JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 5, pp. 17-22, May 2021
https://doi.org/10.9708/jksci.2021.26.05.017

A Study on the Processing Method of pseudonym information considering the scope of data usage

Youn-A Min*

*Professor, Dept. of Applied Software Engineering, Hanyang Cyber University, Seoul, Korea

[Abstract]

With the application of the Data 3 method, the scope of the use of pseudonym information has expanded. In the case of pseudonym information, a specific individual can be identified by linking and combining with various data, and personal information may be leaked due to incorrect use of the pseudonym information. In this paper, we propose the scope of use of data is subdivided and a differentiated pseudonym information processing method according to the scope. For the study, the formula was modified by using zero-knowledge proof among the pseudonym information processing methods, and when the proposed formula was applied, it was confirmed that the performance improved by an average of 10% in terms of verification time compared to the case of applying the formula of the existing zero-knowledge proof.

▸ **Key words**: data de-identification, pseudonym information, zero knowledge proof

[요 약]

데이터3법의 적용과 더불어 가명정보 활용의 범위가 넓어졌다. 가명정보의 경우 다양한 데이터와의 연계 및 결합에 의하여 특정개인을 식별할 수 있으며 가명정보의 잘못된 활용으로 인하여 개인정보가 유출될 수 있다. 본 논문에서는 데이터의 사용범위를 세분화하고 해당범위에 따른 차별화된 가명정보 처리방법을 제안하였다. 연구를 위하여 가명정보 처리방법 중 영지식증명의 수식을 활용하여 해당 수식을 수정하였으며 제안한 수식을 적용한 경우 기존 영지식증명의 수식을 적용한 경우보다 Verification time 측면에서 평균 10%정도 성능 향상을 확인하였다.

▸ **주제어**: 데이터 비식별화, 가명정보, 영지식증명

• First Author: Youn-A Min, Corresponding Author: Youn-A Min
*Youn-A Min (yah0612@hycu.ac.kr), Dept. of Applied Software Engineering, Hanyang Cyber University
• Received: 2021. 03. 11, Revised: 2021. 05. 03, Accepted: 2021. 05. 06.
• This paper is an extension of the paper ("Protocol sequence processing in blockchain-based zero-knowledge proof") presented at the 63rd Winter Conference of the Korea Computer Information Society in 2021.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

2020년 하반기에 개정된 개인정보보호법인 데이터 3법의 개정이 시행과 더불어 2016년 발표한 데이터 비식별화 가이드라인에 대한 실질적인 적용이 가능하고 다양한 분야에서 데이터의 실효성 있는 활용이 가능하게 되었다[1][2][3]. 데이터 3법 개정을 통하여 익명정보와 가명정보의 선별적 활용이 허용되었고 이에 따라 기업, 정부 등에서 개인정보의 폭넓은 활용이 기대된다[1][2].

익명정보란 특정개인을 식별할 수 없는 정보이며 가명정보란 추가정보 없이는 특정 개인을 식별할 수 없도록 데이터를 가명 처리한 정보이다[1][2]. 가명정보의 경우 개인의 식별정보나 속성 등을 추가함으로써 특정 개인을 유추할 수 있으므로 가명정보가 무분별하게 활용되는 경우 개인의 민감 정보 침해 및 개인정보유출의 위협이 발생할 수 있다[1][2]. 이에 가명정보를 통한 특정 개인의 데이터유추가 불가하도록 하는 다양한 방법이 연구 중이다.

본 논문에서는 개인정보 등 민감정보의 효율적 활용을 위하여 데이터 범위에 따른 차별화된 가명정보 처리방법을 제안하였다.

II. Background

1. alias information

정보통신의 발달에 힘입어 네트워크를 통하여 거래되는 데이터의 양이 증가하고 있다. 2019년 발표한 데이터산업 백서에 의하면 국내의 데이터 전체 시장규모는 2010년 이후 7% 이상의 지속적 성장세를 보이는 것으로 조사되었다[8]. 특히 개인정보의 가치에 대한 인식이 증가하며 개인정보의 폭넓은 활용을 위한 시도가 지속되고 있으며 이에 따라 데이터의 도용과 남용 등의 불법 활용에 따른 개인정보 침해사례도 지속적으로 증가하고 있다[4].

Table 1은 2019년 기준 공공. 민간분야 법령에 대한 개인정보침해요인과 개인정보침해의 신고센터 상담·접수 건수를 나타낸 것이다[1][4]. Table 1과 같이 개인정보의 무단수집 및 무단이용제공과 주민번호 등 개인의 민감 정보에 대한 도용 등의 사례는 꾸준히 증가하고 있음을 알 수 있다[5][6].

Table 1. Number of counseling cases for reporting personal information infringement [5][6]

[unit : Number of cases]

Division	´15	´16	´17	´18	´19
Unauthorized collection of personal information	2,442	2,568	1,876	2,764	3,237
Provision of unauthorized use of personal information	3,585	3,141	3,881	6,457	6,055
Theft of other people's information such as resident number	77,598	48,557	63,189	111,483	134,271
Refusal to withdraw membership or request for correction	957	855	862	1,149	1,292
etc	7,089	4,850	4,342	5,488	5,655

2. Data de-identification and alias processing

4차 산업혁명의 주요 키워드인 빅데이터, 인공지능 등의 발전에 힘입어 다양한 데이터를 활용한 데이터 산업 발전의 노력이 증가하며 지난 2020년 1월에 개인정보보호, 신용정보법, 정보통신망법의 개인정보 관련 3개 법령의 내용이 담긴 데이터 3법에 대한 개정안이 발의되었으며 2020년 8월부터 개정된 개인정보보호법이 시행되며 민간·정부· 산업에서의 개인정보가 활발히 이용가능하게 되었다[1][4]. 개정된 개인정보보호법에 의해 익명정보뿐 아니라 가명정보의 선별적 활용까지 가능하게 되었다[4].

가명정보는 데이터 비식별 처리에 의하여 처리된 데이터이다. 기존의 데이터 비식별 처리는 주민등록번호, 이름 등 개인을 식별할 수 있는 식별정보 삭제 등을 통한 낮은 수준의 처리방법과 관련 데이터에 대한 추가정보를 결합하여도 개인정보에 대한 유추가 전혀 불가능한 높은 수준의 처리방법으로 나뉜다.

2020년부터 시행되는 데이터 3법을 기반으로, 국내에서는 통계작성, 과학적 연구 및 공익적 기록보존 등의 연구목적 내에서 정보의 주체자의 동의 없이 가명정보를 비식별화 하여 활용할 있다[1][8].

데이터 비식별을 위하여 식별자(Identifiers)와 속성자(Attribute value)의 개념을 활용한다. 식별자란 개인을 직접적으로 식별할 수 있는 고유 식별정보와 상세주소 등의 정보이며 속성자란 해당 데이터 자체는 식별정보가 아니지만 다른 정보들을 통하여 정보주체를 추론할 수 있는 데이터이다[8][14].

현재 시행중인 데이터 비식별 처리방법은 가명처리(Pseudonymization), 총계처리(Aggregation), 데이터 삭제(Data Reduction), 데이터 범주화(Data Suppression), 데이터 마스킹(Data Masking)로 구분할 수 있다[9][10][11][14].

데이터 비식별화 방법 중 가명정보 처리방법은 휴리스틱 가명화와 암호화 방법으로 구분 되며 이 중 암호화 방법의 대표적 처리 방법으로 동형암호(Homomorphic Encryption)와 영지식증명(Zero-Knowledge Proof), 차등 정보보호(Differential Privacy),다자간 계산(Secure Multiparty Computation)을 들 수 있다[11][12][13].

Table 2. Encryption method for pseudonymization [7][10][11][12][13][14]

Cryptography for data de-identification	Characteristic
Homomorphic Encryption	Fast processing through parallel processing
Zero-Knowledge Proof	Can be used for anonymous authentication
Differential Privacy	Personal information can be protected through appropriate noise
Secure Multiparty Computation	A technology that can process a number of information anonymously to process a common goal

Table 2의 내용 중 동형암호는 데이터를 암호화 한 채 덧셈 및 곱셈의 연산을 보존하여 지원하는 기술로써 최근에는 머신러닝 및 딥러닝의 데이터 분석 및 활용을 유용하게 하는 기술로 활용되고 있다[11][12][13][14]. 영지식증명은 증명자(Prover)가 자신이 가지고 있는 비밀에 대한 정보를 노출하지 않고 자신이 비밀스러운 정보를 가지고 있다는 것을 확인자(verifier)에게 증명하는 기술이다. 최근에는 자신이 가지고 있는 비밀스러운 정보의 범위를 증명하는 range proof 방식을 적용하도록 연산 처리되어 고도의 익명성을 요구하는 데이터의 비식별화에 많이 활용되고 있다[11][12][13]. 차등정보호호란 데이터베이스(A)안의 그룹 B와 나머지 그룹 C에 대하여 개인의 정보가 B또는 C에 포함되어 있음을 명확히 보여주지 않고 정보의 포함여부에 대하여 적절한 노이즈를 섞어 주어 정보를 보호하는 기법으로 구글 및 애플 등에서 활용되는 기술이다[7][11][12][13]. 다자간 계산이란 인터렉티브 프로토콜(Interactive Protocol) 기반에서 공동의 목표를 완성하기 전까지 모든 노드가 온라인 상태를 유지하는 것을 기반으로 동형암호를 사용할 수도 있고 직접 원하는 연산을 구현하여 수행하기도 한다[11][12][13][14].

III. Research Proposal

1. de-identification data processing method considering the scope of use of pseudonym information

개정된 개인정보보호법에서 가명정보는 개인정보를 가명 처리하여 추가정보의 사용이나 결합 없이는 특정인을 알아볼 수 없는 정보라고 명시되어 있다 [1][11][14].

다양한 기관이 활용할 수 있는 개인정보의 범위가 확대됨에 따라 데이터에 연계기술 발전 및 추가정보의 다양한 결합에 의하여 가명정보의 내용이 유추가능해질 수 있는 상황을 가정하여 사용 범위에 따른 차별화된 데이터비식별화방법이 필요하다.

본 논문에서는 개인정보를 사용하는 경우 사용 범위에 따라 가명정보에 대한 보안을 강화할 수 있는 데이터 비식별화 방식을 제안한다.

본 논문에서 사용한 주요 알고리즘은 영지식증명이며 기존 영지식증명을 기반으로 타원곡선알고리즘의 접점 범위를 조정하여 변형된 타원곡선 알고리즘의 수식을 제안하였다.

타원곡선의 접점 범위 암호 설정의 경우 가명정보의 사용 범위를 구분한 변수를 적용하고 이에 따라 연산난이도를 높이기 위한 방법으로 전개한다.

2.. Proposal for a method of processing pseudonym information

2.1 Algorithms for processing alias information

본 논문에서 제안한 가명정보 처리 알고리즘 처리과정을 상세히 설명한다. 먼저, 유한그룹(F)를 포함하는 타원곡선(E)의 식을 데이터의 크기와 사용범위에 따라 설정한다. 해당 알고리즘에서 사용하는 p는 타원곡선 알고리즘을 토대로 접점의 범위 내의 가장 큰 소수로 설정한다. 해당 알고리즘에서 사용되는 x는 가명정보의 사용범위를 고려하기 위한 데이터의 일부로써 감추고자 하는 민감 정보이다. 본 논문에서는 기본적인 설정을 위하여 기존의 영지식증명 수식과 타원곡선 알고리즘의 방정식을 수정 활용한다[14]. F를 범위로 하여 가명 처리할 정보를 암호화 하며 암호 처리 시 p를 매개로 하는 그룹 내의 자연수를 임의로 추출한다. 이 후 기본적 영지식증명의 식을 통하여 x를 매개변수로 속성을 유추하려 하여도 유추가 불가능하도록 처리한다. 만일 x를 매개로 하여 데이터 속성 유추가 가능할 경우 p의 범위를 조정한다.

Fig.1은 데이터 사용범위에 따른 가명정보 처리의 수식 및 적용 절차이다.

① Applying DLP (Discrete Logarithm Problem) using finite F
② Arbitrarily set the equation of the elliptic curve E on F.
Applying the equation of weierstrass equation for the experiment of this paper
E=$y^2 + a_1xy + a_3y = x^6 + a_2x^2 + a_4x + a_6 \in F$
③ Extraction of natural numbers within a finite group with F as a range and the personal information-related variable p as a medium
F=<f> --> cycle group of order n (finite), x∈F
Finding the only x=f^d
d(mod n) (d=logfx)
At this time, the range for the finite group F is designated as an elliptic curve with the encrypted bit of personal information as a variable, and the number of elements of the elliptic curve is set as a natural number in
$p+1-2\sqrt{p}$ ~ $p+1+2\sqrt{p}$

④ Use of pseudonym information in which x is impossible to infer through basic zero-knowledge proof
y: given value, p: large prime (decimal), g: generator
Prover : g^x (mod p) = y , y
setup(share) y value to Verifier)
Prover : A = g^r (mod p), Setup (share) A value to Verifier
Verifier : (x+r) (mod (p-1)) or r challenge
$(A.y) \bmod p \equiv g^{(x+r) \bmod (p-1)} \bmod p$

$A \equiv g^r \bmod p$
Depending on the range of p, the difficulty of inference for x varies.

Fig. 1. Procedure for applying formula to de-identify pseudonym information

Fig.1에서 가명정보에 대한 안정성 평가를 위한 ④의 식을 통하여 p 범위의 적정성이 의심되면 p의 범위를 수정하여 암호처리의 난이도를 강화함으로써 가명처리의 안정성을 높인다.

2.2 Categorized by Data Usage

가명정보 처리를 위하여 데이터 사용범위에 따라 Table 3과 같이 분류하였다.

Table 3. Classification of data use range

division	p setting (de-identification bit)
Functional application stage: For simple research and statistical use	x<=32 Adjust p value within x bit range
Linked application steps: To link data between institutions	x<=64
Guaranteed use steps: For data linkage between multiple organizations	x>64

Table 3에서 제시한 사용범위 구분은 기능적 활용단계와 연계형, 보장형 활용단계로 구분한다.

기능적 활용단계는 단순 연구 및 통계활용으로 구분되는 경우이며 연계형은 단일 기관 내의 비식별 처리방법으로, 여러 기관 간 데이터 연계 시 보장형 활용단계의 범주로 구분한다.

모든 단계에서 알고리즘 적용 시 Table 3의 x의 범위를 고려하였다. Table 3에서 제한한 x는 데이터 중 임의의 비트(bit)수이며 데이터의 크기에 비례하여 적절히 x의 크기를 변형할 수 있다.

각 단계에서 명정보 처리를 위한 알고리즘은 데이터일부에 대한 암호화 및 암호화 데이터일부의 텍스트 대체를 통한 특정 변수(p)를 제시함으로써 데이터 범위 설정 및 해당 내용을 통한 비식별화 처리가 가능하도록 한다. 특정 변수 p는 데이터 중 일부 비트(x) 내의 소수이며 p를 매개로 하는 그룹 내의 자연수를 추출할 때 사용 되며 영지식 증명 수식의 기본이 되는 타원곡선알고리즘의 원소개수 선정을 위한 범위로 활용한다.

기능적 활용단계에서는 단순한 통계 분석. 연구데이터로 활용되기 위한 가명처리를 원칙으로 하며 빠른 처리 및 재활용 가능성을 처리의 주안점으로 한다.

p의 설정을 통하여 데이터 사용범위에 따른 효율적 연산 처리 및 병렬처리가 가능하도록 하며 변수 자체에 대한 유추 불가를 통하여 가명 정보의 속성정보 유추가능성이 낮아진다.

연계형 활용단계에서는 단일 기관 내 데이터 공유 시 추가정보에 의한 개인정보 유추가 거의 불가능하도록 하는 것을 목표로 하며 동일 기관이라도 공유되는 데이터에 대한 개인정보 유추 및 개인정보 악용 불가능하도록 한다. 비식별화 알고리즘은 암호화된 데이터의 일부를 재 암호화한 후 변수를 사용한 함수식을 통하여 p의 범위를 처리하고 해당 단계의 처리를 통하여 빠른 연산 처리 및 수치데이터에 대한 빠른 처리가 가능하다.

보장형 활용단계는 다기관간 데이터 공유를 통하여 데이터 패턴 분석 등이 시행될 경우 가명정보 처리 후 추가데이터의 결합을 통하여 개인정보 유추가 불가능하도록 하여 산업적 활용이 가능하다.

3. Experiment and performance evaluation

임의로 생성한 데이터 셋 10개에 대하여 가정정보의 verification time을 측정하기 위하여 실험환경과 성능평가환경은 Fig.2와 같다.

Experiment environment :
Raspberry Pi 4 / 1.5GHz quad-core CPU
SDRAM 2GB / Data format: CSV
comparison target :
A simple zero-knowledge proof formula is applied in the case of the same range of p
Same data format as other experimental environment
Alias data set for performance evaluation: 10
Prediction dataset for performance evaluation: 25 ~ 30
Performance evaluation factor: Verification Time
Performance evaluation analysis tool: R studio

Fig. 2. Experiment and performance evaluation environment

Fig.2의 내용을 토대로 동일한 환경에서의 5회의 반복 수행을 하였으며 해당 내용에 대한 평균 측정시간을 Table 4와 같이 정리할 수 있다.

Table 4. Verification Time Comparison Measurement
[unit : data set/ms]

Division	1	3	5	8	10
Suggested formula	692	712	756	846	929
Comparison formula	731	795	845	962	1019

본 연구의 실험은 실험환경의 과부하에 따른 오류를 제거하고자 데이터의 수를 작게 하였다.

데이터 분석 툴인 R을 통하여 실험의 결과를 토대로 데이터 변동 및 성능 추이를 식으로 나타낼 수 있다. x1=c(1,3,5,8,10)의 값을 토대로 코드 식 coef(m1)을 수행하고 fitted(m1)를 수행한 후 predict(m1,newdata=new1)와 predict(m2,newdata=new2)를 수행하여 Table 5의 식을 산출하였다.

Table 5. Comparison and measurement of verification time according to data set increase (prediction)
[unit : data set/ms]

Data set /ms	20	30	40
Suggested formula y≒26.5x+642	1052.323	1413.887	1604.451
Comparison formula y≒30.85x+701	1391.350	1625.808	1934.267

Table 5와 같이 제안 수식을 통하여 데이터 셋이 20~40개까지 증가한 경우를 유추한 결과 데이터 셋이 20인 경우 11052ms 이며 기존 영지식증명의 수식은 1391ms의 성능을 보인다. 제안 수식 및 비교 수식에 대한 성능평가에 대하여 Table 6과 같은 성능향상을 추측할 수 있으며 이에 대한 그래프를 Fig.3과 같이 나타낼 수 있다.

Table 6. Degree of performance improvement
[unit : data set/%]

Division	1	3	5	8	10	20	30	40
Performance improvement	9%	9%	9%	10%	10%	11%	11%	12%

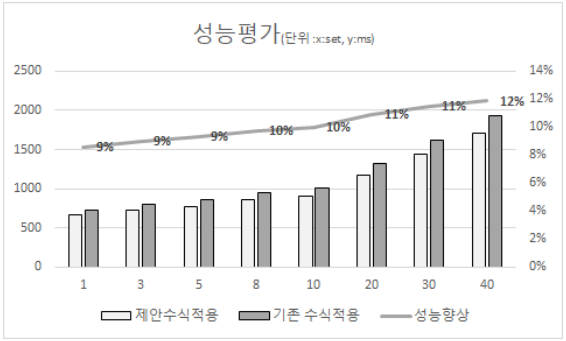

Fig. 3. Performance evaluation result

Table 6의 수치와 같이 제안 수식은 기존 영지식증명 수식 대비 평균 10%의 성능향상을 보였으며 데이터 셋의 증가에 따라 점점 성능이 향상됨을 알 수 있다.

Table 6의 제안 수식에 대한 유의미성을 측정하기 위하여 Null hypothesis: μ=0 (the prediction is unreliable) 및 Alternative hypothesis: μ>0 (There is confidence in the prediction)의 같은 가설을 세우고 Table 5의 추이 예측 수식을 통하여 비교한 결과 제안수식과 비교수식의 t-value가 각각 8.532, 19.89이고 유의수준 0.05 기준 p-value가 각각 0.00338, 0.000278임을 파악하였으며 이를 통하여 귀무가설이 기각되고 제안내용이 측정 및 예측 가능한 수식으로써 유의미성을 가짐을 입증할 수 있다.

IV. Conclusions

데이터의 가치에 대한 인식이 높아지며 정부와 기업 등에서 개인정보를 활용하고자 하는 노력이 지속되고 있다. 2020년 8월에 데이터 3법이 통과됨에 따라 개인정보에 대하여 익명화된 정보와 가명정보의 부분적 활용이 가능하게 되었다.

가명정보의 경우 데이터 연결 및 결합을 통하여 특정 개인의 정보를 유추할 수 있다. 정보 활용을 위한 다양한 가명정보 처리방법이 제시되고 있지만 데이터 사용범위가 다양화됨에 따라 데이터 사용범위를 고려한 가명정보 처리방법이 필요한 상황이다. 본 논문에서는 데이터의 사용범위를 고려하여 해당 그룹에서의 개인정보 중요도에 따

라 가명처리 범위를 임의로 설정하고 영지식증명 수식을 수정 적용하였다.

제안 내용에 대한 성능평가를 위하여 샘플데이터를 생성하고 해당 데이터 셋에 대하여 기존 영지식증명 수식과 제안한 수식을 각각 적용하여 Verification Time을 측정하였으며 이를 통하여 제안 수식의 안정성과 효율성 측면을 확인하였다. 또한 R을 통하여 데이터 셋 증가를 고려한 추이수식을 생성하여 데이터 셋 증가상황에서도 해당 제안이 효율적으로 작동하는지를 확인하였다.

본 논문에서 제안한 방법을 통하여 가명정보의 데이터의 사용자 범위를 고려한 안정성확보를 위한 처리가 가능하고 처리시간의 효율성을 증대할 수 있다. 향후 네트워크 환경에서의 다양한 이벤트 발생을 고려한 안정성 측면을 고려한 연구를 지속할 예정이다.

REFERENCES

[1] 2020 Regular Assembly Report, “Annual Report on Personal Information Protection”, 2020

[2] Choi Ji-seon, Lee Ye-won, Oh Yong-seok, and Lim Hyeong-jin, "International Standardization Trends in Non-identification Processing", Journal of the Korea Information Security Society, Vol 29, No 4, pp 13-18,2019

[3] Kwang-hee Choi, Sang-Jun Lee, "A Study on the Risk Calculation Method for Using Pseudonymized Information", Korean Enterprise Architecture Society, vol.17, no.2, pp. 167-177, 2020

[4] Researchers of Korea Data Industry Promotion Agency, "2019 Data Industry White Paper", Korea Data Industry Promotion Agency, Vol 22, 2019.

[5] Personal Information Infringement Report Center Application Materials, https://privacy.kisa.or.kr/main.do,2020

[6] Number of personal information infringement cases, http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1366,2020

[7] Chun, Heuiju, Yi, Hyun Jee, Yeon, Kyupil, Kim Dongrae, "Data Quality Measurement on a De-identified Data Set Based on Statistical ", The Journal of the Korea Contents Association, Vol. 19, No. 5, pp.553-561,2019

[8] Soohyun Um, Inkyung Lee, Woogi Lee, "Big Data-based Trend of Personal Information De-idendification,", Informatization Research, Vol. 15, No. 4, pp.545-552,2018

[9] Gyu-seong No and Ju-yeon Lee, “A Study on the Analysis of Differences in Perceptions of Big Data in the Age of Convergence”, Vol.13, No.10,pp. 305-312, 2015

[10] Hyun-cheol Yang, Young-ju Lee, Shin-Gon Kim, "Effects of Application Level of Personal Information De-identification Technology on Intention to Use Big Data", Journal of Informatization Research, Vol. 13, No. 3, pp. 395-404, 2016

[11] Seung-Hwan Kim and Seong-Hae Jeon, “Big Data Integration Using Data De-identification”, Journal of the Korean Intelligent Systems Society, Vol. 29, No. 3, pp. 235-241, 2019

[12] Cha, S.-C.1,,Hsu, T.-Y.1, Xiang, Y.2, Yeh, K.-H.3,4, “Privacy enhancing technologies in the internet of things: Perspectives and challenges”, IEEE Internet of Things Journa, Vol.6, No.2, pp.2159-2187, 2019

[13] Ministry of Government Administration and Home Affairs, “Guideline for non-identification measures for personal information-Guidelines for non-identification measures and support and management system”, 2016.

[14] Hongjin Kim, "A Scheme to authenticate and manage based on zero knowledge proof for IoT device on blockchain", Sogang University's Electronic Engineering thesis, 2018

Authors

Youn-A Min received a Ph.D. in computer science from Dongguk University, Korea, in 2008, 2013. Dr. Youn-A Min is a professor of applied software engineering at Hanyang Cyber University, 2020.

She is also a visiting professor at Hanyang University. She is interested in embedded system security and blockchain.