JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 6, pp. 69-76, June 2024
https://doi.org/10.9708/jksci.2024.29.06.069

An Enhanced Data Utility Framework for Privacy-Preserving Location Data Collection

Jong Wook Kim*

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

[Abstract]

Recent advances in sensor and mobile technologies have made it possible to collect user location data. This location information is used as a valuable asset in various industries, resulting in increased demand for location data collection and sharing. However, because location data contains sensitive user information, indiscriminate collection can lead to privacy issues. Recently, geo-indistinguishability (Geo-I), a method of differential privacy, has been widely used to protect the privacy of location data. While Geo-I is powerful in effectively protecting users' locations, it poses a problem because the utility of the collected location data decreases due to data perturbation. Therefore, this paper proposes a method using Geo-I technology to effectively collect user location data while maintaining its data utility. The proposed method utilizes the prior distribution of users to improve the overall data utility, while protecting accurate location information. Experimental results using real data show that the proposed method significantly improves the usefulness of the collected data compared to existing methods.

▸ **Key words**: Location Data Collection, Data Utility, Data Privacy, Differential Privacy

[요 약]

최근 센서 기술과 모바일 기술의 급속한 발전으로 인하여 사용자 위치 데이터 수집이 가능해졌다. 사용자 위치 정보는 다양한 산업에서 중요한 자산으로 활용되고 있으며, 그 결과 위치 데이터의 수집 및 공유에 대한 수요가 증가하고 있다. 그러나 위치 정보에는 사용자의 민감한 데이터가 포함되어 있으므로, 무분별한 수집은 프라이버시 침해 문제를 일으킬 수 있다. 최근에는 차분 프라이버시의 한 방법으로 Geo-Indistinguishability (Geo-I)가 위치 데이터의 프라이버시 보호에 활용되고 있다. Geo-I는 사용자의 위치를 효과적으로 보호할 수 있는 강력한 방법을 제공하지만, 데이터 변조로 인해 수집된 위치 데이터의 유용성이 감소하는 문제가 있다. 따라서, 본 논문에서는 Geo-I 기술을 활용해 사용자 위치 데이터를 효과적으로 수집하면서 데이터의 유용성을 유지할 수 있는 방법을 제안한다. 제안 기법은 사용자의 사전 분포 정보를 활용하여 정확한 위치 정보를 보호하면서도 데이터의 전체적인 유용성을 향상시킨다. 실데이터를 이용한 실험 결과는 제안 기법이 기존 방법보다 수집된 데이터의 유용성을 상당히 향상시킬 수 있음을 보여준다.

▸ **주제어**: 위치 데이터 수집, 데이터 유용성, 개인정보 보호, 차분 프라이버시

• First Author: Jong Wook Kim, Corresponding Author: Jong Wook Kim
*Jong Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2024. 05. 02, Revised: 2024. 05. 29, Accepted: 2024. 05. 29.

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

I. Introduction

최근들어 센서 기술과 모바일 기술이 급속도로 발전함에 따라, 사용자의 위치 데이터를 수집하는 것이 가능해졌다. 스마트폰, 웨어러블 디바이스, GPS 장치와 같은 기기들을 통해 사용자의 정확한 위치 정보를 실시간으로 수집할 수 있게 되었다. 이러한 위치 정보는 맞춤형 마케팅, 실시간 교통 상황 분석, 개인화된 추천 서비스 등 다양한 방면에서 기업들에게 중요한 자산으로 활용되고 있다 [1,2,3,4]. 특히 소매업, 광고, 헬스케어, 교통 시스템 관리와 같은 산업에서 위치 데이터의 수집과 활용은 큰 경쟁력을 가지며, 사용자 경험을 향상시키고 비즈니스 성과를 극대화하는 데 필수적이다. 따라서 사용자 위치 데이터의 수집 및 공유에 대한 수요는 지속적으로 증가하고 있다.

그러나 사용자의 위치 정보는 사용자의 매우 민감한 정보(예, 주거지, 직장, 그리고 일상의 이동 패턴)를 포함하고 있기때문에, 이를 무분별하게 수집하는 것은 프라이버시 침해 문제를 발생시킬 수 있다 [5,6]. 위치 데이터가 잘못 사용되거나 유출될 경우, 사용자의 사생활이 침해받을 뿐만 아니라, 심각한 경우 안전까지 위협받을 수 있다. 가령, 개인의 이동 패턴을 분석하여 그 사람이 자주 방문하는 장소나 특정 시간대에 집을 비우는 패턴을 파악할 수 있으며, 이 정보가 범죄로 이어질 수 있다. 또한, 광고업체들이 동의없이 개인의 위치 정보를 활용하여 개인화된 광고를 보내는 등의 방식으로 프라이버시를 침해할 수 있다.

이러한 문제를 해결하기 위해, 프라이버시를 보존하면서 사용자의 위치 데이터를 수집하기 위한 많은 연구가 진행되었다. 전통적인 방법으로는 데이터의 익명화 기술 [7]과 암호화 기법 [8] 등이 있다. 최근에는 차분 프라이버시 (Differential Privacy) [9,10] 가 데이터 수집 시 프라이버시를 보존하는 데 중요한 기술로 자리잡게 되었다. 차분 프라이버시는 개인 정보가 포함된 데이터베이스에 잡음을 추가하여 개인을 식별할 수 없게 하면서도 데이터의 전체적인 패턴은 유지할 수 있도록 하는 기술이다. 이 기법은 통계적 분석과 데이터 공유에 있어서 개인의 프라이버시 침해 가능성을 크게 줄여준다.

위치 데이터의 프라이버시를 보호하기 위해 차분 프라이버시에 거리 개념을 확장하여 적용하는 연구가 활발히 진행되고 있다. 이 중 대표적인 기술로 Geo-Indistinguishability (Geo-I) [11,12]가 있다. Geo-I는 사용자의 위치 데이터에 의도적으로 잡음을 추가함으로써 사용자의 정확한 위치를 숨기는 동시에 위치 기반 서비스 (location based service, LBS)가 요구하는 기본적인 유용성을 유지할 수 있게 하는 기술이다. Geo-I는 사용자와 관련된 위치 데이터 포인트에 대해, 그 거리가 증가함에 따라 식별 가능성이 감소하도록 잡음을 조정한다 [11]. 이를 통해, 사용자가 실제로 위치한 지점과 가까운 다른 지점들 사이에서 구별이 어려워지게 만드는 방식으로 사용자의 정확한 위치 정보를 보호한다.

Geo-I는 LBS 환경뿐만 아니라 사용자 위치 데이터의 수집에도 효과적으로 활용될 수 있는 기술이다. Geo-I는 차분 프라이버시의 원칙을 바탕으로 하며, 사용자의 실제 위치 데이터에 잡음을 추가하여 위치 정보를 변조함으로써 개인의 위치 프라이버시를 보호한다. 이러한 접근 방식은 사용자의 위치를 외부로부터 보호할 수 있는 강력한 방법을 제공하지만, 데이터 변조로 인하여 수집된 위치 데이터의 정확성이 감소하는 문제가 있다 (그림 1). 이는 차분 프라이버시 기법이 공통적으로 가지는 한계이다.

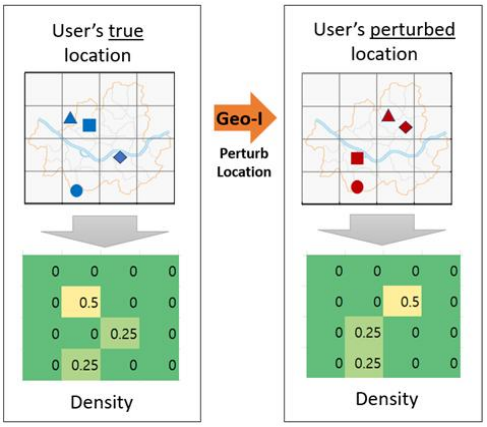


Fig. 1. Reduced accuracy resulting from location data collection under Geo-I

그러므로 본 논문에서는 Geo-I 기술을 활용하여 사용자의 위치 데이터를 효과적으로 수집하면서도 데이터 유용성을 보존할 수 있는 방법을 제안한다. 본 연구는 사용자의 사전 분포(prior distribution) 정보를 활용하여, 사용자의 정확한 위치는 보호하면서도, 전체적인 데이터 유용성(예, 사용자 밀집도 같은 통계적 정보)을 향상 시키는 방법을 제안한다.

본 논문은 다음과 같이 구성된다. 2장에서는 본 논문과 관련된 선행 연구들을 살펴보고, 3장에서는 본 연구의 이론적 배경에 대하여 설명한다. 4장에서는 본 논문에서 제안하는 기술에 대해 자세히 논의하며, 5장에서는 실데이터를 이용하여 제안 기술의 성능을 실험적으로 평가한다. 마지막으로, 6장에서는 연구 결과를 요약하고 결론을 맺는다.

II. Related Work

최근 사용자의 위치 정보를 프라이버시를 보호하면서 수집하기 위한 다양한 연구가 진행되었다. [13]은 지역 차분 프라이버시를 활용하여 사용자의 실내 위치 데이터를 수집하는 방법을 제안하였다. 제안 기법은 주어진 데이터 분석 워크로드의 정확성을 최대화하는 위치 데이터 수집 방법을 개발하였다. [14,15]는 모바일 크라우드 환경에서 작업자의 위치 데이터를 보호하기 위해 Geo-I를 사용하였다. 제안 방법은 Geo-I를 이용해 작업자의 변조된 위치 데이터를 수집한 후, 이 데이터를 활용하여 작업자들에게 작업을 할당하였다. EGeoIndis [16]은 Geo-I 기반의 차량 위치 프라이버시 보호 프레임워크로, 교통 밀도를 추정하는 데 사용된다. EGeoIndis은 교통 밀도 추정 과정에서 차량의 위치 정보의 프라이버시를 보호하기 위해 Geo-I 기술을 적용하였다. [17]에서는 COVID-19 증상을 보이는 환자들의 위치를 프라이버시를 보호하면서 수집하기 위해 Geo-I를 이용하였다. Geo-I를 통해 수집된 환자의 위치 데이터는 COVID-19 감염병 지도를 만드는 데 활용되었다.

위치 데이터 이외에 다양한 분야에서 프라이버시를 보호하면서 사용자의 민감한 데이터를 수집하기 위한 다양한 연구가 진행되었다. [18]은 Geo-I를 활용하여 사용자의 민감한 텍스트 마이크로데이터를 안전하게 수집하는 방법을 제안하였다. 제안 기법은 텍스트 데이터의 각 단어를 워드 임베딩을 통해 벡터로 변환하고, ε-Geo-I를 만족하도록 이 벡터 데이터에 잡음을 추가한다. 실제 단어는 잡음이 추가된 벡터와 가장 근접한 위치에 해당하는 단어로 대체되며, 이 대체 단어를 데이터 수집 서버로 전송한다. [19]는 지역 차분 프라이버시를 이용하여 사용자로부터 텍스트 데이터를 수집하기 위한 방법을 제안하였다. 이렇게 수집된 데이터는 인공지능 모델 학습에 사용된다. [20]은 Geo-I를 활용하여 환자들의 병명 데이터를 안전하게 수집하는 기법을 제안하였다. [20]에서는 병명과 같은 민감한 데이터를 효과적으로 수집하기 위해 ε-Geo-I를 만족하는 효과적인 데이터 변조 방법을 제안하였다. [21]에서는 스마트워치 사용자의 심박수와 누적 걸음수와 같은 건강 데이터를 지역 차분 프라이버시를 사용하여 프라이버시를 보호하면서 수집하는 방법을 제안하였다. 제안 기법은 전체 데이터를 직접 수집하는 대신, 먼저 데이터에서 중요한 특징점을 추출한 후, 지역 차분 프라이버시를 이용하여 추출된 특징점들은 수집한다. 데이터 수집 서버는 수집한 특징점들을 기반으로 전체 데이터를 복원한다.

III. Background

Geo-I는 차분 프라이버시의 한 방식으로, 사용자의 위치 데이터에 잡음을 추가하여 변조함으로써 공격자가 사용자의 정확한 위치를 파악하는 것을 어렵게 하는 프라이버시 보호 기술이다.

정의 1. (ε-Geo-Indistinguishability) 사용자의 실제 위치 데이터 집합을 X, 사용자가 서버에 전송한 변조된 위치 데이터 집합을 Y라고 각각 가정하자. M을 임의의 매커니즘이라고 가정하자. 이때, X의 임의의 데이터 x_1, x_2와 M으로부터 생성되는 모든 결과값 $y \in Y$에 대하여 다음 식을 만족하면, M은 ε-Geo-I를 만족한다.

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

식(1)에서 $d(x_1,x_2)$는 x_1과 x_2 사이의 거리에 해당한다. 또한, $M(x_1)(y)$는 매커니즘 M이 입력 x_1으로부터 y를 무작위로 생성하는 과정을 의미한다. ε는 차분 프라이버시 종류의 기법에서 사용하는 프라이버시 예산(privacy budget)이다. Geo-I는 차분 프라이버시에 거리 개념을 통합한 기법으로, $d(x_1,x_2)=1$인 경우, 식(1)은 차분 프라이버시 정의와 같아진다.

Geo-I를 구현하는 데에는 두 가지 주요 방법이 있다. 첫 번째 방법은 라플라스 매커니즘을 활용하는 것이다. 이 기법은 사용자의 실제 위치 데이터에 라플라스 분포를 따르는 잡음을 추가하여 위치 데이터를 변조한다 [11]. 라플라스 매커니즘은 구현이 간단하지만, 위치 데이터 변조 과정에서 상대적으로 많은 잡음이 추가될 수 있어 데이터의 유용성이 낮아질 수 있다.

두 번째 방법은 최적화 기법[12,22]을 사용하는 것이다. 이 기법은 사용자의 사전 분포(prior distribution)가 주어진 경우 라플라스 매커니즘에 비해 데이터 유용성이 상대적으로 높다는 장점이 있다. 최적화 방식은 전체 영역을 격자로 나누고, 사용자의 위치를 이 격자 상에 표현한다. 이후 선형 프로그래밍을 통해 사용자 위치 정보를 확률적으로 변조하기 위한 변조 행렬을 생성한다. 최적화 기법은 데이터 유용성을 높이는 장점이 있지만, 변조 행렬을 생성하는 데 선형 프로그램에 의존하므로, 계산 복잡도가 높다는 단점이 있다. 그러므로 격자의 개수가 많아지면(즉, 정밀하게 사용자의 위치를 수집하는 경우) 최적화 기법을 적용하기 어려운 문제가 발생한다.

[20]의 제안 기법은 최적화 기법과 유사하게 사용자의 사전 분포 정보를 활용하지만, 선형 프로그래밍을 사용하

지 않기 때문에 계산 복잡도가 낮은 장점이 있다. 이로 인해 격자 수가 많은 경우에도 적용이 가능하다. [20]의 제안 기법은 전체 영역이 m개의 격자 $G=\{g_1, g_2, ..., g_m\}$로 구성된 경우, $m \times m$ 크기의 변조 행렬 O를 다음과 같이 정의한다.

$$O[i,j] = \frac{\Theta(p_{g_i}) \times e^{-\frac{\epsilon}{2} \times d(g_i, g_j)}}{\sum_{g_k \in G} \Theta(p_{s_k}) \times e^{-\frac{\epsilon}{2} \times d(g_i, g_k)}} \quad \text{식(2)}$$

이때, $O[i,j]$는 변조된 위치 데이터 g_j가 실제 위치 데이터 g_i로부터 무작위로 생성될 확률을 나타낸다 (즉 $O[i,j] = M(g_i)(g_j)$). 또한, p_{g_i}는 격자 g_i의 사전 분포 정보를 나타내며, $\Theta(\bullet)$는 임의의 단조 증가(monotone increasing) 함수에 해당한다.

그림 2에서 볼 수 있듯이, LBS 서버는 먼저 변조 행렬 O를 생성하고 이를 사용자에게 배포한다. 사용자는 배포된 변조 행렬을 이용하여 자신의 실제 위치 데이터를 변조한 후, 변조된 데이터를 서버에 전송한다. 본 연구에서는 [20]의 제안 기법을 바탕으로 사전 분포 지식을 활용하여 사용자의 위치 정보를 효과적으로 수집하는 방법을 제안한다.

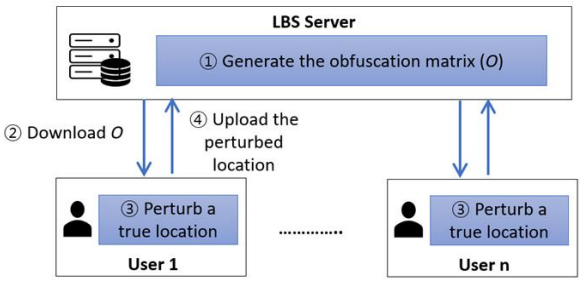


Fig. 2. Privacy-preserving location data collection using Geo-I

IV. Proposed Method

본 장에서는 제안 기법을 설명한다. 제안 기법은 위치 데이터 수집 과정에서 얻은 변조된 위치 데이터를 기반으로 사용자의 분포에 대한 사전 분포 정보를 추출하고, 이 정보를 다시 데이터 수집 과정에 활용한다.

4.1 Estimation Prior Distribution

특정 응용 프로그램의 경우 사용자 분포에 대한 사전 분포 정보가 미리 제공될 수 있다. 그러나 모든 응용프로그램 환경에서 이러한 정보가 항상 제공되는 것은 아니다. 따라서 본 연구에서는 데이터 수집 과정 중에 얻은 변조된 사용자 위치 정보로부터 사용자의 분포에 대한 사전 분포 정보를 추출하는 방법을 사용한다.

그러나 데이터 수집 과정에서 얻은 위치 데이터만을 활용하여 사용자의 분포를 추정하는 것은 Geo-I에 의한 데이터 변조로 인해 정확한 측정이 어려울 수 있다. 그러므로 본 논문에서는 변조 행렬에 인코딩된 실제 위치 데이터와 변조된 위치 데이터 간의 확률적 매핑 정보를 활용하여 변조된 위치 데이터 집합으로부터 사용자 밀도 분포를 측정한다.

전체 영역이 m개의 격자 $G=\{g_1, g_2, ..., g_m\}$로 구성되어 있다고 가정하자. 특정 시점에 데이터 수집 과정 Geo-I를 이용하여 수집한 변조된 위치 데이터 집합을 DB라고 가정하자. 또한, $cnt(g_i, DB)$는 격자 g_i가 DB에 나타나는 빈도수를 의미한다고 가정하자. 이때, 격자 g_i의 밀도율 p_{g_i}는 변조된 위치 데이터와 실제 위치 데이터 간의 매핑 확률 정보를 이용하여 다음과 같이 구할 수 있다.

$$p_{g_i} = \frac{\sum_{g_j \in G} (O[i,j] \times cnt(g_j, DB))}{\sum_{g_k \in G} p_{g_k}} \quad \text{식(3)}$$

식(3)에서 분모는 밀도율의 합을 1로 만들기 위한 정규화 요소에 해당한다. 식(3)은 변조된 위치 데이터 g_j가 실제 위치 데이터 $g_1, g_2, ..., g_m$로부터 변조 행렬에 인코딩된 특정 확률로 각각 생성된다는 특성을 활용한 것이다.

4.2 Privacy-preserving Location Data Collection Leveraging Prior Distribution

본 절에서는 사용자 분포에 대한 사전 분포 정보를 활용한 프라이버시 보존 위치 데이터 수집 기법에 관하여 설명한다.

4.2.1 Using Latest Prior Information

첫 번째 방식은 가장 최근에 수집된 사용자의 변조된 위치 데이터로부터 추출한 밀도 분포를 사전 분포 정보로 활용하는 것이다. 알고리즘 1은 첫 번째 기법의 의사코드를 나타낸다. 초기에 사용자 밀도 분포는 균일하게 설정된다. 3번에서는 현재 밀도 분포를 사전 분포 정보로 활용하여 변조 행렬 O를 식(2)를 이용해 계산한다. 이후, 변조 행렬 O를 사용자에게 배포하고, 사용자로부터 변조된 위치 데

이터를 Geo-I를 만족하도록 수집한다 (4~5번). 6번에서는 이전 타임 스탬프에서 수집한 변조된 위치 데이터 DB를 사용하여 식(3)에 따라 사용자 밀도 분포를 업데이트한다. 업데이트된 밀도 분포는 다음 타임 스탬프 동안의 사용자 위치 데이터 수집시 사전 분포로 활용된다.

Algorithm 1. Privacy-preserving location data collection using the latest prior distribution

```
1: [pg1, pg2, ...,pgm] = [1/m, 1/m, ..., 1/m]
2: while (true)
3:     O = Perturbation_Matrix([pg1, pg2, ...,pgm])
4:     Distribute_Matrix_Users(O)
5:     DB = Collect_Data()
6:     [pg1, pg2, ...,pgm] = Compute_Prior(DB)
```

4.2.2 Using Cumulative Prior Information

두 번째 방식은 현재까지 수집된 누적된 사용자의 변조 위치 데이터로부터 밀도 분포를 추출하고, 이를 사전 분포 정보로 활용하는 방법이다. 알고리즘 2는 이 기법의 의사 코드를 보여준다. 초기에는 사용자 밀도 분포를 균일하게 설정되고, DB는 공집합으로 초기화한다. 알고리즘 1과 유사하게 Geo-I를 만족하도록 사용자의 위치 데이터를 수집한 후, 이를 $DB_{current}$에 임시로 저장한다(4~6번). 7번에서 $DB_{current}$를 DB에 추가한 후, 8번에서 축적된 사용자 위치 데이터 집합 DB를 이용하여 사용자 밀도 분포를 업데이트한다.

Algorithm 2. Privacy-preserving location data collection using the cumulative prior distribution

```
1: [pg1, pg2, ...,pgm] = [1/m, 1/m, ..., 1/m]
2: DB = ∅
3: while (true)
4:     O = Perturbation_Matrix([pg1, pg2, ...,pgm])
5:     Distribute_Matrix_Users(O)
6:     DBcurrent = Collect_Data()
7:     DB = DB ∪ DBcurrent
8:     [pg1, pg2, ...,pgm] = Compute_Prior(DB)
```

4.2.3 Using KL Divergence

알고리즘 1과 2는 각 타임 스탬프마다 사용자 밀도 분포를 업데이트하고, 이를 바탕으로 변조 행렬을 업데이트한다. 반면, 알고리즘 3은 현재 사용 중인 사용자 밀도 분포와 새롭게 계산된 사용자 밀도 분포의 차이를 Kullback-Leibler divergence (KLD)를 사용하여 측정한 후, 이 차이가 사전에 설정된 임계값을 초과할 경우에만 사용자 밀도 분포와 변조 행렬을 업데이트한다. KLD를 활용함으로써, 밀도 분포 간의 차이가 클 때만 변조 행렬을 재계산하여 사용자에게 배포한다. 그렇지 않을 경우, 기존의 변조 행렬을 다음 데이터 수집 단계에 계속 사용한다. 이 접근법은 사용자 밀도 분포 변화가 미미할 때 불필요한 계산을 줄이고 효율성을 높일 수 있다.

10번에서 지금까지 수집한 사용자의 변조된 위치 데이터를 기반으로 사용자 밀도 분포 $Q=[p'_{g_1}, p'_{g_2}, ..., p'_{g_m}]$를 계산한다. 이어서 11번에서는 현재 사용 중인 사용자 밀도 분포 $P=[p_{g_1}, p_{g_2}, ..., p_{g_m}]$와 새롭게 구한 Q 사이의 KLD를 계산한다. 두 밀도 분포간의 KLD를 계산하는 방법은 다음과 같다.

$$D_{KL}(P\|Q)=\sum_{g_i\in G} p_{g_i}\log\frac{p_{g_i}}{p'_{g_i}} \qquad \text{식(4)}$$

만약 두 밀도 분포 간의 KLD가 사전에 설정한 임계값 θ를 초과하는 경우, 사용자 밀도 분포를 업데이트하고, DB를 공집합으로 초기화한 후, 'update' 변수를 true로 설정한다(12~15번). 'update' 변수가 true로 설정되었기 때문에, 다음 타임 스탬프에서는 변조 행렬을 업데이트하고, 새로운 변조 행렬을 사용하여 사용자 위치 데이터의 수집을 계속 진행한다.

Algorithm 3. Privacy-preserving location data collection using KL Divergence

```
 1: [pg1, pg2, ...,pgm] = [1/m, 1/m, ..., 1/m]
 2: DB = ∅
 3: update = true
 4: while (true)
 5:     if update=true
 6:         O = Perturbation_Matrix([pg1, pg2, ...,pgm])
 7:         Distribute_Matrix_Users(O)
 8:     DBcurrent = Collect_Data()
 9:     DB = DB ∪ DBcurrent
10:     [p'g1, p'g2, ...,p'gm] = Compute_Prior(DB)
11:     div = KL([pg1,pg2,...,pgm], [p'g1,p'g2,...,p'gm])
12;     if (div > θ)
13:         [pg1,pg2,...,pgm] = [p'g1,p'g2,...,p'gm]
14:         DB = ∅
15:         update=true
16:     else
17:         update=false
```

V. Experiments and Results

5.1. Experiment Setup

본 연구에서는 성능 평가를 위해 Porto 택시 데이터셋[22]을 사용하였다. 실험을 위해 Porto 택시 데이터셋에서 50,000개의 이동 경로 데이터를 추출하였으며, 각 이동 경로는 30개의 위치 정보로 구성되어 있다. 따라서 실험에서는 각 이동 경로로부터 총 30번의 위치 데이터를 수집하였다. 또한, 실험에 사용된 격자의 수(m)는 1040개이다. 성

능 비교를 위해 다음과 같은 기법들을 사용하여 결과를 비교하였다.

- P_Last: 4.2.1절에서 설명한 가장 최근에 수집된 변조 위치 데이터를 활용하는 방법
- P_Cum: 4.2.2절에서 설명한 누적된 변조 위치 데이터를 활용하는 방법
- P_KL: 4.2.3절의 설명한 KLD를 기반으로 하는 방법
- NA: 균등 사전 분포 정보를 활용하는 단순 기법(즉, $[p_{g_1}, p_{g_2}, ..., p_{g_m}] = [1/m, 1/m, ..., 1/m]$으로 항상 설정함)

P_KL의 경우 임계값(θ)으로 0.1을 사용하였다. 즉, KLD 값이 0.1를 초과하는 경우, 사용자 밀도 분포 및 변조 행렬을 업데이트 한다. 실험에서는 성능 측정을 위해 평균 절대 오차(Mean Absolute Error, MAE)를 사용했다. MAE는 다음과 같이 정의된다.

$$MAE = \frac{1}{m} \times \sum_{g_i \in G} \left| f_i^{true} - f_i^{perturb} \right| \qquad 식(4)$$

이때, f_i^{true}는 격자 g_i에 위치한 실제 택시의 수를 나타내고, $f_i^{perturb}$는 변조된 위치 데이터를 수집한 후, 이를 바탕으로 계산한 격자 g_i에 위치한 택시 수를 나타낸다.

5.2. Experimental Results

그림 3은 Geo-I의 프라이버시 예산(ε) 변화에 따른 MAE 변화를 보여준다. 실험에서는 ε 값으로 0.5, 1.0, 2.0, 3.0을 사용하였다. 그림 3의 실험 결과에서 알 수 있듯이, ε 값이 낮아질수록 MAE 값이 증가하는 경향이 있음을 알 수 있다. 이는 ε 값이 낮아질수록 사용자 프라이버시 보호가 강화되며, 이에 따라 원본 위치 데이터에 대한 변조가 증가하여 수집한 위치 데이터 정확도가 감소하는 것을 의미한다. 반대로 ε 값이 증가할 때, MAE 값이 줄어드는 것을 관찰할 수 있는데, 이는 프라이버시 보호 수준이 낮아지면서 원본 위치 데이터의 변조가 줄어들기 때문이다. 이러한 경향은 차분 프라이버시를 적용하는 경우 일반적으로 관찰되는 현상이며, 본 연구의 제안 기법 또한 같은 경향을 나타내고 있다.

그림 3의 실험 결과를 통해 확인할 수 있듯이, 본 논문의 제안 기법인 P_Last, P_Cum, P_KL은 모든 프라이버시 예산에 대하여 단순 기법인 NA에 비해 뛰어난 성능을 보여주고 있다. 이 결과는 제안 기법들이 사용자의 위치 데이터 수집에 있어 사전 분포 정보를 효과적으로 활용하고 있음을 입증한다. 또한 세 가지 제안 기법 중 P_Last가 가장 우수한 성능을 나타내는 반면, P_Cum은 가장 낮은 성능을 보여주고 있다. 이러한 결과는 사용자의 현재 위치는 바로 직전의 위치에 크게 의존하기 때문이다. 그러므로 가장 최근의 데이터를 통해 얻은 사용자 밀도 분포 정보를 사전 분포 정보로 활용하는 것이 효과적임을 보여주고 있다.

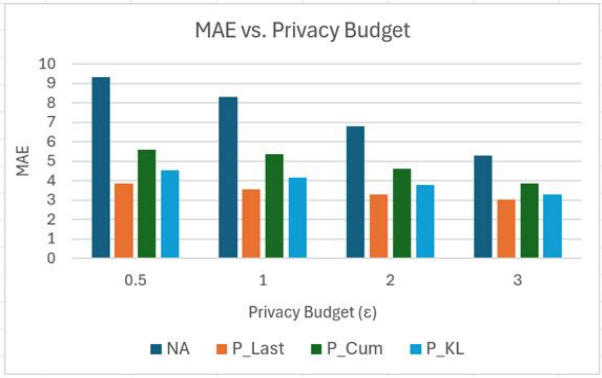
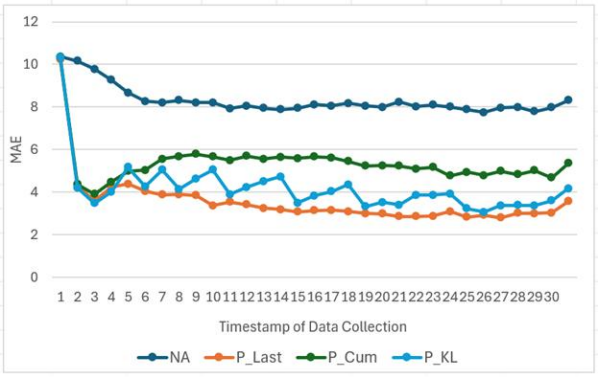


Fig. 3. MAE for varying privacy budget (ε)

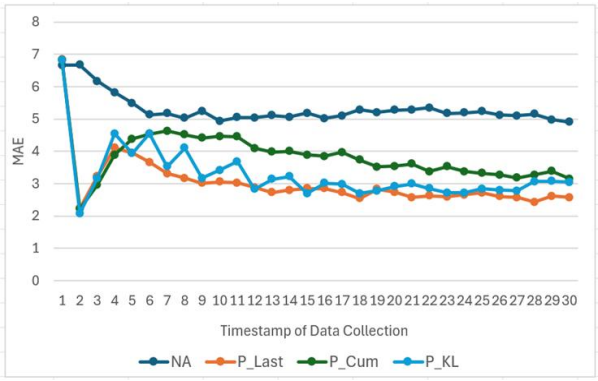
Table 1. The number of updates to the perturbation matrix during 30 data collection periods

	P_Last	P_Cum	P_KL
ε = 0.5	30	30	9
ε = 1.0	30	30	10
ε = 2.0	30	30	10
ε = 3.0	30	30	11

그림 3의 실험 결과에 따르면, P_KL 방법은 P_Last에 비해 다소 낮은 성능을 보이지만, 여전히 근접한 수준의 성능을 제공하고 있다. P_KL의 주요 장점 중 하나는 변조 행렬을 지속적으로 업데이트할 필요가 없다는 것으로, 이를 통해 보다 효율적으로 사용자 데이터를 수집할 수 있다. 이 점은 표 1을 통해 확인할 수 있다. 표 1은 30회의 데이터 수집 동안 변조 행렬이 몇 번 업데이트 되었는지를 나타낸다. P_Last와 P_Cum은 각 수집마다 변조 행렬을 업데이트해야 하므로 총 30회의 업데이트가 필요하다. 반면 P_KL은 KLD 값을 기준으로 현재의 사용자 밀도 분포와 새로 계산된 분포 사이의 차이가 설정된 임계값을 초과할 경우에만 업데이트를 진행한다. 본 실험에서는 프라이버시 예산에 따라 9~11회 사이의 업데이트가 이루어졌다. 그러므로 시간적 측면에서 P_KL은 P_Last와 P_Cum보다 효율적임을 알 수 있다. 그림 3의 결과에서 P_KL이 P_Last와 비슷한 수준의 성능을 보여준다는 점을 고려하면, 계산 복잡도가 중요한 요인이 되는 응용 프로그램에서는 P_Last 대신 P_KL를 사용하여 계산 비용을 절감하고 일정 수준의 데이터 유용성을 유지할 수 있다.



(a) Privacy budget (ε) = 1.0



(b) Privacy budget (ε) = 3.0

Fig. 4. MAE for each data collection

그림 4는 각 위치 데이터 수집 시점별 MAE 값의 변화를 보여주고 있다. 실험 결과에 따르면, 본 논문의 제안 기법인 P_Last, P_Cum, P_KL이 모든 경우에 걸쳐 단순 방법인 NA보다 우수한 성능을 보여주고 있다. 또한 그림 3의 실험 결과와 마찬가지로, 제안 기법중에서는 P_Last가 가장 높은 성능을 보이는 반면, P_Cum은 가장 낮은 성능을 보이고 있다. P_KL은 P_Last와 P_Cum의 중간 수준의 성능을 나타내며, 변조 행렬을 업데이트할 때 P_Last와 비슷한 성능을 보이고 있음을 알 수 있다.

실험 결과를 통해 제안 기법이 단순 방법보다 수집된 데이터의 유용성이 더 높음을 입증하였다. 특히, P_Last가 제안 기법들 중 가장 우수한 성능을 보이는 것으로 나타났다. 또한, P_KL는 P_Last와 유사한 성능을 제공하면서 계산 복잡도를 줄일 수 있는 장점이 있음을 확인할 수 있었다.

VI. Conclusions

본 논문에서는 Geo-I 기술을 활용하여 사용자의 위치 데이터를 효과적으로 수집하고 데이터의 유용성을 유지하는 방법을 제안하였다. 특히, 사용자의 사전 분포 정보를 활용하여 사용자의 정확한 위치를 보호하면서 전체 데이터의 유용성을 향상시키는 방법을 개발하였다. 실데이터를 이용한 실험을 통해, 제안 기법이 단순 방법보다 우수한 성능을 보이고 있음을 입증하였다. 특히, 실험을 통하여 P_Last가 제안 방법 중에서 가장 우수한 성능을 보였으며, P_KL는 P_Last와 유사한 성능을 제공하면서 계산 복잡도를 낮출수 있음을 입증하였다.

향후 연구에서는 사용자의 이동 패턴이 복잡한 상황에서 제안 기법의 유용성을 검증할 필요가 있다. 이를 위해 다양한 복잡한 이동 시나리오를 설정하고, 제안 기법이 이러한 상황에서도 데이터를 효과적으로 보호하고 유용성을 유지할 수 있는지 평가할 예정이다.

REFERENCES

[1] P. Xie, T. Li, J. Liu, S. Du, X. Yang, and J. Zhang. Urban flow prediction from spatiotemporal data using machine learning: A survey. Information Fusion, vol. 59, 2020. DOI: 10.1016/j.inffus.2020.01.002

[2] Ni. G. Polson and V. O. Sokolov. Deep learning for short-term traffic flow prediction. Transportation Research Part C: Emerging Technologies, vol. 79, 2017. DOI: 10.1016/j.trc.2017.02.024

[3] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang. A hybrid deep learning based traffic flow prediction method and its understanding. Transportation Research Part C: Emerging Technologies, vol. 90, 2018. DOI: 10.1016/j.trc.2018.03.001

[4] A. Almeida, S. Bras, I. Oliveira, and S. Sargento. Vehicular traffic flow prediction using deployed traffic counters in a city. Future Generation Computer Systems, vol. 128, Mar. 2022. DOI: 10.1016/j.future.2021.10.022

[5] G. Ghinita. Privacy for location-based services. Synthesis Lectures on Information Security, Privacy, and Trust, vol. 17, no. 10, 2013.

[6] H. Jiang, J. Li, P. Zhao, F. Zeng, Z. Xiao, and A. Iyengar. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. ACM Computing Surveys, vol. 54, no.4, 2021. DOI: 10.1145/3423165

[7] L. Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, vol. 10, no. 5, 2002. DOI: 10.1142/S0218488502001648

[8] S . Mascetti, D. Freni, C. Bettini, X. Wang, and S. Jajodia. Privacy in geo-social networks: Proximity notification with untrusted service providers and curious buddies. The International Journal on Very Large Data Bases, vol. 20, no. 4, 2011. DOI: 10.1007/

s00778-010-0213-7

[9] C. Dwork. Differential privacy. in Proceedings of the International Conference on Automata Languages Program, Venice, Italy, 2006. DOI: 10.1007/11787006_1

[10] J. W. Kim, K. Edemacu, J. S. Kim, Y. D. Chung, and B. Jang. A survey of differential privacy-based techniques and their applicability to location-Based services. Computers & Security, vol. 111, 2021. DOI: 10.1016/j.cose.2021.102464

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

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

[13] J. W. Kim and B. Jang. Workload-aware indoor positioning data collection via local dierential privacy. IEEE Communications Letters, vol. 23, no. 8, 2019. DOI: 10.1109/LCOMM.2019.2922963

[14] L. Wang, D. Yang, X. Han, T. Wang, D. Zhang, and X. Ma. Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation. in Proceedings of the International Conference on World Wide Web, Perth, Australia, 2017. DOI: 10.1145/3038912.3052696

[15] W. Jin, M. Xiao, L. Guo, L. Yang, and M. Li. ULPT: A user-centric location privacy trading framework for mobile crowd sensing. IEEE Transactions on Mobile Computing, Early Access, 2021. DOI: 10.1109/TMC.2021.3058181

[16] W. Ren and S. Tang. EGeoIndis: An effective and efficient location privacy protection framework in traffic density detection. Vehicular Communications, vol. 21, 2020. DOI: 10.1016/j.vehcom.2019.100187

[17] R. Chen, L. Li, J. J. Chen, R. Hou, Y. Gong, Y. Guo, and M. Pan. COVID-19 vulnerability map construction via location privacy preserving mobile crowdsourcing. in Proceedings of IEEE Conference and Exhibition on Global Telecommunications, 2020. DOI: 10.1109/GLOBECOM42002.2020.9348141

[18] O. Feyisetan, B. Balle, T. Drake, and T. Diethe, Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In Proceedings of the International Conference on Web Search and Data Mining, Houston, TX, USA, February 2020. DOI: 10.1145/3336191.3371856

[19] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S.M., Chow, Differential privacy for text analytics via natural text sanitization. arXiv2021, arXiv:2106.01221. DOI: 10.48550/arXiv.2106.01221

[20] S. Song and J.W. Kim, Adapting Geo-Indistinguishability for Privacy-Preserving Collection of Medical Microdata, Electronics, vol. 12, 2023. DOI: 10.3390/electronics12132793

[21] J. W. Kim, B. Jang, and H. Yoo. Privacy-preserving aggregation of personal health data streams. PLoS ONE, vol. 13, no. 11, 2018. DOI: 10.1371/journal.pone.0207639

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

[23] Porto Taxi Trajectory Data, https://www.kaggle.com/datasets/crailtap/taxi-trajectory/data.

Authors

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.