(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년09월14일
(11) 등록번호 10-2442737
(24) 등록일자 2022년09월07일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) *G06F 21/60* (2013.01)
(52) CPC특허분류
G06F 21/6254 (2013.01)
G06F 21/602 (2013.01)
(21) 출원번호 10-2016-0136415
(22) 출원일자 2016년10월20일
심사청구일자 2021년10월19일
(65) 공개번호 10-2017-0052465
(43) 공개일자 2017년05월12일
(30) 우선권주장
14/931,774 2015년11월03일 미국(US)
(56) 선행기술조사문헌
US20020169793 A1
(뒷면에 계속)

(73) 특허권자
팔로 알토 리서치 센터 인코포레이티드
미국 캘리포니아주 94304 팔로 알토 코요테 힐 로드 3333
(72) 발명자
줄리엔 프러이디거
미합중국 94043 캘리포니아주 마운틴 뷰 락 스트리트 2309
알레잔드로 이. 브리토
미합중국 94040 캘리포니아주 마운틴 뷰 오르테가 애비뉴 163
(뒷면에 계속)
(74) 대리인
장훈

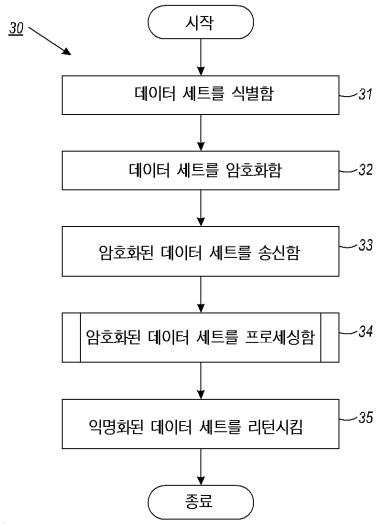
전체 청구항 수 : 총 20 항 심사관 : 구대성

(54) 발명의 명칭 **암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템 및 방법**

(57) 요 약

암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템 및 방법이 제공된다. 적어도 하나의 속성은 익명화를 위해 데이터세트 내에서 식별되고 복수의 데이터 값들과 연관된다. 각각의 데이터 값은 암호화된 데이터 값들의 순서를 유지하면서 각각의 식별된 속성에 대해 암호화된다. 암호화된 값들은 순서화되고 순서화된 암호화된 데이터 값들은 암호화된 데이터 값들의 순서에 기초하여 2개 이상의 클래스(class)들로 세그먼트된다. 세그먼트된 클래스들의 각각 클래스내 암호화된 데이터 값들의 범위가 식별되고 클래스들 중 하나의 클래스의 범위는 해당 클래스 내 각각의 암호화된 데이터 값에 익명화된 데이터로서 할당된다.

대 표 도 - 도2



(72) 발명자

샨타누 라네

미합중국 94025 캘리포니아주 멘로 파크 샤론 파크 드라이브 675 아파트먼트 201

에르신 우준

미합중국 95008 캘리포니아주 캠벨 카프리 드라이브 1186

(56) 선행기술조사문헌

US20050147246 A1

US20120121080 A1

EP02752786 A1

EP02228735 A1

명 세 서

청구범위

청구항 1

암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템에 있어서:

익명화를 위해 데이터세트 내 적어도 하나의 속성을 식별하기 위한 식별 모듈로서, 각각의 속성은 복수의 데이터 값들과 연관되는, 상기 식별 모듈;

상기 암호화된 데이터 값의 순서를 유지하면서 각각의 식별된 속성에 대한 각각의 데이터 값을 암호화하기 위한 암호화 모듈;

상기 암호화된 값들을 순서화하기 위한 순서 모듈;

상기 암호화된 데이터 값들의 순서에 기초하여 상기 순서화된 암호화된 데이터 값들을 2개 이상의 클래스(class)들로 세그먼트하는 세그멘테이션 모듈(segmentation module);

상기 세그먼트된 클래스들의 각각의 클래스내의 상기 암호화된 데이터 값들의 범위를 결정하기 위한 결정 모듈; 및

상기 클래스들 중 하나의 범위를 해당 클래스 내의 각각의 암호화된 데이터 값에 익명화된 데이터로서 할당하는 할당 모듈을 포함하고,

상기 모듈들은 프로세서를 통해 실행되는, 컴퓨터 구현 시스템.

청구항 2

제 1 항에 있어서,

신뢰되지 않은 제 3 자에 상기 익명화된 데이터 값들을 제공하기 위한 전달 모듈을 더 포함하는, 컴퓨터 구현 시스템.

청구항 3

제 1 항에 있어서,

익명화를 위해 데이터 세트 내에서 추가 속성을 식별하고 상기 추가 속성과 연관된 데이터 항목(item)들의 각각에 마스킹(masking)된 값을 랜덤으로 할당하기 위한 마스킹 모듈(masking module)을 더 포함하는, 컴퓨터 구현 시스템.

청구항 4

제 1 항에 있어서,

상기 적어도 하나의 속성으로 익명화를 위해 상기 데이터 세트 내에서 추가 속성을 식별하기 위한 결합 세그멘테이션 모듈을 더 포함하고, 상기 적어도 하나의 속성에 대한 각각의 데이터 값은 상기 추가 속성에 대한 추가 데이터 값에 대응하고, 상기 추가 데이터 값들을 암호화하고, 상기 세그먼트된 클래스들의 각각에 대한 상기 추가 암호화된 데이터 값들을 추가 세그먼트된 클래스들로 세그먼트하며, 상기 적어도 하나의 속성 및 상기 추가 속성에 대한 클래스들의 그룹들을 생성하는, 컴퓨터 구현 시스템.

청구항 5

제 1 항에 있어서,

n개의 세그먼트된 클래스들을 식별하는 클래스 식별자를 더 포함하는, 컴퓨터 구현 시스템.

청구항 6

제 5 항에 있어서,

상기 세그멘테이션 모듈은 상기 순서화된 암호화된 데이터 값들을 상기 n개의 세그먼트된 클래스들로 랜덤하게 세그먼트하는, 컴퓨터 구현 시스템.

청구항 7

제 5 항에 있어서,

상기 세그멘테이션 모듈은 클래스 당 최소 k개의 암호화된 데이터 값들을 선택하고, 상기 식별된 n개의 세그먼트된 클래스들이 도달하고 상기 세그먼트된 클래스들의 각각이 적어도 k개의 암호화된 데이터 값들을 포함할 때까지 상기 순서화된 암호화된 데이터 값들을 세그먼트하는, 컴퓨터 구현 시스템.

청구항 8

제 1 항에 있어서,

상기 세그멘테이션 모듈은 세그먼테이션을 위한 반복 횟수를 지정하고 상기 순서화된 암호화된 데이터 값들을 지정된 반복 횟수가 수행될 때까지 세그먼트하는, 컴퓨터 구현 시스템.

청구항 9

제 1 항에 있어서,

상기 세그멘테이션 모듈은 트리를 통해 상기 순서화된 암호화된 데이터 값들의 세그먼테이션을 수행하는, 컴퓨터 구현 시스템.

청구항 10

제 1 항에 있어서,

상기 익명화된 값들 및 다른 속성들의 데이터 값들을 갖는 데이터 세트 및 상기 익명화된 값들 및 상기 다른 속성들의 암호화된 데이터 값들을 갖는 데이터 세트 중 적어도 하나를 상기 제 3 자에게 제공하는 데이터 편집 모듈을 더 포함하는, 컴퓨터 구현 시스템.

청구항 11

암호화된 데이터를 익명화하기 위한 컴퓨터 구현 방법에 있어서:

신뢰된 서버에 의해 익명화를 위해 데이터세트 내 적어도 하나의 속성을 식별하는 단계로서, 각각의 속성은 복수의 데이터 값들과 연관되고, 상기 신뢰된 서버는 중앙 처리 장치, 메모리, 입력 포트, 및 출력 포트를 포함하는, 식별하는 단계;

상기 신뢰된 서버에 의해 상기 암호화된 데이터 값의 순서를 유지하면서 각각의 식별된 속성에 대한 각각의 데이터 값을 암호화하는 단계;

익명화기를 통해 상기 암호화된 값들을 순서화하는 단계;

상기 익명화기에 의해 상기 암호화된 데이터 값들의 순서에 기초하여 상기 순서화된 암호화된 데이터 값들을 2개 이상의 클래스(class)들로 세그먼트하는 단계;

상기 익명화기에 의해 상기 세그먼트된 클래스들의 각각의 클래스내의 상기 암호화된 데이터 값들의 범위를 식별하는 단계; 및

상기 익명화기에 의해 상기 클래스들 중 하나의 범위를 해당 클래스 내의 각각의 암호화된 데이터 값에 익명화된 데이터로서 할당하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 12

제 11 항에 있어서,

신뢰되지 않은 제 3 자에 상기 익명화된 데이터 값들을 제공하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 13

제 11 항에 있어서,

익명화를 위해 데이터 세트 내에서 추가 속성을 식별하는 단계; 및

상기 추가 속성과 연관된 데이터 항목들의 각각에 마스킹된 값을 랜덤으로 할당하기 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 14

제 11 항에 있어서,

상기 적어도 하나의 속성으로 익명화를 위해 상기 데이터 세트 내에서 추가 속성을 식별하는 단계로서, 상기 적어도 하나의 속성에 대한 각각의 데이터 값은 상기 추가 속성에 대한 추가 데이터 값에 대응하는, 추가 속성을 식별하는 단계;

상기 추가 데이터 값들을 암호화하는 단계;

상기 세그먼트된 클래스들의 각각에 대한 상기 추가 암호화된 데이터 값들을 추가 세그먼트된 클래스들로 세그먼트하는 단계; 및

상기 적어도 하나의 속성 및 상기 추가 속성에 대한 클래스들의 그룹들을 생성하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 15

제 11 항에 있어서,

n개의 세그먼트된 클래스들을 식별하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 16

제 15 항에 있어서,

상기 순서화된 암호화된 데이터 값들을 상기 n개의 세그먼트된 클래스들로 랜덤하게 세그먼트하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 17

제 15 항에 있어서,

클래스 당 최소 k개의 암호화된 데이터 값을 선택하는 단계; 및

상기 식별된 n개의 세그먼트된 클래스들이 도달하고 상기 세그먼트된 클래스들의 각각이 적어도 k개의 암호화된 데이터 값들을 포함할 때까지 상기 순서화된 암호화된 데이터 값들을 세그먼트하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 18

제 11 항에 있어서,

세그먼테이션을 위한 반복 횟수를 지정하는 단계; 및

상기 순서화된 암호화된 데이터 값들을 지정된 반복 횟수가 수행될 때까지 세그먼트하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 19

제 11 항에 있어서,

트리를 통해 상기 순서화된 암호화된 데이터 값들의 세그먼테이션을 수행하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 20

제 11 항에 있어서,

상기 데이터 세트 내의 다른 속성들의 데이터 값들을 암호화하는 단계 및 상기 익명화된 값들 및 암호화된 값들을 갖는 데이터 세트를 상기 제 3 자에게 제공하는 단계; 및

상기 익명화된 값들 및 상기 다른 속성들의 데이터 값들을 갖는 데이터 세트를 상기 제 3 자에게 제공하는 단계 중 하나 이상의 단계를 수행하는 단계를 더 포함하는, 컴퓨터 구현 방법.

발명의 설명

기 술 분 야

[0001] 본 출원은 일반적으로, 중요한(sensitive) 데이터를 보호하는 것에 관한 것이고 특히, 암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 비즈니스에 대한 일반 과정 동안, 회사들은 많은 양의 데이터를 축적한다. 최근에, 일부 회사들은 그들의 데이터를 광고주들, 연구원들, 또는 협업 파트너들과 같은, 제 3 자들과 공유함으로써 이 데이터를 현금화하기 시작했다. 제 3 자들은 통화 수수료를 지불하고 그 대신, 데이터 소유자로부터 관련 데이터를 수신한다. 제 3 자는 그 다음, 광고를 목적으로 하거나 연구를 행하기 위해 데이터를 사용할 수 있다. 그러나, 제 3 자들에 의해 요청된 데이터는 종종, 데이터가 수집되는 하나 이상의 개인들에 개인적인 정보를 포함한다.

[0003] 하나의 예에서, 병원은 환자 식별, 나이, 거주지 주소, 사회 보장 번호, 및 의학적 진단들을 포함하는 환자 기록들을 유지한다. 당뇨병에 관한 연구를 행하는 제 3 자는 나이가 40세 미만의 환자들에 대해 최대 그리고 최소 수의 유형 II 당뇨병 진단들을 갖는 미국의 지역들을 식별하기를 원한다. 요청된 데이터를 전송하기 이전에, 데이터 소유자는, 제공될 데이터가 신뢰되지 않은 제 3 자가 개인의 개인 정보에 액세스하거나 개인의 계정(identity)를 결정하는 것을 허용하지 않음을 보장해야 한다.

[0004] 데이터 익명화는 요청하는 제 3 자가 데이터를 사용하는 것을 허용하는 피쳐(feature)들을 유지하면서, 중요한 정보를 보호하기 위해 데이터의 변경을 포함한다. 데이터 변경은 잡음을 부가하거나, 데이터의 정밀도를 감소시키거나, 데이터의 부분들 그 자체를 삭제하는 것을 포함할 수 있다. 일반적으로, 데이터 소유자들은 익명화에 관한 충분한 지식을 갖지 않고, 따라서 제 3 자에 데이터를 제공하기 이전에 그들의 데이터를 익명화하기 위해 제 3 자들에 의존한다. 하나의 접근법은 데이터 익명화에 도움을 주기 위해 개인 인력을 제공하는 익명화 서비스 제공자에 접촉하는 것을 포함한다. 익명화에 할당된 인력은 신뢰되지 않은 제 3 자임에도 불구하고 데이터에 액세스한다. 현재, 많은 회사들은 익명화 서비스에, 데이터가 익명화되기 이전 및 익명화된 후에 데이터를 보호하기 위해 양해각서(Memorandum of Understanding) 또는 기밀 유지 협약서(Non-Disclosure Agreement)와 같은, 비밀보장 협약서들에 서명할 것을 요청한다.

[0005] 데이터 익명화를 수행하기 위한 종래의 방법들이 존재하지만, 신뢰되지 않은 제 3 자에 의한 익명화의 문제를 해결하는데 실패한다. 스위니(Sweeney)에 대한 미국 특허 번호 제 7,269,578 호에서, 표의 엔트리(entry)들은 특정 필드들 및 기록들, 수신자 프로파일, 및 최소 익명성 레벨과 같은, 사용자 명세들에 기초하여 변경된다. k에 대한 값이 컴퓨팅(computing)되고 속성들의 그룹에 걸쳐 할당된 동일한 값들을 갖는 k개의 튜플(tuple)들인 유사(quasi) 식별자들이 배포(release)를 위해 식별된다. 각각의 속성의 감도(sensitivity)가 결정되고 단방향 해싱을 포함하는, 동치류 치환(equivalence class substitution), 또는 일반화된 대체와 같은, 대체 전략이 각각의 중요한 속성에 대해 결정된다. 일반화된 대체는 최고 수의 별개의 값들을 갖는 속성을 식별하고 그 값을 위해 제공된 정보의 양을 감소시킴으로써 그 속성에 대한 각각의 값을 일반화하는 것을 포함한다. 예를 들면, 월, 일 및 년을 가지는 날짜들은 월 및 년, 단지 년, 또는 년들의 범위로 일반화될 수 있다. 그러나, 스위니는 익명화가 신뢰되지 않은 당사자에 의해 수행될 수 있고 따라서, 익명화될 데이터의 어떠한 보호도 제공하지 않음을 고려하는데 실패한다. 게다가, 스위니는 익명화될 데이터 세트가 나누어져야 하는 복수의 클래스(class)들을 식별하고 그 데이터 값이 속하는 클래스에 기초하여 각각의 데이터 값을 익명화하는데 실패한다.

[0006] 게다가, 르페브르(LeFevre), 등에 의한 명칭이 "몬드리안 다차원 K-익명성(Mondrian Multidimensional K-Anonymity)"인 논문은 각각의 지역이 k개 이상의 지점들을 포함하도록, 단일 차원 또는 다차원 파티션

(partition)들을 사용하여 데이터 세트를 분할하는 것을 설명한다. 하나의 예에서, 분할은 중앙값 분할을 사용하여 발생할 수 있다. 그러나, 르페브르 논문은 신뢰되지 않은 제 3 자가 데이터를 익명화하는 경우에, 익명화 이전에 데이터를 보호하기 위한 단계들을 설명하는데 실패한다. 게다가, 르페브르는 익명화를 위해 속성들을 자동으로 식별하기 위해 데이터 감도의 측정치들을 제공하는데 실패하며 또한, 마스킹(masking)하기를 고려하는데 실패한다.

[0007] 따라서, 데이터가 수집되는 개인들의 프라이버시를 손상시키지 않고 제 3 자 익명화를 위해 중요한 데이터를 이용가능하게 하기 위한 접근법에 대한 필요성이 존재한다.

발명의 내용

해결하려는 과제

[0008] 회사들이 그들의 데이터를 상업화하는 것을 허용하기 위해, 데이터의 소유자는, 데이터 세트 내의 중요한 정보가 보호됨을 보장해야 한다. 데이터를 보호하기 위해, 데이터 세트 내의 각각의 속성은 그 속성의 감도를 결정하도록 분석된다. 속성의 감도에 기초하여, 예를 들면 일반화 또는 마스킹에 의해, 데이터의 익명화는 데이터를 애매하게 만들기 위해 사용될 수 있다. 익명화가 선택될 때, 속성에 대한 데이터 값들은 암호화되고 신뢰되지 않은 데이터 익명화기(anonymizer)로 송신된다. 익명화기는 속성에 대한 암호화된 데이터 값들을 클래스들로 분할하고 각각의 클래스 내의 암호화된 값들의 범위들을 식별함으로써 데이터를 익명화한다. 마지막으로, 각각의 클래스의 범위는 그 클래스에 속하는 각각의 암호화된 데이터 값에 할당된다.

과제의 해결 수단

[0009] 일 실시예는 암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템 및 방법을 제공한다. 적어도 하나의 속성은 익명화를 위해 데이터세트 내에서 식별되고 복수의 데이터 값들과 연관된다. 각각의 데이터 값은 암호화된 데이터 값들의 순서를 유지하면서 각각의 식별된 속성에 대해 암호화된다. 암호화된 값들은 순서화되고 순서화된 암호화된 데이터 값들은 암호화된 데이터 값들의 순서화에 기초하여 2개 이상의 클래스들로 분할된다. 분할된 클래스들의 각각 내에서 암호화된 데이터 값들의 범위가 식별되고 클래스들 중 하나의 범위는 그 클래스 내의 각각의 암호화된 데이터 값에 익명화된 데이터로서 할당된다.

발명의 효과

[0010] 여전히, 본 발명의 다른 실시예들은 다음의 상세한 설명으로부터 당업자들에게 용이하게 명백해질 것이고, 본 발명의 실시예들은 본 발명을 실행하기 위해 고려된 최상의 모드를 도시함으로써 설명된다. 실현될 바와 같이, 본 발명은 다른 그리고 상이한 실시예들을 가능하게 하고 그것의 몇몇 상세들은 모두 본 발명의 사상 및 범위를 벗어나지 않고, 다양한 명백한 측면들에서 수정할 수 있다. 그에 따라, 도면들 및 상세한 설명은 제한적인 것으로서가 아닌 사실상 예시적인 것으로서 간주될 것이다.

도면의 간단한 설명

[0011] 도 1은 하나의 실시예에 따른, 암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템을 도시하는 블록도.

도 2는 하나의 실시예에 따른, 암호화된 데이터를 익명화하기 위한 컴퓨터 구현 방법을 도시하는 흐름도.

도 3은 순차적(order preserving) 암호화를 위한 프로세스를 예로서 도시하는 블록도.

도 4는 일반화에 의해 데이터를 익명화하기 위한 프로세스를 예로서 도시하는 흐름도.

도 5는 일반화에 의한 익명화를 위한 데이터 값들의 표를 예로서 도시하는 블록도.

도 6은 암호화된 데이터 값들을 분할하기 위한 프로세스를 예로서 도시하는 흐름도.

도 7은 일반화를 통해 단일 속성에 대한 데이터 값들을 익명화하기 위한 프로세스를 예로서 도시하는 흐름도.

도 8은 단일 속성의 암호화된 데이터 값들에 대한 트리를 예로서 도시하는 블록도.

도 9는 일반화를 통해 2개 이상의 속성들의 데이터를 결합적으로 익명화하기 위한 프로세스를 예로서 도시하는 흐름도.

도 10은 제 2 속성에 대한 데이터 값들의 분할을 예로서 도시하는 블록도.

도 11은 속성들 중 2개에 대한 일반화된 익명화 값들을 갖는 도 5의 데이터 세트를 예로서 도시하는 블록도.

도 12는 마스킹을 통해 데이터 값들을 익명화하기 위한 프로세스를 예로서 도시하는 흐름도.

도 13은 마스킹된 익명화 데이터 값들을 갖는 도 11의 데이터 세트를 예로서 도시하는 블록도.

발명을 실시하기 위한 구체적인 내용

[0012] 회사들의 데이터를 현금화하는데 관심 있는 회사들의 증가는 데이터 보호가 극단적으로 중요함을 보장한다. 일반적으로, 회사들은 데이터 프라이버시에 관한 충분한 지식을 갖지 않고 요청하는 제 3 자에 데이터를 제공하기 전에 데이터를 익명화하기 위해 외부 비즈니스들을 유지한다. 그러나, 외부의 익명화 회사들은 종종 신뢰되지 않고 기밀 유지 협약서들과 같은, 비밀보장 협약서들은 이전 데이터를 보호하려는 시도로 활용된다. 데이터 보호를 보장하기 위해, 비밀보장 협약서들 대신에, 익명화의 적절한 레벨이 결정되고 요구되면, 익명화는 맹목적으로(blindly) 수행될 수 있다. 맹목적 익명화 동안, 데이터는 제 3 자 익명화 회사로 데이터를 송신하기 전에 암호화된다. 암호화된 데이터는 그 다음, 클래스들로 분할되고 각각의 클래스의 암호화된 범위들이 식별된다. 범위들은 그 클래스 내의 각각의 암호화된 데이터 항목(item)에 일반화된 익명화된 값으로서 할당된다.

[0013] 맹목적 익명화는 데이터 보안을 보장하고 회사들이 데이터가 수집되는 개인들의 프라이버시를 손상시키지 않고 외부 요청자들에 그들의 데이터를 제공하는 것을 허용한다. 도 1은 하나의 실시예에 따른, 암호화된 데이터를 익명화하기 위한 컴퓨터 구현 시스템을 도시하는 블록도이다. 데이터 소유자는 시간에 걸쳐 수집된 많은 양의 데이터(22)를 유지한다. 데이터(22)는 비즈니스의 데이터 소유자의 장소에 또는 원격으로 위치된 신뢰된 서버(17)에 상호접속된 데이터베이스(21)에 저장될 수 있다. 대안적으로, 데이터는 다수의 서버들 상의 풀(pool)들을 포함하는, 클라우드(cloud)에 저장될 수 있다. 데이터 소유자는 데스크탑(11a) 또는 랩탑(11b) 컴퓨터를 통해, 또는 모바일 컴퓨팅 디바이스(도시되지 않음)와 같은, 다른 유형들의 컴퓨팅 디바이스들을 통해 데이터(22)에 액세스할 수 있다. 저장된 데이터(22)는 하나 이상의 데이터 세트들(22)을 포함할 수 있고, 각각의 데이터 세트는 하나 이상의 개인들에 대한 복수의 속성들과 연관된다. 게다가, 각각의 속성은 개인들의 각각에 대한 데이터 값들과 연관된다.

[0014] 신뢰된 서버(17)는 감도 모듈(18), 암호화기(19) 및 송신기(20)를 포함한다. 감도 모듈(18)은 잠재적으로 중요하고 익명화를 요구할 수 있는 요청된 데이터 세트 내에서 그들 속성들을 식별한다. 일단 식별되면, 암호화기(19)는 잠재적으로 중요한 속성에 대한 각각의 데이터 값을 암호화할 수 있다. 후속적으로, 송신기(20)는 익명화를 위해 신뢰되지 않은 제 3 자 익명화기(12)로 모든 암호화된 데이터 값들을 송신한다. 익명화기(12)는 감도 할당 모듈(13), 순서 모듈(14), 클래스 생성기(15), 및 라벨 할당(16)을 포함한다. 감도 할당 모듈(13)은 신뢰된 서버(17)로부터 수신된 암호화된 데이터 값들에 대한 각각의 속성을 분석하고 각각의 속성에 감도 값을 할당한다. 감도 값은, 데이터가 사회 보장 번호, 신용 카드 번호, 또는 주소와 같은, 개인에 대한 개인 정보를 포함하는지의 여부와 같은, 데이터 감도의 측정치를 제공한다. 다른 유형들의 개인 정보가 가능하다. 속성에 할당된 감도 값에 의존하여, 그 속성에 대한 데이터 값들의 익명화는 제 3 자가 개인 데이터에 액세스하는 것을 방지하는데 필요할 수 있다. 익명화는 데이터의 일반화 또는 마스킹을 포함할 수 있다. 감도 값에 기초하여 속성에 대한 적절한 유형의 익명화를 결정하는 것은 도 2에 관하여 하기에 또한 설명된다.

[0015] 일반화된 익명화에 대해, 순서 모듈(14)은 낮은 것으로부터 높은 것으로 또는 높은 것으로부터 낮은 것으로 속성에 대한 암호화된 데이터 값들을 순서화한다. 그 다음, 클래스 생성기(15)는 순서화된 데이터를 2개 이상의 클래스들로 분할하고, 라벨 할당 모듈(16)은 각각의 클래스에 대해 암호화된 데이터 값들의 범위를 결정하며 그 클래스 내의 각각의 암호화된 데이터 값에 범위를 익명화된 값으로서 할당한다. 익명화된 데이터는 그 다음, 신뢰된 서버(17)로 송신되고, 데이터베이스(21)에 저장되거나 요청하는 제 3 자에 제공된다. 대안적으로, 속성의 데이터 값들은 각각의 데이터 값을 마스킹함으로써 익명화될 수 있다. 마스킹 모듈(24)은 예를 들면, 해싱을 사용하여 중요한 속성들을 의사 랜덤(pseudo-random) 값들로 대체한다.

[0016] 컴퓨팅 디바이스들 및 서버들은 각각, 중앙 처리 장치, 랜덤 액세스 메모리(RAM), 하드 드라이브 또는 CD ROM 드라이브와 같은, 비 휘발성 부 저장장치, 네트워크 인터페이스들, 및 키보드 및 디스플레이와 같은 사용자 인터페이싱 수단을 포함하는, 주변 디바이스들을 포함할 수 있다. 소프트웨어 프로그램들을 포함하는, 프로그램 코드, 및 데이터는 CPU에 의해 실행 및 프로세싱하기 위해 RAM에 로딩되고 결과들은 디스플레이, 출력, 송신, 또는 저장하기 위해 생성된다.

[0017] 모듈들은 컴퓨터 프로그램 또는 종래의 프로그래밍 언어로 소스 코드로서 기록된 절차로서 구현될 수 있고 중앙 처리 장치에 의해 객체(object) 또는 바이트 코드로서 실행하기 위해 제공된다. 대안적으로, 모듈들은 또한, 집

적 회로로서 또는 판독 전용 메모리 구성요소들로 버닝(burning)되는 것으로서, 하드웨어로 구현될 수 있고, 컴퓨터 디바이스들의 각각 및 서버는 특수화된 컴퓨터의 역할을 할 수 있다. 예를 들면, 모듈들이 하드웨어로서 구현될 때, 그 특정한 하드웨어는 메시지 우선순위화를 수행하기 위해 특수화되고 다른 컴퓨터들은 사용될 수 없다. 부가적으로, 모듈들이 판독 전용 메모리 구성요소들로 버닝될 때, 판독 전용 메모리를 저장하는 컴퓨팅 디바이스 또는 서버는, 다른 컴퓨터들이 할 수 없는 메시지 우선순위화를 수행하기 위해 특수화된다. 다른 유형들의 특수화된 컴퓨터들이 가능하다. 게다가, 관리 시스템은, 관리 시스템이 구현되는 특정 클라이언트들 및 특정 하드웨어로 제한될 뿐만 아니라, 단지 가입하는 클라이언트들에 대한 가입 서비스에 의해 제한될 수 있다. 소스 코드 및 객체 및 바이트 코드들의 다양한 구현들은 플로피 디스크, 하드 드라이브, 디지털 비디오 디스크(DVD), 랜덤 액세스 메모리(RAM), 판독 전용 메모리(ROM) 및 유사 저장 매체들과 같은, 컴퓨터 판독가능한 저장 매체 상에서 보유될 수 있다. 다른 유형들의 모듈들 및 모듈 기능들 뿐만 아니라, 다른 물리적 하드웨어 구성요소들이 가능하다.

[0018] 마스킹 데이터는 이전 연구들이 익명화해제 마스킹된 데이터베이스들의 확률에 대해 보여준바와 같이, 일반화보다 낮은 프라이버시 보호를 제공한다. 마스킹은 따라서, 홀로 사용되어서는 안되고, 일반화와 조합되어야 한다. 데이터 일반화는 연구 또는 광고를 위해 일반화된 데이터를 활용하기 위해 제 3 자들에 대한 데이터의 유용성을 동시에 유지하면서, 데이터의 일부 보호를 제공한다. 데이터 일반화가 수행될 때, 속성에 대한 데이터 값들은 데이터 보호의 부가적인 계층을 부가하기 위한 익명화 이전에 암호화된다. 도 2는 하나의 실시예에 따른, 암호화된 데이터를 익명화하기 위한 컴퓨터 구현 방법을 도시하는 흐름도이다. 데이터 소유자는 특정한 유형의 데이터를 위한 요청을 수신하고 요청된 데이터를 갖는 데이터 세트가 식별된다(블록(31)). 데이터 세트는 개인들의 목록 및 개인들의 각각과 연관된 속성들에 대한 데이터 값들을 포함할 수 있다. 데이터 소유자는 속성들 중 하나 이상이 중요하거나, 개인 데이터를 잠재적으로 포함하도록 결정되는지의 여부 및 그렇다면, 데이터가 얼마나 중요한지를 결정하기 위해 데이터 세트를 분석한다. 중요한 데이터는 특정 개인을 식별할 수 있는 정보를 포함하고, 이는 다른 사람들에게 폭로되면, 보안의 손실을 야기할 수 있다. 후속적으로, 잠재적으로 중요한 속성들과 연관된 데이터 값들이 암호화된다(블록(32)). 또 다른 실시예에서, 데이터 세트에서의 모든 데이터 값들은 익명화기에 제공하기 위해 암호화되고, 상기 익명화기는 단독으로 어떤 속성들이 중요한지를 결정할 수 있다.

[0019] 하나의 실시예에서, 순차적 암호화는 데이터 값들이 암호화되지 않은 데이터 값들의 순서를 유지하면서 각각의 데이터 값에 암호화된 값들을 할당하는 의사 랜덤 생성기를 통해 실행된다. 진화된 암호화 표준과 같은, 대부분의 암호화 알고리즘들과 달리, 순차적 암호화는 그들의 암호화된 형태로 평문 데이터의 순서화를 유지한다. 도 3은 순차적 암호화를 위한 프로세스를 예로서 도시하는 블록도이다. 평문 버전의 데이터 값들(41)은 숫자들(0 내지 5)을 포함한다. 평문 데이터 값들(41)의 순서는 순서화된 암호문 데이터 값들(42)을 생성하기 위해 순차적 암호화 시에 유지된다. 구체적으로, 의사 랜덤 생성기는 암호문 값들로서 순차적 암호화 값들에 대한 평문 값들의 의사 랜덤 매핑들을 수행한다. 순차적 암호화를 위한 유일한 기준은, 평문 값들 및 암호문 값들의 순서들이 교차할 수 없다는 것이다. 하나의 예에서, 숫자(0)는 457에 매핑되는 반면에, 1은 473에 매핑되고, 2는 510에 매핑되고, 3은 625에 매핑되고, 4는 635에 매핑되며, 5는 1001에 매핑된다. 제로(Zero)는 최소 평문 데이터 값이고 제로의 암호화된 값(457)은 데이터세트의 최소 암호문 값이다. 부가적으로, 평문 값(5)은 최대인 반면에, 5에 대한 암호화된 값(1001)은 또한 최대이다.

[0020] 데이터 값들이 컬러들과 같은 텍스트, 질병들 또는 다른 유형들의 정량가능하지 않은 값들을 표현할 때, 수치들은 암호화 이전에 데이터 값들의 각각에 할당될 수 있다. 예를 들면, 수치들은 상태의 심각도 또는 이환율(morbidity)에 기초하여 질병들에 할당되고, 후속적으로 암호화될 수 있다. 암호화된 값들의 순서를 보존하는 것은 데이터가, 암호화됨에도 불구하고 여전히 유용함을 보장하는데 도움을 준다. 다른 암호화 알고리즘들이 가능하지만; 최소에서, 암호화된 데이터 값들의 순서화는 평문 값들과 일관되게 보존되어야 한다.

[0021] 도 2로 리턴하면, 암호화된 데이터 세트는 그 다음, 데이터의 프로세싱(블록(34))을 위해 신뢰되지 않은 익명화기로 송신된다(블록(33)). 구체적으로, 익명화기는, 익명화가 감도 값들에 기초하여 필요한지의 여부를 결정할 수 있고 그렇다면, 변화의 양은 중요한 속성의 데이터를 보호하기 위해 필요했다. 구체적으로, 속성이 사용자들을 식별하기 위해 사용될 수 있으면, 속성은 중요한 것으로서 자격이 있고, 유사 식별자로서 공지된다. 속성의 감도를 결정하기 위해, 익명화기는 속성을 익명화하고 2015년 11월 3일에 출원되고, 명칭이 "익명화를 위해 속성들을 자동으로 식별하기 위한 컴퓨터 구현 시스템 및 방법(Computer-Implemented System and Method for Automatically Identifying Attributes for Anonymization)"이며, 대리인 사건 번호 제 022.1454.US.UTL 호인 미국 특허 출원 일련 번호 제 14/931,802 호에서 또한 설명된 바와 같이 감도 값을 할당한다. 하나의 실시예에서, 속성들은 감도 값을 결정하는 것을 돕기 위해 사용될 수 있는 가중치와 연관될 수 있다. 데이터 익명화의

유형은 데이터 일반화 또는 데이터 마스킹을 포함하는, 각각의 속성의 감도 값에 기초하여 결정될 수 있다. 하나의 예에서, 각각의 감도 값은 그 범위가 0부터 1까지일 수 있고, 제로는 중요하지 않은 데이터를 표현하고 1은 극단적으로 중요한 데이터를 표현한다. 속성에 대한 감도의 레벨이 매우 높을 때, 데이터 값들의 마스킹은 유일한 이용가능한 옵션일 수 있다. 그러나, 감도 값들이 낮거나 중간일 때, 속성은 유사 식별자로서 지정될 수 있고, 이는 속성이 저절로 연관된 개인을 식별할 수 없지만, 하나 이상의 유사 식별자들과 같은, 다른 정보와 협력하여 개인이 식별될 수 있음을 나타낸다. 예를 들면, 나이는 일반적으로, 사람을 식별하지 않지만; 함께 취해진 나이, 성 및 주소는 가능하게 연관된 사람을 식별할 수 있다. 유사 식별자들에 대해, 일반화와 같은, 익명화 기술들은 데이터를 보호하고 폭로를 방지하기 위해 적절할 수 있다.

[0022] 데이터 프로세싱 동안, 일반화되거나 마스킹된 값들은 도 4 및 도 12에 관하여 하기에 또한 설명된 바와 같이, 중요한 속성과 연관된 각각의 데이터 값을 위해 결정된다. 후속적으로, 익명화된 데이터 값들은 제 3 자에 선택적으로 제공하기 위해 데이터 소유자로 송신될 수 있다(블록(35)). 일단 수신되면, 데이터 소유자는 데이터 세트에서의 평문 데이터 값들을 익명화기로부터의 익명화되거나 마스킹된 값들로 대체할 수 있다. 또 다른 실시예에서, 일반화된 값들에 대해, 데이터 소유자는 일반화된 범위의 암호화된 데이터 값들을 복호화하고 암호문보다 평문으로, 각각의 데이터 값에 할당된 일반화된 범위들을 제공할 수 있다. 예를 들면, 익명화기는 데이터 소유자로 일반화된 범위들의 암호화된 데이터 값들을 송신한다. 데이터 소유자는 그 다음, 요청하는 제 3 자에 제공되는 평문 값들의 범위를 얻기 위해 익명화된 범위들을 복호화할 수 있다. 이 실시예에서, 범위는 데이터 값의 일반화를 제공하는 반면에, 평문은 제 3 자가, 일반화된 범위들이 암호문으로 제공될 때보다 데이터에 더 액세스하는 것을 허용한다.

[0023] 일반화 동안, 데이터 값들은 그룹핑(grouping)되고 각각의 그룹에 대한 값들의 범위는 그 그룹 내의 데이터 값들에 익명화된 값들로서 할당된다. 도 4는 일반화에 의해 데이터를 익명화하기 위한 프로세스를 예로서 도시하는 흐름도이다. 익명화는 하나 이상의 중요한 속성들에 대한 암호화된 데이터 값들을 수신하고 단일 속성에 대한 암호화된 데이터 값들을 순서화한다(블록(51)). 익명화기는 그 다음, 순서화된 데이터를 2개 이상의 클래스들로 분할한다(블록(52)). 분할은 도 6에 관하여 하기에 또한 설명된 바와 같이, 암호화된 데이터 값들의 중간값을 사용하여, 또는 미리 결정된 n-값에 기초하여 랜덤으로 발생할 수 있다. 대안적으로, 데이터 소유자는 익명화기에 암호문 범위들을 제공할 수 있고, 익명화기는 그 다음, 암호화된 데이터 값들을 범위에 기초하여 그룹들에 둔다. 일단 암호화된 데이터 값들이 클래스들로 분할되거나 넣어졌으면, 암호화된 데이터 값들의 범위는 각각의 클래스에 대해 결정되고(블록(53)) 결정된 범위들은 그 범위에 대응하는 클래스에 속하는 각각의 암호화된 데이터 값에 익명화된 값들로서 할당된다(블록(54)).

[0024] 하나의 예에서, 광고 연구 회사는 고객 나이, 집 코드, 및 구매 금액을 포함하는, 고객 비용에 관한 데이터를 위해 백화점에 접촉한다. 백화점은 요청된 데이터가 유지되는 데이터 세트를 식별하고 이에 액세스한다. 도 5는 익명화를 위한 데이터세트(60)를 예로서 도시하는 블록도이다. 데이터 세트(60)는 x축을 따라 열거된 속성들(61 내지 66) 및 각각의 열거된 속성 아래의 열들을 채우는 데이터 값들을 갖는 차트를 포함한다. 속성들은 이름(first name)(61), 성(last name)(62), 나이(63), 계좌 번호(64), 집 코드(65), 및 구매 금액(66)을 포함한다. 데이터 소유자는, 속성들 중 임의의 속성이 중요할 수 있거나 중요한지의 여부를 예비적으로 결정하기 위해 데이터세트를 검토할 수 있다. 이 예에서, 데이터 소유자는, 나이 속성이 다소 중요한 것으로 결정되고 익명화가 아마 필요하다고 결정한다. 나이 속성에 대한 데이터 값들은 각각, 순차적 암호화를 사용하여 암호화된다. 또 다른 예에서, 데이터 소유자는 데이터 세트에서 모든 속성들에 대한 데이터 값들을 암호화하고 익명화기에 어떤 속성들이 연관된 데이터 값들의 익명화를 요구할지를 결정하도록 요청한다.

[0025] 일단 암호화되면, 백화점은 익명화기로 암호화된 데이터를 전송할 수 있거나 대안적으로, 백화점은 암호화된 데이터 값들에 직접적으로 액세스하기 위해 익명화기에 가상 사설 네트워크 접속부를 제공할 수 있다. 나이 속성에 대한 암호화된 데이터 값들의 수신 시에, 익명화기는 순서화된 암호화된 데이터 값들을 클래스들로 분할하고 그 클래스 내의 각각의 암호화된 데이터 값에 클래스 라벨을 할당함으로써 암호화된 데이터 값들을 익명화한다.

[0026] 암호화된 데이터 값들을 분할하기 위한 다른 방법들이 가능하다. 도 6은 암호화된 데이터 값들을 분할하기 위한 프로세스(70)를 예로서 도시하는 흐름도이다. 암호화된 데이터 값들에 적용될 분할 방법이 결정되고(블록(71)) 분할 중지 지점이 선택된다(블록(72)). 암호화된 데이터 값들은 선택된 중지 지점이 만족될 때까지(블록(74)), 선택된 분할 방법을 사용하여 클래스들로 나누어진다(블록(73)).

[0027] 상이한 분할 방법들은 랜덤 분할, n-분할, 및 중간값 분할을 포함할 수 있다. 그러나, 분할을 위한 다른 방법들이 가능하다. 랜덤 분할은 분할 중지 지점에 도달할 때까지 암호화된 데이터 값들을 랜덤으로 분할하는 것을 포

함한다. 분할들 및 클래스들은 균일하거나 균일하지 않을 수 있고, 이는 익명화기에 의해 결정되거나 데이터 소유자에 의해 요청될 수 있다. 대안적으로, 암호화된 데이터 값들은 미리 결정된 n-값 및 중지 지점을 포함하는, 하나 이상의 분할 파라미터들을 통해 분할될 수 있다. n-분할 값에 대해, n=2이면, 암호화된 데이터 값들은 각각의 분할, 또는 데이터 값들의 구분에서 절반으로 나누어진다. 게다가, n=3일 때, 암호화된 데이터 값들은 각각의 분할에서 3개의 그룹핑들로 나누어진다. n개의 분할들은 중지 지점에 만족될 때까지 발생한다. 데이터 소유자, 데이터 소유자와 연관된 개인, 또는 익명화기는 n에 대한 값 뿐만 아니라, 선택된 중지 지점을 결정할 수 있다. 마지막으로, 중간값 분할은 암호화된 데이터 값들을 중간값 암호화된 데이터 값(n=2에 등가임)에 기초하여 2개의 클래스들로 나누는 것을 포함한다. 암호화된 데이터 값들의 분할은 중지 지점이 만족될 때까지 데이터 값들을 중간값으로 나눔으로써 계속된다.

[0028] 하나의 실시예에서, 중지 지점은 모든 클래스들이 k개의 요소들을 가질 때이고, 여기서 k는 원하는 레벨의 프라이버시이다. 또 다른 실시예에서, 중지 지점은 클래스들의 미리 결정된 수이며, 이는 분할이 종료하기 위해 만족되어야 한다. 또 다른 실시예에서, 분할은 미리 결정된 수의 분할들이 완료되었을 때, 종료된다. 데이터 소유자 또는 익명화기는 미리 결정된 중지 지점을 결정할 수 있다. 익명화기는 예를 들면, 이전 데이터 설정들, 속성들, 및 난독화(obfuscation)의 레벨들에 기초하여, 중지 지점을 산출하기 위해 분야 전문성(domain expertise)을 사용한다. 형성될 클래스들의 수는 데이터 익명화의 영향력(strength)에 관련된다. 예를 들면, 더 적은 수의 클래스들이 생성될 때, 클래스들의 각각에 대한 익명화 값이 더 익명으로 되는데, 이는 각각의 클래스 내의 데이터 값들의 범위 및 수가 더 크기 때문이다. 대조적으로, 더 많은 클래스들이 생성될 때, 익명화 값들은 덜 익명이 된다. 게다가, 분할 중지 지점은 제 3 자에 의해 요청된 데이터의 특수성에 의존할 수 있다. 더 특정한 데이터가 요구될 때, 클래스들의 수는 덜 특정한 데이터가 요구될 때보다 높게 설정될 수 있다.

[0029] 분할은 순서화된 암호화된 데이터 값들의 목록에 관해, 또는 대안적으로, 순서화된 암호화된 데이터 값들의 트리를 통해 수행될 수 있다. 도 5의 데이터 세트에 관하여 상기 예로 리턴하면, 나이 속성이 익명화를 위해 선택된다. 도 7은 단일 속성에 대한 데이터 값들을 익명화하기 위한 프로세스를 예로서 도시하는 흐름도이다. 나이 속성(81)은 백화점의 고객들에 대한 개인 나이들을 데이터 값들로서 열거한다. 후속적으로, 데이터 값들은 개별적으로, 암호화된 데이터 값들(82)을 생성하기 위해 순차적 암호화를 사용하여 암호화된다. 암호화된 데이터 값들(82)은 그 다음, 예를 들면, 최저로부터 최고 값까지 또는 최고로부터 최저 값까지 순서화된다(83). 다음, 순서화된 암호화된 데이터 값들은 분할된다(84a 내지 84b).

[0030] 하나의 실시예에서, 순서화된 암호화된 데이터 값들은 분할이 발생하는 목록을 형성할 수 있다. 중지 지점들로서 또한 공지된, 결과로 발생하는 클래스들의 수를 표현하는 n-분할 값, 및 암호화된 데이터 값들이 나누어질 클래스 당 k개의 요소들 또는 데이터 값들을 포함하는 분할 파라미터들이 하나의 예에서, 분할을 행하기 위해 제공된다. n-분할 값은 n=3으로 설정되고, 데이터 값들의 수는 k=2로 설정된다. k-값은 데이터 값들의 최소 수이고, n개의 세그먼트들 또는 클래스들 중 임의의 것은 최소 k개의 데이터 값들 또는 그 이상을 포함할 수 있다.

[0031] 순서화된 목록에 대하여, 제 1 분할은 암호화된 데이터 값들을 2개의 클래스들로 나누어서, 암호화된 데이터 값들(0857, 1056, 2764, 및 4798)이 하나의 클래스에 있고 암호화된 데이터 값들(6321, 7744, 8791 및 9921)이 또 다른 클래스에 있도록 한다. n=3이기 때문에, 도 다른 분할이 3개의 클래스들을 생성하도록 요구된다. 또 다른 분할은 0857 내지 1056이 제 1 클래스를 형성하고, 2764 내지 4798이 제 2 클래스를 형성하며, 6321 내지 9921이 제 3 클래스를 형성하도록, 최저 순서화된 암호화된 값들을 갖는 클래스 내에서 행해진다. 또 다른 실시예에서, 최저 순서화된 값들보다는, 최고 순서화된 암호화된 값들이 나누어질 수 있다. 분할의 순서는 미리 결정되거나, 사용자에 의해 입력되거나, 자동으로 결정될 수 있다.

[0032] 또 다른 실시예에서, n 값은 설명된 바와 같이, 최종 세그먼트들 또는 클래스들의 수 보다는, 데이터 값들에 적용될 분할들의 수를 표현할 수 있다. 예를 들면, n이 분할들의 수를 3으로서 표현할 때, 제 1 분할은 0857, 1056, 2764, 및 4798이 하나의 클래스에 있고 6321, 7744, 8791 및 9921이 또 다른 클래스에 있는, 2개의 클래스들을 형성하기 위해 행해진다. 제 2 분할은 그 다음, 0857 및 1056이 하나의 클래스에 있고 2764 및 4798이 또 다른 클래스에 있도록, 최저 순서화된 값들로 구성될 수 있다. 마지막으로, 제 3 분할은 6321 및 7744가 하나의 클래스에 있고 8791 및 9921이 또 다른 클래스에 있도록, 구성될 수 있다. 따라서, 3개의 분할들이 발생한 후에, 4개의 클래스들이 생성된다.

[0033] 또 다른 실시예에서, 트리는 속성에 대한 암호화된 데이터 값들을 분할하기 위해 사용될 수 있다. 도 8은 단일 속성의 암호화된 데이터 값들에 대한 트리(90)를 예로서 도시하는 블록도이다. 순서화된 암호화된 데이터 값들

은 트리(90)의 위에서 노드(96)에 의해 표현된다. 순서화된 암호화된 데이터를 2개의 그룹들로 분할하는 제 1 구분(91)이 발생하고, 암호화된 데이터 값들(0857, 1056, 2764, 및 4798)은 하나의 그룹(97)에 있고 암호화된 데이터 값들(6321, 7744, 8791, 및 9921)이 또 다른 그룹(95)에 있다. 클래스들의 미리 결정된 수가 3이기 때문에 또 다른 구분(92)이 요구된다. 따라서, 더 낮은 값들을 갖는 그룹은 암호화된 데이터 값들의 총 3개의 클래스들(93 내지 95)을 형성하기 위해 2개의 그룹들로 또한 나누어지며, 0857 및 1056은 하나의 그룹(93)에 있고 2764 및 4798은 상이한 그룹(94)에 있다.

[0034] 도 7에 대한 논의로 리턴하면, 각각의 클래스에 대한 범위들(85)이 분할 후에 결정된다. 할당된 범위들은 범위가 또 다른 클래스의 범위와 중첩하지 않는 한, 각각의 클래스에서 최저 및 최고 암호화된 데이터 값들로부터 형성될 수 있거나 대안적으로, 범위들은 부가적인 데이터 값들을 커버하기 위해 확장될 수 있다. 예를 들면, 클래스 I에 대한 범위는 데이터 값들에 기초하여 0857 내지 1056일 수 있지만; 범위는 또한, ≤1056 또는 ≤2700일 수 있다. 다른 범위들이 가능하다. 클래스 II에 대해, 암호화된 데이터 값들은 그 범위가 2764로부터 4798까지이고, 클래스 III에 대해, 암호화된 데이터 값들은 그 범위가 6321로부터 9921까지이다. 그러나, 범위들은 또한, 클래스 II에 대해 2705 내지 6320과 같고 클래스 III에 대해 6321 및 그보다 더 높은 것과 같이 더 넓을 수 있다. 다른 범위들이 가능하다. 범위 결정을 위한 방법은 미리 결정되고, 자동으로 결정되거나, 사용자에 의해 선택될 수 있다.

[0035] 일단 결정되면, 각각의 클래스의 범위는 그 다음, 그 클래스에 속하는 각각의 암호화된 데이터 값에 익명화된 값으로서 할당된다. 따라서, 나이(62)는 8791의 암호화된 데이터 값을 갖고 클래스 III에 속한다. 6321 내지 9921의 범위는 나이(62)에 익명화된 값으로서 할당된다. 부가적으로, 나이(27)는 1056의 암호화된 값을 갖고 클래스 I에 속하며, 따라서 범위(0857 내지 1056)는 나이(27)에 익명화된 데이터 값(86)으로서 할당된다.

[0036] 일단 하나 이상의 속성들에 대한 데이터 값들이 익명화되었으면, 데이터세트는 프로세싱 또는 분석을 위해 제 3 자에 제공될 수 있다. 순차적 암호화로 인해, 제 3 자는 암호화된 데이터 값들의 범위들에 기초하여 데이터세트의 젊은, 늙은, 그리고 중년의 인구들을 식별하는 것과 같은, 연구를 위해 데이터를 활용할 수 있을 것이다. 예를 들면, 더 낮은 순서화된 범위들은 더 낮은 나이들과 연관되는 반면에, 더 높은 순서화된 범위들은 더 높은 나이들과 연관된다.

[0037] 또 다른 실시예에서, 데이터 값들은 트리에 기초하여 상향식(bottom up) 방법을 사용하여 클래스들에 형성되고, 여기서 트리의 얼마나 많은 리프(leaf) 노드들이, n개의 클래스들 및 각각의 클래스에 대한 k개의 데이터 값들이 만족되는지를 보장하기 위해 조합되어야 하는지에 관한 결정이 행해진다.

[0038] 또 다른 실시예에서, 속성들 중 2개 이상에 대한 데이터 값들은 속성들의 데이터 값들 사이의 관계가 유지됨을 보장하기 위해 함께 익명화될 수 있다. 도 9는 2개 이상의 속성들의 데이터를 결합적으로 익명화하기 위한 프로세스(100)를 예로서 도시하는 흐름도이다. 2개의 중요한 속성들이 익명화를 위해 선택된다. 익명화를 위한 속성들은 감도 값, 데이터 값 콘텐트, 또는 중요도에 기초하여 뿐만 아니라, 다른 인자들에 기초하여 선택될 수 있다. 예를 들면, 2개의 속성들의 데이터 값들은 순차적 암호화를 사용하여 개별적으로 암호화된다(블록(101)). 다음, 속성들 중 하나는 익명화 프로세싱을 위해 선택된다(블록(102)). 몇몇 가능한 선택 전략들이 존재한다. 예를 들면, 감도 값에 기초할 때, 최고 감도를 갖는 속성은 제 1 속성으로서 선택될 수 있고, 그 역도 마찬가지이다. 또 다른 확률은 랜덤으로 또는 속성들이 취하는 고유 값들의 수에 기초하여 속성을 선택하는 것이다. 그러나, 선택을 위한 다른 전략들이 가능하다. 선택된 속성의 데이터 값들은 그 다음, n개의 최종 클래스들의 값을 만족시키는 것과 같은, 중지 지점 조건이 만족될 때까지, 클래스들로 분할된다(103). 속성 데이터의 분할에 기초하여, 클래스들은 제 1 선택된 속성에 대한 암호화된 데이터 값들로부터 형성되고(블록(104)) 각각의 클래스에 대한 범위들이 결정된다. 중지 지점이 만족되고 데이터 값들이 더 이상 분할되지 않을 때, 각각의 클래스에 대한 범위들이 결정되고 그 다음, 그 클래스에 포함되는 암호화된 데이터 값들에 익명화된 데이터 값들로서 할당된다(블록(105)). 다음, 임의의 또 다른 속성들이 프로세싱을 위해 남아 있는지의 여부에 관한 결정이 행해지고(블록(106)) 그렇다면, 속성 선택(블록(102)) 및 분할(블록(103)) 뿐만 아니라, 분할로부터 생성된 클래스에 대한 라벨들의 할당(블록(104))이 발생한다. 그러나, 그렇지 않으면, 2개 이상의 속성들에 대해 발생된 클래스들은 조합된 클래스들로 그룹핑되고(블록(107)), 도 10에 관하여 하기에 또한 설명된 바와 같이, 조합된 속성 값들의 익명화된 값들에 대한 라벨들이 생성된다. 일단 익명화 값들이 결정되었으면, 결정된 익명화 값들은 그 다음, 요청하는 제 3 자에 데이터를 제공하기 이전에, 데이터 세트에서 조합된 속성들의 각각의 데이터 값에 대한 평문 데이터 값들을 대체할 수 있다.

[0039] 2개 이상의 속성들에 대한 조합된 클래스들의 그룹핑은 랜덤으로, 또는 미리 결정된 기반으로, 또는 사용자에

의해 지시된 바와 같이 발생할 수 있다. 도 10은 2개의 상이한 속성들에 대한 클래스들의 차트(120)를 도시하는 블록도이다. 속성 1과 같은, 하나의 속성(121)에 대한 클래스들은 차트의 상부 행을 따라 열거되는 반면에, 속성 2와 같은, 다른 속성(122)의 클래스들은 차트(120)의 열을 따라 열거된다. 상이한 속성들에 대한 클래스들의 그룹핑은 미리 결정된 n-분할 값에 기초한다. 이 예에서, n=3이고 따라서, 2개의 속성들의 클래스들은 3개의 그룹들로 조합되어야 한다. 분할은 미리 결정된 방법과 같이, 또는 사용자에 의해 지시된 바와 같이 랜덤으로 수행될 수 있다. 이 예에서, 총 3개의 그룹들은 도 10에 표시된 바와 같이 형성된다. 이것은 n개의 컬러들로 표의 셀들을 칠하는 것과 같고, 여기서 n은 원하는 그룹들 또는 세그먼트들의 수이다. 후속적으로, 최종 그룹들에 기초하여, 익명화 값들은 각각의 그룹에 할당된다. 이전에 설명된 실시예와 유사하게, 사용자는 그룹 인덱스들에 또는 그룹 인덱스들의 수학 함수에 순차적 암호화를 적용함으로써 익명화에 대한 값들을 생성할 수 있다.

[0040] 결합 익명화는 제 3 자가 다수의 속성들을 만족시키는 개인들을 식별하기를 원할 때, 이롭다. 예를 들면, 연구자는 캘리포니아에 사는 62 내지 80세의 나이들 사이의 알츠하이머 환자들의 수를 결정하기 위해 병원으로부터 데이터를 얻는다. 이 예에서, 질병에 대한 데이터 값들은 평문으로서 남을 수 있는 반면에, 나이 및 집 코드들 값들은 익명화된다. 하나의 실시예에서, 나이가 먼저 익명화되고 그 다음, 각각의 나이 값과 연관된 집 코드들이 각각의 식별된 집 코드에 사는 62 및 81의 나이들 사이의 알츠하이머 환자들을 식별하기 위해 익명화된다. 캘리포니아에 사는 62 및 81의 나이들 사이의 총 수의 알츠하이머 환자들은 그 다음, 캘리포니아에서 각각의 집 코드에 대한 환자들의 수를 합산함으로써 식별된다. 또 다른 실시예에서, 나이에 대한 데이터 값들이 먼저 익명화되고 그 다음, 각각의 집 코드와 연관된 주들이 캘리포니아에 사는 62 및 81의 나이들 사이의 알츠하이머 환자들 대 캘리포니아에 살지 않는 62 및 81의 나이들 사이의 그들 알츠하이머 환자들을 식별하기 위해 익명화된다.

[0041] 게다가, 상기 백화점 예로 리턴하면, 제 3 자 데이터 요청자는 평균 고객 비용을 결정하거나 가장 많은 돈을 쓰고, 백화점을 가장 자주 방문하며, 백화점에 의해 발행된 신용 카드들을 갖는 고객들의 나이 그룹을 발견하기를 원한다. 이 예에서, 제 3 자는 더 늙은 나이 또는 더 젊은 나이의 개인들이 도 5에 도시된 바와 같이, 익명화된 나이 데이터 및 평문 구매 금액들의 조합을 사용하여 더 많은 돈을 썼는지의 여부를 식별할 수 있을 것이다. 게다가 여전히, 제 3 자는 또한, 가장 많은 돈을 쓰는 개인들이 사는 지역들을 결정하기를 원한다. 제 3 자에 요청된 데이터를 제공하기 위해, 데이터 값들의 관계는 익명화 동안 유지되어야 한다. 관계들은 상기 설명된 바와 같이, 제 1 속성에 대한 암호화된 데이터 값들의 그룹들에 익명화 값들을 할당하고 제 2 속성을 분리하며 그 다음, 제 1 및 제 2 속성의 클래스들을 그룹핑함으로써 유지될 수 있다.

[0042] 일단 결정되고 할당되면, 익명화 값들은 그 다음, 데이터 세트에서 중요한 속성들의 평문 데이터 값들을 대체하기 위해 사용된다. 도 11은 익명화된 값들을 갖는 도 5의 데이터 세트를 예로서 도시하는 블록도이다. 도 5에서와 같이, 데이터 세트는 이름, 성, 나이, 계좌 번호, 집 코드, 및 구매 금액에 대한 속성들(61)을 포함한다. 각각의 속성은 데이터 세트에 의해 표현된 개인들에 대한 평문 데이터 값들(62)과 연관된다. 나이 및 집 코드 속성들에 대한 평문 데이터 값들은 신뢰되지 않은 제 3 자가 데이터 세트 내에서 데이터 값들에 의해 표현된 특정 개인을 식별하는 것을 방지하기 위해 익명화된 데이터 값들(63)과 대체되었다. 예를 들면, 익명화된 데이터 값들은 도 7 및 도 8에 관하여 상기 설명된 바와 같이, 나이 속성들에 대해 결정되었다.

[0043] 이 데이터 세트 내에서, 신뢰되지 않은 제 3 자는 아마도 여전히 개인들을 식별할 수 있고, 이는 이름 및 성이 각각의 개인에 제공되기 때문이다. 개인들의 계정들을 보호하기 위해, 이름 및 성에 대한 데이터 값들이 마스킹될 수 있다. 마스킹은 데이터가 중요할 때 또는 주 또는 연방 규칙들에 의해 지시될 때, 특정한 속성들을 위해 요구될 수 있다. 또 다른 옵션은 그들 속성들을 단순하게 억제하는 것이다. 도 12는 데이터 값들을 마스킹하기 위한 프로세스(140)를 예로서 도시하는 흐름도이다. 데이터 세트 내의 적어도 하나의 속성은 중요한 것으로 결정되거나(블록(141)) 주 또는 연방 규정들 하에서 마스킹되도록 요구된다. 하나의 예에서, 데이터 소유자는 익명화기에 데이터 세트를 익명화하도록 지시하고, 상기 데이터 세트는 건강 보험 양도 및 책임에 관한 법(HIPAA; Health Insurance Portability and Accountability Act) 하에 속하는 데이터를 포함한다. HIPAA 하에서, 특정 속성들에 대한 데이터 값들은 완전하게 애매하게 만들도록 요구된다. 익명화기는 HIPPA의 규정들로 프로그래밍되고 HIPPA 요구조건들에 기초하여 마스킹하기 위해 그들 속성들을 식별한다.

[0044] 중요한 속성에 대한 데이터 값이 선택되고(블록(142)) 선택된 데이터 값에 마스킹 값을 할당함으로써 랜덤으로 마스킹된다(블록(143)). 마스킹은 치환되는 데이터 값과 비슷한 또 다른 값과의 데이터 값의 치환, 암호 해시와 같은, 의사 랜덤 함수를 사용한 데이터 함수의 치환, 또는 암호화를 포함할 수 있다. 다른 유형들의 마스킹이 가능하다. 마스킹되지 않은 데이터 값들이 속성에 대해 남아 있으면(블록(145)), 또 다른 데이터 값이 마스킹을

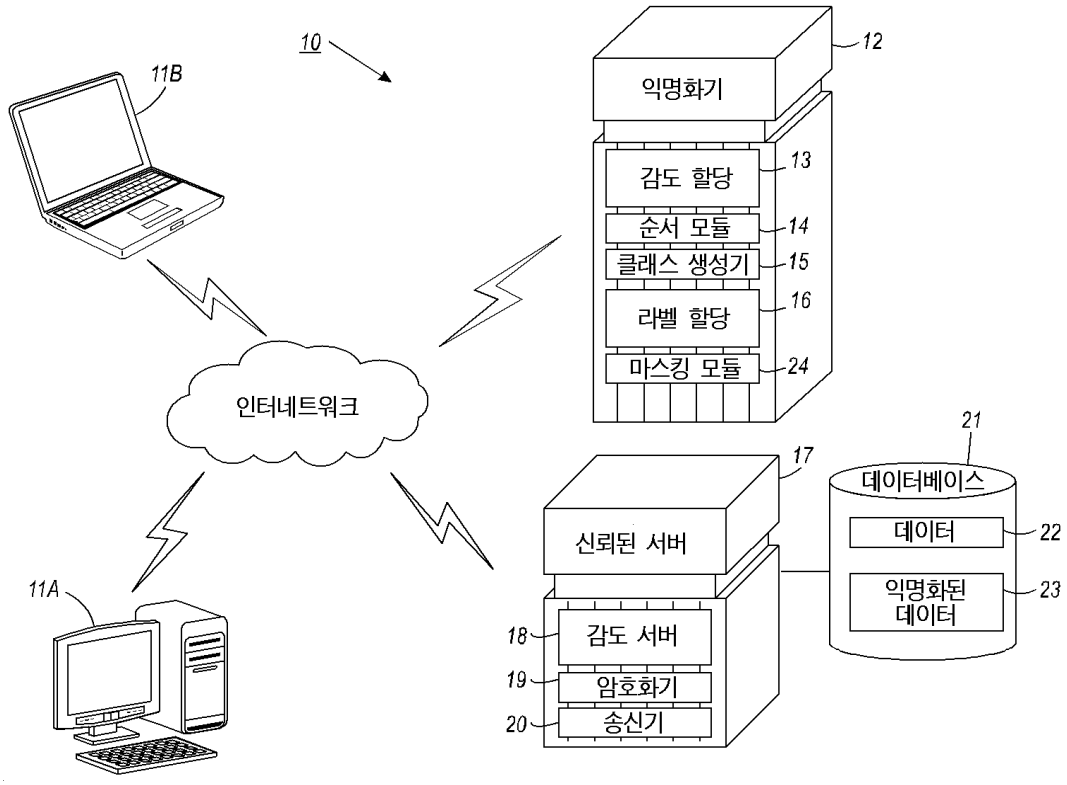
위해 선택된다(블록(142)).

[0045] 마스킹된 데이터 값들은 중요한 속성에 대한 평문 데이터 값을 데이터 세트 내의 익명화된 데이터로서 대체한다. 도 13은 속성들 중 2개의 속성들에 대한 마스킹된 데이터 값들을 갖는 도 11의 데이터 세트(150)를 예로서 도시하는 블록도이다. 이 데이터 세트(150)에서, 이름 및 성에 대한 속성들(151)은 중요한 것으로서 식별되었고 데이터 값들은 그들의 프라이버시 감도에 따라 프로세싱되었다. 구체적으로, 이름 및 성 속성들에 대한 데이터 값들은 해싱에 의해 마스킹되었다. 부가적으로, 유사 식별자들로 고려되는 나이 및 집 코드 속성들은 일반화에 의해 익명화되었다. 데이터 세트에 남아 있는 평문 데이터 값들은 평문으로 유지될 수 있거나 제 3 자에 데이터세트를 제공하기 이전에 암호화될 수 있다.

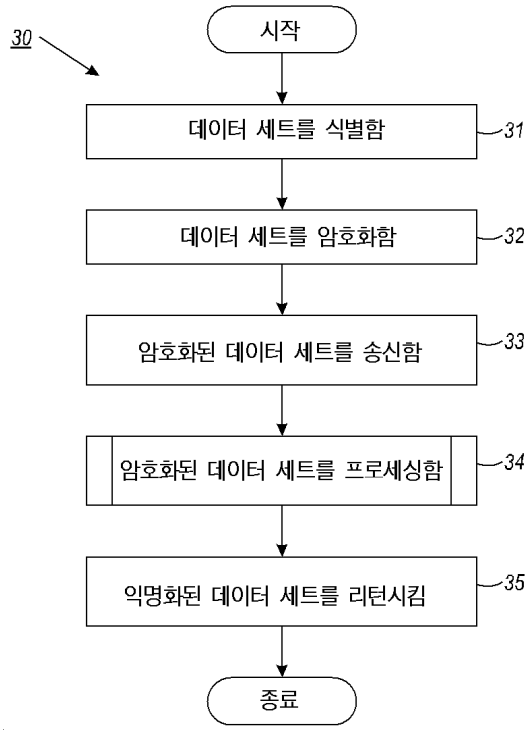
[0046] 일반화와 마찬가지로, 데이터 소유자는 중요한 속성들을 암호화하고 익명화기에 그들을 마스킹하도록 요청할 수 있다. 이 경우에 익명화기는 암호화된 데이터에 대해 동작한다.

도면

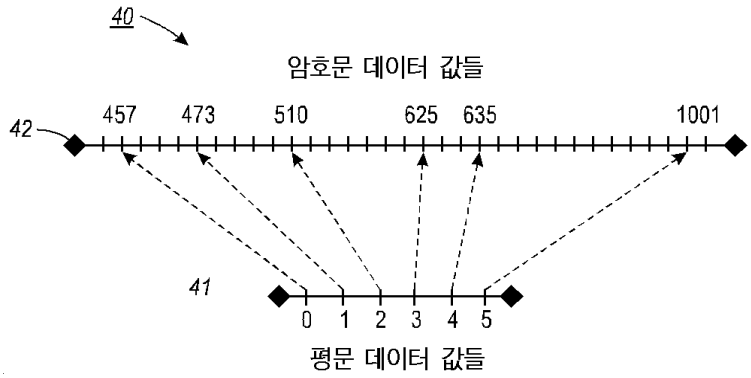
도면1



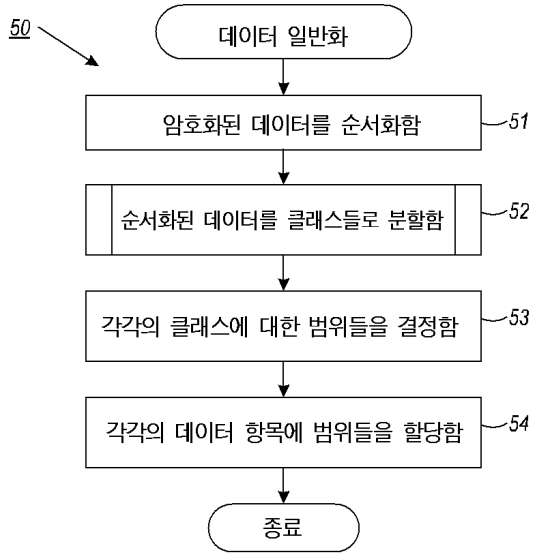
도면2


30
시작
데이터 세트를 식별함
31
데이터 세트를 암호화함
32
암호화된 데이터 세트를 송신함
33
암호화된 데이터 세트를 프로세싱함
34
익명화된 데이터 세트를 리턴시킴
35
종료

도면3


40
암호문 데이터 값들
42
457
473
510
625
635
1001
41
0
1
2
3
4
5
평문 데이터 값들

도면4

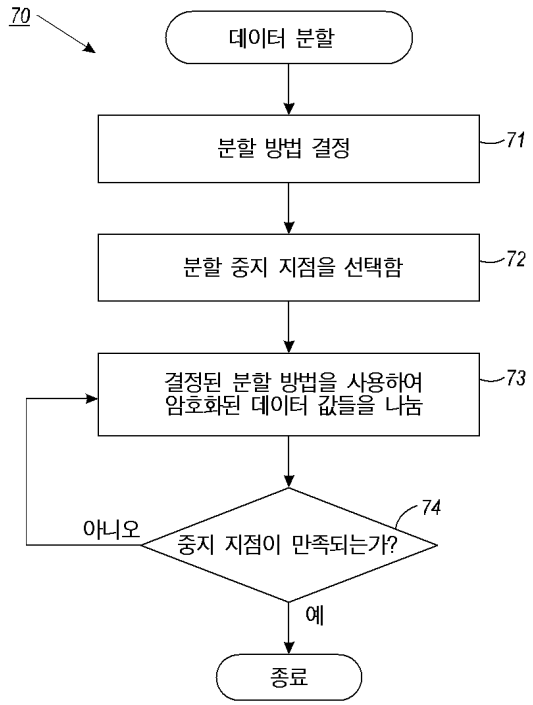


도면5

60

61 이름	62 성	63 나이	64 계좌 번호	65 집 코드	66 매입액
베티	우	62	14578	21768	$ 1,057.21
존	아일랜드	68	16743	01764	$ 4,109.15
모리스	토마스	27	09673	94602	$ 11,012.17
러셀	스미스	36	37810	98125	$ 19,429.08
수잔	야마모토	45	41137	14557	$ 7,120.54
마숀	셔먼	43	28759	98109	$ 3,235.07
조	리	52	84221	57214	$ 2,250.14
브레일런	윌슨	22	02356	78265	$ 9,283.16

도면6



도면7

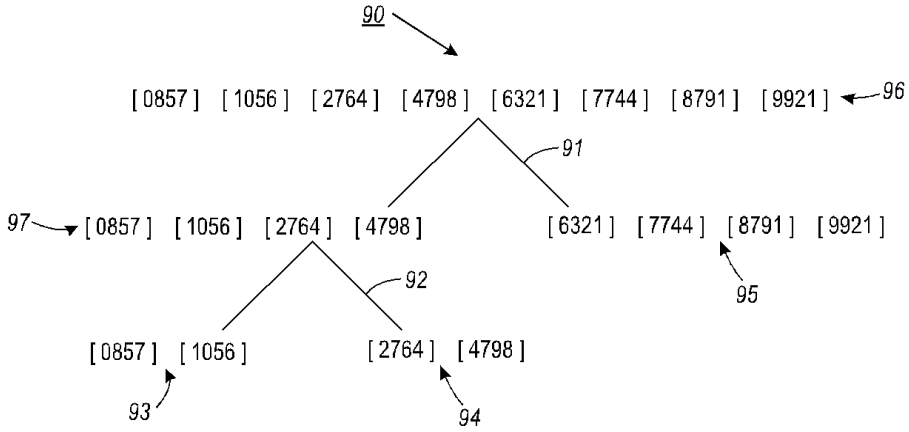
80

나이	암호화된 값들	순서	분할		범위	익명화된 라벨
62	[8791]	[0857]	[0857]	[0857]	클래스 I	[6321]-[9921]
68	[9921]	[1056]	[1056]	[1056]	[0857]-[1056]	[6321]-[9921]
27	[1056]	[2764]	[2764]	[2764]	클래스 II	[0857]-[1056]
36	[2764]	[4798]	[4798]	[4798]	[2764]-[4798]	[2764]-[4798]
45	[6321]	[6321]	[6321]	[6321]		[6321]-[9921]
43	[4798]	[7744]	[7744]	[7744]	클래스 III	[2764]-[4798]
52	[7744]	[8791]	[8791]	[8791]	[6321]-[9921]	[6321]-[9921]
22	[0857]	[9921]	[9921]	[9921]		[0857]-[1056]

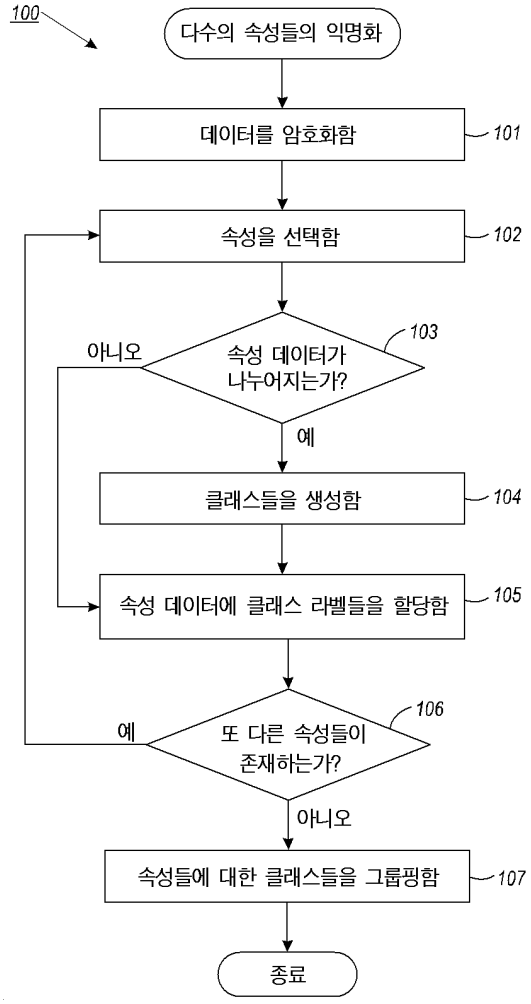
81 82 83 84a 84b 85 86

n=2
클래스들=3

도면8


90
[0857] [1056] [2764] [4798] [6321] [7744] [8791] [9921]
96
91
97
[0857] [1056] [2764] [4798]
[6321] [7744] [8791] [9921]
95
92
[0857] [1056]
[2764] [4798]
93
94

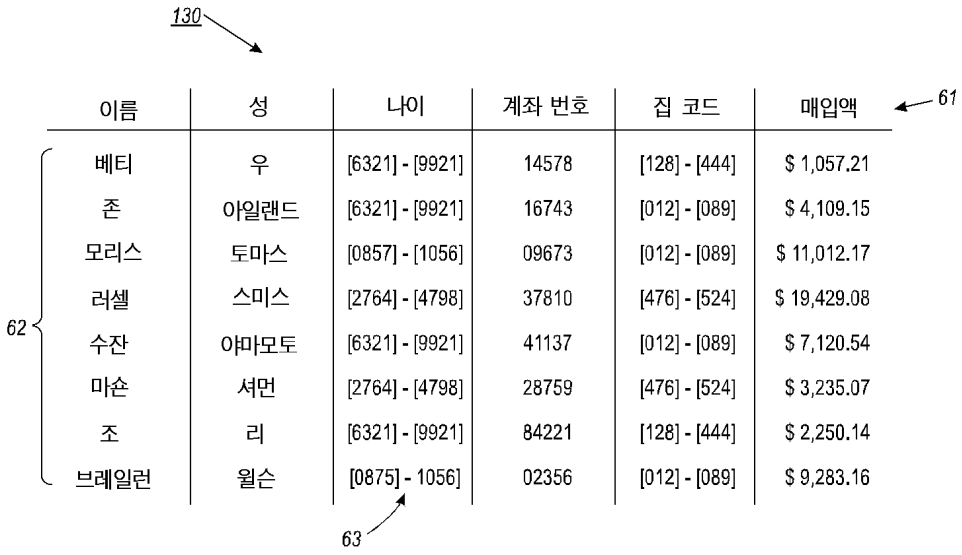
도면9


100
다수의 속성들의 익명화
데이터를 암호화함
101
속성을 선택함
102
속성 데이터가 나누어지는가?
103
아니오
예
클래스들을 생성함
104
속성 데이터에 클래스 라벨들을 할당함
105
또 다른 속성들이 존재하는가?
106
예
아니오
속성들에 대한 클래스들을 그룹핑함
107
종료

도면10

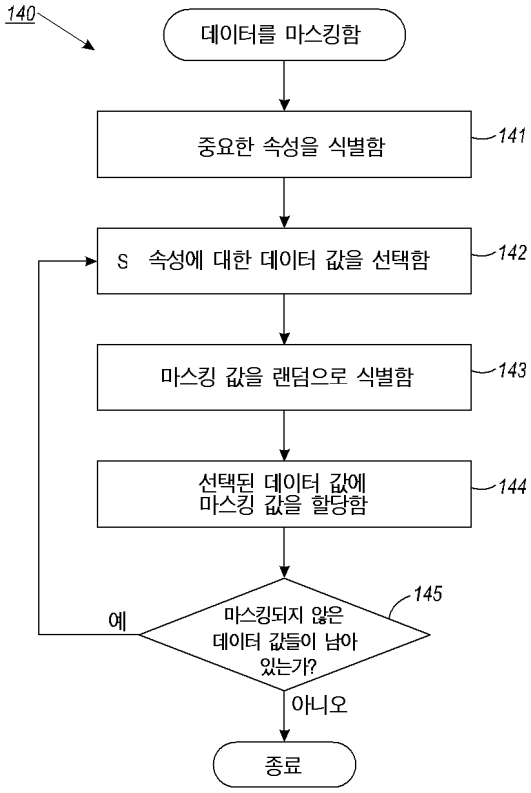
	속성 1 클래스$_1$	속성 1 클래스$_2$	속성 1 클래스$_3$	속성 1 클래스$_4$
속성 2 클래스$_1$	그룹 3	그룹 1	그룹 2	그룹 2
속성 2 클래스$_2$	그룹 3	그룹 1	그룹 2	그룹 2
속성 2 클래스$_3$	그룹 3	그룹 3	그룹 1	그룹 1

도면11



이름	성	나이	계좌 번호	집 코드	매입액
베티	우	[6321] - [9921]	14578	[128] - [444]	$ 1,057.21
존	아일랜드	[6321] - [9921]	16743	[012] - [089]	$ 4,109.15
모리스	토마스	[0857] - [1056]	09673	[012] - [089]	$ 11,012.17
러셀	스미스	[2764] - [4798]	37810	[476] - [524]	$ 19,429.08
수잔	야마모토	[6321] - [9921]	41137	[012] - [089]	$ 7,120.54
마숀	셔먼	[2764] - [4798]	28759	[476] - [524]	$ 3,235.07
조	리	[6321] - [9921]	84221	[128] - [444]	$ 2,250.14
브레일런	윌슨	[0875] - 1056]	02356	[012] - [089]	$ 9,283.16

도면12



도면13

150

이름	성	나이	계좌 번호	집 코드	매입액
[22568893]	[48589921]	[6321] - [9921]	14578	[128] - [444]	$ 1,057.21
[58965899]	[64785031]	[6321] - [9921]	16743	[012] - [089]	$ 4,109.15
[00589758]	[47023568]	[0857] - [1056]	09673	[012] - [089]	$ 11,012.17
[13214485]	[08856123]	[2764] - [4798]	37810	[476] - [524]	$ 19,429.08
[98865218]	[74002441]	[6321] - [9921]	41137	[012] - [089]	$ 7,120.54
[23569109]	[99982032]	[2764] - [4798]	28759	[476] - [524]	$ 3,235.07
[11458011]	[13120336]	[6321] - [9921]	84221	[128] - [444]	$ 2,250.14
[85623410]	[38556011]	[0875] - 1056]	02356	[128] - [444]	$ 9,283.16

151

152