ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 44, No. 1, pp. 21-26, 2017. 1
https://doi.org/10.5626/JOK.2017.44.1.21

결정트리 기반의 기계학습을 이용한 동적 데이터에 대한 재익명화기법

(Re-anonymization Technique for Dynamic Data Using Decision Tree Based Machine Learning)

김 영 기 †　　홍 충 선 ††
(Young Ki Kim)　　(Choong Seon Hong)

요 약 사물인터넷, 클라우드 컴퓨팅, 빅데이터 등 새로운 기술의 도입으로 처리하는 데이터의 종류와 양이 증가하면서, 개인의 민감한 정보가 유출되는 것에 대한 보안이슈가 더욱 중요시되고 있다. 민감정보를 보호하기 위한 방법으로 데이터에 포함된 개인정보를 공개 또는 배포하기 전에 일부를 삭제하거나 알아볼 수 없는 형태로 변환하는 익명화기법을 사용한다. 그러나 준식별자의 일반화 수준을 계층화하여 익명화를 수행하는 기존의 방법은 데이터 테이블의 레코드가 추가 또는 삭제되어 k-익명성을 만족하지 못하는 경우에 더 높은 일반화 수준을 필요로 한다. 이와 같은 과정으로 인한 정보의 손실이 불가피하며 이는 데이터의 유용성을 저해하는 요소이다. 따라서 본 논문에서는 결정트리 기반의 기계학습을 적용하여 기존의 익명화방법의 정보손실을 최소화하여 데이터의 유용성을 향상시키는 익명화기법을 제안한다.

키워드: 민감정보, 익명화, k-익명성, 결정트리, 기계학습

Abstract In recent years, new technologies such as Internet of Things, Cloud Computing and Big Data are being widely used. And the type and amount of data is dramatically increasing. This makes security an important issue. In terms of leakage of sensitive personal information. In order to protect confidential information, a method called anonymization is used to remove personal identification elements or to substitute the data to some symbols before distributing and sharing the data. However, the existing method performs anonymization by generalizing the level of quasi-identifier hierarchical. It requires a higher level of generalization in case where k-anonymity is not satisfied since records in data table are either added or removed. Loss of information is inevitable from the process, which is one of the factors hindering the utility of data. In this paper, we propose a novel anonymization technique using decision tree based machine learning to improve the utility of data by minimizing the loss of information.

Keywords: sensitive information, anonymization, k-anonymity, decision tree, machine learning

· 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(R0126-16-1009, ICBMS 플랫폼 간 정보 모델 연동 및 서비스 매쉬업을 위한 스마트 중재 기술 개발)

† 학생회원 : 경희대학교 컴퓨터공학과
qoo0144@khu.ac.kr
†† 종신회원 : 경희대학교 컴퓨터공학과 교수
(Kyung Hee Univ.)
cshong@khu.ac.kr
(Corresponding author임)

논문접수 : 2016년 8월 18일
(Received 18 August 2016)
논문수정 : 2016년 10월 17일
(Revised 17 October 2016)
심사완료 : 2016년 10월 26일
(Accepted 26 October 2016)

정보과학회논문지 제44권 제1호(2017. 1)

1. 서 론

최근에는 사물인터넷, 빅데이터 등 새로운 기술의 도입으로 처리하는 데이터의 종류와 양이 점차 증가하면서, 개인의 민감한 정보가 유출되는 것에 대한 보안이슈가 더욱 중요시되고 있다. 특히, 기업에서는 광고를 목적으로 방대한 양의 개인정보를 수집하고 있다. 또한 학교, 정부 등의 다양한 기관에서 수집된 개인정보는 민감한 정보를 포함하고 있기 때문에, 유출될 경우 각종 범죄에 악용될 수 있는 여지가 많다. 실제로 금융정보와 같은 개인정보를 노린 각종 보안 위협이 꾸준히 늘어나는 추세이다.

따라서 개인의 민감한 정보를 보호하기 위한 방법으로 데이터에 포함된 개인정보를 공개하거나 배포하기 전에 일부를 삭제하거나 알아볼 수 없는 형태로 변환하는 익명화기법을 사용한다.

한편, 기존의 익명화 방법에서는 데이터의 준식별자에 해당되는 값을 알아볼 수 없는 형태로 변환하는 일반화의 수준(level of generalization)을 계층화(hierarchy)하는 방법으로 익명화를 수행한다. 그러나 기존의 익명화 방법에서 사용자가 요구하는 K-익명성을 만족하지 않도록 레코드가 추가/삭제된 경우 더 높은 준식별자의 수준으로 익명화를 수행해야 한다. 이 과정에서 불필요한 정보의 손실을 피할 수 없으며 이는 데이터의 유용성을 저해하는 요소이다. 따라서 본 논문에서는 익명화 기법에 결정트리기반의 기계학습을 적용하여 정보의 손실을 최소화하여 익명화를 수행할 수 있는 방안을 제안한다.

2장에서는 데이터 테이블의 구조와 익명화, K-익명성, 기계학습의 관련 연구를 분석하고, 3장에서는 기존 익명화 방법의 문제점과 제안하는 익명화 기법에 대해 기술한다. 4장에서는 실험 결과 및 성능평가를 분석하고, 마지막으로 5장에서는 본 논문의 결론에 대하여 언급한다.

2. 관련 연구

2.1 데이터 테이블의 구조

일반적으로 수집된 데이터 테이블은 그림 1[1]과 같이 식별자(identifier), 준식별자(quasi-identifier), 민감정보(sensitive attribute)로 구성된다. 이름이나 주민등록번호와 같이 특정 데이터를 다른 데이터와 구분할 수 있는 속성을 식별자라 하며 나이, 성별 등 직접적으로 대상을 알 수는 없지만 해당 데이터의 특징을 나타내는 속성을 준식별자라 한다. 또한 급여, 질병과 같이 개인의 민감한 속성을 포함한 데이터를 민감정보라 한다.

Identifier Quasi-Identifier Sensitive Attribute

Name	Age	Gender	Zip code	Disease
Amy	27	F	482010	AIDS
Bobby	22	M	482750	cancer
Carol	37	F	420760	cold
David	39	F	420880	diabetes

그림 1 데이터 테이블의 예
Fig. 1 An Example of Data Table

표 1 데이터 익명화의 예
Table 1 Example of Data Anonymization

Age	Gender	Zip code	Disease
20-29	*	482***	AIDS
20-29	*	482***	cancer
30-39	*	420760	cold
30-39	*	420880	diabetes

표 1은 익명화된 데이터 테이블의 일례이다. 익명화는 일반적으로 데이터 테이블에서 식별자를 삭제하고 준식별자에 해당되는 데이터를 치환함으로써 프라이버시 보호를 수행한다.

2.2 k-익명성(k-anonymity)

K-익명성은 수집된 데이터 테이블에 익명화를 수행하여 적어도 k개 이상의 준식별자 속성값들이 동일한 값을 갖도록 하는 것이다. Latanya Sweeney는 [2]에서 K-익명성을 다음과 같이 정의했다.

정의. K-익명성 K-anonymity)

$RT(A_1, \cdots, A_n)$는 데이터 테이블이고 QI_{RT}는 데이터 테이블 RT의 준식별자(quasi-identifier)일 때, RT가 K-익명성을 만족한다는 뜻은 $RT[QI_{RT}]$의 모든 기록들이 적어도 K번 $RT[QI_{RT}]$에서 나타난다는 의미이다.

표 2는 k=2를 만족하도록 데이터 테이블을 익명화한 예이다.

표 2 k-익명성의 예(k=2)
Table 2 Example of k-anonymization(k=2)

Age	Gender	Zip code	Disease
20-29	*	482***	AIDS
20-29	*	482***	cancer
30-39	F	420880	cold
30-39	F	420880	cold

2.3 기계학습 (Machine Learning)

본 논문에서는 개인정보를 보호하기 위한 익명화 기법으로 준식별자 값의 레코드가 정보의 손실을 최소화하는 단계로 일반화하는 것을 기준으로 결정트리를 구성하여 기계학습에 적용한다. 그림 2는 표 2의 데이터

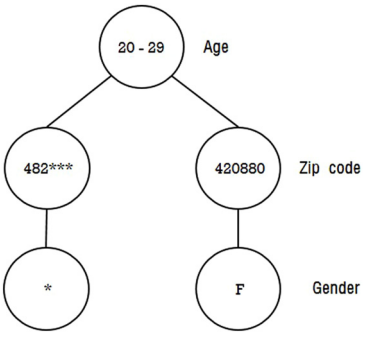

그림 2 익명화된 테이블로 구성한 결정트리
Fig. 2 Decision Tree Consisting of Anonymized Table

테이블을 활용하여 기계학습에 필요한 결정트리를 구성하는 예이다.

각 노드는 익명화된 준식별자의 레코드값을 바탕으로 구성되며 같은 트리레벨에는 하나의 준식별자에 대한 데이터들이 나열된다. 이와 같이 구성된 결정트리의 리프노드(leaf node)의 수는 익명화된 데이터 테이블의 동질집합의 수와 같다.

2.4 결정트리 학습법(Decision Tree Learning)

결정트리 학습법은 기계학습의 한 종류로, 어떤 항목에 대한 관측 값과 목표 값을 연결시켜주는 예측 모델로써 결정트리를 사용한다. 이는 통계학과 데이터 마이닝, 기계학습에서 사용하는 예측 모델링 방법 중 하나이다.

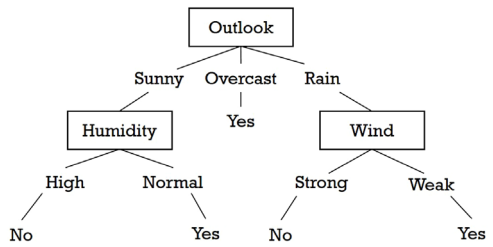

그림 3 결정트리 학습법의 예
Fig. 3 An Example of Decision Tree Learning

그림 3은 결정트리 학습법의 예이다. 학습을 위한 기존의 데이터로 날씨, 습도, 바람의 세기를 분류기준으로 하여 실외활동을 했는지 여부를 판단한다. 이처럼 결정트리 학습법은 기존의 데이터를 바탕으로 분류체계를 구성하고, 새로운 데이터가 추가되었을 때 그 결과를 쉽게 예측할 수 있는 모델이다.

3. 기존 연구의 문제점 및 제안사항

[3]에서는 동질집합을 구성하는 과정에서 군집화를 적용하여 손실되는 정보를 최소화 하는 익명화 기법을 제안하였다. 또한, [4]에서는 유전자 알고리즘 기반의 군집화를 적용한 익명화기법을 제안하였다. 제안된 익명화기법은 군집화 과정에 휴리스틱(heuristic)을 적용하여 발생하는 정보의 손실을 최소화하는 것을 목적으로 한다.

이와 같이 기존의 k-익명성을 활용하여 정보의 손실을 최소화하는 익명화기법에 대한 연구는 활발하게 진행되고 있지만, 새롭게 추가/삭제되는 데이터의 익명화에 대한 연구는 부족한 현실이다. 또한, 앞서 언급한 익명화기법들은 데이터가 추가/삭제되어 k-익명성을 만족하지 못하는 경우 모든 레코드를 도메인으로 다시 익명화를 수행해야 하는 문제점이 존재한다.

따라서 본 논문에서는 결정트리를 적용하여 새롭게 추가/삭제되는 동적 데이터에 대한 효율적인 익명화기법을 제안한다.

준식별자의 일반화 수준을 계층화하여 익명화를 수행하는 기존의 방법은 데이터 테이블의 레코드가 추가 또는 삭제되어 더 이상 사용자가 요구하는 k-익명성을 만족하지 못하는 경우에, 더 높은 일반화 수준으로 데이터를 익명화해야 한다. 이와 같은 과정에서 발생되는 정보의 손실은 불가피하며 이는 데이터의 유용성을 저해하는 주요 요인이다[5]. 그림 4와 그림 5는 데이터 테이블의 레코드가 추가/삭제됨에 따라 발생되는 정보손실의 예이다.

그림 4와 같이 데이터 테이블에 추가된 레코드를 k=2를 만족하도록 익명화하기 위해 두 번째 동질집합의 레코드들이 더 높은 수준의 일반화가 적용되었다. 이 과정에서 우편번호 데이터가 추가적으로 손실되었다.

마찬가지로 그림 5에서는 데이터 테이블의 레코드가 삭제되면서 수행되는 재익명화 과정에서 다른 레코드들의 정보가 손실되었다.

위와 같이 재익명화를 수행하는 과정에서 정보의 손실은 불가피하다[6]. 본 논문에서는 위와 같은 과정에서 발생하는 불가피한 정보의 손실을 최소화하기 위한 방법으로 데이터를 익명화하는 과정에 결정트리기반의 기계학습을 적용한다.

익명화된 데이터를 학습 집단(training set)으로 결정트리를 구성하고 이를 통해 사용자가 요구하는 k-익명성을 만족시키지 못하는 경우 일반화의 수준을 기준으로 정보의 손실을 최소화 하도록 한다. 제안사항은 최적의 성능을 보장하기 위해 다음 두 가지의 제약사항을 만족해야 한다.

조건 1. 결정트리를 구성할 때는 일반화의 계층이 많이 존재 하는 준식별자를 우선적으로 분류한다.

조건 2. 데이터의 추가 또는 삭제로 k-익명성을 만족할 수 없는 경우 결정트리에서 넓이우선탐색을 수행하여 정보의 손실이 최소가 되는 노드를 찾는다.

Age	Gender	Zip code	Disease
20-29	*	482***	AIDS
20-29	*	482***	cancer
20-29	F	420880	cold
20-29	F	420880	cold
23	F	418000	diabete

Age	Gender	Zip code	Disease
20-29	*	482***	AIDS
20-29	*	482***	cancer
20-29	F	*	cold
20-29	F	*	cold
20-29	F	*	diabete

그림 4 데이터 추가에 따른 정보손실

Fig. 4 Loss of Information According to Additional Data

Age	Gender	Zip code	Disease
20-29	*	482***	AIDS
20-29	*	482***	cancer
20-29	F	420880	cold
20-29	F	420880	cold

Age	Gender	Zip code	Disease
20-29	*	*	cancer
20-29	*	*	cold
20-29	*	*	cold

그림 5 데이터 삭제에 따른 정보손실

Fig. 5 Loss of Information According to Data Removal

조건 1과 2를 이용하여 구성된 결정트리를 바탕으로 재익명화가 수행된다고 가정했을 때, 루트노드에 위치할수록 많은 정보손실이 발생한 준식별자 값을 고려할 수 있다. 또한 변경된 데이터가 결정트리의 어느 노드에 위치하는지 넓이우선탐색을 수행하여 정보손실을 최소화하는 노드를 찾는다.

그림 6은 결정트리를 이용하여 재익명화가 수행되는 과정을 나타낸다. 익명화된 데이터와 일반화의 수준을 계층화한 정보를 바탕으로 결정트리를 구성하고, 추가/삭제되는 데이터의 해당되는 준식별자 값이 조건을 만족하면 다음 자식노드들을 확인하여 정보손실이 최소화되도록 트리노드와 데이터 테이블을 변경하며 익명화를 수행한다.

그림 7은 익명화를 수행하기 위해 사용되는 일반화의 계층을 바탕으로 구성한 분류트리이다. 표 1에서 존재하는 우편번호 데이터를 바탕으로 일반화의 수준이 낮은 데이터를 리프노드로 하고, 모든 값이 특정 문자로 치환되어 일반화의 수준이 높은 데이터를 루트노드로 한다.

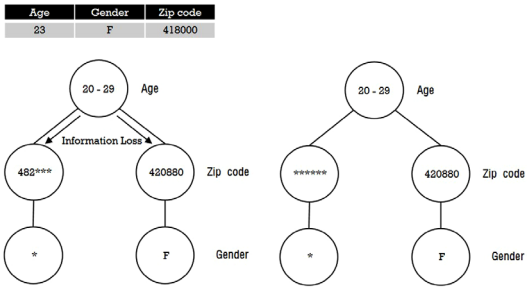

그림 6 결정트리를 이용한 재익명화

Fig. 6 Re-anonymization Using Decision Tree

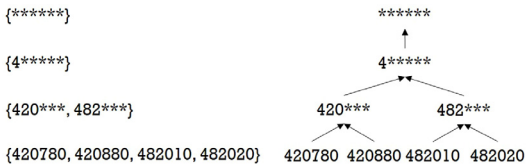

그림 7 일반화의 계층으로 구성한 분류트리

Fig. 7 Classification Tree Based on Generalization Hierarchy

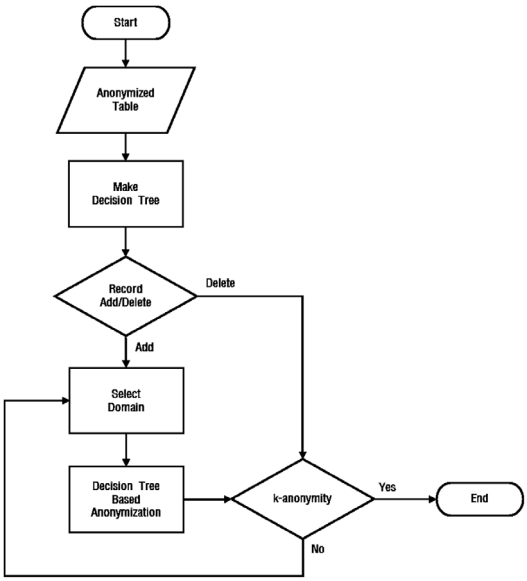

그림 8 제안하는 익명화 기법의 순서도

Fig. 8 Flowchart of Proposed Anonymization Scheme

정보의 손실률은 익명화 하고자 하는 두 데이터의 최소공통조상을 찾아 서브트리의 루트노드로 설정하고, 전체트리의 높이를 서브트리의 높이로 나눈 값으로 측정한다[7].

그림 8은 본 논문에서 제안하는 익명화 기법의 순서도이다. 먼저 익명화된 테이블을 바탕으로 결정트리를

구성하고 레코드의 추가 또는 삭제가 발생하면 주어진 과정에 따라 익명화를 수행한다. 새로운 레코드가 추가된 경우에는 테이블에서 재익명화를 수행할 대상을 선정하고 결정트리를 기반으로 익명화를 수행한다. 또한 익명화된 테이블이 사용자가 요구하는 k-익명성을 만족하는지 여부를 확인한다. 테이블에서 레코드가 삭제된 경우에는 먼저 k-익명성을 만족하는지 확인하고, 그렇지 않은 경우에는 레코드가 추가될 때와 마찬가지로 테이블에서 대상을 선정하고 익명화를 수행한다.

4. 성능 평가

본 논문에서 수행한 실험을 위한 환경은 다음과 같다.

크기가 정해진 익명화된 데이터 테이블에 50개의 레코드가 추가 또는 삭제됨에 따른 기존의 익명화 기법과 제안하는 기법의 정보의 손실률을 비교하여 측정한다. 측정을 위한 계산방법은 앞서 그림 7을 설명하면서 언급한 바와 같다.

본 논문에서는 연구를 수행하기 위한 익명화 도구로 ARX[8] 라이브러리를 사용하였으며, Java-ML[9] 라이브러리를 활용하여 결정트리를 구성하고 익명화모듈에 기계학습을 적용한다.

그림 9는 익명화한 데이터 테이블에 표 3의 환경에서 언급한 시나리오를 바탕으로 결과를 분석한 결과이다. 평가기준은 준식별자가 최대로 일반화된 것을 정보의 손실이 가장 많이 발생한 것으로 하며, 일반화 단계의 수만큼 가중치를 달리하여 모든 준식별자의 정보손실을 계산하고 원본 데이터와 비교한다.

데이터의 추가/삭제가 발생한 경우에 익명화에 필요한 k값을 달리하여 비교해 보았을 때, 기존의 익명화기법보다 발생하는 정보의 손실이 더 적은 것을 확인할 수 있었다.

그림 10은 기존의 익명화기법과 제안하는 기법이 재익명화에 소요되는 시간을 측정한 결과이다. 결과에서 확인할 수 있듯이 제안하는 익명화기법에서는 정보의 손실을 최소화하기 위해 추가적인 연산을 필요로 하므로,

표 3 실험 환경

Table 3 Experimental Environment

Table Size	100 × 6
Total Record	99
Num. of Added Records	10 - 50
Num. of Deleted Records	10 - 50
Total Attribute	6
Num. of Identifier	1
Num. of Q-Identifier	4
Num. of Sentitive Attribute	1
k-anonymity	k = 2

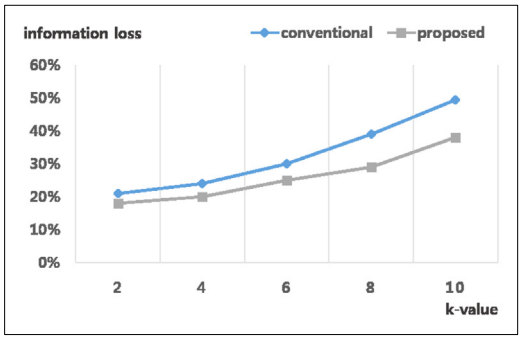

그림 9 데이터 추가/삭제에 따른 정보손실

Fig. 9 Information Loss Ratio According to Data Addition and Removal

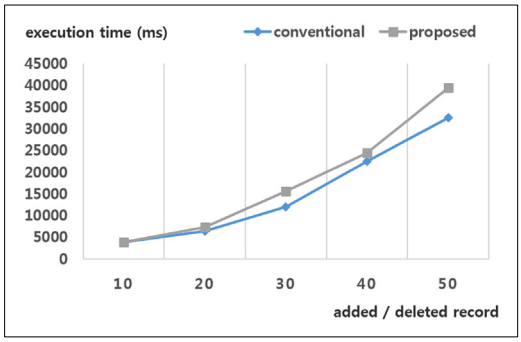

그림 10 재익명화에 소요되는 실행시간

Fig. 10 Execution Time of Re-anonymization

기존의 기법보다 실행시간이 더 많이 걸린다. 추가/삭제되는 데이터가 증가할수록 결정트리를 구성하고 비교하는 연산의 횟수가 많아지므로 실행시간의 차이는 더욱 커질 것으로 예상된다.

5. 결 론

본 논문은 데이터 테이블의 레코드가 추가 또는 삭제되어 더 이상 사용자가 요구하는 k-익명성을 만족하지 못하는 경우에, 더 높은 일반화 수준으로 데이터를 익명화하는 기존 방법의 문제점을 해결하기 위한 방법으로 결정트리기반의 기계학습을 적용했다. 4장에서 성능평가를 통해 추가/삭제되는 데이터의 양이 증가함에 따른 정보손실을 분석하여 제안하는 익명화 기법의 성능을 검증했다. 그러나 본 논문에서 제안하는 익명화기법은 정보의 손실측면에서는 기존 방법보다 우수하지만, 결정트리를 구성하고 비교를 위한 추가적인 연산을 필요로 하므로 실행시간이 더 오래 걸린다. 따라서 향후에는 연산과정을 최적화하여 재익명화에 필요한 실행시간을 개선하는 연구를 수행할 것으로 예상된다.

References

[1] Chikwang Hwang, Jongwon Choe, Choong Seon Hong, "A Study on Service-based Secure Anonymization for Data Utility Enhancement," *Journal of KIISE*, Vol. 42, No. 5, pp. 681-689, May. 2015. (in Korean)
[2] L. Sweeny, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp. 557-570, 2012.
[3] Byun, Ji-Won, et al., "Efficient k-anonymization using clustering techniques," *International Conference on Database Systems for Advanced Applications*, pp. 188-200, Apr. 2007.
[4] Lin, Jun-Lin, and Meng-Cheng Wei, "Genetic algorithm-based clustering approach for k-anonymization," *Journal of Expert Systems with Applications*, Vol. 36, No. 6, pp. 9784-9792, Dec. 2009.
[5] Li, Tiancheng, and Ninghui Li, "On the tradeoff between privacy and utility in data publishing," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 517-526, Jun. 2009.
[6] H. Kim, "Privacy Preserving for Statistical Anonymity," *NIA Privacy Issues*, No. 2, Jun. 2012. (in Korean)
[7] Xiaoshuang Xu, Masayuki Numao, "An Efficient Clustering Method for k-Anonymization," *International Symposium on Computing and Networking*, pp. 499-502, Dec. 2015.
[8] ARX, [Online], Available: http://arx.deidentifier.org/
[9] Java-ML, [Online], Available: http://java-ml.sourceforge.net/

김 영 기
2016년 경희대학교 컴퓨터공학과(공학사)
2016년 3월부터 현재까지 경희대학교 컴퓨터공학과 석사과정. 관심분야는 네트워크 보안, 프라이버시 보호

홍 충 선
1983년 경희대학교 전자공학과(공학사)
1985년 경희대학교 전자공학과(공학석사)
1997년 Keio University, Department of Information and Computer Science (공학박사). 1988년~1999년 한국통신통신망연구소 수석연구원/네트워킹 연구실장. 1999년~현재 경희대학교 컴퓨터공학과 교수. 관심분야는 인터넷 서비스 및 망 관리구조, 미래인터넷, IP mobility, Sensor Networks, Network Security