Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.5, Oct. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.5.915

결합키 생성항목의 갱신에 강건한 결합키 생성 기법*

장 호 빈,[1†] 노 건 태,[2] 정 익 래,[3] 천 지 영[2‡]
[1,3]고려대학교 (대학원생, 교수), [2]서울사이버대학교 (교수)

Combination Key Generation Scheme Robust to Updates of Personal Information*

Hobin Jang,[1†] Geontae Noh,[2] Ik Rae Jeong,[3] Ji Young Chun[2‡]
[1,3]Korea University (Graduate student, Professor),
[2]Seoul Cyber University (Professor)

요 약

개인정보 보호법과 가명정보 처리 가이드라인에 따르면, 서로 다른 결합신청자들이 결합을 희망할 때 Salt값을 포함한 결합키 생성항목의 해시값으로 매핑을 진행한다. 결합키 생성항목의 예시로는 성명, 전화번호, 생년월일, 주소 등의 개인정보가 될 수 있으며, 해시 함수의 특성상 서로 다른 결합신청자들이 이들의 항목을 정확히 동일한 형태로 저장하고 있을 때 문제없이 결합을 진행할 수 있다. 하지만 이러한 기법은 서로 다른 결합신청자들의 데이터베이스 갱신 시점이 달라서 발생하는 주소 변경, 개명 등의 시나리오에서의 결합은 취약하다. 따라서 본 연구에서 우리는 주소 변경, 개명 등의 결합키 생성항목이 갱신된 시나리오에서도 개인정보보호를 만족하는 강건한 결합키 생성 기법을 확률적 자료 연계를 통한 임계값을 바탕으로 제안하며, 본 연구 결과를 활용한 국내 빅데이터 및 인공지능 사업의 발전에 기여하고자 한다.

ABSTRACT

According to the Personal Information Protection Act and Pseudonymization Guidelines, the mapping is processed to the hash value of the combination key generation items including Salt value when different combination applicants wish to combine. Example of combination key generation items may include personal information like name, phone number, date of birth, address, and so on. Also, due to the properties of the hash functions, when different applicants store their items in exactly the same form, the combination can proceed without any problems. However, this method is vulnerable to combination in scenarios such as address changing and renaming, which occur due to different database update times of combination applicants. Therefore, we propose a privacy preserving combination key generation scheme robust to updates of items used to generate combination key even in scenarios such as address changing and renaming, based on the thresholds through probabilistic record linkage, and it can contribute to the development of domestic Big Data and Artificial Intelligence business.

Keywords: Combination Key Generation, Record Linkage, Pseudonymization, Big Data, Artificial Intelligence

Received(08. 05. 2022), Modified(09. 19. 2022), Accepted(09. 21. 2022)

* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2021R1F1A1063992).

† 주저자, jhb0858@korea.ac.kr
‡ 교신저자, jychun@iscu.ac.kr(Corresponding author)

I. 서 론

4차 산업혁명의 핵심 기술인 빅데이터와 인공지능을 활용하여 수많은 기업들은 개인 맞춤형 광고, 금융 상품 추천 등의 다양한 서비스를 제공하고 있다. 다양한 빅데이터 및 인공지능 기반의 서비스 제공을 위해 유럽, 호주 등 다수의 국가에서는 마이데이터 사업 및 정책을 시행하고 있으며, 이를 위한 빅데이터 구축 및 데이터에 대한 개인정보보호가 필수적이다[1]. 하지만 빅데이터 구축을 위한 데이터베이스 결합 시, 개인정보 보호법에 따라 이름, 주민등록번호 등 특정 개인을 식별할 수 있는 정보를 사용할 수 없다.

따라서 현재 국내 빅데이터 및 인공지능 사업의 기반이 되는 데이터 결합 시, 가명정보 처리 가이드라인 등을 준수하여 개인 식별이 불가능한 가명 데이터를 생성해야 하며, 결합키관리기관과 결합전문기관을 통해 데이터베이스 결합이 진행되어야 한다[2]. 결합키관리기관과 결합전문기관을 통해서 결합신청자들의 데이터베이스를 결합할 때, 데이터의 구분자 역할을 하는 결합키가 매핑의 도구로 사용된다. 결합키 생성 시 Salt값을 포함한 성명, 전화번호, 생년월일, 주소 등과 같은 결합키 생성 항목의 해시값을 사용하는데, 해시 함수의 특성상 결합키 생성항목의 데이터가 정확히 일치하는 경우에만 매핑이 성공적으로 진행된다. 따라서 결합키 생성항목에 속하는 데이터의 형식 일치 등의 전처리가 필요할 수 있으며, 결합신청자들의 결합키 생성항목에 대한 데이터베이스 갱신 시점이 다른 경우, 서로 다른 데이터로 판단될 가능성이 존재한다.

따라서 본 논문에서 우리는 국내 데이터베이스 결합 시 발생 가능한 주소 변경, 전화번호 변경, 개명 등과 같은 갱신이 서로 다른 시점에 진행된 서로 다른 결합신청자들의 데이터베이스에서도 동일 개체임을 확인할 수 있는 강건한 결합키 생성 방안을 제시하고자 한다. 이를 위해 유럽, 미국 등 해외에서 연구되고 있는 오탈자 및 약어 표기 등이 이루어진 데이터베이스 간의 결합을 위한 자료 연계(record linkage) 기법을 활용한다. 자료 연계는 데이터베이스의 결합 시 자료(record)의 공통 속성의 유사도를 이용하여 서로 다르게 표현된 자료들 사이에서 동일 개체의 자료인지를 판별하는 방법이다[3, 4]. 자료 연계로 초기에 제안된 Fellegi-Sunter 모델 이후 블룸 필터 등을 이용한 자료 연계 방안들이 연구되고 있다[5, 6]. 하지만 지금까지 연구된 대부분의 자료 연계 방안은 유럽, 미국 등에서 자주 발생하는 “John”과 “Jon”과 같은 오탈자, “Street”와 “St.”과 같이 약어가 포함된 주소나 성명 등에 강건한 자료 연계이다. 기존의 자료 연계 방안을 바탕으로 본 논문에서는 결합키 생성항목의 갱신 시점이 서로 다른 데이터베이스 결합을 위한 결합키 생성기법 알고리즘을 제안한다.

또한, 제안하는 기법의 강건함을 보이기 위해 공공데이터를 기반으로 한 다양한 시나리오로 이루어진 가상의 갱신된 데이터베이스를 구축하여 실험을 진행하였다. 최종적으로 결합하려는 데이터베이스의 자료에 대한 유사도를 기준으로 시나리오별 결합키 생성에 대한 적절한 임계값인 0.8, 0.9를 제시하고, 해당 임계값에 대한 자료 연계 가능성을 판단한다. 즉, 기존의 Salt값을 포함한 결합키 생성항목의 해시값으로 매핑을 진행할 때에는 불가능했던 경우에 대해서도 동일한 개체인 경우에는 자료 연계가 가능함을 확인하였으며, 현재 결합키관리기관과 결합전문기관을 통한 데이터 결합 시 발생 가능한 문제점 중 하나인 결합키 생성을 위한 결합키 생성항목의 데이터 전처리와 해당 데이터가 갱신된 경우에 대한 결합 불가능성 등을 최소화하였다.

본 논문에서 제안하는 기법은 기존의 가명정보 처리 가이드라인을 준수하면서도 서로 다른 결합신청자들의 데이터베이스 갱신 시점이 달라도 가명정보 결합이 가능한 결합키 생성기법이다.

II. 배경 지식

2.1 개인정보 보호법

개인정보 보호법은 데이터 3법 중 하나로써, 2020년 2월 개정된 개인정보 보호법에서 정의하는 “개인정보”는 살아 있는 개인에 관한 정보로서 이름, 주민등록번호 등 개인을 식별할 수 있는 정보 및 다른 정보와 쉽게 결합하여 특정 개인을 유추할 수 있는 정보 등을 의미한다. 또한 “가명 처리”는 개인정보의 일부 또는 전체를 삭제하는 방법으로 추가 정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 의미한다[2].

개인정보 보호법 개정으로 인해 “가명정보”를 이용한 개인정보의 활용 및 확대, 마이데이터 사업 등이 시작되었으며, 빅데이터를 이용한 상품 추천 등 다양

한 데이터 관련 사업이 진행되고 있다. 하지만 데이터 활용에 따른 개인정보 침해의 우려가 대두되고 있으며, 개인정보보호의 중요성이 대두되고 있다[7].

2.2 가명정보 처리 가이드라인

2022년 4월 발간된 가명정보 처리 가이드라인은 개정된 개인정보 보호법에 맞는 가명 처리 및 데이터 결합 절차를 안내한다. 가명 처리를 통해 달성하고자 하는 목적에 따른 가명 처리 대상 항목 선정, 개인 식별 위험성 등의 가이드라인을 제시하며, 가명 처리된 정보에 대한 데이터 결합 과정을 나타낸다[2].

현재 데이터 결합 과정은 Fig. 1.과 같으며 "결합신청", "결합 및 추가처리", "반출 및 활용", "안전한 관리"의 과정으로 진행되며 다음과 같다.

① 결합 신청: 결합신청자는 신청자 간에 가명정보 결합에 필요한 사전 준비사항을 확인하고 결합전문기관에 결합을 신청한다.

② 결합 및 추가 처리: 가명정보를 제공하는 결합신청자는 결합키 관리기관으로부터 결합키 생성에 이용되는 정보(Salt)를 수신하여 결합키를 생성하고, 결합에 필요한 정보를 각 기관에 전송한다.

③ 반출 및 활용: 결합정보 또는 분석결과 등을 반출하려는 경우, 결합전문기관에 반출을 신청한다.

④ 안전한 관리: 결합정보를 이용하는 결합신청자는 반출한 결합정보를 목적에 따라 처리하고, 안전조치 의무 등을 준수한다.

"결합 및 추가처리" 과정 시, 결합키를 통해 데이터의 가명처리가 이루어진다. 결합신청자는 결합키 관리기관으로부터 Salt값을 전송받아 SHA256 등의 결합키 생성 알고리즘를 이용해 결합키를 생성한다. 결합키 생성을 위한 결합키 생성항목은 "이름", "전화번호" 등 결합신청자 간에 동일하게 가지고 있는 속성을 결합신청자 간의 협의를 통해 선정한다.

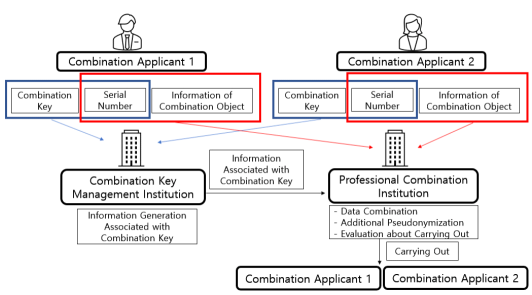

Fig. 1. Combination and carrying out pseudonymized information

이후, 결합전문기관은 결합키 관리기관으로부터 받은 결합키 연계정보를 토대로 가명처리된 데이터를 결합한다. 그리고 결합된 데이터를 이용하고자 하는 결합신청자에게 반출 및 데이터에 대한 안전한 관리를 진행한다.

2.3 Fellegi-Sunter 모델

1969년 Fellegi와 Sunter가 발표한 Fellegi-Sunter 모델은 확률적 자료 연계에 대한 기본 개념을 제시하였다[4]. 오탈자, 약어 등이 포함된 서로 다른 데이터셋(dataset)에 대해 자료를 구성하는 "이름", "주소" 등의 속성(field)에 가중치를 부여한다. 그리고 부여된 가중치를 바탕으로 자료 간의 유사도를 판별하여 자료를 연계한다. Fellegi-Sunter 모델은 두 데이터베이스 간에 공통 속성이 있음을 가정하며, 공통 속성의 가중치에 따라 연계(link), 비연계(non-link), 연계 가능성(possible-link)을 판정한다.

2.4 블룸 필터 (Bloom Filter)

블룸 필터는 1970년 Bloom에 의해 제안된 자료구조로써, 원소의 집합에 대한 포함 여부를 확인하는 확률적 자료 구조이며, 블룸 필터 길이만큼의 균등한 확률을 출력하는 해시 함수를 사용하여 집합에 대한 원소의 포함 여부를 확인할 수 있다[8]. 집합에 포함되지 않는 원소는 블룸 필터 적용 시 0인 bit가 존재하며, 모두 1인 bit를 가질 경우, 해당 집합에 포함되는 원소라 추측할 수 있다.

예를 들어, Fig. 2.와 같이 $n=3$개의 원소를 갖는 집합 $S=\{x,y,z\}$의 각 원소에 대해 $k=2$개의 서로 다른 해시 함수를 사용하여 길이 $m=8$의 블룸 필터를 생성한다고 가정하자. w_1과 같이 블룸 필

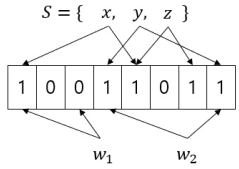

Fig. 2. Example of Bloom Filter

터 결과가 0인 bit를 가질 경우, 해당 원소는 S의 원소가 아님을 확신할 수 있다. 또한 w_2와 같이 모든 bit가 1인 경우, 해당 원소는 S의 원소가 될 가능성이 존재한다.

III. 관련 연구

EM 알고리즘은 기댓값 최대화 알고리즘으로 통계 모델의 수식을 정확히 해결할 수 없을 때, 원하는 값들이 최대로 나올 가능성인 최대가능도를 구하는 데 사용된다. 2000년 Winkler는 EM 알고리즘을 이용하여 Fellegi-Sunter 모델의 자료 연계를 위한 속성의 가중치를 계산하는 방법을 제안하였다[9].

2009년 Schnell 등의 연구에서는 블룸 필터와 암호학적 해시 함수를 이용한 개인정보보호 자료 연계 방법을 제안하였다[5]. 해당 연구에서는 독일의 전화번호부 데이터와 가상의 데이터베이스를 활용하여 오탈자, 약어 등이 포함된 데이터셋을 구성하였다. 구성된 데이터셋의 "이름", "주소" 등의 속성을 기반으로 블룸 필터를 생성하는 방법을 제시하였으며, 블룸 필터 생성을 위한 암호학적 해시 함수의 개수 등 최적의 블룸 필터 매개변수를 제시하였다.

2014년 Durham 등의 연구에서는 [5, 9]의 연구를 바탕으로 강화된 개인정보보호 자료 연계 방법을 제안하였다[6]. 해당 연구에서는 미국 North Carolina 주의 투표자 데이터를 바탕으로 오탈자, 약어 등이 포함된 데이터셋을 구성하였다. 데이터셋 속성의 가중치 계산을 위해 EM 알고리즘을 이용한 Fellegi-Sunter 모델을 적용하였으며, 개인정보보호 자료 연계를 위해 암호학적 해시 함수를 이용한 속성 기반의 블룸 필터와 데이터셋을 구성하는 자료 기반의 블룸 필터를 계산하는 방법을 제안하였다. 또한, 블룸 필터 인코딩 결과에 대한 사전 공격을 방지하기 위한 치환(permutation) 모델을 제안하였다.

2018년 Christen 등의 연구에서는 블룸 필터를 활용한 개인정보보호 자료 연계에 대한 공격을 제시하였다[10]. 공격자가 자료 연계에 사용된 데이터베이스에 접근 가능함을 가정하며, 해당 연구에서 사용된 데이터셋은 [6]의 연구와 동일한 데이터셋을 사용한다. 자료 연계에 사용된 고정된 블룸 필터의 길이, 해시 함수의 개수 등의 매개 변수에 대해 개인정보보호가 이루어지지 않을 수 있는 값을 제시하였다.

이처럼 기존의 개인정보보호 자료 연계 방안은 모두 유럽, 미국 등의 오탈자, 약어 표기가 이루어진 데이터베이스 간의 결합을 위한 연구이다. 이를 바탕으로 국내에서 발생 가능한 결합키 생성항목의 갱신 시점이 다른 데이터베이스 간의 결합 시나리오 및 기존의 가명정보 처리 가이드라인을 준수하는 결합키 생성기법을 제안한다.

IV. 개인정보보호 자료 연계 모델

4.1 기본 모델

기존 가명정보 처리 가이드라인 "결합 및 추가처리" 과정의 Salt를 이용해 결합키를 생성하는 과정 대신, [6]에서 제안한 개인정보보호 자료 연계 모델을 이용하여 가명정보 결합을 진행하고자 한다. [6]의 모델을 이용하기 위해 필요한 파라미터들과 [6]의 마지막 과정인 자료 기반 블룸 필터 치환(4.2.5절)에서 필요한 비밀 정보는 결합키관리기관이 결합신청자에게 제공한다.

자료 연계를 위한 데이터베이스는 속성, 데이터, 자료, 데이터셋으로 구분된다. 데이터베이스 Attribute에 해당하는 속성은 데이터를 분류하는 기준이다. 자료는 데이터베이스 Tuple에 해당하며, 데이터는 자료의 속성별 값이다. 데이터셋은 분석하고자 하는 자료의 모음이다. Fig. 3.은 임의의 데이터베이스 A, B에 대한 자료 연계 예시이며, 이름이 같은 자료는 연계를 시킬 수 있음을 의미한다.

[6]에서 제안한 블룸 필터를 이용한 개인정보보호 자료 연계 모델은 "속성 기반 블룸 필터 생성"(FBF (Field level Bloom Filter) parameterization & generation), "적합한 비트 식별"(eligible bit identification), "속성 가중치 선정"(field weighting), "자료 기반 블룸 필터 생성"(RBF (Record level Bloom Filter) parameterization & generation), "자료 기반 블룸 필터 치환"(RBF permutation)의 과정을 가진다. 전체적인 모델은 Fig. 4.와 같다.

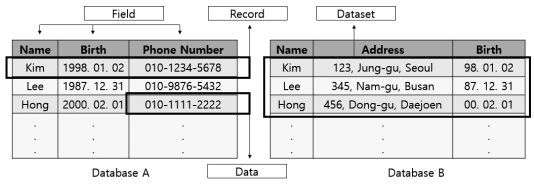

Fig. 3. Example of record linkage database

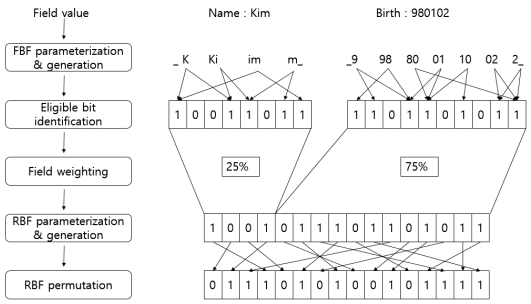

Fig. 4. Privacy-preserving record linkage model using Bloom Filter

"속성 기반 블룸 필터 생성"은 각 데이터의 블룸 필터 길이를 생성한다. "적합한 비트 식별"은 데이터의 블룸 필터 길이를 이용해 속성을 기준으로 블룸 필터 길이를 계산한다. 그리고 계산된 속성별 블룸 필터 길이를 갖는 데이터의 블룸 필터를 계산한다. 블룸 필터 계산 시, 암호학적 해시 함수를 사용하여 데이터에 대한 개인정보보호를 이룬다. "속성 가중치 선정"은 각 데이터셋에 맞는 속성 가중치를 선정하는 것이다. "자료 기반 블룸 필터 생성"은 선정된 속성 가중치에 맞게 자료의 블룸 필터를 생성한다. "자료 기반 블룸 필터 치환"은 자료의 블룸 필터에 대한 사전 공격을 방지하기 위해 두 데이터셋 간에 합의된 치환을 진행한다. 최종적으로 자료의 블룸 필터에 대한 상관관계를 분석한다.

4.2 단계별 매개변수 선정

4.2.1 속성 기반 블룸 필터 생성

집합 $S=\{s_1, s_2, ..., s_n\}$을 k개의 해시 함수를 이용하여 길이 m의 블룸 필터 인코딩을 진행한다고 가정하자. 이 때, 블룸 필터 인코딩 bit가 0이 나올 확률 p는 (1)과 같다[11]. [6]에서 제안된 Dynamic FBF는 속성 기반 블룸 필터를 각 속성에 적용하였을 때 블룸 필터 길이이며 (2)와 같다.

$$p=(1-\frac{1}{m})^{kn}\approx e^{-kn/m} \quad (1)$$

$$m=1/(1-\sqrt[kn]{p}) \quad (2)$$

예를 들어, "이름" 속성의 "홍길동" 값에 Dynamic FBF를 적용한다고 하자. 오타 발생 가능성에 의해 "홍길동"의 값을 갖는 데이터를 q글자 단위로 분리한다[5]. $q=2$일 때, $S=\{-홍, 홍길, 길동, 동-\}$ ($-$: blank), $n=4$임을 알 수 있다. 그리고 15개의 해시 함수를 이용한 블룸 필터 인코딩 bit의 0, 1 비율을 50%로 가정하면, $k=15, p=0.5$이다. 이를 계산 시 $m\approx 87$이다.

4.2.2 적합한 비트 식별

분석하고자 하는 두 개의 데이터셋에 대해 (2)를 적용하기 위한 속성별 블룸 필터 길이를 선정하는 과정이다. 데이터를 q글자 단위로 분리한 집합 S의 크기 n의 평균과 k, p를 이용해 구한 m을 속성별 블룸 필터 길이로 한다.

예를 들어, Fig. 3.의 공통 속성인 "이름", "생년월일"에 대해 $q=2$, $k=15$, $p=0.5$일 때 Dynamic FBF를 이용해 속성별 블룸 필터 길이를 구하면 Table 1.과 같다.

Table 1. Example of eligible bit identification

Field	Name	Birth
average size of S	4	8
Dynamic FBF	87	174

4.2.3 속성 가중치 선정

i번째 속성의 가중치 $w[i]$는 일치 가중치 $w_a[i]$, 비일치 가중치 $w_d[i]$를 이용하여 계산한다[6].

$$range[i]=w_a[i]-w_d[i] \quad (3)$$

$$w[i]=\frac{range[i]}{\sum range[i]} \quad (4)$$

일치 가중치 $w_a[i]$와 비일치 가중치 $w_d[i]$는 Fellegi-Sunter 모델과 EM 알고리즘을 이용하여 계산하며, 계산 상의 편의를 위해 로그를 사용한다. m_i는 두 데이터셋의 일치하는 자료 쌍 M개 중 i번째 속성의 데이터 쌍이 일치할 확률이다. u_i는 두 데이터셋의 일치하지 않는 자료 쌍 U개 중 i번째 속성

의 데이터 쌍이 일치하지 않을 확률이다. i번째 속성의 데이터가 일치하는 데이터 쌍의 집합을 M_i, 일치하지 않는 데이터 쌍의 집합을 U_i라 할 때, $w_a[i], w_d[i], m_i, u_i$는 (5), (6)과 같다.

$$m_i = \frac{|M_i|}{|M|},\ u_i = \frac{|U_i|}{|U|} \tag{5}$$

$$w_a[i] = \log_2\left(\frac{m_i}{u_i}\right),\ w_d[i] = \log_2\left(\frac{1-m_i}{1-u_i}\right) \tag{6}$$

데이터가 일치하는 기준은 Jaro-Winkler similarity를 이용하여 판단한다[12]. Jaro-Winkler similarity는 Jaro similarity를 Winkler가 발전시킨 것으로, Jaro similarity sim_j는 (7)과 같다. m은 문자열 s_1, s_2의 일치하는 문자의 개수, t는 $\lfloor \max(|s_1|,|s_2|)/2 \rfloor - 1$ 범위에서 m에 포함되지 않으며 일치하는 문자의 개수이다.

Jaro-Winkler similarity sim_w는 s_1, s_2의 공통 접두사 길이 $l\ (0 \le l \le 4)$과 similarity 조정을 위한 scaling factor $p\ (0 < p \le 0.25)$에 대해 (8)과 같다. p의 기본값은 0.1이며, p가 클수록 일치하는 문자가 적은 경우에도 문자열 s_1, s_2가 일치한다고 판단한다.

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{else} \end{cases} \tag{7}$$

$$sim_w = sim_j + lp(1 - sim_j) \tag{8}$$

4.2.4 자료 기반 블룸 필터 생성

자료 기반 블룸 필터는 속성별로 선정된 가중치 비율에 맞게 병합될 블룸 필터의 길이를 선정한다. 그리고 속성별로 동일한 치환을 적용하여 블룸 필터를 병합한다.

예를 들어, Table 1.과 같이 Dynamic FBF로 계산된 "이름", "생년월일"의 블룸 필터 길이가 87, 173이며, Fig. 4.와 같이 속성의 가중치가 25%, 75%라 가정하자. "이름" 속성의 RBF 길이는 $87 \times (1/0.25) = 348$이며, "생년월일" 속성의 RBF 길이는 $174 \times (1/0.75) = 232$이다. 그리고 속성별 RBF 길이 최대값을 기준으로 가중치 비율에 맞게 각 속성의 블룸 필터 길이를 계산한다. 따라서 $\max(348, 232) = 348$이므로 "이름", "생년월일"의 블룸 필터 길이는 $87, 261\ (348 \times 0.25, 348 \times 0.75)$이다.

그리고 Fig. 5.와 같이 속성별로 동일한 치환을 적용하여 블룸 필터를 병합한다. 치환은 랜덤하며, 비교하고자 하는 데이터셋의 동일한 속성 간에 동일한 치환을 적용한다.

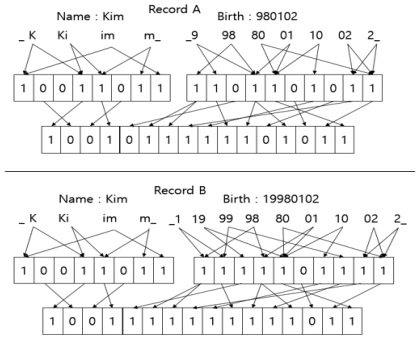

Fig. 5. Example of RBF

4.2.5 자료 기반 블룸 필터 치환

자료 기반 블룸 필터로 병합된 블룸 필터에 대해 Fig. 6.와 같이 두 데이터셋 간에 합의된 치환을 적용하여 블룸 필터에 대한 사전 공격을 방지한다. 그리고 최종적으로 계산된 블룸 필터에 대해 (9) 식과 같이 Dice Coefficient를 계산함으로써 두 자료 간의 상관관계를 계산한다. 자료 간의 일치하는 문자가 많을수록 Dice Coefficient는 1에 근사하며, 그렇지 않을 경우, 0에 근사한다. 비트스트링으로 표현된 블룸 필터의 동일 위치의 bit가 일치하는 경우 (9) 식의 교집합의 원소가 된다.

$$\text{Dice Coefficient} = \frac{2(X \cap Y)}{|X| + |Y|} \tag{9}$$

예를 들어, Fig. 6.에 대한 Dice Coefficient 계산 시, Record A, B는 각 16bit이며, 일치하는 bit가 14개이므로 Dice Coefficient는 $\frac{2(\text{RecordA} \cap \text{RecordB})}{|\text{RecordA}| + |\text{RecordB}|} = \frac{2 \times 14}{16 + 16} = 0.875$이다.

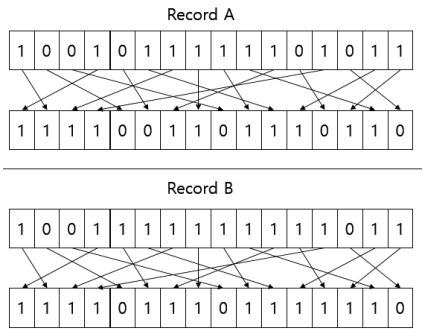

Fig. 6. Example of RBF permutation

4.3 연구 모델

기본 모델(4.1절, 4.2절)을 제안하고자 하는 결합키 생성 기법에 적용하기 위해 다음과 같은 알고리즘을 제시한다. 사용되는 용어는 Table 2.와 같으며, 단계별 매개변수 선정(4.2절)에서 사용된 용어는 동일하게 사용한다. 주요 용어는 다음과 같다.

*CK_Fields*는 결합키 생성항목인 속성의 집합이며, *FBF_Len*는 적합한 비트 식별(4.2.2절) 과정 이후 생성된 *CK_Fields*에 포함된 각 결합키 생성 속성 별 블룸 필터의 길이이다. *RBF_Len*는 자료 기반 블룸 필터 생성(4.2.4절) 과정 이후 생성된 모든 자료의 블룸 필터 길이이며, FBF_DS_i는 i번 데이터셋의 결합키 생성 속성 별 블룸 필터 결과, RBF_DS_i는 i번 데이터셋의 최종 결합키이다.

그리고 *weights*는 속성 가중치 선정(4.2.3절)을 통해 계산된 각 속성 별 가중치이며, *DetailCheck*는 임계값(*threshold*) 부근의 Dice Coefficient (*Dice*)를 갖는 경우(*scope*에 포함되는 경우), 자세한 조사를 수행하는 함수이다. *Combine*은 *CK_Fields*에 포함된 결합키 생성 속성을 제외하고 자료 쌍을 결합하는 함수이다.

*CK_Fields*에 포함된 결합키 생성항목인 속성에 대한 블룸 필터 계산 과정은 Fig. 7.와 같다. 기본 모델의 적합한 비트 식별 과정까지 거친 뒤 계산된 *FBF_Len*에 맞는 블룸 필터를 계산하는 과정으로 결합키 생성을 위한 초기 단계이다.

최종적인 결합키 생성은 Fig. 8.와 같다. 속성 가중치 선정 과정을 통해 계산된 *weights*와 각 속성의 블룸 필터에 대한 치환인 *F_Perm*을 이용하여 자료 기반 블룸 필터 생성을 진행한다. 그리고 기존 결합키 생성 방법의 Salt를 대체하기 위해 자료에 대한 치환인 *R_Perm*을 이용하여 최종 결합키를 생성한다.

데이터셋 DS_1,DS_2를 결합한 데이터셋 *CDS*는

Table 2. Notation for algorithms

Notation	
DS_1, DS_2	first dataset and second dataset what want to combine
CDS	combined dataset
$DS_i Len$	the number of records in DS_i
CK_Fields	the fields used to generate the combination key
EBI	"Eligible bit identification" stage
FBF_Len	bloom filter length by each field after *EBI*
RBF_Len	bloom filter length of all record after "RBF parameterization & generation"
FBF_DS_i	dataset with bloom filter by each field in *CK_Fields* after *EBI*
RBF_DS_i	dataset with combination key after "RBF permutation"
FBF_BF	bloom filter of each field in FBF_DS_i
ck	the combination key of each record in DS_i
weights	the fields' weight after "field weighting"
Dice	dice coefficient of record pair
Combine	combine records except for fields in *CK_Fields*
DetailCheck	detail check between records
F_Perm	permutation used when combine *FBF_BF*
R_Perm	permutation used when "RBF permutation"
scope	ambiguous scope requiring *DetailCheck*
H_i	ith hash used for bloom filter
threshold	the threshold that can combine the records
$D[i], d[i]$	ith record / value of dataset D / data d
$D[i].field$	the value corresponding *field* of $D[i]$

Algorithm1 :
FBF Combination Key Generation

Input: $DS_1, DS_2, k, p, q, CK\_Fields$
Output: FBF_DS_1, FBF_DS_2
1. $FBF\_Len = EBI(DS_1, DS_2, k, p, q, CK\_Fields)$
2. Create FBF_DS_1, FBF_DS_2
3. For $i=1$ to 2 :
4. For $idx=1$ to DS_iLen :
5. $record = DS_i[idx]$
6. For $field$ in CK_Fields :
7. $len = FBF\_Len.field$
8. $data = record.field$
9. $bf = \underbrace{0.....0}_{len}$
10. For $h=1$ to k :
11. $bf[H_h(data)\%len]=1$
12. End For
13. $FBF\_DS_i[idx].field = bf$
14. End For
15. End For
16. return FBF_DS_1, FBF_DS_2

Fig. 7. Algorithm1 : FBF Combination Key Generation

Algorithm2 :
RBF Combination Key Generation

Input:
$FBF\_DS_1, FBF\_DS_2, weights, F\_Perm, R\_Perm, CK\_Fiel$
Output: RBF_DS_1, RBF_DS_2
1. Create RBF_DS_1, RBF_DS_2
2. $RBF\_Len = 0$
3. For $field$ in CK_Fields :
4. $w = weights.field$
5. $RBF\_Len = \max(RBF\_Len, FBF\_Len \times (1/w))$
6. End For
7. For $i=1$ to 2 :
8. For $idx=1$ to DS_iLen :
9. rbf = empty bit string
10. For $field$ in CK_Fields :
11. $len = RBF\_Len \times weights.field$
12. concatenate
$F\_Perm(FBF\_DS_i[idx].field, len)$ to rbf
13. End For
14. $RBF\_DS_i[idx] = R\_Perm(rbf)$
15. End For
16.End For
17.return RBF_DS_1, RBF_DS_2

Fig. 8. Algorithm2 : RBF Combination Key Generation

Algorithm3 : Dataset Combination

Input: $RBF\_DS_1, RBF\_DS_2, CK\_Fields, threshold, scope$
Output: CDS
1. Create CDS
2. For $idx_1=1$ to DS_1Len :
3. For $idx_2=1$ to DS_2Len :
4. $ck_1 = RBF\_DS_1[idx_1]$
5. $ck_2 = RBF\_DS_2[idx_2]$
6. If $Dice(ck_1, ck_2) > threshold$ and not in $scope$:
7. $CDS[idx_1, idx_2] =$
$Combine(DS_1[idx_1], DS_2[idx_2], CK\_Fields)$
8. Else If $Dice(ck_1, ck_2)$ is in $scope$:
9. If pass $DetailCheck(DS_1[idx_1], DS_2[idx_2])$:
10. $CDS[idx_1, idx_2] =$
$Combine(DS_1[idx_1], DS_2[idx_2], CK\_Fields)$
11. End If
12. End If
13. End For
14.End For
15.For $idx_1=1$ to DS_1Len :
16. If idx_1 not in CDS :
17. $CDS[idx_1, \bullet] = Combine(DS_1[idx_1], CK\_Fields)$
18. End If
19.End For
20.For $idx_2=1$ to DS_2Len :
21. If idx_2 not in CDS :
22. $CDS[\bullet, idx_2] = Combine(DS_2[idx_2], CK\_Fields)$
23. End If
24.End For
25.return CDS

Fig. 9. Algorithm3 : Dataset Combination

$RBF_DS_1, RBF_DS_2, threshold$를 이용하여 Fig. 9.와 같이 생성된다. 결합키(ck) 간의 $Dice$값을 비교하여 임계값보다 크고 임계값 부근의 범위인 $scope$에 포함되지 않아 $DetailCheck$가 필요하지 않은 경우, 해당 자료 쌍은 결합되어 CDS에 추가된다. $scope$에 포함되어 $DetailCheck$가 필요한 자료 쌍의 경우, $DetailCheck$를 통과하였을 때 CDS에 추가된다. 이외의 경우, 결합될 수 있는 자료 쌍이 존재하지 않으므로 개별 자료로써 CDS에 추가된다.

V. 연구 결과

5.1 실험 개요

본 연구를 위한 실험에 사용된 가명정보 결합 모

델은 기본 모델(4.1절)과 연구 모델(4.3절)을 이용한다. “속성 기반 블룸 필터 생성” 및 “Algorithm1” 과정 시, [5]에서 제시된 개인정보보호를 이루는 최적의 블룸 필터 매개변수를 사용한다. 데이터를 2글자 단위로 분리하며($q=2$), 블룸 필터 인코딩 bit의 0, 1 비율을 50%($p=0.5$), 15개의 해시 함수를 사용하여($k=15$) Dynamic FBF를 실시한다.

블룸 필터에 적용된 15개의 해시 함수는 두 개의 암호학적 해시 함수를 이용해 동일한 성능을 갖도록 (10) 식을 이용해 생성한다[13]. m은 블룸 필터의 길이, i는 $1 \le i \le 15$이며, 실험에 사용된 암호학적 해시 함수는 h_1:SHA256, h_2:KECCAK256이다.

$$g_i(x) = h_1(x) + ih_2(x) \pmod m \quad (10)$$

“속성 가중치 선정” 시, scaling factor p에 대해 $p \in \{0.1, 0.15, 0.2, 0.25\}$의 평균 sim_w를 계산한다. 평균 sim_w가 0.9 이상인 경우, 일치하는 데이터로 판단하여 M_i에 포함한다. 평균 sim_w가 0.9미만인 경우, 불일치하는 데이터로 판단하여 U_i에 포함한다. 그리고 “자료 기반 블룸 필터 생성” 및 “Algorithm2” 과정에서 각 속성별 동일한 랜덤 치환을 적용하며, “자료 기반 블룸 필터 치환” 및 “Algorithm2” 과정에서는 분석하고자 하는 두 개의 데이터셋에 동일한 랜덤 치환을 적용한다.

실험의 결과와 “Algorithm3”를 바탕으로 개인정보가 갱신된 데이터셋 간의 가명정보 결합 가능성을 판단한다.

5.2 실험 시나리오

실험은 총 9가지의 시나리오로 구성된다. 각 시나리오 모두 2개의 데이터셋(데이터셋 1, 2)을 비교하며, 각 시나리오별 데이터셋은 모두 “이름”, “주소”, “전화번호”, “생년월일” 속성을 가진다. 또한 “속성 가중치 선정” 과정에 따라 시나리오별로 다른 속성 가중치를 가진다.

“주소 변경”, “전화번호 변경” 등 동일 인물임을 분석하고자 하는 경우, 데이터셋 1, 2의 동일 인덱스는 동일 인물임을 의미한다. “동명이인”, “다른 인물이지만 전화번호가 동일한 경우” 등 다른 인물임을 분석하고자 하는 경우, 데이터셋 1, 2의 동일 인덱스는 특정 속성에 같은 데이터를 갖지만 다른 인물임을 의미한다. 시나리오별 상세 내용은 다음과 같다.

Table 3. Scenarios for study

Scenario	Content
①	change to arbitrary address
②	change an address in the same region
③	change to arbitrary phone number
④	change a phone number maintaining the last 4 digits
⑤	rename
⑥	namesake
⑦	different people, but the phone number is the same
⑧	mixed data for the same person
⑨	all different people

① 주소가 변경된 경우

주소가 변경되었지만 데이터베이스에 아직 주소 변경 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 주소는 이전 주소이며, 데이터셋 2의 주소는 갱신된 주소이다.

② 주소가 변경된 경우 (동일 지역)

동일 지역으로 주소가 변경되었지만 데이터베이스에 아직 주소 변경 내용이 적용되지 않은 시나리오이다. 동일 지역의 기준은 “OO시 OO구” 또는 “OO도 OO시”까지 일치하는 경우 동일 지역으로 가정한다.

①, ② 시나리오 모두 데이터셋 1, 2의 동일 인덱스의 자료는 “주소” 속성의 데이터가 다르며, “이름”, “전화번호”, “생년월일” 속성의 데이터는 동일하다.

③ 전화번호가 변경된 경우

전화번호가 변경되었지만 데이터베이스에 아직 전화번호 변경 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 전화번호는 이전 전화번호이며, 데이터셋 2의 전화번호는 갱신된 전화번호이다.

④ 전화번호가 변경된 경우 (뒷 4자리 동일)

전화번호가 가운데 4자리만 변경되었지만 데이터베이스에 아직 전화번호 변경 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 전화번호는 이전 전화번호이며, 데이터셋 2의 전화번호는 갱신된 전화번호이다.

③, ④ 시나리오의 전화번호는 모두 휴대전화번호를 기준으로 한다. 그리고 데이터셋 1, 2의 동일 인

덱스의 자료는 “전화번호” 속성의 데이터가 다르며, “이름”, “주소”, “생년월일” 속성의 데이터는 동일하다.

⑤ 개명의 경우

개명을 하였지만 데이터베이스에 아직 개명 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 이름은 이전 이름이며, 데이터셋 2의 이름은 개명된 이름이다. 개명은 성은 동일하며, 이름만 개명된 경우로 가정한다. 데이터셋 1, 2의 동일 인덱스의 자료는 “이름” 속성의 데이터가 다르며, “주소”, “전화번호”, “생년월일” 속성의 데이터는 동일하다.

⑥ 동명이인의 경우

데이터셋 1, 2의 동일 인덱스의 자료는 동명이인으로 “이름” 속성의 데이터가 동일하다. 그리고 “주소”, “전화번호”, “생년월일” 속성의 데이터는 모두 다르다.

⑦ 다른 인물이지만 전화번호 동일한 경우

이전 전화번호 소유자의 변경된 전화번호를 데이터베이스에 갱신하지 않은 경우이다. 데이터셋 1, 2의 동일 인덱스의 자료는 다른 인물이며 “전화번호” 속성의 데이터가 동일하다. 그리고 “이름”, “주소”, “생년월일” 속성의 데이터는 모두 다르다.

⑧ 동일 인물의 데이터가 섞인 경우

① ~ ⑤ 시나리오의 모든 데이터가 섞인 경우로서 데이터셋 1, 2에 대해 특정 속성만이 갱신된 경우가 아닌 임의의 속성이 갱신된 경우이다. 데이터셋 1, 2의 동일 인덱스의 자료는 동일 인물이다.

⑨ 모든 정보가 다른 경우

“이름”, “주소”, “전화번호”, “생년월일” 속성이 모두 다른 데이터셋 1, 2에 대한 경우이다. 데이터셋 1, 2의 자료는 모두 동일하지 않은 인물이다.

5.3 데이터셋 구성

기본 데이터는 공공데이터포털의 “건강보험심사평가원 요양기관 개설 현황” 데이터베이스를 활용하였다[14]. 건강보험심사평가원에서 제공하는 2021년 기준 운영 중인 요양기관 개설 현황 공공데이터로 총 99205건의 데이터를 제공한다. 해당 데이터베이스에는 “요양기관명”, “우편번호”, “주소”, “전화번호”, “개설일자” 등의 속성을 가진다.

실험에 사용할 데이터셋 구성을 위해 데이터베이스의 속성 중 “요양기관명”, “주소”, “전화번호”, “개설일자”를 사용하였다. “요양기관명”은 “이름” 속성으로 변경하였으며, “개설일자”는 “생년월일” 속성으로 변경하였다.

“이름” 속성의 데이터는 “성”과 “이름”으로 구분하여 공공데이터를 활용해 가명을 생성하였다. “성”은 통계청의 “통계지리정보서비스”에서 제공하는 “성씨·본관별 인구” 데이터를 활용하였다[15]. “이름”은 대한민국 법원의 “전자가족관계등록시스템”의 “상위 출생신고 이름 현황” 데이터를 활용하였다[16].

“성씨·본관별 인구” 데이터는 2015년까지 전국 성씨에 대한 총 조사이다. 이 중 Table 4.와 같이 전국 단위 상위 50개의 성씨를 비율에 맞게 99205건을 생성하였다. “상위 출생신고 이름 현황” 데이터는 2008년부터 현재까지 출생신고된 이름 순위 데이터이다. 이 중 Table 5.와 같이 상위 1000개의 이름을 비율에 맞게 99205건을 생성하였다. 생성된 99205건의 성씨와 이름을 랜덤하게 매칭하여 “이름” 속성의 데이터를 생성하였으며, “요양기관명” 속성의 데이터를 “이름” 속성의 데이터로 변경하였다.

“전화번호” 속성은 휴대전화번호 형식인 “OOOO-OOOO”의 형식으로 일치시켰다. 기존 “전화번호” 속성의 데이터가 위 형식과 맞지 않을 시, 가운데 4자리는 “1001”~“9999” 중, 뒷 4자리는 “0001”~“9999” 중 랜덤하게 선택하였다.

“건강보험심사평가원 요양기관 개설 현황”의 “주소” 속성의 99205건 데이터 모두 도로명 주소로 구축되

Table 4. The nation's top 50 family names

Rank	Family Name	Counts
1	Kim	10689959
2	Lee	7306828
~		
50	Gong	91869

Table 5. Top 1000 names born since 2008

Rank	Name	Counts
1	Ji Woo	44128
2	Min Jun	39435
~		
1000	Si Jin	775

어 있다. “OO동”과 같은 부가 정보는 제외하여 데이터셋을 구성하였다.

따라서 실험에 사용된 데이터베이스는 1번 자료 “이름: 백재이, 주소: 경기도 시흥시 오이도어시장로 19, 전화번호: 3371-1030, 생년월일: 2013-04-30”부터 99205번 자료 “이름: 이다인, 주소: 경기도 연천군 백학면 두일로 157, 전화번호: 6768-5010, 생년월일: 1984-10-10”까지이다.

실험을 위해 500건의 자료로 이루어진 데이터셋 1, 2를 각 시나리오별로 구성하였다. 또한 도출된 임계값이 올바른지 확인하기 위한 100건의 자료로 이루어진 검증 데이터셋 1, 2를 구성하였다.

①, ② 시나리오의 경우, 데이터셋 2의 “주소” 속성의 데이터는 데이터셋 1의 “주소” 속성의 데이터에 사용되지 않은 98605건의 “주소” 속성의 데이터 중 랜덤하게 선택하여 구성하였다. ② 시나리오의 경우, 98605건의 “주소” 속성의 데이터 중 동일 지역 조건에 맞는 데이터를 랜덤하게 선택하였다.

③, ④ 시나리오의 경우, 데이터셋 2의 “전화번호” 속성의 데이터는 데이터셋 1의 “전화번호” 속성의 데이터에 사용되지 않은 98605건의 “전화번호” 속성의 데이터 중 랜덤하게 선택하여 구성하였다. ④ 시나리오의 경우, 98605건의 “전화번호” 속성의 데이터 중 가운데 4자리만을 랜덤하게 선택하였다.

⑤ 시나리오의 경우, 데이터셋 2의 “이름” 속성의 데이터는 데이터셋 1의 “이름” 속성의 데이터에 사용되지 않은 98605건의 “이름” 속성의 데이터 중 성을 제외한 이름만을 랜덤하게 선택하였다.

⑥ 시나리오의 경우, 데이터셋 1, 2의 “이름” 속성의 데이터만이 일치하며, ⑦ 시나리오의 경우, 데이터셋 1, 2의 “전화번호” 속성의 데이터만이 일치한다.

⑧ 시나리오의 경우, ① ~ ⑤ 시나리오에 사용된 자료 중, 각 시나리오별로 100건의 자료를 랜덤하게 선택하여 데이터셋 1, 2을 구성하였다. 검증 데이터셋은 각 시나리오별로 20건의 자료를 랜덤하게 선택하여 검증 데이터셋 1, 2를 구성하였다.

⑨ 시나리오의 경우, 99205건의 자료 중 랜덤하게 1200건의 자료를 선택하여 500건의 자료로 구성된 데이터셋 1, 2와 100건의 자료로 구성된 2개의 검증 데이터셋 1, 2를 구성하였다.

각 시나리오별 속성 가중치 계산 결과는 Table 6.와 같다. ⑧, ⑨ 시나리오의 경우, 특정 속성을 기준으로 자료에 대한 일치, 불일치를 판단할 수 없으므로 25%의 균등한 속성 가중치를 가진다.

Table 6. Field weights by scenarios

	Weights			
Scenario	Name	Address	Mobile	Birth
①	30%	10%	33%	27%
②	30%	7%	36%	27%
③	30%	32%	11%	27%
④	30%	35%	8%	27%
⑤	11%	30%	34%	25%
⑥	13%	26%	28%	33%
⑦	27%	31%	10%	32%
⑧	25%	25%	25%	25%
⑨	25%	25%	25%	25%

① ~ ⑨ 시나리오의 데이터셋에 포함된 자료의 예시는 다음과 같다. 데이터셋 1, 2 순으로 “이름”, “전화번호”, “주소”, “생년월일”을 의미하며, ① ~ ⑤, ⑧ 시나리오의 예시는 동일 인물, ⑥, ⑦, ⑨ 시나리오의 예시는 다른 인물을 나타낸다.

①: “정지호, 서울특별시 강서구 수명로 76, 7569-3333, 2008-04-23”, “정지호, 전라북도 완주군 화산면 화산로 866, 7569-3333, 2008-04-23”

②: “김소윤, 부산광역시 동래구 아시아드대로 191, 9592-5303, 2019-05-01”, “김소윤, 부산광역시 동래구 충렬대로348번길 23, 9592-5303, 2019-05-01”

③: “김은서, 충청남도 천안시 서북구 월봉로 83, 1190-8986, 2000-03-03”, “김은서, 충청남도 천안시 서북구 월봉로 83, 2589-5554, 2000-03-03”

④: “장승우, 서울특별시 동작구 상도로 146, 8638-8112, 1993-12-31”, “장승우, 서울특별시 동작구 상도로 146, 3149-8112, 1993-12-31”

⑤: “이하준, 인천광역시 계양구 계산로 89, 2642-6698, 2020-06-30”, “이시은, 인천광역시 계양구 계산로 89, 2642-6698, 2020-06-30”

⑥: “임효린, 서울특별시 송파구 양재대로 932, 5632-5487, 2010-08-09”, “임효린, 서울특별시 마포구 마포대로 68, 8909-7575, 2021-11-09”

⑦: “이예준, 서울특별시 강남구 압구정로 151, 9041-1101, 2015-01-30”, “전태연, 전라북도 남원시 주천면 정령치로 55, 9041-1101, 1995-01-01”

⑧: “김다온, 서울특별시 광진구 용마산로 8, 8730-8575, 2005-03-02”, “김윤서, 서울특별시 광진구 용마산로 8, 8730-8575, 2005-03-02”

⑨: “이형우, 충청남도 공주시 번영1로 87, 5410-0075, 2004-05-27”, “윤시원, 부산광역시 부산진구 중앙대로 736, 7269-9778, 2011-05-23”

5.4 실험 결과

5.4.1 실험 시나리오 순서

우선 특정 속성 기준이 없는 시나리오인 “⑧ 동일 인물의 데이터가 섞인 경우”, “⑨ 모든 정보가 다른 경우”에 대한 결과를 분석한다. 해당 결과를 토대로 임의의 데이터에 대한 임계값을 계산한다. 다음으로 ⑧, ⑨ 시나리오를 토대로 계산된 임계값을 기반으로 특정 속성을 기준으로 잡을 수 있는 나머지 ① ~ ⑦ 시나리오에 대한 임계값을 계산한다. 해당 임계값을 토대로 실험 결과에 따른 최종 결론을 도출한다.

그래프에 표현된 “(A, B) : C”는 데이터셋 1의 A번 자료와 데이터셋 2의 B번 자료의 Dice Coefficient가 C임을 의미한다. 동일인물임을 판단하는 경우, 동일 인물 자료 쌍의 최소 Dice Coefficient와 다른 인물 자료 쌍의 최대 Dice Coefficient를 그래프에 표시한다. 다른 인물임을 구분하는 경우, 다른 인물 자료 쌍의 최대 Dice Coefficient를 그래프에 표시한다. 그래프의 x축은 데이터셋 1의 인덱스이며, y축은 데이터셋 1의 A번 자료와 데이터셋 2의 B번 자료의 Dice Coefficient이다.

5.4.2 특정 속성 기준이 없는 시나리오

특정 속성 기준이 없는 시나리오의 경우, “이름”, “주소”, “전화번호”, “생년월일” 속성의 가중치를 모두 25%로 가진다. ⑧ 시나리오에 대한 결과는 Fig. 10.과 같다. 동일 인물 자료 쌍에 대한 최소 Dice Coefficient는 데이터셋 1, 2의 204번 자료로 약 0.8269이다. 다른 인물 자료 쌍에 대한 최대 Dice Coefficient는 데이터셋 1의 282번 자료와 데이터셋 2의 155번 자료로 약 0.7974이다.

해당 결과를 바탕으로 특정 속성을 기준으로 잡을 수 없을 때 임계값을 0.8로 잡을 수 있다. 이를 검증 데이터셋에 적용한 결과는 Fig. 11.과 같다.

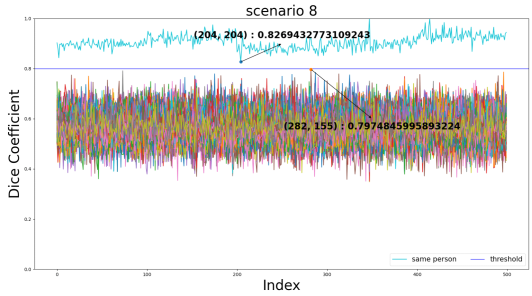

Fig. 10. Result of ⑧ scenario

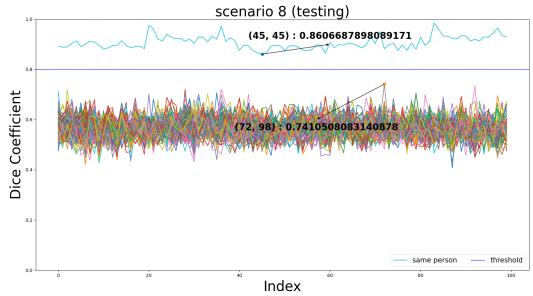

Fig. 11. Testing result of ⑧ scenario

⑧ 시나리오 검증 데이터셋에 대한 동일 인물 자료 쌍의 최소 Dice Coefficient는 검증 데이터셋 1, 2의 45번 자료로 약 0.8606이다. 다른 인물 자료 쌍의 최대 Dice Coefficient는 검증 데이터셋 1의 72번 자료와 검증 데이터셋 2의 98번 자료로 약 0.7410이다. 임계값으로 정한 0.8 이상의 다른 인물 자료 쌍의 Dice Coefficient가 존재하지 않으며, 동일 인물 자료 쌍의 Dice Coefficient가 모두 0.8 이상이므로 임계값 0.8은 유효한 결과로 추정할 수 있다.

이를 바탕으로 ⑨ 시나리오에 임계값 0.8을 적용한 결과는 Fig. 12.와 같다. 총 4개의 다른 인물 자료 쌍이 임계값으로 정한 0.8 이상의 결과를 가진다. 다른 인물 자료 쌍에 대한 최대 Dice Coefficient는 데이터셋 1의 14번 자료와 데이터셋 2의 480번 자료로 약 0.8259이다.

총 250,000개의 다른 인물 자료 쌍 중 4개의 다른 인물 자료 쌍만이 임계값 0.8 이상의 Dice Coefficient를 가지므로 0.0016%의 확률로 임계값 이상의 결과를 가짐을 알 수 있다. ⑨ 시나리오의 검증 데이터셋에 대한 결과는 Fig. 13.과 같다.

⑨ 시나리오의 검증 데이터셋에 대한 다른 인물 자료 쌍의 최대 Dice Coefficient는 검증 데이터

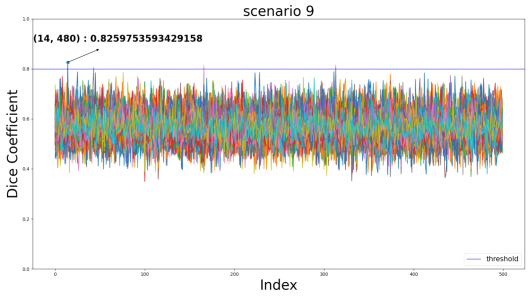

Fig. 12. Result of ⑨ scenario

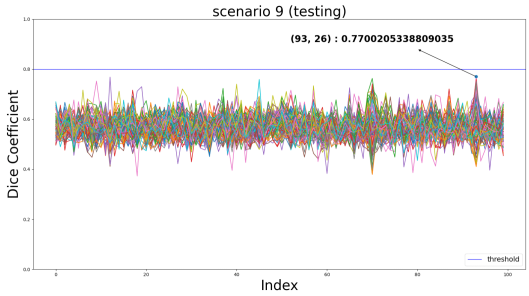

Fig. 13. Testing result of ⑨ scenario

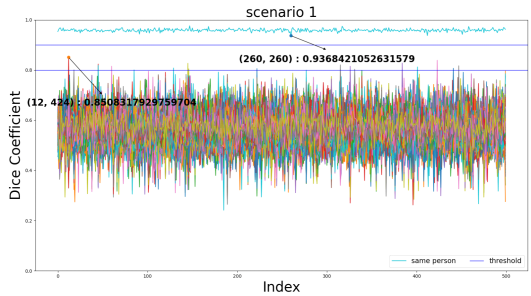

Fig. 14. Result of ① scenario

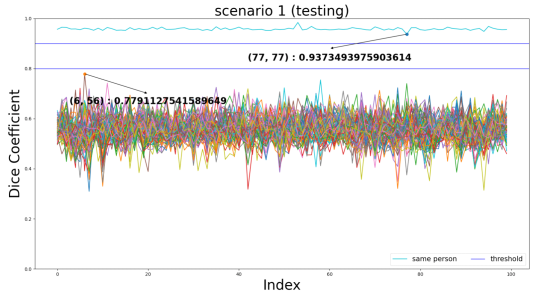

Fig. 15. Testing result of ① scenario

셋 1의 93번 자료와 검증 데이터셋 2의 26번 자료로 약 0.7700이다. 모든 자료 쌍에 대해 임계값 0.8 이상의 결과를 갖지 않는다. 따라서 특정 속성 기준이 없는 가명정보 결합의 경우, 임계값 0.8에 따라 가명정보 결합이 가능하다.

5.4.3 특정 속성 기준 동일 인물 결합 시나리오

"① 주소가 변경된 경우", "② 주소가 변경된 경우 (동일 지역)", "③ 전화번호가 변경된 경우", "④ 전화번호가 변경된 경우 (뒷 4자리 동일)", "⑤ 개명의 경우" 시나리오는 특정 속성을 기준으로 동일 인물에 대한 가명정보 결합이 가능한 시나리오이므로 앞서 정한 임계값 0.8 이상의 임계값을 추가로 정하고자 한다.

① 시나리오는 "주소" 속성의 가중치가 낮음으로써, "주소" 이외의 속성에 따라 가명정보 결합이 가능하다. ① 시나리오의 결과는 Fig. 14.과 같다.

동일 인물 자료 쌍의 최소 Dice Coefficient는 데이터셋 1, 2의 260번 자료로 약 0.9368이다. 다른 인물 자료 쌍의 최대 Dice Coefficient는 데이터셋 1의 12번 자료와 데이터셋 2의 424번 자료로 약 0.8508이다. 다른 인물 자료 쌍 총 19개는 앞서 정한 임계값 0.8 이상의 Dice Coefficient를 가지므로, 약 0.0076%의 확률로 임계값 0.8 이상의 결과를 가짐을 알 수 있다.

따라서 동일 인물 자료 쌍의 모든 Dice Coefficient가 0.9 이상, 다른 인물 자료 쌍의 Dice Coefficient가 모두 0.9 미만이므로 특정 속성 기준의 가명정보 결합 시 임계값을 0.9로 정할 수 있다. 이를 검증 데이터셋에 적용한 결과는 Fig. 15.와 같다.

① 시나리오의 검증 데이터셋에 대한 다른 인물 자료 쌍의 최대 Dice Coefficient는 검증 데이터셋 1의 6번 자료와 검증 데이터셋 2의 56번 자료로 약 0.7791이다. 동일 인물 자료 쌍의 최소 Dice Coefficient는 검증 데이터셋 1, 2의 77번 자료로 약 0.9373이다. 이를 바탕으로 특정 속성 기준의 가명정보 결합 시 임계값 0.9가 유효함을 알 수 있다.

임계값 0.8, 0.9에 대한 ② ~ ⑤ 시나리오에 대한 결과는 Fig. 16. ~ Fig. 19.와 같다. ② ~ ⑤ 시나리오에 대한 다른 인물 자료 쌍의 최대 Dice Coefficient는 모두 임계값 0.9 미만의 값을 가진다. 그리고 임계값 0.8 이상의 값을 갖는 다른 인물 자료 쌍은 ② 시나리오 18개, ③ 시나리오 27개, ④ 시나리오 6개, ⑤ 시나리오 3개이다. 따라서 ③ 시

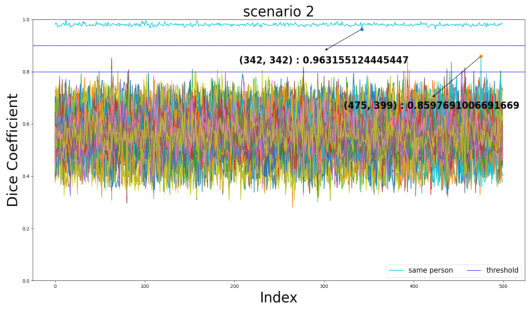

Fig. 16. Result of ② scenario

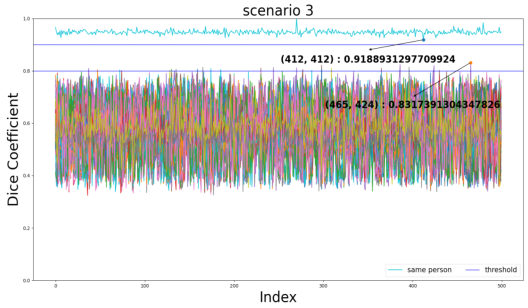

Fig. 17. Result of ③ scenario

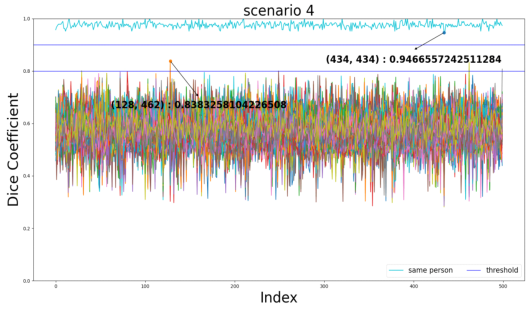

Fig. 18. Result of ④ scenario

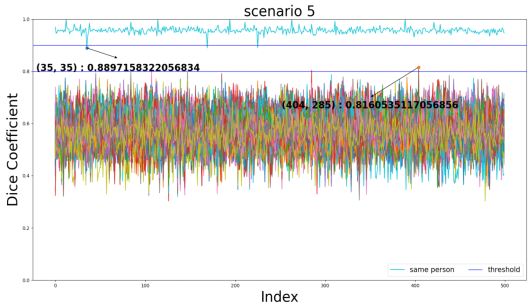

Fig. 19. Result of ⑤ scenario

나리오와 같이 최대 약 0.0108%의 확률로 임계값 0.8 이상의 Dice Coefficient를 얻을 수 있으며, 임계값 0.9 이상의 결과는 존재하지 않는다.

동일 인물 자료 쌍의 최소 Dice Coefficient 중 임계값 0.9 미만의 값을 갖는 경우는 ⑤ 시나리오의 데이터셋 1, 2의 35번 자료로 약 0.8897이다. 또한 ⑤ 시나리오의 동일 인물 자료 쌍 중 3개는 임계값 0.9 미만의 Dice Coefficient를 가진다. 따라서 0.6%의 확률로 임계값 0.9 미만의 동일 인물 자료 쌍이 존재하며, 임계값 0.9에 대해 특정 속성 기준의 동일 인물의 가명정보 결합이 가능하다.

5.4.4 특정 속성 기준 다른 인물 구분 시나리오

"⑥ 동명이인의 경우", "⑦ 다른 인물이지만 전화번호 동일한 경우"의 시나리오는 특정 속성을 기준으로 다른 인물임을 구분하는 시나리오이다. 데이터셋 1, 2의 동일 인덱스 자료는 특정 속성의 데이터를 제외하고 모두 다른 값을 가진다. 따라서 동일 인덱스 자료 간의 Dice Coefficient를 구함으로써 앞서 정한 임계값 0.8, 0.9 미만의 결과를 만족하는지 확인한다. ⑥, ⑦ 시나리오에 대한 결과는 Fig. 20., Fig. 21.과 같다.

⑥ 시나리오의 다른 인물 자료 쌍의 최대 Dice

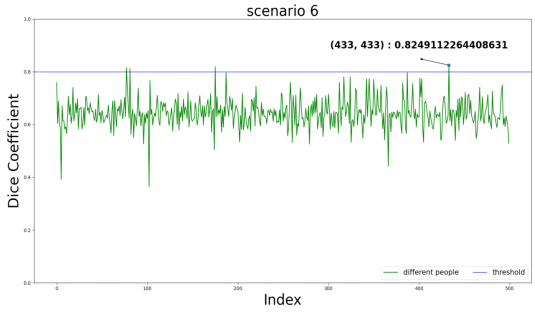

Fig. 20. Result of ⑥ scenario

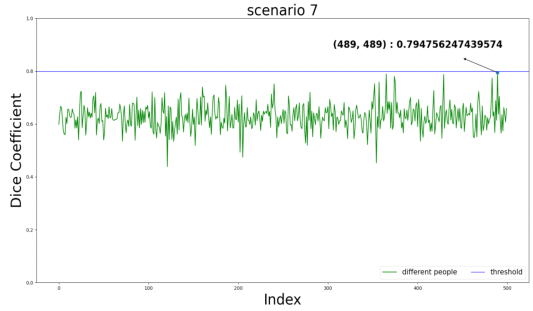

Fig. 21. Result of ⑦ scenario

Table 7. Result of each scenarios

Ch.	Scn.	$\#SR$	$\#DR$	$\#\{SR \le t_1\}$	$\#\{SR \le t_2\}$	$\#\{DR \ge t_1\}$	$\#\{DR \ge t_2\}$	$Min\{SR\}$	$Max\{DR\}$
5.4.2	⑧	500	249,500	0	•	0	0	0.8269	0.7974
	⑨	•	250,000	•	•	4	0	•	0.8259
5.4.3	①	500	249,500	0	0	19	0	0.9368	0.8508
	②	500	249,500	0	0	18	0	0.9631	0.8597
	③	500	249,500	0	0	27	0	0.9188	0.8317
	④	500	249,500	0	0	6	0	0.9466	0.8383
	⑤	500	249,500	0	3	3	0	0.8897	0.8160
5.4.4	⑥	•	500	•	•	4	0	•	0.8249
	⑦	•	500	•	•	0	0	•	0.7974

Ch. : chapter, Scn. : scenario, t_1 : threshold 0.8, t_2 : threshold 0.9
SR/DR : set of same person / different people record pair's dice coefficient
: the number of elements in the set, Min, Max : minimum, maximum value of the set

Coefficient는 데이터셋 1, 2의 433번 자료로 약 0.8249이다. ⑦ 시나리오의 다른 인물 자료 쌍의 최대 Dice Coefficient는 데이터셋 1, 2의 489번 자료로 약 0.7947이다.

⑥ 시나리오의 경우, 임계값 0.8 이상의 Dice Coefficient를 갖는 다른 인물 자료 쌍은 총 4개이며, 0.8%의 확률로 임계값 0.8 이상의 다른 인물 자료 쌍이 존재한다. ⑦ 시나리오의 경우, 임계값 0.8 이상의 Dice Coefficient를 갖는 가른 인물 자료 쌍은 존재하지 않는다. 임계값 0.9 이상의 Dice Coefficient를 갖는 다른 인물 자료 쌍은 ⑥, ⑦ 시나리오 모두 존재하지 않는다.

5.4.5 실험 결과

실험에 사용된 9가지의 시나리오에 대한 동일 인물 자료 쌍의 임계값 이하에 대한 결과 및 최소 Dice Coefficient와 다른 인물 자료 쌍의 임계값 이상에 대한 결과 및 최대 Dice Coefficient는 Table 7.와 같다.

그리고 "자료 쌍의 Dice Coefficient가 임계값 이상이면 동일 인물이다."라는 가설 1(hypothesis 1), "자료 쌍의 Dice Coefficient가 임계값 미만이면 다른 인물이다."라는 가설 2(hypothesis 2)에 대해 실험 결과를 적용하면 다음과 같다.

⑧ 시나리오와 같이 가명정보 결합 과정에서 특정 속성을 이용할 수 없는 경우, 임계값 0.8에 대해 가설 1, 2를 100% 만족한다. ① ~ ⑤ 시나리오와 같이 특정 속성을 기준으로 가명정보 결합이 가능한 경우, 임계값 0.9에 대해 가설 1을 최소 99.4% 만족하며, 가설 2를 최소 99.9892% 만족한다.

⑨ 시나리오와 같이 다른 인물 자료 구분에 특정 속성을 이용할 수 없는 경우, 임계값 0.8에 대해 가설 2를 99.9984% 만족한다. ⑥, ⑦ 시나리오와 같이 특정 속성을 기준으로 다른 인물 자료 구분이 가능한 경우, 임계값 0.9에 대해 가설 2를 최소 99.2% 만족한다. ⑥, ⑦, ⑨ 시나리오는 동일 인물 자료 쌍이 존재하지 않으므로 가설 1에 대한 검증은 불가능하다.

따라서 해당 결과를 통해 모든 시나리오에 대해 가설 1, 2를 99% 이상의 정확도로 만족함을 Table 8.와 같이 알 수 있으며, 특정 임계값을 기준으로 갱신된 데이터베이스 간의 가명정보 결합이 가능함을 확인할 수 있다. 그리고 임계값과 근사한 Dice Coefficient의 결과를 갖는 소수의 자료 쌍은 가설을 만족하지 않을 수 있으므로 추가적인 상세 분석이 필요하다.

Table 8. Accuracy for hypothesis 1, 2

Accuracy	Hypothesis 1 (%)		Hypothesis 2 (%)	
scenario	threshold		threshold	
	0.8	0.9	0.8	0.9
①	100	100	99.9924	100
②	100	100	99.9928	100
③	100	100	99.9892	100
④	100	100	99.9976	100
⑤	100	99.4	99.9988	100
⑥	•	•	99.2	100
⑦	•	•	100	100
⑧	100	•	100	100
⑨	•	•	99.9984	100

5.5 안전성 분석

본 연구에서 우리는 가명정보 처리 가이드라인의 가명정보 결합 및 반출 방법을 준수하였다. 즉, 기존의 시스템을 그대로 사용하면서도 데이터베이스의 갱신 시점이 다른 동일 개체를 찾아낼 수 있다.

다만, 기존의 시스템을 그대로 사용하기 위해 본 논문에서 우리가 제시한 방법을 사용하면 결합신청자들은 결합키관리기관으로부터 파라미터값들과 치환에 필요한 비밀 정보를 제공받기 때문에, 기존의 방법과 동일하게 결합키관리기관에 의한 사전 공격(dictionary attack)이 가능하다.

[17]에 따르면 유명 인사의 병원 기록이 환자의 동의 없이 접근되어지는 경우가 많은데, 예를 들어 Britney Spears의 경우 첫 아이를 낳았을 때의 기록과 정신 감정 기록이 동의 없이 열람되었다. 만약 결합신청자가 정신병원이었다면 공격자는 자신이 알고 있는 지인의 정보를 가지고 결합키를 생성해봄으로써 지인이 정신병원에 방문한 적이 있는지 알아낼 수 있을 것이다. 이와 동일한 방법으로 결합키관리기관은 결합신청자들이 생성한 결합키를 이용하여 자신이 원하는 사람에 대한 정보가 있는지를 확인해볼 수 있을 것이다.

따라서 결합키관리기관에 의한 사전 공격을 막기 위해서는 결합신청자들끼리 파라미터값들과 치환에 필요한 비밀 정보를 결합키관리기관을 배제한 채 공유하면 되나, 기존의 가명정보 처리 가이드라인에서 제시한 방법에 변화는 불가피하다. 따라서 결합키관리기관을 신뢰하는 경우, 우리가 제시한 방법을 사용하며, 결합키관리기관을 신뢰하지 않아 결합키관리기관으로부터의 사전 공격을 막고자 하는 경우, 필요한 비밀 정보를 결합신청자들끼리만 공유하도록 시스템을 수정할 수 있다.

VI. 결 론

현재 국내에서 데이터 결합 시 가명 처리는 필수적이며, 기존의 결합키 생성 기법을 사용하면 결합키 생성항목의 형태가 정확히 일치하는 경우에만 동일한 개체로 매핑이 가능하였다. 하지만 본 연구에서 우리가 제안한 기법을 활용하면 데이터베이스의 갱신 시점이 다르더라도 동일 개체라면 강건하게 매핑에 성공할 수 있음을 보였다. 특히, 시나리오별로 실험을 진행하였으며, 이에 따른 적합한 임계값 역시 제시하였다.

또한, 결합키 생성 시, 결합키관리기관으로부터 사전 공격이 우려되는 경우, 결합신청자 간의 합의된 치환을 사용함으로써 이를 방지할 수 있다.

우리의 결과를 통해 국내 데이터 결합 시 결합속성 관련 데이터 전처리 과정 및 해당 데이터의 갱신에 따른 결합 불가능성 등의 문제를 해결할 수 있을 것으로 기대한다.

References

[1] Hyeon-ju Noh, "MyData Business Status and Insurance Company Implications", Research Report, Korea Insurance Research Institute, 2021(4), June 2021.

[2] Personal Information Protection Commission, "Pseudonymization Guideline", Feb. 2020.

[3] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, "Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges", Handbook of Big Data Technologies, Springer, pp. 851-895, Feb. 2017.

[4] I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage", Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183-1210, Dec. 1969.

[5] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters", BMC Medical Informatics and Decision Making, vol. 9, no. 41, Aug. 2009.

[6] E.A. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom Filters for Secure Record Linkage", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 12, pp. 2956-2968, Dec. 2014.

[7] Young-im Bae, Hye-ri Shin, "Three Revised Act, The Beginning of the

Data Economy", Issue & Analysis, Gyeonggi Research Institute, 2020(405), Feb. 2020.

[8] B.H. Bloom, "Space/Time Trade-offs in Hash Coding with Allowable Errors", Communications of the ACM, vol. 13, no. 7, pp. 422-426, July 1970.

[9] W.E. Winkler, "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage", Statistical Research Division, Methodology and Standards Directorate, U.S. Bureau of the Census, Oct. 2000.

[10] P. Christen, T. Ranbaduge, D. Vatsalan, and R. Schnell, "Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage", IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 11, pp. 2164-2177, Nov. 2019

[11] A. Broder and M. Mitzenmacher, "Network Applications of Bloom Filters: A Survey", Internet Mathematics, vol. 1, no. 4, pp. 485-509, Dec. 2004.

[12] W.E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", Statistical Research Division, U.S. Bureau of the Census, 1990.

[13] A. Kirsch and M. Mitzenmacher, "Less Hashing, Same Performance: Building a Better Bloom Filter", Lecture Notes in Computer Science, ESA 2006, vol. 4168, pp. 456-467, 2006.

[14] Public Data Portal, "Status of establishment of nursing care institutions, Health Insurance Review & Assessment Service ", https://www.data.go.kr/data/15051057/fileData.do, Feb. 2022.

[15] Statistics Korea, "Family name census, Statistical Geographic Information Service", https://sgis.kostat.go.kr/statbd/family_01.vw, Jan, 2017.

[16] Court of Korea, "Top birth registration name status, Family Relations Registration System", https://stfamily.scourt.go.kr/st/StFrrStatcsView.do?pgmId=090000000025, 2022.

[17] J. Pepper, The Electronic Health Record for the Physician's Office: For Simchart for the Medical Office, 3rd Ed., Elsevier, Sep. 2019.

〈저자소개〉

장 호 빈 (Hobin Jang) 학생회원
2022년 2월: 서울시립대학교 수학과 졸업
2022년 3월~현재: 고려대학교 정보보호대학원 융합보안학과 석사과정
〈관심분야〉 정보보호, 데이터 보안, 프라이버시 향상 기술

노 건 태 (Geontae Noh) 종신회원
2008년 2월: 고려대학교 산업시스템정보공학과 졸업
2010년 2월: 고려대학교 정보경영공학과 석사
2014년 8월: 고려대학교 정보보호학과 박사
2014년 9월~2017년 2월: 고려대학교 정보보호연구원 박사후 연구원, 연구교수
2017년 2월~현재: 서울사이버대학교 빅데이터·정보보호학과 조교수
2020년 3월~현재: 서울사이버대학교 빅데이터·AI센터 센터장
〈관심분야〉 프라이버시 향상 기술, 암호 이론, 데이터 보안, 블록체인, 인공지능

정 익 래 (Ik Rae Jeong) 종신회원
1998년 2월: 고려대학교 전산학과 졸업
2000년 2월: 고려대학교 전산학과 석사
2004년 8월: 고려대학교 정보보호학과 박사
2006년 3월~2008년 2월: 한국전자통신연구원 암호기술연구팀 선임연구원
2008년 3월~현재: 고려대학교 정보보호대학원 교수
〈관심분야〉 암호 이론, 프라이버시 향상 기술, 데이터베이스 보안, 생체인증

천 지 영 (Ji Young Chun) 종신회원
1997년 2월: 이화여자대학교 수학과 졸업
2006년 2월: 고려대학교 정보보호학과 석사
2011년 8월: 고려대학교 정보경영공학과 박사
2012년 8월~2014년 3월: University of Illinois at Urbana-Champaign 박사후 연구원
2021년 2월~현재: 서울사이버대학교 빅데이터·정보보호학과 조교수
2022년 2월~현재: 서울사이버대학교 빅데이터·AI센터 부센터장
〈관심분야〉 데이터 보안, 인공지능, 연합학습, 프라이버시 향상 기술