도로 네트워크 환경을 위한 궤적 클러스터링의 성능 평가

Performance Evaluation of Trajectory Clustering in Road Network Environment

백지행*, 원정임, 김상욱
Ji-Haeng Baek*, Jung-Im Won, Sang-Wook Kim
한양대학교 전자컴퓨터통신공학과
hanghang1010@hanmail.net*, {jiwon, wook}@hanyang.ac.kr

요 약

최근 궤적 정보를 이용한 많은 연구들이 진행되고 있으나, 이들 대부분의 연구는 유클리드 공간 내의 궤적들을 대상으로 하고 있다. 그러나 실제 응용에서 대부분의 이동 객체들은 도로 네트워크 공간상에 존재하므로, 유클리드 공간을 대상으로 한 연구들은 도로 네트워크 공간에 적용시키는 것은 적합하지 않다. 본 논문에서는 도로 네트워크 내 이동 객체들의 대용량 궤적 정보를 대상으로 제안된 선행 연구의 클러스터링 기법을 다양한 실험을 통하여 그 정확도를 검증한다. 실험 결과에 따르면 제안된 기법은 사람에 의하여 유사 궤적들을 클러스터링한 결과와 비교하여 95%이상의 높은 정확도를 보였다.

1. 서론

이동 객체의 궤적 정보를 효율적으로 저장 및 관리하는 기법들에 관한 많은 연구 결과들이 보고되고 있으며, 몇몇 연구에서는 주어진 이동 객체의 궤적과 유사한 궤적을 검색 또는 클러스터링하고, 이를 도로 정보 및 사용자 정보 등과 연계하여 분석하는 것을 시도되고 있다.

이들 궤적은 궤적이 생성되는 공간에 따라 유클리드(Euclidean) 공간상의 궤적과 도로 네트워크(road network) 공간상의 궤적으로 구분할 수 있다. 대부분의 궤적에 대한 연구는 유클리드 공간상의 이동 객체의 궤적 정보를 대상으로 수행되어 왔다. 궤적 간 유사도 측정 방식으로는 주로 유클리드 거리를 사용한다.

그러나 실제 텔레매틱스 응용에서 대부분의 이동 객체들은 도로 네트워크 공간상에 존재하며, 이동 객체의 공간 정보를 파악하기 위하여는 1차원의 도로 정보가 사용자에게 보다 직관적이고, 유용한 정보를 제시할 수 있다.

최근, 유클리드 공간상에서의 연구들이 도로 네트워크 공간상의 연구로 전환되고 있으며, 이동이 제한된 도로 네트워크 공간내의 이동 객체의 궤적 정보를 효과적으로 표현, 저장, 인덱싱 하고자 하는 몇몇 연구가 시도된 바 있다. 그러나 유사 궤적 검색 및 클러스터 기법에 관한 연구는 아직 미흡한 상태이다[Hwa06].

본 논문의 선행 연구 [Bae06] 에서는 이동 객체의 궤적을 하나의 이동 객체가 지나온 도로 세그먼트들의 연속으로 표현하고, 도로 세그먼트 길이 정보를 이용한 유사도 측정 방식과 이를 이용한 클러스터링 기법을 제안한 바 있다.

본 논문에서는 선행 연구에서 제안된 클러스터링 기법의 정확도를 검증하기 위하여 다양한 성능 평가를 수행하고, 그 결과를 분석하여 제안된 기법의 우수성을 보인다.

2. 유사 궤적 클러스터링 방안

이동 객체의 궤적이란 세그먼트의 식별자와 길이의 리스트이며, $T_i = \{(S_1, L_1), \ldots, (S_n, L_n)\}$으로 표현한다. 여기서, T_i은 궤적의 식별자이고, $S_j (1 \le j \le n)$는 세그먼트의 식별자, L_j는 세그먼트의 길이를 나타낸다. 본 장에서는 선행 연구 [Bae06] 에서 제안된 유사 궤적 클러스터링 방

안에 대하여 간략하게 설명한다.

2.1 유사도 측정 함수

궤적 클러스터링을 수행하기 위하여 먼저 궤적들 간의 유사도 측정 함수를 정의하여야 한다. 궤적은 문자열로 표현되는 세그먼트 식별자들의 리스트로 구성되므로, 유사도 측정을 위하여 문자열간의 거리 함수로 많이 사용되는 ED(edit distance)방식을 이용할 수 있다. 그러나 비교되는 두 궤적의 세그먼트 개수가 서로 다른 경우 유사도는 세그먼트 개수가 많은 궤적에 의하여 영향을 받는다. 따라서 ED 방식에 의하여 유사 궤적을 검색하는 것은 적합하지 않다.

선행 연구에서는 이러한 문제점을 해결하기 위하여 세그먼트의 개수에 영향을 받지 않으면서 보다 정확한 유사도 측정을 지원하는 세그먼트의 길이 정보를 이용한 유사도 측정 함수 DSL(dissimilarity with length)을 사용한다.

$$DSL(T_i, T_j) = \frac{(T_i와\ T_j의\ 비공통\ 세그먼트의\ 길이의\ 합)}{(T_i의\ 세그먼트길이의\ 합 + T_j의\ 세그먼트길이의\ 합)}$$

2.2 궤적 클러스터링

궤적 클러스터링이란 궤적간의 유사도를 이용하여 전체 궤적들을 그룹화하는 것을 말하며, 일반적인 클러스터링 방법에서는 클러스터의 무게 중심이라 할 수 있는 중심점의 반복적인 변경에 의해 클러스터를 구성한다. 궤적간의 상대적인 거리로 클러스터를 구성한다면 중심점이라는 기준이 모호해 진다. 이러한 문제점을 해결하기 위하여 선행 연구에서는 FastMap [Fal95]을 이용하여 각 궤적을 k차원 공간 상의 한 점으로 표현한 후, 전체 궤적들과 대응되는 점들을 대상으로 클러스터링을 수행한다. 이때, 서로 다른 길이를 갖는 궤적들을 하나의 차원으로 매핑시키기 위하여 제안된 DSL 방식에 의해 측정된 두 궤적간의 유사도 값을 이용한다.

3. 성능 평가

3.1 실험 환경

본 논문에서는 실험을 위하여 개발된 도로 네트워크 기반의 궤적 데이터 생성기[Bae07]를 이용하여 100개와 10,000개의 궤적 데이터를 생성하여 사용하였다. 도로 네트워크 데이터로는 [Fh05]에서 7,035개의 세그먼트로 구성된 올덴버그(Oldenburg) 도로 네트워크를 다운로드 받아 사용하였다. 궤적 데이터 생성을 위한 방법은 [Bae07]에서 제안된 방법을 사용하였고, 궤적 데이터 생성을 위하여 출발지 노드와 목적지 노드는 임의 선택하여 총 100개를 사용하였으며, 생성된 궤적 데이터들의 평균 세그먼트의 개수는 25개이며, 평균 궤적의 길이는 3000이다.

FastMap[Fal95]을 통하여 k차원 상의 점들로 변환된 궤적들을 대상으로 계층 클러스터링(hierarchical clustering) 방식[Han01]을 이용하여 클러스터링 수행한다. 계층 클러스터링 방식은 가장 유사한 객체들을 동일 클러스터로 그룹화하는 방식으로 단계적인 병합 과정을 수행하여 모든 객체들이 하나의 클러스터에 속할 때까지 클러스터링 과정을 반복하는 방식이다. 제안된 기법의 성능 평가를 위하여 유사도 함수의 오차율과 클러스터링의 정확도를 비교 분석한다. 정확도 분석을 위한 비교 대상으로는 사람에 의하여 구성된 클러스터 집합을 사용한다.

3.2 실험 결과

실험 1과 실험 2에서는 클러스터링을 수행하기 전에 FastMap을 위한 최적의 차원수를 결정하는 실험을 수행한다. 먼저 실험 1에서는 제안된 유사도 함수 DSL을 이용하여 측정된 실제 궤적간의 거리 차이와 FastMap을 이용하여 궤적을 k차원 공간상의 점으로 변환했을 때 궤적들 간의 거리 차이를 이용한 오차율을 다음과 같이 정의하여 사용한다.

$$오차율 = \frac{\sum_{i=1}^{n}\sum_{j=1(\neq i)}^{n}(DSL(T_i, T_j) - EU(T'_i, T'_j))^2}{n(n-1)}$$

위의 식에서 n은 전체 궤적의 개수를 의미하며, $n(n-1)$는 이들 궤적을 조합하여 생성 가능한 전체 궤적 쌍의 개수를 의미한다. $DSL(T_i, T_j)$는 제안된 DSL 함수

를 이용하여 궤적 T_i와 T_j 간의 유사도를 측정한 값이며, T'_i와 T'_j는 궤적 T_i와 T_j를 FastMap을 이용하여 변환된 k차원 공간상의 점을 말하며 $EU(T'_i, T'_j)$는 변환된 두 점간의 유클리드 거리를 의미한다.

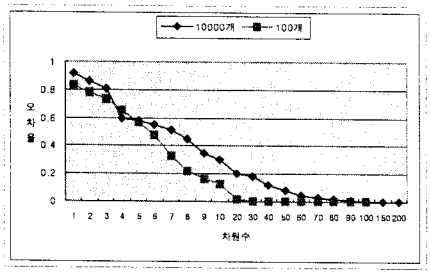

그림 1. FastMap의 차원수 변화에 따른 오차율.

그림 1은 FastMap의 차원수 변화에 따른 오차율을 측정한 결과를 보인다. 실험 데이터로는 100개와 10,000개의 궤적 데이터를 사용하였으며, X축은 FastMap의 변환 차원 수, Y축은 오차율을 나타낸다. 실험 결과에 따르면, 차원이 커질수록 오차율이 작아지는 것을 볼 수 있다. 100개의 궤적 데이터인 경우에는 10차원 이상에서 0.1이하의 오차율을 보이며, 10,000개의 궤적 데이터의 경우에는 50차원 이상에서 0.1이하의 오차율을 보였다.

실험 2에서는 실제 궤적들 간의 유사도 순위가 FastMap으로 변환 후에 변화가 있는지를 실험한다. 이를 위하여 참고 문헌 [Hav02]에서 정의된 KSim을 사용한다.

수식은 실제 궤적 T_i와 궤적 T_j 간의 유사도 순위 $order(T_i, T_j)$가 이들 두 궤적을 FastMap에 의하여 k차원 공간상의 한 점으로 변환한 궤적 T'_i와 궤적 T'_j간의 유사도 순위 $order(T'_i, T'_j)$와 달라진 궤적들의 비율을 의미한다.

$$KSim = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \begin{cases} 0 & \text{if } order(T_i, T_j) = order(T'_i, T'_j), T_i \neq T_j \\ 1 & \end{cases}$$

그림 2는 FastMap의 차원수 변화에 따른 궤적간의 유사도 순위 변화를 측정한 결과를 보인다. X축은 FastMap의 차원수를 나타내며, Y축은 유사도 순위가 변하는 궤적의 비율을 나타낸다. 실험 결과에 따르면, 변환 차원이 증가됨에 따라 유사도 순위가 변하는 궤적의 수가 확연히 감소하는 것을 알 수 있다.

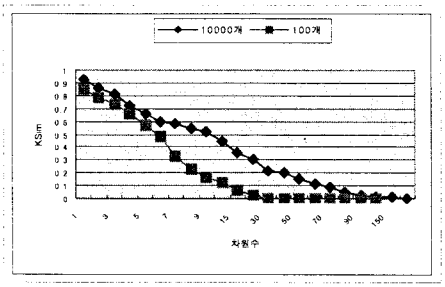

그림 2. FastMap의 차원수 변화에 따른 궤적간의 유사도 순위 변화.

실험 1과 실험 2의 결과에 따라 본 실험에서는 FastMap의 변환 차원을 50차원으로 고정하고 이후 클러스터링 실험을 수행한다.

실험 3에서는 정답 집합과 제안하는 기법을 이용하여 얻어진 클러스터링 결과 집합을 비교 분석한다. 정답 집합은 일반인 5명을 대상으로 본 실험에서 사용하고 있는 10,000개의 궤적 데이터를 5개부터 50개까지의 클러스터로 구성하게 하여 얻어진 집합이다. 구성된 각 정답 집합의 클러스터내에 포함되는 궤적들은 대상자들마다 약간의 차이가 발생할 수 있으므로 이 경우 더 많은 사람들이 선호하는 클러스터로 궤적을 포함시킨다. 또한, 제안된 기법의 클러스터링 결과에 대한 정확도 측정을 위한 척도로 참고 문헌 [Gol06]에서 제안된 정답 집합과 구성된 클러스터들간의 중심 값들을 비교하는 clusterdistance를 사용한다.

$$clusterdistance(A,B) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1(\neq i)}^{n} \min[dist(\overline{a_i}, \overline{b_j})]$$

수식은 본 논문에서 제안된 클러스터링 기법을 사용하여 구성된 클러스터 집합 $A = (\overline{a_1}, \overline{a_2}, ..., \overline{a_k})$와 사람에 의하여 구성된 클러스터 집합 $B = (\overline{b_1}, \overline{b_2}, ..., \overline{b_k})$간의 거리를 구하는 것으로, 여기서 $\overline{a_i}$와 $\overline{b_j}$는 클러스터 집합 A와 B를 구성하는 각 클러

스터들의 중심값이다. $dist(\overline{a_i}, \overline{b_j})$는 두 클러스터 중심값 $\overline{a_i}$와 $\overline{b_j}$간의 유클리드 거리를 의미한다.

그림 3은 10,000개의 궤적 데이터에 대하여 정확도를 측정한 결과를 보인다. 실험 결과에 따르면 클러스터의 개수가 작을수록 100%에 가까운 높은 정확도를 보였으며, 클러스터의 개수가 많아지면 정확도가 다소 낮아짐을 알 수 있다. 그러나 제안하는 기법은 95%이상의 높은 정확도를 나타냄으로써, 비교적 정확한 클러스터링의 결과를 얻을 수 있음을 알 수 있다.

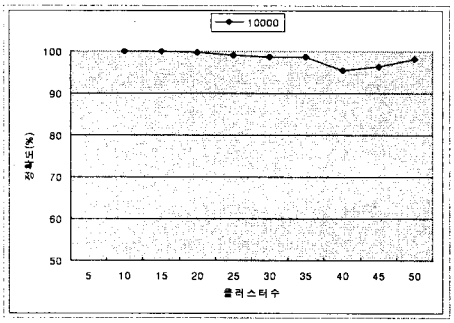

그림 3. 클러스터의 개수 변화에 따른 정확도의 변화.

4. 결론

본 논문에서는 실험을 통하여 선행 연구 [Bae06]에서 제안된 도로 네트워크 공간상에서의 이동객체의 궤적들을 대상으로 하는 클러스터링 기법의 정확성을 검증하였다. 실험 결과에 따르면, 제안된 클러스터링 기법은 사람에 의하여 유사 궤적들을 클러스터링한 결과와 비교하여 95%이상의 높은 정확도를 보였다.

감사의 글

본 논문은 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었습니다.
(IITA-2008-C1090-0801-0040)

참고 문헌

[Bae06] 백 지행, 원 정임, 김 상욱, "도로 네트워크에서의 유사 궤적 클러스터링," *한국정보과학회 추계학술 발표논문집*, Vol. 33, No. 2(C), pp. 256-260, 2006년.

[Bae07] J. Baek et al., "Generating Trajectories on Road Networks," In *Proc. Int'l. Conf. on Mechatronics and Information Technologys*, ICMIT, pp. 31-31, 2007.

[Fh05] FH Oldenburg/Ostfriesland/Wilhelmshaven, Network-based Generator of Moving Objects, http://www.fh-oow.de/institute/iapg/personen/brinkhoff/generator/, 2005.

[Fal95] C. Faloutsos and K. Lin, "Fast map: A Fast Algorithm for Indexing, Data-Mining, and Visualization of Traditional and Multimedia Datasets," In *Proc. Int'l. Conf. on Management of Data*, ACM SIGMOD, pp. 163-174, 1995.

[Gol06] D. Goldin, R. Mardales, and G. Nagy, "In Search of Meaning for Time Series Subsequence Clustering: Matching Algorithms Based on a New Distance Measure," In *Proc. Int'l. Conf. on Information and Knowledge Management*, pp. 347-356, 2006.

[Han01] J. Han and M. Kamber, Data Mining: Concepts and Techniques, *Academic Press*, 2001.

[Hav02] T. Haveliwala, "Topic-Sensitive PageRank," In *Proc. Int'l. Conf. on World Wide Web*, WWW, pp. 517-526, 2002.

[Hwa06] 황 정래, 강 혜영, 이 기준, "시공간 유사성을 이용 도로 네트워크 상의 유사한 궤적 검색," *정보처리학회 논문지*, Vol. 13-D, No. 3, pp. 337-346, 2006년.