



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0030895
(43) 공개일자 2024년03월07일

(51) 국제특허분류(Int. Cl.)
H04L 9/08 (2006.01) G06F 16/28 (2019.01)
G06F 17/15 (2006.01) G06F 17/18 (2006.01)
G06F 21/62 (2013.01)
(52) CPC특허분류
H04L 9/0861 (2013.01)
G06F 16/288 (2019.01)
(21) 출원번호 10-2022-0132902
(22) 출원일자 2022년10월17일
심사청구일자 2022년10월17일
(30) 우선권주장
1020220109703 2022년08월31일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정익래
서울특별시 광진구 아차산로69길 19, 904동 2205호(광장동, 현대아파트)
이연주
충청남도 천안시 서북구 백석로 136, 113동 1102호(백석동, 백석현대아파트)
(뒷면에 계속)
(74) 대리인
윤귀상

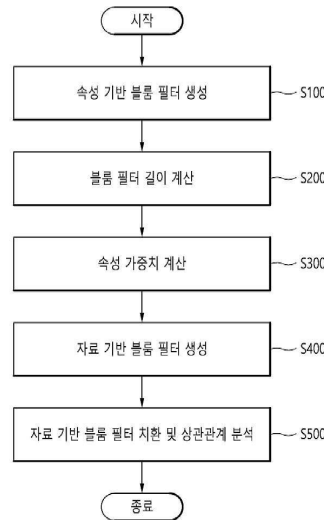
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 개인정보 보호를 위한 결합기 생성 시스템 및 방법

(57) 요약

본 발명은 결합기 생성 방법으로, 데이터를 속성에 기반하여 분류하고 상기 속성에 기반하여 분류된 데이터 각각에 대한 제1 블룸 필터를 생성하는 단계, 상기 속성에 기반하여 분류된 데이터를 글자 단위로 분리한 집합의 크기, 복수의 해시 함수들 및 블룸 필터의 인코딩 비트가 0이 나올 확률에 기초하여 상기 속성에 기반하여 분류된 데이터 각각에 대한 제1 블룸 필터의 길이를 계산하는 단계, 상기 속성에 기반하여 분류된 데이터 각각에 대하여 속성 가중치를 계산하는 단계, 상기 속성에 기반하여 분류된 데이터 각각에 대하여 상기 속성 가중치를 고려한 제2 블룸 필터를 생성하고 상기 제2 블룸 필터의 길이를 계산하는 단계 및 상기 속성에 기반하여 분류된 데이터 각각에 대한 치환을 수행하고, 상기 각 제2 블룸 필터에 대한 상관관계를 분석하는 단계를 포함하는 것을 특징으로 한다.

대표도 - 도1



(52) CPC특허분류

G06F 17/15 (2013.01)

G06F 17/18 (2013.01)

G06F 21/6218 (2013.01)

(72) 발명자

장호빈

서울특별시 동대문구 전농로38길 84, 303호(
회경동, 락랑빌딩)

천지영

서울특별시 성북구 오패산로16가길 40, 109동 304
호(하월곡동, 꿈의숲 푸르시오)

노건태

경기도 하남시 위례중양로 215, 6401동 306호(학암
동, 위례롯데캐슬)

명세서

청구범위

청구항 1

결합키 생성 방법으로,

데이터를 속성에 기반하여 분류하고 상기 속성에 기반하여 분류된 데이터 각각에 대한 제1 블록 필터를 생성하는 단계;

상기 속성에 기반하여 분류된 데이터를 글자 단위로 분리한 집합의 크기, 복수의 해시 함수들 및 블록 필터의 인코딩 비트가 0이 나올 확률에 기초하여 상기 속성에 기반하여 분류된 데이터 각각에 대한 제1 블록 필터의 길이를 계산하는 단계;

상기 속성에 기반하여 분류된 데이터 각각에 대하여 속성 가중치를 계산하는 단계;

상기 속성에 기반하여 분류된 데이터 각각에 대하여 상기 속성 가중치를 고려한 제2 블록 필터를 생성하고 상기 제2 블록 필터의 길이를 계산하는 단계; 및

상기 속성에 기반하여 분류된 데이터 각각에 대한 치환을 수행하고, 상기 각 제2 블록 필터에 대한 상관관계를 분석하는 단계를 포함하는 것을 특징으로 하는, 결합키 생성 방법.

청구항 2

청구항 1에 있어서,

상기 제1 블록 필터의 길이는 하기 (수학식 1)의 과정으로 계산되는 것을 특징으로 하는, 결합키 생성 방법.

(수학식 1)

$$m = \frac{1}{1 - \sqrt[kn]{p}}$$

(여기서, m 은 제1 블록 필터의 길이이고, p 는 블록 필터의 인코딩 비트가 0이 나올 확률이고, n 은 데이터를 글자 단위로 분리한 집합의 크기의 평균이며, k 는 복수의 해시 함수들의 개수임)

청구항 3

청구항 2에 있어서,

상기 속성 가중치는, 하기 (수학식 2)의 과정으로 계산되는 것을 특징으로 하는, 결합키 생성 방법.

(수학식 2)

$$w[i] = \frac{range[i]}{\sum range[i]}$$

(여기서, $w[i]$ 는 속성 가중치이고, $range[i] = w_a[i] - w_d[i]$ 이고, $w_a[i]$ 는 일치 가중치로

$w_a[i] = \log_2\left(\frac{m_i}{u_i}\right)$ 이고, $w_d[i]$ 는 비일치 가중치이로 $w_d[i] = \log_2\left(\frac{1-m_i}{1-u_i}\right)$ 이고, $m_i = \frac{|M_i|}{|M|}$ 로 상

기 속성에 기반하여 분류된 데이터 쌍 M 개 중 i 번째 속성의 데이터 쌍이 일치할 확률이고 ,

$$u_i = \frac{|U_i|}{|U|}$$

로 상기 속성에 기반하여 분류된 데이터 쌍 U 개 중 i 번째 속성의 데이터 쌍이 일치하지 않을 확률임)

청구항 4

청구항 3에 있어서,

상기 제2 블록 필터의 길이는,

상기 각각의 제1 블록 필터의 길이와 상기 각각의 제1 블록 필터에 부합하는 속성 가중치의 역수 값의 곱으로 계산된 값들 중 가장 큰 값에 상기 속성 가중치를 곱하여 계산되는 것을 특징으로 하는, 결합키 생성 방법.

청구항 5

청구항 4에 있어서,

상기 각 제2 블록 필터에 대한 상관관계는,

하기 (수학식 3)의 과정으로 계산되는 것을 특징으로 하는, 결합키 생성 방법.

(수학식 5)

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

(여기서, Dice Coefficient는 상관관계 지수이고, $|X|$, $|Y|$ 는 각 제2 블록 필터의 길이이고, $X \cap Y$ 는 각 제2 블록 필터에서 일치하는 비트의 개수임)

발명의 설명

기술 분야

[0001] 본 발명은 개인정보 보호를 위한 결합키 생성 시스템 및 방법에 관한 것으로, 보다 상세하게는, 확률적 자료 연계 방법을 통한 임계 값을 바탕으로 주소 변경, 개명 등 다양한 시나리오에서도 강건한 결합키를 생성할 수 있는 시스템 및 방법에 관한 것이다.

배경 기술

[0002] 4차 산업혁명의 핵심 기술인 빅데이터와 인공지능을 활용하여 다수의 기업들은 개인 맞춤형 광고, 금융 상품 추천과 같은 다양한 서비스를 제공하고 있다. 빅데이터 및 인공지능 기반의 다양한 서비스를 제공하기 위하여 유럽, 호주 등 다수의 국가에서는 마이 데이터(MyData) 사업 및 정책을 시행하고 있는데, 이를 위하여 빅데이터 구축 및 데이터에 대한 개인정보 보호가 필수적으로 이루어져야 한다. 하지만 빅데이터 구축을 위한 데이터베이스 결합 시, 개인정보 보호법에 따라 이름, 주민등록번호 등 특정 개인을 식별할 수 있는 정보를 사용할 수 없는 문제가 있다.

[0003] 현재 국내 빅데이터 및 인공지능 사업의 기반이 되는 데이터 결합 시, 가명 정보 처리 가이드라인 등을 준수하여 개인 식별이 불가능한 가명 데이터를 생성해야 하며, 결합키 관리기관과 결합 전문기관을 통해 데이터베이스 결합이 진행되어야 할 수 있다.

[0004] 한편, 결합키 관리기관과 결합 전문기관을 통해서 결합신청자들의 데이터베이스를 결합할 때, 데이터의 구분자 역할을 하는 결합키가 매핑의 도구로 사용될 수 있다. 결합키 생성 시 Salt 값을 포함한 성명, 전화번호, 생년월일, 주소 등과 같은 결합키 생성항목의 해시값이 사용되는데, 해시 함수의 특성상 결합키 생성항목의 데이터

가 정확히 일치하는 경우에만 매핑이 성공적으로 수행될 수 있다. 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 수 있다.

[0005] 따라서 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 방법이 필요한 실정이다.

발명의 내용

해결하려는 과제

[0007] 상기와 같은 문제점을 해결하기 위한 본 발명의 갱신이 서로 다른 시점에 진행된 서로 다른 결합신청자들의 데이터베이스에서도 동일 개체임을 확인할 수 있는 강건한 결합키 생성 시스템 및 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 결합키 생성 방법은, 데이터를 속성에 기반하여 분류하고 상기 속성에 기반하여 분류된 데이터 각각에 대한 제1 블록 필터를 생성하는 단계, 상기 속성에 기반하여 분류된 데이터를 글자 단위로 분리한 집합의 크기, 복수의 해시 함수들 및 블록 필터의 인코딩 비트가 0이 나올 확률에 기초하여 상기 속성에 기반하여 분류된 데이터 각각에 대한 제1 블록 필터의 길이를 계산하는 단계, 상기 속성에 기반하여 분류된 데이터 각각에 대하여 속성 가중치를 계산하는 단계, 상기 속성에 기반하여 분류된 데이터 각각에 대하여 상기 속성 가중치를 고려한 제2 블록 필터를 생성하고 상기 제2 블록 필터의 길이를 계산하는 단계 및 상기 속성에 기반하여 분류된 데이터 각각에 대한 치환을 수행하고, 상기 각 제2 블록 필터에 대한 상관관계를 분석하는 단계를 포함하는 것을 특징으로 한다.

[0010] 또한, 상기 제1 블록 필터의 길이는 하기 (수학식 1)의 과정으로 계산되는 것을 특징으로 한다.

[0011] (수학식 1)

$$m = \frac{1}{1 - \sqrt[kn]{p}}$$

[0012]

[0013] (여기서, m 은 제1 블록 필터의 길이이고, p 는 블록 필터의 인코딩 비트가 0이 나올 확률이고, n 은 데이터를 글자 단위로 분리한 집합의 크기의 평균이며, k 는 복수의 해시 함수들의 개수임)

[0014] 또한, 상기 속성 가중치는, 하기 (수학식 2)의 과정으로 계산되는 것을 특징으로 한다.

[0015] (수학식 2)

$$w[i] = \frac{range[i]}{\sum range[i]}$$

[0016]

[0017] (여기서, $w[i]$ 는 속성 가중치이고, $range[i] = w_a[i] - w_d[i]$ 이고, $w_a[i]$ 는 일치 가중치로

$w_a[i] = \log_2\left(\frac{m_i}{u_i}\right)$ 이고, $w_d[i]$ 는 비일치 가중치이로 $w_d[i] = \log_2\left(\frac{1-m_i}{1-u_i}\right)$ 이고, $m_i = \frac{|M_i|}{|M|}$ 로 상

기 속성에 기반하여 분류된 데이터 쌍 M 개 중 i 번째 속성의 데이터 쌍이 일치할 확률이고 ,

$$u_i = \frac{|U_i|}{|U|}$$
 로 상기 속성에 기반하여 분류된 데이터 쌍 U 개 중 i 번째 속성의 데이터 쌍이 일치하지 않을 확률임)

[0018] 또한, 상기 제2 블록 필터의 길이는, 상기 각각의 제1 블록 필터의 길이와 상기 각각의 제1 블록 필터에 부합하는 속성 가중치의 역수 값의 곱으로 계산된 값들 중 가장 큰 값에 상기 속성 가중치를 곱하여 계산되는 것을 특징으로 한다.

[0019] 또한, 상기 각 제2 블록 필터에 대한 상관관계는, 하기 (수학식 3)의 과정으로 계산되는 것을 특징으로 한다.

[0020] (수학식 5)

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

[0021]

[0022] (여기서, Dice Coefficient는 상관관계 지수이고, $|X|$, $|Y|$ 는 각 제2 블록 필터의 길이이고, $X \cap Y$ 는 각 제2 블록 필터에서 일치하는 비트의 개수임)

발명의 효과

[0024] 본 발명에 따른 개인정보 보호를 위한 결합키 생성 시스템 및 방법은 가명 정보 처리 과정에서 서로 다른 갱신 시점 간 결합키 생성항목 결합 시 발생하는 취약점을 보완할 수 있다.

[0025] 또한, 기존의 Salt 값을 포함한 결합키 생성항목의 해시값으로 매핑을 진행할 때에는 불가능했던 경우에 대해서도 동일한 개체인 경우에는 자료 연계가 가능하도록 할 수 있다. 또한, 결합키 생성을 위한 결합키 생성항목의 데이터 전처리와 해당 데이터가 갱신되는 경우에 있어, 결합 불가능성 등의 문제를 최소화할 수 있다. 또한 기존의 가명정보 처리 가이드라인을 준수하는 강건한 결합키 생성 방법 및 시스템을 제공할 수 있다.

[0026] 다만, 본 발명의 실시 예들에 따른 개인정보 보호를 위한 결합키 생성 시스템 및 방법이 달성할 수 있는 효과는 이상에서 언급한 것들로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0028] 본 발명에 관한 이해를 돕기 위해 상세한 설명의 일부로 포함되는, 첨부도면은 본 발명에 대한 실시예를 제공하고, 상세한 설명과 함께 본 발명의 기술적 사상을 설명한다.

도 1은 본 발명에 따른 결합키 생성 방법을 도시한 흐름도이다.

도 2는 본 발명에 따른 결합키 생성 방법을 도시한 개념도이다.

도 3은 본 발명에 따른 결합키 생성 방법에서 레코드 연결 데이터베이스(record linkage database)를 도시한 것이다.

도 4는 본 발명에 따른 결합키 생성 방법에서 블록 필터의 일 실시예를 도시한 것이다.

도 5는 본 발명에 따른 결합키 생성 방법에서 블록 필터가 생성되는 일 실시예를 도시한 것이다.

도 6은 본 발명에 따른 결합키 생성 방법에서 블록 필터가 치환되는 과정을 도시한 것이다.

도 7 내지 도 15는 각 시나리오에서 본 발명에 따른 결합키 생성 방법을 적용한 결과를 도출한 그래프이다.

발명을 실시하기 위한 구체적인 내용

- [0029] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0030] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [0031] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [0032] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0033] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0034] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.
- [0035] 도 1은 본 발명에 따른 결합키 생성 방법을 도시한 흐름도이고, 도 2는 본 발명에 따른 결합키 생성 방법을 도시한 개념도이다.
- [0036] 도 1 및 도 2를 함께 참조하면, 본 발명에 따른 결합키 생성 방법은 데이터를 속성에 기반하여 분류하고, 속성에 기반하여 분류된 데이터 각각에 대한 제1 블룸 필터를 생성하는 단계(S100), 속성에 기반하여 분류된 데이터를 글자 단위로 분리한 집합의 크기, 복수의 해시 함수들 및 블룸 필터의 인코딩 비트가 0이 나올 확률에 기초하여 속성에 기반하여 분류된 데이터 각각에 대한 제1 블룸 필터의 길이를 계산하는 단계(S200), 속성에 기반하여 분류된 데이터 각각에 대하여 속성 가중치를 계산하는 단계(S300), 속성에 기반하여 분류된 데이터 각각에 대하여 속성 가중치를 고려한 제2 블룸 필터를 생성하고 제2 블룸 필터의 길이를 계산하는 단계(S400) 및 속성에 기반하여 분류된 데이터 각각에 대한 치환을 수행하고, 각 제2 블룸 필터에 대한 상관관계를 분석하는 단계(S500)를 포함할 수 있다.
- [0037] 즉, 본 발명에 따른 블룸 필터를 이용한 결합키 생성 방법은 속성 기반 블룸 필터를 생성하는 과정(FBF (Field level Bloom Filter) parameterization & generation), 적합한 비트를 식별하는 과정(eligible bit identification), 속성 가중치를 선정하는 과정(field weighting), 자료에 기반하여 블룸 필터를 생성하는 과정(RBF (Record level Bloom Filter) parameterization & generation), 자료에 기반하여 블룸 필터를 치환하는 과정(RBF permutation)을 포함하는 방법으로 수행될 수 있다. 한편, 본 발명에 따른 S100 내지 S500의 과정은 특정한 장치에 포함된 논리적 구성이 주체를 통하여 수행될 수 있다.
- [0038] S100 및 S200은 속성에 기반하여 블룸 필터를 생성하고, 적합한 비트를 식별하여 블룸 필터의 길이를 생성하는 과정일 수 있다. 예를 들어, 집합 $S = \{s_1, s_2, \dots, s_n\}$ 에서 k 개의 해시 함수를 이용하여 길이 m 의 블룸 필터 인코딩을 진행하는 경우, 블룸 필터 인코딩 비트(bit)가 0이 나올 확률 p 는 하기 (수학식 1)과 같을 수 있다.

수학식 1

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

[0039]

[0040]

Dynamic FBF는 속성 기반 블록 필터를 각 속성에 적용하였을 때 블록 필터 길이일 수 있고, 이는 하기 (수학식 2)이 표현될 수 있다.

수학식 2

$$m = \frac{1}{1 - \sqrt[kn]{p}}$$

[0041]

[0042]

예를 들어, “이름” 속성의 “홍길동” 값에 Dynamic FBF를 적용하는 경우, 오타 발생 가능성에 의해 “홍길동”의 값을 갖는 데이터를 가지는 데이터를 q 글자 단위로 분리할 수 있다. $q=2$ 일 때, $S = \{-\text{홍}, \text{홍길}, \text{길동}, \text{동}-\}$, $n=4$ 임을 알 수 있다. 그리고 15개의 해시 함수를 이용한 블록 필터 인코딩 bit의 0, 1 비율을 50%로 가정하면 $k=15$, $p=0.5$ 일 수 있다. 이를 계산하면 $m \approx 87$ 일 수 있다.

[0043]

이후, 분석하고자 하는 두 개의 데이터셋에 대하여 속성 별 블록 필터의 길이를 계산할 수 있다. 데이터를 q 글자 단위로 분리한 집합 S 의 크기 n 의 평균과 k , p 를 이용하여 계산한 m 을 속성 별 블록 필터의 길이로 정할 수 있다.

[0044]

도 3은 본 발명에 따른 결합키 생성 방법에서 레코드 연결 데이터베이스(record linkage database)를 도시한 것이다. 도 3을 참조하면, 공통 속성인 “이름”, “생년월일”에 대하여 $q=2$, $k=15$, $p=0.5$ 일 때, Dynamic FBF를 이용해 속성별 블록 필터 길이를 구하면 하기 표 1과 같을 수 있다.

표 1

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

[0045]

[0046]

한편, 도 4는 본 발명에 따른 결합키 생성 방법에서 블록 필터의 일 실시예를 도시한 것이다. 도 4를 참조하면, 블록 필터는 원소의 집합에 한 포함 여부를 확인하는 확률적 자료 구조일 수 있고, 블록 필터 길이만큼의 균등한 확률을 출력하는 해시 함수를 사용하여 집합에 대한 원소의 포함 여부를 확인할 수 있다[8]. 집합에 포함되지 않는 원소는 블록 필터 적용 시 0인 bit가 존재하며, 모두 1인 bit를 가질 경우, 해당 집합에 포함되는 원소라 추측할 수 있다.

[0047]

예를 들어, 도 4와 같이 $n=3$ 개의 원소를 가지는 집합 $S = \{x, y, z\}$ 의 각 원소에 대해 $k=2$ 개의 서로 다른 해시 함수를 사용하여 길이 $m=8$ 의 블록 필터를 생성한다고 가정한다. 이 때, w_1 과 같이 블록

필터 결과가 0인 bit를 가질 경우, 해당 원소는 S 의 원소가 아님을 확신할 수 있다. 또한 w_2 와 같이 모든 bit가 1인 경우, 해당 원소는 S 의 원소가 될 가능성이 존재할 수 있다.

[0048] 다시 도 1 및 도 2를 참조하면, S300은 각 데이터셋에 맞는 속성 가중치를 선정하는 과정일 수 있다. 이 때, 속성 가중치는 $w[i]$ 는 하기 (수학식 3)의 과정으로 계산될 수 있다.

수학식 3

$$w[i] = \frac{range[i]}{\sum range[i]}$$

[0049] 여기서, $w[i]$ 는 속성 가중치이고, $range[i] = w_a[i] - w_d[i]$ 이고, $w_a[i]$ 는 일치 가중치로

$$w_a[i] = \log_2\left(\frac{m_i}{u_i}\right) \text{ 이고, } w_d[i] \text{ 는 비일치 가중치이고 } w_d[i] = \log_2\left(\frac{1-m_i}{1-u_i}\right) \text{ 이고, } m_i = \frac{|M_i|}{|M|} \text{ 로}$$

상기 속성에 기반하여 분류된 데이터 쌍 M 개 중 i 번째 속성의 데이터 쌍이 일치할 확률이고 ,

$$u_i = \frac{|U_i|}{|U|} \text{ 로 상기 속성에 기반하여 분류된 데이터 쌍 } U \text{ 개 중 } i \text{ 번째 속성의 데이터 쌍이 일치하지 않을 확률이다.}$$

[0051] 일치 가중치 $w_a[i]$ 와 비일치 가중치 $w_d[i]$ 는 Fellegi-Sunter 모델과 EM 알고리즘을 이용하여 계산될 수 있고, 계산 상의 편의를 위해 로그가 사용될 수 있다. Fellegi-Sunter 모델은 확률적 자료 연계에 대한 모델로, 오타자, 약어 등이 포함된 서로다른 데이터셋(dataset)에 대해 자료를 구성하는 “이름”, “주소” 등의 속성(field)에 가중치를 부여할 수 있다. 그리고 부여된 가중치를 바탕으로 자료 간의 유사도를 판별하여 자료를 연계할 수 있다. Fellegi-Sunter 모델은 두 데이터베이스 간에 공통 속성이 있음을 가정하며, 공통 속성의 가중치에 따라 연계(link), 비연계(non-link), 연계 가능성(possible-link)을 판정할 수 있다.

[0052] 데이터의 일치 기준 sim_j 는 하기 (수학식 4)로 계산될 수 있다.

수학식 4

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

[0053] 여기서, m 은 s_1 , s_2 의 일치하는 문자의 개수, t 는 $\lfloor \max(|s_1|, |s_2|)/2 \rfloor - 1$ 범위에서 m 에 포함되지 않으며 일치하는 문자의 개수일 수 있다.

[0055] 또 다른 실시예로서 데이터의 일치 기준 sim_w 는 하기 (수학식 5)로 계산될 수 있다.

수학식 5

$$sim_w = sim_j + lp(1 - sim_j)$$

[0056]

[0057] sim_w 는 s_1 , s_2 의 공통 접두사 길이 l ($0 \leq l \leq 4$) 과 유사도 조정을 위한 스케일 팩터(scaling factor) p ($0 < p \leq 0.25$) 일 수 있다. p 의 기본값은 0.1이며, p 가 클수록 일치하는 문자가 적은 경우에도 문자열 s_1 , s_2 가 일치한다고 판단할 수 있다.

[0058]

S400은 선정된 속성 가중치에 맞는 자료의 블록 필터를 생성하는 과정일 수 있다. 즉, 자료 기반 블록 필터는 속성별로 선정된 가중치 비율에 맞게 병합된 블록 필터의 길이를 선정하고, 속성별로 동일한 치환을 적용하여 블록 필터를 병합할 수 있다. 이 때, 가중치에 부합하는 자료의 제2 블록 필터는 앞서 S100 및 S200에서 계산된 제1 블록 필터의 길이와 상기 각각의 제1 블록 필터에 부합하는 속성 가중치의 역수 값의 곱으로 계산된 값들 중 가장 큰 값에 상기 속성 가중치를 곱하여 계산될 수 있다.

[0059]

예를 들어, 상기 (표 1)과 같이 Dynamic FBF로 계산된 “이름”, “생년월일”의 블록 필터 길이가 87,173이고, 속성의 가중치가 25%, 75%라 가정하는 경우, “이름” 속성의 RBF 길이는 $87 \times (1/0.25) = 348$ 이며, “생년월일” 속성의 RBF 길이는 $174 \times (1/0.75) = 232$ 일 수 있다. 그리고 속성별 RBF 길이 최대값을 기준으로 가중치 비율에 부합하게 각 속성의 블록 필터 길이를 계산할 수 있다. 이 때, $\max(348, 232) = 348$ 이므로 “이름” “생년월일”의 블록 필터의 길이는 $348 \times 0.25 = 87$, $348 \times 0.75 = 261$ 일 수 있다.

[0060]

도 5는 본 발명에 따른 결합키 생성 방법에서 블록 필터가 생성되는 일 실시예를 도시한 것이다. 도 5를 참조하면, 속성 별로 동일한 치환을 적용하여 블록 필터를 병합하는 과정이 수행될 수 있다. 치환은 랜덤하게 수행될 수 있고, 비교하고자 하는 데이터셋의 동일한 속성 간에 동일한 치환이 적용될 수 있다.

[0061]

다시 도 1 및 도 2를 참조하면, S500은 자료의 블록 필터에 대한 사전 공격을 방지하기 위해 두 데이터셋 간에 합의된치환을 진행하고, 최종적으로 자료의 블록 필터에 대한 상관관계를 분석하는 과정일 수 있다.

[0062]

도 6은 본 발명에 따른 결합키 생성 방법에서 블록 필터가 치환되는 과정을 도시한 것이다. 도 6을 참조하면, 자료 기반 블록 필터로 병합된 블록 필터에 대해 도 6에 도시된 바와 같이 두 데이터셋 간에 합의된 치환을 적용하여 블록 필터에 대한 사전 공격을 방지할 수 있다. 그리고 최종적으로 계산된 블록 필터에 대해 하기 (수학식 6)과 같이 Dice Coefficient를 계산함으로써 두 자료 간의 상관관계를 계산할 수 있다.

수학식 6

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

[0063]

[0064]

자료 간의 일치하는 문자가 많을수록 Dice Coefficient는 1에 근사할 수 있고, 그렇지 않을 경우, 0에 근사할 수 있다. 비트스트링으로 표현된 블록 필터의 동일 위치의 bit가 일치하는 경우, 상기 (수학식 6)의 교집합의 원소가 될 수 있다. 예를 들어, 도 6에서 Dice Coefficient 계산 시, Record A, B는 각 16 비트이며, 일치하는 비트가 14개이므로, Dice Coefficient는

$$\frac{2(\text{Record A} \cap \text{Record B})}{|\text{Record A}| + |\text{Record B}|} = \frac{2 \times 14}{16 + 16} = 0.875$$

일 수 있다.

[0065] 도 7 내지 도 15는 각 시나리오에서 본 발명에 따른 결합키 생성 방법을 적용한 결과를 도출한 그래프이다.

[0066] 실험 개요

[0067] 본 발명에서 속성 기반 블록 필터 생성 시 데이터를 2글자 단위로 분리하며($q=2$), 블록 필터 인코딩 비트의 0.1 비율을 50%($p=0.5$), 15개의 해시함수를 사용하여($k=15$) Dynamic FBF를 실시한다.

[0068] 블록 필터에 적용된 15개의 해시 함수는 두 개의 암호학적 해시 함수를 이용해 동일한 성능을 갖도록 하기 (수학식 7)을 이용하여 생성한다.

수학식 7

$$g_i(x) = h_1(x) + ih_2(x) \pmod{m}$$

[0069]

[0070] m 은 블록 필터의 길이, i 는 $1 \leq i \leq 15$ 이며, 실험에 사용된 암호학적 해시함수는 $h_1 : \text{SHA256}$, $h_2 : \text{KECCAK256}$ 이다.

[0071] “속성 가중치 선정” 시, scaling factor p 에 대하여 $p \in \{0.1, 0.15, 0.2, 0.25\}$ 의 평균 sim_w 를 계산한다. 평균 sim_w 가 0.9 이상인 경우, 일치하는 데이터로 판단하여 M_i 에 포함한다. 평균 sim_w 가 0.9미만인 경우, 불일치하는 데이터로 판단하여 U_i 에 포함한다. 그리고 “자료 기반 블록 필터 생성” 시, 각 속성별 동일한 랜덤 치환을 적용하며, “자료 기반 블록필터 치환” 시, 분석하고자 하는 두 개의 데이터셋에 동일한 랜덤 치환을 적용한다. 실험의 결과를 토대로 개인정보가 갱신된 데이터셋 간의 가명정보 결합 가능성을 판단한다.

[0073] 실험 시나리오

[0074] 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. 이는 하기 (표 2)와 같다.

표 2

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

[0075]

- [0076] “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.
- [0077] ① 주소가 변경된 경우
- [0078] 주소가 변경되었지만 데이터베이스에 아직 주소 변경 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 주소는 이전 주소이며, 데이터셋 2의 주소는 갱신된 주소이다.
- [0079] ② 주소가 변경된 경우 (동일 지역)
- [0080] 동일 지역으로 주소가 변경되었지만 데이터베이스에 아직 주소 변경 내용이 적용되지 않은 시나리오이다. 동일 지역의 기준은 “00시 00구” 또는 “00도 00시” 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.
- [0081] ③ 전화번호가 변경된 경우
- [0082] 전화번호가 변경되었지만 데이터베이스에 아직 전화번호 변경 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 전화번호는 이전 전화번호이며, 데이터셋 2의 전화번호는 갱신된 전화번호이다.
- [0083] ④ 전화번호가 변경된 경우 (뒷 4자리 동일)
- [0084] 전화번호가 가운데 4자리만 변경되었지만 데이터베이스에 아직 전화번호 변경 내용이 적용되지 않은 시나리오이다. 데이터셋 1의 전화번호는 이전 전화 번호이며, 데이터셋 2의 전화번호는 갱신된 전화번호이다. ③, ④ 시나리오의 전화번호는 모두 휴대전화번호를 기준으로 한다. 그리고 데이터셋 1, 2의 동일 인덱스의 자료는 “전화번호” 속성의 데이터가 다르며, “이름”, “주소”, “생년월일” 속성의 데이터는 동일하다.
- [0085] ⑤ 개명의 경우
- [0086] 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.
- [0087] ⑥ 동명이인의 경우
- [0088] 데이터셋 1, 2의 동일 인덱스의 자료는 동명이인으로 “이름” 속성의 데이터가 동일하다. 그리고 “주소”, “전화번호”, “생년월일” 속성의 데이터는 모두 다르다.
- [0089] ⑦ 다른 인물이지만 전화번호가 동일한 경우
- [0090] 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.
- [0091] ⑧ 동일 인물의 데이터가 섞인 경우
- [0092] ① 내지 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.
- [0093] ⑨ 모든 정보가 다른 경우
- [0094] “이름”, “주소”, “전화번호”, “생년월일” 속성이 모두 다른 데이터셋 1, 2에 대한 경우이다. 데이터셋 1, 2의 자료는 모두 동일하지 않은 인물이다.

[0096] 데이터셋 구성

[0097] 기본 데이터는 공공데이터포털의 “건강보험심사평가원 요양기관 개설 현황” 데이터베이스를 활용하였다. 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.

[0098] “이름” 속성의 데이터는 “성”과 “이름”으로 구분하여 공공데이터를 활용해 가명을 생성하였다. “성”은 통계청의 “통계지리정보서비스”에서 제공하는 “성씨 및 본관별 인구” 데이터를 활용하였다[15]. “이름”은 대한민국 법원의 “전자가족관계등록시스템”의 “상위 출생신고 이름 현황” 데이터를 활용하였다. “성씨·본관별 인구” 데이터는 2015년까지 전국 성씨에 대한 총 조사이다. 이 중 (표 3)과 같이 전국 단위 상위 50개의 성씨를 비율에 맞게 99205건을 생성하였다.

표 3

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

[0099]

[0101] “상위 출생신고 이름 현황” 데이터는 2008년부터 현재까지 출생신고된 이름 순위 데이터이다. 이 중 (표 4)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.

표 4

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

[0102]

[0103] “전화번호” 속성은 휴대전화번호 형식인 “0000-0000”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 구축되어 있다. “00동”과 같은 부가 정보는 제외하여 데이터셋을 구성하였다.

[0104] 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.

[0105] 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.

- [0106] 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.
- [0107] ⑤ 시나리오의 경우, 데이터셋 2의 “이름” 속성의 데이터는 데이터셋 1의 “이름” 속성의 데이터에 사용되지 않은 98605건의 “이름” 속성의 데이터 중 성을제외한 이름만을 랜덤하게 선택하였다.
- [0108] ⑥ 시나리오의 경우, 데이터셋 1, 2의 “이름” 속성의 데이터만이 일치하며, ⑦ 시나리오의 경우, 데이터셋 1, 2의 “전화번호” 속성의 데이터만이 일치한다.
- [0109] ⑧ 시나리오의 경우, ① ~ ⑤ 시나리오에 사용된 자료 중, 각 시나리오별로 100건의 자료를 랜덤하게 선택하여 데이터셋 1, 2를 구성하였다. 검증 데이터 셋은 각 시나리오별로 20건의 자료를 랜덤하게 선택하여 검증 데이터셋 1, 2를 구성하였다.
- [0110] 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.
- [0111] 각 시나리오별 속성 가중치 계산 결과는 하기 (표 5)와 같다.

표 5

Scenario	Weights			
	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%

- [0112]
- [0113] ⑧, ⑨ 시나리오의 경우, 특정 속성을 기준으로 자료에 대한 일치, 불일치를 판단할 수 없으므로 25%의 균등한 속성 가중치를 가진다. 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.
- [0114] ①: “정지호, 서울특별시 강서구 수명로 76, 7569-3333, 2008-04-23”, “정지호, 전라북도 완주군 화산면 화산로 866, 7569-3333, 2008-04-23”
- [0115] ②: “김소윤, 부산광역시 동래구 아시아드대로 191, 9592-5303, 2019-05-01”, “김소윤, 부산광역시 동래구 충렬대로348번길 23, 9592-5303, 2019-05-01”
- [0116] ③: “김은서, 충청남도 천안시 서북구 월봉로 83, 1190-8986, 2000-03-03”, “김은서, 충청남도 천안시 서북구 월봉로 83, 2589-5554, 2000-03-03”
- [0117] ④: “장승우, 서울특별시 동작구 상도로 146, 8638-8112, 1993-12-31”, “장승우, 서울특별시 동작구 상도로 146, 3149-8112, 1993-12-31”
- [0118] ⑤: “이하준, 인천광역시 계양구 계산로 89, 2642-6698, 2020-06-30”, “이시은, 인천광역시 계양구 계산로 89, 2642-6698, 2020-06-30”
- [0119] ⑥: “임효린, 서울특별시 송파구 양재대로 932, 5632-5487, 2010-08-09”, “임효린, 서울특별시 마포구 마포대로 68, 8909-7575, 2021-11-09”

- [0120] ⑦: “이예준, 서울특별시 강남구 압구정로 151,
- [0121] 9041-1101, 2015-01-30”, “전태연, 전라북도 남원시 주천면 정령치로 55, 9041-1101, 1995-01-01”
- [0122] ⑧: “김다은, 서울특별시 광진구 용마산로 8, 8730-8575, 2005-03-02”, “김윤서, 서울특별시 광진구 용마산로 8, 8730-8575, 2005-03-02”
- [0123] ⑨: “이형우, 충청남도 공주시 변영1로 87, 5410-0075, 2004-05-27”, “윤시원, 부산광역시 부산진구 중앙대로 736, 7269-9778, 2011-05-23”
- [0125] **실험 결과**
- [0126] **실험 시나리오 순서**
- [0127] 우선 특정 속성 기준이 없는 시나리오인 “⑧ 동일인물의 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.
- [0128] **특정 속성 기준이 없는 시나리오**
- [0129] 특정 속성 기준이 없는 시나리오의 경우, “이름”, “주소”, “전화번호”, “생년월일” 속성의 가중치를 모두 25%로 가진다. ⑧ 시나리오에 대한 결과는 도 7의 (a)와 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 도 7의 (b)와 같다.
- [0130] 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.
- [0131] 이를 바탕으로 ⑨ 시나리오에 임계값 0.8을 적용한 결과는 도 8의 (a)와 같다. 총 4개의 다른 인물 자료쌍이 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.
- [0132] 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 도 8의 (b)와 같다.
- [0133] ⑨ 시나리오의 검증 데이터셋에 대한 다른 인물자료 쌍의 최대 Dice Coefficient는 검증 데이터셋 1의 93번 자료와 검증 데이터셋 2의 26번 자료로 약 0.7700이다. 모든 자료 쌍에 대해 임계값 0.8 이상의 결과를 갖지 않는다. 따라서 특정 속성기준이 없는 가명정보 결합의 경우, 임계값을 0.8에 따라 가명정보 결합이 가능하다.
- [0134] **특정 속성 기준 동일 인물 가명정보 결합 시나리오**
- [0135] “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 정하고자한다. ① 시나리오는 “주소” 속성의 가중치가 낮음으로써, “주소” 이외의 속성에 따라 가명정보 결합이 가능하다. ①

시나리오에 대한 결과는 도 9의 (a)와 같다.

- [0136] 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.
- [0137] 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 도 9의 (b)와 같다.
- [0138] 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 도 10 내지 도 13과 같다.
- [0139] 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 시나리오와 같이 최대 약 0.0108%의 확률로 임계값 0.8 이상의 Dice Coefficient를 얻을 수 있으며, 임계값 0.9 이상의 결과는 존재하지 않는다.
- [0140] 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 자료쌍이 존재하며, 임계값 0.9에 대해 특정 속성 기준의 동일 인물의 가명정보 결합이 가능하다.
- [0141] **특정 속성 기준 다른 인물 구분 시나리오**
- [0142] “⑥ 동명이인의 경우”, “⑦ 다른 인물이지만 전화번호 동일한 경우”의 시나리오는 특정 속성을 기준으로 다른 인물임을 구분하는 시나리오이다. 데이터셋1, 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 도 14 및 도 15와 같다.
- [0143] ⑥ 시나리오의 다른 인물 자료 쌍의 최대 Dice 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 인물자료 쌍이 존재한다. ⑦ 시나리오의 경우, 임계값 0.8 이상의 Dice Coefficient를 갖는 다른 인물 자료 쌍은 존재하지 않는다. 임계값 0.9 이상의 Dice Coefficient를 갖는 다른 인물 자료 쌍은 ⑥, ⑦ 시나리오 모두 존재하지 않는다.
- [0144] **실험 결과**
- [0145] “자료 쌍의 Dice Coefficient가 임계값 이상이면 동일 인물이다.”라는 가설 1(hypothesis 1), “자료쌍의 Dice Coefficient가 임계값 미만이면 다른 인물이다.”라는 가설 2(hypothesis 2)에 대해 실험결과를 적용하면 다음과 같다. ⑧ 시나리오와 같이 가명정보 결합 과정에서 특정속성을 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.
- [0146] 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 인물자료 쌍이 존재하지 않으므로 가설 1에 대한 검증은 불가능하다. 따라서 해당 결과를 통해 모든 시나리오에 대해 가설 1, 2를 99% 이상의 정확도로 만족함을 (표 6)과 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.

표 6

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

[0147]

[0148]

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 발명에서 우리가 제안한 기법을 활용하면 데이터베이스의 갱신 시점이 다르더라도 동일 개체라면 강건하게 매핑에 성공할 수 있음을 확인하였다. 특히, 시나리오별로 실험을 진행하였으며, 이에 따른 적합한 임계값 역시 제시하였다. 본 발명을 통하여 국내 데이터 결합 시 결합속성 관련 데이터 전처리 과정 및 해당 데이터의 갱신에 따른 결합 불가능성 등의 문제를 해결할 수 있다.

[0149]

본 발명의 실시예에 따른 방법의 동작은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

[0150]

또한, 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0151]

본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시예에서, 가장 중요한 방법 단계들의 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0152]

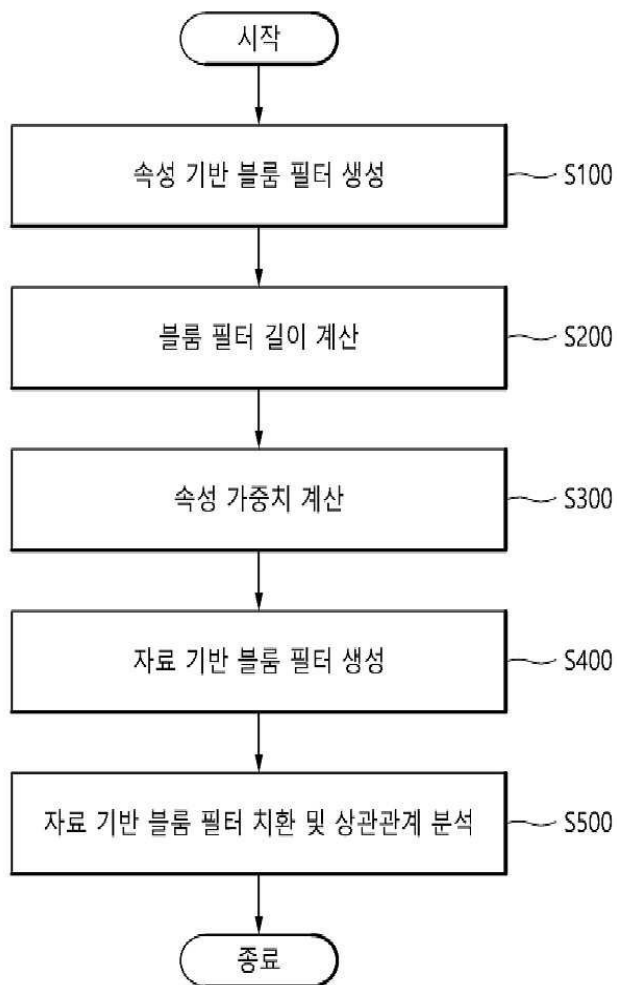
실시예들에서, 프로그램 가능한 로직 장치(예를 들어, 필드 프로그램블 게이트 어레이)가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시예들에서, 필드 프로그램블 게이트 어레이는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

[0153]

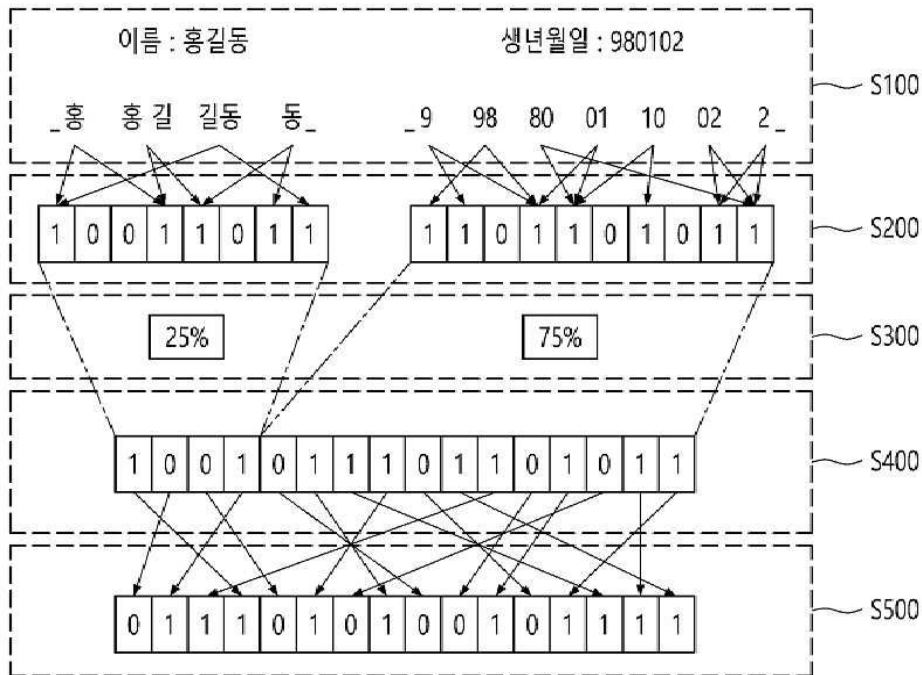
이상 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

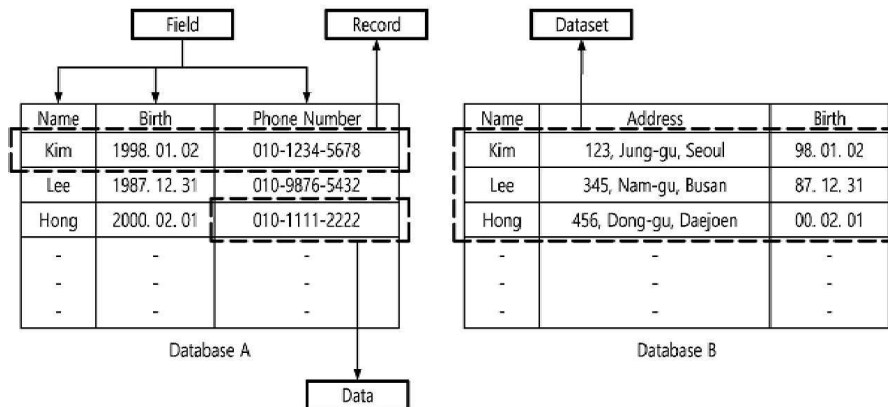
도면1



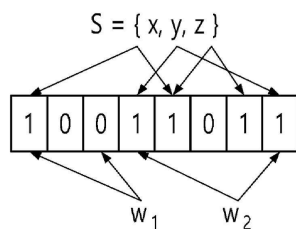
도면2



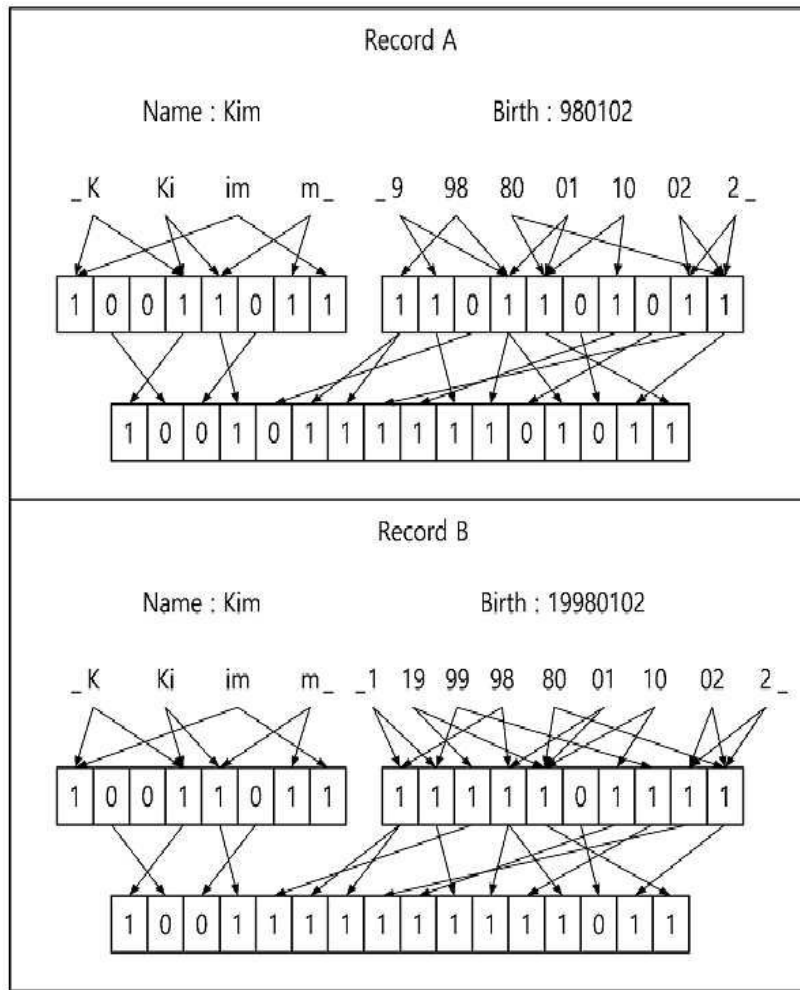
도면3



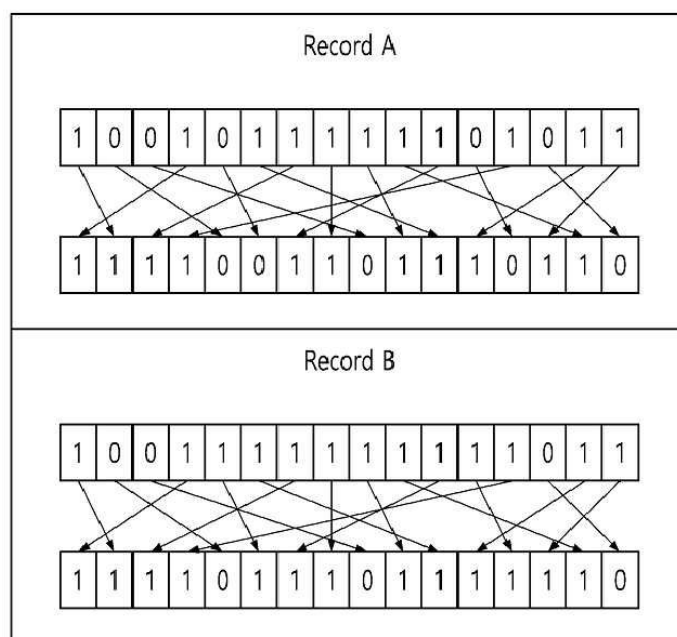
도면4



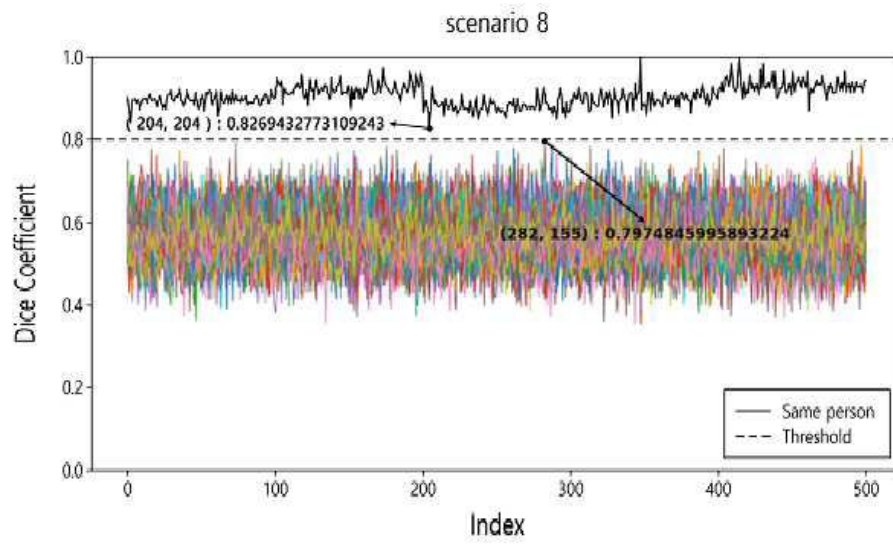
도면5



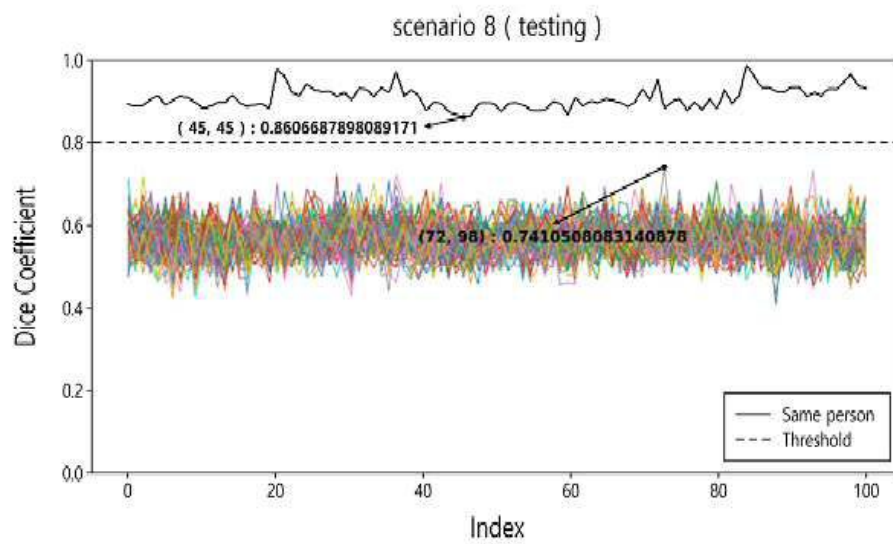
도면6



도면7

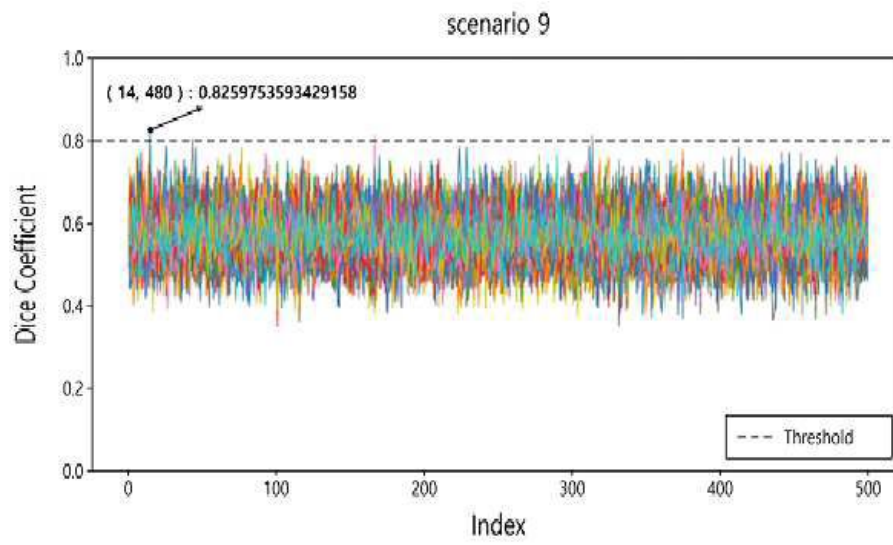


(a)

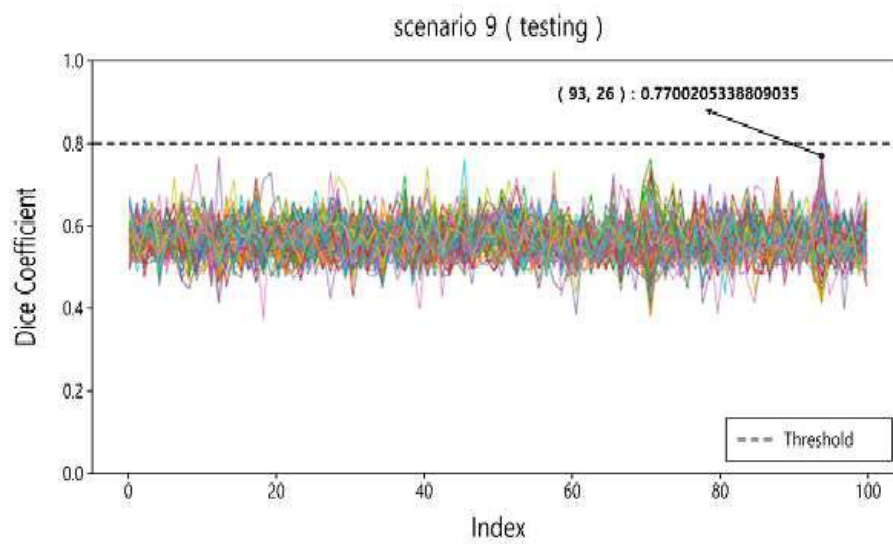


(b)

도면8

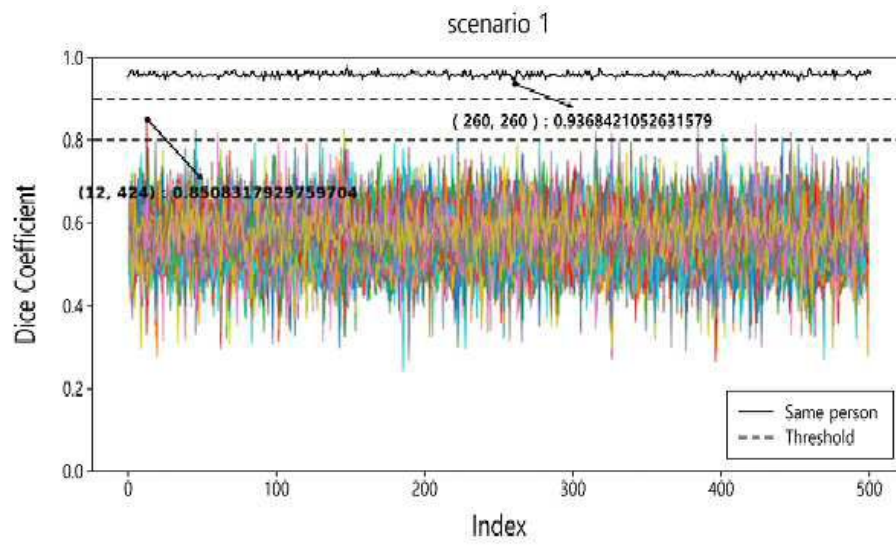


(a)

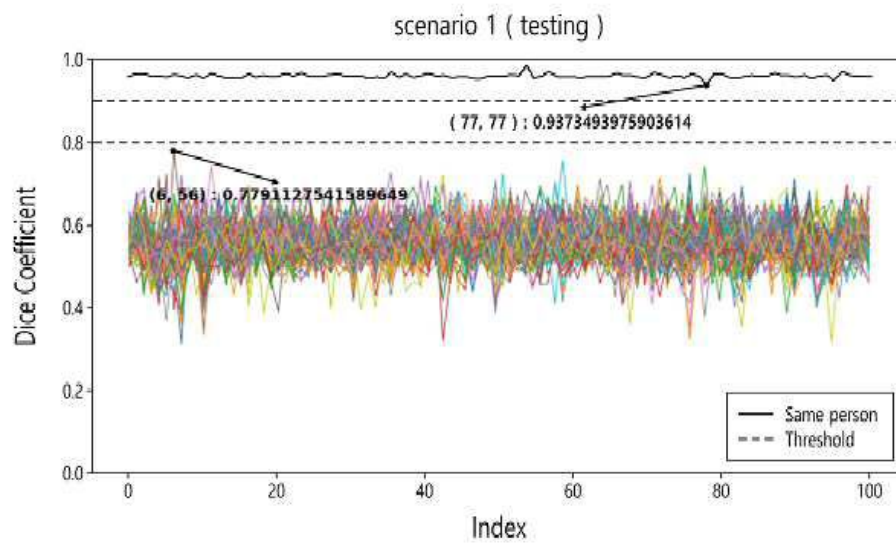


(b)

도면9

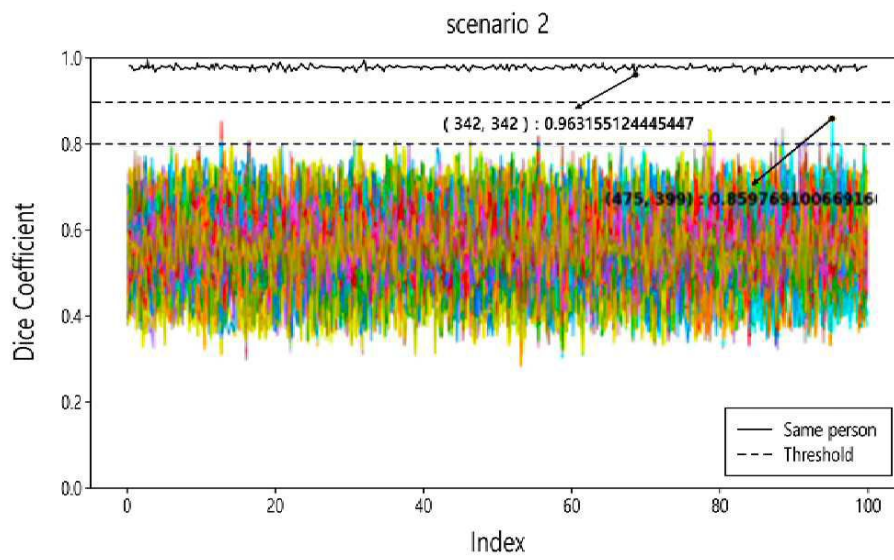


(a)

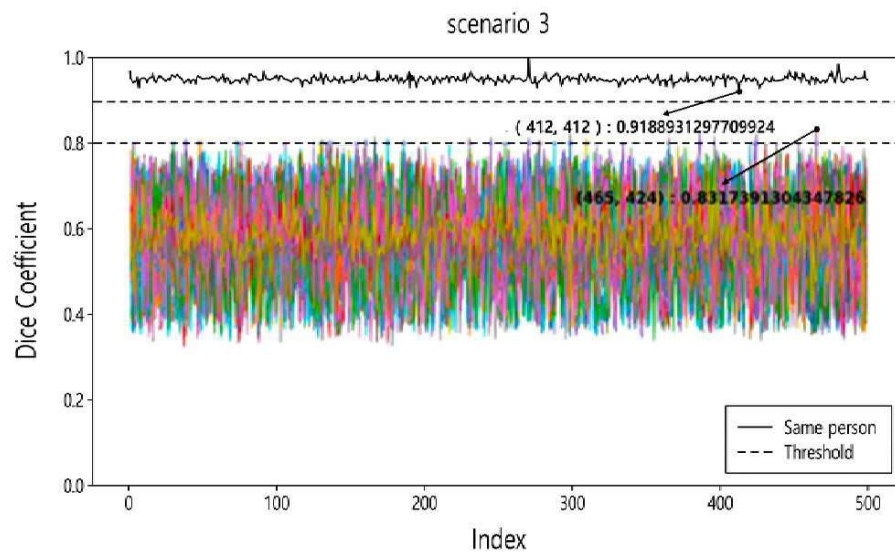


(b)

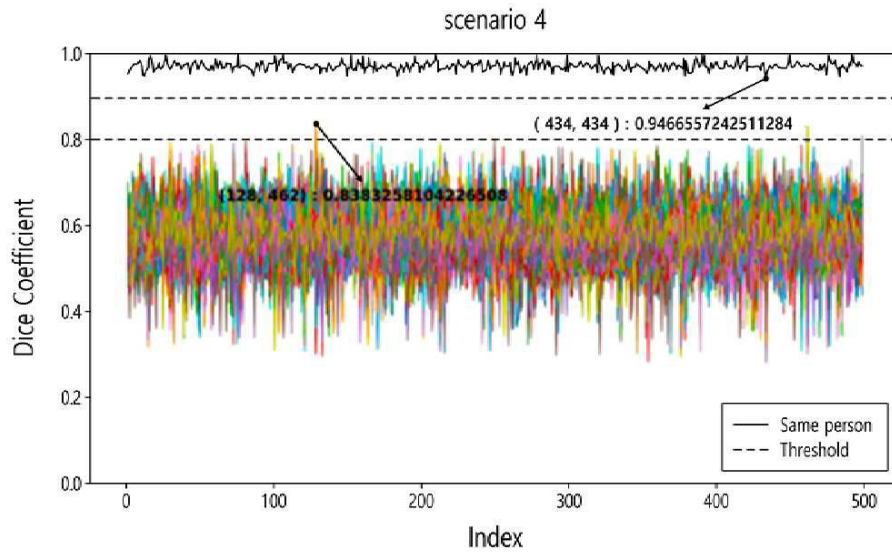
도면10



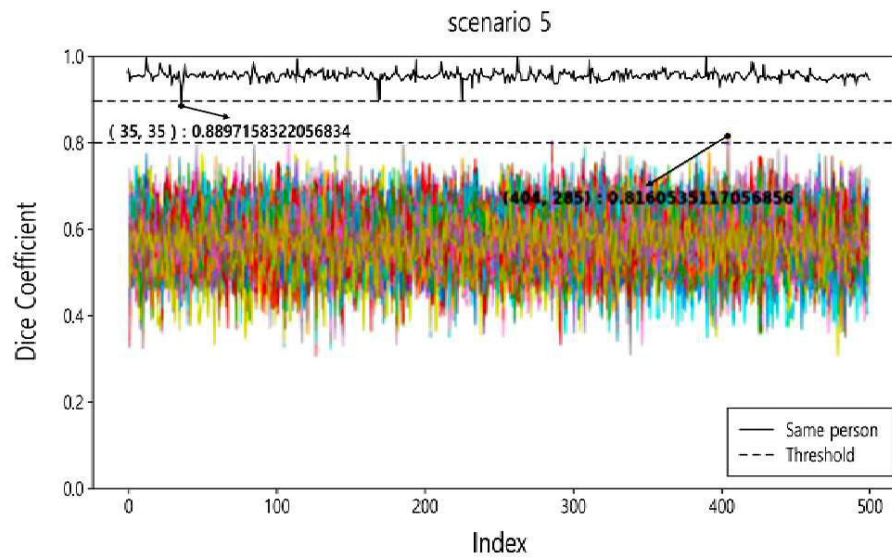
도면11



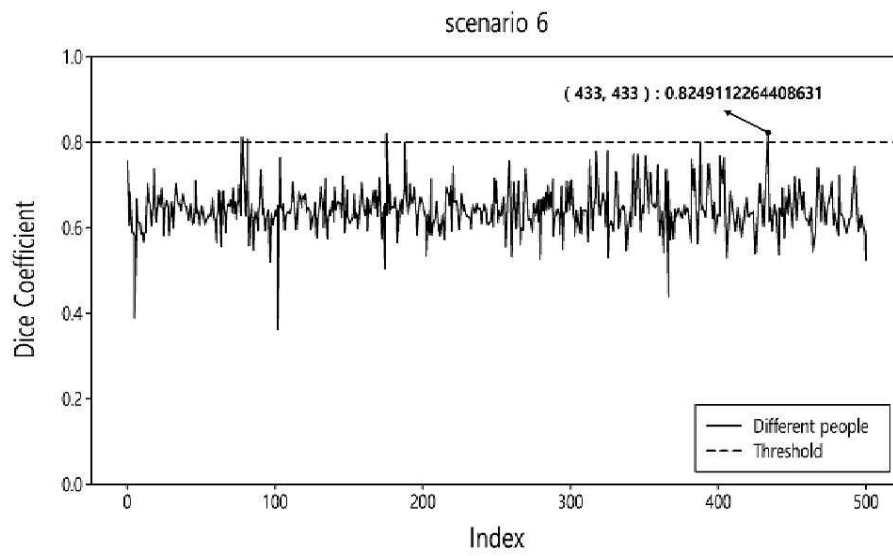
도면12



도면13



도면14



도면15

