JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 5, pp. 93-100, May 2024
https://doi.org/10.9708/jksci.2024.29.05.093

Privacy-Preserving Collection and Analysis of Medical Microdata

Jong Wook Kim*

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

[Abstract]

With the advent of the Fourth Industrial Revolution, cutting-edge technologies such as artificial intelligence, big data, the Internet of Things, and cloud computing are driving innovation across industries. These technologies are generating massive amounts of data that many companies are leveraging. However, there is a notable reluctance among users to share sensitive information due to the privacy risks associated with collecting personal data. This is particularly evident in the healthcare sector, where the collection of sensitive information such as patients' medical conditions poses significant challenges, with privacy concerns hindering data collection and analysis. This research presents a novel technique for collecting and analyzing medical data that not only preserves privacy, but also effectively extracts statistical information. This method goes beyond basic data collection by incorporating a strategy to efficiently mine statistical data while maintaining privacy. Performance evaluations using real-world data have shown that the propose technique outperforms existing methods in extracting meaningful statistical insights.

▸ **Key words**: Disease Data Collection, Data Privacy, Differential Privacy, Aggregate Analysis

[요 약]

4차 산업혁명의 도래와 함께 인공지능, 빅데이터, 사물인터넷, 클라우드 컴퓨팅 등의 첨단 정보기술이 다양한 산업 분야에서 혁신을 이끌고 있다. 이 기술들은 방대한 양의 데이터를 생성하고 있으며, 많은 기업들이 이를 활용하고 있다. 그러나 개인 데이터 수집 시 발생할 수 있는 프라이버시 침해 위험으로 인해 사용자들은 민감한 정보 제공을 망설이고 있다. 특히 의료 분야에서는 환자의 병명과 같은 민감한 정보 수집이 큰 도전이 되고 있으며, 프라이버시 문제가 데이터 수집과 분석의 장애가 되고 있다. 본 연구는 프라이버시 보호를 유지하면서도 통계적 정보를 효과적으로 추출할 수 있는 의료 데이터 수집 및 분석 기법을 제안한다. 제안 기법은 기존의 단순한 데이터 수집을 넘어서, 프라이버시를 보장하면서 수집된 데이터에서 통계적 정보를 효과적으로 추출하는 방법을 포함한다. 실제 데이터를 이용한 성능 평가에서는 제안된 기법이 기존 방법보다 더 효과적으로 프라이버시를 보존하며 통계적 정보를 도출할 수 있음을 입증한다.

▸ **주제어**: 병명 데이터 수집, 개인정보 보호, 차분 프라이버시, 집계 분석

• 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. 04. 24, Revised: 2024. 05. 23, Accepted: 2024. 05. 23.

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

I. Introduction

4차 산업혁명의 도래는 인공지능, 빅데이터, 사물인터넷, 클라우드 컴퓨팅 등의 첨단 정보 기술을 중심으로 수많은 산업 분야에서 혁신적인 변화를 촉진하고 있다. 이러한 기술들은 건강관리, 제조업, 금융 서비스, 교육, 농업 등 다양한 분야에 광범위하게 적용되며, 이 과정에서 대량의 데이터가 생성되고 있다 [1,2]. 가령, 제조업 분야에서는 스마트 팩토리의 도입이 진행되고 있다. 스마트 팩토리는 센서와 연결된 기계들을 통해 생산 과정의 모든 단계를 모니터링하며, 수집된 데이터를 분석해 생산 효율을 최적화하고 고장을 예방하고 있다 [3].

현대 기업들은 개인화된 서비스 제공, 시장 트렌드 예측 등을 목적으로 다양한 개인 데이터를 수집하고 있다. 그러나 이러한 데이터 수집은 종종 사용자의 동의 없이 이루어지기도 하며, 수집된 데이터의 보안이 충분히 확보되지 않는 경우가 많다. 특히 민감한 데이터를 수집할 때는 개인의 프라이버시를 심각하게 위협할 수 있다. 예를 들어, 위치 데이터와 같은 민감한 정보를 통해 개인의 일상적인 패턴, 거주 지역, 심지어 건강 상태까지 예측할 수 있으며, 이로 인해 프라이버시 침해의 우려가 있다 [4,5]. 이러한 프라이버시 우려 때문에 많은 사용자들이 자신의 민감한 정보를 기업에 제공하는 것에 주저하고 있으며, 이로 인해 기업들은 사용자 데이터 수집 및 분석에 많은 어려움을 겪고 있다.

의료 분야는 사용자 데이터 수집 및 분석에서 특별히 어려움을 겪고 있는 주요 응용 분야이다. 환자의 병명과 같은 의료 정보는 건강 상태와 직접 연결되어 있으며, 이는 질병의 발생과 추이를 분석하는 데 중요한 역할을 한다. 의료 데이터는 극도로 민감한 정보에 해당하므로, 이를 취급할 때는 환자의 개인정보 보호가 매우 중요하다 [6]. 의료 데이터가 부적절하게 공개되면, 환자는 사생활 침해뿐 아니라 사회적, 경제적 피해를 입을 위험이 크다. 따라서 의료 데이터의 수집과 관리에는 철저한 개인정보 보호 조치를 적용해야 한다.

최근 사용자의 개인정보를 보호하면서 의료 데이터, 특히 병명 정보를 수집하기 위한 많은 연구가 제안되었다. 이중 차분 프라이버시(Differential Privacy)[7]를 활용하여 개인의 병명 데이터를 수집하는 방법이 크게 주목받고 있습니다. 예를 들어, [8]은 차분 프라이버시의 한 종류인 Geo-Indistinguishability (Geo-I) [9,10]를 적용해 개별 사용자의 병명 데이터를 안전하게 수집하는 기법을 제안하였다. 그러나 이러한 방법들은 사용자의 마이크로 데이터(microdata)를 보호하면서 수집하는 데는 효과적이지만, 수집된 데이터를 이용해 유용한 집계 분석(aggregate analysis)을 수행하는 데는 한계가 있다. 반면 데이터 분석가들은 단순한 데이터 수집을 넘어서 수집된 데이터로부터 정확하고 유의미한 정보를 추출하는 데 더 큰 관심을 보이고 있다. 이에 따라, 단순한 의료 데이터 수집 방법을 발전시켜 보다 심도 있는 집계 분석이 가능한 연구가 필요하다.

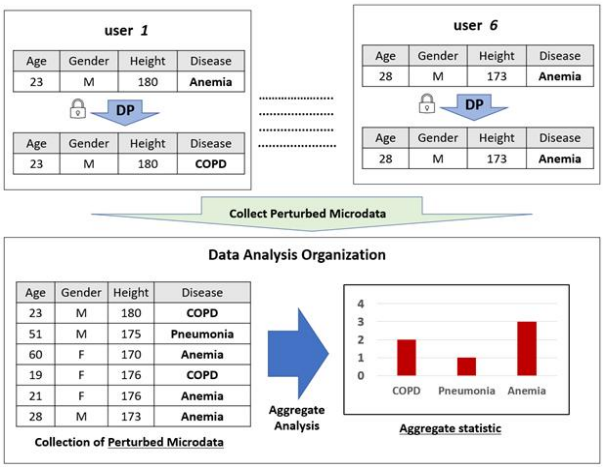


Fig. 1. Privacy-preserving collection of medical microdata and aggregate analysis

데이터 분석에서는 경우에 따라 개별 마이크로 데이터를 활용하기도 하지만, 대부분은 집계 정보와 같은 통계적 데이터를 필요로 한다. 가령, 데이터 분석가가 프라이버시를 보호하며 수집한 사용자의 병명 데이터를 바탕으로 각 병명의 빈도수를 계산하려 한다고 가정하자(그림 1). 이 경우 기존 방법들은 변조된 마이크로 데이터를 이용하여 직접 집계 정보를 추출하는데, 이는 정확도가 떨어질 수 있다. 이에 따라, 본 연구에서는 의료 데이터 수집 및 분석을 보다 효과적으로 수행할 수 있는 새로운 기법을 제안한다. 특히, 기존 연구[8]는 단순한 프라이버시 보존 데이터 수집에 중점을 두었지만, 본 연구에서는 프라이버시를 유지하면서도 개별 사용자로부터 병명 데이터를 수집하고, 이를 바탕으로 각 병명의 발병 빈도수를 효과적으로 계산할 수 있는 방법을 제안한다

본 연구의 기여는 다음과 같다

- 프라이버시를 유지하며 민감한 사용자 병명 데이터를 안전하게 수집하기 위한 프레임워크를 개발한다.
- 수집된 데이터로부터 효과적으로 통계적 분석 결과(예, 질병 빈도수)를 도출하기 위한 방법을 제안한다. 특히, 제안 방법은 Expectation-Maximization

(EM) 기법을 활용하여 프라이버시를 보존하면서 수집된 변조된 병명 데이터로부터 효과적으로 통계적 정보를 추출한다.
- 마지막으로, 실제 데이터를 이용해 제안 기법의 성능을 평가한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문과 연관된 선행 연구에 대해서 설명하고, 3장에서는 배경 지식에 대하여 설명한다. 4장에서는 제안 기법에 대하여 설명한다. 5장에서는 제안 기법의 성능을 평가를 수행한 후, 6장에서 결론을 맺는다.

II. Related Work

최근 개인정보를 보호하면서 사용자의 마이크로 데이터를 수집하기 위한 연구가 활발히 진행되고 있다. [11]은 Geo-I를 활용하여 사용자의 민감한 텍스트 마이크로데이터를 수집하는 프라이버시 보호 기법을 제안하였다. 이 기법은 먼저 텍스트 데이터의 각 단어를 워드 임베딩을 통해 벡터 x로 변환한다. 그 후, ε-Geo-I를 만족시키기 위해 x에 노이즈 n을 추가한다 (즉, v=x+n). 마지막으로 단어는 변조된 벡터 v와 가장 가까운 임베딩을 가진 다른 단어로 대체되고, 이 대체된 단어를 데이터 수집 서버에 제공한다. 데이터 수집 서버는 이렇게 변조된 단어만을 수집하므로, 사용자의 원본 데이터가 외부로 노출될 위험이 없다. [12]는 프라이버시 보존 자연어 처리 기법을 제안하였다. 제안 기법은 지역 차분 프라이버시를 활용하여 사용자의 텍스트 데이터를 변조하는 방식으로 수집한다. 변조된 데이터를 수집한 후, 이를 활용하여 인공지능 모델을 학습시키고, 학습된 모델로 데이터 분석을 진행한다. 제안 기법은 변조된 데이터만을 수집하기 때문에 사용자의 프라이버시를 보호할 수 있다. [8]은 Geo-I를 이용하여 개별 사용자의 병명 데이터를 안전하게 수집하는 기법을 제안하였다. 제안된 기법은 병명과 같은 마이크로데이터를 효과적으로 수집하기 위해 ε-Geo-I를 충족하는 데이터 변조 방법을 도입하였다. 본 연구는 [8]의 기법을 발전시켜 단순히 병명 마이크로데이터를 수집하는 것뿐만 아니라, 수집된 데이터를 활용하여 집계 분석을 보다 효율적으로 수행할 수 있는 방법을 제안한다.

최근 들어 차분 프라이버시 기법을 활용하여 사용자의 민감한 데이터를 안전하게 수집하기 위한 다양한 연구가 진행되었다 [13,14,15]. 특히, [16]은 스마트워치 사용자의 심박수와 누적 걸음수 같은 건강 데이터를 프라이버시를 보존하면서 차분 프라이버시를 사용하여 수집하는 방법을 제안하였다. [17]은 포그-클라우드 환경에서 사용자의 민감한 데이터를 보호하며 효과적으로 수집하기 위한 기법을 개발했습니다. 추가로, 데이터 보안성을 강화한 분산 차분 프라이버시(Distributed Differential Privacy) 기법도 프라이버시 보호 데이터 수집에 활용되고 있다. [18]은 분산 차분 프라이버시를 사용하여 민감한 데이터를 안전하게 수집하는 방법을 제안하다. 또한, Arachchige et al. [19]은 딥러닝에서 사용되는 데이터의 개인정보 유출을 방지하기 위해 LATENT라는 새로운 차분 프라이버시 기반 알고리즘을 도입하였다.

III. Background and Problem Statement

1. Geo-Indistinguishability

Geo-I는 차분 프라이버시의 일종으로, 사용자의 위치 데이터에 노이즈를 추가해 변조함으로써 공격자가 사용자의 정확한 위치를 유추하는 것을 방지하는 프라이버시 보호 기법이다.

정의 1. (ε-Geo-Indistinguishability) X를 사용자의 실제 위치 데이터 집합, Z를 사용자가 서버에 전송한 변조된 위치 데이터 집합이라고 각각 가정한다. M을 임의의 매커니즘이라고 가정하자. 이때, X의 임의의 데이터 x_1과 x_2에 대하여 (즉, $x_1, x_2 \in X$), M으로부터 생성되는 모든 결과값 $z \in Z$에 대하여 다음 식을 만족하면 M은 ε-Geo-I를 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 거리에 해당한다. 또한, $M(x_1)(z)$은 매커니즘 M이 입력 x_1으로부터 z를 무작위로 생성하는 것을 의미한다. Geo-I는 차분 프라이버시에 거리 개념을 도입한 기법이다. 즉, $d(x_1,x_2)=1$이면, 식(1)은 차분 프라이버시 정의와 동일해 진다.

2. Problem Statement

$U=\{u_1, u_2, ..., u_n\}$을 사용자들의 집합이라고 가정한다. 이때 n은 전체 사용자를 나타내고, $u_i \in U$는 i번째 사용자를 나타낸다. 본 연구에서는 사용자들이 데이터 수집자에게 데이터를 제공할 때, 프라이버시 보호를 위해 ε-Geo-I 기법을 사용하여 변조된 병명 마이크로데이터를

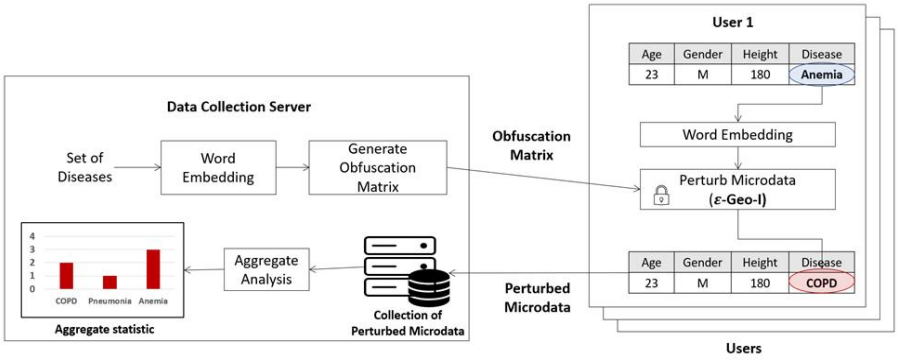


Fig. 2. Overview of the proposed framework for collecting and analyzing medical microdata

사용한다고 가정하자. $S=\{s_1, s_2, ..., s_m\}$은 m개의 서로 다른 병명 데이터 집합을 나타낸다. i번째 사용자 $u_i \in U$의 병명 데이터는 $su_i \in S$로 표현하며, 이는 데이터 수집가가 사용자로부터 수집하는 데이터이다.

데이터 수집가가 n명의 사용자로부터 ε-Geo-I를 이용하여 수집한 변조된 병명 데이터 집합을 $DB=\{su'_1, su'_2, ..., su'_n\}$라고 가정하자. 이때, $su'_i \in S$는 i번째 사용자의 변조된 병명 데이터에 해당된다 (본 연구에서는 $su_i \in S$를 i번째 사용자 실제 병명 데이터, $su'_i \in S$를 동일 사용자의 변조된 병명 데이터를 각각 표현하기 위해 사용한다). 본 연구에서는 ε-Geo-I를 만족하면서 수집한 병명 데이터로부터 각 질병의 빈도수 집합 $F=\{f_1, f_2, ..., f_m\}$을 효과적으로 구하는 것을 목적으로 한다. 이때, $f_i \in F$는 병명 s_i의 빈도수를 나타낸다.

IV. Proposed Framework

4장에서는 본 논문의 제안 기법을 설명한다. 그림 2는 제안 기법의 구성도에 해당한다.

- 프라이버시 보존 데이터 수집: 데이터 수집 서버는 ε-Geo-I를 만족하는 변조 행렬(obfuscation matrix)를 생성 후, 이를 각각의 사용자에게 배포한다. 각각의 사용자의 자신의 병명 데이터를 변조 행렬을 이용하여 변조한 후, 변조된 데이터를 데이터 수집 서버에 전송한다.
- 프라이버시 보존 집계 분석: 데이터 수집 서버는 ε-Geo-I를 이용하여 수집한 변조된 병명 데이터를 데이터베이스에 저장한다. 또한 데이터베이스에 저장된 데이터를 이용하여 집계 분석을 수행하여 유용한 통계적 정보(예, 각 질병의 빈도수)를 추출한다.

1. Privacy-preserving Medical Data Collection

본 연구의 병명 데이터 수집 방법은 선행 연구[8]에 기반을 두고 있으며, 본 절에서는 그 방법을 간략하게 설명한다.

워드 임베딩: Geo-I는 벡터 공간상에서 표현된 사용자 위치 정보를 보호하도록 설계된 방법이므로, 텍스트 형태의 병명 데이터에는 직접 적용이 불가능하다. 이러한 문제를 해결하기 위해, 각 병명 데이터를 먼저 고차원 벡터로 변환할 필요가 있다. 워드 임베딩(word embedding) 기술인 Word2Vec [20] 및 BERT [21]는 텍스트 기반 데이터를 고차원 공간에서 연속 벡터로 표현하는 데 사용된다. 본 연구에서는 Word2Vec을 활용하여 병명 데이터를 고차원 벡터로 변환한다. 즉, 병명 데이터 $s_i \in S$가 주어졌을 때, v_i는 Word2Vec을 사용하여 얻은 해당 t-차원 병명 벡터 데이터에 해당한다.

변조 행렬: 병명 데이터 집합 $S=\{s_1, s_2, ..., s_m\}$이 주어졌을 때, 이 데이터를 앞에서 설명한 워드 임베딩 기법을 사용해 벡터로 변환한 집합은 $V=\{v_1, v_2, ..., v_m\}$으로 표현한다. 이때, $m \times m$ 크기의 변조 행렬을 O로 나타내며, $O[i,j]$는 변조된 병명 데이터 s_j가 실제 병명 데이터 s_i로부터 무작위로 생성될 확률을 나타낸다 (즉 $O[i,j]=M(s_i)(s_j)$). 특히, 선행연구[8]에서 제안된 ε-Geo-I 기준을 만족하는 변조 행렬에서는 $O[i,j]$는 다음과 같이 정의된다.

$$O[i,j]=\frac{e^{-\frac{\epsilon}{2}\times d(v_i,v_j)}}{\sum_{s_k \in S} e^{-\frac{\epsilon}{2}\times d(v_i,v_k)}} \quad \text{식(2)}$$

여기서 $d(v_i,v_j)$는 두 벡터 v_i와 v_j 사이의 거리를 나타낸다. 변조 행렬 O를 정의한 후, 데이터 수집 서버는 이를 각각의 사용자에게 배포한다.

데이터 변조: 데이터 수집 서버로부터 변조 행렬 O를 받은 후, 각 사용자는 행렬에 포함된 확률에 따라 실제 병명 데이터를 변조한다. 사용자 u_i의 실제 병명 데이터가 $s_k \in S$라고 가정하면(즉, $s_k = su_i$), 변조 행렬 O의 k번째 행(즉, $O[k,1], O[k,2], ..., O[k,m]$)에 인코딩된 확률에 따라 사용자가 변조된 병명 데이터를 생성하고, 이를 데이터 수집 서버로 전송한다. 이러한 방식으로 데이터 변조는 사용자의 단말기에서 이루어지므로, 사용자의 민감한 병명 정보가 외부에 노출되는 것을 방지할 수 있다.

2. Effective Aggregate Analysis with Privacy-preserved Medical Microdata

본 절에서는 프라이버시를 보존하면서 수집된 변조된 병명 데이터로부터 유용한 통계적 정보(즉, 각 질병의 빈도수)을 얻기 위한 분석 방법을 제안한다.

4장 1절에서 설명한 ε-Geo-I를 이용하여 수집한 변조된 병명 데이터 집합을 $DB = \{su'_1, su'_2, ..., su'_n\}$라고 가정하자. 병명 s_i의 빈도수 f_i는 선행 연구[8]의 기법을 이용하여 다음과 같이 구할 수 있다.

$$f_i = cnt(s_i, DB) \qquad \text{식(3)}$$

이때, 함수 $cnt(s_i, DB)$는 병명 s_i가 DB에 나타나는 빈도수를 의미한다. 그러나 이 방식은 ε-Geo-I의 데이터 변조 방식을 고려하지 않고 있기 때문에, 정확한 빈도수를 계산하기에는 어려움이 있다. 그러므로 본 연구에서는 ε-Geo-I의 데이터 변조 확률을 고려한 두 가지 기법(확률화 방식, EM 기반 방식)을 제안한다.

2.1 Probabilistic Approach

이전 방식보다 더 효과적인 접근은 변조 행렬 O에 인코딩된 실제 데이터와 변조 데이터 간의 확률적 매핑 정보를 활용하는 것이다. 변조 행렬의 정의에 따르면, 모든 $s_i \in S$에 대해 $O[i,j]$는 실제 데이터 s_i로부터 무작위로 변조된 데이터 s_j가 생성될 확률을 나타낸다. 그러므로 변조된 데이터와 실제 데이터 간의 매핑 확률 정보를 이용하면, 빈도수 f_i를 다음과 같이 추정할 수 있다.

$$f_i = \sum_{s_j \in S} (O[i,j] \times cnt(s_j, DB)) \qquad \text{식(4)}$$

즉, 확률화 기법은 빈도수 f_i를 계산할 때, 변조 행렬에 인코딩된 확률을 고려한다. 이 기법은 변조된 데이터 s_j가 실제 데이터 $s_1, s_2, ..., s_m$에서 각각 변조 행렬에 인코딩된 특정 확률로 생성된다는 특성을 활용해 빈도수를 계산한다.

2.2 Expectation-Maximization-based Approach

Expectation-Maximization (EM) 기법은 잠재 변수를 포함하는 복잡한 확률 모델의 파라미터 추정에 사용되는 방법이다. 본 논문에서는 EM 기법을 활용하여, 초기에 모든 질병의 빈도수를 동일하게 가정하고, 이후 이를 점진적으로 최적화하는 방식으로 질병의 빈도수를 예측한다. EM 알고리즘은 초기화, Expectation 단계, Maximization 단계로 구성된다.

- 초기화: 초기 질병 빈도수를 다음과 같이 동일하게 설정한다.

$$f_k^{(0)} = \frac{|DB|}{m}, \ 1 \le k \le m \qquad \text{식(5)}$$

- Expectation 단계: 현재의 빈도수 $f_1^{(t)}, f_2^{(t)}, ..., f_m^{(t)}$를 사용하여 사후 확률 $P(s_k|su'_i)$를 구한다. 이때 $P(s_k|su'_i)$는 i번째 사용자로부터 수집한 변조된 병명 데이터 su'_i가 실제 병명 데이터 s_k로부터 생성되었을 확률을 나타내며, 다음과 같이 베이지 정리를 이용하여 구할 수 있다.

$$\begin{aligned} P(s_k|su'_i) &= \frac{P(s_k) \times P(su'_i|s_k)}{\sum_{j=1}^{m} P(s_j) \times P(su'_i|s_j)} \\ &= \frac{\frac{f_k^{(t)}}{|DB|} \times P(su'_i|s_k)}{\sum_{j=1}^{m} \frac{f_j^{(t)}}{|DB|} \times P(su'_i|s_j)} \\ &= \frac{f_k^{(t)} \times P(su'_i|s_k)}{\sum_{j=1}^{m} f_j^{(t)} \times P(su'_i|s_j)} \end{aligned} \qquad \text{식(6)}$$

여기서, $P(su'_i|s_k)$는 변조 행렬를 이용하여 구할 수 있다. 만약 su'_i가 s_g이면(즉, $su'_i = s_g$), $P(su'_i|s_k)$는 $O[k,g]$에 해당한다.

- Maximization 단계: 매 반복마다 $f_k^{(t)}$는 다음과 같이 업데이트 된다.

$$f_k^{(t+1)} = |DB| \times \sum_{i=1}^{n} \left(P(s_k \mid su'_i) \right) \quad \text{식(7)}$$

위에서 설명한 Expectation 단계와 Maximization 단계는 사전에 정한 횟수만큼 반복하거나, 파라미터 간의 차이가 수렴할 때까지 계속 반복한다. 위의 EM 알고리즘 종료 후, 빈도수는 $f_k = f_k^{(t+1)}$에 의해 구한다.

EM 기법은 Expectation 단계와 Maximization 단계를 반복적으로 적용하기 때문에 대규모 데이터를 처리할 때 효율성 저하 문제가 발생할 수 있다. 이를 해결하기 위해 EM 기법을 병렬 처리하여 효율성을 높일 수 있으며, 이와 관련한 다양한 연구가 이루어졌다 [22].

V. Experiments and Results

1. Experiments

본 연구에서는 성능 평가를 위해 먼저 위키피디아의 질병 목록에서 61개의 병명 데이터를 수집하였다. 그런다음 61,000명의 환자 데이터를 생성하고, 각 환자에게 61개의 병명 중 하나를 할당하였다. 따라서 실험에 사용된 전체 환자 수(n)는 61,000명, 병명 수(m)는 61개이다.

성능 비교를 위해, IV.2절에서 설명한 세 가지 접근법인 선행 연구[8]의 단순 기법(NA), 확률화 기법(PA), EM 기반 기법(EM)을 사용하여 결과를 비교하였다. 실험에서는 평균 절대 오차(Mean Absolute Error, MAE)를 사용하여 성능을 측정했으며, MAE는 다음과 같이 정의된다.

$$MAE = \frac{1}{m} \times \sum_{s_i \in S} \left| f_i^{true} - f_i^{estimation} \right| \quad \text{식(8)}$$

이때, f_i^{true}는 병명 s_i의 실제 빈도수를 나타내며, $f_i^{estimation}$는 프라이버시를 보존하면서 수집한 변조된 병명 데이터로부터 추정한 병명 s_i의 빈도수를 나타낸다.

2. Experimental Results

그림 3의 실험은 프라이버시 예산 ε 값이 예측 결과에 미치는 영향을 보여준다. 실험에서 ε 값으로 0.3, 0.5, 0.7, 1.0, 1.5, 2.0을 사용하였다. 그림에서 알 수 있듯이 모든 ε 값에 대하여 제안 기법인 확률화 방식과 EM 기반 방식이 단순 기법보다 우수한 성능을 보여주고 있다. 특히, EM 기반 방식은 단순 기법 대비 매우 우수한 성능을 보이고 있으며, ε 값이 증가함에 따라 두 기법간의 성능 차이가 커지는 것을 알 수 있다.

또한, 그림 3의 실험에서 ε 값이 감소할수록 MAE 값이 증가하는 것을 알 수 있다. 이는 ε 값이 낮아짐에 따라 사용자의 프라이버시 보호 수준이 향상되고, 원본 데이터에 대한 변조가 심하기 발생하기 때문이다. 반대로 ε 값이 높아질수록 예측 값이 실제 값과 더 유사해지는 것을 볼 수 있습니다. 이는 ε 값이 증가할수록 사용자의 프라이버시 보호 수준이 감소하고, 원본 데이터에 대한 변조가 줄어들기 때문입니다. 이는 차분 프라이버시를 적용할 때 일반적으로 나타나는 현상이며, 본 연구에서 제안한 기법 또한 이와 같은 결과를 나타내고 있다.

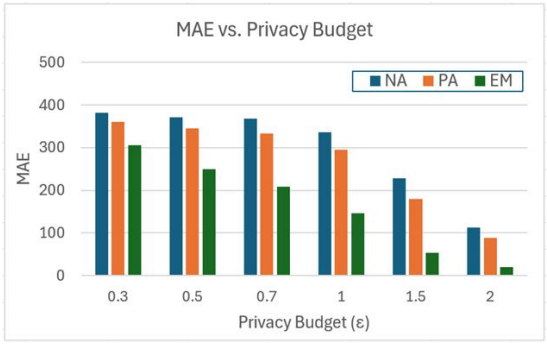


Fig. 3. MAE for varying privacy budget (ε)

그림 4의 실험은 EM 기반 기법의 반복 횟수에 따른 MAE 결과 값을 보여준다. 실험에서는 ε값으로 0.5, 1.0, 2.0을 사용하였으며, EM 기반 기법의 Expectation 단계와 Maximization 단계는 최대 200번까지 반복하였다. 실험 결과에서 볼 수 있듯이, 초기 반복 단계에서 MAE 값이 급격히 감소하는 것을 확인할 수 있다. 특히, ε 값이 높을수록 초기 반복에서 빠르게 최적값으로 수렴하는 경향을 보이고 있다. 대부분의 경우, 처음 50번의 반복 후에는 MAE 값이 안정적으로 수렴하는 것을 확인할 수 있다. 이 실험 결과를 통해 본 연구에서 제안한 EM 기반 방식이 초기 반복을 통해 효과적으로 최적값으로 수렴한다는 것을 확인할 수 있다.

그림 5의 실험에서는 각 병명별 빈도수 그래프를 나타낸다. 그림에서 x축은 실험에 사용한 61개 병명의 색인에 해당하며, y축은 각 병명의 빈도수를 나타낸다. 비교 목적으로 그림 5에는 프라이버시 보존 기법을 사용하지 않고 수집한 원본 병명 데이터를 이용한 실험 결과도 포함되어 있다(그림에서 'Ground'로 표시됨). 실험에서는 ε값으로 1.0, 1.5, 2.0을 사용하였으며, 단순 기법과 EM 기반 기법의 성능을 비교하였다. 모든 ε 값에서 EM 기반 기법이 단순 기법보다 실제 빈도수에 더 근접한 결과를 보여주고 있다. 특히, ε값이 1.5 이상인 경우, EM 기반 기법은 실제 빈

도수와 매우 유사한 결과를 나타낸다. 이를 통해 본 연구의 제안 기법이 프라이버시를 보존하면서도 병명 데이터의 효과적인 집계 분석을 가능하게 함을 확인할 수 있다.

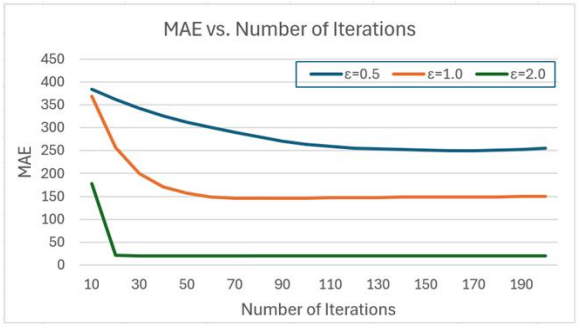
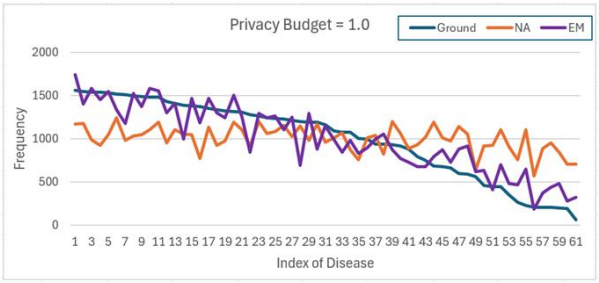
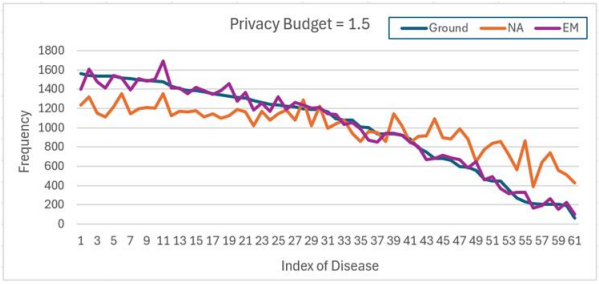


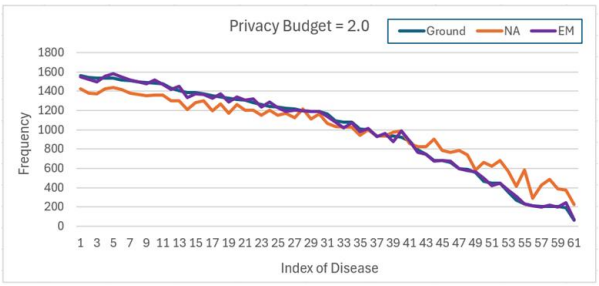
Fig. 4. MAE for varying the number of iteration for EM based approach



(a) Privacy budget (ε) = 1.0



(b) Privacy budget (ε) = 1.5



(c) Privacy budget (ε) = 2.0

Fig. 5. Comparison between EM-based approach and navie approach for the computation of each disease frequency

VI. Conclusions and Future Work

최근 의료 분야에서 민감한 환자 정보의 수집과 활용이 중요한 도전 과제로 부상하고 있다. 이를 해결하기 위해 본 연구는 프라이버시를 보호하면서도 통계적으로 가치 있는 정보를 효과적으로 추출할 수 있는 새로운 데이터 수집 및 분석 방법을 제안하였다. 제안 기법은 차분 프라이버시의 한 형태인 Geo-I를 활용하여 사용자의 병명 마이크로데이터를 안전하게 수집하고, 이 수집된 변조 데이터를 바탕으로 효율적인 집계 분석을 수행하는 방법을 제안하였다. 또한, 실험을 통해 본 연구에서 제안한 EM 기반 기법이 집계 분석에 있어서 기존 방식보다 우수한 성능을 나타내는 것을 확인할 수 있다.

향후 연구에서는 실시간 데이터 수집 및 분석 환경에서 제안 기법의 적용 가능성을 검토하고, 효율적인 알고리즘 최적화를 수행해야 한다. 또한, 본 연구의 제안 기법을 금융 데이터와 위치 데이터와 같은 다른 민감한 정보 유형에도 적용하여 그 유효성을 평가하는 연구가 필요하다.

REFERENCES

[1] S.R.E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur, A survey of vision-based traffic monitoring of road intersections, IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 10, 2016. DOI: 10.1109/TITS.2016.2530146

[2] 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

[3] M. Soori, B. Arezoo, and R. Dastres, Internet of things for smart factories in industry 4.0, a review, Internet of Things and Cyber-Physical Systems, vol 3, 2023. DOI: 10.1016/j.iotcps.2023.04.006

[4] 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

[5] J.W. Kim, K. Edemacu, and B. Jang. Privacy-preserving mechanisms for location privacy in mobile crowdsensing: A survey. Journal of Network and Computer Applications, vol. 200, 2022. DOI: 10.1016/j.jnca.2021.103315

[6] S. Dash, S.K. Shakyawar, M. Sharma, and S. Kaushik. Big data in healthcare: Management, analysis and future prospects. Journal of Big Data, vol. 6, 2019. DOI: 10.1186/s40537-019-0217-0

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

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

[9] 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

[10] 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

[11] 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 onWeb Search and Data Mining, Houston, TX, USA, February 2020. DOI: 10.1145/3336191.3371856

[12] 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

[13] R. Chen, L. Li, Y. Ma, Y. Gong, Y. Guo, T. Ohtsuki, and Mi. Pan. Constructing Mobile Crowdsourced COVID-19 Vulnerability Map With Geo-Indistinguishability. IEEE Internet of Things Journal, vol. 9, no. 18, pp. 17403-17416, 2022. DOI: 10.1109/JIOT.2022.3158895

[14] M. Yang, T. Guo, T. Zhu, I. Tjuawinata J. Zhao, and K-Y. Lam. Local differential privacy and its applications: A comprehensive survey. Computer Standards & Interfaces, vol. 89, 2024. DOI: 10.1016/j.csi.2023.103827

[15] 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, 2435 no. 3, 2020. DOI: 10.1109/TCE.2019.2924466

[16] Su-Mee Moon, Jong-Wook Kim. Privacy-Preserving Method to Collect Health Data from Smartband. Journal of the Korea Society of Computer and Information 25(4), 113-121. 2020. DOI: 10.9708/jksci.2020.25.04.113

[17] Jong-Hyun Lim, Jong Wook Kim. Privacy-Preserving IoT Data Collection in Fog-Cloud Computing Environment. Journal of the Korea Society of Computer and Information 24(9), 43-49, 2019. DOI: 10.9708/jksci.2019.24.09.043

[18] Jong-Hyun Lim, Jong Wook Kim. Privacy-Preserving Aggregation of IoT Data with Distributed Differential Privacy. Journal of the Korea Society of Computer and Information 25(6), 65-72, 2020. DOI: 10.9708/jksci.2020.25.06.065

[19] P.C.M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, M. Local differential privacy for deep learning. IEEE Internet of Things Journal, 7(7), 2019. DOI:

[20] B. Jang, I. Kim, and J.W. Kim. Word2vec convolutional neural networks for classification of news articles and tweets. Plos One, vol. 14, no. 8, 2019. DOI: 10.1371/journal.pone.0220976

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805. DOI: 10.48550/arXiv.1810.04805

[22] J. Prakash, U. Agarwal, and P. K. Yalavarthy. Multi GPU parallelization of maximum likelihood expectation maximization method for digital rock tomography data. Scientific Reports, vol. 11, 2021. DOI: 10.1038/s41598-021-97833-z

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.