(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0076304
(43) 공개일자 2019년07월02일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) *H04W 12/02* (2009.01)
(52) CPC특허분류
G06F 21/6245 (2013.01)
H04W 12/02 (2019.01)
(21) 출원번호 10-2017-0177999
(22) 출원일자 2017년12월22일
심사청구일자 2017년12월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정연돈
서울특별시 중구 청계천로 400, 105-812(황학동, 롯데캐슬베네치아)
김종선
서울특별시 성북구 삼선교로20길 52-4, B102
(74) 대리인
김등용, 김홍석

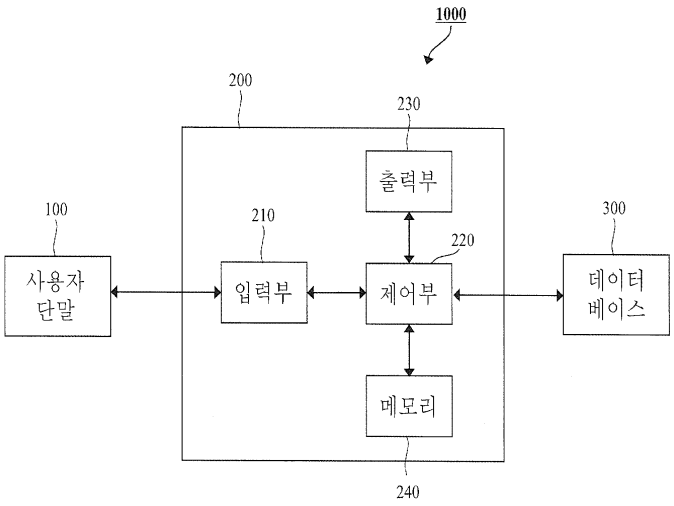
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **데이터베이스 기반 프라이빗 공간 분해 수행 방법 및 장치**

(57) 요 약

본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD: Private Spatial Decomposition)를 수행하는 방법은, 2차원 위치 데이터셋(two-dimensional location dataset)에 기반하여, 전체영역에서 객체들의 밀도가 높은 서브영역들인 핫스팟(hotspot)들을 검출하는 핫스팟 검출과정; 및 상기 검출된 핫스팟들에 기반하여, 상기 전체영역을 복수의 그리드로 분할(split)하는 그리드 파티셔닝을 통해 상기 핫스팟들의 경계를 검출하는 핫스팟 경계 검출 과정을 포함하여, 핫스팟에 기반하여 위치 기반 서비스를 제공하면서 개인 위치 정보를 보호할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 2221/2111 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 2015-0-00579-003
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터
연구사업명 SW컴퓨팅산업원천기술개발
연구과제명 빅데이터 환경에서 비식별화 기법을 이용한 개인정보보호 기술 개발
기 여 율 1/1
주관기관 고려대학교 산학협력단
연구기간 2017.06.01 ~ 2018.03.31

명 세 서

청구범위

청구항 1

데이터베이스 기반 프라이빗 공간 분해(PSD: Private Spatial Decomposition)를 수행하는 방법에 있어서,

2차원 위치 데이터셋(two-dimensional location dataset)에 기반하여, 전체영역에서 객체들의 밀도가 높은 서브영역들인 핫스팟(hotspot)들을 검출하는 핫스팟 검출과정; 및

상기 검출된 핫스팟들에 기반하여, 상기 전체영역을 복수의 그리드로 분할(split)하는 그리드 파티셔닝을 통해 상기 핫스팟들의 경계를 검출하는 핫스팟 경계 검출 과정을 포함하는, 프라이빗 공간 분해 수행 방법.

청구항 2

제1 항에 있어서,

상기 핫스팟 경계 검출 과정은,

상기 핫스팟에 존재하는 n개의 객체들 O = {o_1, o_2, …, o_n}의 리스트를 2차원 좌표상에서 정렬하는 객체 정렬 과정;

상기 정렬된 좌표의 인터벌에 스코어를 랭크(rank)하고, 상기 인터벌에 상기 랭크된 스코어에 비례하는 확률값을 도출하는 확률값 도출 과정; 및

상기 인터벌 중 특정 인터벌(I_k)을 선택하고, 상기 선택된 특정 인터벌(I_k)에 잡음 값을 부가하는 잡음 값 부가 과정을 포함하는, 프라이빗 공간 분해 수행 방법.

청구항 3

제2 항에 있어서,

상기 서브영역(R)과 상기 서브영역(R)을 둘러싸는 상기 전체영역(B)에 대하여, 상기 서브영역(R)의 상대 밀도(relative density)는 $RelativeDensity_B(R) = \frac{density(R)}{density(B)}$ 에 의해 결정되는 것을 특징으로 하는, 프라이빗 공간 분해 수행 방법.

청구항 4

제3 항에 있어서,

상기 전체영역을 상기 검출된 핫스팟을 이용하여 상기 객체들로 채운 후, 각각의 핫스팟은 균일한 상기 복수의 그리드로 분할되고,

상기 그리드의 크기(m)은 $m = \sqrt{\frac{N\varepsilon}{c}}$ 으로 결정되고,

N은 상기 핫스팟 내의 객체들의 개수이고, ε은 프라이버시 버짓(privacy budget)이고, c는 상수 값인 것을 특징으로 하는, 프라이빗 공간 분해 수행 방법.

청구항 5

제4 항에 있어서,

상기 복수의 그리드의 각각에 대하여, 상기 객체의 실제 카운트(actual count)에 라플라시안 잡음 $\chi \sim Lap\left(\frac{1}{\varepsilon^c}\right)$ 을 부가하는 것을 특징으로 하고,

ε^c는 각각의 그리드의 실제 카운트 값에 상기 라플라시안 잡음을 부가하는데 사용되는 파라미터와 연관된, 프라이빗 공간 분해 수행 방법.

청구항 6

제4 항에 있어서,

상기 핫스팟의 경계(boundary)들에 해당하는 좌측-하단(left-bottom) 및 우측-상단(top-right) 좌표들은 상기 각각의 핫스팟 내의 상기 객체들의 위치에 의해 결정되고,

상기 객체들의 특정 위치는 상기 핫스팟의 구조로부터 유추(infer)되는 것을 특징으로 하는, 프라이빗 공간 분해 수행 방법.

청구항 7

제5 항에 있어서,

상기 프라이버시 버짓은,

상기 잡음이 부가된 상기 핫스팟의 경계들을 추출하는데 사용되는 제1파라미터에 기반하여 추정된 제1프라이버시 버짓(ε^h); 및

상기 각각의 그리드의 상기 실제 카운트 값에 상기 라플라시안 잡음을 부가하는데 사용되는 제2파라미터에 기반하여 추정된 제2프라이버시 버짓(ε^c)을 모두 포함하는 총 프라이버시 버짓인 것을 특징으로 하는, 프라이빗 공간 분해 수행 방법.

청구항 8

제7 항에 있어서,

상기 핫스팟 경계 검출 과정을 수행하는 제어부의 입력은 상기 2차원 위치 데이터셋 및 상기 총 프라이버시 버짓(ε)이고, 상기 제어부의 출력은 상기 위치 데이터셋의 스큐-인지 그리트 파티셔닝(SAGA: Skew-aware Grid pArtitioning) 히스토그램인 것을 특징으로 하는, 프라이빗 공간 분해 수행 방법.

청구항 9

제8 항에 있어서,

상기 핫스팟 검출 과정 이전에,

사용자 단말로부터의 입력에 기반하여 질의 범위를 결정하고, 상기 결정된 질의 범위에 따라 상기 객체들의 개수를 추정하는 객체 추정 과정을 더 포함하는, 프라이빗 공간 분해 수행 방법.

청구항 10

제9 항에 있어서,

상기 핫스팟 검출과정에서 사용되는 상기 2차원 위치 데이터셋은,

상기 전체영역 중 상기 결정된 질의 범위에 대응하는 상기 서브영역들을 결정하여, 상기 서브영역별로 상기 2차원 위치 데이터셋의 분포가 결정되는 것을 특징으로 하는, 프라이빗 공간 분해 수행 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 데이터베이스 기반 프라이빗 공간 분해 수행 방법 및 장치에 관한 것이다. 보다 상세하게는, 차분 프라이빗 공간 스큐-인지 공간 분해 방법 및 장치에 관한 것이다.

배 경 기 술

[0003] 기술의 급격한 발전에 따라, 다양한 모바일 기기와의 상호 작용이 이루어지고 있다. 이러한 기기들은 GPS를 구비하고 있고, 기기를 소지한 개인들의 위치를 파악할 수 있다. 이와 관련하여, 위치 데이터는 도시 플래닝, 교통 모니터링, 및 위치-기반 서비스(LBS) 등 다양한 목적을 위하여 수집될 수 있다. 예를 들어, 서울 심야 버스 경로 최적화는 도시 플래닝의 일 예가 될 수 있다. 심야에 버스 경로를 최적으로 플래닝하기 위해, 각 지역에서의 버스 수요를 파악하기 위해 심야 모바일 폰 데이터를 활용할 필요가 있다. 이러한 예에서, 대부분의 위치-기반 분석은 "How many people are within the given regeion?"과 같은 질의를 요구할 수 있다. 그러나, 적절한 프라이버시 보호가 없다면, 프라이버시 침해를 유발할 수 있다는 문제점이 있다.

[0004] 차분 프라이버시(DP: Differential Privacy)는 프라이버시 보호 데이터 분석을 위한 사실상의 표준이다. DP는 분석 결과를 랜덤화하여 프라이버시 보호를 제공한다. 하지만, k-익명성(anoynmity), l-다양성(diversity) 및 t-근접성(closeness)와 같은 기존의 익명성 기반 프라이버시 모델은 k, l, t 값들보다 큰 백그라운드 지식을 갖는 경우에 있어 프라이버시를 보호할 수 없다는 문제점이 있다.

발명의 내용

해결하려는 과제

[0006] 따라서, 본 발명에서 해결하고자 하는 과제는, 프라이빗 공간 분해(PSD)를 통해 프라이버시 보호가 가능한 위치 데이터 셋을 표시하는 방법을 제공하는 것을 목적으로 한다.

[0007] 또한, 본 발명에서 해결하고자 하는 과제는, 스큐-인지 그리드 파티셔닝 (SAGA)을 통해 프라이버시 보호가 가능한 위치 데이터 셋을 표시하는 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0009] 상기와 같은 과제를 해결하기 위한 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD: Private Spatial Decomposition)를 수행하는 방법은, 2차원 위치 데이터셋(two-dimensional location dataset)에 기반하여, 전체영역에서 객체들의 밀도가 높은 서브영역들인 핫스팟(hotspot)들을 검출하는 핫스팟 검출과정; 및 상기 검출된 핫스팟들에 기반하여, 상기 전체영역을 복수의 그리드로 분할(split)하는 그리드 파티셔닝을 통해 상기 핫스팟들의 경계를 검출하는 핫스팟 경계 검출 과정을 포함하여, 핫스팟에 기반하여 위치 기반 서비스를 제공하면서 개인 위치 정보를 보호할 수 있다. 또한, 핫스팟 경계 검출 과정 이후에, 상기 검출된 핫스팟 및 상기 분할된 그리드에 기반하여, 상기 데이터셋을 2차원 위치 공간상에 표시하는 데이터 표시 과정이 더 수행될 수 있다.

[0010] 일 실시 예에서, 상기 핫스팟 경계 검출 과정은, 상기 핫스팟에 존재하는 n개의 객체들 $O = \{o_1, o_2, \ldots, o_n\}$의 리스트를 2차원 좌표상에서 정렬하는 객체 정렬 과정; 상기 정렬된 좌표의 인터벌에 스코어를 랭크(rank)하고, 상기 인터벌에 상기 랭크된 스코어에 비례하는 확률값을 도출하는 확률값 도출 과정; 및 상기 인터벌 중 특정 인터벌(I_k)을 선택하고, 상기 선택된 특정 인터벌(I_k)에 잡음 값을 부가하는 잡음 값 부가 과정을 포함할 수 있

다.

[0011] 일 실시 예에서, 상기 서브영역(R)과 상기 서브영역(R)을 둘러싸는 상기 전체영역(B)에 대하여, 상기 서브영역(R)의 상대 밀도(relative density)는 $RelativeDensity_B(R) = \frac{density(R)}{density(B)}$ 에 의해 결정될 수 있다.

[0012] 일 실시 예에서, 상기 전체영역을 상기 검출된 핫스팟을 이용하여 상기 객체들로 채운 후, 각각의 핫스팟은 균일한 상기 복수의 그리드로 분할되고, 상기 그리드의 크기(m)은 $m = \sqrt{\frac{N\varepsilon}{c}}$ 으로 결정될 수 있다. 이때, N은 상기 핫스팟 내의 객체들의 개수이고, ε은 프라이버시 버짓(privacy budget)이고, c는 상수 값이다.

[0013] 일 실시 예에서, 상기 복수의 그리드의 각각에 대하여, 상기 객체의 실제 카운트(actual count)에 라플라시안 잡음 $X \sim Lap\left(\frac{1}{\varepsilon^c}\right)$ 을 부가하는 것을 특징으로 할 수 있다. 이때, ε^c는 각각의 그리드의 실제 카운트 값에 상기 라플라시안 잡음을 부가하는데 사용되는 파라미터와 연관될 수 있다.

[0014] 일 실시 예에서, 상기 핫스팟의 경계(boundary)들에 해당하는 좌측-하단(left-bottom) 및 우측-상단(top-right) 좌표들은 상기 각각의 핫스팟 내의 상기 객체들의 위치에 의해 결정되고, 상기 객체들의 특정 위치는 상기 핫스팟의 구조로부터 유추(infer)되는 것을 특징으로 할 수 있다.

[0015] 일 실시 예에서, 상기 프라이버시 버짓은, 상기 잡음이 부가된 상기 핫스팟의 경계들을 추출하는데 사용되는 제1파라미터에 기반하여 추정된 제1프라이버시 버짓(ε^h); 및 상기 각각의 그리드의 상기 실제 카운트 값에 상기 라플라시안 잡음을 부가하는데 사용되는 제2파라미터에 기반하여 추정된 제2프라이버시 버짓(ε^c)을 모두 포함하는 총 프라이버시 버짓일 수 있다.

[0016] 일 실시 예에서, 상기 핫스팟 경계 검출 과정을 수행하는 제어부의 입력은 상기 2차원 위치 데이터셋 및 상기 총 프라이버시 버짓(ε)이고, 상기 제어부의 출력은 상기 위치 데이터셋의 스큐-인지 그리트 파티셔닝(SAGA: Skew-aware Grid pArtitioning) 히스토그램일 수 있다.

[0017] 일 실시 예에서, 상기 핫스팟 검출 과정 이전에, 사용자 단말로부터의 입력에 기반하여 질의 범위를 결정하고, 상기 결정된 질의 범위에 따라 상기 객체들의 개수를 추정하는 객체 추정 과정을 더 포함할 수 있다.

[0018] 일 실시 예에서, 상기 핫스팟 검출과정에서 사용되는 상기 2차원 위치 데이터셋은, 상기 전체영역 중 상기 결정된 질의 범위에 대응하는 상기 서브영역들을 결정하여, 상기 서브영역별로 상기 2차원 위치 데이터셋의 분포가 결정되는 것을 특징으로 할 수 있다.

발명의 효과

[0020] 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해 수행 방법 및 장치를 제공하여 위치 기반 서비스에서 개인의 위치 정보를 보호할 수 있다는 장점이 있다.

[0021] 또한, 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해 수행 방법 및 장치는, 개인의 특정 위치에 랜덤화된 잡음을 부가하여 개인의 위치 정보를 보호하면서, 위치 기반 서비스를 제공할 수 있다는 장점이 있다.

도면의 간단한 설명

[0023] 도 1은 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD: Private Spatial Decomposition)를 수행하는 장치를 포함하는 PSD 시스템의 구성도를 나타낸다.

도 2는 본 발명에 따른 데이터베이스 기반 차분 프라이버시의 개념도를 나타낸다.

도 3은 본 발명에 따른 분석 결과에 랜덤 잡음을 부가하는 변조 방법의 개념도를 나타낸다.

도 4는 본 발명에 따른 가능한 출력 셋으로부터 분석 결과를 랜덤하게 도출하는 변조 방법의 개념도를 나타낸다.

도 5는 본 발명에 따른 공간 인덱스 형태의 프라이빗 공간 분해(PSD) 수행 방법의 개념도를 도시한다.

도 6은 본 발명에 따른 서브영역(R)과 상기 서브영역(R)을 둘러싸는 전체영역(B)으로 이루어진 구조에서 상대 밀도를 구하는 개념도를 나타낸다.

도 7a 및 도 7b는 본 발명에 따른 차분 프라이빗 핫스팟 검출 알고리즘에 따른 동작을 나타낸다.

도 8은 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD) 방법의 흐름도를 나타낸다.

도 9는 본 발명에 따른 핫스팟 경계 생성 알고리즘에 따른 동작을 나타낸다.

도 10은 본 발명에 따른 차분 프라이빗 SAGA 히스토그램 구축의 알고리즘을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0024] 상술한 본 발명의 특징 및 효과는 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해 질 것이며, 그에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다. 본 발명은 다양한 변경을 가할 수 있고 여러 가지 형태를 가질 수 있는바, 특정 실시 예들을 도면에 예시하고 본문에 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 개시형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 본 명세서에서 사용한 용어는 단지 특정한 실시 예들을 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다.

[0025] 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용한다.

[0026] 제1, 제2등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0027] 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.

[0028] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미가 있다.

[0029] 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않아야 한다.

[0030] 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈", "블록" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.

[0031] 이하, 본 발명의 바람직한 실시 예를 첨부한 도면을 참조하여 당해 분야에 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 설명한다. 하기에서 본 발명의 실시 예를 설명함에 있어, 관련된 공지의 기능 또는 공지의 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다.

[0032] 이하, 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD: Private Spatial Decomposition)를 수행하는 방법 및 장치에 대해 살펴보면 다음과 같다.

[0033] 이와 관련하여, 도 1은 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD: Private Spatial Decomposition)를 수행하는 장치를 포함하는 PSD 시스템 의 구성도를 나타낸다. 이러한, PSD 수행 장치와 관련하여, 데이터베이스를 기반으로 하는 서버 또는 단말 모두 PSD 수행 장치가 될 수 있다.

[0034] 도 1을 참조하면, PSD 시스템(1000)은 사용자 단말(100), PSD 수행 장치(200), 및 데이터 베이스(300)를 포함한다. 한편, PSD 수행 장치(200)는 입력부(210), 제어부(220), 및 출력부(230), 및 메모리(240)를 포함한다. 입력부(210)는 하나 이상의 사용자 단말(100)로부터 질의를 수신하도록 구성된다. 제어부(200)는 상기 질의와 연관된 데이터셋(dataset)에 기반하여 처리된 결과가 출력부(230)를 통해 출력되도록 제어한다. 한편, 상기 데이터셋은 상기 질의에 따라 외부의 데이터베이스(300)로 요청된 사항에 대하여 응답에 기반하여 생성될 수 있다.

또는, PSD 수행 장치(200)가 내부에 데이터베이스를 구비한 경우에는 외부의 데이터베이스(300)로 별도로 요청하지 않고, 내부적으로 데이터셋을 생성할 수 있다.

[0035] 한편, 프라이버시 보호와 함께 위치 서비스를 제공하는 경우에, 사용자 단말(100)은 위치 데이터와 관련 정보를 제공할 수 있다. 입력부(210)는 하나 이상의 사용자 단말(100)로부터 위치 데이터와 관련 정보를 수신하도록 구성된다. 또한, 입력부(210)는 이러한 위치 데이터를 요구하는 서비스 제공자 단말 (미도시)로부터 위치 데이터 요청에 관한 질의를 수신할 수 있다. 이때, 제어부(200)는 상기 질의와 연관된 데이터셋(dataset)에 기반하여 처리된 결과가 서비스 제공자 단말의 출력부(미도시)를 통해 출력되도록 제어한다. 이때, 상기 처리된 결과는 위치 서비스를 제공하는 데 필요한 위치 데이터 정보이지만, 개인들의 정확한 위치 정보를 알 수 없는 형태이어야 한다.

[0036] 이러한 처리 결과와 관련하여, 그 형태는 히스토그램 형태일 수 있고, 본 발명에서는 이를 스큐-인지 그리드 파티셔닝(SAGA: Skew-aware Grid pArtitioning) 히스토그램으로 지칭할 수 있다. 즉, 제어부(220)는 상기 질의와 연관된 데이터셋을 이용하여 상기 출력부(230)의 공간상에 출력 결과가 위치 기반으로 표시되도록 제어할 수 있다. 메모리(240)는 데이터셋과 연관된 데이터 및 데이터셋으로부터 처리된 결과와 관련된 정보를 저장하도록 구성된다.

[0037] 따라서, 본 발명의 목표(objective)는 전술된 SAGA에 의해 PSD를 수행하는 것이며, PSD는 차분 프라이빗 공간 히스토그램 퍼블리싱(publishing) 방법이다. 즉, 차분 프라이버시(differential privacy)는 프라이버시 보호를 위한 평가기준(criteria)가 된다. 한편, 스큐 데이터 분포를 고려하면, 기존의 방법은 히스토그램을 구축 시에 데이터 분포에 대하여 거의 고려하지 않았다. 따라서, 기존 방법들은 데이터셋이 스큐 분포를 갖는 경우에 정확하지 않은 히스토그램을 제공한다는 문제점이 있다.

[0038] 한편, 본 발명과 관련하여, 차분 프라이버시에 대해 살펴보기로 하자. 차분 프라이버시는 프라이버시-보존 데이터 분석을 위한 프라이버시 모델이다. 차분 프라이버시는 분석 결과들을 랜덤화하여 프라이버시 보호를 제공한다. 차분 프라이버시와 관련하여, 임의의 메커니즘$\mathcal{A}$ 은 아래의 수학식 1을 만족하고, 이를 ε-차분 프라이버시라고 지칭할 수 있다.

수학식 1

[0039] $$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq \exp(\varepsilon) \times \Pr[\mathcal{A}(\mathcal{D}') \in S], S \in Range(\mathcal{A})$$

[0040] 여기서, $\mathcal{D}$와 $\mathcal{D}'$는 하나의 레코드만 다른 임의의 인접한 데이터베이스들이다. 이와 관련하여, 도 2는 본 발명에 따른 데이터베이스 기반 차분 프라이버시의 개념도를 나타낸다. 도 2를 참조하면, $\mathcal{D}$와 $\mathcal{D}'$는 하나의 레코드만 다른 임의의 인접한 데이터베이스들에 해당한다. 예를 들어, 데이터베이스$\mathcal{D}$는 Alice, Bob, …, Frank에 대해 Age, …, Disease를 레코드로 갖는다. 반면에, $\mathcal{D}'$는 Alice, Bob, …, Gary에 대해 Age, …, Disease를 레코드로 갖는다. 한편, 입력부(210)가 사용자 단말(100)로부터 질의(query)를 수신하고, 제어부(220)는 질의에 대한 결과(query result)를 획득한다. 예를 들어, "How many person have AIDS?"라는 질의에 대해, 질의 결과에 대한 변조(perturbation)가 발생할 수 있고, 이에 따라, 잡음성 질의 응답(noisy query result)이 생성될 수 있다.

[0041] 한편, 변조 방법과 관련하여 분석 결과에 잡음이 부가되는 메커니즘에 대해 살펴보면 다음과 같다. 도 3은 본 발명에 따른 분석 결과에 랜덤 잡음을 부가하는 변조 방법의 개념도를 나타낸다. 이와 같은 랜덤 잡음을 랜덤 라플라스 잡음(random Laplace noise)로 지칭할 수 있고, 변조 방법(perturbing method)은 분석 결과에 랜덤 라플라스 잡음을 부가하는 방법이다.

[0042] 구체적으로, 1) 단계에서 질의의 민감도(sensitivity) ΔF 를 연산한다. 이때, 민감도는 임의의 레코드가 질의 결과에 최대로 얼마나 많이 영향을 미칠 수 있는지를 의미한다. 또한, 2) 단계에서 평균(mean)=0 및 스케일 ΔF/ε을 갖는 라플라스 분포로부터 샘플링된 랜덤 잡음 χ를 부가한다. 한편, "What is the average net worth of database D"라는 질의에 대해 데이터베이스 $\mathcal{D}$에 기반하여, 응답(answer)을 구할 수 있다. 이때, Bill

Gates, Warren Buffett, …, David Koch의 Net worth의 산술 평균과 질의의 민감도에 따른 랜덤 잡음 χ의 합이 응답(answer)이 될 수 있다. 여기서, 랜덤 잡음 χ는 $\Delta F / \varepsilon$을 갖는 라플라스 분포일 수 있다.

[0043] 다음으로, 본 발명에 따른 지수 메커니즘(Exponential mechanism)에 대해 살펴보면 아래와 같다. 이와 관련하여, 도 4는 본 발명에 따른 가능한 출력 셋으로부터 분석 결과를 랜덤하게 도출하는 변조 방법의 개념도를 나타낸다. 즉, 지수 메커니즘은 1) 단계에서 우선적으로 각 레코드에 스코어를 부여한다. 다음으로, 주어진 스코어에 비례하는 확률을 갖는 출력을 랜덤하게 도출한다. 즉, 스코어가 높을 수록 결과는 더욱 매력적일 수 있다. 도 4를 참조하면, "Who is the richest person in database D"라는 질의에 대해 데이터베이스 $\mathcal{D}$에 기반하여, 질의에 대한 확률(probalitiy)을 Name 레코드의 인물별로 구할 수 있다. 따라서, 높은 확률을 갖는 레코드일수록 질의에 대한 정답에 해당할 가능성이 높다.

[0044] 다음으로, 본 발명에 따른 PSD와 관련하여, 2차원 위치 데이터셋(two-dimensional location dataset) 관점에서 살펴보기로 하자. 도 5는 본 발명에 따른 공간 인덱스 형태의 프라이빗 공간 분해(PSD) 수행 방법의 개념도를 도시한다. 이와 관련하여, PSD는 Kd-트리, 그리드, 쿼드트리 등과 같은 공간 인덱스 형태를 갖는다. 한편, 각 셀은 객체의 잡음 수(noisy number)와 서브-영역(sub-region)의 경계를 포함한다. 예를 들어, 도 5에 도시된 바와 같이, 서울에 있는 사람들의 예시적인 분포(Example distribution of people in Seoul)에 따라 2차원 위치 데이터셋이 생성될 수 있다. 한편, 이러한 2차원 위치 데이터셋을 Kd-트리, 그리드, 쿼드트리 등과 같은 공간 인덱스 형태로 분류할 수 있다. 예를 들어, 2차원 위치 데이터셋이 m×n 그리드 PSD에 대응될 수 있다. 예컨대, 2차원 위치 데이터셋이 3×3 그리드 PSD에 대응될 수 있고, 각 셀에서 객체의 수는 실제 카운트(actual count) 및/또는 잡음성 카운트(noisy count)를 포함할 수 있다. 예를 들어, 특정 셀에 '2(1)'로 표시된 경우, 실제 카운트(actual count) 및/또는 잡음성 카운트(noisy count)는 각각 2와 1에 해당한다.

[0045] 도 5에서의 예시에 따라, PSD에 대해 상세하게 살펴보면 다음과 같다. 임의의 범위의 질의들에서 객체들의 수를 예측하기 위해 PSD가 사용될 수 있다. 예를 들어, 도 5에서 인구에 대응하는 객체, 예컨대, 점선 사각형(dotted rectangle)은 세 개의 상부(top) 및 중간(middle) 셀들을 포함하는 질의 범위를 나타낸다. 이후, 예측되는 카운트는 잡음성 카운트에 따라 (2 + 6 + 1 + 3 + 4 + 3) = 19일 수 있다. 한편, 실제 카운트는 16일 수 있다.

[0046] 본 발명에 따른 PSD에 있어서의 목표는 잡음성 카운트를 정확히 반영하는 히스토그램을 구축하는 것이다. 한편, 본 발명에 따른 스큐-인지 그리드 파티셔닝(SAGA: Skew-aware Grid pArtitioning)에 대해 입력/출력 등 문제 정의(problem definition)에 대해 살펴보면 다음과 같다. 즉, 입력은 2차원 위치 데이터셋과 총 프라이버시 버짓(ε)일 수 있고, 출력은 위치 데이터셋의 SAGA 히스토그램일 수 있다. 즉, 도 1을 참조하면, 핫스팟 경계 검출을 수행하는 제어부(220)의 입력은 2차원 위치 데이터셋 및 총 프라이버시 버짓(ε)이고, 제어부(220)의 출력은 위치 데이터셋의 스큐-인지 그리트 파티셔닝(SAGA: Skew-aware Grid pArtitioning) 히스토그램이다.

[0047] 한편, 스큐-인지 그리드 파티셔닝(SAGA: Skew-aware Grid pArtitioning)에 대해 단계적으로 상세하게 살펴보면 다음과 같다. SAGA 히스토그램은 2 단계(two phases)로 이루어진다. 도 5를 참조하면, 1) 단계에서, 고밀도를 갖는 서브-영역들인 핫스팟을 검출한다. 다음으로, 2) 단계에서, 핫스팟은 그리드, 예컨대 균일(uniform) 그리드로 더 분할될 수 있다. 이와 관련하여, 순차적 합성(sequential composition)에 따라, 총 프라이버시 버짓(ε)은 전술된 1) 단계 및 2) 단계에서의 버짓의 합인, $\varepsilon = \varepsilon^h + \varepsilon^c$로 결정될 수 있다. 즉, 1) 단계에서 ε^h는 잡음성 핫스팟 경계들을 도출하는데 사용된다. 또한, 2) 단계에서 ε^c는 각 그리드 셀의 실제 카운트 값에 라플라스 잡음을 부가하는데 사용된다.

[0048] 이상에서 전술된 동작 및 핫스팟 영역과 경계 생성과 관련하여 구체적인 내용에 대해 살펴보면 다음과 같다. 먼저, 핫스팟은 특정 영역을 의미하고, 상기 특정 영역에서 객체들은 높은 상대 밀도를 가지고 함께 근접하게 배치된다. 이와 관련하여, 서브영역(R)과 서브영역(R)을 둘러싸는 전체영역(B)에 대하여, 서브영역(R)의 상대 밀도(relative density)는 $RelativeDensity_B(R) = \frac{density(R)}{density(B)}$ 에 의해 결정될 수 있다. 이와 관련하여, 도 6은 본 발명에 따른 서브영역(R)과 상기 서브영역(R)을 둘러싸는 전체영역(B)으로 이루어진 구조에서 상대 밀도를 구하는 개념도를 나타낸다. 도 6을 참조하면, 전체영역(B)에 하나 이상의 서브영역(R_1 및 R_2)이 존재함을 알 수 있다. 예를 들어, 하나 이상의 서브영역(R_1 및 R_2)의 크기가 각각 B/32, 및 B/16라고 하고, 전체영역(B) 내의 객체의 총 개수가 20개라고 가정하자. 이때, R_1의 상대밀도와 R_2의 상대밀도는 각각 7.2와 8로 계산된다. 즉, 제1 서브영역(R_1)내의 객체 수가 제2 서브영역(R_2) 내의 객체 수보다 더 적지만, 서브 영역들의 면적을 고려하면 제1 서

브영역(R_1)의 상대밀도가 더 높음을 알 수 있다.

[0049] 다음으로, 차분 프라이빗 핫스팟 검출 알고리즘에 대해 상세하게 살펴보면 다음과 같다. 이와 관련하여, 도 7a 및 도 7b는 본 발명에 따른 차분 프라이빗 핫스팟 검출 알고리즘에 따른 동작을 나타낸다. 도 7a 및 도 7b를 참조하면, 차분 프라이빗 핫스팟 검출과 관련하여, 입력은 프라이빗 버짓(ε^h), 공간 도메인(D), 로컬 지점들의 세트(O), 사이즈 파라미터(s), 및 주파수 파라미터(f)이다. 이러한 입력에 따른 출력은 핫스팟들의 세트(H)이다. 이와 관련하여, 도 1을 참조하면, 제어부(220)는 프라이빗 버짓(ε^h), 공간 도메인(D), 로컬 지점들의 세트(O), 사이즈 파라미터(s), 및 주파수 파라미터(f)에 기반하여, 데이터셋 내의 객체 위치 기반 핫스팟을 검출한다. 이러한 핫스팟 검출과 관련하여, 핫스팟 경계 생성이 무엇보다 중요하다.

[0050] 한편, 핫스팟 경계 생성과 관련하여, 핫스팟 경계들, 예컨대 핫스팟의 좌측-하단(left-bottom)과 상부-우측(top-right)은 각 핫스팟 내의 객체들의 위치에 기반하여 결정된다. 한편, 핫스팟들의 구조로부터 객체의 특정 위치를 유추(infer)할 수 있다. 따라서, 핫스팟 경계들은 차분 프라이빗 방법으로 구조화(formulate)될 수 있다.

[0051] 한편, 도 1을 참조하고, 전술된 내용에 기반하여, 데이터베이스 기반 프라이빗 공간 분해(PSD) 방법에 대해 살펴보면 다음과 같다. 이와 관련하여, 도 8은 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD) 방법의 흐름도를 나타낸다. 이와 관련하여, 도 1을 참조하면, 제어부(220)는 데이터베이스 기반 PSD 방법에 따라 데이터셋에 대한 핫스팟 검출 및 히스토그램 구축을 수행할 수 있다. 즉, 제어부(220)는 객체 추정 과정(S100), 핫스팟 검출과정(S200), 및 핫스팟 경계 검출 과정(S300)을 수행할 수 있다.

[0052] 객체 추정 과정(S100)에서, 사용자 단말(100)로부터의 입력에 기반하여 질의 범위를 결정하고, 상기 결정된 질의 범위에 따라 상기 객체들의 개수를 추정할 수 있다.

[0053] 또한, 핫스팟 검출과정(S200)에서, 2차원 위치 데이터셋(two-dimensional location dataset)에 기반하여, 전체영역에서 객체들의 밀도가 높은 서브영역들인 핫스팟(hotspot)들을 검출하는 핫스팟 검출과정(S200)을 수행한다. 또한, 상기 검출된 핫스팟들에 기반하여, 상기 전체영역을 복수의 그리드로 분할(split)하는 그리드 파티셔닝을 통해 상기 핫스팟들의 경계를 검출하는 핫스팟 경계 검출 과정(S300)을 수행할 수 있다.

[0054] 한편, 제어부(220)는 상기 검출된 핫스팟 및 상기 분할된 그리드에 기반하여, 상기 데이터셋을 2차원 위치 공간상에 표시하는 데이터 표시 과정(S400)을 더 수행할 수 있다.

[0055] 한편, 핫스팟 검출과정(S200)에서 사용되는 2차원 위치 데이터셋은, 전체영역 중 결정된 질의 범위에 대응하는 서브영역들을 결정하여, 상기 서브영역별로 상기 2차원 위치 데이터셋의 분포가 결정되는 것을 특징으로 할 수 있다.

[0056] 한편, 핫스팟 경계 검출 과정(S300)은 객체 정렬 과정(S310), 확률값 도출 과정(S320), 및 잡음 값 부가 과정(S330)을 포함할 수 있다. 이때, S310 내지 S330 과정은 응용에 따라 일부 단계만 수행되거나, 단계의 순서를 변경하여 수행될 수 있다. 객체 정렬 과정(S310)에서, 핫스팟에 존재하는 n개의 객체들 $O = \{o_1, o_2, \cdots, o_n\}$의 리스트를 2차원 좌표상에서 정렬한다. 예를 들어, 객체들의 리스트를 아래의 수학식 2와 같은 별개의 좌표 리스트들로 나눌 수 있다. 이때, x_{list} 및 y_{list} 의 좌표들은 경도(longitude) 및 위도(latitude)에 해당할 수 있다. 한편, 이러한 리스트는 올림 차순(ascending order)으로 정렬될 수 있다.

수학식 2

[0057] $$O = \{o_1, o_2, \cdots, o_n\} \rightarrow X_{list} = \{x_1, x_2, \cdots, x_n\} \,\&\, Y_{list} = \{y_1, y_2, \cdots, y_n\}$$

[0058] 확률값 도출 과정(S320)에서, 정렬된 좌표의 인터벌에 스코어를 랭크(rank)하고, 상기 인터벌에 상기 랭크된 스코어에 비례하는 확률값을 도출한다. 구체적으로, 좌측 경계를 결정하는 경우, 각 좌표 인터벌에 아래의 수학식 3과 같이 스코어를 랭크 (부여)할 수 있다.

수학식 3

[0059] $$rank(I_0 = [x_0, x_1)) = 0, rank(I_1 = [x_1, x_2)) = 1, \cdots, rank(I_{n-1} = [x_{n-1}, x_n)) = n - 1$$

[0060] 본 발명의 일 실시 예에 따라 지수 메커니즘을 적용하는 경우, 랭크 스코어에 비례하는 확률로 인터벌을 도출할 수 있다. 예를 들어, 오리지널 핫스팟 경계에 근접할 수록, 선택은 더 매력적일 수 있다.

[0061] 또한, 잡음 값 부가 과정(S330)에서, 인터벌 중 특정 인터벌(I_k)을 선택하고, 상기 선택된 특정 인터벌(I_k)에 잡음 값을 부가할 수 있다. 이때, 특정 인터벌(I_k)이 선택된 이후에, 잡음성 좌표 값(noisy coordinate value)이 특정 인터벌(I_k)로부터 랜덤하면서 균일하게 선택될 수 있다.

[0062] 한편, 전술된 핫스팟 경계 생성 알고리즘에 대해 살펴보면 다음과 같다. 이와 관련하여, 도 9는 본 발명에 따른 핫스팟 경계 생성 알고리즘에 따른 동작을 나타낸다. 도 9를 참조하면, 정렬된 좌표 리스트 V, 프라이버시 버짓(ε^h), 및 방향(direction, d)를 입력으로 하여, 잡음성 좌표 값(c)을 출력한다. 도 9에 도시된 바와 같이, 좌표 인터벌에 따라 랭크를 부여하고, 아래의 수학식 4에 비례하는 확률을 갖는 특정 인터벌(Ik)를 선택한다. 이후, 선택된 특정 인터벌(Ik)로부터 랜덤 값인 잡음성 좌표 값(c)를 선택한다.

수학식 4

[0063] $$\text{probability} \propto |I_k| \times exp(\varepsilon^{\tilde{h}} \times rank(I_k)/2)$$

[0064] 한편, 전술한 바와 같이, 핫스팟 검출과정(S200)에서 검출된 핫스팟으로 전체영역을 채운 후, 각각의 핫스팟은 균일 그리드로 분할될 수 있다. 이때, 그리드의 크기(m)는 아래의 수학식 5로 결정될 수 있다.

수학식 5

[0065] $$m = \sqrt{\frac{N\varepsilon}{c}}$$

[0066] 여기서, N은 상기 핫스팟 내의 객체들의 개수이고, ε는 프라이버시 버짓이고, c는 상수 값이다. 특히, 프라이버시 버짓 ε은 잡음이 부가된 상기 핫스팟의 경계들을 추출하는데 사용되는 제1파라미터에 기반하여 추정된 제1프라이버시 버짓(ε^h)일 수 있다. 한편, 각 그리드 셀에 대해, 라플라시안 잡음 $\chi \sim Lap\left(\frac{1}{\varepsilon^c}\right)$이 실제 카운트에 부가될 수 있다.

[0067] 한편, 응용(application)에 따라 상기 프라이버시 버짓 ε은 상기 잡음이 부가된 상기 핫스팟의 경계들을 추출하는데 사용되는 제1파라미터에 기반하여 추정된 제1프라이버시 버짓(ε^h), 및 상기 각각의 그리드의 상기 실제 카운트 값에 상기 라플라시안 잡음을 부가하는데 사용되는 제2파라미터에 기반하여 추정된 제2프라이버시 버짓(ε^c)을 모두 포함하는 총 프라이버시 버짓을 모두 포함할 수 있다.

[0068] 다음으로, 도 10은 본 발명에 따른 차분 프라이빗 SAGA 히스토그램 구축의 알고리즘을 도시한다. 도 10을 참조하면, 핫스팟의 세트(H), 및 프라이버시 버짓(ε^c)를 입력으로, m×m 균일 그리드 H_{ug}를 출력할 수 있다.

[0069] 이상에서는 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해(PSD) 수행 장치 및 방법에 대해 살펴보았다. PSD 수행 장치와 PSD 수행 방법에 대한 내용은 상호 결합되어 이용될 수 있음은 물론이다.

[0070] 한편, 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해 수행 방법 및 장치의 적어도 일 실시 예에

따르면, 위치 기반 서비스에서 개인의 위치 정보를 보호할 수 있다는 장점이 있다.

[0071] 또한, 본 발명에 따른 데이터베이스 기반 프라이빗 공간 분해 수행 방법 및 장치의 적어도 일 실시 예에 따르면, 개인의 특정 위치에 랜덤화된 잡음을 부가하여 개인의 위치 정보를 보호하면서, 위치 기반 서비스를 제공할 수 있다는 장점이 있다.

[0072] 소프트웨어적인 구현에 의하면, 본 명세서에서 설명되는 절차 및 기능뿐만 아니라 각각의 구성 요소들은 별도의 소프트웨어 모듈로도 구현될 수 있다. 상기 소프트웨어 모듈들 각각은 본 명세서에서 설명되는 하나 이상의 기능 및 작동을 수행할 수 있다. 적절한 프로그램 언어로 쓰여진 소프트웨어 어플리케이션으로 소프트웨어 코드가 구현될 수 있다. 상기 소프트웨어 코드는 메모리에 저장되고, 제어부(controller) 또는 프로세서(processor)에 의해 실행될 수 있다.

부호의 설명

[0074] 1000: PSD 시스템

100: 사용자 단말

200: PSD 수행 장치

210: 입력부

220: 제어부,

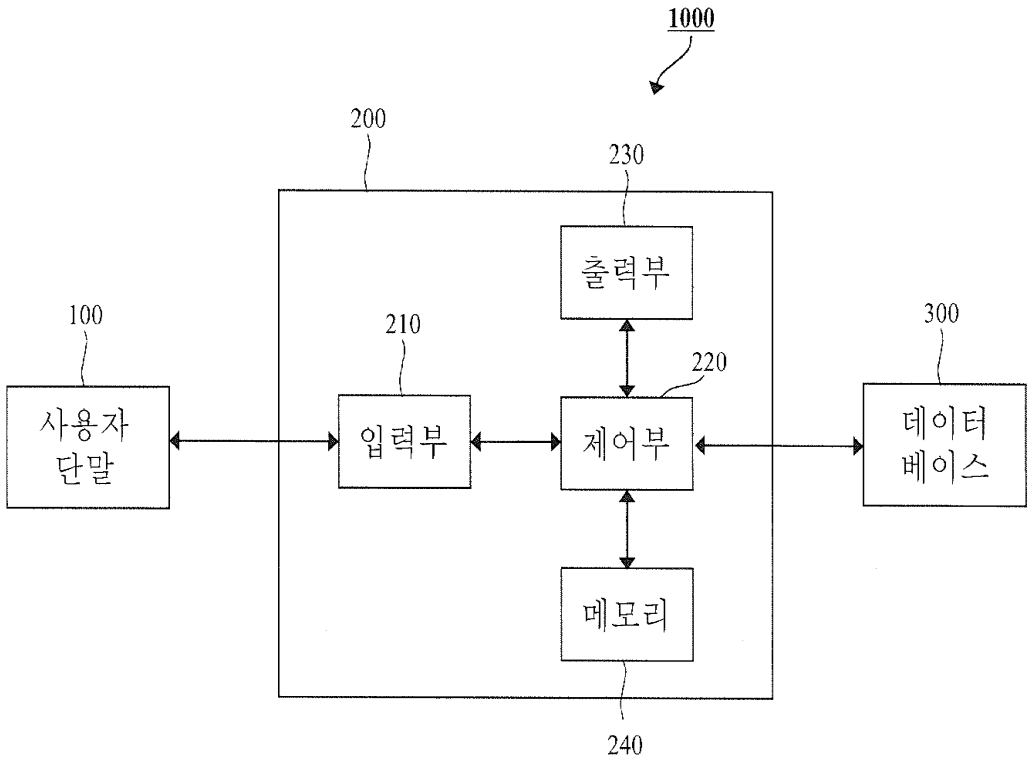
230: 출력부

240: 메모리

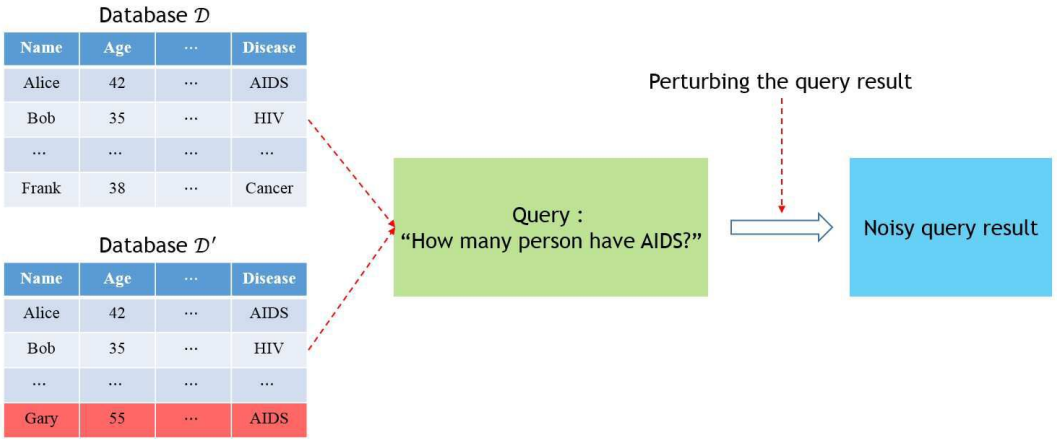
300: 데이터 베이스

도면

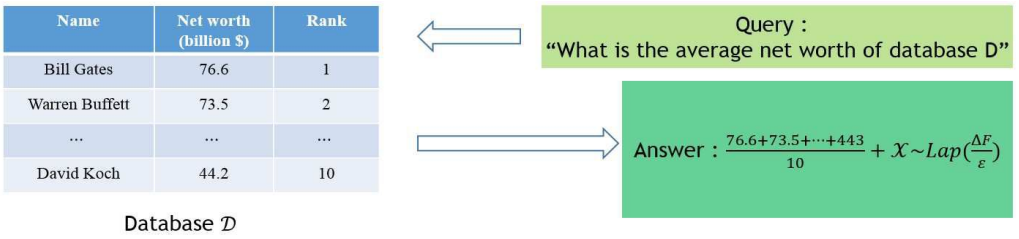
도면1



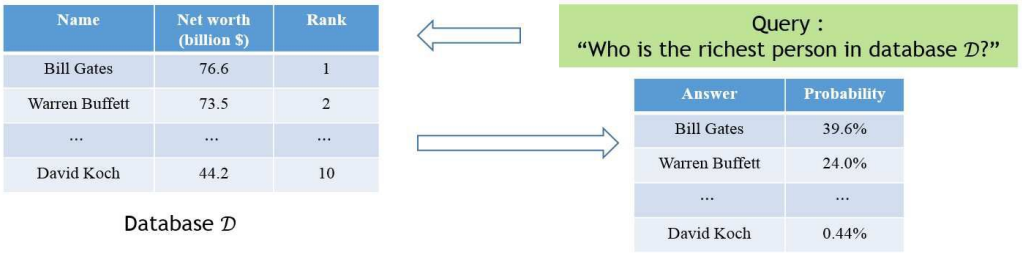
도면2



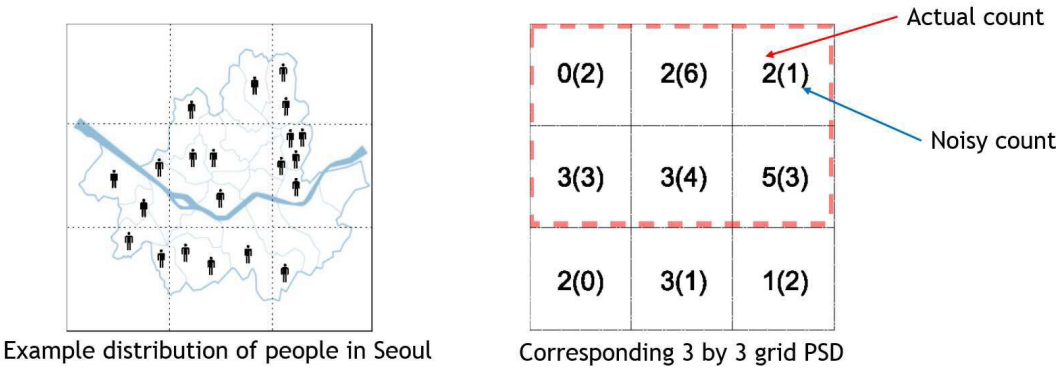
도면3



도면4



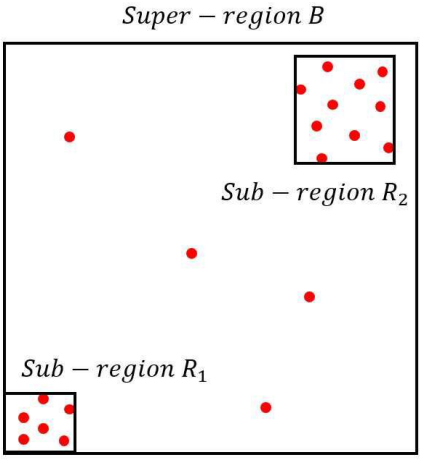
도면5



Example distribution of people in Seoul

Corresponding 3 by 3 grid PSD

도면6



도면7a

Algorithm 1 Differentially Private Hotspot Detection

Input: privacy budget ε^h, spatial domain D, a set of location points O, size parameter s, frequency parameter f

Output: a set of hotspots H

1: $SortedListX = SortAscend(O,$ x) $\leftarrow$ sort points in O based on x-values in ascending order

2: $SortedListY = SortAscend(O,$ y) $\leftarrow$ sort points in O based on y-values in ascending order

3: $w = \frac{width(D)}{\sqrt{s}} \leftarrow$ the width of the hotspot

4: $h = \frac{height(D)}{\sqrt{s}} \leftarrow$ the height of the hotspot

5: $o_i \leftarrow$ the first point in $SortedListX$ which has the coordinates (x_i, y_i)

6: **while** o_i exists **do**

7: $CandiListX \leftarrow$ A list of points in $SortedListX$ which have a x-coordinate value in $[x_i, x_i + w)$

8: **if** $|CandiListX| \geq \frac{|O|}{f}$ **then**

9: $(x_{left}, x_{right}) = (DPBG(CandiListX.x, \varepsilon^h/4, left), DPBG(CandiListX.x, \varepsilon^h/4, right))$

10: $SortedCandiListY = SortAscend(CandiListX,$ y) $\leftarrow$ sort points in $CandiListX$ based on y-values in ascending order

11: $o_j \leftarrow$ the first point in $SortedCandiListY$ which has the coordinates (x_j, y_j)

12: **while** o_j exists **do**

13: $CandiListY \leftarrow$ A list of points in $SortedCandiListY$ which have a y-coordinate value in $[y_j, y_j + h)$

14: **if** $|CandiListY| \geq \frac{|O|}{f}$ **then**

15: $(y_{bottom}, y_{top}) = (DPBG(CandiListY.y, \varepsilon^h/4, bottom), DPBG(CandiListY.y, \varepsilon^h/4, top)$

16: $R \leftarrow$ A rectangle whose bottom-left and top-right coordinates are (x_{left}, y_{bottom}) and (x_{right}, y_{top})

17: **if** R does not overlap any hotspots in H **then**

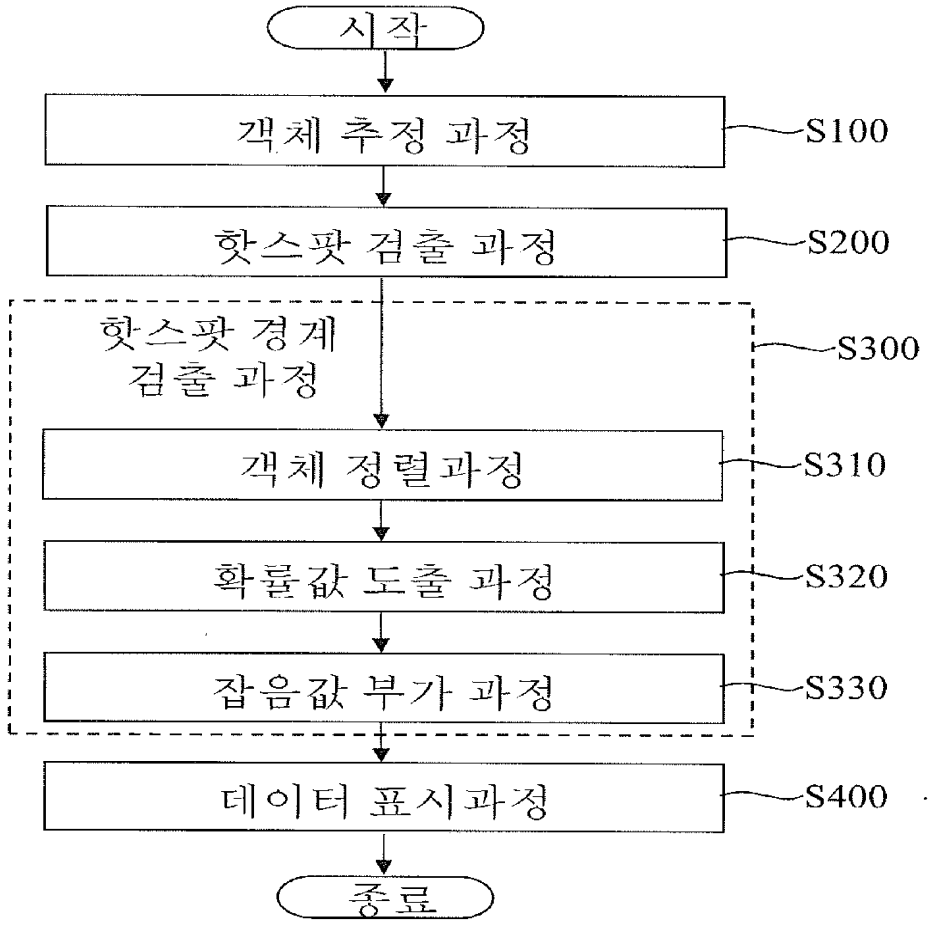
18: Add R into H

19: Remove the points in $CandiListY$ from $SortedListX$ and $SortedListY$

도면7b

20: $o_j \leftarrow$ the next point in $SortedCandiListY$ which has a y-coordinate value greater than $y_j + h$
21: **end if**
22: **else**
23: $o_j \leftarrow$ the next point in $SortedCandiListY$
24: **end if**
25: **end while**
26: **else**
27: $o_i \leftarrow$ the next point in $SortedListX$
28: **end if**
29: **end while**

도면8



도면9

Algorithm 2 Differentially Private Boundary Generation

Input: a sorted coordinate value list V, privacy budget $\varepsilon^{\tilde{h}}$, direction d

Output: a noisy coordinate value c

for $i = 0$ to $|V| - 2$ **do**
  $I_i = [V[i], V[i+1]) \leftarrow$ a coordinate interval
  if $d == left$ or $d == bottom$ **then**
    $rank(I_i) = i$
  else
    $rank(I_i) = |V| - (i+2)$
  end if
end for
Choose an interval I_k with probability $\propto |I_k| \times exp(\varepsilon^{\tilde{h}} \times rank(I_k)/2)$
Choose a random value c from I_k
return c

도면10

Algorithm 3 Differentially Private SAGA Histogram Construction

Input: a set of hotspots H, privacy budget ε^{c}

Output: a set of hotspots with $m \times m$ uniform grid H_{ug}

$h_i \leftarrow$ the first hotspot in H
while h_i exists **do**
  $N_i \leftarrow$ the number of points in h_i
  Compute the grid size $m_i = \sqrt{\frac{N_i \varepsilon^{c}}{c}}$
  Construct $m_i \times m_i$ uniform grid in h_i
  for point $p \in h_i$ **do**
    Increase the count of a grid cell that p falls in
  end for
  Perturb the cell counts by adding Laplace noise $X \sim Lap(\frac{1}{\varepsilon^{c}})$
  $h_i \leftarrow$ the next hotspot in H
end while
return H_{ug}