Journal of Next-generation Convergence Information Services Technology Vol.10, No.4, August (2021), pp.347-360
ISSN 2384-101X(p) 2672-1163(e) *http://dx.doi.org/10.29056/jncist.2021.08.01*

도로 네트워크 상에서 대용량 궤적 데이터 정보 보호를 위한 l-warping 기반 궤적 k-익명화 기법

l-warping-based trajectory k-anonymization method for protecting large trajectory information on road network

김용기[1]

Yong-Ki Kim[1]

요 약

이동 객체 궤적 데이터에 대한 데이터 마이닝 결과의 신뢰성을 향상시키기 위해 궤적 데이터 사이의 유사도 측정 시 발생하는 오차를 감소시키는 연구가 필수적이다. 한편, 이동 객체 궤적 데이터는 사용자의 위치 등 민감한 개인 정보를 포함하기 때문에, 이를 보호하기 위한 기법이 요구된다. 이를 위해, K개의 유사한 궤적 그룹화 기반의 궤적 익명화 기법을 이용하여 대표 궤적을 생성하는 연구가 수행되었다. 본 논문에서는 데이터 마이닝 응용의 신뢰성 향상을 위해, 길이 기반 워핑 기법을 사용하여 효율적인 궤적 k-익명화 기법을 제안한다. 제안하는 기법은 첫째, l-warping 기법을 통해 노드간 중복을 허용함으로써, 궤적 간 거리 측정 시 일부 구간이 삭제되는 문제점을 해결한다. 둘째, 궤적 구간별 출현 빈도를 고려한 궤적 재구성을 통해 원본과 재구성된 데이터 사이의 유사도 및 데이터 마이닝 적합도를 향상시킨다. 마지막으로 성능평가를 통해 제안하는 기법이 궤적 정보 보호의 정도, 수행 시간 효율성, 재구성된 데이터의 정확성 측면에서 기존의 기법보다 우수함을 나타낸다.

핵심어 : 궤적 익명화, 데이터보호, 이동 객체, 도로 네트워크, 빅 데이터

Abstract

To improve the reliability of the trajectory data mining result, it is necessary to reduce errors while calculating similarity measures. On the other hand, protecting users' privacy is the important issue. For this, trajectory anonymization schemes were proposed to make a representative trajectory by grouping K similar trajectories. We, in this paper, propose an efficient trajectory anonymization scheme using length-based warping (l-warping) technique for data mining applications. To reducing errors of calculating distances and similarities among trajectories, we, first, propose a representative trajectory generation algorithm based on our l-warping distance. Our scheme can reduce the number of eliminated parts when calculating the distances between two trajectories. Secondly, we propose a trajectory reconstruction algorithm which is suitable for applications which need the partial path information of trajectories such as mobile commerce advertisement system. Finally, we show from performance analysis that our scheme outperforms the existing schemes in terms of the anonymization efficiency and the accuracy of reconstruction data.

Keyword : Trajectory Anonymization, Privacy Protection, Moving Object, Road Networks, Big Data

1 Department of IT Convergence System, Vision College of Jeonju, jeonju, Korea [Professor]
e-mail: kimyk@jvision.ac.kr

Received(July 11, 2021), Review Result(1st: July 28, 2021), Accepted(August 13, 2021), Published(August 31, 2021)

1. 서론

최근 무선 통신 및 모바일 위치 측정 기술의 발달과 이동 단말기의 보편화로 인하여, 위치 기반 서비스(Location Based Services)의 이용이 확산되었다. LBS 서비스 제공자는 유, 무선 통신망을 통해 얻은 위치정보를 다른 유용한 정보와 실시간으로 결합하여, 친구 찾기, 인접한 POI(Point of Interest) 찾기 등, 사용자가 필요로 하는 부가적인 응용 서비스를 제공한다. 아울러 사용자의 위치 정보를 통해 얻은 궤적을 바탕으로 데이터 마이닝을 수행하여, 정체 구간 분석, 시간 별 도로 이용 패턴 분석 등의 교통 정보 서비스를 지원한다 [1][2]. 그러나 사용자가 위치기반 서비스를 제공받기 위해서 정확한 위치 정보를 데이터베이스 서버에 전송하는 것은 심각한 개인 정보 누출 문제를 발생시킬 수 있다. 예를 들어 악의적인 상대방(adversary)이 사용자의 위치 정보를 바탕으로 자주 방문하는 장소 혹은 자주 이용하는 도로 구간 등의 정보를 분석하여, 사용자의 생활패턴, 종교 등의 사생활 정보를 획득할 수 있다 [3-10]. 따라서 모바일 사용자의 안전하고 편리한 위치기반 서비스 사용을 위한 개인 정보 보호 방법이 요구된다.

위치기반 응용 서비스 지원 시 개인 정보 보호를 위한 연구 중, 대표적인 연구로는 K-anonymity를 적용한 연구가 존재한다 [11-13]. K-anonymity는 특정 사용자의 정보가 다른 K-1명의 사용자 정보와 동일하여 해당 정보만으로는 사용자를 식별하거나 사용자의 개인 정보를 유추할 수 없는 특성을 의미한다. 이를 위해 기존 연구에서는 K개 이하의 노드를 지나는 궤적 데이터를 삭제하고 궤적 집합을 생성하거나 [11], 클러스터링을 통해 유사 궤적 집합을 생성하거나 [12], 혹은 가까운 두 궤적을 나타내기 위한 대표 궤적을 생성하여 K를 만족할 때까지 점진적으로 궤적 집합을 증가시켜 가며 궤적 집합을 생성하는 기법을 제안하였다 [13]. 그러나 기존 기법들은 다음과 같은 단점을 지니고 있다. 첫째, 이동에 제약이 없는 유클리디언(Euclidean) 공간을 가정하고 있기 때문에, 도로 등의 네트워크상에서 이동하는 사용자가 지니는 제약을 고려하지 못한다. 둘째, 궤적이 지닌 위치 정보를 삭제하기 때문에, 궤적 데이터의 왜곡이 발생한다. 따라서 본 논문에서는 도로 네트워크 환경에서 궤적 데이터의 왜곡 정도를 감소시킬 수 있는 궤적 정보 보호 기법을 제안한다.

제안하는 기법은 첫째, 궤적의 시공간 거리에 기반한 클러스터링을 통하여 유사 궤적 집합을 구성한다. 궤적의 시공간 거리는 네트워크 거리에 기반한 시공간 유사도 측정 기법으로, 도로 네트워크 환경에 적용이 용이하다. 둘째, 길이 기반 warping 기법을 이용하여 궤적 간 유사도 계산 시 궤적에 포함된 위치 정보의 삭제를 방지한다. 이를 통해 궤적 데이터의 왜곡 정도를 감소한다.

본 논문의 구성은 다음과 같다. 2장에서는 K-anonymity 기반의 궤적 정보 보호를 지원하는 관련 연구를 소개한다. 3장에서는 제안하는 궤적 정보 보호 기법에 대해 기술하고, 4장에서는 실험을 통한 성능 분석 결과를 제시한다. 마지막으로 5장에서는 결론 및 향후 연구를 제시한다.

2. 관련연구

궤적 정보 보호를 위해 다양한 k-anonymity 기반 궤적 보호 기법들이 연구되었다[11-13]. 이 중 a generalization based trajectory anonymization method (General-Based Approach: GBA) [13] 기법은 궤적 데이터베이스의 K-anonymity를 만족시키기 위해, 점진적인 일반화를 통한 궤적 그룹핑 (Grouping) 및 대표 궤적 생성을 수행한다. 더불어, 데이터 마이닝(data-mining)을 지원하기 위한 궤적 재구성을 수행한다. GBA에서 유사한 k개의 궤적을 찾기 위해 궤적간 거리 계산을 수행하는데, 이때, 궤적간의 거리는 가장 가까운 두 포인트 간 거리의 합으로 정의된다. 두 궤적 tr1과 tr2의 가장 가까운 포인트의 쌍을 포인트 링크(Point Link: PL) 라 하고, 포인트 링크의 집합을 포인트 매칭(Point Matching: PM)이라 한다. 포인트 링크 및 포인트 매칭의 정의는 정의 1과 같다.

정의 1. 포인트 링크(point-Link) 및 포인트 매칭(Point Matching)

k개의 유사 궤적으로 구성된 유사 궤적 집합 TR={tr_1, tr_2, ..., tr_K}에 대해,

- i번째 Point Link(PL_i): PL_i={$p1_i$, $p2_i$, ..., pK_i}, $p_{ji} \in tr_j$
- Point Matching (PM): PM= {PL_1, PL_2, ..., PL_m}, if i<j, then $PL_{i\cdot t} < PL_{j\cdot t}$

단, m은 TR에 속한 궤적 데이터 중 가장 많은 노드를 지나는 궤적의 전체 노드 수를 의미한다 (m=max(|tr_1|, |tr_2|, ..., |tr_K|)).

한편, GBA는 크게 유사 궤적 거리 계산 및 대표 궤적 생성의 두 단계로 이루어진다. 유사 궤적 거리 계산을 위해 알고리즘에서는 두 개의 궤적에 대해 PL 및 PM을 이용한 거리 계산을 수행한다. 이때 포인트 매칭의 크기가 두 궤적 사이의 거리를 나타내며, 이는 두 궤적 중 길이가 짧은 궤적의 길이와 같다. 아래는 거리 계산을 위한 식을 나타낸다.

$$\sum_{i=0}^{m} dist(p_1, p_2) + |PM^c| \times MAXdist, (|PM| = m, p_{ij} \in PL_i)$$

궤적 정보 보호 과정은 다음과 같다. 첫째, 기준 궤적과 다른 모든 궤적 간의 거리를 계산하여 가장 가까운 1개의 궤적을 선정한다. 이때, 처음 기준 궤적은 임의로 선정된다. 둘째, 기준 궤적과 선정된 궤적의 가장 가까운 포인트 쌍을 포인트 링크로 구성하고, 포인트 매칭에 포함되지 않은 포인트는 삭제한다. 셋째, 각 포인트 링크에 대해서 최소경계영역(Minimum Bounding Box: MBB)을 설정한다. 넷째, MBB의 중심점을 계산하여 대표 궤적을 생성한다. 마지막으로, 그룹에 포함된 궤적의 수가 K이상이 될 때까지 생성된 대표 궤적을 기준 궤적으로 하여 위 과정을 반복한다.

궤적 정보 보호 단계를 거친 결과로 생성되는 대표 궤적은 실제 방문 포인트 등의 원 궤적의 특징을 지니고 있지 않으므로, 방문 빈도가 높은 구간을 찾아내는 등의 궤적 데이터 마이닝 응용에는 적합하지 않다. 이를 위해 GBA 기법은 포인트 링크를 통해 궤적을 재구성하는 알고리즘을

제공한다. GBA 기법의 궤적 재구성 과정은 다음과 같다. 첫째, 각 포인트 링크에서 임의로 하나의 포인트를 선정한다. 둘째, 선정된 포인트를 포인트 링크 순서에 따라 연결하여 하나의 궤적으로 구성한다. 셋째, 이미 선정된 포인트를 제외하고, 위와 같은 과정을 반복 수행한다.

3. 길이 기반 warping 기법을 이용한 궤적 k-익명화 알고리즘

3.1 Preliminary

도로 네트워크상에서 사용자의 실제 이동 경로를 표현하기 위해, 도로 네트워크상에서의 거리를 정의한다. 일반적으로 도로 네트워크는 교차로를 의미하는 노드(node)의 집합 V와 도로를 의미하는 에지(edge)의 집합 E로 표현되는 그래프 G(V, E)로 표현된다. 따라서 도로 네트워크상에서 이동하는 이동객체는 통과한 노드의 정보를 이용하여 궤적을 표현한다. 즉, 궤적은 이동객체가 통과한 노드 및 해당 노드를 통과할 때의 시간에 대한 정보의 집합이다. 이는 시간을 기준으로 정렬되어 있으며, 연속한 두 노드는 하나의 에지로 연결되어 있다. 정의 2는 도로 네트워크 상에서의 궤적을 나타낸다.

정의 2. 도로 네트워크 상에서의 궤적

도로 네트워크의 노드 $v \in V$, 에지 $e \in E$ 이면,

- $tr=\{(v1, t1), ..., (vn, tn)\}$, if $i<j$ then $ti< tj$, if $v1 \in e$ then $vi+1 \in e$

이때, ti 는 이동 객체가 노드 vi 를 통과할 때의 시간 정보를 의미한다.

한편, 도로 네트워크 상에서 궤적 간 유사도를 측정하기 위해, 기존 연구 [14]를 참조하여 시공간 거리를 정의한다. 정의 3 은 서로 다른 두 궤적 사이의 시공간 거리를 나타낸다.

정의 3. 두 궤적 사이의 시공간 거리

각각 n, m 개의 노드를 지나는 두 궤적 t_{ra} 와 t_{rb}에 대해, w_{spt}가 공간 거리에 대한 가중치, w_{time}가 시간 거리에 대한 가중치라 하면, 공간거리(SS), 시간거리(TS), 시공간거리(STS)는 다음과 같다.

공간거리 : $$SS(tr_a, tr_b) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} D_{nt}(tr_{ai}, tr_{bi})$$

시간거리 : $$TS(tr_a, tr_b) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} D_{time}(tr_{ai}, tr_{bi})$$

시공간거리 : $$STS(tr_a, tr_b) = w_{spt} \times SS(tr_a, tr_b) + w_{time} \times TS(tr_a, tr_b)$$

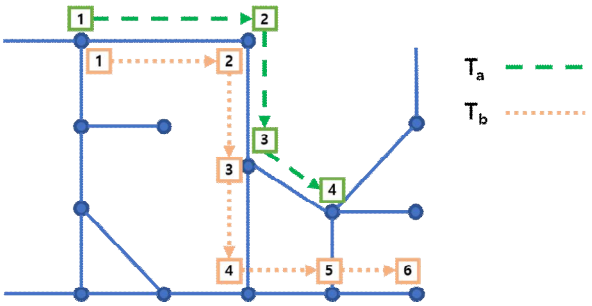
한편, 기존 기법의 경우 궤적 간 유사도 측정을 위해 길이가 긴 궤적의 노드를 일부 삭제하여

궤적이 왜곡되는 문제점이 발생한다. 이를 해결하기 위해, 길이 기반 warping(length-based warping: l-warping) 기법 [15]을 사용하여 일부 궤적 데이터를 중복시켜 유사도를 측정한다. L-warping 기법은 포인트 매칭 시, 포인트의 중복을 허용함으로써 포인트의 삭제를 방지하고 각 포인트 링크에 모든 궤적의 포인트가 포함되는 것을 보장한다. 그러나 기존 l-warping 기법은 도로 네트워크를 고려하지 않기 때문에, 제안하는 연구에서는 정의 4와 같이 도로 네트워크를 위한 새로운 l-warping 기반의 시공간 거리를 정의한다.

정의 4. 두 점 사이의 길이 기반 warping(Length-based warping: l-warping)을 이용한 시공간 거리 측정 길이가 각각 n, m (n>m)인 두 궤적 데이터 tra 과 trb에서, 두 점 tra(i+df)와 trb(i) (tra(i+df)∈ tra, trb(i)∈ trb)에 대한 l-warping 기반 시공간 거리 (Dl)는 다음과 같다:

$$D_l(tr_{a(i+df)}, tr_{b(i)}) = \begin{cases} w_{spt} \times D_{nt}(tr_{a(i+df)}, tr_{bi}) + w_{time} \times D_{time}(tr_{a(i+df)}, tr_{bi}) \\ \qquad ,where\ 0 \le df \le n-m \text{ and} 1 \le i \le m \\ \infty \\ \qquad ,otherwise \end{cases}$$

3.2 도로 네트워크 상에서의 궤적 k-anonymization 기법



(a) 두 궤적 Ta, Tb

Tb \ Ta	1	2	3	4
1	0	2	3.5	4.6
2	-	0	1.5	2.6
3	-	-	0	1.1
4	-	-	-	2
5	-	-	-	1
6	-	-	-	2

(b) l-warping 시공간 거리 측정 (Wspt=1, Wtime=0)

[그림 1] l-warping 시공간 거리를 이용한 포인트 매칭

[Fig. 1] Point matching using l-warping spatio-temporal distance

제안하는 궤적 k-anonymization 알고리즘은 크게 궤적 클러스터링 단계, 대표 포인트 생성 단계, 궤적 재구성 단계로 구성된다. 첫째, 동일한 궤적으로 표현될 K개의 유사 궤적을 묶어 하나의 클러스터를 구성한다. 이를 위해 l-warping 기반 시공간 거리를 이용하여 궤적간 유사도를 측정하여 클러스터링을 수행한다. 아울러 클러스터링을 수행함으로써, 대표 포인트 생성을 위한 포인트 링크 집합을 생성한다. 둘째, 각 포인트 링크에서는 평균 이동 속도 및 벡터합을 이용한 이동 방향을 분석함으로써, 링크별 대표 포인트(representative point)를 생성한다. 생성된 대표 포인트 집합은 클러스터 내 모든 궤적을 대표하는 이동 패턴을 나타낸다. 생성된 대표 포인트를 도로 네트워크 맵에 매핑하여 실제 응용에 적합한 궤적 익명화 데이터를 생성한다.

```
Algorithm 1. Trajectory clustering algorithm
input :   trajectories - set of all trajectories
          numTraj - number of Trajectories
          K - for anonymity
output :  ptLinks - point matching for a trajectory set
01. numGroup = numTraj/K;
02. groups = extractSeeds(numGroup);
03. for(tri in trajectoies) {
04.     for(each group gj) {
05.         dist =distNearestTraj(tri, gj);
06.         if(mindist > dist && numElement(gj) < K) {
07.             mindist = dist;
08.             nearest = gj; } }
09.     putTraj(tri, nearest);}
10.  standTraj = selectStandTraj(group);
11.  numPtLink = length(StanTraj);
12.  ptLinks = makeLink(numPtLink);
13.  for(tri in group) {
14.     distTable = (tri, StandTraj);
15.     currGIdx = 0;
16.     for(each point ptj in tri) {
17.         lastGIdx = (numPtLink-length(tri))+j;
18.         nextGIdx = getNearestpoint(stj, standTraj, currGIdx, lastGIdx);
19.      while(currGIdx > nextGIdx){
20.         putpoint(ptLinkscurrGidx, ptj);
21.         currLinkIdx++;}}}}
22. return ptLinks;
```

[그림 2] 궤적 클러스터링 알고리즘

[Fig. 2] Trajectory Clustering Algorithm

제안하는 궤적 클러스터링 단계에서는, 첫째, k-anonymity를 만족시기키 위해, 전체 궤적 수를 k로 나눈 몫을 바탕으로 초기 클러스터 수를 결정하고 이와 동일한 수의 Seed trajectory들을 임의로 추출한다. 둘째, seed trajectory와 가장 가까운 다른 k-1개의 궤적을 선정하기 위해, 제안하는

l-warping 시공간 거리를 측정하여 포인트 매칭(point matching: PM)을 수행한다. 포인트 매칭 수행 결과, 이동객체의 이동 순서에 따른 포인트 링크가 생성된다. 이때, l-warping 기법의 특성에 따라 생성되는 포인트 링크 집합의 수는 궤적의 길이가 가장 긴 궤적의 노드 수와 동일하다. [그림 1]은 포인트 매칭의 예를 나타낸다. 두 궤적 T_a, T_b에 대해, 모든 포인트 사이의 l-warping 거리를 측정한다 [그림 1(b)]. T_a의 궤적 길이는 4이고, T_b의 궤적 길이는 6 이기 때문에, T_b가 기준 궤적으로 선정된다. 전체 포인트 링크 쌍의 수는 T_b의 궤적 길이와 동일한 6개가 생성된다. 다음, T_b의 각 노드를 기준으로 가장 작은 l-warping 거리를 지닌 포인트 링크 쌍을 선택하여 포인트 매칭을 수행한다. [그림 2]는 제안하는 궤적 클러스터링 단계의 알고리즘을 나타낸다.

대표 포인트(representative points) 생성 단계는 클러스터 생성 시 포인트 매칭이 수행되며, 포인트 매칭 결과를 바탕으로 대표 이동객체의 움직임을 구성할 수 있다. 첫째, 첫 포인트 링크에 포함된 포인트들의 위치 정보를 이용하여 중점을 계산한다. 둘째, 이전 포인트에서 현재 포인트로의 위치 변화 및 시간 변화를 이용하여 각 궤적의 속도를 계산한 후, 해당 포인트 링크의 평균 속도 및 평균 시간을 계산한다. 아래 식은 평균 속도(V_{tr}) 를 계산하기 위한 식을 나타낸다.

$$V_{tr_i}(i) = \frac{distance(location_{tr_i}(i), location_{tr_i}(i-1))}{T_{tr_i}(i) - T_{tr_i}(i-1)}, where\ i > 1$$

셋째, 유클리디언 공간을 가정하고 대표 이동객체가 이전 포인트 링크의 위치에서 평균 속도로 평균 시간만큼 이동하였을 경우, 대표 이동객체의 현재 위치를 계산한다. 아래의 식은 대표 이동객체의 현재 위치를 계산하는 식이다.

$$REPLocation n_i = REPLocation n_{i-1} + V_{AVG}(i) \times (T_{AVG}(i) - T_{AVG}(i-1), where\ i > 1$$

생성된 대표 포인트는 실제 도로 네트워크에 적합하지 않기 때문에, 가까운 노드를 탐색하여 적합한 에지에 매핑을 수행해야 한다. 그러나 단순히 가장 가까운 노드에 'i'번째 포인트를 매핑할 경우, 'i-1' 번째 포인트 또는 'i+1'번째 포인트에 연결이 되지 않는 문제점이 발생할 수 있다. 이를 해결하기 위해 노드 간 연결 및 궤적 출현 빈도를 고려하여 맵 매핑을 수행한다. 'i'번째 대표 포인트에서 유클리디안 거리를 바탕으로 범위 탐색을 수행하여 가장 가까운 n개의 노드를 탐색한다. 탐색된 n개의 노드 중, 대표 포인트의 좌표와 동일한 좌표를 지니는 노드가 존재한다면 해당 노드에 매핑한다. 동일한 좌표를 지닌 노드가 존재하지 않는다면, 'i-1'번째 노드와 연결되는 노드가 존재하는지 탐색한다. 'i-1'번째 노드와 연결되는 노드가 여러 개일 경우, 클러스터 내 포함된 궤적의 출현 빈도수가 더 높은 노드를 선택하며, 궤적 출현 빈도까지 동일할 경우, 다른 노드로의 연결수가 더 많은 노드를, 매핑을 수행할 기준 노드로 선정한다. 기준 노드가 선정되면, 'i-1'번째 노드와 선정된 노드를 연결하는 에지 상에 이동속도를 고려하여 위치를 결정함으로써 맵 매핑을 종료한

다. [그림 3]은 대표 포인트 선정 단계의 알고리즘을 나타낸다.

```
Algorithm 2. Representative point generation algorithm
input :  ptLinks - Point links for a trajectory set
output : reppoints - A representative points for each point group
01.  for(each point link ptgi)
02.    if(i==0) reppointsi = calcFirstpoint(ptgi);
03.    else {
04.        avgTime = calcAvgTime(ptgi);
05.        velocity = calcVelocity(ptgi-1, ptgi, timevlaue);
06.        timevalue = avgTime - lasttime;
07.        reppointsi = calcCurrpoint (reppointi-1, velocity, timevalue);}
08.    lasttime = avgTime;}
09.  for(each point sti in reppoints){
10.    NCross = calcNearCross(sti, n);
11.    for(each Cross point ncp in NCross) {
12.          getPriority(sti, ncp);}}
13.    SelNC = NCross.pop();
14.    sti = mapping(SelNC);
15. return reppoints;
```

[그림 3] 대표 포인트 선정 알고리즘

[Fig. 3] Representative point selection algorithm

궤적은 데이터 마이닝을 통해 교통량 측정, 시간대 별 주요 이용 도로 분석 등 다양한 응용에 활용된다. 그러나 궤적 정보 보호를 수행하면 K개의 궤적이 동일한 궤적으로 표현되기 때문에 궤적 데이터 마이닝에 사용하기에는 적합하지 않다. 따라서 k-익명화 된 결과를 궤적 데이터로 재구성하는 작업인 궤적 재구성 단계가 필수적이다. 궤적 재구성 시 생성된 궤적과 원본 궤적 사이의 왜곡도를 감소시키기 위해, 포인트 링크 내에서의 출현 빈도(궤적 통과 수)를 고려한 포인트 선정을 수행한다. 각 포인트의 출현 빈도는 궤적의 교차를 의미하기 때문에, 출현 빈도가 가장 작은 순으로 다음 포인트를 선정한다. 즉, i+1번째 포인트 링크에서 i번째 포인트와 에지로 연결된 포인트가 둘 이상이라면, 둘 중 출현 빈도가 작은 포인트를 i+1번째 포인트로 선정한다. 출현 빈도가 높은 포인트는 해당 포인트로 연결되는 궤적의 수가 많음을 의미하기 때문에, 출현 빈도가 높은 포인트를 남기면 다음 궤적 재구성 시, 연결될 포인트가 없는 경우의 발생 확률을 감소할 수 있다. 이때, l-warping 알고리즘에 의해 중복된 포인트의 경우, 실제 출현 빈도보다 많은 출현 빈도를 지니게 된다. 따라서 출현 빈도에 음수를 표시함으로써 중복된 포인트임을 명시한다. 중복된 포인트를 선택한 경우, 다음 포인트 링크에서도 동일한 포인트를 선택하도록 함으로써 궤적 데이터의 왜곡을 최소화한다. [그림 4]는 궤적 재구성 단계의 알고리즘을 나타낸다.

```
Algorithm 3. Reconstruction Algorithm
input : stGroups - groups of points for a trajectory set
output : reconTrajs - K reconstructed trajectories
01. for (j<K)
02.    for (each group stgi)
03.       if(i==0) currpoint = selectMinCount(stgi);
04.       else
05.          if(isWarped(reconpointi-1))
06.             currpoint = findSamepoint(stgi, reconpointsi-1);
07.          if(!isWarped(reconpointi-1) || currpoint not existed)
08.             connectedpoints = connectedpoint(stgi, reconpointsi-1);
09.             for(each point cst in connectedpoints)
10.                if(isWarpedpoint(cst))
11.                   deletepoints(connectedpoints);
12.                currpoint = selectMinCount(connectedpoints);
13.          reconpointsj = currpoints;
14.       reconTrajsj = makeTraj(reconpoints);
15. return reconTrajs;
```

[그림 4] 궤적 재구성 알고리즘

[Fig. 4] Trajectory reconstruction algorithm

4. 성능평가

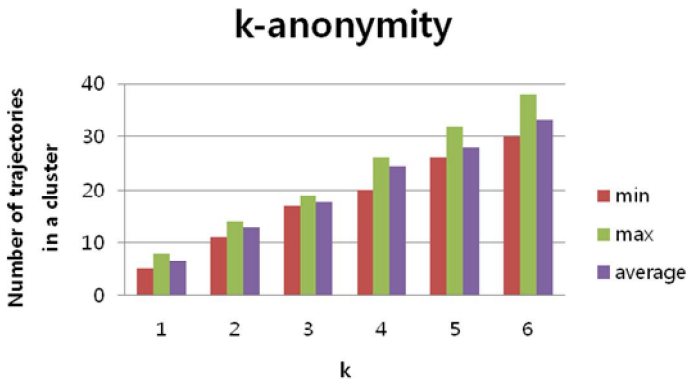
제안하는 l-warping 기법 기반의 궤적 k-익명화 알고리즘의 성능 평가를 수행하였다. 성능 평가 환경은 표 1과 같다. 성능 평가를 위해 사용된 데이터는 17만개의 노드와 22만개의 에지로 구성된 샌프란시스코 만의 도로 네트워크 데이터이며, Brinkhoff [16] 알고리즘을 사용하여 이동객체 궤적을 생성하였다. 이 때, 궤적 길이의 차이에 따른 성능을 측정하기 위해 궤적 길이의 표준 편차가 0, 5, 10, 15, 20, 25, 30인 궤적 데이터를 각각 1000개씩 생성하였고, 데이터의 평균 궤적 길이는 70이다. 또한, 제안하는 기법과 기존 기법 모두 유사도 계산 시 적용되는 시간/공간 가중치 비율을 0.5/0.5로 고정하였으며, 제안하는 기법의 전처리 과정으로 각 노드마다 1000개의 최근접 노드와의 네트워크 거리를 메모리에 적재 하였다. 성능 분석은 궤적 정보 보호 알고리즘과 궤적 재구성 알고리즘을 제공하는 GBA 기법 [13]과 성능을 비교한다.

[표 1] 성능평가 환경

[Table 1] Performance evaluation environment

Hardware	HP ML 150 G3 Server
CPU	Intel Zeon 3.0GHz dual
Memory	2GB
HDD	250GB
OS	Window 2003 server

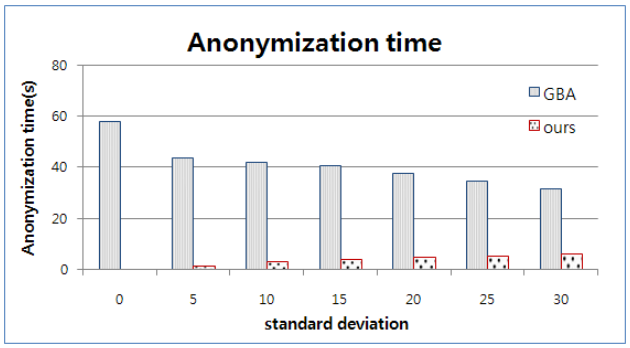
첫째, 궤적 데이터의 정보 보호도를 확인하기 위해, 각 클러스터 내의 궤적 수를 측정하였다. 만약 클러스터 내에 k개 미만의 궤적이 포함되는 경우, k-익명화를 수행하지 못하기 때문에 정보 보호도가 감소한다. [그림 5]는 k 증가에 따른 k-익명화 수행 후 클러스터 내 궤적 데이터 수를 나타낸다. 제안하는 기법의 경우, 초기 클러스터 생성시 k개 미만의 궤적 데이터를 포함하는 경우, 가장 가까운 클러스터와 병합을 수행하여 모든 클러스터가 k개 이상의 궤적 데이터를 포함한다. 이를 통해 궤적 노출 위험도가 언제나 1/k보다 낮은 장점이 존재한다.



[그림 5] 클러스터 내 궤적 데이터 수

[Fig. 5] The number of trajectory in cluster

둘째, 수행 시간 측면에서 제안하는 기법의 효율성을 증명하기 위해, GBA와 궤적 정보 보호 단계의 수행 시간을 비교하였다.



[그림 6] 궤적 길이의 분산에 따른 궤적 k-익명화 수행 시간(k=20)

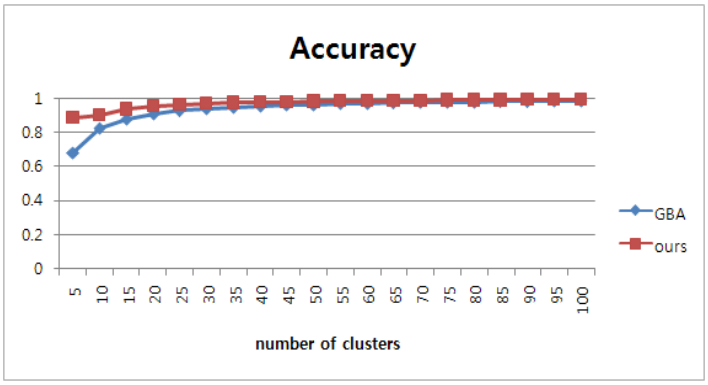
[Fig. 6] Execution time according to standard deviation of trajectory for trajectory k-anonymization

[그림 6]은 궤적 길이의 분산에 따른 궤적 정보 보호 수행 시간을 나타낸다. K값을 20으로 고정하고 궤적 길이의 표준 편차를 0~30으로 변경하여 수행한 결과, 표준 편차가 0일 때와 30일 때, 제

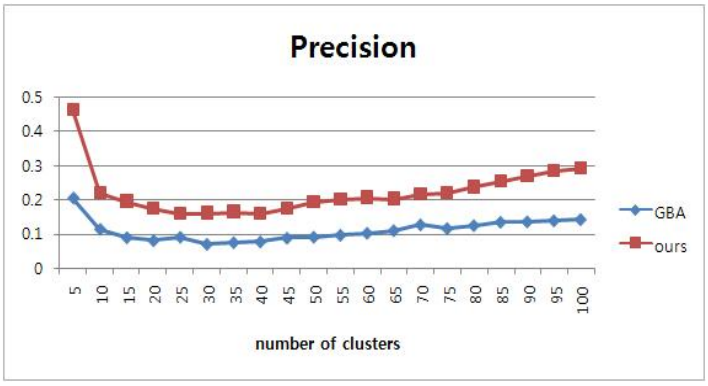
안하는 기법은 각각 0.5초, 6.4초, GBA 기법은 각각 58초, 32초의 성능을 보여, 제안하는 기법이 GBA 기법보다 약 100배에서 5.5배 좋은 성능을 보인다. 한편, 표준 편차가 커짐에 따라 제안하는 기법의 궤적 정보 보호 수행 시간은 증가하고, GBA 기법의 궤적 정보 보호 수행 시간은 감소하였다. 제안하는 기법은 l-warping 기법을 적용하여 궤적의 포인트를 보존하기 때문에, 포인트가 중복하여 저장된다. 이로 인해 궤적 길이의 편차가 클 경우, 포인트 매칭 및 대표 궤적 생성 시에 계산량이 증가한다. 반면, GBA 기법은 다른 궤적과 포인트 링크로 연결되지 않은 포인트를 삭제하기 때문에 포인트의 수가 감소한다. 따라서 궤적 길이의 편차가 클수록 많은 포인트가 삭제되므로 포인트 매칭 및 대표 궤적 생성 시에 계산량이 감소한다.

재구성된 궤적의 데이터마이닝 적합성은 재구성된 데이터와 원본 데이터에 대해 각각 클러스터링을 수행하여, 각 클러스터에 포함된 궤적을 비교하였다. 비교한 결과를 4개의 경우 (true positive(TP), true negative(TN), false positive(FP), false negative(FN))로 나누어 accuracy, precision, recall을 측정한다 [17][18]. 궤적 클러스터링은 complete-link 알고리즘을 이용하였으며, 클러스터링을 위한 궤적간의 거리는 시계열 데이터 분석에 사용되는 ERP distance metric [19]을 적용하여 계산하였다. [그림 7]은 클러스터의 수에 따른 재구성된 궤적의 적합성을 나타낸다. K값 및 궤적 길이의 편차를 각각 20, 10으로 고정하여 성능평가를 수행하였다. Accuracy의 경우, 두 기법 모두 클러스터 개수가 10 이상일 때, 제안하는 기법은 0.9이상, GBA 기법은 0.8이상의 정확도를 보이며, 클러스터의 수가 많아질수록 성능 향상을 보인다 [그림 7(a)]. 이는 클러스터의 수가 많아질수록 클러스터에 속한 궤적의 수가 적어지므로, 두 경우 모두 특정 클러스터에 속하지 않은 데이터의 수(TN)가 증가하기 때문이다. Precision의 경우, 클러스터의 수가 5일 때, GBA와 제안하는 기법이 각각 0.2, 0.45의 정확도를 보이고, 클러스터의 수가 100일 때, 각각 0.15, 0.3의 정확도를 보인다 [그림 7(b)]. 제안하는 기법이 GBA기법 보다 약 2배의 성능 향상을 보이는데, 이는 제안하는 기법이 궤적 데이터의 삭제를 방지함으로써 GBA보다 원본 궤적과 유사한 재구성 궤적을 생성하기 때문이다. 한편, 두 기법 모두 클러스터의 수가 5일 때와 45 이상일 때, 좋은 성능을 보인다. 클러스터 수가 5인 경우에는 하나의 클러스터에 포함되는 궤적의 수가 매우 많으므로, 재구성 후 유사도가 높지 않은 궤적도 동일한 클러스터에 포함될 확률이 증가한다. 반면, 클러스터의 수가 45이상인 경우에는 하나의 클러스터에 포함되는 궤적의 수 k가 20개 이하이기 때문에, 동일한 클러스터에 포함될 확률이 증가한다. 즉, 원본 궤적의 클러스터링 시에 하나의 클러스터에 포함되었던 궤적이 재구성 후 동일한 클러스터에 포함될 확률이 높으므로 precision이 증가한다. Recall의 경우, 클러스터의 수가 5일 때, GBA와 제안하는 기법이 각각 0.25, 0.47의 정확도를 보이고, 클러스터의 수가 100 일 때, 각각 0.46, 0.48의 정확도를 보인다 [그림 7(c)]. 특히, 클러스터의 수가 40 이상인 경우, 두 기법 모두 약 0.35~0.5의 정확도로 유사한 결과를 나타낸다. 결론적으로 제안하는 기법은 유사한 원본 궤적이 재구성 후에도 유사하게 판단되는 비율은 GBA와 거의 동일하거나 약간 향상되었지만, 유사

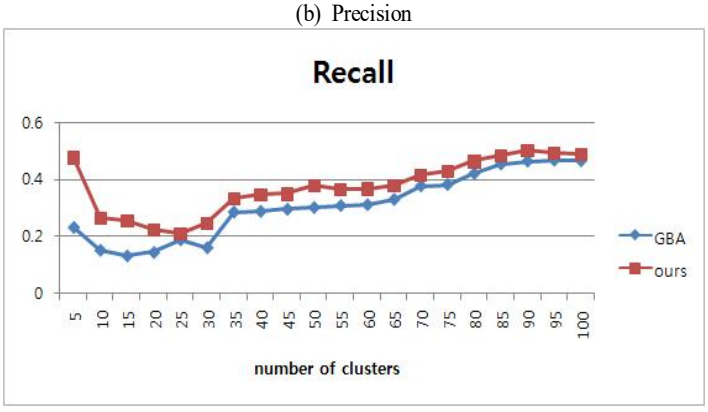
하지 않은 궤적이 재구성 후에도 유사하지 않게 판단되는 비율은 GBA보다 향상되었음을 알 수 있다.



(a) Accuracy



(b) Precision



(c) Recall

[그림 7] 재구성된 궤적의 성능 비교

[Fig. 7] Comparison of Reconstructed Trajectories

5. 결론

사용자의 위치 및 시간 정보를 포함하는 이동객체의 궤적 데이터는 다양한 위치 기반 서비스에 사용될 수 있다. 이때, 궤적 데이터에 포함된 사용자의 위치 및 시간 정보는 중요한 개인 정보이기 때문에, 이를 보호하기 위한 효율적인 궤적 k-익명화 기법이 필요하다. 이를 위해 본 논문에서는 도로 네트워크에서 l-warping 기반의 효율적인 궤적 anonymization 기법을 제안하였다. 첫째, 도로 네트워크의 특성을 반영하기 위해 도로 네트워크 상에서의 이동 객체 간 거리를 정의하고, 이를 통해 궤적 간의 유사도를 측정한다. 둘째, l-warping 기법을 사용함으로써, k-익명화 수행 중 삭제되는 부분 궤적의 비율을 줄임으로써, 궤적 k-익명화에 의한 데이터 왜곡도를 감소시킨다. 셋째, 제안한 대표 궤도 생성 알고리즘을 통해, 실제 도로 네트워크 상에서의 이동 경로를 반영하는 k-익명화된 궤적 데이터를 생성한다. 이를 통해 다양한 LBS 응용에서 사용자 정보를 보호함과 동시에 높은 정확도를 지원하는 것이 가능하다. 마지막으로 제안한 궤적 재구성 알고리즘을 통해 원본 궤적 데이터와 유사한 궤적을 재구성함으로써 데이터 마이닝 응용에 적합하다. 성능평가 결과 궤적 k-익명화 수행 시간 및 재구성 궤적의 정확성 측면에서 제안하는 기법이 기존 GBA 기법에 비해 우수한 실험 결과를 나타낸다. 향후 연구로는 시간대에 따른 트래픽 패턴 분석을 위해 지능형 교통 시스템에 제안하는 기법을 적용할 계획이다.

References

[1] D. Pfoser, C. S. Jensen, Y. Theodoridis, “Novel Approaches to the indexing of Moving Object Trajectories”, The 26th International Conference on Very Large Databases, September 10-14, 2000, Cairo, Egypt, pp. 395-406.

[2] F. Giannotti, D. Pedreschi, “Mobility, data mining and privacy: Geographic knowledge discovery”, Springer Science & Business Media, 2008

[3] R. J. Bayardo, R. Agrawal, “Data privacy through optimal k-anonymization”, The 21st International conference on data engineering, April 5-8, 2005, Tokyo, Japan, pp. 217-228, doi: 10.1109/ICDE.2005.42.

[4] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, “Mondrian multidimensional k-anonymity”, The 22nd International conference on data engineering, April 3-7, 2006, Atlanta, GA, USA, pp. 25-25, doi: 10.1109/ICDE.2006.101.

[5] M. Gruteser, D. Grunwald, “Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking”, The 1st international conference on Mobile systems, applications and services, May 5-8, 2003, San Francisco, California, USA, pp. 31-42, doi: 10.1145/1066116.1189037.

[6] B. Gedik, L. Liu, “Location-Privacy in Mobile Systems: A personalized Anonymization Model”, The 25th IEEE International Conference on Distributed Computing Systems, June 6-10, 2005, Columbus, OH, USA,

pp. 620-629, doi: 10.1109/ICDCS.2005.48.

[7] M. F. Mokbel, C. Chow, W. G. Aref, “The new Casper: query processing for location services without compromising privacy”, The 32nd international conference on Very large databases, September 12-15, 2006, Seoul, Korea, pp. 763-774.

[8] P. Kalnis, G. Ghinita, K. Mouratidis, D. Papadias, “Preventing Location-Based Identity Inference in Anonymous Spatial Queries”, IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 12, December 2007, pp. 1719-1733, doi: 10.1109/TKDE.2007.190662.

[9] V. Torra, G. Navarro-Arribas, “Big data privacy and anonymization”, IFIP International Summer School on Privacy and Identity Management, August 21-26, 2016, Karlstad, Sweden, pp. 15-26, doi: 10.1007/978-3-319-55783-0_2.

[10] P. Khan, Y. Khan, S. Kumar, “Single Identity Clustering-Based Data Anonymization in Healthcare”, Computationally Intelligent Systems and their Applications, April 2021. pp. 1-9. doi: 10.1007/978-981-16-0407-2_1.

[11] M. Terrovitis, N. Mamoulis, “Privacy preservation in the publication of trajectories”, The International Conference on Mobile Data Management, April 27-30, 2008, Beijing, China, pp. 65-72, doi: 10.1109/MDM.2008.29.

[12] O. Abul, F. Bonchi, M. Nanni, “Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases”, IEEE international conference on data engineering, April 7-12, 2008, Cancun, Mexico, pp. 376-385, doi: 10.1109/ICDE.2008.4497446.

[13] M. E. Nergiz, M. Atzori, Y. Saygin, “Towards Trajectory Anonymization: a Generalization-Based Approach”, International Workshop on Security and Privacy in GIS and LBS, November, 4, 2008, Irvine, California, USA, pp. 52-61, doi: 10.1145/1503402.1503413.

[14] Y. C. Kim, J. W. Chang, “A New Similar Trajectory Search Algorithm Based on Spatio-Temporal Similarity Measure for Moving Objects in Road Networks”, IEICE TRANSACTIONS on Information and Systems, vol. E92-D, no. 2, February 2009, pp. 327-331, doi: 10.1587/transinf.E92.D.327.

[15] C. B. Shim, J. W. Chang, “A new similar trajectory retrieval scheme using k-warping distance algorithm for moving objects”, International Conference on Web-Age Information Management, August, 17-19, 2003, Chengdu, China, pp. 433-444, doi: 10.1007/978-3-540-45160-0_43.

[16] T. Brinkhoff, “A Framework for Generating Network-Based Moving Objects”, GeoInfomatica, vol. 6, no. 2, June 2002, pp. 153-180, doi: 10.1023/A:1015231126594.

[17] G. Salton, “Automatic text processing: the transformation, analysis, and retrieval of information by computer”, Published by Addison-Wesley Company Inc., 1989.

[18] R. R. Korfhage, “Information Storage and Retrieval”, John Wiley & Sons Inc., 1997.

[19] L. Chen, R. Ng, “On the marriage of lp-norms and edit distance”, The Thirtieth international conference on Very large databases, August 31-September 3, 2004, Toronto, Canada, pp. 792-803, doi: 10.1016/B978-0120884469-8.50070-X.