 데이터 검색자에 대한 키들의 위치가 드러나는 문제점을 가지고 있다. 예를 들면, k번째 사용자가 서버에 자신의 접근키 ak_k와 자신이 선택한 키워드 $kw_{k'}$를 사용하여 암호화된 질의 Q_k를 서버에 보냈다고 하자. 이 경우 서버는 ak_k와 $kw_{k'}$가 포함되어 있는 문서들의 리스트에서 키들의 위치를 파악할 수 있다. 그리고 그 이후에 k번째 사용자가 서버에 자신의 접근키 ak_k와 자신이 선택한 예전과는 다른 키워드 $kw_{k''}$를 사용하여 암호화된 질의 Q_k'를 서버에 보낸 경우, 서버는 이미 접근키 ak_k에 해당하는 문서들의 리스트를 일정량 알고 있기 때문에, 이러한 것들을 토대로 서버는 암호화된 질의 Q_k와 Q_k'를 생성한 사용자가 같은지 다른지를 판단할 수 있는 정보를 얻게 된다.

이러한 문제점을 개선하기 위해, 우리는 PIR 기법[11-15]을 사용하여 사용자의 인덱스를 서버에 유출시키지 않고 원하는 데이터의 검색이 가능하게 하였다. Yau 등의 기법에서는 매치질의 단계에서 검증 알고리즘이 어떤 i에 대해서 1을 출력하면 데이터 검색자에게 암호화키의 리스트 C에서 i번째 키를 전송하는 방법을 사용함으로써 검색자의 인덱스 정보를 서버에게 노출시켰다. 제안하는 기법에서는 암호화키의 리스트 C에서 i번째 키를 검색자에게 전송할 때 PIR 기법을 사용하기 때문에 서버는 검색자의 인덱스 정보를 알 수 없다. 제안하는 기법이 문헌 [11]의 PIR 기법을 사용하였을 경우, 만약 g^{r_m/ak_i}가 b 비트라고 하면 Ip로부터 g^{r_m/ak_i}를 베이비-스텝 자이언트-스텝 알고리즘과 같은 이산 대수 계산 방법에 의해 $O(2^{b/2})$ 시간 안에 찾을 수 있다.

V. 결 론

본 논문에서는 기존 연구에서 신뢰하지 않는 서버에 검색자의 프라이버시가 노출되는 문제를 해결한 프라이버시를 보호하는 접근제어가 가능한 키워드 검색 기법을 제안하였다. 제안된 기법을 통해 데이터 공급자와 데이터 검색자는 신뢰하지 않는 공개 서버에 데이터를 저장하고, 프라이버시를 노출시키지 않으면서 데이터를 안전하게 공유할 수 있다.

참 고 문 헌

[1] 김선영, 서재우, 이필중, "검색 가능 암호 기술의 연구 동향," 정보보호학회지, 19(2), pp. 63-73,

2009년 4월.

[2] D.X. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," IEEE Computer Society, IEEE Symposium on Security and Privacy, pp. 44-55, May 2000.

[3] E.J. Goh, "Secure Indexes," Technical report 2003/216, In IACR ePrint Cryptography Archive, Oct. 2003.

[4] Y.C. Chang and M. Mitzenmacher, "Privacy Preserving Keyword Searches on Remote Encrypted Data," ACNS, LNCS 3531, pp. 442-455, 2005.

[5] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," ACM Conference on Computer and Communications Security, pp. 79-88, Oct. 2006.

[6] D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public Key Encryption with Keyword Search," EUROCRYPT, LNCS 3027, pp. 506-522, 2004.

[7] P. Golle, J. Staddon, and B.R. Waters, "Secure Conjunctive Keyword Search over Encrypted Data," ACNS, LNCS 3089, pp. 31-45, 2004.

[8] D. Boneh and B. Waters, "Conjunctive, Subset, and Range Queries on Encrypted Data," TCC, LNCS 4392, pp. 535-554, 2007.

[9] J. Katz, A. Sahai, and B. Waters, "Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products," EUROCRYPT, LNCS 4965, pp. 146-162, 2008.

[10] S.S. Yau and Y. Yin, "Controlled privacy preserving keyword search," AISACCS, pp. 321-324, Mar. 2008.

[11] D. Boneh, E.J. Goh, and K. Nissim, "Evaluating 2-DNF Formulas on Ciphertexts," TCC, LNCS 3378, pp. 325-341, 2005.

[12] E. Kushilevitz and R. Ostrovsky, "Replication is NOT Needed: SINGLE Database, Computationally-Private Information Retrieval," FOCS, pp. 364-373, Oct. 1997.

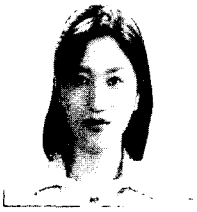
[13] C. Cachin, S. Micali, and M. Stadler, "Computationally Private Information Retrieval with Polylogarithmic Communication," EUROCRYPT, pp. 402-414, May 1999.

[14] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch Codes and Their Applications," STOC, pp. 373-382, June 2004.

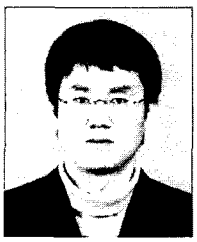
[15] C. Gentry and Z. Ramzan, "Single-Database Private Information Retrieval with Constant Communication Rate," ICALP, pp. 803-815, July 2005.

〈著者紹介〉

노 건 태 (Geon Tae Noh) 학생회원
2008년 2월: 고려대학교 산업시스템정보공학과 학사 졸업
2008년 3월 ~ 현재: 고려대학교 정보경영공학과 석사과정
〈관심분야〉 암호 이론, 프라이버시향상기술(PET), 유비쿼터스 보안



천 지 영 (Ji Young Chun) 학생회원
1997년 2월: 이화여자대학교 수학과 학사 졸업
2006년 2월: 고려대학교 정보경영공학과 석사 졸업
2006년 3월 ~ 현재: 고려대학교 정보경영공학과 박사과정
〈관심분야〉 암호 이론, 프라이버시향상기술(PET), 유비쿼터스 보안



정 익 래 (Ik Rae Jeong) 정회원
1998년 2월: 고려대학교 전산학과 학사 졸업
2000년 2월: 고려대학교 전산학과 석사 졸업
2004년 8월: 고려대학교 정보보호대학원 박사 졸업
2006년 6월 ~ 2008년 2월: 한국전자통신연구원 암호기술연구팀 선임연구원
2008년 3월 ~ 현재: 고려대학교 정보경영공학부 조교수
〈관심분야〉 프라이버시향상기술(PET), 데이터베이스 암호, 암호 이론

이 동 훈 (Dong Hoon Lee) 종신회원
1983년 8월: 고려대학교 경제학과 학사 졸업
1987년 12월: Oklahoma University 전산학과 석사 졸업
1992년 5월: Oklahoma University 전산학과 박사 졸업
1993년 3월 ~ 1997년 2월: 고려대학교 전산학과 조교수
1997년 3월 ~ 2001년 2월: 고려대학교 전산학과 부교수
2001년 3월 ~ 현재: 고려대학교 정보경영공학부 교수
〈관심분야〉 암호 프로토콜, 암호 이론, USN 이론, 키 교환, 익명성 연구, PET 기술