JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 12, pp. 161-169, December 2022
https://doi.org/10.9708/jksci.2022.27.12.161

Privacy-Preserving Estimation of Users' Density Distribution in Location-based Services through Geo-indistinguishability

Seung Min Song*, Jong Wook Kim*

*Student, Dept. of Computer Science, Sangmyung University, Seoul, Korea
*Professor, Dept. of Computer Science, Sangmyung University, Seoul, Korea

[Abstract]

With the development of mobile devices and global positioning systems, various location-based services can be utilized, which collects user's location information and provides services based on it. In this process, there is a risk of personal sensitive information being exposed to the outside, and thus Geo-indistinguishability (Geo-Ind), which protect location privacy of LBS users by perturbing their true location, is widely used. However, owing to the data perturbation mechanism of Geo-Ind, it is hard to accurately obtain the density distribution of LBS users from the collection of perturbed location data. Thus, in this paper, we aim to develop a novel method which enables to effectively compute the user density distribution from perturbed location dataset collected under Geo-Ind. In particular, the proposed method leverages Expectation-Maximization(EM) algorithm to precisely estimate the density disribution of LBS users from perturbed location dataset. Experimental results on real world datasets show that our proposed method achieves significantly better performance than a baseline approach.

▸ **Key words**: Location-based services, Location data, Data Privacy, Geo-indistinguishability

[요 약]

최근 들어 모바일 디바이스와 GPS(Global Positioning System)의 발전으로 다양한 위치 기반 서비스(Location-Based Servises, LBS)를 활용할 수 있게 되었다. LBS 사용자는 서비스를 이용하기 위해 자신의 위치 정보를 서비스 제공자에게 노출한다. 이 과정에서 개인의 민감한 정보를 침해할 가능성이 있으므로 사용자의 위치 데이터를 변조하여 프라이버시를 보존할 수 있는 Geo-indistinguishability(Geo-Ind) 기법이 많이 활용되고 있다. 그러나 Geo-Ind 기법으로 인하여 사용자로부터 변조된 데이터를 수집하는 경우, LBS 제공자는 사용자 분포에 대한 정확한 정보를 얻을 수 없다. 그러므로 본 논문에서는 Geo-Ind 기법을 이용하여 사용자로부터 수집한 변조된 위치 데이터로부터 사용자 분포에 대한 정보를 정확하게 계산하기 위한 방법을 제안한다. 특히, Expectation-Maximization(EM) 기법을 이용하여 변조된 데이터로부터 사용자의 위치 분포를 정확하게 예측하기 위한 기법을 제안한다. 또한 실제 데이터를 이용해 제안 기법의 우수성을 입증한다.

▸ **주제어**: 위치 기반 서비스, 위치 데이터, 데이터 프라이버시, Geo-indistinguishability

• First Author: Seung Min Song, Corresponding Author: Jong Wook Kim
*Seung Min Song (hungryhappy@naver.com), Dept. of Computer Science, Sangmyung University
*Jong Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2022. 10. 20, Revised: 2022. 12. 09, Accepted: 2022. 12. 09.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

4차 산업혁명이 도래한 이후로 인공지능, 사물 인터넷, 빅데이터, 모바일 등 다양한 분야에서 물리적, 생물학적, 디지털 세계를 연결하며 혁신적인 변화가 나타나고 있다. 특히 기존 산업혁명보다 진화한 첨단 지능정보기술이 사용되어 더 넓은 범위에 더 생산적인 영향을 주고 있는 것으로 알려져 있다. 매년 디지털 환경에서 사람과 산업을 연결하는 방대한 자료가 생성되는데 이것을 효과적으로 보관, 정리, 분석, 가공하여 다른 산업에 적용하는 것이 빅데이터 사회에서의 중요한 과제이다. 사용자가 생성하는 자료 중에서 위치 데이터는 여러 기술의 발전과 함께 더 많은 사용가치를 가지게 되었다.

정보통신 사회를 기반으로 사물인터넷(Internet of Things, IoT) 기술은 생활 전반에 걸쳐 밀접하게 사용되고 있다. IDC(International Data Corporation)는 많은 IoT 기기들이 향후 몇 년간 기하급수적인 빅데이터를 생성할 것이라는 통계자료를 발표했다. 2025년에는 73.1 ZB(1ZB=10억 TB)에 달할 것으로 예상되며 이는 2019년에 발생한 데이터의 422%에 해당한다 [1]. IoT 산업의 발달과 더불어 대부분의 모바일 기기는 위치를 인식할 수 있는 GPS(Global Positioning System) 기능을 탑재하고 있다.

모바일 디바이스와 GPS의 발전으로 인해 위치기반서비스(Location-Based Servises, LBS)를 사용할 수 있게 되었다. 위치기반서비스는 이동통신망이나 GPS를 통해 얻은 정보를 바탕으로 사람이나 사물의 위치를 정확하게 파악하여 사용자에게 응용 서비스를 제공하는 것이다 [2]. LBS의 형태는 목적에 따라 크게 두 가지로 구분할 수 있다. 첫 번째로 공적 서비스는 사용자가 위험한 상황에서 조난 구조 서비스나 긴급 경보 시스템과 같은 것을 제공하는 서비스를 말한다. 두 번째로 사적 서비스는 사용자의 편의를 위하여 제공하는 서비스를 말한다. 사용자들은 자신의 위치를 기반으로 한 네비게이션 시스템, 기상 정보 등 유용한 정보를 제공 받을 수 있다. 예를 들어, 가장 가까운 ATM 기기를 찾거나 주변 가게의 쿠폰을 받을 수도 있다. 또한 배달 주문 어플에서는 주변에서 주문 가능한 식당을 보여줄 뿐만 아니라, 배달원이 주문 목적지까지 이동하는 경로를 실시간으로 지도에 나타내어 보여주고, 목적지까지의 예상시간을 알려준다.

LBS는 주로 이동통신 업체와 자동차 업체를 중심으로 발전되고 있다. 그런데 위치 정보를 활용하는 업체들이 사용자의 정보를 수집하고 서비스를 제공하는 과정에서 사용자의 프라이버시를 침해할 가능성이 존재한다. 많은 LBS에서는 사용자가 인식하지 못하는 순간에도 방대한 양의 사용자 위치 정보를 수집하고 이 정보를 외부에 제공하는 경우가 있다. 예를 들어 미국의 이동통신 회사 AT&T, T-Mobile, Sprint는 사용자들의 실시간 위치 데이터를 수집하여 데이터 브로커에게 유출하였다 [3]. 개인의 위치 정보는 민감한 정보(예, 직장 주소, 병원 방문 기록)를 담고 있으므로, 활용하는 과정에서 프라이버시 보호 기법을 사용하는 것이 필요하다.

LBS 환경에서 사용자의 위치 정보에 대한 프라이버시를 보존하면서, 서비스를 제공하기 위한 다양한 기법이 연구되어 왔다. 전통적인 방식으로는 dummy location [4], cloaking [5] 기법과 같이 익명화 기법을 사용하는 방식이 있다. 최근들어 차분 프라이버시(Differential Privacy, DP) 기법이 프라이버시 보존 데이터 수집 및 처리에 있어서 표준으로 사용되고 있다 [6, 7, 8, 9, 10]. 이에 따라 DP 기법을 사용자의 위치 데이터에 적용하려는 다양한 방법이 제안되었다. 그중 Geo-indistinguishability (Geo-Ind) [11, 12, 13, 14]는 위치 데이터에 적용 가능한 가장 대표적인 DP 기반 기법에 해당한다.

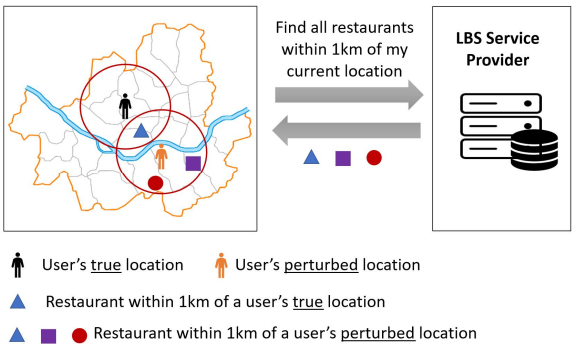

Fig. 1. Protecting location privacy of users in LBS using Geo-Ind

그림 1과 같이 사용자는 Geo-Ind 기법을 이용하여 위치 데이터에 대한 변조를 수행한 후, 변조된 위치 데이터와 서비스 요청을 LBS 서버에 전송한다. LBS 서버는 사용자의 변조된 데이터를 기반으로 사용자에게 서비스를 제공한다. 이 과정에서 사용자의 원본 데이터는 외부에 노출되지 않기 때문에, 사용자의 위치 정보에 대한 프라이버시를 보존할 수 있다.

일반적으로 LBS 제공자는 사용자로부터 수집한 대용량 위치 데이터로부터 사용자 분포도를 추출한 후, 이를 서비스 개선에 활용한다. 그러나 Geo-Ind 기법으로 인하여 사용자로부터 변조된 데이터를 수집하는 경우, LBS 제공자는 사용자 분포에 대한 정확한 정보를 얻을 수 없다. 그러

므로 본 논문에서는 Geo-Ind 기법을 이용하여 사용자로부터 수집한 변조된 위치 데이터로부터 사용자 분포에 대한 정보를 정확하게 계산하기 위한 방법을 제안한다. 본 논문의 제안 기법을 통하여 사용자 분포에 대한 예측 정확도를 높임으로써 LBS 제공자는 제공하는 서비스 품질 향상을 기대할 수 있다. 이로 인하여 서비스 사용자는 보다 개선된 서비스를 제공받을 수 있는 혜택을 누릴 수 있다.

본 연구의 기여는 다음과 같다. 먼저 사용자로부터 수집한 변조된 데이터로부터, Expectation-maximization (EM) 기법을 이용하여 사용자의 위치 분포를 정확하게 예측하기 위한 기법을 제안한다. 또한 실제 데이터를 이용해 제안 기법의 우수성을 평가한다. 본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문과 연관된 선행 연구에 대해서 설명하고, 3장에서는 배경 지식인 Geo-Ind에 대해 설명한다. 4장에서는 제안 기법에 대하여 설명한다. 5장에서는 제안 기법의 성능을 평가를 수행한 후, 6장에서 결론을 맺는다.

II. Related Work

프라이버시를 보존하며 사용자 분포도를 예측하기 위한 다양한 연구가 이루어져 왔다. 가장 간단하면서도 널리 사용된 방법은 히스토그램을 사용하여 분포도를 예측하는 것이다 [15]. 하지만 해당 방식의 기법을 사용하기 위해서는 현실적이지 않은 이산화된 공간을 가정하게 되어 분포도가 적절하지 않다. 또한 커널 밀도 추정(Kernel Density Estimation)도 대표적인 분포 추정 기법이다 [16]. 전통적인 이산 히스토그램을 부드럽게 근사하여 Kernel 함수를 얻고 이를 활용하여 분포도를 추정하는 방법이다. 파라미터를 사용하지 않는 방법들은 데이터 전체를 저장하는 것이 필요한데 결점이 많아서 사용하지 않는다. 최근에는 딥러닝을 기반으로 한 기법들이 다양하게 연구되고 있는데 자기회귀변환(autoregressive transformations)을 사용하는 기법들도 있다. 결합 확률 분포를 조건부 확률의 곱으로 나타낼 때 연쇄법칙을 사용하는 변환 방식이다. 이러한 모델들에는 Masked Autoregressive Flow (MAF) [17], 밀도 추정을 위해 최적화된 Real NVP 가 있다. 또한 Inverse Autoregressive Flow [18] 는 다른 기법들에 비해 변분추론(variational inference)에 최적화되어있다 [19, 20, 21].

사용자의 위치 정보를 변조하여 프라이버시를 보존하는 방식인 Geo-Ind는 다양한 분야에서 사용되고 있다. 공간 클라우드 소싱(Spatial crowdsourcing, SC)에서는 개인이나 그룹이 자신들의 주변에 있는 정보들을 수집하여 SC 서버에 전송하는 역할을 수행한다. 이때, 이들은 자신들의 위치 정보를 SC 서버와 공유하게 되며, 이로인하여 위치 정보에 대한 프라이버시 침해 문제가 발생할 수 있다. 그러므로 사용자들의 실제 위치 대신에 Geo-Ind를 이용하여 변조된 위치 정보를 SC 서버에 제공하기 위한 다양한 연구가 진행되고 있다. Wang et al. [22]는 Geo-Ind를 활용하여 SC 플랫폼 상에서의 사용자들의 위치 데이터를 보호하기 위한 방법을 제안하였다. 공유 주행 서비스는 유사한 경로로 이동하는 사용자들끼리 차량을 공유하고 비용을 절감하게 해주는 서비스로서, 교통 완화, 에너지 절약, 환경 보호 등의 부가적인 장점이 있다. Tong et al. [23]는 Geo-Ind를 활용하여 공유 주행 서비스 환경에서 사용자의 위치 정보를 보호하기 위한 기법을 제안하였다. 위치 인식 소셜 네트워크(Location-aware social networks)에서도 Geo-Ind를 활용하여 위치 정보에 대한 프라이버시를 보존하기 위한 다양한 연구가 진행되어 있다. 가장 대표적으로는 Location Based Social Discovery(LBSD) 서비스이다. 이는 사용자 주변의 사람들을 인식하여 사용자 주변 사람들의 리스트를 제공한다. 사용자는 LBSD를 통하여 주변에 사는 친구를 만나거나 새로운 친구를 사귀는 기회를 가질 수 있다. 하지만 이 과정에서 사용자의 위치 정보에 대한 프라이버시가 침해될 우려가 있으며, 이를 방지하기 위하여 Geo-Ind 기법이 활용된다 [24, 25, 26, 27].

사용자의 분포도를 예측하는 과정에서 정확도를 높이기 위하여 EM 알고리즘 기법을 사용할 수 있다. EM 알고리즘이란 E-step (Expectation)과 M-step (Maximization)을 반복적으로 시행하여 최대가능도(Maximum Likelihood)를 최대화하는 확률 모델이다[28, 29]. 주로 완전하지 않은 관측 데이터셋를 가지고 있을 때 잠재적인 수치를 고려하여 완전한 데이터셋을 추정할 때 사용할 수 있다. EM 알고리즘에 관한 연구도 다양한 분야에서 지속적으로 진행되어왔다. 금속공학분야에서 효율적인 추가 재료 구매를 위한 의사결정을 할 때 사용되었다 [30]. 신호처리분야에서 코히어런트 광통신과 직접위치결정기법의 정확도를 높이기 위해 사용되었다 [31, 32].

III. Background

Geo-Ind란 사용자의 위치 데이터에 노이즈를 추가하는 방식으로 변조하여 공격자가 사용자의 위치를 정확하게

유추하는 것을 방지하는 기법이다.

정의 1. (ε-Geo-Indistinguishability) X를 사용자의 실제 위치 데이터 집합이라 가정하고, Z를 사용자가 서버에 전송한 변조된 위치 데이터 집합이라고 가정한다. M는 임의의 매커니즘이라고 가정할 때 아래의 조건을 만족하면 M는 ε-Geo-Ind를 만족한다.

(1) X의 임의의 데이터 x_1과 x_2에 대하여

(2) M로부터 생성되는 모든 결과값 $z \in Z$에 대하여 다음 식을 만족한다.

$$M[x_1, z] \leq e^{\epsilon \times d(x_1, x_2)} \times M[x_2, z] \qquad \text{식(1)}$$

이때, $d(x_1,x_2)$는 x_1과 x_2 사이의 거리이다.

Geo-Ind를 구현하기 위한 두 가지 방법이 있다. 그중 하나인 라플라스 매커니즘(Laplace mechanism) [6, 7]은 사용자의 실제 데이터에 라플라스 분포상의 노이즈를 추가하는 방식으로 변조하는 기법이다. 이는 간단한 방식이지만, 위치 데이터 변조 과정에서 노이즈가 많이 추가될 가능성이 있어서 정확도가 떨어지고, 데이터 유용성이 낮아질 우려가 있다. 반면 최적화(Optimization) 매커니즘은 이전에 언급한 라플라스 기법에 비해 사용자의 실제 위치 데이터에 대한 변조가 적게 발생하므로, 데이터 유용성이 높은 기법이다. 그러므로 본 논문에서는 최적화 매커니즘을 이용한다.

X를 사용자의 실제 위치 데이터 집합이라 가정하고, Z를 사용자가 서버에 전송한 위치 데이터 집합이라고 가정한다. π는 사전 확률 분포(prior probability distribution)로 가정할 때 최적화 매커니즘 M는 아래의 조건을 만족한다.

$$\begin{aligned} &\text{min: } \sum_{x \in X, z \in Z} \pi \cdot M[x, z] \cdot d(x, z) \\ &\text{s.t.: } M[x_1, z] \leq e^{\varepsilon \cdot d(x_1, x_2)} \times M[x_2, z] \\ &\qquad\qquad x_1, x_2 \in X,\ z \in Z \qquad \text{식(2)} \\ &\sum_{z \in Z} M[x, z] = 1 \qquad x \in X \\ &M[x, z] > 0 \qquad x \in X, z \in Z \end{aligned}$$

이때, π는 $\frac{1}{|X|}$인 균등분포이다.

그림 2에서 보이듯 최적화 매커니즘에서는 먼저 LBS 서버가 Geo-Ind를 만족하는 변조화 행렬 M을 생성한 후 사용자에게 배포한다. 사용자는 배포된 변조화 행렬을 내려받아 자신의 실제 위치 데이터를 변조한다. 그리고 다시 서버에게 변조된 데이터를 전송한다.

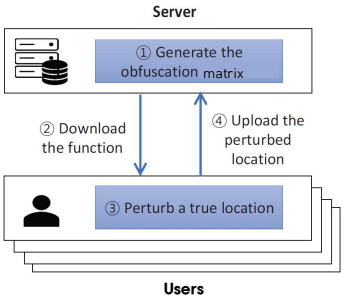

Fig. 2. Optimization Mechanism

IV. The Proposed Scheme

본 장에서는 본 논문의 제안 기법을 설명한다. 그림 3은 제안 기법의 구성도에 해당한다.

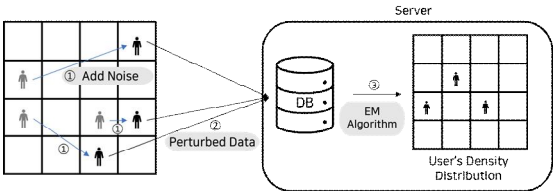

Fig. 3. The architecture of the proposed approach

- 사용자는 Geo-Ind의 변조화 기법에 따라 실제 위치 데이터를 변조한다.
- 사용자는 변조된 위치 데이터를 서버에 전송한다.
- 사용자드로부터 프라이버시를 보존하며 수집한 대용량 위치 데이터로부터 EM 알고리즘을 이용하여 사용자의 분포도를 예측한다.

4.1 Problem Definition and Naive Solution

위치 데이터를 수집하고자 하는 지역을 m개의 겹치지 않는 격자로 분할하였다고 가정하면, 전체 지역을 $G = \{g_1, g_2, \cdots, g_m\}$ 으로 표현할 수 있다. 또한 k명의 사용자가 있다고 가정하면, 사용자의 집합을 $U = \{u_1, u_2, \cdots, u_k\}$ 로 표현할 수 있다. 이때 i-번째 사용자 u_i의 위치를 $g_{l_i} \in G$라고 하면, k명의 사용자의 위치 데이터는 $\{g_{l_1}, g_{l_2}, \cdots, g_{l_k}\}$에 해당한다. $g_{l_i}' \in G$를 Geo-Ind의 변조 기법에 의해 생성된 사용자 u_i의 변조된 위치 데이터라 가정하자. 이때 사용자의 실제 위치 데이터 g_{l_i}가 Geo-Ind의 변조기법에 의해서 g_{l_i}'로 변조될 확률은 $M[g_{l_i}, g_{l_i}']$에 해당한다. 3.1절에서 언급하였듯이 변조화

행렬 M은 식(2)에 의해 계산되어 사용자에게 배포된다.

LBS 서버가 t명의 사용자($u_1, u_2, \cdots, u_k$)로부터 전송받아 데이터베이스에 저장한 변조된 데이터를 $DB = \{g_{l_1}', g_{l_2}', \cdots, g_{l_k}'\}$라 가정하자. 이러한 변조된 데이터를 이용하여 사용자의 분포를 예측하기 위해서 사용할 수 있는 단순 기법은 다음과 같다. 사용자가 전체 구역 m개 중 r-번째 격자 위치 g_r에 위치할 확률(즉, 분포율)을 $Fq(g_r)$이라 하자. 이때 단순기법에 의해서 분포율은 다음과 같이 계산할 수 있다.

$$Fq(g_r) = \frac{Num(g_r, DB)}{k} \quad \text{식(3)}$$

이때 k는 전체 사용자의 수에 해당하고, $N(g_r, DB)$는 g_r이 DB에 나타나는 빈도수에 해당한다.

위의 방식은 단순하지만, 변조화 행렬 M을 고려하지 않으므로 사용자의 분포도를 정확히 예측할 수 없다. 그러므로 다음 절에서는 Geo-Ind 기법에 의해 변조된 위치 데이터로 부터 사용자의 분포도를 정확하게 예측할 수 있는 방법을 제안한다.

4.2 Estimating the User Distribution Using EM Algorithm

Geo-Ind에 의해 변조화된 위치 데이터로 인하여 사용자들의 분포도를 정확하게 측정하기는 쉽지 않다. 그러므로 본 연구에서는 프라이버시를 보존하기 위한 데이터 변조로 인해 발생한 불확실성을 해결하기 위하여 EM 알고리즘을 사용한다. EM 알고리즘을 이용하여 데이터의 분포도를 처음엔 특정 분포(예, 균등 분포)로 가정한 후, 이를 반복적으로 최적화 시키는 방법을 이용하여 전체의 분포도를 예측한다.

서버가 사용자로부터 전송받아 데이터베이스에 저장한 $DB = \{g_{l_1}', g_{l_2}', \cdots, g_{l_k}'\}$ 중에서 r-번째 사용자로부터 받은 변조된 데이터는 $g_{l_r}' \in DB$이다. 이때 이 r-번째 사용자의 원래 위치가 $g_i \in G$ 이었을 확률은 조건부 확률 $P(g_i \mid g_{l_r}')$로 표현 가능하고, 이는 베이즈 정리에 의해 다음과 같다.

$$P(g_i \mid g_{l_r}') = \frac{P(g_i \cap g_{l_r}')}{P(g_{l_r}')} = \frac{P(g_i)P(g_{l_r}' \mid g_i)}{P(g_{l_r}')} = \frac{P(g_i)P(g_{l_r}' \mid g_i)}{\sum_{j=1}^{m} P(g_j)P(g_{l_r}' \mid g_j)} \quad \text{식(4)}$$

이때 변조화 행렬 M의 정의에 의해 $P(g_{l_r}' \mid g_i) = M[g_i, g_{l_r}']$ 이므로, 식(4)는 다음과 같이 정리할 수 있다.

$$P(g_i \mid g_{l_r}') = \frac{P(g_i)M[g_i, g_{l_r}']}{\sum_{j=1}^{m} P(g_j)M[g_j, g_{l_r}']} \quad \text{식(5)}$$

4.1장에서 언급한 분포율은 $Fq(g_i) = P(g_i)$에 해당한다. 그런데 이는 변조된 데이터 만으로는 정확히 측정할 수 없으므로, 본 논문에서는 EM 알고리즘을 이용하여 실제와 유사한 분포율을 계산하는 방법을 제안한다. EM 알고리즘을 이용하여 분표율 $P(g_1), P(g_2), ..., P(g_m)$를 구하기 위한 방법은 다음과 같이 4단계로 구성된다.

- 초기화: 초기 파라미터의 값은 균등분포를 가정하여 다음과 같이 설정한다.

$$\theta_i^{(0)} = P(g_i)^{(0)} = \frac{1}{m},\ 1 \le i \le m \quad \text{식(6)}$$

즉, 전체 영역이 m개의 격자로 구성되어 있으므로, 각각 영역의 분포율은 $\frac{1}{m}$로 균등하게 초기화된다 (즉, $\sum_{i=1}^{m}\theta_i^{(0)} = \sum_{i=1}^{m} P(g_i)^{(0)} = 1$).

- E-step: 현재의 파라미터 $\theta^{(t)}$를 사용하여 사후 확률 $P(g_k \mid g_{l_i}')$를 구한다. $P(g_k \mid g_{l_i}')$는 식(4)에 의해 다음과 같다.

$$P(g_i \mid g_{l_r}' : \theta^{(t)})$$

$$= \frac{P(g_i)^{(t)} M[g_i, g_{l_r}']}{\sum_{j=1}^{m} P(g_j)^{(t)} M[g_j, g_{l_r}']} \quad \text{식(7)}$$

$$= \frac{\theta_i^{(t)} M[g_i, g_{l_r}']}{\sum_{j=1}^{m} \theta_j^{(t)} M[g_j, g_{l_r}']}$$

▪ M-step: 매 반복마다 $\theta_i^{(t)}$는 다음과 같이 업데이트 된다.

$$\theta_i^{(t+1)} = P(g_i)^{(t+1)}$$

$$= \sum_{r=1}^{k} \left(P(g_i \mid g'_{l_r} : \theta_i^{(t)}) \right) \quad \text{식(8)}$$

식(8)에서 사후 확률 $P(g_k \mid g_{l_i}' : \theta_i^{(t)})$는 이전 E-step에서 구한 값을 사용한다. 모든 지역의 분포율의 합은 1이므로, $\theta_i^{(t+1)}$은 다음과 같이 정규화된다.

$$\theta_i^{(t+1)} = \frac{\theta_i^{(t+1)}}{\sum_{j=1}^{m} \theta_j^{(t+1)}} \quad \text{식(9)}$$

▪ 위의 E-step과 M-step은 미리 지정한 횟수 만큼 반복 후, 종료한다. 또는 각 분포율을 나타내는 파라미터의 차이가 수렴할 때까지 반복한다. 즉, 다음의 조건을 만족할 때까지 위의 E-step과 M-step을 반복한다.

$$\max \mid \theta_i^{(t+1)} - \theta_i^{(t)} \mid < \gamma \quad \text{식(10)}$$

이때, γ는 사용자가 지정한 임계값에 해당한다.

위의 EM 알고리즘 종료 후, 분포율은 $Fq(g_i) = P(g_i)^{(t+1)} = \theta_i^{(t+1)}$에 의해 구할 수 있다.

V. Experiment

5.1 Experiment Setup

본 연구에서는 베이징에서 운행하는 택시의 이동경로를 기록한 T-Drive 데이터[33]를 사용하여 성능 평가를 수행하였다. T-Drive 데이터에서 택시의 위치는 위도/경도로 표현된다. 먼저, 전체 영역을 위도/경도에 의해 10×10, 15×15, 20×20개의 격자 구역으로 나누었다. 전체 T-Drive 데이터 중에 85,707개의 택시를 임의로 추출하였다. 또한, 각 택시별로 이동경로 중에 있는 10개의 위치를 임의로 추출하여 실험에 사용하였다. 본 실험에서는 Geo-Ind의 프라이버시 예산 ε 값으로 0.5, 1.0, 2.0을 사용하였다. 변조화 행렬은 Gurobi 라이브러리를 사용하여 구현하였다.

제안 기법의 성능을 평가하기 위해 원본 위치 데이터로부터 추출한 사용자 분포도와 변조된 위치 데이터로부터 추출한 사용자 분포도 사이의 평균 절대 오차 (Mean Absolute Error, MAE)를 사용하였다. 전체 m개의 구역에서 각 구역의 실제 분표율을 $Fq_{org}(g_i)$이라고 하고, 변조된 데이터를 바탕으로 예측한 분포율을 $Fq_{est}(g_i)$이라고 가정하자 (단, $1 \le i \le m$). 이때 MAE는 다음과 같이 정의된다.

$$MAE = \frac{1}{m} \times \sum_{i=1}^{m} \mid Fq_{org}(g_i) - Fq_{est}(g_i) \mid \quad \text{식(11)}$$

본 실험에서는 4.1절의 단순 기법(NA)과 4.2절의 제안 기법(EM) 간의 비교 성능 평가를 수행한다.

5.2 Results

그림 4는 단순 기법과 제안 기법의 성능 평가 결과를 나타낸다. 이 실험은 프라이버시 예산 ε 값이 예측 결과에 미치는 영향을 보여준다. 이 실험에서 세 개의 서로 다른 ε 값이 사용되었으며, 10×10 격자가 사용되었다. 제안 기법의 EM 알고리즘 반복 횟수는 10으로 설정하였다.

그림에서 알 수 있듯이 ε 값이 감소할수록 MAE 값이 증가하는 것을 알 수 있다. 이는 ε 값이 감소할수록 사용자의 프라이버시 보호 수준이 증가하여, 원본 데이터에 대한 심한 변조가 발생하기 때문이다. 반면에 ε 값이 증가할수록 예측 값이 실제 값과 유사해짐을 알 수 있다. 이는 ε 값이 증가할수록 사용자의 프라이버시 보호 수준이 감소하여, 원본 데이터에 대한 변조가 작게 발생하기 때문이다. 그림에서 알 수 있듯이 모든 ε 값에 대하여 제안 기법이 단순 기법보다 우수한 성능을 보이고 있다. 특히 ε 값이 감소할수록 단순 기법과 제안 기법 간의 성능 차이가 커지는 것을 알수 있다.

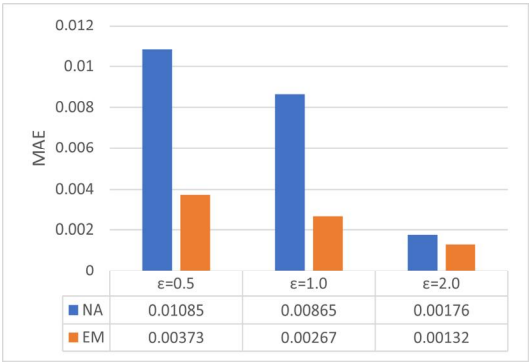

Fig. 4. Comparison between NA and EM for varying ε

그림 5는 세 개의 서로 다른 격자에 대한 실험 결과이다. 이 실험에서 ε 값은 1.0으로 설정되었으며, 제안 기법의 EM 알고리즘 반복 횟수는 10으로 설정하였다.그림 4의 실험 결과와 유사하게 모든 격자 크기에 대하여 본 논문의 제안 기법이 단순 기법보다 우수한 성능을 보인다. 그림 4와 그림 5의 실험 결과는 본 논문에서 제안하는 EM 알고리즘 기반 기법의 우수성을 입증한다.

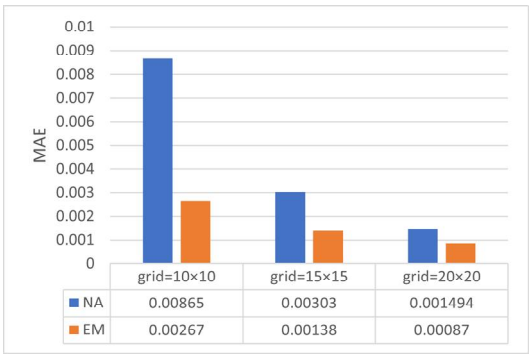

Fig. 5. Comparison between NA and EM for varying grid sizes

그림 6은 제안 기법의 EM 알고리즘 반복 횟수에 따른 MAE 값의 변화를 보여준다. 그림 6-(a)의 실험에서는 ε 값을 0.5에서 2.0으로 설정하였으며, 10×10 격자가 사용되었다. 그림 6-(b)의 실험에서는 ε 값을 1.0으로 설정하였으며, 세 개의 서로 다른 격자 10×10, 15×15, 20×20가 사용되었다. 이 실험에서 EM 알고리즘의 최대 반복 횟수는 50으로 설정하였다. 그림에서 보이듯이, MAE 값은 초기에 급격하게 감소한 후, 서서히 증가하거나 혹은 안정화되는 것을 알 수 있다. 그림 6의 실험 결과는 EM 알고리즘의 적은 반복 횟수만으로도 좋은 예측 결과를 얻을 수 있음을 보여준다. 이러한 결과는 EM 알고리즘의 반복으로 인한 추가 과부화(overhead)가 적게 발생함을 보여준다.

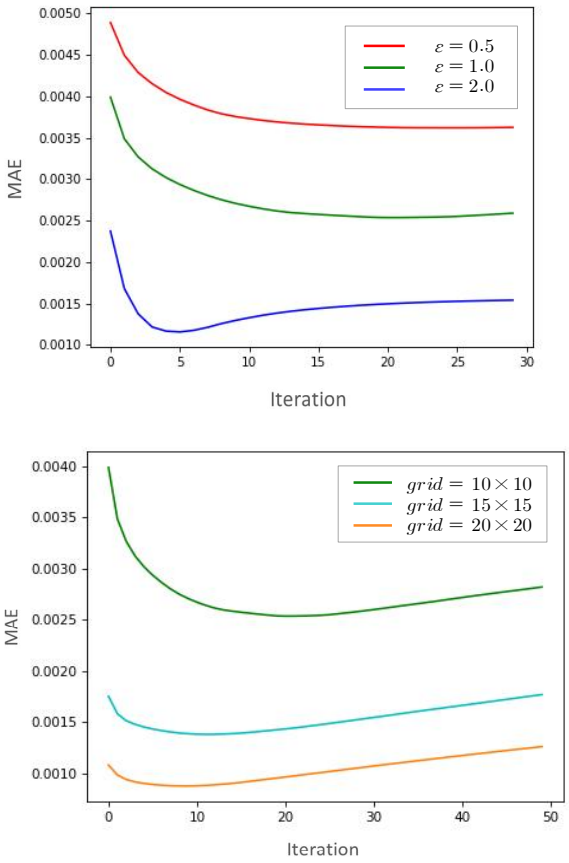

Fig. 6. Iteration vs. MAE for varying (a) ε and (b) grid sizes

VI. Conclusions

4차 산업혁명 이후 사물 인터넷, 모바일, GPS 등의 분야가 비약적으로 발전하면서 언제 어디서나 엄청난 양의 데이터를 수집할 수 있게 되었다. 이 데이터를 적절히 가공하고 분석하여 활용할 수 있는 대표적인 서비스는 LBS이다. 이때 사용자의 위치는 민감한 개인정보에 해당하므로 사용자의 프라이버시를 보호하기 위해서는 데이터를 변조하여 서버에 전송해야 한다. 하지만 변조된 데이터를 사용하면 원래 사용자 분포도와 차이가 크기 때문에 데이터 유용성이 낮아진다. 본 연구에서는 사용자의 프라이버시를 보존하면서도 원래 사용자의 분포도에 가깝게 예측하여 데이터 유용성을 높이기 위한 방법을 제안하였다. 또한, 실제 데이터를 이용한 성능 평가를 통하여 본 논문에서 제안한 EM 알고리즘 기반 기법의 우수성을 입증하였다.

ACKNOWLEDGEMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2021-0-00884,Development of Integrated Platform for Untact Collaborative Solution and Blockchain Based Digital Work). This research was also supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF-2020R1F1A1072622).

REFERENCES

[1] Bojan Jovanovic, “Internet of Things statistics for 2022 - Taking Things Apart”, https://dataprot.net/statistics/iot-statistics/

[2] Nam Jiyeon, Rha Jong Youn, “Location Privacy Concern and its impact on consumers’ intention to use Location-Based Services”, Consumer Policy And Education Review, Vol. 5, No. 2, pp. 81-102, June 2009.

[3] Matt Kapko, AT&T, T-Mobile, Sprint seek to make amends over location breach, https://www.fiercewireless.com/wireless/u-s-carriers-caught-selling-customers-location-data-to-third-parties-again

[4] Kido H, Yanagisawa Y, Satoh T, “Protection of Location Privacy Using Dummies for Location-based Services”, Proceedings of the International Conference on Data Engineering Workshops, pp.1248-1248, Tokyo, Japan, April 2005. DOI: 10.1109/ICDE.2005.269.

[5] Gruteser MO, Grunwald D., “Anonymous Usage of Location-based Services through Spatial and Temporal Cloaking”, Proceedings of the International Conference on Mobile Systems, Applications and Services pp. 31–42. San Francisco, CA, USA, 2003. DOI: 10.1145/1066116.1189037

[6] C. Dwork. “Differential privacy”, Proceedings of the International Colloquium on Automata, Languages, pp. 1-12, Venice, Italy, 2006. DOI: 10.1007/11787006_1

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis”, Proceedings of the Third conference on Theory of Cryptography, pp. 265-284, New York, NY 2006. DOI: 10.1007/11681878_14

[8] U. Erlingsson, V. Pihur, and A. Korolova, “RAPPOR: Randomized aggregatable privacy-preserving ordinal response”, Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 1054-1067, Scottsdale, AZ, USA, Nov. 2014. DOI: 10.48550/arXiv.1407.6981

[9] J. W. Kim and B. Jang. “Workload-aware indoor positioning data collection via local differential privacy”, IEEE Communications Letters, Vol. 23, No. 8, pp. 1352-1356, August 2019. DOI: 10.1109/LCOMM.2019.2922963

[10] J. W. Kim, J. H. Lim, S. M. Moon, and B. Jang, “Collecting health lifelog data from smartwatch users in a privacy-preserving manner”, IEEE Transactions on Consumer Electronics, Vol. 65, No. 3, pp. 369-378, Aug. 2019. DOI: 10.1109/TCE.2019.2924466

[11] M. E. Andres, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. “Geo-indistinguishability: Differential privacy for location-based systems”, Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 901-914, Berlin, Germany, Nov. 2013. DOI: 10.1145/2508859.2516735

[12] R. Ahuja, G. Ghinita, and C. Shahabi, “A utility-preserving and scalable technique for protecting location data with geo-indistinguishability”, Proceedings of the International Conference on Extending Database Technology, pp. 210-231, Lisbon, Portuga, April 2019.

[13] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidess, “Optimal geo-indistinguishable mechanisms for location privacy”, Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 251-262, New York, NY, USA, Nov. 2014. DOI: 10.1145/2660267.2660345

[14] K. Chatzikokolakis, E. ElSalamouny, and C. Palamidessi, “Efficient utility improvement for location privacy”, Proceedings on Privacy Enhancing Technologies, pp. 210-231, Minneapolis, USA, July 2017. DOI: 10.1515/popets-2017-0035

[15] Chawla S., Dwork C., McSherry F., and Talwar K, “On the Utility of Privacy-Preserving Histograms”, Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI, July 2005. DOI: 10.48550/arXiv.1207.1371

[16] Pozi, M.S.M. and Mohd. Hasbullah Omar, “A Kernel Density Estimation Method to Generate Synthetic Shifted Datasets in Privacy-Preserving Task”, Journal of Internet Services and Information Security, 10(4), pp. 70-89, Nov. 2020. DOI:10.22667/JISIS.2020.11.30.070

[17] Papamakarios G., Pavlakou T., and Murray I, “Masked Autoregressive Flow for Density Estimation”, Advances in Neural Information Processing Systems, Vol. 30, pp. 2338–2347, May 2017. DOI: 10.48550/arXiv.1705.07057

[18] Kingma D. P., Salimans T., Jozefowicz R., Chen X., Sutskever I., and Welling M, “Improved Variational Inference with Inverse Autoregressive Flow”, Advances in Neural Information Processing Systems, Vol. 29, pp. 4743-4751, 2016. DOI: 10.48550/arXiv.1606.04934

[19] Fakoor R., Chaudhari P., Mueller J., and Smola A. J, “TradE: Transformers for Density Estimation”. ArXiv preprint, April 2020. DOI: 10.48550/arXiv.2004.02441

[20] Huang C.-W., Krueger D., Lacoste A., and Courville A, “Neural

Autoregressive Flows", Proceedings of the 35th International Conference on Machine Learning, Vol. 80, pp. 2078-2087, 2018. DOI: 10.48550/arXiv.1804.00779

[21] Oliva J., Dubey A., Zaheer M., Poczos B., Salakhutdinov R., Xing E., and Schneider J, "Transformation Autoregressive Networks", Proceedings of the 35th International Conference on Machine Learning, Vol. 80, pp. 3898-3907, 2018. DOI: 10.48550/arXiv.1801.09819

[22] Wang L, Yang D, Han X, Wang T, Zhang D, Ma X. "Location Privacy-preserving Task Allocation for Mobile Crowdsensing with Differential Geo-obfuscation", Proceedings of the International Conference on World Wide Web, pp. 627-636, Perth, Australia, April 2017. DOI: 10.1145/3038912.3052696

[23] Tong W, Hua J, Zhong S, "A jointly differentially private scheduling protocol for ridesharing services", IEEE Transaction on Information Forensics and Security, Vol. 12, No. 10, pp. 2444-2456, 2017. DOI: 10.1109/TIFS.2017.2707334

[24] Ma C, Chen CW, "Nearby friend discovery with geo-indistinguishability to stalkers", Procedia Computer Science, Vol. 34, pp. 352-359, 2014. DOI: 10.1016/j.procs.2014.07.036

[25] Pan X, Zhang J, Wang F, Yu PS, "DistSD: Distance-based Social Discovery with Personalized Posterior Screening", Proceedings of the IEEE International Conference on Big Data, p. 1110-1119. Washington, DC, USA, Dec. 2016. DOI: 10.1109/BigData.2016.7840714

[26] Xue M, Liu Y, Ross KW, Qian H, "I Know Where You Are: Thwarting Privacy Protection in Location-based Social Discovery Services" Proceedings of the IEEE Conference on Computer Communications Workshops, pp. 179-184, Hong Kong, China, 2015. DOI: 10.1109/infcomw.2015.7179381

[27] Huang C, Lu R, Zhu H, Shao J, Alamer A, Lin X. "EPPD: Efficient and Privacy-preserving Proximity Testing with Differential Privacy Techniques", Proceedings of the IEEE International Conference on Communications, pp. 1-6, Kuala Lumpur, Malaysia, 2016. DOI: 10.1109/ICC.2016.7511194

[28] Ng, S.K., Krishnan, T., McLachlan, G.J., "The EM Algorithm", Springer, Berlin, Heidelberg, Dec 2011. DOI: 10.1007/978-3-642-21551-3_6

[29] T. K. Moon, "The expectation-maximization algorithm", in IEEE Signal Processing Magazine, Vol. 13, no. 6, pp. 47-60, Nov 1996. DOI: 10.1109/79.543975

[30] Kujawińska, A., M. Rogalewicz, and M. Diering. "Application of expectation maximization metod for purchase decision-making support in welding branch,‖ Management and Production Engineering Review", Vol. 7, No. 2, pp. 29-33, 2016. DOI: 10.1515/mper-2016-0014

[31] D. Zibar, O. Winther, R. Borkowski, I. T. Monroy, L. Carvalho and J. Oliveira, "Applications of expectation maximization algorithm for coherent optical communication", 22nd European Signal Processing Conference, pp. 1890-1894, 2014

[32] Elad Tzoreff, Anthony J. Weiss, "Expectation-maximization algorithm for direct position determination", Signal Processing, Vol. 133, pp. 32-39, 2017. DOI: 10.1016/j.sigpro.2016.10.015

[33] Yu Zheng, T-Drive trajectory data sample, https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample

Authors

Seung Min Song received the B.S. degree from Sangmyung University in 2020, where she is currently pursuing the master's degree with the Department of Computer Science. Her research mainly focuses on data privacy and Artificial Intelligence.

Jong Wook Kim received the Ph.D. degree in Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013.

Dr. Kim is currently an Associate Professor with the Department of Computer Science at Sangmyung University. His primary research interests include the area of data privacy, distributed databases, and query optimization.