(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년02월19일
(11) 등록번호 10-2218374
(24) 등록일자 2021년02월16일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) *G06F 16/215* (2019.01)
G06F 16/22 (2019.01)
(52) CPC특허분류
G06F 21/6254 (2013.01)
G06F 16/215 (2019.01)
(21) 출원번호 10-2019-0045149
(22) 출원일자 2019년04월17일
심사청구일자 2019년04월17일
(65) 공개번호 10-2020-0122195
(43) 공개일자 2020년10월27일
(56) 선행기술조사문헌
JP2017228255 A*
KR1020180034108 A*
KR1020180119104 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
이원석
서울특별시 영등포구 여의동로 143(여의도동, 대우트럼프월드)
(74) 대리인
특허법인우인

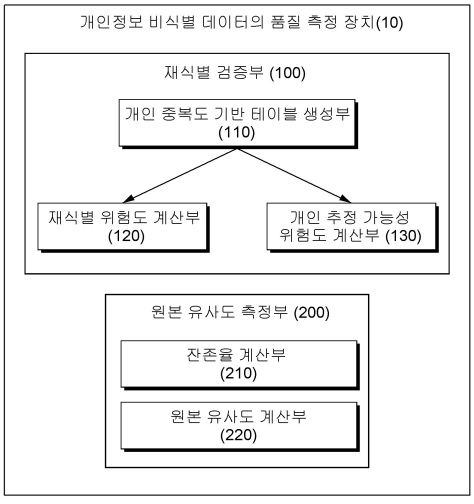
전체 청구항 수 : 총 10 항 심사관 : 구대성

(54) 발명의 명칭 **비정형 트랜잭션 비식별 데이터의 품질 측정 방법 및 장치**

(57) 요 약

본 실시예들은 복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 여부를 검증하고, 원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도 여부를 측정함으로써, 비식별화된 트랜잭션 데이터와 원본 데이터 간의 유사도에 대한 지표와 재식별 가능성에 대한 측정 지표를 제시하는 개인정보 비식별 데이터의 품질 측정 방법 및 장치를 제공한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/22 (2019.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 2015-0-00579
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기술진흥센터
연구사업명 정보통신방송연구개발사업
연구과제명 빅데이터 환경에서 비식별화 기법을 이용한 개인정보보호 기술 개발
기 여 율 1/1
과제수행기관명 (주)이지서티
연구기간 2018.04.01 ~ 2019.05.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

컴퓨팅 디바이스에 의한 개인정보 비식별 데이터의 품질 측정 방법에 있어서,

복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인의 식별여부를 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 위험성을 검증하는 단계; 및

원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도를 측정하는 단계를 포함하고,

상기 활용 품질 지표를 이용하여 원본 유사도를 측정하는 단계는 트랜잭션 레코드의 각 항목에 대한 1-항목 원본 유사도를 형성하고, 각 항목 별 원본 유사도와 원본 데이터 세트를 비교하여 상기 트랜잭션 레코드의 원본 유사도를 계산하는 단계를 포함하고,

상기 1-항목 원본 유사도는 항목 전체의 도메인 크기 대비 항목 계층 정보를 나타내는 자식 노드의 수로 계산을 하며, 상기 원본 유사도와 정보 손실이 반비례하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 방법.

청구항 2

제1항에 있어서,

상기 개인 중복도 모델을 이용하여 재식별 위험성을 검증하는 단계는,

개인 정보의 항목이 나타난 상기 트랜잭션을 발생시킨 각 개인들의 총 합인 개인 중복수를 지지도의 개념으로 사용하여 개인 중복도 기반 빈발 항목 집합을 찾아 개인 중복도 기반 테이블을 생성하는 단계; 및

상기 개인 중복도 기반 테이블을 기반으로, 상기 개인 중복도 모델에 따라 개인 중복도를 검증하는 단계를 포함하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 방법.

청구항 3

제2항에 있어서,

상기 개인 중복도를 검증하는 단계는,

상기 개인 중복도 모델이 1일 경우, 개인 최소 지지도를 1로 두고 상기 개인 최소 지지도를 넘지 못하는 항목 집합을 갖는 재식별 위험이 있는 레코드를 기반으로 재식별 위험도를 계산하는 단계; 및

상기 개인 중복도 모델이 2 이상일 경우, 특정 개인을 추정하는 가능성을 확률로 평가하여 상기 트랜잭션 데이터베이스에 따른 개인 추정 가능성 위험을 통해 개인 추정 가능성을 계산하는 단계를 더 포함하는 개인정보 비식별 데이터의 품질 측정 방법.

청구항 4

제1항에 있어서,

상기 활용 품질 지표를 이용하여 원본 유사도를 측정하는 단계는,

원본 데이터를 비식별 데이터로 변환하는 과정에서 비식별 처리에 위반하는 레코드를 제거하여 상기 원본 데이터 수보다 적은 수의 상기 비식별 레코드를 형성하며, 상기 비식별 레코드의 수 및 상기 원본 레코드의 수를 이용하여 잔존율을 계산하는 단계를 포함하는 개인정보 비식별 데이터의 품질 측정 방법.

청구항 5

삭제

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 트랜잭션 레코드의 원본 유사도는 상기 트랜잭션에 해당하는 각 항목들이 상기 항목들의 전체 크기로 가중치를 가지며, 상기 1-항목 원본 유사도에 상기 가중치를 부여하여 계산하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 방법.

청구항 8

제7항에 있어서,

상기 트랜잭션 레코드의 원본 유사도는 원본 레코드 세트와 비식별 결과 레코드 세트 사이의 유사도를 나타내며, 상기 비식별 결과 레코드 세트의 각각의 레코드에 대한 상기 원본 레코드와의 유사도를 계산하고, 상기 유사도를 평균 내어 결과 유사도를 산출하며,

상기 결과 유사도는 비식별 조치에 대한 상기 통계적 유사성 및 상기 활용 품질 지표로 활용하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 방법.

청구항 9

복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인의 식별여부를 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 위험성을 검증하는 재식별 검증부; 및

원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도를 측정하는 원본 유사도 측정부를 포함하고,

상기 원본 유사도 측정부는 트랜잭션 레코드의 각 항목에 대한 1-항목 원본 유사도를 형성하고, 각 항목 별 원본 유사도와 원본 데이터 세트를 비교하여 트랜잭션 레코드의 원본 유사도를 계산하는 원본 유사도 계산부를 포함하고,

상기 1-항목 원본 유사도는 항목 전체의 도메인 크기 대비 항목 계층 정보를 나타내는 자식 노드의 수로 계산을 하며, 상기 원본 유사도와 정보 손실이 반비례하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 장치.

청구항 10

제9항에 있어서,

상기 원본 유사도 측정부는,

원본 데이터를 비식별 데이터로 변환하는 과정에서 비식별 처리에 위반하는 레코드를 제거하여 상기 원본 데이터 수보다 적은 수의 상기 비식별 레코드를 형성하며, 상기 비식별 레코드의 수 및 상기 원본 레코드의 수를 이용하여 잔존율을 계산하는 잔존율 계산부를 포함하는 개인정보 비식별 데이터의 품질 측정 장치.

청구항 11

삭제

청구항 12

삭제

청구항 13

제9항에 있어서,

상기 트랜잭션 레코드의 원본 유사도는 상기 트랜잭션에 해당하는 각 항목들이 상기 항목들의 전체 크기로 가중치를 가지며, 상기 1-항목 원본 유사도에 상기 가중치를 부여하여 계산하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 장치.

청구항 14

제9항에 있어서,

상기 트랜잭션 레코드의 원본 유사도는 원본 레코드 세트와 비식별 결과 레코드 세트 사이의 유사도를 나타내며, 상기 비식별 결과 레코드 세트의 각각의 레코드에 대한 상기 원본 레코드와의 유사도를 계산하고, 상기 유사도를 평균 내어 결과 유사도를 산출하며,

상기 결과 유사도는 비식별 조치에 대한 통계적 유사성 및 활용성 품질 지표로 활용하는 것을 특징으로 하는 개인정보 비식별 데이터의 품질 측정 장치.

발명의 설명

기 술 분 야

[0001] 본 실시예가 속하는 기술 분야는 비정형 트랜잭션 비식별 데이터의 품질을 측정하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 최근에 들어 정보 통신 기술의 급속한 발전과 폭발적인 성장으로 인하여 데이터의 양이 늘어나고 이에 따라서 고도화된 분산 처리 시스템을 기반으로 하여 대용량 데이터의 수집 및 관리 기술과 분석 기술이 성숙해지고 있다.

[0004] 유통되는 대용량 데이터 중에 비정형 데이터인 트랜잭션 데이터는 개개인의 슈퍼마켓 구매 정보 및 병원의 진단 내역을 포함하여 다양한 정보를 가지며 다양한 형태의 항목 집합 형태로 나타난다. 또한, 트랜잭션 데이터는 마케팅 및 의약품과 같은 영역에서 굉장히 많은 새로운 지식을 발견하는데 도움을 주기 때문에 트랜잭션 데이터에 대한 연구 및 유통이 필수적이다.

[0005] 그러나 트랜잭션 데이터를 유통할 때, 트랜잭션 데이터 안에 포함되어 있는 개개인의 기록을 신원 정보에 연결하여 개인 정보 침해가 발생할 위험도 존재한다. 이와 같은 개인 정보의 침해를 막기 위해 개인 정보보호 관련 법률이 발의되면서 개인의 사생활 정보 유출 관련 이슈가 발생하지 않도록 현행 법규에서 제시하는 개인 식별 가능 정보가 포함된 데이터 활용 시 데이터의 비식별화 조치가 필수적으로 요구된다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국공개공보 제10-2018-0119104호 (2018.11.01)

발명의 내용

해결하려는 과제

[0007] 본 발명의 실시예들은 비식별 데이터와 원본 데이터 간의 유사도에 대한 지표와 재식별 가능성에 대한 측정 지표를 제시하는 데 발명의 주된 목적이 있다.

[0008] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0009] 본 실시예의 일 측면에 의하면, 컴퓨팅 디바이스에 의한 개인정보 비식별 데이터의 품질 측정 방법에 있어서, 복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 여부를 검증하는 단계 및 원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도 여부를 측정하는 단계를 포함하는 개인정보 비식별 데이터의 품질 측정 방법을 제공한다.

[0010] 본 실시예의 다른 측면에 의하면, 하나 이상의 프로세서 및 상기 하나 이상의 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로세서는 복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 여부를 검증하는 재식별 검증부 및 원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도 여부를 측정하는 원본 유사도 측정부를 포함하는 개인정보 비식별 데이터의 품질 측정 장치를 제공한다.

발명의 효과

[0011] 이상에서 설명한 바와 같이 본 발명의 실시예들에 의하면, 재식별 위험도 및 개인 추정 가능성은 여러 명의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 p 명의 해당 항목 집합에 나타나 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 '개인 중복도(p)' 모델을 검증 모델로 제시한다. 원본 유사도는 원본 트랜잭션 데이터와 비식별화 시킨 트랜잭션 데이터의 유사도가 얼마나 되는지 정량적으로 평가함에 따라 원본과 매우 유사하여 특정 개인이 재식별되는 문제가 있거나 원본과 지나치게 상이하여 통계적 유사성이 떨어지는 데이터의 활용 방지를 막을 수 있는 효과가 있다.

[0012] 여기에서 명시적으로 언급되지 않은 효과라 하더라도, 본 발명의 기술적 특징에 의해 기대되는 이하의 명세서에서 기재된 효과 및 그 잠정적인 효과는 본 발명의 명세서에 기재된 것과 같이 취급된다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일 실시 예에 따른 개인정보 비식별 데이터의 품질 측정 장치를 예시한 블록도이다.

도 2는 본 발명의 일 실시 예에 따른 비식별 위험 품질 지표의 흐름도이다.

도 3은 본 발명의 일 실시 예에 따른 잔존율 원본 유사도를 나타내는 흐름도이다.

도 4는 본 발명의 일 실시 예에 따른 항목 기반 원본 유사도의 흐름도이다.

도 5는 본 발명의 일 실시 예에 따른 표 2의 항목 계층 정보를 나타내는 도면이다.

도 6은 본 발명의 일 실시 예에 따른 개인정보 비식별 데이터의 품질 측정 방법을 예시한 흐름도이다.

도 7은 본 발명의 일 실시 예에 따른 개인정보 비식별 데이터의 품질 측정 방법을 예시한 흐름도이다.

도 8은 실시예들에서 사용되기에 적합한 컴퓨팅 디바이스를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0014] 이하, 본 발명을 설명함에 있어서 관련된 공지기능에 대하여 이 분야의 기술자에게 자명한 사항으로서 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하고, 본 발명의 일부 실시예들을 예시적인 도면을 통해 상세하게 설명한다.

[0015] 도 1은 개인정보 비식별 데이터의 품질 측정 장치를 예시한 블록도이다. 도 1에 도시한 바와 같이, 개인정보 비식별 데이터의 품질 측정 장치(10)는 재식별 검증부(100) 및 원본 유사도 측정부(200)를 포함한다. 개인정보 비식별 데이터의 품질 측정 장치(10)는 도 1에서 예시적으로 도시한 다양한 구성요소들 중에서 일부 구성요소를 생략하거나 다른 구성요소를 추가로 포함할 수 있다.

[0016] 개인정보 비식별 데이터의 품질 측정 장치(10)는 비식별 처리된 트랜잭션 데이터들의 재식별 위험성 및 개인 추

정도와 원본 유사도를 기반으로 비식별 데이터 세트에 대한 품질 평가 지표를 제시한다.

[0017] 트랜잭션 데이터는 미리 정의된 데이터 모델이 없거나 정형화 되지 않은 비정형 데이터의 한 형태로서, 개인의 슈퍼마켓 구매 정보 및 병원의 진단 내역을 포함하여 다양한 정보를 가지고 있으며, 다양한 항목 집합 형태로 나타난다. 이는 마케팅 및 의약품과 같은 영역에 도움을 줄 수 있다.

[0018] 비식별화는 특정 개인을 식별할 수 없도록 개인 정보의 일부 또는 전부를 변환하는 과정으로, 개인 식별 데이터를 다른 값으로 변환하거나 대체한다.

[0019] 재식별 검증부(100)는 개인 중복도 기반 테이블 생성부(110), 재식별 위험도 계산부(120) 및 개인 추정 가능성 위험도 계산부(130)를 포함한다.

[0020] 재식별 검증부(100)는 복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 여부를 검증한다.

[0021] 개인 중복도 기반 테이블 생성부(110)는 개인 정보의 항목이 나타난 트랜잭션을 발생시킨 각 개인들의 총합인 개인 중복수를 지지도의 개념으로 사용하여 개인 중복도 기반 빈발 항목 집합을 찾아 개인 중복도 기반 테이블을 생성한다.

[0022] 재식별 위험도 계산부(120)는 개인 정보 모델이 1인 경우, 개인 정보가 포함된 트랜잭션 레코드에 해당 트랜잭션 항목들이 특정 개인에게만 발생하여 특정 개인이 트랜잭션인지를 식별할 수 있는 트랜잭션 레코드를 검사하여 재식별 위험도를 계산한다.

[0023] 개인 추정 가능성 위험도 계산부(130)는 개인 정보 모델이 2 이상일 경우, 특정 개인을 추정하는 가능성을 확률로 평가하여 트랜잭션 데이터베이스에 따른 개인 추정 가능성 위험을 통해 개인 추정 가능성을 계산한다.

[0024] 원본 유사도 측정부(200)는 잔존율 계산부(210) 및 원본 유사도 계산부(220)를 포함한다.

[0025] 원본 유사도 측정부(200)는 원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도 여부를 측정한다.

[0026] 잔존율 계산부(210)는 원본 데이터를 비식별 데이터로 변환하는 과정에서 비식별 처리에 위반하는 레코드를 제거하여 원본 데이터 수보다 적은 수의 비식별 레코드를 형성하며, 비식별 레코드의 수 및 원본 레코드의 수를 이용하여 잔존율을 계산한다.

[0027] 원본 유사도 계산부(220)는 트랜잭션 레코드의 각 항목에 대한 1-항목 원본 유사도를 형성하고, 각 항목 별 원본 유사도와 원본 데이터 세트를 비교하여 트랜잭션 레코드의 원본 유사도를 계산한다. 1-항목 원본 유사도는 항목 전체의 도메인 크기 대비 항목 계층 정보를 나타내는 자식 노드의 수로 계산을 하며, 원본 유사도와 정보 손실이 반비례한다. 트랜잭션 레코드의 원본 유사도는 트랜잭션에 해당하는 각 항목들은 항목들의 전체 크기로 가중치를 가지며, 1-항목 원본 유사도에 상기 가중치를 부여하여 계산한다.

[0028] 트랜잭션 레코드의 원본 유사도는 상기 원본 레코드 세트와 비식별 결과 레코드 세트 사이의 유사도를 나타내며, 비식별 결과 레코드 세트의 각각의 레코드에 대한 원본 레코드와의 유사도를 계산하고, 유사도를 평균 내어 결과 유사도를 산출하며, 결과 유사도는 비식별 조치에 대한 통계적 유사성 및 활용 품질 지표로 활용한다.

[0029] 이하에서는 개인정보 비식별 데이터의 품질 측정 장치(10)가 비정형 트랜잭션을 통해 재식별 검증 기법을 설명하기로 한다.

[0030] 개인 중복도(p-중복성)는 여러 명의 개인들에 대한 개인정보를 담고 있는 항목들의 트랜잭션 데이터베이스에서 동일한 항목들이 최소 서로 다른 P(P는 자연수)명의 트랜잭션에 나타날 때 해당 항목들을 갖게 하여 특정 개인을 식별하지 못하게 한다.

[0031] 개인 정보를 담고 있는 트랜잭션 데이터베이스에서의 p-중복성은 동일한 생활정보에 대해서 최소 서로 다른 P명의 개인들이 해당 생활정보를 갖고 있어야 한다.

[0032] 아래의 표 1은 개인 정보 신상 테이블을 나타내고, 표 2는 개인별 원본 트랜잭션 테이블을 나타낸다.

표 1

[0033]

USER_ID	NAME	INCOME
UID1	John	3,400,000
UID2	Alice	1,500,000
UID3	Bob	5,640,000

표 2

[0034]

TRANSACTION	USER_ID	ITEMS
TID1	UID1	우유, 계란, 버터, 식빵
TID2	UID2	우유, 계란, 식빵
TID3	UID3	우유, 버터
TID4	UID1	우유, 계란, 커피
TID5	UID2	식빵, 라면
TID6	UID1	식빵, 버터

[0035] 상술한 표 1 및 표 2는 공격자가 John이 계란과 버터를 구입했다는 사실을 알고 있으며, 표 2의 개인별 트랜잭션 데이터베이스를 검색할 때 TID1이 John의 구입내역이라는 사실을 알게 된다면 John이 우유와 식빵도 같이 구입했다는 사실을 알 수 있다. 이는 개인의 구매 항목에 대하여 정보가 유출될 위험이 있다는 것을 안다.

[0036] 따라서, 트랜잭션 비식별 조치에 대한 위험 품질 평가 지표는 이러한 비식별 처리된 트랜잭션 데이터 세트에 대해 개인 중복도(p-중복성) 모델을 이용하여 비식별 트랜잭션 데이터 세트에 대한 재식별 위험도 및 개인 추정 가능성을 측정하여 사용한다.

[0037] 개인 중복도 검증은 개인 중복도 기반 빈발 항목 집합을 통해 검증한다. 개인 중복도 기반 빈발 항목 집합을 구하는 과정은 Apriori 방법과 유사하지만 항목이 나타난 트랜잭션을 발생시킨 각 개인들의 총 합인 개인 중복수(Personal Support)를 지지도 개념으로 사용한다.

[0038] 아래의 표 3은 표 2의 개인 구매 내역 테이블을 나타낸 것이다.

표 3

[0039]

USER_ID	NAME	ITEM SET
UID1	John	우유, 식빵, 버터, 계란, 라면
UID2	Alice	우유, 식빵, 계란, 커피
UID3	Bob	우유, 버터

[0040] 아래의 표 4는 표 3을 개인 중복수 기반으로 지지도를 구한 것으로 모든 항목 집합에 대해 나타낸 것이며, 트랜잭션 테이블의 개인 중복수 기반 테이블을 나타낸 것이다.

표 4

[0041]

1-ITEM	PS	ITEM SET	PS	ITEM SET	PS	ITEM SET	PS
우유	3	우유, 계란	2	계란, 커피	1	우유, 계란, 커피	1
계란	2	우유, 버터	2	식빵, 버터	1	우유, 버터, 식빵	1
버터	2	우유, 식빵	2	식빵, 커피	1	우유, 계란, 버터, 식빵	1
식빵	2	우유, 커피	1	식빵, 라면	1		
커피	1	계란, 버터	1	우유, 계란, 식빵	2		
라면	1	계란, 식빵	2	우유, 계란, 버터	1		

[0042] 상술한 표 4를 참조하면 전체 항목(집합)에 대한 개인 중복수는 총 3명(John, Alice, Bob)의 고객 중 몇 명의 개인 트랜잭션에 나타난 항목인지를 개인 중복수를 통해 확인할 수 있다.

[0043] 비식별 위험 품질 평가 지표는 생성된 개인 중복도 기반 빈발 항목 집합 테이블과 해당 비식별 트랜잭션 테이블

을 이용하여 개인 중복도 검증에서 발견된 레코드들을 사용할 수 있다.

[0044] 개인 중복도 검증은 개인 중복도 모델 즉, p 값에 따라서 검증하려는 지표가 두 가지로 나누어진다. 상기 p 값이 1인 경우는 재식별 위험도를 나타내며, p 값이 2 이상인 경우는 개인 추정 가능성을 계산한다. 개인 중복도 검증은 비식별 트랜잭션에 대한 개인 중복도 기반 테이블을 생성한 다음 이루어진다. 개인 중복도 p 값을 개인 최소 지지도(Minimum Personal Support)로 정의하고 개인 최소 지지도를 넘지 못하는 항목 집합을 갖는 트랜잭션 레코드의 수를 구한다.

[0045] 재식별 위험도와 개인 추정 가능성의 비율은 표준적인 측정 방법으로 활용하기 위해 전체 데이터 세트의 크기 대비로 계산하여 재식별 위험도 및 개인 추정 가능성의 크기가 항상 [0,1] 사이의 값을 갖게 한다.

[0046] 이하에서는 재식별 위험도와 개인 추정 가능성 위험도를 비식별 위험 품질 지표로 사용하는 것에 대해 설명하기로 한다. 도 2는 비식별 위험 품질 지표의 흐름도이다.

[0047] 개인 중복도(p)는 분석되는 데이터의 성격에 따라 그 값을 특정하여 사용할 수 있으며, p 값이 커질수록 비식별화 정도는 높아지지만, p 값이 무한대라면 데이터베이스의 모든 내용의 구분이 불가능할 것이며, 반대로 p 값이 1이라면, 기존의 데이터베이스와 동일한 형태로 모든 값의 구별이 가능하게 되므로, 분석되는 데이터의 성격에 따라 적절한 값을 설정한다.

[0048] 개인 중복도 검증은 개인 중복도(p)의 값이 1 또는 2 이상일 때로 나뉘며, p=1일 경우, 재식별 위험도를 계산하며, p≥2일 경우, 개인 추정 가능성을 계산한다.

[0049] 1-중복도 검증은 개인 정보가 담겨있는 트랜잭션 레코드에 해당 트랜잭션의 항목들이 특정 개인 한명에게만 발생하여 특정 개인의 트랜잭션인지를 식별할 수 있게 하는 트랜잭션 레코드를 검사한다. 즉, 재식별 위험성이 존재하는 트랜잭션 레코드를 검증한다. 검증 방법은 개인 최소 지지도(Minimum Personal Support)인 p 값을 1로 두어 그 지지도를 넘지 못하는 항목 집합을 가지는 트랜잭션 레코드의 수를 구한다.

[0050] 1-중복도 검증에서 식별되는 레코드들은 항목 집합의 조합이 유일한 값 조합이므로 해당 레코드들을 재식별 위험이 있는 레코드로 정의할 수 있다. 재식별 위험도를 표준적인 측정 방법으로 활용하기 위하여 재식별 위험도를 다음과 같이 정의한다.

수학식 1

[0051] $$Risk_{uniq} = \frac{RC.count(1)}{Count(|T|)}$$

[0052] 상술한 수학식 1을 참조하면, 개인 중복도가 1인 경우, $RC.count(1)$ 은 재식별 위험성이 있는 트랜잭션 레코드, $Count(|T|)$ 는 비식별 트랜잭션 전체 레코드를 의미하며, 재식별 위험도를 [0,1]로 상기 수학식 1로 정의할 수 있다. 이때, T는 검사 할 비식별 트랜잭션이다.

[0053] 개인 중복도(p)가 2 이상일 경우, p-중복성에서 검증된 트랜잭션 데이터는 모든 항목 집합에 대하여 유일한 항목 집합이 존재하지 않아 개인정보를 담고 있는 항목들을 통해 특정 개인을 식별할 수 있는 데이터가 존재하지 않는다. 특정 개인을 재식별 할 수는 없지만 개인 중복도 값을 2이상으로 두어 검증을 할 때에는 특정 개인을 추정할 수 있는 가능성을 평가할 수 있다.

[0054] 아래의 표 5는 원본 트랜잭션 데이터베이스인 표 2를 항목 계층 구조를 이용하여 일반화 조치가 일어난 비식별 트랜잭션 데이터베이스이며, 표 2의 비식별 조치된 트랜잭션 테이블을 나타낸다.

표 5

[0055]

TRANSACTION	USER_ID	ITEMS
TID1	UID1	우유, 계란, 식빵
TID2	UID2	우유, 계란, 식빵
TID3	UID3	계란, 식빵
TID4	UID1	우유, 계란
TID5	UID2	식빵, 음식

TID6	UID1	식빵, 음식

[0056] 상술한 표 5의 트랜잭션은 모두 1-중복도 검증에서 발견되는 레코드가 존재하지 않는다. 하지만 2-중복도 검증을 하면 4개의 레코드(TID1, TID2, TID5, TID6)가 발견된다. 이는 <우유, 계란, 식빵>의 항목 집합을 가지는 트랜잭션 데이터가 (TID1, UID1)과 (TID2, UID2)만이 존재하고 <식빵, 음식>의 항목 집합을 갖는 트랜잭션 데이터가 (TID5, UID2)와 (TID6, UID1)이기 때문이다.

[0057] TID3의 경우 해단 항목 집합<계란, 식빵>을 가지는 레코드가 (TID1, UID1)과 (TID2, UID2)가 있어 개인 중복수가 3이기 때문에 발견되지 않는다. 이와 같이, 재식별 위험성을 가지는 레코드는 아니지만 낮은 p 값을 가지는 개인 중복도 검증에서 발견되는 레코드들을 특정 개인을 추정할 수 있는 가능성이 높은 레코드들임을 알 수 있다.

[0058] 따라서, 개인 추정 가능성은 특정 개인을 추정할 수 있는 확률로 개인 중복도 검증으로 검증을 하게 되면 트랜잭션 데이터 베이스에서 동일한 항목들이 최소 서로 다른 p명의 트랜잭션에 나타나기 때문에 $\frac{1}{P}$ 만큼의 개인 추정 가능성을 갖는다.

수학식 2

[0059] $$Risk_{presume} = \frac{RC.count(p)}{Count(|T|)}$$

[0060] 상술한 수학식 2를 참조하면, 개인 중복도가 p, $RC.count(p)$는 개인 추정 가능성이 있는 트랜잭션 레코드, $Count(|T|)$ 비식별 트랜잭션 전체 레코드를 의미하며, 개인 추정 가능성 위험도를 [0,1]로 수학식 2를 정의한다. 이때, T는 검사 할 비식별 트랜잭션이다.

[0061] 따라서, 표 5의 개인 추정 가능성을 상술한 수학식 2를 사용하여 계산하면 2-중복성 검증을 하였을 때, 전체 트랜잭션 데이터베이스의 $\frac{2}{3}$ 만큼의 데이터가 $\frac{1}{2}$ 확률로 개인 추정 가능성 위험이 존재한다고 할 수 있다.

[0062] 비식별 모델에 관한 검증 방법으로 재식별 위험도 및 개인 추정 가능성은 레코드의 유통에 대한 위험 품질 지표일 수 있다. 유통에 대한 위험 품질 지표와 반대로 원본 유사도는 원본 레코드 세트와 결과 레코드 세트 간의 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표이다. 즉, 생성된 데이터의 활용성을 정량적으로 평가한다. 원본 유사도 측정은 크기가 항상 [0,1] 사이의 값을 가지게 하여 정량적으로 평가할 수 있는 일반적인 품질 지표로 쓰일 수 있다.

[0063] 활용 품질 지표는 두 가지로 나누어 원본 유사도를 계산 및 평가한다. 첫 번째는 잔존율로 비식별 조치 과정에서 삭제되는 데이터의 비율을 지표로 분석함으로써 비식별 조치된 유통 데이터의 활용성을 평가하고, 두 번째는 항목 기반 원본 유사도이다.

[0064] 이하에서는 해당 비식별 프라이버시 모델에 위반하는 레코드를 제거하여 비식별 데이터를 형성하는 과정을 설명하기로 한다. 도 3에서는 잔존율 원본 유사도를 나타내는 흐름도가 도시되어 있다.

[0065] 잔존율은 비식별 변환 과정에서 해당 비식별 처리에 위반하는 레코드를 제거하여 이에 따라 원본 레코드 수보다 적은 수의 결과 레코드를 보유하고 있다. 이렇게 제거되는 레코드의 수가 많을수록 데이터 세트의 활용성은 떨어지게 되므로 원본 레코드 수 대비 결과 레코드 수를 비율로 잔존율을 정의하여 데이터 활용성에 대한 지표로 평가한다. 도 3은 잔존율을 원본 유사도로 계산하는 설명에 대한 도면이다.

수학식 3

[0066] $$SIM_{rem}(T) = \frac{|\text{비식별 레코드 수}|}{|\text{원본 레코드 수}|}$$

[0067] 상술한 수학식 3을 참조하면, 트랜잭션 데이터 세트(T)에 대하여 잔존율은 비식별 레코드의 수를 원본 레코드의 수로 나누어 구할 수 있다. 잔존율은 비식별된 레코드의 수가 클수록 크며, 이는 비식별 처리에 위반하는 레코드가 적게 제거될수록 잔존율이 크다. 따라서, 잔존율이 클수록 데이터 세트의 활용성도 크다.

[0068] 이하에서는 전체 비식별 처리된 트랜잭션 데이터에 대한 항목 기반 원본 유사도를 구하는 프로세서에 대해 설명하기로 하며, 트랜잭션 레코드의 각 1-항목에 대한 항목 유사도에 대해 정의하고 최종적으로 전체 트랜잭션 데이터 세트의 원본 유사도까지 설명한다. 도 4에서는 항목 기반 원본 유사도의 흐름도가 도시되어 있다.

[0069] 항목 기반 원본 유사도 흐름도와 같이 항목 기반 원본 유사도의 계산은 트랜잭션 레코드의 각 1-항목에 대한 항목 유사도를 정의하고, 이를 기준으로 트랜잭션 레코드에 해당하는 항목의 전체 크기로 각 항목마다 가중치를 부여하여 트랜잭션 레코드에 대한 원본 유사도를 정의한다. 이 후 모든 트랜잭션 레코드 유사도 값을 평균내어 전체 비식별 트랜잭션 데이터 세트의 원본 유사도 값을 정의한다.

[0070] 트랜잭션 데이터의 원본 유사도는 트랜잭션 데이터가 항목 집합의 형태로 나타나게 되어 각 항목 별로 원본 유사도를 원본 데이터 세트와 비교하여 구할 수 있다.

[0071] 항목을 구성하는 데이터는 항목에 대한 계층 정보가 포함되어 있는데 이 계층 정보는 트랜잭션 데이터의 비식별 조치가 이루어 질 때 사용되는 정보이다. 개인의 기록을 신원 정보에 연결하여 개인 정보 침해가 일어날 위험이 있는 항목에 대해 계층 정보를 이용하여 일반화(Generalization)를 한다. 일반화는 초기 도메인의 다른 값이 대상 도메인의 단일 값에 매핑이 되도록 초기 도메인에서 다른 도메인으로 값이 매핑되는 것을 의미한다. 얼마만큼의 일반화가 이루어졌는지 그 양을 계산하여 정보 손실의 양을 계산한다.

[0072] 1-항목은 트랜잭션 데이터를 구성하는 단위로 1-항목 집합 I = {i1,i2....,ii}는 트랜잭션 전체의 도메인으로 정의할 수 있다. ($|I|$ = n, $|I|$는 I 집합의 전체 원소 개수) 1-항목으로 이루어진 원본 트랜잭션을 T = {t1,t2....,tt}, 비식별이 이루어진 트랜잭션을 P = {p1, p2,.....,pk}, 비식별이 이루어진 트랜잭션의 항목($\{x'_k | x'_k \in P\}$)은 각 트랜잭션마다 일대일로 대응된다.

[0073] 비식별 조치가 이루어진 일반화된 1-항목에 대하여 원본과 비교하였을 때, 원본과의 차이를 정량적으로 계산하기 위해 근접 상위 노드를 정의한다. 예를 들어, a1,.....,al이 조상 노드이고, u는 a1,.....,al의 자식 노드 외에 다른 자식 노드를 갖지 않는다. 이때, u는 a1,.....,al의 근접 상위 노드라 말한다. 근접 상위 노드는 계층 정보 트리에서, 입사귀 노드 a1,a2....,al만을 자식 노드로 갖는 상위 노드를 S에 대한 근접 상위 노드라 한다.

[0074] 근접 상의 노드는 1-항목 원본 유사도(1-item Similarity)를 이용하여 계산할 수 있다.($|u|$= 입사귀 노드 a1,.....,al의 총 개수 = l) 원본 트랜잭션 항목($\{x_k | x_k \in T\}$)의 근접 상위 항목을 ux 로 표현하면 두 가지 경우로 나누어 1-항목 원본 유사도를 계산할 수 있다.

[0075] 이하에서는 표 2가 표 5로 비식별 되는 트랙잭션을 구성하는 항목들에 대한 계층 정보를 설명한다. 도 5는 표 2의 항목 계층 정보를 나타내는 도면이다.

[0076] 첫 번째로 원본 유사도가 0이 되는 경우는 비식별화된 황목의 일반화가 계층 정보의 최고 계층 노드(root node)까지 이루어져 그 모든 정보가 손실되었을 때 발생하는데 이는 도 5로 설명할 수 있다.

[0077] 도 5 및 표 2와 표 5에서 나타낸 바와 같이 (TID1, UID1)의 항목인 '버터'는 최고 계층 노드(root node)인 '*'로 일반화가 되었다. 이는 트랜잭션 TID1이 비식별 조치에 대하여 만족하지 못해 항목 '버터'의 일반화가 일어나야 하는데 항목 '버터'의 근접 상위 노드인 항목 '육류'로 일반화가 되어도 비식별 조치에 만족하지 못한다. 같은 방식으로 '육류'의 근접 상위 노드로 일반화하여 '음식'으로 일반화가 되어도 비식별 조치에 만족하지 못하기 때문에 최종적으로 최고 계층 노드인 '*'로 일반화가 된 것이다. 항목의 일반화가 계층 정보의 최고 계층 노드(*)까지 이루어지면 이 항목을 전지 작업(Pruning)시키며, TID1의 비식별 트랙잭션은 {우유, 계란,

식빵}이 된다. 이는 항목에 대한 모든 정보가 손실되었기 때문에 TID1의 항목 '버터'에 대한 원본 유사도는 0으로 계산한다.

[0078] Pruning은 규칙의 일부가 잘리거나 무시되어도 좋은가를 결정하여 탐색 공간을 줄여줄 수 있다.

[0079] TID3의 항목 {우유, 계란, 커피} 중 1-항목 '커피'는 위와 같은 방식으로 근접 상위 노드로 일반화가 일어나다가 모든 상위 계층 항목이 비식별 조치에 만족하지 못하여 최종적으로 최고 계층 노드로 일반화가 되고 Pruning되어 원본 유사도가 0이 된다.

[0080] 두 번째로 원본 유사도를 근접 상위 노드의 크기를 이용하여 계산하는 경우는 TID5의 항목 집합 {식빵, 라면} 중 1-항목 '라면'이 '음식'으로 일반화된 것을 예로 들 수 있다. 이는 원본 트랜잭션 1-항목 $\{x_k | x_k \in T\}$가 최고 계층 항목이 아닌 근접 상위 항목인 ux 보다 더 상위 항목으로 일반화 된 경우에 대한 원본 유사도 계산 방식이다. '간식'은 항목 '라면'의 근접 상위 노드이므로 '간식'의 크기를 1이라 할 수 있다. 최종적으로 '간식'에서 '음식'으로 일반화가 되었기 때문에 '음식'의 크기를 계산해야 한다. '음식'의 크기는 4로 계산하는데 이는 '음식'이 1-항목 {계란, 버터, 식빵, 라면}을 자식 입사귀 노드로 갖기 때문이다.

[0081] 1-항목 원본 유사도는 전체 항목 도메인의 크기 대비로 계산하여 항상 그 범위가 [0,1] 사이의 값을 갖는다. 항목이 일반화가 일어났을 때의 원본 유사도는 일반화가 일어나지 않았을 때의 원본 유사도 1을 기준으로 항목 도메인 크기 대비 얼마만큼의 항목에 대한 일반화가 일어나 결과의 정보 손실이 발생하였는지 계산한다

[0082] 따라서, (라면 -> 음식)의 원본 유사도는 전체 항목의 도메인의 크기($|T|$ = 6)를 분모로 항목 '간식'의 근접 상위 노드의 크기($|u_x|$ = 4)를 분자로 계산하여 총 $\frac{2}{3}$ 만큼의 정보의 손실이 일어났다고 계산을 하여 트랜잭션 TID5의 (라면 -> 음식) 원본 유사도는 1- $\frac{2}{3}$ = $\frac{1}{3}$ 로 계산할 수 있다. 같은 방식으로 트랜잭션 TID6 (버터 -> 음식) 원본 유사도는 (라면 -> 음식)과 같이 1- $\frac{2}{3}$ = $\frac{1}{3}$ 로 계산할 수 있다.

[0083] 아래의 표 6 및 표 7은 서로 다른 크기의 근접 상위 항목으로 일반화가 일어난 예시를 나타낸 것이다. 표 6은 개인별 원본 트랜잭션 테이블2이며, 표 7은 표 6의 비식별 조치된 트랜잭션 테이블이다.

표 6

[0084]

TRANSACTION	USER_ID	ITEMS
TID1	UID1	우유, 계란, 버터, 식빵
TID2	UID2	우유, 계란, 식빵
TID3	UID3	계란, 우유
TID4	UID1	우유, 계란, 커피
TID5	UID2	식빵, 라면
TID6	UID1	식빵, 버터

표 7

[0085]

TRANSACTION	USER_ID	ITEMS
TID1	UID1	우유, 계란, 식빵
TID2	UID2	우유, 계란, 식빵
TID3	UID3	계란, 음료
TID4	UID1	우유, 계란
TID5	UID2	식빵, 음식
TID6	UID1	식빵, 음식

[0086] 상술한 표 6 및 표 7에 따르면, TID3, TID4는 각 1-항목 {우유, 커피}가 같은 두 단계 상위 항목인 '음식'로 일반화가 되어 비식별 조치가 만족된 트랜잭션이 되었다. TID5, TID6의 각 항목 {라면, 버터}는 같은 두 단계 상위 항목인 '음식'으로 일반화가 되어 비식별 트랜잭션이 되었다.

[0087] TID3의 항목 '우유'에 대한 원본 유사도를 계산하게 되면 항목 전체 도메인의 크기($|T|$ = 6)를 분모로 항목 '우유'의 두 단계 상위 노드의 크기($|u_x|$ = 2)를 분자로 계산하여 총 $\frac{1}{3}$ 만큼의 정보의 손실이 일어났다고 계산을 하여 TID3의 항목 (우유 -> 음료) 원본 유사도는 1- $\frac{1}{3}$ = $\frac{2}{3}$ 로 계산할 수 있다. 반면 TID5의 항목 (라면 -> 음식)에 대한 원본 유사도는 상술하듯이 1- $\frac{2}{3}$ = $\frac{1}{3}$ 가 나온다.

[0088] 서로 다른 근접 상위 노드를 갖는 각 1-항목에 대하여 같은 단계 크기만큼 일반화가 되었다고 해도 근접 상위 노드의 크기가 다르면 변환된 결과 1-항목의 정보 손실은 다르게 계산된다.

[0089] 따라서, 원본 유사도 계산 방식을 정리하면 1-항목 원본 유사도의 계산은 해당 항목이 얼마나 일반화가 되어 정보 손실이 일어났는지를 항목 전체의 도메인 크기 대비 자식 잎사귀 노드의 수로 계산을 하여 구한 뒤 원본 유사도와 정보 손실이 반비례함을 이용하여 계산한다. 이를 원본 유사도(SIMitem)라 정의하고 다음과 같이 정의한다.

수학식 4

[0090]
$$SIM_{item}(x) = 1 - \frac{|u_x|}{|T|}$$

[0091] 상술한 수학식 4를 참조하면, 1-항목 원본 유사도(1-item Similarity)는 원본의 1-항목이 비식별 과정을 거치며 얼마나 많은 정보 손실을 하였는지 계산하고 정보 손실의 양과 반비례을 이용하여 1-항목에 대한 원본 유사도를 측정한다.

[0092] 이하에서, 트랜잭션 레코드 원본 유사도는 특정 원본 레코드와 비식별 트랜잭션 레코드 쌍 사이의 유사성을 의미하며 비식별 트랜잭션 레코드가 가지고 있는 각각의 항목에 대한 원본 항목과의 유사도를 계산하고, 그 값을 기준으로 계산한다. 트랜잭션에 해당하는 각 항목들은 동일한 가중치(Weight)를 갖는다. 도 6은 트랜잭션 원본 유사도의 흐름도이다.

[0093] 예를 들어 i번째 트랜잭션 Ti의 항목 집합이 {a, b, c, d, e, f}라 하였을 때, 각 항목들은 총 $\frac{1}{6}\frac{1}{|Ti|}$ 의 가중치를 갖게 된다. 각 항목의 동일한 가중치와 수학식 5를 이용하여 1-항목 유사도를 곱한 값을 각 항목 당 해당 트랜잭션 항목 집합에서 영향을 준 지표로 계산하여 그 합을 해당 트랜잭션 레코드의 원본 유사도로 계산한다.

[0094] 표 6 및 표 7을 기준으로 트랜잭션(TID4, UID1)의 원본 유사도를 계산하게 되면 각 항목은 트랜잭션의 항목 집합의 크기인 3(우유, 계란, 커피)을 동일한 가중치를 갖는다. 항목 '우유, 계란'은 일반화 되지 않고 원본을 유지하여 1-항목 원본 유사도의 값은 1이고, 항목 '커피'의 경우 상위 항목인 '음료'로 일반화되어 원본 유사도의 값이 $\frac{2}{3}$ 이다. 그러므로 트랜잭션(TID4, UID1)의 레코드 원본 유사도는 $\frac{1}{3}$ (1+1+ $\frac{2}{3}$) = $\frac{8}{9}$ 의 값을 갖는다. 위의 계산 방식에 따라서 트랜잭션 레코드 원본 유사도는 아래 수학식 5와 같이 나타낼 수 있다.

수학식 5

[0095]
$$SIM_{item} = \frac{1}{|T|}\sum_{a_I \in T_I}\left(1 - \frac{|u_{x_k}|}{|I|}\right)$$

[0096] 상술한 수학식 5를 참조하면, I는 전체 도메인, T는 트랜잭션의 항목 집합의 크기이다. 수학식 5는 1-항목 원본

유사도에 동일한 가중치를 적용하여 항목의 원본 유사도 값을 모두 더한 식이다.

[0097] 결론적으로 전체 비식별 트랜잭션 테이블의 원본 유사도는 원본 레코드 세트와 비식별 결과 레코드 세트 사이의 유사도를 의미하며, 비식별 결과 레코드 세트의 각각의 레코드에 대해 원본 레코드와의 레코드 유사도를 계산하고, 그 값을 평균내에 계산한다. 그 결과 값을 비식별 조치에 대한 통계적 유사성 및 활용성 품질 지표로 사용한다.

[0098] 도 7은 본 발명의 다른 실시예에 따른 개인정보 비식별 데이터의 품질 측정 방법을 예시한 흐름도이다. 개인정보 비식별 데이터의 품질 측정 방법은 컴퓨팅 디바이스에 의하여 수행될 수 있으며, 개인정보 비식별 데이터의 품질 측정 장치가 수행하는 동작에 관한 상세한 설명과 중복되는 설명은 생략하기로 한다.

[0099] 단계 S710에서, 컴퓨팅 디바이스는 복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 여부를 검증한다.

[0100] 단계 S710은 개인 정보의 항목이 나타난 트랜잭션을 발생시킨 각 개인들의 총 합인 개인 중복수를 지지도의 개념으로 사용하여 개인 중복도 기반 빈발 항목 집합을 찾아 개인 중복도 기반 테이블을 생성하고, 생성된 개인 중복도 기반 테이블을 기반으로, 개인 중복도 모델에 따라 개인 중복도를 검증한다.

[0101] 개인 중복도 검증에 발견되는 레코드를 계산하는 것은 의사코드로 표현하면 다음과 같다.

[0102] for each item set in in T

[0103] RC.count ← 0

[0104] if(for all record of PFI.ps p)

[0105] for n=0; n<T.item.size; n++ do

[0106] if in $\in$ PFI of the record then

[0107] end if

[0108] if in $\notin$ PFI of the record then

[0109] RC.count ++ // Uniqueness record count

[0110] end if

[0111] end for

[0112] return RC.count

[0113] 개인 중복도 검증 알고리즘에서, T는 검사 트랜잭션 테이블, PFI는 개인 중복도 기반 빈발 항목 집합 테이블, p는 개인 최소 지지도를 의미한다. 개인 중복도 검증 알고리즘은 T, PFI 및 p가 입력되어 p-중복도 검증 레코드 카운트(RC)가 출력된다.

[0114] 개인 중복도 검증은 개인 중복도 모델이 1일 경우, 개인 최소 지지도를 1로 두고 개인 최소 지지도를 넘지 못하는 항목 집합을 갖는 재식별 위험이 있는 레코드를 기반으로 재식별 위험도를 계산하고, 개인 중복도 모델이 2 이상일 경우, 특정 개인을 추정하는 가능성을 확률로 평가하여 트랜잭션 데이터베이스에 따른 개인 추정 가능성 위험을 통해 개인 추정 가능성을 계산한다.

[0115] 단계 S720에서, 컴퓨팅 디바이스는 원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도 여부를 측정한다.

[0116] 원본 유사도는 원본 트랜잭션 데이터와 비식별화 시킨 트랜잭션 데이터의 유사도가 얼마나 되는지를 평가한다. 원본 유사도가 매우 유사한 경우, 특정 개인의 정보가 재식별되는 문제가 있으며, 원본 유사도가 지나치게 상이한 경우, 통계적 유사성이 떨어지는 데이터를 활용하는 문제가 있다.

[0117] 활용 품질 지표를 이용하여 원본 유사도를 여부를 측정하는 단계(S720)는 원본 데이터를 비식별 데이터로 변환하는 과정에서 비식별 처리에 위반하는 레코드를 제거하여 원본 데이터 수보다 적은 수의 상기 비식별 레코드를 형성하며, 비식별 레코드의 수 및 원본 레코드의 수를 이용하여 잔존율을 계산한다.

[0118] 또한, 활용 품질 지표를 이용하여 원본 유사도를 여부를 측정하는 단계(S720)는 트랜잭션 레코드의 각 항목에 대한 1-항목 원본 유사도를 형성하고, 각 항목 별 원본 유사도와 원본 데이터 세트를 비교하여 트랜잭션 레코드의 원본 유사도를 계산한다. 1-항목 원본 유사도는 항목 전체의 도메인 크기 대비 항목 계층 정보를 나타내는 자식 노드의 수로 계산을 하며, 원본 유사도와 정보 손실이 반비례한다.

[0119] 트랜잭션 레코드의 원본 유사도는 트랜잭션에 해당하는 각 항목들이 항목들의 전체 크기로 가중치를 가지며, 상기 1-항목 원본 유사도에 가중치를 부여하여 계산한다. 트랜잭션 레코드의 원본 유사도는 원본 레코드 세트와 비식별 결과 레코드 세트 사이의 유사도를 나타내며, 비식별 결과 레코드 세트의 각각의 레코드에 대한 원본 레코드와의 유사도를 계산하고, 유사도를 평균 내어 결과 유사도를 산출하며, 결과 유사도는 비식별 조치에 대한 통계적 유사성 및 활용 품질 지표로 활용한다.

[0120] 도 7에서는 각각의 과정을 순차적으로 실행하는 것으로 개재하고 있으나 이는 예시적으로 설명한 것에 불과하고, 이 분야의 기술자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 도 7에 기재된 순서를 변경하여 실행하거나 또는 하나 이상의 과정을 병렬적으로 실행하거나 다른 과정을 추가하는 것으로 다양하게 수정 및 변형하여 적용 가능할 것이다.

[0121] 도 8은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 디바이스를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술되지 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0122] 도시된 컴퓨팅 환경은 개인정보 비식별 데이터의 품질 측정 장치(10)를 포함한다. 일 실시예에서, 개인정보 비식별 데이터의 품질 측정 장치(10)는 타 단말과 신호를 송수신하는 모든 형태의 컴퓨팅 디바이스일 수 있다.

[0123] 개인정보 비식별 데이터의 품질 측정 장치(10)는 적어도 하나의 프로세서(810), 컴퓨터 판독 가능한 저장매체(820) 및 통신 버스(860)를 포함한다. 프로세서(810)는 개인정보 비식별 데이터의 품질 측정 장치(10)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(810)는 컴퓨터 판독 가능한 저장 매체(820)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(810)에 의해 실행되는 경우 개인정보 비식별 데이터의 품질 측정 장치(10)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0124] 컴퓨터 판독 가능한 저장 매체(820)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(820)에 저장된 프로그램(830)은 프로세서(810)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독한 가능 저장 매체(820)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 개인정보 비식별 데이터의 품질 측정 장치(10)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0125] 통신 버스(860)는 프로세서(810), 컴퓨터 판독 가능한 저장 매체(820)를 포함하여 개인정보 비식별 데이터의 품질 측정 장치(10)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0126] 개인정보 비식별 데이터의 품질 측정 장치(10)는 또한 하나 이상의 입출력 장치(미도시)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(840) 및 하나 이상의 통신 인터페이스(850)를 포함할 수 있다. 입출력 인터페이스(840) 및 통신 인터페이스(850)는 통신 버스(860)에 연결된다. 입출력 장치(미도시)는 입출력 인터페이스(840)를 통해 개인정보 비식별 데이터의 품질 측정 장치(10)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(미도시)는 개인정보 비식별 데이터의 품질 측정 장치(10)를 구성하는 일 컴포넌트로서 개인정보 비식별 데이터의 품질 측정 장치(10)의 내부에 포함될 수도 있고, 개인정보 비식별 데이터의 품질 측정 장치(10)와는 구별되는 별개의 장치로 컴퓨팅 디바이스와 연결될 수도 있다.

[0127] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 매체에 기록될 수 있다. 컴퓨터 판독 가능한 매체는 실행을 위해 프로세서에 명령어를 제공하는데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는

이들의 조합을 포함할 수 있다. 예를 들면, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0128] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

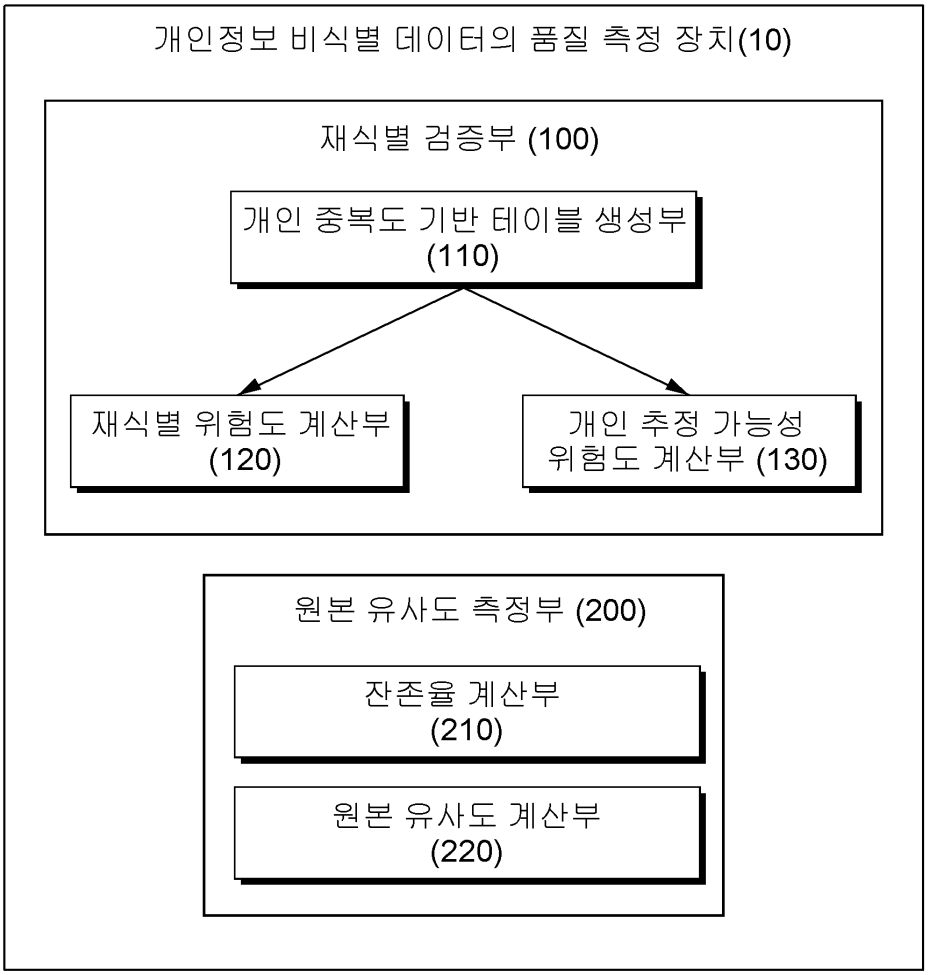
[0129] 10: 개인정보 비식별의 품질 측정 장치

100: 재식별 검증부

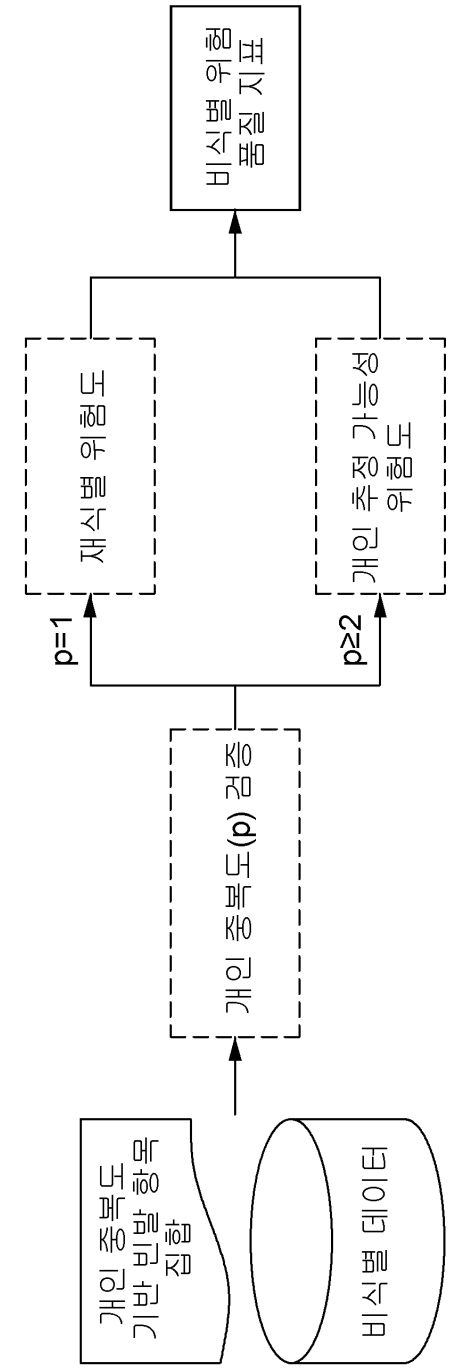
200: 원본 유사도 측정부

도면

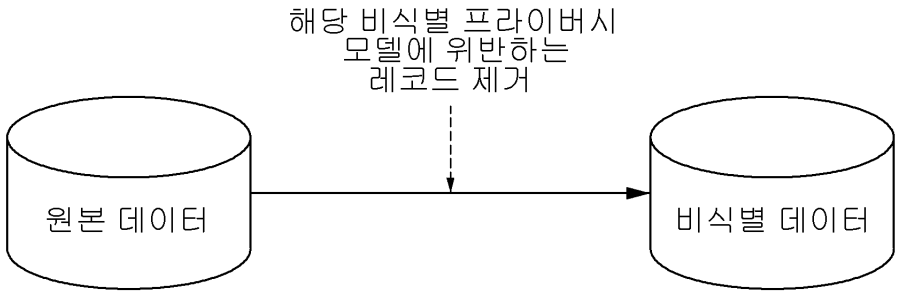
도면1



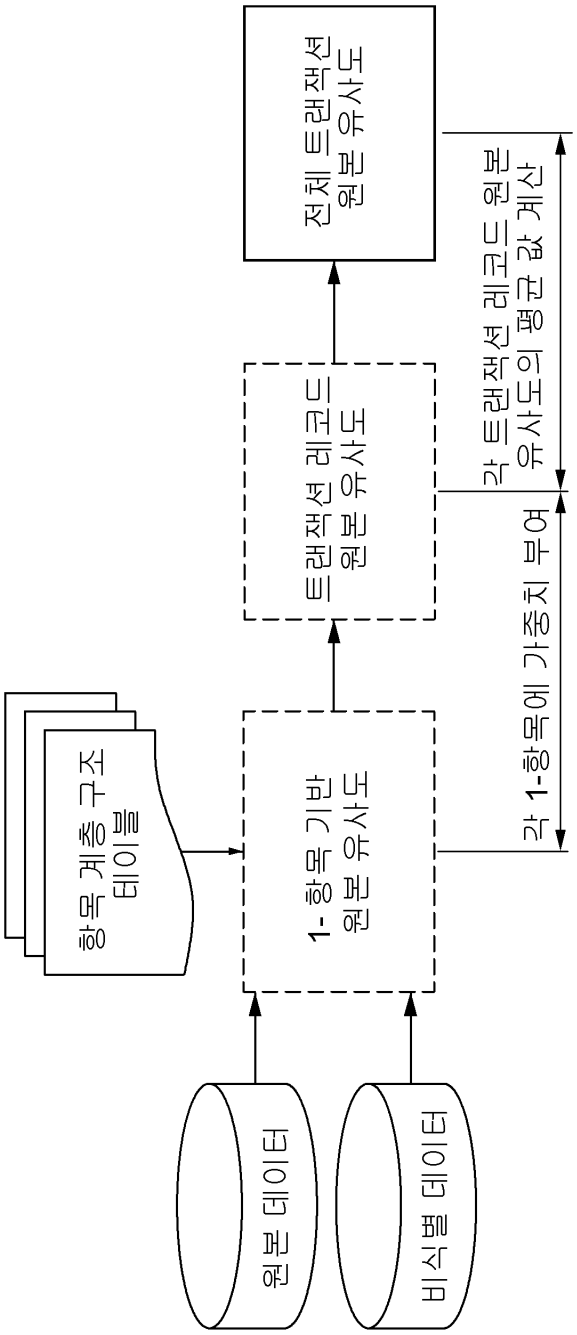
도면2


비식별 데이터
개인 중복도 기반 빈발 항목 집합
개인 중복도(p) 검증
p=1
p≥2
재식별 위험도
개인 추정 가능성 위험도
비식별 위험 품질 지표

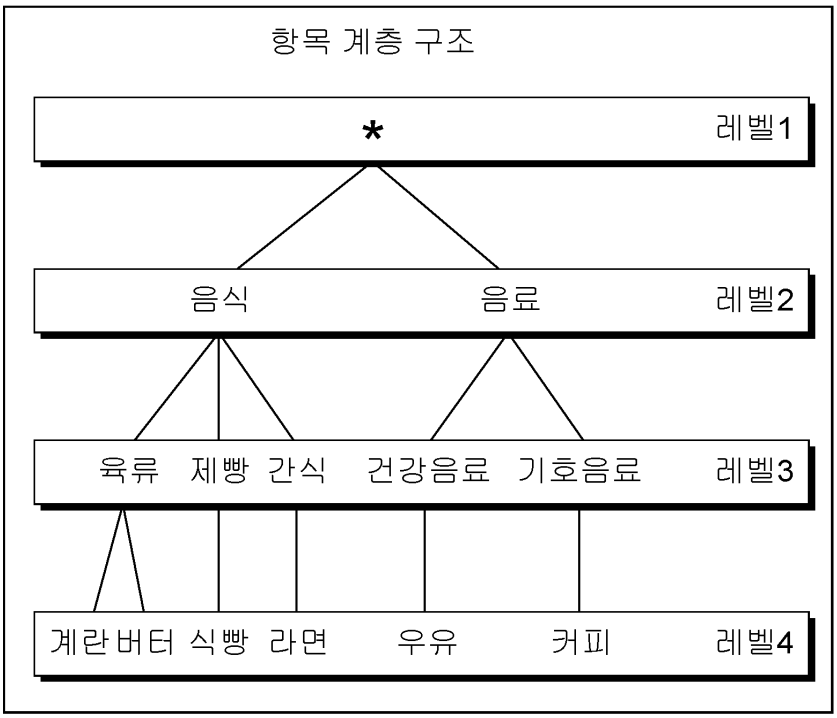
도면3



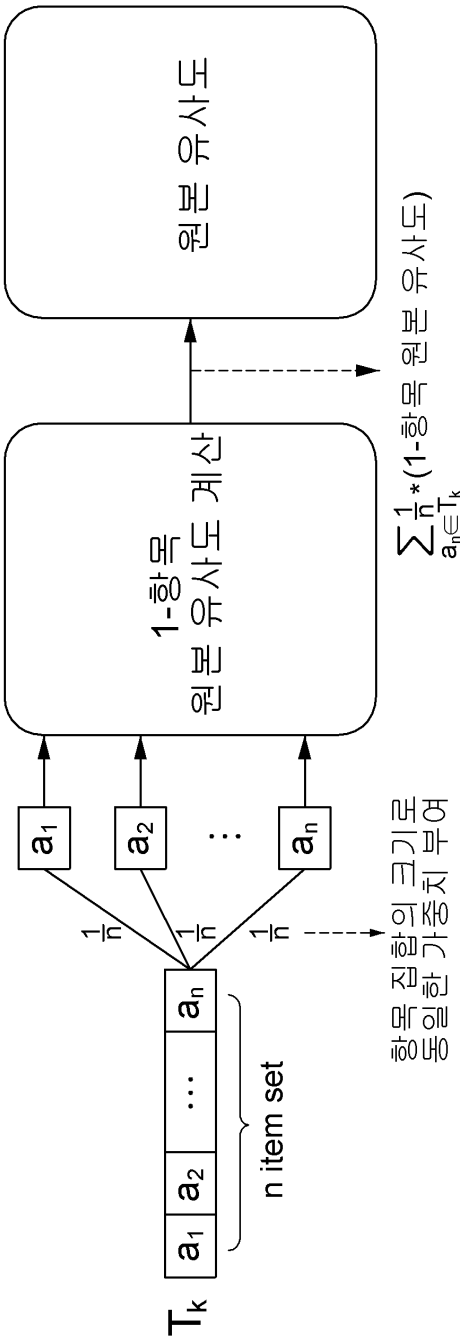
도면4


전체 트랜잭션
원본 유사도
각 트랜잭션 레코드 원본
유사도의 평균 값 계산
트랜잭션 레코드
원본 유사도
각 1-항목에 가중치 부여
1- 항목 기반
원본 유사도
항목 계층 구조
테이블
원본 데이터
비식별 데이터

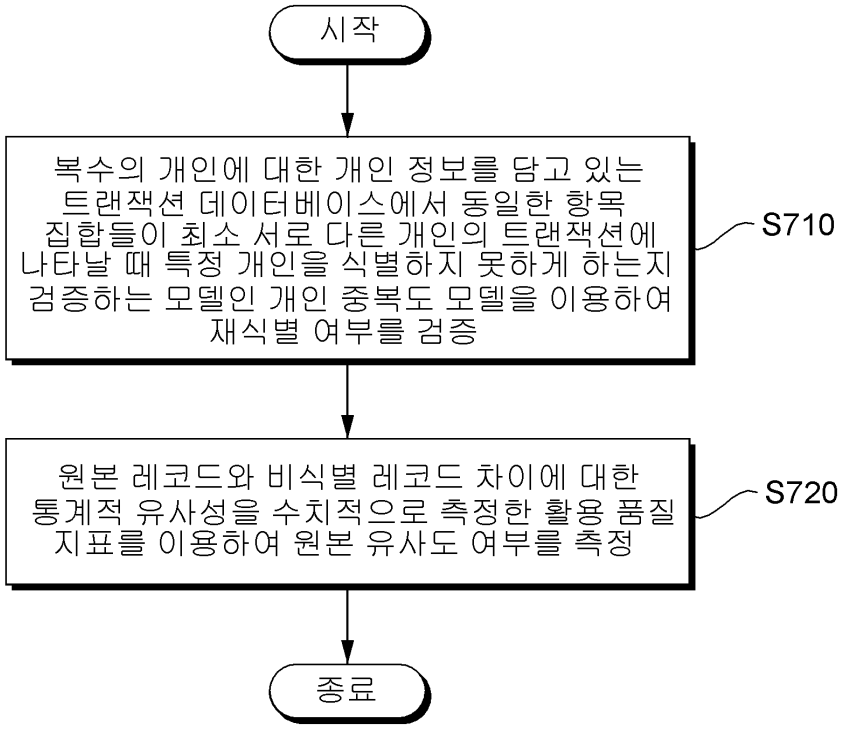
도면5



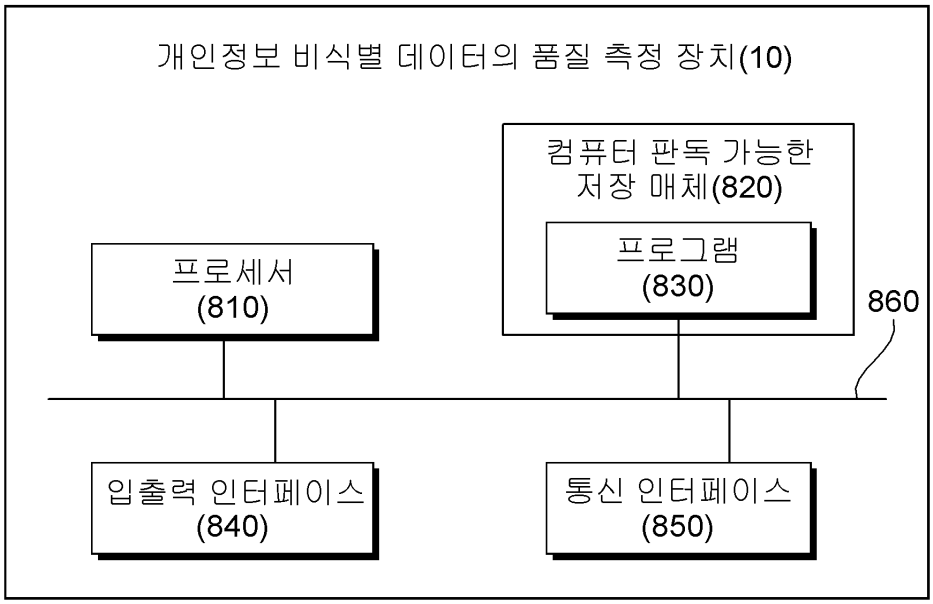
도면6


원본 유사도
1-항목 원본 유사도 계산
$\sum \frac{1}{n} * (1\text{-항목 원본 유사도})$
$a_n \in T_k$
a_1
a_2
...
a_n
$\frac{1}{n}$
$\frac{1}{n}$
$\frac{1}{n}$
항목 집합의 크기로 동일한 가중치 부여
T_k
a_1
a_2
...
a_n
n item set

도면7


시작
복수의 개인에 대한 개인 정보를 담고 있는 트랜잭션 데이터베이스에서 동일한 항목 집합들이 최소 서로 다른 개인의 트랜잭션에 나타날 때 특정 개인을 식별하지 못하게 하는지 검증하는 모델인 개인 중복도 모델을 이용하여 재식별 여부를 검증
S710
원본 레코드와 비식별 레코드 차이에 대한 통계적 유사성을 수치적으로 측정한 활용 품질 지표를 이용하여 원본 유사도 여부를 측정
S720
종료

도면8


개인정보 비식별 데이터의 품질 측정 장치(10)
컴퓨터 판독 가능한 저장 매체(820)
프로그램 (830)
프로세서 (810)
860
입출력 인터페이스 (840)
통신 인터페이스 (850)