다양한 공간객체의 데이터 마이닝을 위한 공간 클러스터링 기법의 설계

문상호*, 최진오*, 김진덕**

Design of Spatial Clustering Method for Data Mining of Various Spatial Objects

Sang-Ho Moon*, Jin-Oh Choi*, Jin-Duk Kim**

요 약

공간 데이터 마이닝을 위한 기존의 클러스터링 기법들은 점 객체만을 대상으로 한다. 즉, 선이나 면 같은 다양한 공간 객체들을 지원하지 못한다. 이것은 클러스터링 과정에서 객체들 간의 거리 계산에 있어서, 점 객체는 용이하지만 선과 면인 경우에는 어렵기 때문이다. 본 논문에서는 이러한 문제점을 해결하기 위하여 균등 격자를 이용한 클러스터링 기법을 설계한다. 세부적으로 이 기법에서는 다각형 객체들 간의 거리 계산을 균등 격자를 이용하여 단순화시킴으로서 거리 계산에 따른 시간과 비용을 줄일 수 있다.

ABSTRACT

Existing Clustering Methods for spatial data mining process only point objects, not spatial objects with polygonometry such as lines and areas. It is because that distance computation between objects with polygonometry for clustering is more complex than distance computation between point objects. To solve this problem, we design a clustering method based on regular grid cell structures. In details, it reduces cost and time for distance computation using cell relationships in grid cell structures.

키워드

공간 클러스터링, 공간데이타 마이닝, 균등격자, 다각형

Ⅰ. 서 론

공간 클러스터링은 공간데이타 마이닝 기법 중의 하나로, 공간객체들에 대하여 공간적 특성을 이용하여 집단화하는 과정이다. 이러한 공간 클러스터링은 다른 마이닝 알고리즘의 전처리 단계로 이용되거나 유사성 검색 등의 많은 응용 분야에 널리 사용되고 있다.

공간데이타 마이닝을 위한 클러스터링 기법은 기존에 많이 연구되어 왔다[1][2][3][4]. 그러나 이 연구들의 대부분이 공간객체들 중에서 점 객체만을 대상으로 하고 있다. 이것은 많은 클러스터링 기법들이 거리 계산을 기반으로 하고 있기 때문에,

*부산외국어대학교 컴퓨터공학부
**동의대학교 컴퓨터공학과
접수일자 : 2004. 5. 28

점 객체인 경우에는 거리 계산이 비교적 간단하지만 선(line), 면(area) 등과 같은 다각형 객체인 경우에는 거리 계산이 매우 복잡하기 때문이다.

본 논문에서는 이러한 문제점을 해결하기 위하여 점 뿐만아니라 선, 면 등과 같은 다각형 객체를 모두 처리할 수 있는 클러스터링 기법을 제시한다. 세부적으로 기존에 제시한 균등격자(grid cell)를 이용한 클러스터링 기법을 확장하고자 한다. 기존에 제시한 클러스터링 기법은 점 객체를 대상으로 균등격자 구조를 생성한 후에, 셀 관련성을 이용하여 클러스터링을 수행하였다[5]. 본 논문에서는 이 클러스터링 기법을 다각형 객체에도 적용할 수 있도록 균등격자 구조, 세부 클러스터링 알고리즘 등을 확장하여 설계한다.

II. 균등격자 구조

선, 면 등과 같은 다양한 다각형 객체들을 포함하는 균등격자를 생성하기 위해서는, 먼저 다각형 객체와 셀 간의 관련성을 파악해야 한다.

2.1 셀과 객체의 관계 유형

그림 1은 이 관련성에 대한 유형을 보여준다. 여기서 P1과 L1 객체는 여러 개의 셀들에 걸쳐 있는 반면에, P2와 L2 객체는 한 셀 내에 포함되어 있다. 후자의 경우에는 한 셀 내에 여러 개의 다각형 객체들이 포함될 수도 있다. 이와 같은 경우에는 기존의 점 객체를 대상으로 설계된 균등격자 구조를 그대로 이용할 수 있다. 그러나 전자와 같이 다각형 객체가 여러 셀에 걸쳐 있는 경우가 대부분이므로, 이러한 경우를 처리할 수 있도록 균등격자 구조를 확장하는 것이 필요하다.

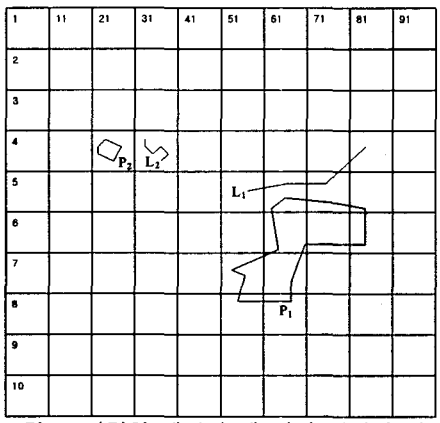

그림 1. 다각형 객체와 셀 간의 관련성 예

균등격자 구조를 설계할 때, 먼저 무엇을 기준으로 할 것인지를 결정해야 한다. 즉, 객체를 기준으로 할 것인지 아니면 셀을 기준으로 할 것인지를 결정해야 한다. 구조적인 측면에서 볼 때는 어떤 것을 기준으로 하더라고 큰 문제는 없다. 실제로 균등격자를 이용한 클러스터링 기법에서는 근본적으로 셀 관련성을 기반으로 한다. 따라서 균등격자 구조의 설계에서는 셀을 기준으로 하는 것이 효율적이다.

2.2 균등격자 구조

본 논문에서 제시하는 균등격자 구조는 전체 공간영역에 대하여 임계값에 따라 나누어진 셀들의 구조를 나타내며, 크게 셀의 기본 정보를 나타내는 셀-헤드와 셀에 속하는 객체들에 대한 정보를 나타내는 셀-디렉토리로 구성된다. 셀-헤드는 셀의 정보를 나타내주는 부분으로서, 전체 셀의 수만큼 저장된다. 내부적으로 셀-디렉토리에서 셀의 시작지점을 나타내는 offset, 셀 내의 객체수, 셀 영역으로 구성된다. 그림 2는 셀-헤드 구조를 나타낸다.

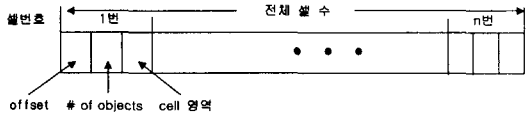

그림 2. 셀-헤드(cell-head) 구조

셀-디렉토리는 셀에 속하는 데이터에 대한 정보를 나타낸다. 세부적으로 셀에 속하는 공간객체들의 식별자들로 구성되며, 실제 객체들의 좌표는 데이터 파일에서 식별자를 이용하여 검색한다. 그림 3은 셀-디렉토리 구조를 나타낸다.

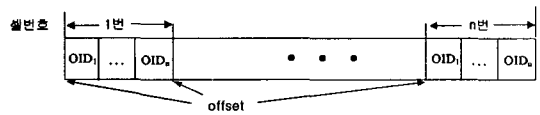

그림 3. 셀-디렉토리(cell-directory) 구조

그림 1에서 C65 셀에 대하여 균등격자 구조를 생성하면 그림 4와 같다. 셀-헤드 구조에서는 셀-디렉토리의 offset, 객체 수(2), 셀 영역을 가지고 있다. 셀-디렉토리 구조에서는 이 셀에 포함되는 2개의 객체인 P1과 L1의 OID 값을 가지며, 이 OID를 이용하여 데이터 파일에서 좌표값을 가져올 수 있다.

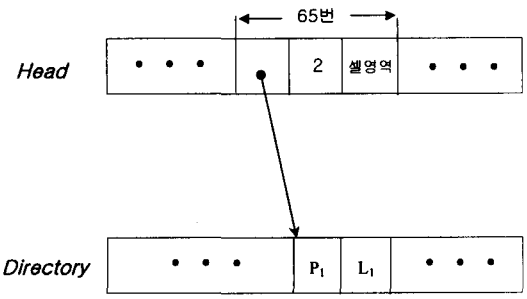

그림 4. 균등격자 구조의 예

여기서 한 가지 주목할 것은 셀을 기준으로 균등격자를 생성하다 보면, 한 객체가 여러 셀들에 중복되어 저장된다는 것이다. 이것은 다각형 객체의 특성상 불가피한 문제지만, 중복 저장에 따른 저장공간의 낭비 문제가 발생할 수도 있다. 그러나 균등격자의 디렉토리 구조에서 저장되는 것은 객체의 좌표값이 아니라, OID값이므로 중복에 따른 공간 낭비는 최소화할 수 있다.

2.3 격자파일 생성 방법

최종적으로 균등격자 구조는 데이터 파일을 기반으로 임계값을 이용하여 격자파일(grid file) 형태로 생성된다. 이 격자파일을 생성하는 과정은 그림 5와 같다. 여기서 InitGridFile() 함수는 임계값을 기준으로 전체영역을 분할하여 셀들로 나누고, Cell_Adjust()는 셀내에 포함되는 객체 와 파일에 저장 위치를 결정하여 격자파일을 생성한다. 여기서 이용되는 구조체(struct)들과 함수들의 세부적인 사항들은 실제 구현상에 구체적인 설계가 이루어진다.

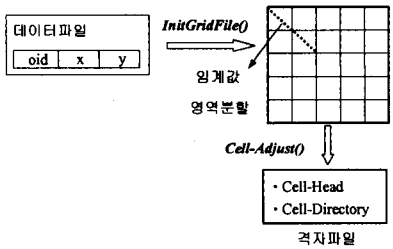

그림 5. 격자파일 생성 과정

III. 공간 클러스터링 기법

본 논문에서 제시하는 클러스터링 기법은 기존의 공간 클러스터링 기법에서 제시된 클러스터 생성 알고리즘과 클러스터 합병 알고리즘을 확장하여 설계한다.

3.1 클러스터 생성 알고리즘

클러스터링 생성 알고리즘은 기본적으로 균등격자 구조에서 셀 관련성만으로 후보 클러스터를 생성한다. 따라서 이 알고리즘은 확장할 필요가 없이 그대로 적용이 가능하다. 다만, 선, 면 등과 같은 다각형 객체를 대상으로 생성된 균등격자 구조를 적용시킨다는 점만 달라질 뿐이다. 클러스터 생성 알고리즘은 그림 6과 같다.

단계 1) 격자구조에서 객체들이 존재하는 셀들을 검색
단계 2) 기준셀(C_i)을 정하고, 직접인접셀($DAC(C_i)$)을 검색
단계 3) 전파가능한 셀들을 찾기 위하여 2단계에서 검색된 셀들($C_j \in DAC(C_i)$)에 대하여 직접인접셀($DAC(C_j)$)을 검색
단계 4) 셀들이 검색될 때까지 2, 3단계를 반복 수행
단계 5) 검색된 셀들을 묶어서 하나의 클러스터를 생성
단계 6) 전체셀에 대해 2~5단계를 반복 수행하여 후보클러스터 찾음

그림 6. 클러스터 생성 알고리즘

3.2 클러스터 합병 알고리즘

클러스터 합병 알고리즘에서는 전단계에서 후보 클러스터들이 셀 관련성만을 이용하여 생성되었으므로, 클러스터들 간의 합병 여부를 확인한다. 이 과정에서 객체들 간의 거리 계산을 최소화하기 위하여 클러스터링 가능셀/범위 정의를 이용한다. 그림 7은 클러스터 합병 알고리즘을 보여준다.

단계 1) 후보 클러스터들에 대하여 클러스터링 가능셀 범위(*ECCA*)를 구함
단계 2) 셀 범위내에서 클러스터링 가능셀(*ECC*)의 존재 여부 확인
단계 3) *ECC*가 존재하면, *ECC* 내부의 객체들간에 거리 계산
단계 4) 객체 거리가 임계값 이내에 있으면 후보 클러스터들을 합병
단계 5) 후보 클러스터들에 대하여 1~4 단계를 반복 수행

그림 7. 클러스터 합병 알고리즘

클러스터 합병 알고리즘도 생성 알고리즘과 마찬가지로 바로 적용이 가능하다. 다만, 단계 4에서 클러스터링 가능셀에 포함된 객체들 간의 거리 계산 부분은 변경해야 한다. 즉, 기존의 합병 알고리즘에서는 점 객체만을 대상으로 하므로 쉽게 거리 계산이 가능하지만, 다각형 객체를 대상으로 하는 경우는 거리 계산이 복잡해진다.

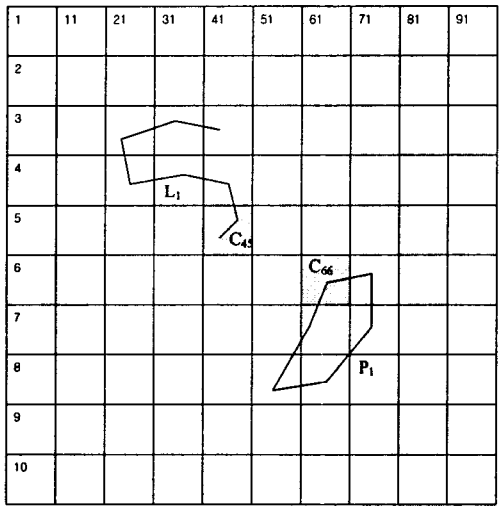

그림 8. 클러스터링 가능셀 내의 객체 거리 계산

예를 들어, 그림 8에서 P1과 L1 객체가 전단계에서 생성된 다른 후보 클러스터에 각각 속해 있고, C45과 C66은 클러스터링 가능셀이다. 그러면 클러스터 합병 알고리즘에서는 이 셀들에 포함된 P1과 L1 객체간의 거리를 계산해야 한다. 그러나 여기서 고려할 사항은 두 객체의 전체 기하(geometry)를 대상으로 거리 계산을 할 필요는 없다. 만약 두 객체의 전체 기하를 대상으로 거리 계산을 하더라도 문제는 없지만, 계산 비용과 시간이 많이 들어 비효율적이기 때문이다. 따라서 본 논문에서는 객체들 간의 거리 계산을 할 때, 클러스터링 가능셀 영역에 포함된 기하를 추출해서, 이 기하만을 대상으로 거리 계산을 수행하게 한다. 이렇게 함으로써 거리 계산에 드는 비용과 시간을 줄일 수 있다.

3.3 출력파일 구조

출력파일은 격자파일과 데이터파일을 기반으로 클러스터링 알고리즘을 적용하여 생성된 클러스터를 저장하기 위한 것이다. 내부적으로 클러스터를 구별하는 플래그(flag)와 각 클러스터는 포함되는 셀 들의 번호를 저장하며, 이 셀 번호를 이용하여 실제 클러스터에 포함되는 객체들을 검색한다. 출력 파일의 구조는 그림 9와 같다.

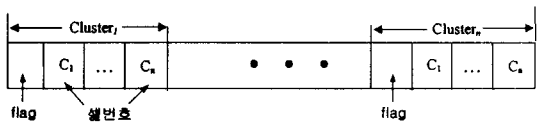

그림 9. 출력(output) 파일 구조

IV. 결 론

본 논문에서는 기존에 제시한 균등격자를 이용한 클러스터링 기법을 점 뿐만아니라 선, 면 등과 같은 다각형 객체를 대상으로 클러스터링이 가능하도록 확장하였다. 세부적으로 다각형 객체를 저장하기 위한 균등격자 구조를 설계하였고, 이 균등격자 구조를 기반으로 클러스터링할 수 있도록 클러스터 생성 및 합병 알고리즘을 확장하였다. 향후 연구로는 설계된 균등격자 구조와 클러스터링 알고리즘을 구현하여 실제 데이터를 이용하여 클러스터링 결과를 시현할 계획이다.

참고문헌

[1] Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Data Mining", 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", Int. Conf. on KDD, pp.226~231, 1996.

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

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

[5] 문상호, 이동규, 서영덕, "공간데이타 마이닝을 위한 효율적인 그리드 셀 기반 공간 클러스터링 알고리즘", 정보처리학회논문지, 10-D권, 4호, 2003.

문상호(Sang-Ho Moon)

한국기계연구원 정보지원실 연구원
부산대학교 컴퓨터공학과 공학석사부산대학교 컴퓨터공학과 공학박사위덕대학교 컴퓨터공학부 조교수
부산외국어대학교 컴퓨터공학부 조교수
※관심분야 : GIS, 공간DB, 데이터마이닝, GIS표준, 정보시스템 감리

최진오(Jin-Oh Choi)

부산대학교 컴퓨터공학과 공학석사
부산대학교 컴퓨터공학과 공학박사
경동대학교 정보통신공학부 전임강사
부산외국어대학교 컴퓨터공학부 조교수
※관심분야 : 공학데이타베이스, 지리정보시스템, 모바일GIS

김진덕(Jin-Duk Kim)

부산대학교 컴퓨터공학과 공학석사
부산대학교 컴퓨터공학과 공학박사
부산정보대학 정보통신계열 전임강사
동의대학교 컴퓨터공학과 조교수
※관심분야 : 객체지향DB, GIS, 공간질의, 공간색인, 이동체데이타베이스, LBS, 텔레매틱스