(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0065073
(43) 공개일자 2023년05월11일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) *G06F 16/26* (2019.01)
G06F 16/901 (2019.01)
(52) CPC특허분류
G06F 21/6245 (2013.01)
G06F 16/26 (2019.01)
(21) 출원번호 10-2021-0150896
(22) 출원일자 2021년11월04일
심사청구일자 2021년11월04일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정연돈
서울특별시 중구 청계천로 400, 105동 812호 (황학동, 롯데캐슬베네치아)
백인철
서울특별시 성북구 안암로9가길 53-1, 104호 (안암동5가)
김종선
서울특별시 성북구 고려대로 29, 404호 (안암동2가)
(74) 대리인
특허법인엠에이피에스

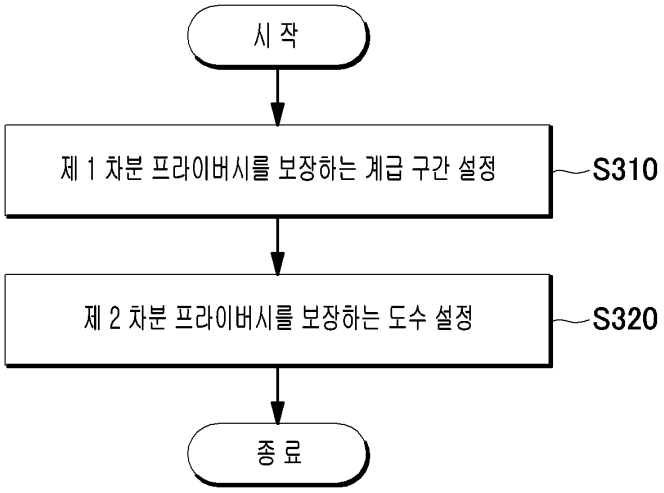
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **차분 프라이버시를 보장하는 히스토그램 생성 장치 및 방법**

(57) 요 약

본 발명의 일 실시예는 차분 프라이버시를 보장하는 히스토그램 생성 방법에 관한 것으로, 관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계 및 상기 설정된 계급 구간 별로 제 2 차분 프라이버시를 보장하는 도수를 산출하는 단계를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G06F 16/9024 (2019.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1345342146
과제번호 2021R1A6A1A13044830
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축(R&D)_대학중점연구소지원사업
연구과제명 융합소프트웨어연구소
기 여 율 2/10
과제수행기관명 고려대학교
연구기간 2021.06.01 ~ 2022.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711132530
과제번호 2019H1D8A2105513
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 차세대 공학연구자 육성사업
연구과제명 D.N.A (Data, Network, AI) 분야 차세대 공학연구자 육성 사업단
기 여 율 2/10
과제수행기관명 고려대학교
연구기간 2019.11.01 ~ 2022.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711134602
과제번호 2021-0-00634-001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보보호핵심원천기술개발(R&D,정보화)
연구과제명 (세부1) 대용량 정형 데이터 대상 개인정보 가명.익명화를 위한 자동처리 기술
기 여 율 5/10
과제수행기관명 주식회사 이지서티
연구기간 2021.04.01 ~ 2024.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711120246
과제번호 2020001819001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 ICT명품인재양성
기 여 율 1/10
과제수행기관명 고려대학교
연구기간 2020.07.01 ~ 2029.12.13

명 세 서

청구범위

청구항 1

차분 프라이버시를 보장하는 히스토그램 생성 방법에 있어서,

관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계 및

상기 설정된 계급 구간 별로 제 2 차분 프라이버시를 보장하는 도수를 산출하는 단계를 포함하는, 히스토그램 생성 방법.

청구항 2

제1항에 있어서,

상기 계급 구간을 설정하는 단계는

휴리스틱 알고리즘에 따라 계급의 개수를 결정하는 단계 및

라플라스 메커니즘을 적용하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계를 포함하는, 히스토그램 생성 방법.

청구항 3

제2항에 있어서,

상기 휴리스틱 알고리즘으로는 제곱근 선택(Square-root choice), 스터지의 공식(Sturges' formula), 라이스 규칙(Rice rule) 중 어느 하나가 선택되는 것인, 히스토그램 생성 방법.

청구항 4

차분 프라이버시를 보장하는 히스토그램 생성 장치에 있어서,

히스토그램 생성 프로그램이 저장된 메모리; 및

상기 메모리에 저장된 프로그램을 실행하는 프로세서를 포함하고,

상기 프로그램은 관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계와, 상기 설정된 계급 구간 별로 제 2 차분 프라이버시를 보장하는 도수를 산출하는 단계를 수행하는, 히스토그램 생성 장치.

청구항 5

제4항에 있어서,

상기 계급 구간을 설정하는 단계는

휴리스틱 알고리즘에 따라 계급의 개수를 결정하는 단계 및

라플라스 메커니즘을 적용하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계를 포함하는, 히스토그램 생성 장치.

청구항 6

제5항에 있어서,

상기 휴리스틱 알고리즘으로는 제곱근 선택(Square-root choice), 스터지의 공식(Sturges' formula), 라이스 규칙(Rice rule) 중 어느 하나가 선택되는 것인, 히스토그램 생성 장치

청구항 7

제1항에 따른 히스토그램 생성 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 비일시적 컴퓨터 판독가능 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 차분 프라이버시를 보장하는 히스토그램 생성 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 데이터의 수집과 분석에 대한 관심이 커지고 있는 가운데 최근에는 데이터에 포함된 개인의 프라이버시를 보호하면서 데이터 분석 결과를 배포 하기 위해 다양한 프라이버시 모델이 제안되었다. 그 중 차분 프라이버시(differential privacy) 는 프라이버시 보호 수준을 수학적으로 보 할 수 있어 사실상의 표준 프라이버시 보호 모델로 받아들여지고 있다.

[0003] 본 발명에서는 특히, 통계적인 정보를 이용하여 생성하는 히스토그램을 생성하과는 과정에서도 개인의 프라이버시를 보호하려고 한다.

[0004] 도 1 은 히스토그램에 의한 프라이버시 침해 예시를 도시한 도면이다.

[0005] 히스토그램을 원본 그대로 공개하면 데이터에 포함된 개인의 프라이버시를 침해할 위험이 있다. 앨리스가 퇴원한 뒤 히스토그램을 다시 생성하면 50세와 60세 구간의 도수가 10 에서 9 로 감소하기에 공격자는 이를 통해 앨리스의 나이가 50세와 60세 사이임을 확인할 수 있게 되는 문제가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국등록특허 제10-2055864호(발명의 명칭: 차분 프라이버시를 적용한 시간 간격 데이터 공개 방법)

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 문제점을 해결하기 위한 것으로, 관계형 데이터에 대한 히스토그램을 생성하는 과정에서 차분 프라이버시를 보장할 수 있는 히스토그램 생성 장치 및 방법을 제공하는 것을 일 기술적 과제로 한다.

[0008] 본 발명이 이루고자 하는 기술적 과제들은 상기한 기술적 과제로 제한되지 않으며, 이하의 설명으로부터 본 발명의 또 다른 기술적 과제들이 도출될 수 있다.

과제의 해결 수단

[0009] 상술한 기술적 과제를 해결하기 위한 기술적 수단으로서, 본 발명의 제1 측면에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법은, 관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계 및 상기 설정된 계급 구간 별로 제 2 차분 프라이버시를 보장하는 도수를 산출하는 단계를 포함한다.

[0010] 또한, 본 발명의 제2 측면에 따른 차분 프라이버시를 보장하는 히스토그램 생성 장치는 히스토그램 생성 프로그램이 저장된 메모리; 및 상기 메모리에 저장된 프로그램을 실행하는 프로세서를 포함하고, 상기 프로그램은 관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계와, 상기 설정된 계급 구간 별로 제 2 차분 프라이버시를 보장하는 도수를 산출하는 단계를 수행한다.

발명의 효과

[0011] 본 발명에 따르면, 차분 프라이버시를 보장하는 히스토그램을 이용하여 개인의 프라이버시를 보호하면서 히스토그램 분석 및 배포를 할 수 있다. 또한, 차분 프라이버시를 보장하는 과정이 자동화되어 비전문가도 쉽게 개인

의 프라이버시를 보호하면서 활용 가능한 히스토그램을 생성할 수 있다.

도면의 간단한 설명

[0012] 도 1 은 히스토그램에 의한 프라이버시 침해 예시를 도시한 도면이다.

도 2는 본 발명의 일 실시예에 따른 히스토그램 생성 장치(100)를 도시한 도면이다.

도 3은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법을 도시한 순서도이다.

도 4는 히스토그램을 정의하는 구간과 도수를 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법을 나타내는 의사코드를 도시한 것이다.

도 6은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법에서 계급 구간을 설정하는 과정의 의사코드를 도시한 것이다.

도 7은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법에서 계급 구간의 개수를 산출하는 휴리스틱 알고리즘을 도시한 것이다.

도 8은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법에서 도수를 설정하는 과정의 의사코드를 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

[0013] 이하에서는 첨부한 도면을 참조하여 본 발명을 상세히 설명하기로 한다. 다만, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예들로 한정되는 것은 아니다. 또한, 첨부된 도면은 본 명세서에 개시된 실시예를 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 명세서에 개시된 기술적 사상이 제한되지 않는다. 여기에 사용되는 기술용어 및 과학용어를 포함하는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 일반적으로 이해하는 의미로 해석되어야 한다. 사전에 정의된 용어들은 관련기술문헌과 현재 개시된 내용에 부합하는 의미를 추가적으로 갖는 것으로 해석되어야 하며, 별도로 정의되지 않는 한 매우 이상적이거나 제한적인 의미로 해석되지 않는다.

[0014] 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 도면에 나타난 각 구성요소의 크기, 형태, 형상은 다양하게 변형될 수 있다. 명세서 전체에 대하여 동일/유사한 부분에 대해서는 동일/유사한 도면 부호를 붙였다.

[0015] 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부" 등은 명세서 작성의 용이함만이 고려되어 부여 되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다. 또한, 본 명세서에 개시된 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 명세서에 개시된 실시 예의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략하였다.

[0016] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결(접속, 접촉 또는 결합)"되어 있다고 할 때, 이는 "직접적으로 연결(접속, 접촉 또는 결합)"되어 있는 경우뿐만 아니라, 그 중간에 다른 부재를 사이에 두고 "간접적으로 연결(접속, 접촉 또는 결합)"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함(구비 또는 마련)"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 "포함(구비 또는 마련)"할 수 있다는 것을 의미한다.

[0017] 본 명세서에서 사용되는 제1, 제2 등과 같이 서수를 나타내는 용어들은 하나의 구성 요소를 다른 구성요소로부터 구별하는 목적으로만 사용되며, 구성 요소들의 순서나 관계를 제한하지 않는다. 예를 들어, 본 발명의 제1구성요소는 제2구성요소로 명명될 수 있고, 유사하게 제2구성요소도 제1구성 요소로 명명될 수 있다. 본 명세서에서 사용되는 단수 표현의 형태들은 명백히 반대의 의미를 나타내지 않는 한 복수 표현의 형태들도 포함하는 것으로 해석되어야 한다.

[0018] 도 2는 본 발명의 일 실시예에 따른 히스토그램 생성 장치(100)를 도시한 도면이다.

[0019] 도 2를 참조하면, 히스토그램 생성 장치(100)는 통신 모듈(110), 메모리(120), 프로세서(130) 및 데이터 베이스(140)를 포함한다.

[0020] 히스토그램 생성 장치(100)는 서버 또는 단말과 같은 장치 형태로 형성될 수 있고, SaaS (Software as a Service), PaaS (Platform as a Service) 또는 IaaS (Infrastructure as a Service)와 같은 클라우드 컴퓨팅 서비스 모델에서 동작 할 수 있다. 또한, 히스토그램 생성 장치(100)는 사설(private) 클라우드, 공용(public) 클라우드 또는 하이브리드(hybrid) 클라우드 시스템과 같은 서버 형태로 구축될 수 있다.

[0021] 통신 모듈(110)은 외부 단말 또는 컴퓨팅 장치와의 정보 송수신을 수행하며, 이를 통해 히스토그램 생성을 위한 관계형 데이터를 수신하거나, 본 발명에 따라 생성된 차분 프라이버시가 보장된 히스토그램을 생성하여 외부 단말 또는 컴퓨팅 장치에 전송한다. 통신 모듈(110)은 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치를 포함할 수 있다.

[0022] 메모리(120)는 차분 프라이버시를 보장하는 히스토그램 생성 프로그램을 저장한다. 히스토그램 생성 프로그램의 명칭은 설명의 편의를 위해 설정된 것으로, 명칭 그 자체로 프로그램의 기능을 제한하는 것은 아니다. 히스토그램 생성 프로그램은 관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계와, 설정된 계급 구간 별로 제 2 차분 프라이버시를 보장하는 도수를 산출하는 단계를 수행하며, 구체적인 구성은 추후 설명하기로 한다.

[0023] 메모리(120)는 통신 모듈(110)로 입력되는 정보 및 데이터, 프로세서(130)에 의해 수행되는 기능에 필요한 정보 및 데이터, 프로세서(130)의 실행에 따라 생성된 데이터 중 적어도 어느 하나 이상을 저장할 수 있다. 메모리(120)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력을 필요로 하는 휘발성 저장장치를 통칭하는 것으로 해석되어야 한다. 또한, 메모리(120)는 프로세서(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행할 수 있다. 메모리(120)는 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치 외에 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0024] 프로세서(130)는 메모리(120)에 저장된 히스토그램 생성 프로그램을 실행하도록 구성된다. 프로세서(130)는 데이터를 제어 및 처리하는 다양한 종류의 장치들을 포함할 수 있다. 프로세서(130)는 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 일 예에서, 프로세서(130)는 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 형태로 구현될 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0025] 데이터베이스(140)는 차분 프라이버시를 보장하는 히스토그램 생성 방법에 따라 생성된 히스토그램에 대한 정보를 관리한다.

[0026] 도 3은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법을 도시한 순서도이고, 도 4는 히스토그램을 정의하는 구간과 도수를 설명하기 위한 도면이고, 도 5는 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법을 나타내는 의사코드(pseudocode)를 도시한 것이고, 도 6은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법에서 계급 구간을 설정하는 과정의 의사코드를 도시한 것이고, 도 7은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법에서 계급 구간의 개수를 산출하는 휴리스틱 알고리즘을 도시한 것이고, 도 8은 본 발명의 일 실시예에 따른 차분 프라이버시를 보장하는 히스토그램 생성 방법에서 도수를 설정하는 과정의 의사코드를 도시한 것이다.

[0027] 먼저, 관계형 데이터에 대하여 제 1 차분 프라이버시를 보장하는 계급 구간을 설정하는 단계를 수행한다(S310).

[0028] 관계형 데이터는 도 1의 (a)에 도시된 바와 같이, 대상체가 갖는 다양한 속성에 대한 정보를 포함하는 것으로서, 예를 들면, 대상이 되는 집합에 속한 각 개인들의 나이, 신체 정보, 소속 정보 들을 서로 연결한 데이터이다. 이러한 정보를 토대로 히스토그램을 생성하게 된다.

[0029] 이때, 본 발명에서는 도 4에 도시된 바와 같이, 히스토그램은 가로축과 세로축에 계급(bins)과 도수(frequecny)를 나타낸 막대 그래프로서, 본 발명에서는 히스토그램을 정의하는 구간과 도수를 각각 설정한다. 계급은 데이터의 어떤 속성이 가질 수 있는 최솟값과 최댓값 사이의 범위를 겹치지 않게 나눈 구간 중 하나이다. 도수는 한 계급 구간에 포함되는 레코드의 개수를 의미한다. 히스토그램에서 레코드의 개수와 속성값의 분포에 따라 데이터 분포를 잘 나타낼 수 있는 계급의 개수를 정하는 것이 중요하다.

[0030] 계급 구간을 설정할때, 휴리스틱 알고리즘에 따라 개급의 개수를 결정한다. 이때, 휴리스틱 알고리즘으로는 제곱근 선택(Square-root choice), 스터지의 공식(Sturges' formula), 라이스 규칙(Rice rule) 중 어느 하나가 선택될 수 있다.

[0031] 데이터베이스 D의 전체 레코드 개수를 n 이라 할 때, 히스토그램 계급의 수 d는 각 휴리스틱 알고리즘에 따라 다음과 같이 정의될 수 있다.

[0032] $$\text{제곱근 선택: } d = \lceil \sqrt{n} \rceil$$
$$\text{스터지의 공식: } d = \lceil \log_2 n \rceil + 1$$
$$\text{라이스 규칙: } d = \lceil 2\sqrt[3]{n} \rceil$$

[0033] 다음으로, 라플라스 메커니즘을 적용하여 제 1 차분 프라이버시(ε_1- 차분 프라이버시)를 보장하는 계급 구간을 설정한다.

[0034] 서로 하나의 레코드만 다른 두 데이터베이스를 각각 D, D' 라고 하고, 데이터 분석 결과를 변조하는 함수는 M이라 할 때, M이 생성할 수 있는 모든 결괏값 S에 대하여 다음의 수학식 1을 만족하면 해당 함수 M은 ε-차분 프라이버시를 보장한다.

[0035] [수학식 1]

[0036] $$\Pr[M(D) = S] \leq \exp(\epsilon) \cdot \Pr[M(D') = S]$$

[0037] 또한, 데이터베이스 D, 함수 f, $z \sim Lap(0, \Delta f/\epsilon)$ 가 주어졌을 때, 라플라스 메커니즘 M은 원본 결괏값 f(D)에 실수 노이즈 z를 더하여 변화한다.

[0038] $$M(D) \triangleq f(D) + z$$

[0039] 이러한 경우, 민감도가 Δf 인 함수 f에 대해서 변조된 결괏값 f(D)+z는 ε-차분 프라이버시를 보장한다.

[0040] 한편, 휴리스틱 알고리즘에 라플라스 메커니즘을 적용하는 과정을 설명하기로 한다.

[0041] 레코드의 전체 개수를 구하는 질의를 f라 할 때, 질의 f의 민감도는 1이다.

[0042] $$\|f(D) - f(D')\|_1 \leq 1$$

[0043] 라플라스 메커니즘 M을 다음과 같이 정의할 때, M(D)는 라플라스 메커니즘이므로, ε_1-차분 프라이버시를 만족한다.

[0044] $$M(D) = f(D) + N, \qquad N \sim Lap(\frac{1}{\epsilon_1})$$

[0045] 앞서 언급한 휴리스틱 알고리즘중 제곱근 선택 알고리즘을 g_1, 스터지의 공식을 g_2, 라이스 규칙을 g_3라고 할 때, g_1, g_2, g_3는 모두 결정론적 함수이고, M(D)는 ε_1-차분 프라이버시를 만족하므로, $g_1(M(D))$, $g_2(M(D))$, $g_3(M(D))$는 모두 ε_1-차분 프라이버시를 만족하게 된다.

[0046] 앞선 단계(S310)의 수행 후, 설정된 계급 구간 별로 제 2 차분 프라이버시(ε_2-차분 프라이버시)를 보장하는 도수를 산출하는 단계를 수행한다(S320).

[0047] 본 단계는 도 8에 도시된 알고리즘에 기초하여 구현될 수 있다.

[0048] 히스토그램의 정의역과, 이 정의역을 d 등분하는 계급 $B_i(B_1, B_2, \ldots, B_d)$가 주어져 있다고 할 때, 레코드 하나만 다른 두 데이터베이스를 각각 D, D' 라고 하고, 각 계급의 도수를 구하는 함수를 $f: D \rightarrow Z^d$ 라고 할 때, 질의 f의 민감도는 1이다.

[0049] $$\|f(D) - f(D')\|_1 \leq 1$$

[0050] 라플라스 메커니즘 M을 다음과 같이 정의할 때, M(D)는 라플라스 메커니즘이므로, ε_2-차분 프라이버시를 만족한다.

[0051] $$M(D) = f(D) + N^d, \qquad N^d \sim Lap(\frac{1}{\epsilon_2})$$

[0052] 한편, 단계(S310)의 라플라스 메커니즘을 M1, 단계(S320)의 라플라스 메커니즘을 M2라고 할 때, 각각의 라플라스 메커니즘은 ε_1-차분 프라이버시와 ε_2-차분 프라이버시를 각각 만족하므로, 이를 순차적으로 적용하는 본 발명의 방법은 $(\varepsilon_1+\varepsilon_2)$-차분 프라이버시를 만족하게 된다. 이는 순차 합성 정리 (sequential composition theorem)에 따라 보장되는 특징에 해당한다.

[0053] 이상 지금까지 설명한 본 발명의 실시예에 따른 히스토그램 생성 방법은, 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0054] 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 상술한 설명을 기초로 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해되어야만 한다. 본 발명의 범위는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다. 본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0055] 100: 히스토그램 생성 장치

110: 통신모듈

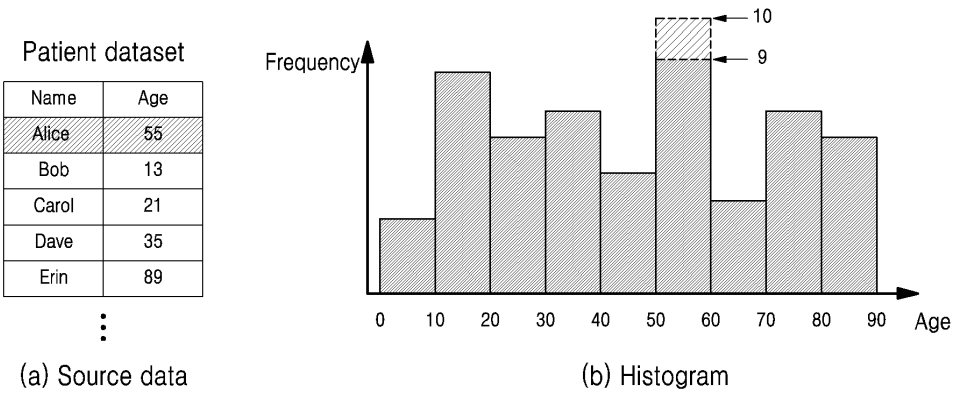
120: 메모리

130: 프로세서

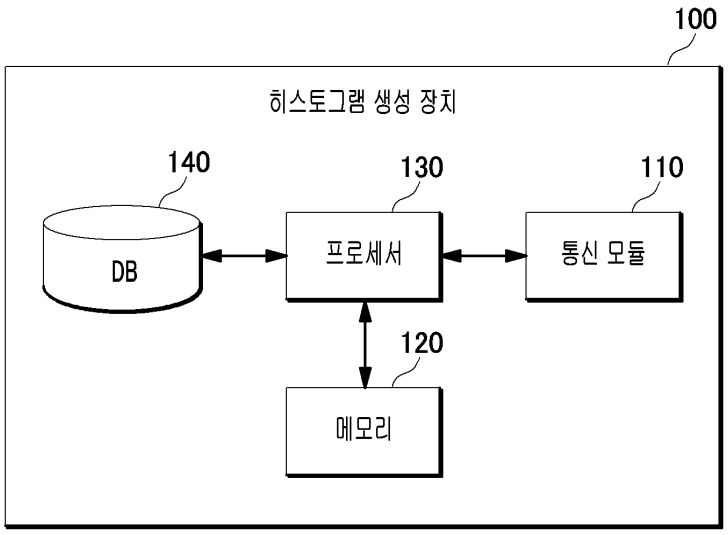
140: 데이터베이스

도면

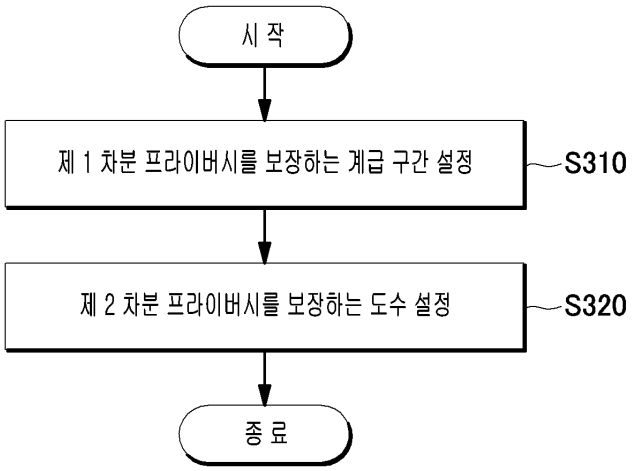
도면1



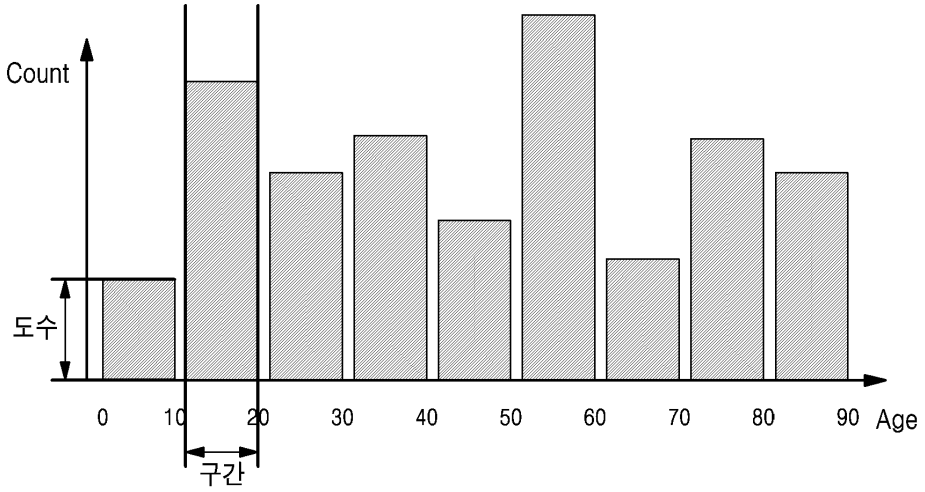
도면2



도면3



도면4



도면5

Algorithm 1 Privacy-preserving histogram construction

Input: data vector $x = (x_1, x_2, \dots, x_n)$, domain of data $D = [m, M]$, privacy budget ϵ, privacy cost ratio $r_1 : r_2$, heuristic algorithm $H \in \{Square-root, Sturge, Rice\}$.

Set privacy costs
$\epsilon_1 \leftarrow r_1 \cdot \epsilon$
$\epsilon_2 \leftarrow r_2 \cdot \epsilon$
Compute differentially private number of bins
$B, \tilde{d} \leftarrow DPBins(D, H, n, \epsilon_1)$
Compute differentially private frequencies
$\tilde{C} \leftarrow DPFrequencies(x, B, \tilde{d}, \epsilon_2)$
Output range of bins B, and perturbed frequencies $\tilde{C}$

도면6

Algorithm 2 DPBins

Input: the number of data n, domain of data $D = [m, M]$, privacy budget ϵ_1, heuristic algorithm $H \in \{Square-root, Sturge, Rice\}$.

Compute differentially private number of bins
$B \leftarrow H(n, \epsilon_1)$
Split bins
for $i = 1$ **to** $i = \tilde{d} - 1$ **do**
$B_i = [m + \frac{(i-1)(M-m)}{\tilde{d}}, m + \frac{i(M-m)}{\tilde{d}})$ /* the range of i-th bin */
$B_{\tilde{d}} = [M - \frac{(M-m)}{\tilde{d}}, M]$
Output range of bins B, the number of bins $\tilde{d}$

도면7

Heuristic Functions

Square-root(n,ϵ_1):
$noise \leftarrow Lap(0, \frac{1}{\epsilon_1})$ // sampling noise from Laplace distribution
return $\lceil \sqrt{n + noise} \rceil$

Sturge(n,ϵ_1):
$noise \leftarrow Lap(0, \frac{1}{\epsilon_1})$
return $\lceil \log_2(n + noise) \rceil + 1$

Rice(n,ϵ_1):
$noise \leftarrow Lap(0, \frac{1}{\epsilon_1})$
return $\lceil 2\sqrt[3]{n + noise} \rceil$

도면8

Algorithm 3 DPFrequencies

Input: data vector $x = (x_1, x_2, \dots, x_n)$, the number of bins $\tilde{d}$, range of bins B, privacy budget ϵ_2.

Compute differentially private frequencies
Initialize $C = [c_1, c_2, \dots, c_{\tilde{d}}] = [0, 0, \dots, 0]$
for $i = 1$ **to** $i = n$ **do**
  for $j = 1$ **to** $j = \tilde{d}$ **do**
    if $x_i \in B_j$ **then**
      $c_j \leftarrow c_j + 1$
      break
for $i = 1$ **to** $i = \tilde{d}$ **do**
  $\tilde{c_i} \leftarrow c_i + Lap(\frac{1}{\epsilon_2})$
Output perturbed frequencies $\tilde{C}$