그리드 셀 기반 공간 클러스터링 방법

이 동규[O+], 정정수[++], 문 상호[+]

위덕대학교 컴퓨터공학과[+], 동명정보대학교 컴퓨터공학과[++]

Grid Cell Based Spatial Clustering Method

Dong-Kyu Lee[O+], Chung-Su Chung[++], Sang-Ho Moon[+]

Dept. of Computer Engineering, Uiduk University[+]

Dept. of Computer Engineering, Tongmyung Univ. of Information Technology[++]

요 약

대용량의 공간 데이터베이스로부터 암시적이고 유용한 지식을 자동적으로 추출하는 공간데이타 마이닝은 데이터 양이 급격히 증가하면서 필요성이 더욱 증대되고 있다. 공간데이타 마이닝에서 데이터를 분석하여 유사한 그룹으로 분류하는 것은 중요한 분야이며, 이를 위해서는 공간 클러스터링 과정이 먼저 수행되어야 한다. 이러한 공간 클러스터링 에서 가장 중요한 점은 클러스터링에 드는 비용의 감소와 점 공간객체에 한정된 클러스터링이 아닌 선 및 다각형 객체들의 클러스터링도 가능해야 한다.

본 논문은 이를 위하여 공간지역성을 보장하는 대표적인 공간분할 방법인 그리드 셀을 이용한다. 기존의 클러스터링에서 사용되는 객체들 간의 거리 계산을 인접한 그리드 셀들 간의 관계 연산으로 대체시키는 것이 핵심 아이디어이다. 이 방법은 기존 클러스터링에서 객체들 간의 거리 계산으로 인한 비용을 현저하게 줄일 수 있고, 선 및 다각형 객체들의 클러스터링도 가능하게 하는 장점이 있다.

1. 서론

공간데이타 마이닝은 일반적인 데이터 마이닝 기법을 공간 도메인에 적용한 경우로, 공간 데이터베이스로부터 암시적이며 잠재적인 지식을 추출하는 과정을 의미한다[1,2,3,4,5,6,7]. 즉, 공간데이타 마이닝은 공간 데이터베이스의 데이터로부터 사용자가 원하는 지식을 자동적으로 추출해 준다. 공간 클러스터링은 공간데이타 마이닝 기법중의 하나로, 공간 객체들에 대하여 공간적 특성을 이용하여 집단화하는 과정이다. 이러한 공간 클러스터링은 일반적인 클러스터링과 마찬가지로 중요한 공간데이타 마이닝 방법으로, 다른 알고리즘의 전처리 단계로 이용되거나 유사성 검색 등의 많은 응용 분야에 널리 사용되고 있다.

데이터 마이닝 작업에서 사용되는 데이터의 크기는 그 특성상 대규모를 이루고 있다. 이러한 대규모의 데이터로부터 지식을 추출하는 작업은 많은 비용이 소모된다. 특히, 공간 데이터베이스는 일반 데이터베이스와는 달리 속성데이타 이외에 점(point), 선(line), 다각형(polygon) 등과 같은 공간데이타를 포함하기 때문에 데이터 양이 방대하다. 이러한 공간데이타의 특성으로 인하여 일반 데이타에 비하여 탐색공간의 복잡도(complexity)가 커진다. 따라서 복잡하고 대용량인 공간데이타를 대상으로 하여 암시적이고 잠재적인 유용한 지식을 발견하기 위해서는 더 많은 비용이 든다.

효율적인 공간데이타 마이닝을 위해서 전체 데이터로부터 의미있는 부분집합(즉, 클러스터)들을 발견하고, 발견된 데이터의 부분집합을 대상으로 유용한 지식을 추출하는 것이 바람직하다. 이러한 이유로 인하여 공간데이타 마이닝에서 공간 클러스터링은 중요한 역할을 담당하며, 탐색공간의 복잡도가 큰 공간데이타 마이닝을 위해서 효율적인 공간 클러스터링 알고리즘의 제시가 필수적이다.

본 논문에서 수행하고자 하는 공간 클러스터링 기법은 앞에서 언급한 기존 공간 클러스터링의 단점을 극복하는데 주안점을 두고자 한다. 특히, 방대한 양의 공간데이타에 대한 효율적인 클러스터링을 위한 비용(계산량) 감소와 점객체 이외의 선 및 다각형 객체의 클러스터링이 가능하게 하는데 목적이 있다. 따라서 본 논문의 내용은 기존의 공간 클러스터링과 전혀 다른 새로운 클러스터링 기법의 제시가 아니라, 공간데이타의 특성으로 인한 탐색공간의 복잡도(complexity)를 효율적으로 처리할 수 있는 클러스터링 기법 제시가 주된 목적이다. 이를 위해서 공간 지역성(spatial locality)을 보장하는 대표적인 공간분할 방법인 그리드 셀을 기반으로 한 공간 클러스터링 기법을 제시한다.

2. 관련연구

기존 연구에서 공간 클러스터링을 위한 방법의 제시가 있었다. 먼저 DBSCAN(Density-based Algorithm for Discovering Clusters in Large Spatial Database with Noise)은 공간데이타 마이닝을 위해 밀도(density)를 기반으로 한 클러스터링 알고리즘이다[2]. 이 알고리즘에서 중요한 매개변수로 Eps와 MinPts를 이용한다. CLARANS(Clustering Large Applications based upon RANdomized Search)는 기존의 데이터 마이닝을 위해 제안된 PAM과 CLARA를 결합하여 제안한 알고리즘이다[1]. 이 방법에서는 데이터 집합의 부분집합만을 검색하며, 검색의 각 단계에서 무작위로 샘플을 생성하여 이용한다. H-SCAN(Hash-based Spatial Clustering Algorithm for kNowledeg Extraction)은 공간데이타를 클러스터링하기 위하여 1-차원을 위한 해시 방법을 확장하여 d-차원 공간에서 사용한 것이다[6]. STING(STatistical INformation Grid) 알고리즘은 그리드 셀 계층(Grid

Cell Hierarchy)을 이용하여 공간 데이터 마이닝을 수행한다[4]. 이 방법의 가장 큰 문제점은 가장 중요한 leaf 노드 셀의 크기을 결정하기 위한 기준값이 제시되지 않고, 단순히 분할 과정에서 셀에 포함되는 객체 수를 비교하여 임의로(눈짐작으로) 값을 결정하는 것이다.

3. 그리드 셀을 이용한 클러스터링

3.1 기존 클러스터링

일반적으로 기존의 클러스터링 과정은 먼저 각 객체들 간의 거리를 계산하여, 임계값(기준값)과 비교하여 거리가 이 값보다 작으면 클러스터에 포함된다. 예를 들어 그림 1과 같은 경우에 객체 *O*를 medoid(기준점)로 하여 각 객체들 간의 거리를 구하면 객체 *A, B*와의 거리 *d1, d2*는 임계치 값내에 있으므로 클러스터에 포함된다. 그러나 객체 *C, D*와의 거리 *d3, d4*는 임계치 값보다 크므로 클러스터에 포함되지 않는다. 따라서 기존 클러스터링 과정에서는 medoid 객체와 다른 객체들 간에 거리 계산에 많은 비용이 든다. 특히 공간 클러스터링인 경우에는 점 객체 이외에 선, 다각형 객체들 간의 거리를 계산해야 하므로 비용이 급격하게 증가한다.

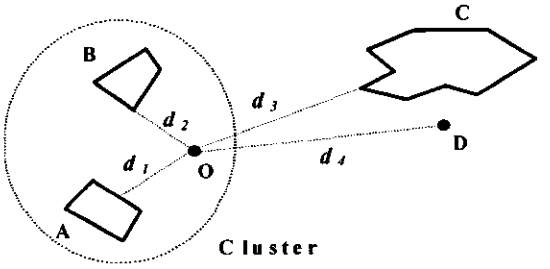


그림 1. 기존 클러스터링 과정

3.2 그리드 셀 기반 클러스터링

기존 클러스터링의 문제점을 극복하기 위하여 공간 지역성(spatial locality)을 보장하는 대표적인 공간분할 방법인 그리드 셀을 기반으로 한 공간 클러스터링 기법을 제시한다.

본 논문에서 제시하는 공간 클러스터링 방법의 핵심 아이디어는 기존 클러스터링 방법에서 객체들 간의 거리 계산을 기반으로 하는 것 대신에 셀들 간의 관계를 이용하여 클러스터링하는 것이다. 이것은 기존 클러스터링 방법에서 거리 계산에 의한 비용을 크게 줄일 수 있는 장점이 있다.

그리드 셀 기반 클러스터링 방법에서 가장 중요한 것은 셀의 크기를 결정하는 것이다. 즉, 셀 관계를 이용한 클러스터링을 위하여 어떻게 전체 공간 영역을 셀로 나누어야 하는 것이다. 여기서는 사용자가 클러스터링을 위한 임계값을 기준으로 하여 셀 크기를 결정한다. 그림 1은 셀 크기 결정 과정을 보여준다.

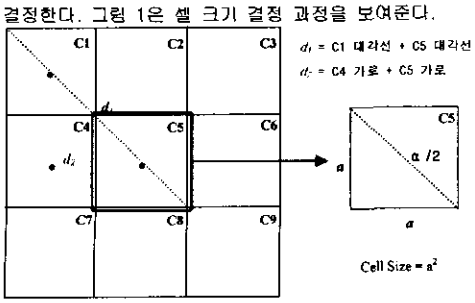


그림 2. 셀 크기 결정

셀 크기를 결정하는데 있어 셀의 대각선의 길이를 임계값/2로 하는 것은 대각선으로 인접한 셀인 경우에, *C1*과 *C5* 셀에 포함된 객체들은 임계 값 내에 있으므로 클러스터로 묶을 수 있다. 이것은 *C1*과 *C5* 셀 간의 가장 먼 거리가 대각선들의 합이므로 셀 내의 모든 객체들은 임계값 내에 있게 된다. 바로 인접한 셀인 경우에, 예를 들어 *C4, C5*인 경우에 가로 길이의 합인 *d2* 값은 *d1*보다 작으므로 이 셀들에 포함된 객체들도 임계값 내에 있으므로 클러스터로 구성할 수 있다.

4. 공간 클러스터링 알고리즘

4.1 셀 관련성 정의

기존의 클러스터링에서 사용되었던 객체들 간의 거리 계산을 인접한 그리드 셀들 간의 관계 연산으로 대체시키기 위하여 다음과 같이 셀 관련성을 정의한다.

정의 1 : Direct Adjacent(Neighboring) Cell (직접인접 셀: DCell)

기준셀에 직접 인접한 셀들로 정의한다. 최대 직접인접 셀들은 8개이다.

정의 2 : Indirect Adjacent(Neighboring) Cell (간접인접 셀: IDCell)

기준셀에 직접 인접하지 않지만, 기준셀과 임계치 내에 부분 포함되어 있어 클러스터링 대상이 되는 셀들로 정의한다.

여기서 간접인접 셀(ID Cell)을 결정하는 방법은 그림 3과 같이 기준셀 내의 객체를 기준으로 하는 것[그림 3(a)]과 기준셀의 4개 꼭지점를 기준으로 정의하는 방법[그림 3(b)]이 있다.

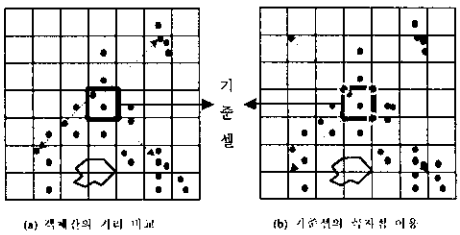


그림 3. 간접인접 셀을 구하는 방법

기준셀 내의 객체를 기준으로 하는 방법은 실제 셀에 포함된 각 객체들에 기준으로 하여 거리를 계산하여야 하기 때문에 시간이 증가하는 단점도 있지만, ID Cell의 수를 줄일 수 있는 장점이 있다. 반면에 기준셀의 꼭지점을 기준으로 하는 방법은 꼭지점에서 임계값 이내의 모든 셀들을 ID Cell로 설정하기 때문에 IDCell의 개수가 증가하는 단점이 있다. 하지만 객체들 간의 거리를 비교없이 ID Cell을 결정하므로 비용이 적게 드는 장점이 있다. 이러한 이유로 인하여 셀 내의 객체를 기준으로 하는 방법보다 꼭지점을 기준으로 하는 방법이 더 낫다고 할 수 있다.

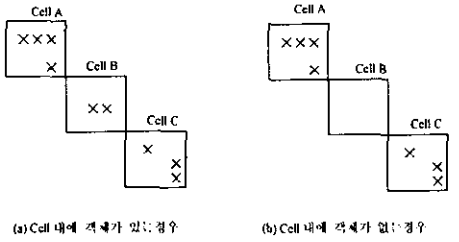
4.2 클러스터링 방법

그리드 셀 기반 공간 클러스터링 방법에서 제일 먼저 수행해야 할 것은 클러스터링을 위한 후보셀들을 결정하는 것이다. 후보셀은 앞에서 정의한 직접인접 셀과 간접인접 셀이 된다. 여기서 직접인접 셀은 임계값을 기준으로 하여 크기를 결정했기 때문에 셀 내의 모든 객체들은 클러스터에 포함된다. 반면에 간접인접 셀은 기준셀과 부분적으로 임계값 내에 있으므로 셀 내의 모든 객체들이 클러스터에

포함되지는 않는다. 따라서 간접인접 셀에 대한 클러스터링 방법의 제시가 필요하다.

간접인접 셀내의 객체들에 대한 클러스터 포함 여부는 기본적으로 기준셀 내의 객체들간의 거리 계산을 통하여 구할 수 있다. 그러나 이 방법은 기준셀 내의 객체들과 간접인접 셀들의 모든 객체들간의 거리 계산을 수반하므로 비용이 많이 든다. 따라서 이러한 문제점을 해결하기 위해서는 간접인접 셀들의 클러스터링에 있어서 비용을 최소화하는 것이 필요하다. 이를 위해서 본 논문에서는 셀들간의 관계를 기반으로 한 전파(propagation) 방법을 제시한다.

간접인접 셀의 유형은 그림 4와 같다. 여기서 Cell A는 기준셀이고, Cell C는 간접인접 셀이다. 그림에서와 같이 그림 4(a)는 기준셀과 간접인접 셀 간에 객체가 포함된 셀이 존재하는 경우이고, 그림 4(b)는 그렇지 않은 경우이다. 여기서 그림 4(a)와 같은 경우는 객체들 간의 거리 계산없이 클러스터링이 가능한 경우이다.



(a) Cell 내에 객체가 있는 경우 (b) Cell 내에 객체가 없는 경우

그림 4. 간접인접 셀 유형

(a)의 경우는 Cell B에 객체가 존재하기 때문에 Cell A와 Cell B가 같은 클러스터에 포함 되고, Cell B와 Cell C도 같은 클러스터에 포함된다. 즉, Cell B가 객체를 포함하고 있으므로 Cell A에서 Cell C로 전파되었고 Cell A, B, C는 동일한 클러스터에 존재하게 된다. 반면에 (b)의 경우는 Cell B에 객체가 존재 하지 않으므로, Cell A의 객체들과 Cell C의 객체들의 거리 계산을 통하여 클러스터에 포함시킬지 여부를 결정한다. 만약 두 셀들 간에 하나의 객체라도 임계값 이내이면 같은 클러스터에 포함시킨다.

그림 5는 공간 클러스터링을 위한 후보셀들을 보여준다. 여기서 후보셀들은 기준셀을 기준으로 직접인접 셀과 간접인접 셀이 된다. 그리고 간접인접 셀은 기준셀의 꼭지점을 이용하여 구한다.

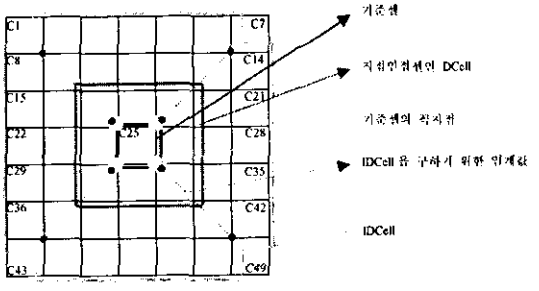


그림 5. 공간 클러스터링을 위한 후보셀

본 논문에서 제시하는 그리드 셀 기반 공간 클러스터링 방법의 세부적인 알고리즘은 다음과 같다.

1. 사용자의 임계값을 기반으로 하여 셀 크기를 결정한다.
2. 결정된 셀 크기를 기준으로 하여 전체 공간 영역을 분할하여, 그리드 셀 구조를 정의한다.
3. 공간 클러스터링을 위한 기준셀을 임의로 결정한다.
4. 기준셀을 기반으로 하여 후보셀을 결정한다. 여기서 후보셀을 위하여
   DCell과 IDCell을 찾는다.
5. DCell에 포함된 모든 객체들을 클러스터에 포함한다.
6. IDCell을 대상으로 하여 전파 방법을 이용하여 해당하는 셀들의 객체들을 클러스터에 포함한다.
7. IDCell 중에서 전파되지 않는 셀들은 기준셀의 객체들과 이 셀들 내의 객체들간의 거리 계산을 통하여 클러스터링을 수행한다.
8. 클러스터링이 종료되면, 후보셀들 중에서 하나의 셀을 기준셀로 선정하여 전체 분할영역이 클러스터링될 때까지 4~8 과정을 반복한다.

5. 결론 및 향후 연구

본 논문에서는 기존 클러스터링의 선 및 다각형 객체 처리의 부적합과 객체들간의 거리계산 방식으로 인한 계산비용이 증가하는 문제점을 극복하기 위하여 객체들 간의 거리를 계산하는 방식을 사용하지 않고 인접한 그리드 셀들간의 관계 연산을 사용하였다. 그래서 방대한양의 공간데이타에 대해 효율적인 클러스터링을 가능하게 하였다.

향후 연구 과제로 클러스터링의 기준이 되는 매개변수 도출에 있어서도 데이터 분포 상태, 밀도 분석을 위하여 반복적인 공간 분할을 이용하는 방법을 제시하고자 한다. 또한 본 논문에서 제시한 공간 클러스터링 방법과 기존 방법과의 성능 평가가 필요하다.

6. 참고 문헌

[1] Ng and J. Han, " Efficient and Effective Clustering Method for Spatial Data Mining" Porc. Of Int. Conf. On VLDB, pp. 144~155, 1994.

[2] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, " A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" Proc. Of the 2nd Int. Conf. On KDD-96, pp. 226~231, 1996.

[3] W. Wang, J. Yang and R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining", Proc. of Int'l Conf. on VLDB, pp. 186-195, 1997.

[4] 오병우, 박기웅, 한기준 " GIS 데이터베이스를 위한 공간 데이타 마이닝" , 한국 정보과학회, 데이터베이스 연구회지 13권 4호 , pp.77~94, 1998.

[5] 윤재관, 오병우, 한기준 " 공간데이타 마이닝을 위한 개방형 객체 관리 시스템의 설계 및 구현 " 개방형 GIS 연구회 논문지 제 1권 제 1호 , pp. 5~18, 1999.

[6] 오병우, 한기준 " H-SCAN : 지식 추출을 위한 해시-기반 공간 클러스터링 알고리즘 " , 한국 정보과학회 논문지 26권 7호, pp. 857~869. 1999.

[7] 진두석, 장재우 " 데이터 마이닝을 위한 대용량 고차원 데이터의 셀~기반 분류방법" , 한국정보과학회 논문지 27권2호, pp. 192~194, 2000.

[8] 장인성, 이기준 " 밀도를 이용한 k-최근접 탐색방법" , 한국정보 과학회 논문지 27호 2권, pp. 80~82, 2000.

[9] 이혜명, 박영배, " 고차원 데이터에서 정진적 프로젝션을 이용한 클러스터링" , 한국정보과학회 논문지 27권 2호, pp. 189~191, 2000.

[10] 김상욱, Aggarwal, Yu, " 고차원 공간에서 최근접 질의를 효과적으로 처리하기 위한 새로운 인덱싱 기법" , 한국정보과학회 논문지 27권 2호, pp. 83~85, 2000.