(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0070534
(43) 공개일자 2021년06월15일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) *G06F 16/21* (2019.01)
G06F 16/23 (2019.01) *G06F 17/18* (2006.01)
G06F 21/62 (2013.01) *G06N 7/00* (2006.01)
(52) CPC특허분류
G06N 20/00 (2019.01)
G06F 16/21 (2019.01)
(21) 출원번호 10-2019-0160372
(22) 출원일자 2019년12월05일
심사청구일자 2019년12월05일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정연돈
서울특별시 중구 청계천로 400, 105-812(황학동, 롯데캐슬베네치아)
정기정
서울특별시 성북구 고려대로13길 18-20, 502호
(74) 대리인
김홍석

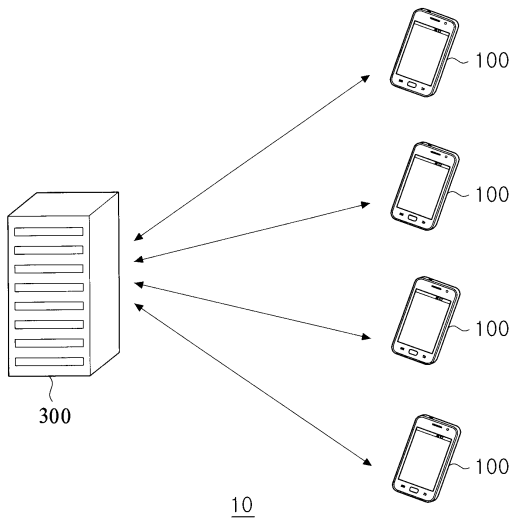
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **지역 차분 프라이버시를 적용한 시계열 위치 데이터의 수집 및 분석 방법**

(57) 요 약

시계열 위치 데이터의 수집 및 분석 시스템과 분석 서버가 개시된다. 상기 수집 및 분석 시스템은 복수의 단말들과 서버를 포함하고, 상기 복수의 단말들 각각은 사용자의 경로 데이터들이 지역 차분 프라이버시(Local Differential Privacy, LDP)를 만족하도록 변조(perturbation)하고 변조된 경로 데이터들을 상기 서버로 송신하고, 상기 서버는 상기 복수의 단말들 각각으로부터 수신된 상기 변조된 경로 데이터들을 이용하여 은닉 마르코프 모델(Hidden Markov Model, HMM)을 모델링하고, 상기 은닉 마르코프 모델의 전이 확률(Transition Probability)을 계산한다.

대 표 도 - 도6



(52) CPC특허분류
G06F 16/23 (2019.01)
G06F 17/18 (2013.01)
G06F 21/6245 (2013.01)
G06N 7/005 (2013.01)
G06F 2221/2111 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711080933
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기술진흥센터
연구사업명 SW컴퓨팅산업원천기술개발
연구과제명 개인정보를 안전하고 편리하게 빅데이터 처리할 수 있는 방법
기 여 율 1/1
과제수행기관명 고려대학교산학협력단
연구기간 2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

복수의 단말들과 서버를 포함하는 시계열 데이터의 수집 및 분석 시스템에 있어서,

상기 복수의 단말들 각각은 사용자의 경로 데이터들이 지역 차분 프라이버시(Local Differential Privacy, LDP)를 만족하도록 변조(perturbation)하고 변조된 경로 데이터들을 상기 서버로 송신하고,

상기 서버는 상기 복수의 단말들 각각으로부터 수신된 상기 변조된 경로 데이터들을 이용하여 은닉 마르코프 모델(Hidden Markov Model, HMM)을 모델링하고, 상기 은닉 마르코프 모델의 전이 확률(Transition Probability)을 계산하는,

시계열 위치 데이터의 수집 및 분석 시스템.

청구항 2

제1항에 있어서,

상기 복수의 단말들 각각은,

사용자의 경로 데이터를 경로의 길이에 대응하는 크기를 갖는 비트 배열로 인코딩하고,

상기 비트 배열로부터 영구 확률 기반 응답(Permanent Randomized Response, PR)을 생성하고,

상기 영구 확률 기반 응답(PR)으로부터 즉석 확률 기반 응답(Instantaneous Randomized Response, IR)을 생성하고,

상기 즉석 확률 기반 응답(IR)을 변조된 경로 데이터로 상기 서버에게 송신하는,

시계열 위치 데이터의 수집 및 분석 시스템.

청구항 3

제1항에 있어서,

상기 서버는 Baum-welch 알고리즘을 이용하여 상기 전이 확률의 업데이트 동작을 반복적으로 수행함으로써 상기 전이 확률을 계산하는,

시계열 위치 데이터의 수집 및 분석 시스템.

청구항 4

제3항에 있어서,

상기 서버는,

상기 복수의 단말들 각각으로부터 상기 변조된 경로 데이터들을 수신하는 데이터 수신부;

상기 변조된 경로 데이터들을 이용하여 상기 은닉 마르코프 모델의 각 변수를 설정하여 상기 은닉 마르크프 모델을 모델링하는 모델 모델링부;

변조된 경로 데이터에 포함된 관측 값, 초기값이 설정된 전이 확률, 및 미리 결정된 관측 확률을 이용하여 사후 확률(Posterior Probability)을 계산하는 사후 확률 도출부; 및

상기 사후 확률을 이용하여 상기 전이 확률을 계산하는 전이 확률 도출부를 포함하는,

시계열 위치 데이터의 수집 및 분석 시스템.

청구항 5

복수의 단말들 각각으로부터 지역 차분 프라이버시를 만족하도록 변조된 경로 데이터들을 수신하는 데이터 수신부;

상기 변조된 경로 데이터들을 이용하여 마르코프 모델의 각 변수를 설정하여 상기 은닉 마르코프 모델을 모델링하는 모델 모델링부;

변조된 경로 데이터에 포함된 관측 값, 초기값이 설정된 전이 확률, 및 미리 결정된 관측 확률을 이용하여 사후 확률을 계산하는 사후 확률 도출부; 및

상기 사후 확률을 이용하여 상기 전이 확률을 계산하는 전이 확률 도출부를 포함하는 시계열 위치 데이터의 분석 서버.

청구항 6

제5항에 있어서,

상기 변조된 경로 데이터는 사용자의 경로 데이터를 경로의 길이에 대응하는 크기를 갖도록 인코딩된 비트 배열로 생성된 영구 확률 기반 응답으로부터 생성된 즉석 확률 기반 응답인,

시계열 위치 데이터의 분석 서버.

청구항 7

제5항에 있어서,

상기 전이 확률 도출부는 상기 전이 확률을 계산하여 업데이트하는 동작을 반복적으로 수행함으로써 상기 전이 확률을 계산하는,

시계열 위치 데이터의 분석 서버.

청구항 8

제7항에 있어서,

전이 확률 도출부에 의해 계산된 전이 확률을 상기 마르코프 모델에 적용하여 사용자들의 시작 위치 분포 또는 각 위치 별 사용자의 이동 경향성을 추정하는 정보 추정부를 더 포함하는,

시계열 위치 데이터의 분석 서버.

발명의 설명

기 술 분 야

[0001] 본 발명은 지역 차분 프라이버시(Local Differential Privacy, LDP)를 적용한 시계열 위치 데이터의 수집 및 분석 방법에 관한 것이다.

배 경 기 술

[0003] 모바일 기기와 IoT(Internet of Things) 기기들은 현재 널리 이용되고 있다. 에릭슨의 보고서(비특허문헌 [1])

에 따르면, 2018년 78억 개의 모바일 기기가 사용되고 있으며 2024년에는 87억 개의 모바일 기기가 사용중일 것으로 예상하고 있다. 모바일 기기의 GPS 센서에서 위치 데이터가 폭발적으로 생성되고 있다. 여러 사용자로부터 위치 데이터를 수집하여 위치 기반 서비스(Location Based Service, LBS) 등 유용한 서비스를 제공하고 있다. 예를 들어, 스마트폰을 이용하여 길안내 서비스를 이용할 경우 목적지까지 거리뿐만 아니라 서비스를 이용하는 다른 사용자의 위치를 활용하여 도로의 혼잡도를 고려한 길 안내가 가능하다.

[0004] 개인의 위치 데이터는 민감한 정보이기 때문에 프라이버시 보호 처리를 하지 않고 그대로 수집하거나 이용한다면 개인의 프라이버시를 침해할 수 있다. 개인의 프라이버시를 보호하면서 사용자의 데이터를 이용할 수 있는 연구가 진행되었다. 가장 흔하게 이용되는 방법은 사용자의 동의를 받은 후 원본 데이터를 수집하는 것이다. 데이터 배포가 필요한 경우 원본 데이터에 프라이버시 보호 처리를 한 후에 데이터를 배포한다. 프라이버시를 충분히 고려하지 않는다면 배포된 데이터로 인하여 개인의 프라이버시가 침해당할 수 있다. 예를 들어, Netflix에서 영화 추천 알고리즘 향상 목적이 대회를 열고 500,000명의 100,000,000개의 영화 평가 데이터를 직접적으로 개인을 식별할 수 있는 요소를 제거한 후 배포하였다. 해당 데이터는 온라인 영화 평가 사이트의 데이터와 결합되어 개인이 식별되었다(비특허문헌 [2]).

[0005] 차분 프라이버시(Differential Privacy, DP)(비특허문헌 [3])는 데이터 프라이버시를 보장하는 가장 강력한 모델이다. 차분 프라이버시는 프라이버시 보호 데이터 마이닝 환경에서 사용하는 모델이다. 통계적으로 공격자가 정해진 프라이버시 보호 수준 이상으로 데이터의 값을 확신할 수 없도록 한다. 가장 대표적으로 차분 프라이버시를 만족하도록 하는 방법으로 라플라스 분포를 따르는 노이즈 추가가 있다. 예를 들어, 판타지 영화를 보는 남성이 평균 연령에 대한 질문을 했을 때에 실제 정답 "30세"에 노이즈가 추가되어 "27세"나 "35세"와 같은 변형된 답을 얻게 된다.

[0006] DP를 적용하기 위해서는 민감도(Sensitivity)를 알아야 한다. 민감도란 데이터셋에서 한 레코드가 없어졌을 때 결과 값에 줄 수 있는 최댓값을 의미한다. 민감도를 알기 위해선 전체 데이터셋(dataset)을 알아야 하는데, 데이터 수집 환경에선 데이터를 제공하는 사용자의 입장에선 다른 모든 사용자의 응답을 알 수 없다. 이러한 이유로 데이터 수집 환경에서 DP를 그대로 적용하기 어렵다.

[0007] DP를 프라이버시 보호 데이터 수집 환경에 적용하기 위해서 지역 차분 프라이버시(Local Differential Privacy, LDP)(비특허문헌 [4])가 제안되었다. Google(비특허문헌 [5]), Uber(비특허문헌 [6]), Microsoft(비특허문헌 [7]), Apple(비특허문헌 [8]), 삼성(비특허문헌 [9, 19]) 등 여러 대기업에서 지역 차분 프라이버시를 이용한 데이터 수집 연구를 진행하고 있다.

[0008] LDP는 데이터를 제공하는 개인이 제공할 수 있는 모든 경우의 수에 대해 통계적으로 공격자가 정해진 프라이버시 보호 수준 이상으로 데이터의 값을 확신할 수 없도록 한다. LDP를 만족하는 가장 대표적인 데이터 수집 방법으로 확률 기반 응답이 있다. 확률 기반 응답은 어떤 질문에 대해 "예" 혹은 "아니오"로 대답하는 경우에 대해 p 확률로 자신의 원래 대답("예" 였다면 "예", "아니오" 였다면 "아니오")을 1-p 확률로 반대 응답("예" 였다면 "아니오", "아니오" 였다면 "예")을 하는 것이다. 이와 같이 데이터를 수집할 경우 데이터 수집자는 한 명의 응답에 대해서 "예"인지 "아니오"인지 알 수 없지만 여러 명으로부터 응답을 수집하여 각 응답의 비율을 추정할 수 있다. 예를 들어, "예" 수집한 답의 비율이 x라고 할 때, 실제 "예"의 응답 비율은 $x \times p+(1-x) \times (1-p)$이다.

[0009] "예", "아니오"가 아닌 여러 종류의 응답이 발생하는 경우에 대한 수집은 다음과 같은 순서로 데이터를 수집 및 분석한다.

[0010] · 데이터 인코딩 (Data encoding)

[0011] · 데이터 변조 (Data perturbation)

[0012] · 총계 (Aggregation)

[0013] 데이터 인코딩이란 사용자의 데이터를 변조하기 쉽게 다른 형태로 변환하는 것이다. 예를 들어, 응답의 경우의 수가 n개 있다고 할 때, 확률 기반 응답을 적용하기 위하여 크기가 n인 비트 배열을 생성하고 각 비트에 응답을 할당한다. 그 후 응답에 해당하는 비트를 1로, 나머지 비트를 0으로 인코딩한다.

[0014] 변조 과정은 지역 차분 프라이버시를 만족하는 변조 방법을 이용하여 인코딩된 데이터를 변조하는 것이다. 데이터 인코딩 과정에서 예시를 이어간다면, 크기가 n인 비트 배열의 각 비트에 확률 기반 응답을 적용한다. 비트는 0 혹은 1이기 때문에 확률 기반 응답을 적용할 수 있다.

[0015] 변조된 데이터는 원본 데이터에 노이즈가 추가되어 있기 때문에 데이터의 유용성이 떨어진다. 따라서 여러 사용

자들로부터 수집된 데이터를 정제하여 유용성을 높인다. 예시를 이어가면, 여러 사용자로부터 수집한 크기가 n인 비트 배열의 각 비트에 대해 실제 1의 비율을 추정할 수 있다. 어떤 응답 k를 나타내는 비트 배역의 한 비트 n_k의 1의 개수가 x라고 할 때, 실제 응답 k의 비율은 $x \times p+(1-x) \times (1-p)$이다.

[0016] 본 발명에서는 지역 차분 프라이버시를 이용한 시계열 위치 데이터의 수집 및 분석을 진행한다. 데이터 수집 환경은 하나의 서버(데이터 수집가)와 복수의 사용자(데이터 제공자)로 구성된다. 사용자는 자신의 위치 데이터를 지역 차분 프라이버시를 만족하도록 변조한 후 주기적으로 자신의 변조된 위치를 데이터 수집가에게 전송한다. 데이터 수집가는 사용자들의 시계열 위치 데이터를 분석하여 빈발 방문 장소 및 이동 패턴 등의 정보를 얻을 수 있다.

선행기술문헌

특허문헌

[0018] (특허문헌 0001) 대한민국 공개특허 제2012-0090812호 (2012.08.17. 공개)

(특허문헌 0002) 대한민국 공개특허 제2013-0142934호 (2013.12.30. 공개)

(특허문헌 0003) 대한민국 등록특허 제1935528호 (2019.01.04. 공고)

비특허문헌

[0019] (비특허문헌 0001) [1] Ericsson Mobility report, June 2019

(비특허문헌 0002) [2] A. Narayanan and V. Shmatikov, “Robust deanonymization of large sparse datasets”, proc. IEEE Symp. Security Privacy, pp. 111-125. , May 2008

(비특허문헌 0003) [3] C. Dwork, “Differential privacy”, Encyclopedia of Cryptography and Security, 338-340, 2011

(비특허문헌 0004) [4] Duchi, John C and Jordan, Michael I and Wainwright, Martin J. “Local privacy and statistical minimax rates”. 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. 429-438. 2013

(비특허문헌 0005) [5] Erlingsson, Ulfar and Pihur, Vasyl and Korolova, Aleksandra. “Rappor: Randomized aggregatable privacy-preserving ordinal response”. Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. 1054-1067. 2014

(비특허문헌 0006) [6] Uber Differential Privacy

(비특허문헌 0007) [7] Ding, Bolin and Kulkarni, Janardhan and Yekhanin, Sergey. “Collecting telemetry data privately”. Advances in Neural Information Processing Systems. 3571-3580. 2017

(비특허문헌 0008) [8] Apple Differential Privacy Team. “Learning with privacy at scale”. Machine Learning Journal. 2017

(비특허문헌 0009) [9] Wang, Ning and Xiao, Xiaokui and Yang, Yin and Hoang, Ta Duy and Shin, Hyejin and Shin, Junbum and Yu, Ge. “PrivTrie: Effective Frequent Term Discovery under Local Differential Privacy”. 2018 IEEE 34th International Conference on Data Engineering(ICDE). 821-832. 2018

(비특허문헌 0010) [10] Shin, Hyejin and Kim, Sungwook and Shin, Junbum and Xiao, Xiaokui. “Privacy enhanced matrix factorization for recommendation with local differential privacy”. IEEE Transactions on Knowledge and Data Engineering 30.9: 1770-1782. 2018

(비특허문헌 0011) [11] Rabiner, Lawrence R. “A tutorial on hidden Markov models and selected applications in speech recognition”. Proceedings of the IEEE 77.2:257-286. 1989

(비특허문헌 0012) [12] Warner, Stanley L. “Randomized response: A survey technique for eliminating

evasive answer bias". Journal of the American Statistical Association 60.309:63-69. 1965

(비특허문헌 0013) [13] Cormode, Graham and Kulkarni, Tejas and Srivastava, Divesh. "Marginal release under local differential privacy". Proceedings of the 2018 International Conference on Management of Data. 131-146. 2018

(비특허문헌 0014) [14] Dwork, Cynthia and Naor, Moni and Pitassi, Toniann and Rothblum, Guy N. "Differential privacy under continual observation". Proceedings of the forty-second ACM symposium on Theory of computing. 715-724. 2010

(비특허문헌 0015) [15] Kellaris, Georgios and Papadopoulos, Stavros and Xiao, Xiaokui and Papadias, Dimitris. "Differentially private event sequences over infinite streams". Proceedings of the VLDB Endowment. 7.12:1155-1166. 2014

(비특허문헌 0016) [16] Cao, Yang and Yoshikawa, Masatoshi and Xiao, Yonghui and Xiong, Li. "Quantifying differential privacy under temporal correlations". 2017 IEEE 33rd International Conference on Data Engineering (ICDE). 821-832. 2017

(비특허문헌 0017) [17] Xiao, Yonghui and Xiong, Li. "Protecting locations with differential privacy under temporal correlations". Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 1298-1309. 2015

(비특허문헌 0018) [18] Dwork, Cynthia and Lei, Jing. "Differential privacy and robust statistics.". STOC.9:371-380. 2009

(비특허문헌 0019) [19] Fanti, Giulia and Pihur, Vasyl and Erlingsson, Ulfar. "Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries". Proceedings on Privacy Enhancing Technologies. 3:41-61. 2016

(비특허문헌 0020) [20] Kim, Jong Wook and Kim, Dae-Ho and Jang, Beakcheol. "Application of Local Differential Privacy to Collection of Indoor Positioning Data". IEEE Access. 2018

(비특허문헌 0021) [21] Jia, Jinyuan and Gong, Neil Zhenqiang. "Calibrate: Frequency Estimation and Heavy Hitter Identification with Local Differential Privacy via Incorporating Prior Knowledge". IEEE International Conference on Computer Communications. 2019

발명의 내용

해결하려는 과제

[0020] 본 발명이 이루고자 하는 기술적인 과제는 지역 차분 프라이버시를 만족하는 시계열 위치 데이터의 수집 및 분석 방법을 제공하는 것이다.

과제의 해결 수단

[0022] 본 발명의 실시예에 따른 시계열 데이터의 분석 서버는 복수의 단말들과 서버를 포함하고, 상기 복수의 단말들 각각은 사용자의 경로 데이터들이 지역 차분 프라이버시(Local Differential Privacy, LDP)를 만족하도록 변조(perturbation)하고 변조된 경로 데이터들을 상기 서버로 송신하고, 상기 서버는 상기 복수의 단말들 각각으로부터 수신된 상기 변조된 경로 데이터들을 이용하여 은닉 마르코프 모델(Hidden Markov Model, HMM)을 모델링하고, 상기 은닉 마르코프 모델의 전이 확률(Transition Probability)을 계산한다.

[0023] 본 발명의 실시예에 따른 시계열 데이터의 분석 서버는 복수의 단말들 각각으로부터 지역 차분 프라이버시를 만족하도록 변조된 경로 데이터들을 수신하는 데이터 수신부, 상기 변조된 경로 데이터들을 이용하여 마르코프 모델의 각 변수를 설정하여 상기 은닉 마르코프 모델을 모델링하는 모델 모델링부, 변조된 경로 데이터에 포함된 관측 값, 초기값이 설정된 전이 확률, 및 미리 결정된 관측 확률을 이용하여 사후 확률을 계산하는 사후 확률

도출부, 및 상기 사후 확률을 이용하여 상기 전이 확률을 계산하는 전이 확률 도출부를 포함한다.

발명의 효과

[0025] 본 발명의 실시 예에 따른 시계열 위치 데이터의 수집 및 분석 방법에 의할 경우, LDP를 만족하는 시계열 위치 데이터로부터 사용자들의 시작 위치 분포, 각 위치 별 사용자의 이동 경향성 등을 추정할 수 있는 효과가 있다.

도면의 간단한 설명

[0027] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 지역 차분 프라이버시를 적용한 시계열 데이터 수집을 은닉 마르코프 모델에 적용하는 모습을 보인다.

도 2는 지역 차분 프라이버시와 은닉 마르코프 모델의 각 요소를 도시한다.

도 3은 사후 확률을 구하기 위한 구체적인 알고리즘을 도시한다.

도 4는 전이 확률을 구하기 위한 구체적인 알고리즘을 도시한다.

도 5는 주어진 프라이버시 파라미터 조합과 경로 개수에 따른 제안 기법과 통계 기반 기법의 첫 번째 위치 분포 오차를 도시한다.

도 6은 본 발명의 일 실시예에 따른 수집 및 분석 시스템을 도시한다.

도 7은 도 6에 도시된 서버의 기능 블럭도이다.

발명을 실시하기 위한 구체적인 내용

[0028] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0029] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0030] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0031] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0032] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0033] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이

속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0034] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시 예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0036] 1.1. 시계열 위치 데이터 수집 및 분석

[0037] 운전자들의 시계열 위치 데이터를 수집하여 분석한다면 길안내 서비스, 대중 교통 계획 등에 활용할 수 있다. 운전자들의 시계열 위치 데이터는 민감한 데이터이기 때문에 원본을 수집할 경우 운전자의 집, 직장 등의 주소가 노출되는 등의 프라이버시 문제가 발생할 수 있다. 이때, 운전자가 자신의 위치를 지역 차분 프라이버시를 만족하는 알고리즘을 이용하여 변조한 후 데이터 수집가에게 전송한다면 개인의 프라이버시를 보호할 수 있다. 즉, 데이터 분석가는 개인의 위치를 정확히 알 수 없지만 분석을 통하여 운전자의 위치 분포, 빈발 위치 등을 알 수 있다.

[0038] 위치 데이터는 주로 GPS 좌표 혹은 로드 네트워크의 한 정점으로 표현된다. 이 데이터를 직접적으로 지역 차분 프라이버시를 만족하도록 변조하기는 어렵기 때문에 변조에 적합한 형태로 데이터를 인코딩한 후 데이터 변조를 하여야 한다. 전체 정점 수가 n개인 로드 네트워크에서 이동하는 사용자의 위치는 크기가 n인 비트 배열로 인코딩한 후 비트 배열의 각 비트에 확률 기반 응답을 적용할 수 있다.

[0040] 1.2. 데이터 수집 시 발생하는 상관관계

[0041] 데이터가 DP 또는 LDP를 만족한다는 의미는 데이터가 실제로 어떤 응답인지 정해진 확률 이상으로 알 수 없음을 의미한다. 그러나, 데이터 간의 상관관계에 의해 정해진 확률 이상으로 어떤 응답인지 추측할 수 있다. 이는 프라이버시 침해로 이어질 수 있다. 따라서, 데이터의 상관관계로 인한 프라이버시 침해를 막기 위하여 데이터의 상관관계를 고려한 DP 연구가 진행되었다. 데이터를 수집할 때 고려해야 할 상관관계는 두 가지가 있다. 먼저, 데이터 수집 과정에서 여러 속성을 동시에 수집할 때, 속성 간의 상관관계로부터 프라이버시 침해가 발생할 수 있다. 두 번째로, 한 사용자로부터 연속적으로 데이터를 수집할 때 시간에 따른 상관관계로부터 프라이버시 침해가 발생할 수 있다.

[0042] 본 발명에서는 한 사용자에게 위치 데이터 하나만 수집하기 때문에 여러 속성을 수집해서 발생하는 상관관계에 의한 프라이버시 침해는 발생하지 않는다. 하지만, 연속적인 데이터 수집에서 발생하는 상관관계는 발생하기 때문에 이에 대한 고려가 필요하다. 본 발명에서는 RAPPOR(Randomized aggregatable privacy-preserving ordinal response, 비특허문헌 [5] 참조)에서 사용한 2단계 확률 기반 응답 기법을 사용한다. 단기간의 프라이버시 예산과 장기간의 프라이버시 예산을 분리하여 시간에 의한 상관관계로부터 발생하는 프라이버시 손실을 제거하는 방법이다.

[0044] 1.3. 은닉 마르코프 모델(Hidden Markov Model, HMM)과 지역 차분 프라이버시 알고리즘

[0045] HMM(비특허문헌 [11] 참조)은 시계열 데이터 값을 추정하는 데에 최적화된 모델이다. HMM은 어떤 연속된 관측값(Observations) $X = \{x^1, x^2, \ldots, x^\tau\}$을 보고 이들이 어떤 은닉 상태(Hidden State) $Y = \{y^1, y^2, \ldots, y^\tau\}$에 있었는지 추측하고자 하는 모델이다. 한 은닉 상태 y_i에서 다른 은닉 상태 y_j로 전이할 확률은 전이 확률(Transition probability) a_{ij}라 하며, 한 은닉 상태 y에서 관측 값 x가 관측될 확률을 관측 확률 $b(y, x)$라 한다.

[0046] 지역 차분 프라이버시를 이용하여 사용자의 위치를 연속적으로 수집하는 알고리즘은 다음과 같은 흐름으로 진행된다. 먼저, 원본 위치를 각 사용자들이 변조하여 변조된 위치 데이터를 생성한다. 그 다음, 변조된 데이터를 데이터 수집가에게 보낸다. 이러한 LDP 알고리즘 요소들은 다음과 같이 HMM의 요소들로 볼 수 있다. 사용자의

원본 데이터를 은닉 상태로 보고, 변조된 데이터를 관측 값으로 본다. 원본 위치에서 다른 위치로 이동할 확률은 전이 확률로 표현할 수 있다. 변조된 데이터는 확률적으로 원본 데이터를 변조한 것이기 때문에, 원본 데이터로부터 변조된 데이터가 생성될 확률을 구할 수 있다. 즉, 관측 확률을 LDP를 만족하는 변조 알고리즘으로 볼 수 있다. 이를 정리하면 표 1과 같다. 또한, 본 발명에서는 관측 확률이 데이터 변조 알고리즘으로부터 정확한 값을 구할 수 있기 때문에 전이 확률만을 최적화한다.

표 1

[0047]

시계열 위치 데이터 분석	은닉 마르코프 모델
사용자의 실제 위치	은닉 상태
변조된 위치 값	관측 값
데이터 변조 알고리즘	관측 확률
사용자의 이동성	전이 확률

[0048] 본 발명에서는 지역 차분 프라이버시를 이용한 시계열 위치 데이터의 수집 및 분석 방법을 제안한다. (1) 데이터를 수집할 때 발생할 수 있는 상관관계를 고려하여 LDP를 만족하도록 시계열 위치 데이터를 수집하고, (2) 수집한 데이터에 대해 HMM을 이용하여 실제 위치 분포, 사용자의 이동성 등을 분석한다. HMM을 이용한 분석은 단순 통계 기법에서 구할 수 없었던 전이 확률 등의 경로 데이터의 시계열적 특성을 이용할 수 있도록 한다. 추가적으로, 두 종류의 로드 네트워크를 이용한 실험을 통하여 제안 기법의 우수성을 보인다.

[0050] 이하에서는, 본 발명에서 사용하는 사전 지식에 대해 기술하기로 한다.

[0051] 2.1. 차분 프라이버시와 지역 차분 프라이버시

[0052] 차분 프라이버시는 프라이버시 데이터 배포 환경에서 이용하는 모델로 가장 강력한 프라이버시 모델이다. 차분 프라이버시는 공격자가 질의 결과를 보고 특정 개인의 존재 유무를 정해진 신뢰도 이상으로 추정할 수 없도록 하여 프라이버시를 보호한다. 차분 프라이버시의 정의는 다음과 같다.

[0053] 정의 1. 차분 프라이버시 (Differential Privacy, DP)

[0054] 최대 하나의 레코드만 다른 데이터베이스 D와 D'에서 함수 A에 대한 결과 O가 다음 식을 만족하면 차분 프라이버시를 만족한다.

[0055] $$\Pr[A(D) = O] \leq e^{\epsilon} \times \Pr[A(D') = O]$$

[0056] 차분 프라이버시의 정의에 따라 최대 하나만 다른 데이터베이스를 알고 있어야 하기 때문에 DP를 적용하기 위해서는 전체 데이터베이스를 알고 있어야 한다. 데이터를 수집하는 환경에서는 한 사용자 입장에서 다른 사용자가 어떤 데이터를 보내는지 알 수 없기 때문에 전체 데이터셋을 알 수 없다. 따라서, 데이터 수집 환경에서 차분 프라이버시를 그대로 사용하기 어렵다. 지역 차분 프라이버시는 데이터 수집 환경에서 차분 프라이버시를 적용하기 위해 제안된 모델로 사용자가 데이터 수집가를 믿을 수 없는 환경에서 프라이버시를 보호하기 위하여 사용한다. 지역 차분 프라이버시의 정의는 다음과 같다.

[0057] 정의 2. 지역 차분 프라이버시 (Local Differential Privacy, LDP)

[0058] 모든 데이터 제공자의 데이터쌍 v_i, v_j에 대해 함수 A에 대한 결과 O가 다음 식을 만족하면 차분 프라이버시를 만족한다.

[0059] $$Pr[A(\ v_i) = O] \leq e^{\epsilon} \times Pr[A(v_j) = O]$$

[0060] 지역 차분 프라이버시를 만족하는 알고리즘을 통해 데이터를 수집하게 되면 데이터 제공자가 v_i를 보냈는지 v_j를 보냈는지 ϵ에 따라 정해진 확률 이상으로 알 수 없기 때문에 데이터 제공자의 프라이버시를 보호할 수 있다.

[0061] 지역 차분 프라이버시를 만족하는 가장 간단한 형태의 데이터 변조 알고리즘은 확률 기반 응답(Randomized response, 비특허문헌 [12] 참조)이다. 확률 기반 응답은 이진 응답(대답의 경우의 수가 2가지인 응답)에서 p의

확률로 원래의 응답을, 1-p의 확률로 반대 응답을 한다. 이때, 확률 기반 응답은 $\epsilon = ln\left(\frac{x}{1-x}\right)$을 만족한다.

[0062] RAPPOR은 지역 차분 프라이버시를 만족하는 데이터 수집 알고리즘으로 해시 함수를 이용하여 응답을 기록한 비트 배열의 각 비트에 확률 기반 응답을 적용하여 지역 차분 프라이버시를 만족하였다. 본 발명의 데이터 수집은 RAPPOR와 같이 데이터를 원본 비트 배열로 인코딩한 후 두 번의 확률 기반 응답 처리를 한다. 사용자의 원본 비트 배열 S는 다음과 같이 정의한다.

[0063] 정의 3. 사용자의 원본 비트 배열 S

[0064] 사용자가 n개의 응답의 경우의 수 중 k번째 응답 n_k를 응답을 한다고 할 때, 사용자의 원본 비트 배열 S의 i번째 비트 S[i]는 다음 조건을 만족한다.

[0065] $$S[i] = \begin{cases} 1, if\ \ k = i \\ 0\ \ otherwise \end{cases}$$

[0066] RAPPOR에서 응답의 경우의 수가 많아질 경우 원본 비트 배열의 크기가 커지는 문제를 해결하기 위하여 블룸 필터 등의 기법을 이용하지만 본 발명에서는 별도의 기법을 이용하지 않고 응답의 경우의 수 만큼의 크기를 가진 원본 비트 배열을 생성한다.

[0067] 다음으로, S의 각 비트에 확률 기반 응답 처리를 하여 영구 확률 기반 응답 PR을 생성한다. 각 사용자들은 자신의 원본 신호 하나에 대하여 하나의 영구 확률 기반 응답을 저장하여 즉석 확률 기반 응답에 이용한다. 영구 확률 기반 응답 PR은 다음과 같이 정의한다.

[0068] 정의 4. 영구 확률 기반 응답 PR (Permanent Randomized Response)

[0069] 사용자의 원본 비트 배열 S에

[0070] 대한 영구 확률 기반 응답 PR의 i번째 비트 PR[i]는 다음 조건을 만족한다.

[0071] $$PR[i] = \begin{cases} 1, & with\ probability\ f/2 \\ 0, & with\ probability\ f/2 \\ S[i], & with\ probaiblity\ 1-f \end{cases}$$

[0072] 사용자는 데이터 수집가에게 데이터를 전송할 때 자신의 원본 위치 데이터 대신 영구 확률 기반 응답 결과를 이용하기 때문에 사용자가 전송한 데이터의 프라이버시가 침해당하더라도 영구 확률 기반 응답이 가지고 있는 프라이버시 수준을 보장한다.

[0073] 사용자는 데이터 수집가에게 영구 확률 기반 응답에 한번 더 확률 기반 처리를 한 즉석 확률 기반 응답을 보낸다. 서버에 지속적으로 같은 응답을 보내더라도 영구 확률 기반 응답 처리된 응답을 이용하여 즉석 확률 기반 응답을 생성하여 보내기 때문에 프라이버시를 보호할 수 있다. 즉석 활률 기반 응답 IR은 다음과 같이 정의한다.

[0074] 정의 5. 즉석 확률 기반 응답 IR (Instantaneous Randomized Response)

[0075] 사용자의 영구 확률 기반 응답 PR에 대한 즉석 확률 기반 응답 IR의 i번째 비트 IR[i]는 다음 조건을 만족한다.

[0076] $$P(IR[i] = 1) = \begin{cases} q, if\ \ PR[i] = 0 \\ p, if\ \ PR[i] = 1 \end{cases}$$

[0077] 데이터 수집가는 결과적으로 IR을 수집하고 이를 분석하게 된다. 영구 확률 기반 응답에 사용하는 파라미터 f, 즉석 확률 기반 응답에 사용하는 파라미터 p, q를 알고 있을 때 수집한 신호에 따른 확률은 다음과 같다.

[0078] 정리 1. 즉석 확률 기반 응답을 이용한 사용자의 확률 분포 추정

[0079] 수집한 응답의 수를 n_{total}, PR[i]=1인 응답의 수를 n_i라고 할 때, S[i]=1인 응답의 추정 수 $num_{est}(i)$는 다음과 같다.

[0080] $$num_{est}(i) = \frac{1}{1-f} \times \left(\frac{n_i - p \times n_{total}}{q-p} - \frac{f \times n_{total}}{2}\right)$$

[0081] 따라서, 각 응답 별 사용자의 비율 $Density_{est}(i)$는 다음과 같다.

[0082] $$Density_{est}(i) = \frac{num_{est}(i)}{\sum num_{est}(n)}$$

[0083] 정리 1과 같이 수집된 데이터를 바탕으로 사용자의 위치 분포를 추정할 수 있다. 그러나, 이는 수집한 데이터의 시공간적 상관관계를 고려하지 않기 때문에 데이터의 특성을 고려한 분석을 할 수 없다.

[0085] 2.2. 데이터의 상관관계와 차분 프라이버시

[0086] 차분 프라이버시를 만족하더라도 데이터의 상관관계에 따라 프라이버시 손실(privacy loss)이 발생할 수 있다. 프라이버시 손실의 정의는 다음과 같다.

[0087] 정의 6. 프라이버시 손실 (Privacy Loss)

[0088] 차분 프라이버시, 지역 차분 프라이버시를 이용한 결과에 대해 보장된 ϵ 보다 높은 확률로 실제 결과를 추측할 수 있을 경우 프라이버시 손실이 발생했다고 정의한다.

[0089] 데이터의 상관관계는 여러 상황에서 발생할 수 있고 각 상황에 따른 프라이버시를 보호하기 위한 연구가 진행되고 있다.

[0090] 한 번에 여러 속성에 대한 수집을 동시에 진행할 경우 속성 사이의 상관관계가 발생한다. 비특허문헌 [13]에서 여러 속성을 동시에 수집할 때 프라이버시와 이들의 상관관계를 분석하는 연구를 진행하였다.

[0091] 지속적인 데이터 수집에 의한 시간적 상관관계가 발생할 수 있다. 비특허문헌 [14], [15]에서 연속적인 데이터 수집 환경의 프라이버시 예산의 관리 방법을 제안하였다.

[0092] 위치 데이터를 연속적으로 수집할 경우 시간적 상관 관계 외에 위치 데이터의 특성으로 인하여 위치적 상관관계가 발생할 수 있다. 비특허문헌 [16]은 시계열 위치 데이터의 총계처리 과정에서 데이터의 상관관계로 발생하는 프라이버시 손실에 관한 논문이다. 비특허문헌 [17]은 δ-location set을 이용하여 위치 상관관계를 고려한 차분 프라이버시 모델을 제안하였다. 위 논문들은 HMM을 이용하지만 모든 파라미터를 이미 알고 있다고 가정한 상태에서 데이터 수집 및 분포를 진행한다. 위와 같은 가정은 데이터 수집 환경에 적합하지 않다. 왜냐하면, HMM의 모든 파라미터를 알고 있다면 원본 데이터의 분포와 특성을 알고 있는 상태라고 볼 수 있다. 따라서, 데이터를 수집할 필요성이 떨어진다. 본 발명에서는 변조된 값을 이용하여 HMM의 파라미터를 학습하여 데이터 분석을 진행하기 때문에 관련 연구들과 차이를 보인다.

[0094] 2.3. EM 알고리즘을 이용하여 LDP를 적용한 수집 데이터 분석

[0095] LDP를 만족하는 데이터 분석 과정은 지역 차분 프라이버시를 만족하도록 데이터를 변조하는 부분과 변조된 데이터를 분석하는 부분으로 나눌 수 있다. 데이터를 프라이버시 침해로부터 안전하게 수집하는 것도 중요하지만 수집한 데이터로부터 유용한 정보를 얻어내는 것도 중요하다. 비특허문헌 [4]에서는 데이터가 변조될 확률을 이용하여 데이터의 실제 분포를 추측하는 방법을 제안한다. 이 외에도 다양한 통계적 기법을 이용하여 수집한 데이터에 대한 분석방법이 제안됐다(비특허문헌 [18] 참조).

[0096] EM 알고리즘은 관측할 수 있는 확률 변수 X, 관측할 수 없는 확률 변수 Z, 모수 Θ가 있을 때, 최대 가능도를 구하는 알고리즘이다. 이 알고리즘은 관측할 수 있는 값을 이용하여 실제 구하고자 하는 값을 추정할 수 있다. 이를 LDP를 이용한 데이터 수집 환경에 적용해본다면 관측할 수 있는 확률 변수 X를 변조된 수집 값, 관측할 수 없는 확률 변수 Z, 모수 Θ가 실제 분석하고자 하는 사용자의 응답 분포라고 할 수 있다. 도 1은 지역 차분 프라이버시를 적용한 시계열 데이터 수집을 은닉 마르코프 모델에 적용하는 모습을 보인다.

[0097] 데이터 분석에 EM 알고리즘을 활용하는 연구가 진행되었다. 비특허문헌 [19], [20]은 조건부 확률을 이용한 EM 알고리즘을 이용하여 확률 기반 응답으로 변조된 데이터의 분포를 추정하는 법을 제안했다. 비특허문헌 [21]은 노이즈 분포, 관측 값의 사전 확률을 특정 확률 분포로 가정한 후 실제 값의 분포를 추정하는 방법을 제안하였다.

[0099] 이하에서는, 제안 기법에 대한 설명을 한다. 제안기법은 사용자들의 LDP를 만족하도록 변조된 경로 데이터를 수집한 후 경로 데이터의 시공간 상관관계를 고려하여 사용자들의 위치 분포와 이동경향성을 분석한다.

[0100] 본 발명에서 제안하는 기법은 클라이언트와 서버를 엔티티로 포함한다. 클라이언트는 원본 데이터에 LDP를 만족하는 변조 알고리즘을 적용하여 서버로 송신한다. 사용자가 보낼 원본 데이터는 자신의 로드 네트워크 상의 위치이다. 자신의 위치를 정해진 시간마다 서버로 전송하여 경로 데이터를 전송한다.

[0101] 변조 알고리즘은 RAPPOR에서 제안된 방법인 원본 비트 배열을 이용하여 영구 확률 기반 응답을 생성한 후 서버에 데이터를 전송할 때 즉석 확률 기반 응답을 생성하여 보내는 방식이 이용될 수 있다. 클라이언트 부분에서 이뤄지는 상세한 과정은 후술하기로 한다. 서버는 여러 클라이언트가 보낸 데이터를 수집 후 분석한다. 이 때 클라이언트로부터 수집한 데이터를 HMM으로 모델링하여 분석을 진행한다. HMM은 은닉 상태, 관측 값, 전이 확률, 관측 확률로 구성된다. LDP를 이용한 데이터 수집 환경에서 사용자의 실제 위치는 직접 관측할 수 없기 때문에 은닉 상태로 모델링한다. 사용자의 변조된 데이터는 사용자의 실제 위치에 따라 확률적으로 생성되는 값이기 때문에 은닉 상태에서 관측되는 값으로 본다. 마지막으로 실제 위치에서 변조된 위치가 발생할 확률은 데이터 변조 확률을 이용할 수 있고, 이는 HMM의 관측 확률로 모델링할 수 있다. 이때, 관측 확률은 정확한 값으로 이미 정확한 관측 확률을 알고 있다고 가정하기 때문에 학습 과정에서 전이 확률만을 업데이트하며 학습을 진행한다. 자세한 모델 학습 과정은 후술한다.

[0103] 3.1. 문제정의

[0104] 제안 기법은 위치를 로드 네트워크(Road network)의 정점으로 표현한다. 로드 네트워크는 n개의 정점과 m개의 간선으로 이루어져 있다. 따라서, 로드 네트워크는 정점 집합 {N_1, N_2, …, N_n}와 간선 집합 {E_1, E_2, …, E_m}으로 표현할 수 있다. 사용자는 정점에 위치하며 정점 사이를 이동할 때 간선을 따라 이동한다.

[0105] 사용자의 원본 위치는 길이가 n인 비트 배열 S로 영구 확률 기반 응답을 PR, 즉석 확률 기반 응답을 IR로 표현하자.

[0106] 경로 데이터는 시간에 따라 이동하는 위치를 표현한 데이터이다. 따라서, 원본 경로 데이터는 로드 네트워크의 정점 집합이고 정점 집합은 원본 비트 배열 S의 집합으로 표현할 수 있다. 본 발명에서 경로의 개수(즉, 클라이언트의 수)는 τ이고, 로드 네트워크에서 이동할 때 방문한 모든 위치에서 응답을 보낸다고 가정한다.

[0107] 사용자 u의 변조된 경로 데이터는 즉석 확률 응답 IR의 집합 $X = \{x_u^1, x_u^2, \ldots, x_u^{t_u}\}$으로 표현한다. t_u는 사용자 u의 경로의 길이이다. 즉, x_u^t는 사용자 u의 경로 중 t 번째 IR이다. 사용자 u의 원본 경로 데이터는 $Y_u = \{y_u^1, y_u^2, \ldots, y_u^{t_u}\}$으로 나타낸다. 즉, y_u^t는 사용자 u의 경로 중 t 번째 원본 위치 배열이다.

[0108] 서버는 사용자의 변조된 경로 데이터를 수집하여 분석에 이용한다. 수집한 전체 경로의 수를 τ라고 하자. 수집한 모든 변조된 경로 데이터 O는 X의 집합 $\left\{ \{x_1^1, x_1^2 \ldots, x_1^{t_1}\}, \ldots, \{x_\tau^1, x_\tau^2 \ldots, x_\tau^{t_\tau}\} \right\} = \{X_1, X_2, \ldots, X_\tau\}$으로 표현한다. 이때, 각 X에 대응하는 원본 비트 배열 H는 $\left\{ \{y_1^1, y_1^2 \ldots, y_1^{t_1}\}, \ldots, \{y_\tau^1, y_\tau^2 \ldots, y_1^{t_\tau}\} \right\} = \{Y_1, Y_2, \ldots, Y_\tau\}$로 나타낸다. HMM의 전이 확률 a_{ij}는 $P(y^{t+1} = N_j | y^t = N_i)$로 정의되면 이는 시간 t에 정점 N_i에 있을 때 시간 t+1에 N_j로 이동할 확률을 의미한다. 관측 확률 $b(x^t, i)$는 $P(x^t | y^t = N_i)$과 같이 정의하며 N_i가 x^t로 변조될 확률을 의미한다.

[0109] 주어진 O, a, b를 이용하여 O를 표현할 수 있는 최적의 a를 찾는 것이 제안 기법의 목표이다.

[0111] 3.2. 클라이언트

[0112] 사용자의 경로 데이터를 가장 안전하게 수집하는 방법은 모든 발생할 수 있는 경로의 수를 응답의 경우의 수로

보고 수집하는 것이다. 그러나, 발생할 수 있는 경로의 수는 $n^{\tau_{Max}}$ 이고 발생할 수 있는 경로의 최대 길이를 알 수 없기 때문에 사실상 무한한 경우의 수가 있다고 볼 수 있다. 응답의 경우의 수가 많아질수록 분석 결과의 품질이 떨어지는 LDP 보호 기법의 특성 상 모든 경로의 수를 고려하여 데이터를 수집하는 것은 불가능하다. 따라서, 본 발명에서는 클라이언트 u는 경로를 따라 움직이면서 t_u 시간 동안 일정한 시간마다 자신의 위치에 대한 IR을 생성하여 서버에게 전송한다. IR을 생성하기 위해서 위치에 대한 원본 신호 S를 생성하고 S에 대한 PR을 생성한 후 저장한다. 두 단계에 걸친 데이터 변조 과정을 통하여 연속적으로 데이터를 수집하는 환경에서 프라이버시를 보장한다.

[0113] 정리. 사용자가 자신의 위치 데이터에 대한 IR을 생성하여 서버에 전송한다면 이는 LDP를 만족한다.

[0114] 증명. 사용자가 원본 위치 n_k에서 S를 생성한 후 변조하여 PR을 생성했다고 하자. S는 $S[n_k]=1$이고 나머지 비트가 모두 0인 비트 배열이고, IR은 0과 1로 구성되어 있는 길이가 n인 비트 배열이다. LDP를 만족하기 위해서 $\frac{\Pr[S[n_k]=1 \mid IR_1]}{\Pr[S[n_k]=1 \mid IR_2]} \leq e^{\epsilon}$ 를 만족해야 한다. e^{ϵ} 의 최댓값을 구하기 위해서 $\Pr[S[n_k]=1 \mid IR]$가 최대가 되는 IR_1과 최소가 되는 IR_2를 구해야 한다. 이를 만족하는 IR_1은 {IR[n_k]=1, else 0}이고, IR_2는 {IR[n_k]=0, else 1}이다.

[0115] 본 발명에서 변조된 데이터를 분석할 때, 사용자의 움직임이 마르코프 과정을 따른다고 가정한다. 이때, LDP를 만족하기 위해선 $\frac{\Pr[S[n_k]=1 \mid IR_1]}{\Pr[S[n_k]=1 \mid IR_2]} \leq e^{\epsilon}$ 를 만족해야 한다.

[0117] 3.3. 서버

[0118] 서버는 여러 클라이언트가 전송한 신호를 수집한 후 분석한다. 본 발명에서는 로드 네트워크 환경에서 수집되는 위치 데이터 간의 상관관계를 고려한 데이터 분석을 통하여 단순 통계 기반 분석 결과보다 유용한 정보를 획득한다.

[0120] 3.3.1. 데이터의 상관관계를 고려한 데이터 분석 방법

[0121] 데이터의 상관관계를 고려한 데이터 분석을 위하여 지역 차분 프라이버시 알고리즘을 은닉 마르코프 모델로 모델링한다. 지역 차분 프라이버시 알고리즘의 각 요소는 표 1과 같이 은닉 마르코프 모델의 요소와 대응시킬 수 있다.

[0122] 모델링한 데이터 분석 알고리즘은 일반적인 은닉 마르코프 모델과 다르게 관측 확률을 정확하게 알 수 있다. 따라서, 분석 과정에서는 전이 확률만을 업데이트한다. 은닉 마르코프 모델의 학습을 위하여 가장 널리 알려진 알고리즘 중 하나인 Baum-welch 알고리즘이 이용될 수 있다. 도 2에는 지역 차분 프라이버시와 은닉 마르코프 모델의 각 요소가 도시되어 잇다.

[0123] 알고리즘을 이용하여 전이 확률을 구하는 과정은 두 단계로 나눠진다. 먼저 관측 값, 전이 확률 및 관측 확률을 이용하여 수집한 신호 데이터에 대한 사후 확률(Posterior Probability)을 계산한다. 사후 확률을 계산하기 쉽도록 전향 확률 α와 후향 확률 β를 계산한다. α와 β는 다음과 같다.

$$\alpha_j^t = \left(\sum_i \alpha_i^{t-1} a_{ij}\right) \mathrm{b_j}(x^t),$$

[0124] $$\beta_j^t = \left(\sum_i \beta_j^{t+1} a_{ij}\right) \mathrm{b_j}(x^{t+1})$$

[0125] 그 다음, 전향 확률, 후향 확률, 주어진 관측 값과 전이 확률, 관측 확률을 이용하여 전이 확률 업데이트에 이용할 사후 확률 γ, ξ를 계산한다. 사후 확률을 구하기 위한 구체적인 알고리즘은 도 3에 도시되어 있다.

[0126] 수집된 신호 데이터들에 대해서 사후 확률 γ, ξ을 이용하여 전이 확률을 업데이트한다. 위 과정을 전이 확률 a가 수렴할 때까지 반복한다. 업데이트되는 일련의 a 값들이 일정 범위(예컨대, c 번째로 도출된 a의 값에서 미리 정해진 변동폭 내에 c+1 번째로 도출된 a의 값 내지 c+d 번째로 도출된 a의 값이 포함되는 경우)에 분포하는 경우에도 수렴하는 것으로 볼 수도 있다. 전이 확률을 구하기 위한 구체적인 알고리즘은 도 4에 도시되어 있다.

[0127] 위 과정을 통하여 전이 확률 a를 구할 수 있다. 전이 확률을 통하여 사용자들의 시작 위치 분포, 각 위치 별 사용자의 이동 경향성 등을 추정할 수 있다.

[0129] 4. 성능 평가

[0130] 제안 기법의 성능 평가을 위하여 두 가지 로드 네트워크에서 발생한 시계열 데이터를 분석하였다. 분석에 사용한 경로의 수에 따른 성능 평가, 프라이버시 파라미터 f, p, q에 따른 성능 평가를 진행하였다. 기법의 성능은 경로의 출발 위치 분포와 전이 확률을 측정하여 원본 데이터의 출발 위치 분포와 전이 확률과의 평균 절대 오차를 구하여 비교하였다.

[0132] 4.1. 실험 환경

[0133] 실험에 사용한 데이터셋과 변수에 대해 설명한다. 데이터셋은 가상 로드 네트워크 하나와 실제 로드 네트워크 하나에서 가상 경로를 생성하여 실험을 진행하였다.

[0134] 실험에 사용하는 변수는 경로의 수 τ, 프라이버시 파라미터 f, p, q가 있다.

[0136] 4.1.1. 데이터셋

[0137] 실제 로드 네트워크 환경에서는 원본 시계열 데이터는 구하기 힘들기 때문에 본 발명에서는 주어진 전이 확률에 따라 로드 네트워크에서 이동하는 가상의 경로를 생성한다.

[0138] 본 실험에서는 Simple Graph와 Seoul Subway라는 두 가지 로드 네트워크를 이용한다. Simple Graph는 실험을 위하여 생성한 가상의 로드 네트워크이며 10개의 노드와 10개의 간선을 가진다. Simple Graph의 전이 확률은 임의로 설정하였다. Seoul Subway는 2017년 기준 수도권 지하철 노선의 1-9호선을 이용하여 생성하였으며 375개의 정점과 421개의 간선을 가진다. Seoul Subway의 전이 확률은 2017년 7월 기준 이용 통계를 이용하여 설정하였다. 실험에 사용하는 경로는 로드 네트워크의 전이 확률을 고려하여 생성하였다.

[0140] 4.1.2. 변수

[0141] 실험에 사용하는 경로의 수 τ는 1,000, 10,000, 100,000을 사용하였다. 이 때 Simple Graph에서 생성한 평균 경로 길이는 3이고 Seoul Subway에서 생성한 평균 경로 길이는 10이다.

[0142] 데이터 변조에 필요한 3개의 변수 f, p, q는 4가지 조합을 사용하였다. 편의상 각 변수 조합에 번호를 붙여 표 2와 같이 정의했다.

표 2

[0143]

번호	f	p	q	위치 하나 당 ϵ
1	0	0.5	0.75	ln 3
2	0	0.25	0.75	ln 9
3	0.5	0.5	0.75	ln 1.71
4	0.5	0.25	0.75	ln 2.78

[0145] 4.2. 실험 결과

[0146] Simple Graph와 Seoul Subway에서 생성한 1,000, 10,000, 100,000개의 경로를 각각 변수 조합 1, 2, 3, 4를 이용하여 수집한 후 분석을 진행하였다. 먼저, 경로의 첫번째 위치 분포를 통계적 기법과 제안 기법을 이용하여 분석한 후 실제 첫 번째 위치 분포와 비교하여 오차를 측정하였다. 다음으로, 경로를 생성할 때 이용한 전이 확률과 제안 기법을 이용하여 추정한 전이 확률 사이의 오차를 측정하였다. 위치 분포의 오차 Err_{π}는 다음과 같이

측정한다.

[0147] $$Err_{\pi} = \frac{\sum|\hat{\pi}_i - \pi_i|}{n}$$

[0148] $\hat{\pi}_i$는 수집한 경로를 바탕으로 추측한 위치 분포, π_i는 실제 위치 분포, n은 로드 네트워크의 정점 수이다.

[0149] 확률 기반 응답에 의해서 실행할 때마다 결과가 달라지기 때문에 10번 시행한 후 평균 오차를 이용하였다.

[0150] 실험 결과, 제안 기법이 기존 기법인 통계 기반 방법보다 전반적으로 더 정확한 분포 추정 능력을 보였다. 경로의 수가 많아질수록 제안 기법과 통계 기반 기법 모두 오차가 줄어드는 경향을 보인다. 또한, ϵ이 낮을수록, 즉 프라이버시 보호가 강력할수록 기존 기법 대비 제안 기법이 더욱 뛰어난 성능을 보임을 알 수 있었다.

[0151] 전이 확률의 경우, 기존 기법에서 측정이 불가능하기 때문에 경로를 생성할 때 이용한 전이 확률과 추정 전이 확률을 비교하였다. 실험 결과 실제 전이 확률과 제안 기법을 이용하여 추정한 전이 확률이 크게 차이 나지 않음을 확인할 수 있었다. 도 5에는 주어진 프라이버시 파라미터 조합과 경로 개수에 따른 제안 기법과 통계 기반 기법의 첫 번째 위치 분포 오차가 도시되어 있다. 즉, 도 5에서 (a-1) 내지 (a-4)는 Simple Graph의 데이터셋을 이용한 결과이고, (b-1) 내지 (b-4)는 Seoul Subway 데이터셋을 이용한 결과이다. 또한, (a-1) 내지 (a-4) 각각은 주어진 파라미터 조합 각각에 대한 실험 결과이며, (b-1) 내지 (b-4) 각각은 주어진 파라미터 조합 각각에 대한 실험 결과이다. 또한, 푸른색의 막대 그래프는 기존의 HMM 기법을 이용한 실험 결과를, 주황색의 막대 그래프는 기존의 통계적 기법을 이용한 실험 결과를, 회색의 막대 그래프가 제안 기법을 이용한 실험 결과를 나타낸다.

[0153] 도 6은 본 발명의 일 실시예에 따른 시계열 위치 데이터의 수집 및 분석 시스템이다.

[0154] 도 6을 참조하면, 수집 및 분석 시스템(10)은 복수의 단말들(100)과 서버(300)를 포함한다. 수집 및 분석 시스템(10)은 데이터 수집 시스템, 데이터 분석 시스템 등 다양한 명칭으로 명명될 수도 있고, 서버(300)는 데이터 수집 서버, 데이터 분석 서버 등 다양한 명칭으로 명명될 수도 있다.

[0155] 수집 및 분석 시스템(10)에 포함되는 복수의 단말들(100) 각각과 서버(300)는 각종 연산 처리 및 신호 생성이 가능한 적어도 하나의 전자 장치를 포함할 수 있다. 여기서, 적어도 하나의 전자 장치는 프로세서 및/또는 프로세서가 설치된 컴퓨팅 장치를 포함할 수 있다. 여기서, 프로세서는, 예를 들어, 중앙 처리 장치(CPU, Central Processing Unit), 마이크로 컨트롤러 유닛(MCU, Micro Controller Unit), 마이컴(Micom, Micro Processor), 애플리케이션 프로세서(AP, Application Processor), 전자 제어 유닛(ECU, Electronic Controlling Unit), 그래픽 처리 장치(GPU, Graphic Processing Unit) 및/또는 각종 연산 처리 및 제어 신호의 생성이 가능한 처리 장치 등을 포함할 수 있다. 이들 처리 장치는 예를 들어 하나 또는 둘 이상의 반도체 칩 및 관련 부품을 이용하여 구현될 수 있다. 또한, 컴퓨팅 장치는, 예를 들어, 데스크톱 컴퓨터, 랩톱 컴퓨터, 서버용 컴퓨터, 스마트 폰, 태블릿 피씨, 스마트 시계, 두부 장착형 디스플레이(HMD, Head Mounted Display) 장치, 휴대용 게임기, 내비게이션 장치, 개인용 디지털 보조기(PDA, Personal Digital Assistant), 인공지능 스피커 장치, 디지털 텔레비전, 셋톱 박스, 로봇, 가전 기기, 기계 장치 및/또는 이외 정보 처리 기능을 수행할 수 있는 적어도 하나의 전자 장치를 포함할 수 있다.

[0156] 복수의 단말들(100) 각각은 사용자의 위치 데이터에 기초하여 원본 신호 S를 생성하고, S에 대한 PR을 생성할 수 있다. 또한, 복수의 단말들(100) 각각은 PR로부터 IR을 생성하고 생성된 IR을 서버(300)로 송신할 수 있다. 여기서, IR은 서버(300)로 송신 직전에 생성될 수도 있으며, 실시예에 따라 미리 생성되어 저장된 후 송신될 수도 있다. 복수의 단말들(100) 각각이 송신하는 IR은 지역 차분 파리이버시를 만족한다.

[0157] 서버(300)는 복수의 단말들(100) 각각으로부터 IR을 수신하고, 수신된 IR에 대한 분석 동작을 수행할 수 있다. 즉, 서버(300)는 수집되는 위치 데이터 간의 상관관계를 고려한 데이터 분석을 통하여 단순 통계 기반 분석 결과보다 유용한 정보를 획득할 수 있다.

[0159] 도 7은 도 6에 도시된 서버의 기능 블럭도이다.

[0160] 도 7을 참조하면, 서버(300)는 데이터 수신부(310), 모델 모델링부(320), 사후 확률 도출부(330), 및 전이 확률

도출부(340)를 포함한다. 실시예에 따라, 서버(300)는 정보 추정부(350) 및/또는 저장부(360)를 더 포함할 수도 있다.

[0161] 서버(300)는 도 7에 도시된 바와 같이 저장부(360)와 데이터 송수신이 가능하도록 직접 또는 간접적으로 연결되어 있을 수 있다. 저장부(360)는 데이터 수신부(310)에 의해 수신된 데이터, 모델 모델링부(320)에 의한 모델링 결과, 사후 확률 도출부(330)에 의한 사후 확률 도출 동작 중 생성되는 데이터, 전이 확률 도출부(340)에 의한 전이 확률 도출 동작 중 생성되는 데이터, 정보 추출부(350)에 의한 정보 추출 동작 중 생성되는 데이터, 및/또는 서버(300)의 제어, 연산 또는 처리 과정에서 생성된 적어도 하나의 데이터를 일시적으로 또는 비일시적으로 저장할 수 있다. 저장부(360)는 주기억장치 및 보조기억장치 중 적어도 하나를 포함할 수 있으며, 주기억장치는 롬(ROM) 및/또는 램(RAM) 등을 포함할 수 있고, 보조기억장치는 플래시 메모리 장치, SD(Secure Digital) 카드, 솔리드 스테이트 드라이브(SSD, Solid State Drive), 하드 디스크 드라이브(HDD, Hard Disc Drive) 컴팩트 디스크(CD), 디브이디(DVD) 및/또는 레이저 디스크 등을 포함할 수 있다.

[0162] 서버(300)는 저장부(360)에 저장된 애플리케이션을 구동시켜, 모델 모델링, 사후 확률 도출, 전이 확률 도출, 정보 추정 등 다양한 동작을 수행할 수도 있다. 여기서, 애플리케이션은, 설계자에 의해 직접 작성되어 저장부(360)에 저장 또는 갱신된 것일 수도 있고, 또는 유선 또는 무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망 등을 통하여 획득 또는 갱신된 것일 수도 있다.

[0163] 데이터 수신부(310)는 복수의 단말들(100) 각각으로부터 지역 차분 프라이버시를 만족하는 IR을 수신하고, 수신된 IR을 저장부(360)에 저장할 수 있다.

[0164] 모델 모델링부(320)는 은닉 마르코프 모델을 이용하여 전이 확률을 도출하기 위하여 지역 차분 프라이버시 알고리즘을 은닉 마르코프 모델로 모델링한다. 즉, 지역 차분 프라이버시 알고리즘의 각 요소를 표 1 및 도 2와 같이 은닉 마르코프 모델의 요소에 대응시킬 수 있다. 보다 구체적으로, 모델 모델링부(320)는 데이터 수신부(310)에 의해 수신된 IR을 이용하여 은닉 마르코프 모델의 각 변수를 설정할 수 있다.

[0165] 사후 확률 도출부(330)는 전이 확률의 업데이트(또는 학습)에 이용할 사후 확률을 계산(또는 도출)할 수 있다. 즉, 사후 확률 도출부(330)는 관측 값, (초기값이 설정된) 전이 확률과 (값이 미리 결정된) 관측 확률을 이용하여 수집한 신호 데이터에 대한 사후 확률을 계산한다. 사후 확률 과정을 보다 용이하게 수행하기 위하여 사후 확률 도출부(330)는 수학식을 이용하여 전향 확률과 후향 확률을 계산한 후에 전향 확률, 후향 확률, 주어진 관측 값과 전이 확률 관측 확률을 이용하여 사후 확률을 계산할 수 있다.

[0166] f, p, q를 이용한 관측확률은 다음과 같이 계산될 수 있다. 우선, 원본 신호 S의 비트에 따른 변조된 신호 IR의 비트 값은 다음과 같은 확률로 변조된다.

[0167]
- $\Pr(IR[i]=0|S[i]=0)=\frac{1}{2}f(1-q)+(1-\frac{1}{2}f)(1-p)=\mathrm{p_{00}}$
- $\Pr(IR[i]=0|S[i]=1)=\left(1-\frac{1}{2}f\right)(1-q)+\frac{1}{2}f(1-p)=\mathrm{p_{01}}$
- $\Pr(IR[i]=1|S[i]=0)=\frac{1}{2}fq+\left(1-\frac{1}{2}f\right)p=\mathrm{p_{10}}$
- $\Pr(IR[i]=1|S[i]=1)=\left(1-\frac{1}{2}f\right)q+\frac{1}{2}fp=\mathrm{p_{11}}$

[0168] 위 식에서, p_{00}은 원본 신호 S의 비트 값이 0일 때 변조된 신호 IR의 비트 값이 0으로 변조될 확률을, p_{01}은 원본 신호 S의 비트 값이 1일 때 변조된 신호 IR의 비트 값이 0으로 변조될 확률을, p_{10}은 원본 신호 S의 비트 값이 0일 때 변조된 신호 IR의 비트 값이 1로 변조될 확률을, p_{11}은 원본 신호 S의 비트 값이 1일 때 변조된 신호 IR의 비트 값이 1로 변조될 확률을 의미한다.

[0169] 또한, 정의에 따라 원본 신호 S는 하나의 비트가 1의 값을 갖고, 나머지 비트가 0의 값을 갖는다. 따라서, 관측확률 b(IR,S)는 2×N의 경우의 수가 존재한다. 즉, IR에서 임의의 비트 x(또는 x번째 비트)의 관측 값 IR[x], IR에서 x번째 비트를 제외한 나머지 비트들 중 관측 값이 0인 비트의 개수 n(IR[i]=0, i≠x), 및 IR에서 x번째 비트를 제외한 나머지 비트들 중 관측 값이 1인 비트의 개수 n(IR[j]=1, j≠x)은 ([0,1], [0:N-1], [N-1:0])의 분포를 갖는다. 결국, 관측확률은 $(p_{10} or\ p_{11}) \times p_{00}^{n(IR[i]=0,i\neq x)} \times p_{01}^{n(IR[i]=0,i\neq x)}$ 이다.

[0170] 전이 확률 도출부(340)는 수집된 신호 데이터들에 대하여 사후 확률을 이용하여 전이 확률을 업데이트(학습)할 수 있다. 전이 확률 업데이트 과정은 전이 확률이 특정 값으로 수렴할 때까지 반복적으로 수행될 수 있다.

[0171] 정보 추출부(350)는 전이 확률 도출부(340)에 의해 도출된 전이 확률을 은닉 마르코프 모델에 적용함으로써 유용한 정보, 예컨대 사용자들의 시작 위치 분포 및/또는 각 위치 별 사용자의 이동 경향성 등을 추정할 수 있다.

[0173] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시 예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0174] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0175] 실시 예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조ㅂ하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시 예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시 예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0177] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0179] 10 : 수집 및 분석 시스템

100 : 단말

300 : 서버

310 : 데이터 수신부

320 : 모델 모델링부

330 : 사후 확률 도출부

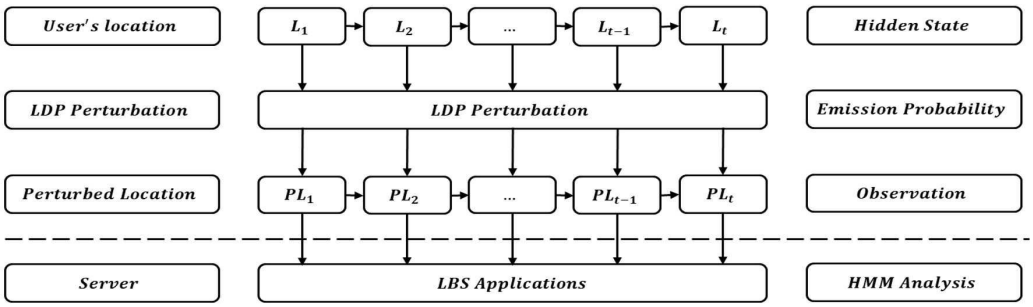
340 : 전이 확률 도출부

350 : 정보 추정부

360 : 저장부

도면

도면1



도면2

- 위치 $N_1, N_2, \ldots, N_n$
- 사용자의 변조된 위치 데이터 x
- 사용자의 원본 위치 데이터 y
- 수집한 신호 데이터 (관측값) $O = \left\{ \{x_1^1, x_1^2 \ldots, x_1^{t_1}\}, \ldots, \{x_\tau^1, x_\tau^2 \ldots, x_\tau^{t_\tau}\} \right\} = \{X_1, X_2, \ldots, X_\tau\}$
- 데이터 O 의 실제 위치 (은닉 상태) $H = \left\{ \{y_1^1, y_1^2 \ldots, y_1^{t_1}\}, \ldots, \{y_\tau^1, y_\tau^2 \ldots, y_1^{t_\tau}\} \right\} = \{Y_1, Y_2, \ldots, Y_\tau\}$
- 전이확률 $a_{ij} = P(y^{t+1} = N_j | y^t = N_i)$
- 관측확률 $b(x^t,\ i) = P(x^t | y^t = N_i)$
- 전향확률 $\alpha_j^t = P(x_{1:t}, y_t = N_i | a, b)$
- 후향확률 $\beta_j^t = P(x^{t+1:t_{end}} | y^t = N_i, a, b)$
- 확률 $\gamma_i^t = P(y^t = N_i | a, b, X)$
- 확률 $\xi_{ij}^t = P(y^t = N_i, y^{t+1} = N_j | a, b, X)$

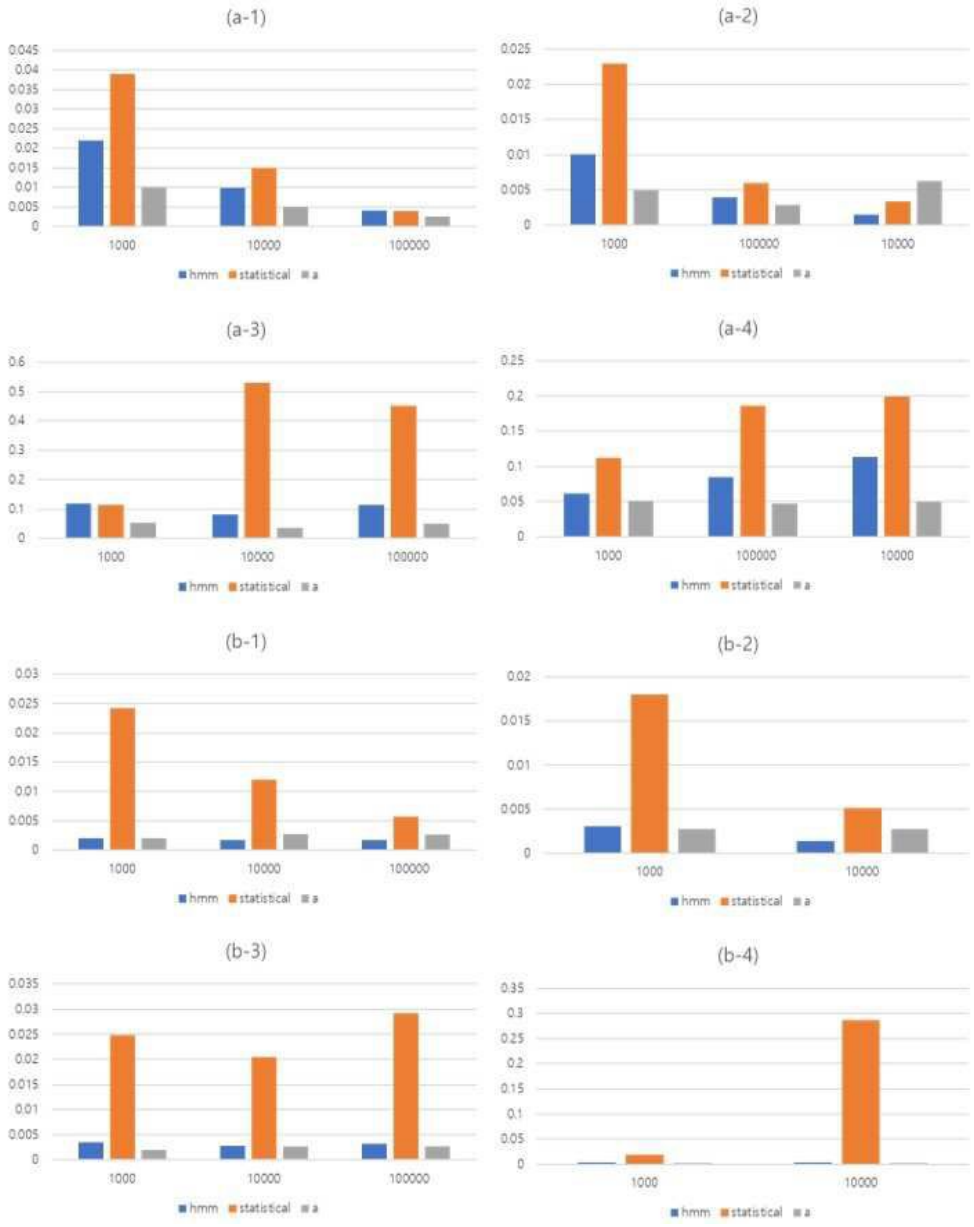
도면3

알고리즘 1. Posterior Probability Estimation(PPE)
Input: $\mathrm{X} = \{\mathrm{x}^1, \mathrm{x}^2, \ldots, \mathrm{x}^T\}, \gamma, \xi, \mathrm{a}, \mathrm{b}$ Output: γ, ξ
1. $\alpha \leftarrow$ forward algorithm(X, a, b) 2. $\beta \leftarrow$ backward algorithm(X, a, b) 3. for each time t 4. for each state i 5. $\gamma_i^t \leftarrow \alpha_i^t \beta_i^t$ 6. $\gamma_i \leftarrow \gamma_i + \gamma_i^t$ 7. if t = 1 then 8. $\xi_{0\mathrm{i}} \leftarrow \xi_{0i} + \gamma_i^1$ 9. end 10. if t ≠ T then 11. for each state j 12. $\xi_{ij}^{\mathrm{t}} \leftarrow \alpha_i^t a_{ij} b(x^{t+1}, j) \beta_j^{t+1}$ 13. $\xi_{ij} \leftarrow \xi_{ij} + \xi_{ij}^{\mathrm{t}}$ 14. end 15. end 16. end 17. end

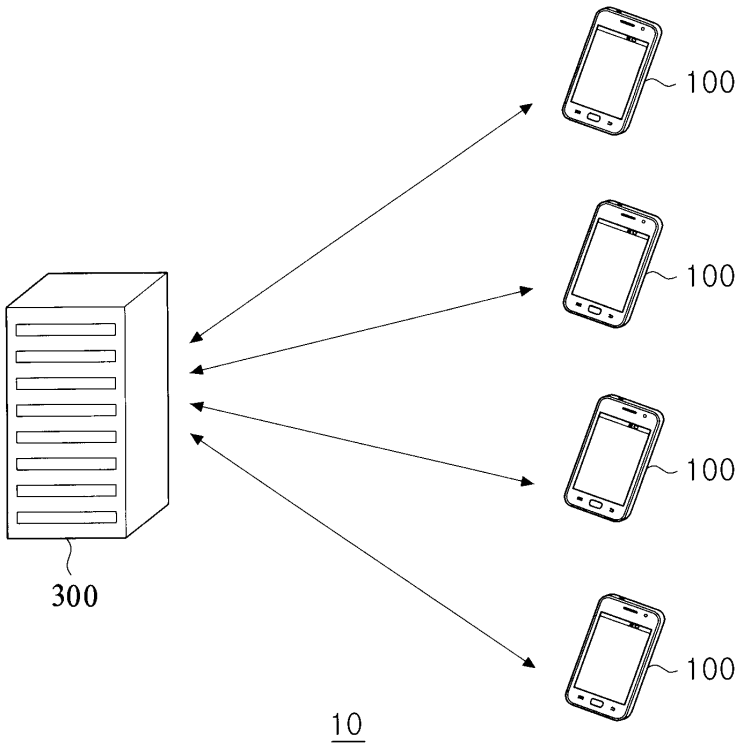
도면4

알고리즘 2. EM algorithm for transition probabilities
Input: $\mathrm{D} = \{X_1, X_2, \ldots, X_\tau\} a, b$ Output: a
1. Repeat 2. Initialize γ, ξ 3. for each tracjectory $X \in D$ 4. $[\gamma, \xi] \leftarrow [\gamma, \xi] + PPE(X, \gamma, \xi, a, b)$ 5. end 6. for each state i 7. $a'_{0i} \leftarrow \xi_{0i}$ 8. for each state j 9. $a'_{ij} \leftarrow \xi_{ij} / \gamma_i$ 10. end 11. end 12. $a \leftarrow a'$ 13. Until converge

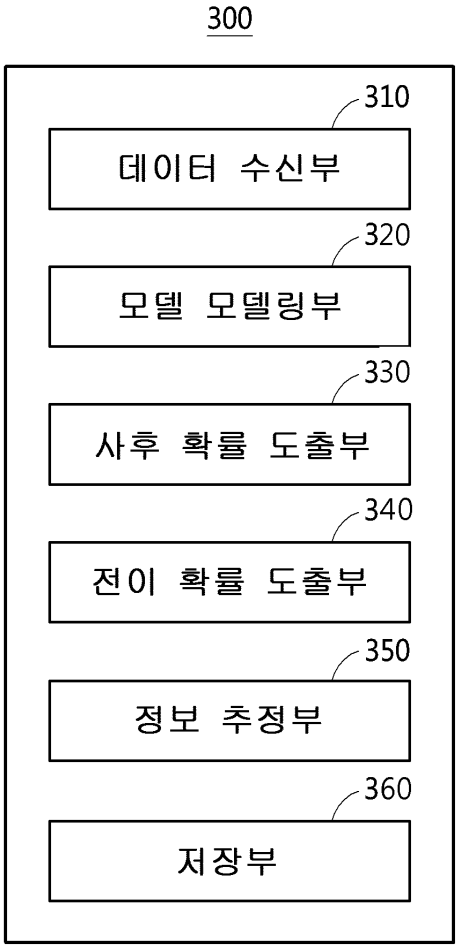
도면5


(a-1)
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
1000
10000
100000
hmm
statistical
a
(a-2)
0.025
0.02
0.015
0.01
0.005
0
1000
100000
10000
hmm
statistical
a
(a-3)
0.6
0.5
0.4
0.3
0.2
0.1
0
1000
10000
100000
hmm
statistical
a
(a-4)
0.25
0.2
0.15
0.1
0.05
0
1000
100000
10000
hmm
statistical
a
(b-1)
0.03
0.025
0.02
0.015
0.01
0.005
0
1000
10000
100000
hmm
statistical
a
(b-2)
0.02
0.015
0.01
0.005
0
1000
10000
hmm
statistical
a
(b-3)
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
1000
10000
100000
hmm
statistical
a
(b-4)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1000
10000
hmm
statistical
a

도면6



도면7


300
310
데이터 수신부
320
모델 모델링부
330
사후 확률 도출부
340
전이 확률 도출부
350
정보 추정부
360
저장부