



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0062764
(43) 공개일자 2019년06월07일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) G06F 16/00 (2019.01)
(52) CPC특허분류
G06F 21/6245 (2013.01)
G06F 16/9024 (2019.01)
(21) 출원번호 10-2017-0161286
(22) 출원일자 2017년11월29일
심사청구일자 2017년11월29일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
박석
서울특별시 종로구 평창12길 16-6, 2동 202호 (평창동, 평창삼호빌라)
강민석
서울특별시 마포구 백범로5길 11, 104호 (노고산동)
(74) 대리인
유미특허법인

전체 청구항 수 : 총 12 항

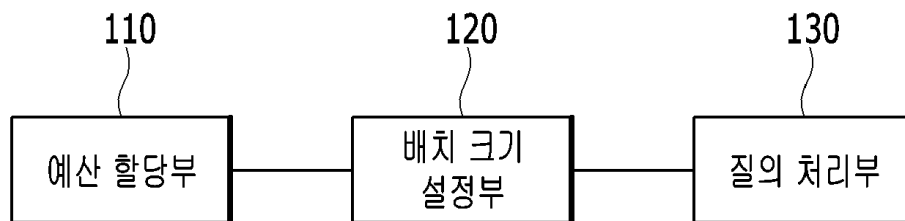
(54) 발명의 명칭 차분 프라이버시 기반 질의 처리 시스템 및 이를 이용한 프라이버시 예산 절약 방법

(57) 요약

질의 처리 시스템이 프라이버시 예산을 절약하는 방법으로서, 수신한 복수의 질의 각각에 대한 최대 갱신 횟수를 계산하고, 계산한 최대 갱신 횟수와 미리 설정된 전체 프라이버시 예산을 기초로 복수의 질의 각각에 대한 예산을 할당한다. 예산이 할당된 복수의 질의 중 적어도 하나의 질의를 하나의 배치에 포함시키고, 하나의 배치에 포함된 적어도 하나의 질의를 재배치한 후 재배치된 적어도 하나의 질의를 처리한다.

대표도 - 도1

100



(52) CPC특허분류

G06F 16/903 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 2017-0-00498

부처명 미래창조과학부

연구관리전문기관 IITP(정보통신기술진흥센터)

연구사업명 정보통신·방송 연구개발 사업 계획서

연구과제명 차분 프라이머시 기반의 비식별화 기술 개발

기 여 율 1/1

주관기관 서강대학교 산학협력단

연구기간 2017.04.01~2017.12.31

명세서

청구범위

청구항 1

질의 처리 시스템이 프라이버시 예산을 절약하는 방법으로서,
수신한 복수의 질의 각각에 대한 최대 갱신 횟수를 계산하는 단계,
상기 계산한 최대 갱신 횟수와 미리 설정된 전체 프라이버시 예산을 기초로 상기 복수의 질의 각각에 대한 예산을 할당하는 단계,
예산이 할당된 복수의 질의 중 적어도 하나의 질의를 하나의 배치에 포함시키는 단계, 그리고
상기 하나의 배치에 포함된 적어도 하나의 질의를 재배치하고, 상기 재배치된 적어도 하나의 질의를 처리하는 단계
를 포함하는 프라이버시 예산 절약 방법.

청구항 2

제1항에 있어서,
상기 최대 갱신 횟수를 계산하는 단계는,
상기 수신한 복수의 질의가 저장된 데이터베이스로부터 원본 히스토그램인 제1 히스토그램과 상이한 균일 분포 형태의 가상 히스토그램인 제2 히스토그램을 생성하는 단계,
상기 제2 히스토그램으로부터 도출된 질의 결과 값인 제1 결과 값과 상기 제1 히스토그램으로부터 도출된 질의 결과 값인 제2 결과 값을 계산하는 단계,
상기 제1 결과 값과 제2 결과 값의 차이가 미리 설정된 임계값을 벗어나는지 확인하는 단계, 그리고
상기 차이가 상기 임계값을 벗어나면, 상기 제2 결과 값에 노이즈를 삽입한 결과를 배포하고, 상기 가상 히스토그램을 갱신하는 단계
를 포함하는 프라이버시 예산 절약 방법.

청구항 3

제2항에 있어서,
상기 최대 갱신 횟수를 계산하는 단계는,
상기 제1 히스토그램과 초기화된 제2 히스토그램 사이의 상대 엔트로피의 상한 값과, 히스토그램 매개 변수, 상기 복수의 질의의 개수, 그리고 상기 임계값을 벗어나는 라플라스 분포 영역을 산출한 값을 기초로 최대 갱신 횟수가 결정되는 프라이버시 예산 절약 방법.

청구항 4

제1항에 있어서,
상기 하나의 배치에 포함시키는 단계는,
포와송 분포에 따라 질의가 진입할 확률이 질의가 임의로 진입할 확률인 균일 분포 값 보다 높아질 때까지 적어도 하나의 질의를 상기 배치에 포함시키는 단계
를 포함하는 프라이버시 예산 절약 방법.

청구항 5

제4항에 있어서,
 상기 배치에 포함시키는 단계는,
 최대 갱신 횟수와 질의 개수 그리고 상기 전체 프라이버시 예산을 토대로 배치 크기의 최대 임계 값을 설정하는 단계
 를 더 포함하는 프라이버시 예산 절약 방법.

청구항 6

제2항에 있어서,
 상기 하나의 질의를 처리하는 단계는,
 상기 제1 결과 값과 제2 결과 값의 차이가 상기 임계값보다 크면, 거리 값을 계산하는 단계, 그리고
 계산한 거리 값을 기준으로 가장 큰 거리 값을 가지는 질의 순으로 재배치하는 단계
 를 포함하고,
 상기 거리 값은 상기 배치에 포함된 적어도 하나의 질의 각각에 대해 상기 원본 히스토그램과 가상 히스토그램
 의 차이로 계산하는 프라이버시 예산 절약 방법.

청구항 7

차분 프라이버시 질의를 처리하는 질의 처리 시스템으로서,
 수신한 복수의 질의 각각에 대한 최대 갱신 횟수를 계산하고, 계산한 최대 갱신 횟수와 미리 설정된 전체 프
 라이버시 예산을 기초로 상기 복수의 질의 각각에 대한 예산을 할당하는 예산 할당부,
 예산이 할당된 복수의 질의 중 적어도 하나의 질의를 하나의 배치에 포함시키는 배치 크기 설정부, 그리고
 상기 하나의 배치에 포함된 적어도 하나의 질의를 처리하는 질의 처리부
 을 포함하는 질의 처리 시스템.

청구항 8

제7항에 있어서,
 상기 예산 할당부는,
 제1 히스토그램과 초기화된 제2 히스토그램 사이의 상대 엔트로피의 상한 값과, 히스토그램 매개 변수, 상기 복
 수의 질의의 개수, 그리고 상기 임계값을 벗어나는 라플라스 분포 영역을 산출한 값을 기초로 상기 최대 갱신
 횟수를 계산하는 질의 처리 시스템.

청구항 9

제8항에 있어서,
 상기 제1 히스토그램은 상기 수신한 복수의 질의가 저장된 데이터베이스로부터 원본 히스토그램이고,
 상기 제2 히스토그램은 제1 히스토그램과 상이한 균일 분포 형태의 가상 히스토그램인 질의 처리 시스템.

청구항 10

제9항에 있어서,
 상기 배치 크기 설정부는,
 포와송 분포에 따라 질의가 진입할 확률이 질의가 임의로 진입할 확률인 균일 분포 값 보다 높아질 때까지 적어
 도 하나의 질의를 상기 배치에 포함시키는 질의 처리 시스템.

청구항 11

제10항에 있어서,

상기 배치 크기 설정부는,

상기 배치의 크기를 최대 갱신 횟수와 질의 개수 그리고 상기 전체 프라이버시 예산을 토대로 최대 임계 값으로 설정하는 질의 처리 시스템.

청구항 12

제11항에 있어서,

상기 배치 크기 설정부는,

상기 배치에 포함된 적어도 하나의 질의 각각에 대해 원본 히스토그램과 가상 히스토그램 사이의 거리 값을 계산하고, 상기 거리 값을 기준으로 가장 큰 거리 값을 가지는 질의 순으로 재배치하는 질의 처리 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 차분 프라이버시 기반 질의 처리 시스템 및 이를 이용한 프라이버시 예산 절약 방법에 관한 것이다.

배경 기술

[0002] 대화형 환경에서 데이터를 공개하거나 임의의 사용자가 질의한 내용에 대한 질의 결과를 제공할 때, 개인 식별을 방지하기 위하여 개인 정보를 보호하는 기법은 필수적이다. 이를 위해 k-익명화 등의 구조적 보호 기법들이 연구되어 왔다. 그러나, 차분 공격을 비롯한 다양한 공격 기법들의 출현으로, 공격자의 사전 지식과 무관하게 일정 수준 이상으로 프라이버시를 보호할 수 있는 차분 프라이버시(differential privacy)가 연구되고 있다.

[0003] 질의 결과에 대해 차분 프라이버시를 만족시키기 위해서는, 일정 수준의 노이즈를 원 데이터에 삽입하여 질의 결과로 제공해야 한다. 이때 원 데이터에 삽입되는 노이즈 값을 도출하는 과정에서, 질의자는 자신에게 할당된 프라이버시 예산(privacy budget)의 일부를 소모한다.

[0004] 예산을 많이 사용할수록 삽입되는 노이즈의 양이 줄어들어, 질의 결과의 정확도는 커진다. 그러나, 질의자의 예산은 한정되어 있기 때문에, 대량의 질의 처리시 각 질의에 적용될 예산을 적절히 분할하여 적용해야 한다. 그리고 예산이 너무 적게 적용될 경우 대량의 노이즈가 삽입되어, 질의 결과의 정확도가 크게 저하되는 문제점이 있다.

발명의 내용

해결하려는 과제

[0005] 따라서, 본 발명은 대화형 환경에서 차분 프라이버시를 적용하여, 선형 질의를 처리하는 질의 처리 시스템 및 이를 이용한 프라이버시 예산을 절약하는 방법을 제공한다.

과제의 해결 수단

[0006] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 하나의 특징인 질의 처리 시스템이 차분 프라이버시 질의를 처리하는 방법으로서,

[0007] 수신한 복수의 질의 각각에 대한 최대 갱신 횟수를 계산하는 단계, 상기 계산한 최대 갱신 횟수와 미리 설정된 전체 프라이버시 예산을 기초로 상기 복수의 질의 각각에 대한 예산을 할당하는 단계, 예산이 할당된 복수의 질의 중 적어도 하나의 질의를 하나의 배치에 포함시키는 단계, 그리고 상기 하나의 배치에 포함된 적어도 하나의 질의를 재배치하고, 상기 재배치된 적어도 하나의 질의를 처리하는 단계를 포함한다.

[0008] 상기 최대 갱신 횟수를 계산하는 단계는, 상기 수신한 복수의 질의가 저장된 데이터베이스로부터 원본 히스토그램인 제1 히스토그램과 상이한 균일 분포 형태의 가상 히스토그램인 제2 히스토그램을 생성하는 단계, 상기 제2 히스토그램으로부터 도출된 질의 결과 값인 제1 결과 값과 상기 제1 히스토그램으로부터 도출된 질의 결과 값인 제2 결과 값을 계산하는 단계, 상기 제1 결과 값과 제2 결과 값의 차이가 미리 설정된 임계값을 벗어나는지 확인하는 단계, 그리고 상기 차이가 상기 임계값을 벗어나면, 상기 제2 결과 값에 노이즈를 삽입한 결과를 배포하

고, 상기 가상 히스토그램을 갱신하는 단계를 포함할 수 있다.

- [0009] 상기 최대 갱신 횟수를 계산하는 단계는, 상기 제1 히스토그램과 초기화된 제2 히스토그램 사이의 상대 엔트로피의 상한 값과, 히스토그램 매개 변수, 상기 복수의 질의의 개수, 그리고 상기 임계값을 벗어나는 라플라스 분포 영역을 산출한 값을 기초로 최대 갱신 횟수가 결정될 수 있다.
- [0010] 상기 하나의 배치에 포함시키는 단계는, 포와송 분포에 따라 질의가 진입할 확률이 질의가 임의로 진입할 확률인 균일 분포 값 보다 높아질 때까지 적어도 하나의 질의를 상기 배치에 포함시키는 단계를 포함할 수 있다.
- [0011] 상기 배치에 포함시키는 단계는, 최대 갱신 횟수와 질의 개수 그리고 상기 전체 프라이버시 예산을 토대로 배치 크기의 최대 임계 값을 설정하는 단계를 포함할 수 있다.
- [0012] 상기 하나의 질의를 처리하는 단계는, 상기 제1 결과 값과 제2 결과 값의 차이가 상기 임계값보다 크면, 거리 값을 계산하는 단계, 그리고 계산한 거리 값을 기준으로 가장 큰 거리 값을 가지는 질의 순으로 재배치하는 단계를 포함하고, 상기 거리 값은 상기 배치에 포함된 적어도 하나의 질의 각각에 대해 상기 원본 히스토그램과 가상 히스토그램의 차이로 계산할 수 있다.
- [0013] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 또 다른 특징인 차분 프라이버시 질의를 처리하는 질의 처리 시스템으로서,
- [0014] 수신한 복수의 질의 각각에 대한 최대 갱신 횟수를 계산하고, 계산한 최대 갱신 횟수와 미리 설정된 전체 프라이버시 예산을 기초로 상기 복수의 질의 각각에 대한 예산을 할당하는 예산 할당부, 예산이 할당된 복수의 질의 중 적어도 하나의 질의를 하나의 배치에 포함시키는 배치 크기 설정부, 그리고 상기 하나의 배치에 포함된 적어도 하나의 질의를 처리하는 질의 처리부를 포함한다.
- [0015] 상기 예산 할당부는, 제1 히스토그램과 초기화된 제2 히스토그램 사이의 상대 엔트로피의 상한 값과, 히스토그램 매개 변수, 상기 복수의 질의의 개수, 그리고 상기 임계값을 벗어나는 라플라스 분포 영역을 산출한 값을 기초로 상기 최대 갱신 횟수를 계산할 수 있다.
- [0016] 상기 배치 크기 설정부는, 포와송 분포에 따라 질의가 진입할 확률이 질의가 임의로 진입할 확률인 균일 분포 값 보다 높아질 때까지 적어도 하나의 질의를 상기 배치에 포함시킬 수 있다.
- [0017] 상기 배치 크기 설정부는, 상기 배치의 크기를 최대 갱신 횟수와 질의 개수 그리고 상기 전체 프라이버시 예산을 토대로 최대 임계 값으로 설정할 수 있다.

발명의 효과

- [0018] 본 발명에 따르면 한정된 프라이버시 예산을 고려하여 배치 처리를 수행하여, 동일한 수의 질의 처리시 보다 적은 예산으로 질의를 처리할 수 있다.

도면의 간단한 설명

- [0019] 도 1은 본 발명의 실시예에 따른 질의 처리 시스템의 구조도이다.
- 도 2는 본 발명의 실시예에 따른 프라이버시 예산 절약 방법에 대한 흐름도이다.
- 도 3은 본 발명의 실시예에 따른 실제 갱신 횟수와 추론된 갱신 횟수를 비교한 예시도이다.
- 도 4는 본 발명의 실시예에 따른 방법과 일반적인 방법에 따른 갱신 횟수를 비교한 예시도이다.
- 도 5는 본 발명의 실시예에 따른 다양한 히스토그램과 라플라스 분포에서 임계값을 벗어난 영역에 대한 예시도이다.

발명을 실시하기 위한 구체적인 내용

- [0020] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

- [0021] 이하, 도면을 참조로 하여 본 발명의 실시예에 따른 차분 프라이버시 기반 질의 처리 시스템 및 이를 이용한 프라이버시 예산 절약 방법에 대해 상세히 설명한다.
- [0022] 도 1은 본 발명의 실시예에 따른 질의 처리 시스템의 구조도이다.
- [0023] 도 1에 도시된 바와 같이, 질의 처리 시스템(100)은 적어도 하나의 프로세서에 의해 동작하고, 예산 할당부(110), 배치 크기 설정부(120), 그리고 질의 처리부(130)을 포함한다.
- [0024] 예산 할당부(110)는 사용자 단말(도면 미도시) 또는 외부로부터 복수의 질의를 수신한다. 이때, 예산 할당부(110)가 수신한 복수의 질의는 선형 질의를 예로 하여 설명한다.
- [0025] 예산 할당부(110)는 수신한 질의의 개수와, 미리 허용된 전체 예산 정보를 토대로, 최대 갱신 횟수를 계산한다. 그리고 계산한 최대 갱신 횟수로 예산을 분할하여, 하나의 질의에 할당될 예산을 결정한다.
- [0026] 즉, 종래에는 질의의 개수를 전체 예산으로 나눈 만큼의 예산이 하나의 질의에 할당하였다. 이 경우, 질의의 개수가 늘어날수록 하나의 질의에 할당되는 예산이 매우 작아지기 때문에, 노이즈가 과도하게 삽입되어 질의의 정확도가 감소되는 단점이 있다.
- [0027] 따라서, 예산 할당부(110)는 단순히 질의의 개수로 전체 예산을 나누는 대신, 먼저 최대 갱신 횟수를 계산한다. 그리고, 계산한 최대 갱신 횟수를 토대로 예산을 분할한다.
- [0028] 배치 크기 설정부(120)는 예산 할당부(110)에서 할당된 예산을 토대로, 배치 크기를 결정한다. 배치 크기는 클수록 질의에 대한 예산 소모를 최적화할 수 있으나, 배치 크기가 지나치게 커지면 각 질의의 반응 시간(response time)이 늘어나는 문제가 발생한다.
- [0029] 따라서, 배치 크기 설정부(120)는 질의의 모수와 포아송 분포에 따라 질의가 진입할 확률을 계산한다. 그리고, 계산한 확률을 토대로 배치 크기 설정부(120)는 배치 크기의 최대 임계 값을 설정한다. 배치 크기 설정부(120)는 설정한 배치 크기의 최대 임계 값을 이용하여, 질의에 대한 배치 크기를 결정한다. 이에 대해서는 이후 상세히 설명한다.
- [0030] 그리고 배치 크기 설정부(120)는 질의가 있을 때 가상 히스토그램으로부터의 질의 결과와 원본 히스토그램으로부터의 질의 결과에 노이즈를 삽입한 값을 비교한다. 그리고, 비교한 값이 미리 설정된 임계값보다 크면 상대 엔트로피가 크다고 판단하여, 질의를 거리 값(distance)이 큰 질의 기준으로 재정렬하여 가상 히스토그램을 갱신한다. 이를 토대로 적은 수의 갱신으로도 가상 히스토그램을 원본 히스토그램과 유사하게 만들어, 프라이버시 예산을 절약한다. 이에 대해서는 이후 상세히 설명한다.
- [0031] 질의 처리부(130)는 예산 할당부(110)가 예산을 분할하고, 배치 크기 설정부(120)가 질의에 대한 배치 크기를 설정하고 질의를 재정렬하면, 분할된 예산과 설정된 배치 크기를 기초로 질의를 처리하여 질의 응답 정보를 생성한다. 질의 처리부(130)가 질의를 처리하는 방법은 다양한 방법으로 수행할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.
- [0032] 또한, 본 발명의 실시예에서는 설명의 편의를 위하여 따로 도시하지 않았으나 질의 처리 시스템(100)을 구동하기 위해 필요한 프로그램들이 저장, 관리하는 메모리를 추가로 포함할 수 있다. 여기서 프로그램의 형태나 질의 응답 정보의 형태는 다양하므로, 본 발명의 실시예에서는 어느 하나의 형태로 한정하지 않는다.
- [0033] 이상에서 설명한 질의 처리 시스템(100)으로 프라이버시 예산을 절약하는 방법에 대해 도 2를 참조로 설명한다.
- [0034] 도 2는 본 발명의 실시예에 따른 프라이버시 예산 절약 방법에 대한 흐름도이다.
- [0035] 도 2에 도시된 바와 같이, 외부로부터 복수의 질의를 수신하면(S100), 질의 처리 시스템(100)은 질의 개수를 확인한다. 그리고, 질의 처리 시스템(100)은 수신한 복수의 질의에서 최대 갱신 횟수를 계산한다(S101).
- [0036] 여기서 질의 처리 시스템(100)이 질의 개수를 확인하는 방법은 다양한 방법으로 확인할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.
- [0037] 또한, 대화형 환경에서 질의 처리 시스템이 선형 질의를 처리할 때, 질의 처리 시스템(100)은 가상 히스토그램을 활용하여 예산의 소모를 줄인다. 이는 가상 히스토그램과 매개 변수들을 활용하여, 일정 조건 하에서 원본 히스토그램이 아닌 가상 히스토그램을 통해 질의 결과를 제공하여 예산의 소모를 줄인다.
- [0038] 이를 통해 질의 처리 시스템(100)은 질의가 저장된 실제 데이터베이스로부터 비롯된 원본 히스토그램과 별개로,

균일 분포 형태의 가상 히스토그램을 생성한다. 이후 가상 히스토그램으로부터 도출된 질의 결과와 원본 히스토그램으로부터 도출된 값에 노이즈를 삽입한 결과를 비교한다. 그리고 차이가 임계값(T) 이하면 가상 히스토그램으로부터 도출된 결과를 배포해 예산을 소모하지 않는다.

- [0039] 반면, 차이가 임계값 이상이면 원본 히스토그램으로부터 도출된 값에 노이즈를 삽입한 결과를 배포한다. 그리고 가상 히스토그램을 갱신하여 원본 히스토그램과 유사하게 만들어 나간다. 원본 히스토그램에 노이즈를 삽입한 결과를 배포할 때만 예산이 소모되며 대화형 환경에서 계속해서 유입되는 질의에 대해 위의 과정을 반복해서 수행한다.
- [0040] 이는 임계값(T)과 히스토그램 갱신 매커니즘을 설정하여, 본 발명의 실시예에 따른 질의 처리 방법이 (ϵ, δ) -차분 프라이버시를 만족시킴을 증명한다. 차분 프라이버시 만족 증명 방법은 다양한 방법으로 증명할 수 있으므로, 본 발명의 실시예에서는 상세한 설명을 생략한다.
- [0041] 그러나 동일한 질의 집합에 대해 이 방법을 수행하더라도, 진입하는 질의의 진입 순서에 따라 가상 히스토그램의 갱신 횟수가 판이하게 달라지며 이는 기법의 성능편차를 매우 크게 만든다. 이에 대해 도 5를 참조로 먼저 설명한다.
- [0042] 도 5는 본 발명의 실시예에 따른 다양한 히스토그램과 라플라스 분포에서 임계값을 벗어난 영역에 대한 예시도이다.
- [0043] 도 5의 (a)에 도시된 바와 같이 질의에 대한 원본 히스토그램이 존재한다고 가정한다. 그러면, 도 5의 (c-1)과 같이 초기값으로 정규화된 가상 히스토그램은 연속적으로 진입하는 질의에 의해 갱신이 일어난다. 만약 첫 번째 질의 Q1이 bin 2에 대한 질의이고, 두 번째 질의 Q2와 세 번째 질의 Q3이 각각 bin 7, bin 10에 대한 질의라고 가정한다.
- [0044] 이 세 질의에 대하여 가상 히스토그램은 도 5의 (c-2), (c-3), (c-4)와 같이 세 번의 갱신이 일어나게 된다. 그러나 만약 일곱 번째 질의 Q7이 bin 7과 bin 10에 대한 질의라면, 도 5의 (b-2)와 같이 Q7를 통한 한 번의 갱신만으로도 Q1, Q2, Q3의 세 번의 갱신의 결과와 유사한 가상 히스토그램을 얻을 수 있다.
- [0045] 배치를 사용하여 배치 내의 질의들을 효율적인 순서로 재배치하면 (ϵ, δ) -차분 프라이버시의 증명에 영향을 주지 않는다. 그리고, 일정 수준의 성능 수준을 유지할 수 있고, 비슷한 가상 히스토그램을 기존의 방법에 비해 더 적은 횟수의 갱신으로도 얻을 수 있기 때문에, 예산 소모도 줄일 수 있다.
- [0046] 상기 도 2를 이어 설명하면, 수신한 복수의 질의에서 최대 갱신 횟수를 계산하기 위하여, 질의 처리 시스템(100)은 초기 가상 히스토그램과 원본 히스토그램 사이의 상대 엔트로피의 상한 값과 질의에 대한 한 번의 히스토그램 갱신이 발생할 때 감소되는 상대 엔트로피 수를 이용한다.
- [0047] 즉, 질의에 대해 한 번의 히스토그램이 갱신될 때 η^2 만큼의 상대 엔트로피 감소가 일어난다. 따라서, 초기 가상 히스토그램(y)와 원본 히스토그램(x) 사이의 상대 엔트로피의 상한은 $\log H$ 가 된다. 이를 통해, 질의 처리 시스템(100)은 초기 가상 히스토그램이 원본 히스토그램과 유사해질 때 까지는 최대 $\log H / \eta^2$ 번 갱신이 시도됨을 계산할 수 있다.
- [0048] 그러나, 최대 갱신 시도 횟수는 라플라스 분포로부터 계산된 노이즈 삽입을 감안하지 않은 결과이다. 따라서 질의 처리 시스템(100)은 가상 히스토그램으로부터 비롯된 질의의 결과값과 원본 히스토그램으로부터 비롯된 값에 노이즈를 삽입한 결과를 비교하여 추가적인 갱신이 발생함을 인지한다. 여기서, 노이즈의 분포는 라플라스 분포를 따르므로, 임계값을 벗어나는 영역을 p라 할 때 다음 수학적 식 1과 같이 구할 수 있다.

수학적 식 1

$$p = 1 - \int_{-T}^T \frac{1}{2b} e^{(-\frac{|x|}{b})} dx$$

[0049]

[0050] 따라서, $N \cdot p$ 만큼의 추가적인 갱신이 발생할 경우, 최대 $\frac{\log H}{\eta^2} + Np$ 만큼의 갱신이 발생한다.

[0051] 질의 처리 시스템(100)은 수학적 1에서 구한 최대 갱신 발생 횟수를 토대로 하나의 질의에 할당할 예산을 결정한다(S102). 즉, 한 번의 갱신 과정에 할당될 수 있는 프라이버시 예산(UE)은 다음 수학적 2와 같다. 질의 처리 시스템(100)은 프라이버시 예산을 갱신할 때 마다 할당한다.

수학적 2

$$UE = \frac{\text{전체 privacy budget}}{\text{최대 갱신 횟수}} = \frac{\epsilon}{\frac{\log H}{\eta^2} + Np}$$

[0052]

[0053] S102 단계를 통해 예산이 결정되면, 질의 처리 시스템(100)은 배치 크기를 설정한다(S103). 배치 크기는 커질수록 질의에 따른 예산 소모를 최적화할 수 있으나, 배치의 크기가 지나치게 커지면 각 질의의 반응 시간이 늘어난다. 따라서, 배치 크기는 적절한 수준을 유지해야 한다.

[0054] 질의 처리 시스템(100)가 배치 크기를 설정하기 위해, 본 발명의 실시예에서는 질의는 모수가 λ 인 포아송 분포에 따라 질의가 진입한다고 가정한다. 분포에 의해 이전 배치가 m 번째 질의까지 처리했을 때, 그 다음 질의부터 n 번째 질의까지 질의가 진입할 확률은 다음 수학적 3과 같다.

수학적 3

$$f(x) = \frac{e^{-\lambda} * \lambda^x}{x!}, \text{ (모수가 } \lambda \text{ 일 때, } x \text{ 번째 사건이 발생할 확률)}$$

$$P(X \leq n) - P(X \leq m) = \sum_{x=0}^n f(x) - \sum_{x=0}^m f(x), \quad n > m$$

[0055]

[0056] 여기서 $f(x)$ 는 모수가 λ 일 때, x 번째 사건이 발생할 확률을 의미한다.

[0057] 질의 처리 시스템(100)은 수학적 3의 포아송 분포에 따라, 질의가 진입할 확률이 질의가 임의로 진입할 확률인 균일 분포 값보다 높아질 때까지 질의들을 하나의 배치에 포함시킨다. 이때, 포아송 분포의 경우 모수에 따라 전체 질의까지 모두 들어올 확률인 $P(X \leq N) = \alpha$ 가 1보다 작을 수 있다. 따라서 포아송 분포 값과 균일 분포 값을 비교하여 다음 수학적 4의 조건이 만족될 때까지 질의들을 배치에 포함시킨다.

수학적 4

$$P(X \leq n) - P(X \leq m) > \frac{(n - m)\alpha}{N}$$

[0058]

[0059] 단, 포아송 분포의 특성상 모수와 차이가 심할수록 확률 값이 급격히 감소하기 때문에, 배치의 크기가 매우 커질 수 있다. 그러므로 일정 수준의 질의 반응시간을 유지하기 위해 배치 크기의 최대 임계 값은 다음 수학적 5와 같이 설정한다.

수학식 5

$$b_{\max} = \frac{N}{\text{최대 갱신 횟수}} \cdot \frac{1}{\epsilon} = \frac{N}{\left(\frac{\log H}{\eta^2} + Np\right)\epsilon}$$

[0060]

[0061]

ϵ 이 커질 때 최대 갱신 횟수는 감소하고, 질의 수가 많아지더라도 최대 갱신 횟수는 증가한다. 따라서 최대 임계 값은 ϵ 나 N 의 변화에도 일정 수준의 질의 반응시간을 유지할 수 있는 배치의 크기를 제공한다. S102 단계에서 결정한 예산과 S103 단계에서 설정한 배치 크기를 고려한 알고리즘은 다음 표 1과 같다.

표 1

[0062]

Require: x, ϵ, Q, H, E, N x: 원본 히스토그램, ϵ: 질의 당 요구되는 프라이버시 수준 y: 가상 히스토그램, T: 결과값 차이 임계값 $Q = \{q_1, q_2, \dots, q_N\}$: 질의 집합, N: 질의 개수 H: 원본 히스토그램 bin 수, B: 배치 크기 E: 전체 프라이버시 예산, n: 데이터베이스 레코드 수 Predefine: η, σ, T $\eta = \frac{\log^{1/4} H}{\sqrt{n}}$: 히스토그램 갱신 변수 $\sigma = \frac{10 \eta \log(1/\epsilon)}{\epsilon}$: 라플라스 분포 분산 $T = 4 \cdot \sigma \cdot \log N$: 결과값 차이 임계값	
<pre> for j = 1, 2, ..., H do y(b_j) = 1/H for each round t = 1, 2, ..., N do if Poisson_accu_dist < Uniform_accu_dist in B ≤ b_{max}: insert q_t in Batch: continue: While B>0: for i = 1, 2, ..., B do A_i ~ Lap(σ). Compute the noisy answer a_i = <q_i, x> + A_i computed the difference d_i = a_i - <q_i, y> ascending_sort_distance(Batch) for i = 1, 2, ..., B do if d_i ≤ T: output <q_i, y> eles: y = U(y), output a_i </pre>	

[0063]

표 1에 나타난 바와 같이, 예산 할당과 배치를 고려한 알고리즘을 실행하기 위해, 원본 히스토그램(x), 질의 당 요구되는 프라이버시 수준(ϵ), 가상 히스토그램(y), 결과값 차이 임계값(T), 질의 집합(Q), 질의 개수(N), 원본 히스토그램 bin 수, 배치 크기(B), 전체 프라이버시 예산(E), 데이터베이스 레코드 수(n)를 입력 정보로 한다. 또한, 사전 정의(predefine) 값으로 히스토그램 갱신 매개 변수(η), 라플라스 분포 분산(σ), 결과값 차이 임계값(T)이 있다.

[0064]

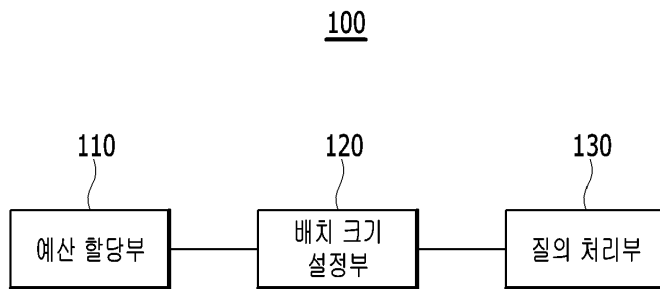
표 1에 도시된 알고리즘을 구동하면서, 질의 처리 시스템(100)은 하나의 배치 내에 포함된 적어도 하나의 질의들을 재정렬한다. 즉, 질의 처리 시스템(100)은 가상 히스토그램의 계속된 갱신을 통해 가상 히스토그램을 원본 히스토그램과 유사하게 만들어 나간다. 이는 가상 히스토그램과 원본 히스토그램의 차이를 상대 엔트로피

(Relative entropy)의 개념으로 나타낼 때, 질의마다 변화시킬 수 있는 상대 엔트로피의 변화량이 다르기 때문이다.

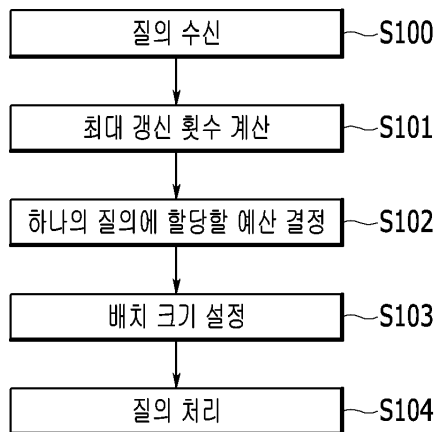
- [0065] 따라서, 하나의 배치 내의 질의 중에서 상대 엔트로피를 가장 크게 변화시키는 질의를 기준으로 질의를 재배치하여, 가상 히스토그램 갱신을 수행한다. 여기서, 질의 재배치를 위해 질의 처리 시스템(100)은 질의를 거리 값(distance)을 계산하는데, 거리 값(d_i)은 i 번째 질의에 대한 원본 히스토그램과 가상 히스토그램의 차이인 $a_i - \langle q_i, y \rangle$ 로 계산하고, $a_i = \langle q_i, x \rangle + A_i$ 이며, A_i 는 노이즈가 삽입된 결과 값을 의미한다. 이를 토대로 적은 수의 가상 히스토그램 갱신으로도 가상 히스토그램을 원본 히스토그램과 유사하게 만들 수 있으며, 이를 통해 프라이버시 예산을 절약한다.
- [0066] 한편 상기 도 2의 S103 단계에서 배치 크기가 설정되면, 질의 처리 시스템(100)은 S100 단계에서 수신한 질의를 처리하여 질의한 사용자의 단말(도면 미도시)에 제공한다(S104).
- [0067] 다음은, 본 발명의 실시예에 따른 실제 갱신 횟수와 추론된 갱신 횟수를 비교한 예에 대해 도 3을 참조로 설명하고, 본 발명의 실시예에 따른 방법과 일반적인 방법에 따른 갱신 횟수의 비교에 대해 도 4를 참조로 설명한다.
- [0068] 도 3은 본 발명의 실시예에 따른 실제 갱신 횟수와 추론된 갱신 횟수를 비교한 예시도이다. 그리고, 도 4는 본 발명의 실시예에 따른 방법과 일반적인 방법에 따른 갱신 횟수를 비교한 예시도이다.
- [0069] 본 발명의 실시예에서 사용된 데이터는 StatLib의 The National Long-Term Case Study(NLTCS)의 데이터로써, 16개의 이진 속성으로 이루어졌으며 21674개의 레코드로 구성되어 있다. 프라이버시 보호수준 ϵ 의 변화에 대한 추론된 갱신횟수의 정확도 측정과, 프라이버시 예산의 소모에 대한 실험이며, δ 는 0.001로 고정한 것을 예로 하여 설명한다.
- [0070] 또한, 본 발명의 실시예에 따른 갱신 횟수는 예산 소모를 의미한다. 포아송 분포의 모수는 질의의 수로 설정하였으며, 10,000개의 임의의 선형 질의가 사용된 것을 예로 하여 설명한다.
- [0071] 먼저 도 3에 도시된 바와 같이, ϵ 가 변화하더라도 실제 갱신횟수와 거의 유사한 값을 추론하는 것을 확인 할 수 있다. 따라서 본 발명의 실시예에 따른 방법으로 계산한 갱신 횟수에 기반하여, 각 갱신 단계에 할당하는 예산이 적합한 것을 확인할 수 있다. 이를 통해 달성하고자 하는 프라이버시 수준(ϵ)에 대하여, 각 갱신단계에 할당할 수 있는 예산을 정해줄 수 있으며 이는 예산이 한정되어 있을 때 매우 유용하다.
- [0072] 그리고 도 4에 도시된 바와 같이, 동일한 프라이버시 성능에 대하여 기존기법의 예산 소모 정도와 제안 기법의 예산 소모 정도를 비교한 예시도를 살펴보면, 모든 구간의 프라이버시 보호 수준에 대하여 배치를 통한 질의의 재배열이 실제로 갱신횟수를 줄이는 데 효과적이라는 것을 확인 할 수 있다. 그리고, 무조건적으로 배치의 크기를 늘린 것이 아니라, 적절한 임계값의 설정을 통해 질의의 반응시간을 과도하게 늘리지 않는다. 이를 통해 본 발명의 실시예에 따른 방법이 기존 방법에 비해 동일 성능의 프라이버시 성능에 대하여 소모되는 예산이 적은 것을 확인할 수 있다.
- [0073] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

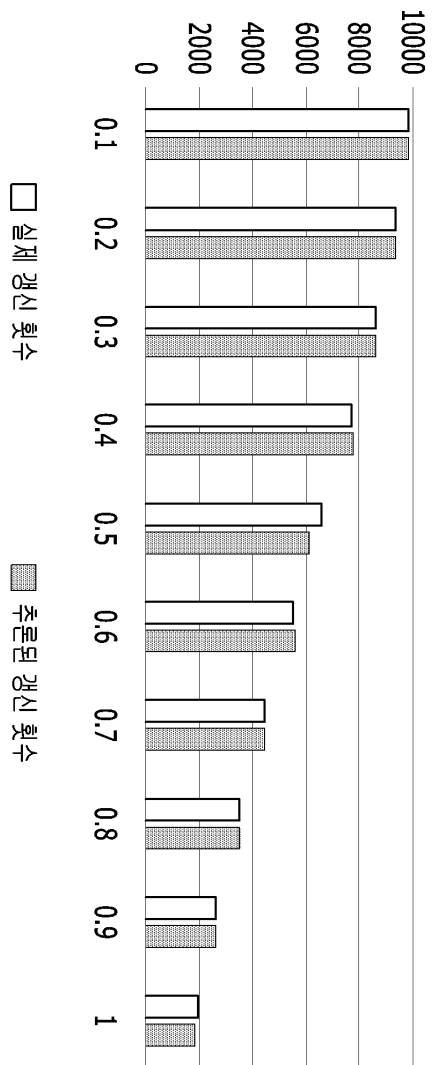
도면1



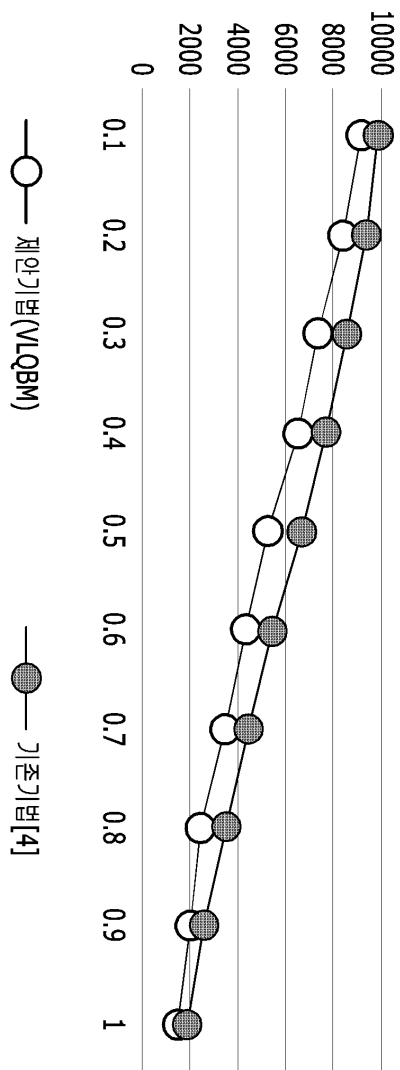
도면2



도면3



도면4



도면5

