데이터 축소와 군집화를 사용하는 시공간 데이터의 이산화 기법*

강주영° 용환승
이화여자대학교 컴퓨터학과
jykang@ewhain.net, hsyong@ewha.ac.kr

Discretizing Spatio-Temporal Data using Data Reduction and Clustering

Juyoung Kang° Hwan-Seung Yong
Department of Computer Science and Engineering, Ewha Womans University

1. 서론

다양한 이동 객체로부터 수집된 궤적데이터를 기반으로 추출된 시공간 순차 패턴은 객체의 이동 패턴을 파악하고 향후 위치를 예측할 수 있도록 함으로써 위치 기반 서비스의 품질을 높이는데 활용될 수 있다. 기존의 순차 패턴 마이닝 기법들은 항목 기반 트랜잭션 데이터베이스를 대상으로 하기 때문에 연속적인 값을 가지는시공간 데이터를 이러한 기법들에 단순하게 적용할 수 없다. 또한 객체들이 동일한 위치를 지나고 있다고 할지라도 움직임의 편차로 인해 위치 측정값이 서로 다르기 때문에 이러한데이터를 기반으로 빈발한 패턴을 발견하는 것이 더욱 어렵게 된다. 따라서 시공간 데이터를 대상으로 순차 패턴 마이닝을 수행하기 위해서는 연속적인시공간 값을 이산화하는 전처리 단계가 필수적이다. 시공간 데이터를 이산화 하는 가장 전형적인 방법은 데이터 공간의 각 차원을 사용자가 정의한 n개의 구간으로 나누는 균일 격자를 사용하는 방법으로 등간격 구간화 EQW(EQual interval Width)의 2차원적 확장이다. 이 방법은 단순하고 직관적이지만 입력 데이터 속성의 분포를 고려하지 않고 고정 크기의 셀에 데이터를 할당하기 때문에 이산화 과정 동안 데이터의 시공간적 상관 정보를 잃을 수 있다. 본 논문에서는 선 단순화(line-simplification)와 군집화를 기반으로 입력 데이터의 시공간적 특성을 고려하여 이산화를 수행하는 이산화 기법을 제안하였다. 제안된 기법은 선 단순화를 통해 원본 궤적에 대한 근사 궤적을 구함으로써데이터의 수를 축소시키고 마이닝 프로세스의 성능을 향상시킨다. 또한 궤적들을 공간적 지역성과 방향 속성을 고려하여군집화함으로써 이산화 과정 동안 원본 데이터의 시공간적 상관 정보를유지한다.

2. 본론

이동 객체의 궤적은 시간에 따라 연속적으로 움직이는 객체의 공간적 위치 이력데이터이다. 이러한 위치 이력데이터는 특정 시각에서 측정된 객체 위치의2차원 좌표 값의 집합 $S=\{(x_1, y_1, t_1), (x_2, y_2, t_2),...,(x_n, y_n, t_n)\}$로 나타낼 수 있다. 이 때 (x_i, y_i)는 시각 t_i에서의 객체의 위치 좌표이다($t_i<t_{i+1}$). 객체의 위치 측정값 (x_i, y_i)는 노이즈를 포함할 뿐 아니라 움직임의 편차로 인해 같은 위치상의 객체라 할지라도 측정된 연속 값이 정확하게 동일한 값을 가지는 경우는 거의 없다. 따라서 이러한 데이터를 기반으로 순차 패턴 마이닝을 수행하기 위해서는 사전에 마이닝 알고리즘이 요구하는 항목 값 형태로 데이터를 이산화해야 한다. 일반적으로 시공간 데이터를 이산화 하는 방법은 t_i 시각의 객체의 위치 좌표 (x_i, y_i)를 객체가 포함되어 있는 특정 공간 영역의 식별자로 변환하는 방법이다. 연속된 측정 시각 사이의 간격은 고정되어 있으므로 결과적으로 원본 이동 궤적은"$C_1C_2C_3\cdots C_n$"와 같은 영역 식별자의 시퀀스 형태로 이산화 된다. 시공간 속성의 연속 값을 시공간 영역 값으로 변환하는 가장 단순한 방법은 데이터 공간을 고정된 크기의 격자로 분할하는 방법이다. x축과 y축 각각을 사용자가 정의한 계수 C_x, C_y를 기반으로 고정된 수의 구간으로 분할하고 데이터를 $C_x \times C_y$개의 격자 셀 중 공간적으로 사상되는 셀의 식별자 값으로 변환한다. 따라서 연속적인 위치 값은 셀 식별자의 시퀀스 $\{C_1C_2...C_n\}$의 형태로 변형된다. 격자 기반 방법은 단순하고 직관적인 반면 몇 가지 한계점을 가지고 있다. 각 차원에 대해 독립적으로 x, y축을 고정된 구간으로 분할하기 때문에 데이터 내에 있는 공간적 상관 정보가 이산화 단계에서 상실될 수 있다. 정해진 셀의 크기가 객체 변위 분포에 비해 너무 큰 경우 셀 내에서의 객체의 움직임 정보를 잃을 수 있으며 반대로 셀의 크기가 지나치게 작은 경우 두 개의 유사한 궤적이 서로 다른 셀로 이산화 될 수 있다. 이러한 결과를 바탕으로 한 순차 패턴 마이닝의 결과 또한 유의미한 패턴을 잃게 된다[1]. 따라서 입력 데이터 내에 숨겨진 의미 있는 패턴들을 잃지 않기 위해서는 이산화 기법이 이산화 과정 동안 객체의 시공간적 변화에 대한 정보를유지해야 한다.

본 논문에서는 입력 데이터의 시공간적 의미를 유지함과 동시에 데이터의 수를 축소시켜 마이닝 프로세스의 효율성을 높이는 이산화 기법인 STEM(Spatio-TEmporal discretization of Moving objects trajectories)을 제안하였다. STEM의 이산화 단계는 크게 세 단계로 나누어진다. 우선 선 단순화를 통해 원본 궤적에 대한 근사 궤적을 구한다. 오랜 기간에 걸친 이동 객체의 위치 이력을 대상으로 순차 패턴 마이닝을 수행하는 경우 시퀀스의 길이에 따라 마이닝 작업의 연산량은 폭발적으로 증가하게 된다. 따라서 이산화 기법은 데이터 내의 시공간 상관 정보를 잃지 않

* 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-511-D00311)

으면서도 원본 데이터의 크기를 축소시킬 수 있어야 한다. 선 단순화 방법은 결정론적(deterministic) 오차 범위 안에서 선(polyline)을 압축하는 방법이다[2]. STEM에서는 선 단순화 알고리즘 중 수학적 우수성이 검증되었으며 원본 데이터들을 가장 압축적으로 표현하는 DP(*Douglas-Peucker*) 알고리즘을 사용하였다. 기본적인 DP 알고리즘은 궤적 T:{$p_1,p_2,\cdots,p_n$}의 점집합을 부분집합 T':{$p'_1, p'_2,\cdots,p'_s$}$\subseteq T$, $s\leq n$ 으로 재귀적으로 분해한다. $\overline{p_1p_n}$ 선분을 시작으로 그 선분으로부터의 수직거리가 가장 먼 점 p_f을 찾고, 선분으로부터 p_f까지의 거리 Dp_f가 사용자가 정의한 임계값 ε(tolerance) 내에 있다면 주어진 선분을 두 선분을 이루는 두 정점사이의 모든 점들에 대한 근사값으로 받아들인다. 만약 Dp_f의 값이 임계값보다 큰 경우 주어진 선분을 p_f를 기준으로 재귀적으로 분할하여 단순화 작업을 수행한다. 결과적으로 원본 궤적의 p'_l에서 p'_r사이의(단, l, r<n) 모든 점들은 수직거리가 최대 ε인 선분으로 단순화된다. 첫 번째 단계에서 구해진 유향선분들을 다차원 특성 벡터의 형태로 표현될 수 있다. p_l에서 p_r을 두 끝점으로 하는 선분의 특성 벡터 v는 두 끝점과 그 선분이 x축과 이루는 각도 값인 Θ로 이루어진다. 각도는 선분의 방향 값 표현하기 위한 것으로, 이 값을 통해 동일한 영역을 지나는 객체들이 움직이는 방향을 구분할 수 있으며 시공간 객체의 움직임을 더욱 정확하게 표현할 수 있다. 두 번째 단계에서는 STEM은 단순화된 궤적 T':{$p'_1,\cdots,p'_m$}로부터 V :{$v_1,\cdots v_k$} (1<k<m-1)를 도출한다. 여기서 v_k = {P_{k-1}, P_k, Θ_k} 이다. 마지막 단계에서 이 특성벡터들을 군집화 함으로써 객체의 이동 변화를 고려한 공간 영역을 도출한다. 이를 위해 군집화 수행 이전에 특성 벡터의 정규화를 수행하여 특성 벡터 값의 범위를 평준화시킨다. STEM은 최대-최소 정규화를 통해 원본 데이터를 [0,1] 사이의 값으로 선형 변환한다. 이동 객체의 시공간적 변화는 데이터 공간 전체에 균일하게 나타나는 것이 아니다. 따라서 마지막 단계에서 STEM은 입력 데이터의 분포를 기반으로 데이터 공간을 분할하는 영역을 찾기 위해 특성 벡터를 유사한 특징을 가지는 그룹들로 군집화 한다. 군집화 과정의 복잡도를 줄이고 근접도에 대한 임계값을 설정하여 군집화를 조절할 수 있도록 하기 위해 BIRCH[3]의 초기군집화(pre-clustering) 단계를 적용한다. BIRCH의 초기군집화 단계는 입력 데이터 크기에 대해 선형 시간 복잡도를 가지는 군집화 알고리즘으로 데이터의 요약 정보를 CF-트리라는 간결한 트리 구조에 저장한다. 이러한 알고리즘을 기반으로 STEM은 주어진 임계값 ε_c 에 대해 유사한 위치, 크기 그리고 각도 속성 값을 가지는 선분들이 분할된 그룹을 형성하도록 다차원 특성 벡터 값을 군집화 한다. 최종적으로 원본 궤적 데이터는 각각의 데이터가 해당되는 군집의 식별자로 이루어진 일련의 시퀀스로 이산화 된다.

STEM의 성능을 평가하기 위하여 합성 데이터를 이용하여 실험을 수행하고 EQW 기법과 마이닝 및 공간 효율성 그리고 마이닝 프로세스에 대한 영향을 분석하였다. 합성 데이터는 G-TERD 시공간 데이터 생성기를 이용하여 생성하였다. 이산화 기법이 마이닝 결과에 미치는 영향을 평가하기 위해 STEM과 EQW의 이산화 결과를 기반으로 *PrefixSpan* 순차 패턴 마이닝을 수행하였다. 마이닝 결과로 도출된 최대 빈발 패턴을 살펴보면 동일한 궤적에 대해 EQW의 경우 동일한 셀 식별자가 반복적으로 나타나는 길이 33인 패턴을 추출한 반면, STEM을 이용한 경우는 최대 길이 8인 패턴을 추출하며 EQW에 비해 더 간결하고 직관적인 패턴을 생성하였다. 마이닝 결과를 가시화하여 각 기법의 마이닝 결과에 대한 영향을 분석 한 결과 STEM은 실제 입력 데이터의 이동 패턴에 변화가 발생한 점을 기준으로 공간 영역을 분할함을 알 수 있었다. 선 단순화를 통한 데이터 축소 효과와 공간 효율성을 평가하기 위해 데이터 축소율을 측정하였다. 선 단순화 임계값을 실제 길이의 2%부터 16%까지 변화시킨 경우 STEM이 원본 데이터의 크기를 50%까지 축소시켜 저장 공간 효율성을 높이는 것을 보였다. 각 이산화 방법에 따른 마이닝 효율성에 대해 이산화와 마이닝 알고리즘의 수행 시간을 합한 전체 수행 시간을 측정하여 평가하였다. EQW의 격자 개수를 총 2500개로 설정하고 데이터 집합 내의 객체의 수를 증가시키며 전체 수행시간을 측정한 결과 STEM이 EQW에 비해 최대 10배까지 마이닝 성능을 향상시킴을 알 수 있었다. STEM의 경우 이산화 단계에서 선 단순화 및 군집화와 관련된 복잡한 연산을 수행하기 때문에 이산화 단계만의 성능은 EQW에 비해 떨어진다. 하지만 마이닝 단계에서 증가된 효율성이 이산화의 느린 수행 속도를 보완하기 때문에 전체적으로는 STEM이 EQW보다 상당히 좋은 성능을 보였다. EQW는 데이터 축소 효과가 없으며 이산화 된 데이터를 표현하는 셀의 개수가 STEM의 경우 보다 많기 때문에 원본 데이터 수가 증가할수록 마이닝 성능이 현저하게 떨어지는 것을 알 수 있다.

3. 결론

본 논문에서는 시공간 데이터를 효율적으로 이산화하기 위한 이산화 기법을 제시하였다. 제안된 기법은 선 단순화를 이용해 원본 궤적에 대한 근사값을 구하고 이를 군집화 함으로써 이산화 과정 동안 데이터의 시공간 속성 사이의 상관 정보를 유지한다. 사용자가 정의한 고정된 수의 격자로 데이터 공간을 분할하는 기존의 접근 방법과는 달리 제안된 기법은 원본 데이터의 시공간 속성 정보를 고려하여 분할 영역을 생성하기 때문에 마이닝 단계에서 더 직관적이고 이해하기 쉬운 패턴을 생성하도록 한다. 또한 데이터 수를 축소시킴으로써 마이닝 단계의 수행 복잡도를 줄인다. 실험 분석을 통해 제안된 기법이 전체 마이닝 프로세스의 효율성을 높이고 추출된 패턴의 해석성을 높이는 것을 보였다.

참고문헌

[1] Cao, H., Mamoulis, N. and Cheung, D. W., *Mining frequent spatio-temporal sequential patterns*., In Proc. of Data Mining, pp. 82-89, Nov., 2005,

[2] Cao, H., Wolfson, O. and Trajcevski. G., *Spatio-temp oral data reduction with deterministic error bounds*, The VLDB Journal, Vol. 15, No. 3, pp. 221 – 228, Sep., 2006.

[3] Zhang, T., Ramakrishnan, R. and Livny. M., *BIRCH: An efficient data clustering method for very large databases.* In Proc. of SIGMOD, pp. 103 – 114, Jun., 1996.