□ 기술해설 □

공간 데이타 마이닝에 관한 고찰†

건국대학교 오병우·이강준·한기준*

1. 서 론

최근에는 정보 기술의 발달로 데이타베이스에서 관리하는 데이타의 양이 급격히 증가하여 사람의 능력으로는 분석할 수 없는 상태에 이르렀다. 이렇게 폭발적으로 늘어나는 데이타로부터 컴퓨터를 사용하여 지식 또는 정보를 자동적으로 발견하고자 하는 필요성이 증대되고 있다. 그리하여, 데이타 마이닝(Data Mining) 또는 데이타베이스에서의 지식 발견(KDD : Knowledge Discovery in Database)이라 불리는 분야가 대두되었다. 데이타 마이닝은 주로 데이타베이스 연구 분야에서 사용되고 KDD는 주로 인공지능 연구 분야에서 사용되는 용어로서 본 논문에서는 데이타 마이닝이라는 용어를 사용한다. 이들은 모두 대용량 데이타베이스로부터 이전에 알려지지 않았던 암시적인 유용한 지식을 자동적으로 발견하는 것으로 정의될 수 있다 [7].

데이타 마이닝은 인공지능, 데이타베이스, 통계학 등이 통합된 분야로 초기에는 주로 관계형 및 트랜잭션 데이타베이스에서 데이타 마이닝에 대한 연구가 활발히 진행되었다[5, 7, 12]. 특히, 바코드를 사용하는 상점에서의 상품 매매 경우와 같이 빈번한 트랜잭션을 처리하는 트랜잭션 데이타베이스에서 유용한 지식을 발견하기 위한 데이타 마이닝은 상업적인 목적에서 실제로 적용되고 있다. 최근에는 관계형 및 트랜잭션 데이타베이스뿐만 아니라 공간 데이타베이스, 시간 데이타베이스, 객체-지향 데이타베이스, 멀티미디어 데이타베이스 등과 같은 다른 응용 가능한 데이타베이스들에서도 데이타 마이닝에 대한 필요성이 증가하고 있다[1, 11, 16, 18, 22].

공간 데이타는 특정한 공간에 위치하는 객체와 관련된 데이타이다. 공간 데이타베이스는 공간 데이타 타입과 공간 객체들간의 위상 관계에 의해 표현되는 공간 데이타를 저장한다. 공간 데이타는 위치 및 위상 정보를 포함하고, 일반적으로 신속한 검색을 위해 공간 인덱스 구조를 갖고 있으며, 공간 접근 방법(SAM : Spatial Access Method)에 의해 접근된다. 공간 데이타 마이닝은 공간 데이타베이스로부터 암시적인 지식, 공간 관계, 또는 공간 데이타베이스에 명시적으로 저장되지 않은 규칙들을 추출하는 것을 의미한다[15]. 즉, 기존의 공간 데이타베이스 시스템은 단순히 데이타의 저장 및 관리를 지원하기 위한 것인 반면에, 공간 데이타 마이닝 시스템은 공간 데이타베이스에 저장 및 관리되는 데이타로부터 사용자에게 원하는 종류의 지식을 자동화된 분석 과정을 거쳐 제공하는 것이다.

기계 학습, 데이타베이스 시스템, 그리고 통계학 분야 등에서의 기존 연구들은 데이타베이스에서 지식 발견을 위한 연구에 기초를 제공하고, 공간 데이타베이스 분야에서 공간 데이타 구조, 공간 추론, 계산적 기하 등에 관한 연구도 공간 데이타 마이닝의 연구를 위한 기반을 제공한다. 공간 데이타의 경우에는 일반적으로 관계형 데이타베이스 시스템에서 처리되는 데이타보다 대용량이고 복잡하므로 공간 데

† 본 과제는 한국과학재단 '97특정기초연구(KOSEF 97-01-02-04-01-3)와 정보통신부 '98대학기초연구(과제번호 : C1-98-5144-00) 과제로부터 부분적으로 지원받았음.
* 종신회원

이타 마이닝에서는 보다 효율적인 알고리즘이 절실히 요구된다. 본 논문에서는 공간 데이타 마이닝에 대해 전반적으로 고찰한다.

본 논문의 구성은 다음과 같다. 제2장은 기존의 공간 데이타 마이닝 시스템에 대해 언급한다. 제3장에서는 공간 데이타베이스로부터 지식을 발견하기 위해 사용되는 알고리즘에 대해 고찰한다. 제4장에서는 공간 데이타 마이닝과 관련된 인터넷상의 URL을 나열한다. 마지막으로, 제5장에서는 고찰에 대한 결론과 앞으로의 공간 데이타 마이닝 연구를 위한 향후 연구 분야를 제시한다.

2. 공간 데이타 마이닝 시스템

2.1 GeoMiner[9]

GeoMiner 시스템은 캐나다의 Simon Fraser 대학에서 관계형 데이타 마이닝 시스템인 DBMiner를 확장하여 공간 데이타를 처리할 수 있도록 개발하고 있는 시스템이다. GeoMiner의 기본이 되는 DBMiner는 데이타 마이닝과 데이타 웨어하우스 기술의 결합으로서 개발되었고, 관계형 데이타 마이닝을 위한 데이타 큐브 구축과 처리, 애트리뷰트-지향 유도, 다중-레벨 조합 분석, 통계적 데이타 분석, 기계 학습 등을 포함하고 있다. GeoMiner는 질의어로서 GMQL(Geo-Mining Query Language)을 제공하고 데이타 마이닝 결과를 테이블, 차트, 지도 등의 형태로 출력하기 위한 대화식 및 그래픽 사용자 인터페이스도 지원한다. 또한, 공간 데이타의 처리를 위해 MapInfo Professional 4.1 GIS를 사용하며, 비공간 데이타, 공간 데이타, 개념 계층을 저장하는 데이타베이스를 관리한다.

GeoMiner에서 비공간 데이타 마이닝은 DBMiner를 직접 사용하고, 공간 데이타 마이닝과 공간 및 비공간 데이타간의 마이닝을 위해서는 5가지 기능 모듈을 사용한다. 지리-특성화(Geo-characterize) 모듈은 개념 계층을 통한 추상화를 사용하여 다중-레벨 및 다중-관점으로 데이타를 분석할 수 있도록 하고, 지리-비교(Geo-comparator) 모듈은 데이타 그룹들간의 대조되는 특성을 사용하여 비교할 수 있도록 하며, 지리-관련(Geo-associator) 모듈은 데이타 그룹들간에 "X→Y(s%, c%)" 형태의 관계를 얻을 수 있도록 한다. 지리-클러스터 분석(Geo-cluster analyzer) 모듈은 CLARANS 알고리즘을 사용하여 공간 클러스터링을 수행한 후 애트리뷰트-유도를 사용해 클러스터의 비공간 특성을 발견하도록 하고, 지리-클래스화(Geo-classifier) 모듈은 일반화-기반 의사-트리 유노 방법을 사용하여 의사 트리를 생성하고 이를 근거로 클래스를 나눌 수 있도록 한다. 그리고, 시간 관련 데이타 마이닝을 포함한 서너개의 추가 데이타 마이닝 모듈들이 현재 연구 개발 중에 있다.

2.2 ASK-ME[20]

ASK-ME(Aspatial, Spatial, and Knowledge Data Mining Engine)는 건국대학교에서 개발하고 있는 공간 데이타 마이닝 시스템으로서 객체 지향 데이타베이스, Geographic Information System(GIS), 비공간 및 공간 클러스터링, 그래픽 사용자 인터페이스, 그리고 데이타 마이닝 기술을 결합하여 개발하고 있다. ASK-ME는 크게 사용자 인터페이스 계층, 공간 데이타 마이닝 계층, 그리고 데이타 관리 계층으로 구성된다.

사용자 인터페이스 계층은 클래스 정의 및 삭제, 객체 삽입 및 검색, 비공간·공간·지식 데이타 표현 등을 지원하며 OGC에서 제시한 표준[21]을 통해 데이타 관리 계층 및 공간 데이타 마이닝 계층에 접속한다.

공간 데이타 마이닝 계층은 공간 데이타베이스로부터 실제로 지식을 추출하는 가장 중요한 계층으로서 데이타의 분포 분석, 클러스터 생성·통합·분할, 공간 유사도 추출, 다양한 지식 추출 기능 등을 제공한다.

공간 데이타 마이닝을 위한 데이타 관리 계층[25]은 공간 데이타를 효율적으로 저장하고 관리하기 위하여 미국 Texas Instrument사에서 개발한 Open OODB[6]를 확장하여 추가로 공간 데이타 처리, 공간 인덱스, SDTS (Standard Data Transfer Standard) 수입/수출 기능 등을 제공한다.

ASK-ME는 효율적인 공간 데이타 마이닝을 위하여 해싱에 근거한 클러스터링 방법을 제공한다. 해싱은 일반적으로 다른 방법보다 효율적인 검색을 제공하지만 삽입 및 갱신시에 매우 복잡한 처리가 요구된다. 그러나, 공간 데이타 마이닝에서는 일반적으로 갱신은 적고 검색이 매우 빈번하므로 해싱을 사용하는 방법이 타당하다. ASK-ME에서는 공간 데이타로서 레이블, 점, 선, 면을 제공한다. 레이블과 점의 경우에는 해당 셀을 찾기 위한 해싱 함수가 간단하지만, 선과 면의 경우에는 해싱을 위해 부가적인 기술이 필요하다. ASK-ME에서는 객체간의 거리를 구하기 위하여 공간 유사 함수를 제공하는데, 선과 면의 해싱을 위해서도 공간 유사 함수를 사용한다.

ASK-ME는 분포, 클러스터링, 일반화, 비교, 그룹화, 특성화, 종속 등의 지식을 추출할 수 있다. 지식을 추출할 때는 데이타의 분포 분석 및 클러스터링을 통해 데이타 자체를 분석하여 지식을 추출하는 방법을 사용한다. 데이타 자체를 분석하면 데이타에 충실한 지식이 추출되는 장점을 갖지만, 경우에 따라서는 사용자의 요구에 따라 분석을 수행하여야 할 때도 있다. 이를 위하여 지식을 추출할 때 사용자가 정의한 개념 계층(concept hierarchy)[10]을 사용하는 기능도 제공한다.

2.3 기타 시스템

SKICAT(The Sky Image Cataloging and Analysis)[4]는 미국 NASA의 Jet Propulsion Laboratory에서 개발되었고, 관측 기구에 의해 입력된 이미지 데이타를 관리하고 우주 객체를 분류하기 위한 지능형 학습 가능 이미지 분석 도구이다. SKICAT은 관측 기구로부터 입력된 이미지를 이미지 프로세싱을 통해 값을 측정하고 의사 결정 트리를 사용해 클래스를 찾는다. 그리고, 전문가가 습득된 데이타를 분석하여 학습 알고리즘을 통해 의사 결정 트리에 feed-back한다.

WEKA(The Waikato Environment for Knowledge Analysis)[8]는 뉴질랜드의 Waikato 대학에서 개발되었고, 실세계의 데이타 집합, 특히 뉴질랜드의 농업 부문의 데이타 집합에 대해 기계 학습 기술을 응용하기 위한 시스템이다. WEKA는 공통 프레임워크와 단일 사용자 인터페이스로서 다른 기계 학습 도구들을 통합한 것이다. 즉, WEKA 사용자 인터페이스, 간단한 규칙을 위한 프로그램, 복잡한 규칙을 위한 유도(induct) 프로그램, 인스턴스-기반 학습을 위한 프로그램, 회귀 모델 트리를 위한 프로그램, 규칙 평가 프로그램 등으로 구성된다.

3. 공간 데이타 마이닝 알고리즘

3.1 일반화-기반 지식 발견

기계 학습에서 광범위하게 사용되는 튜플-지향 알고리즘은 샘플의 개수에 지수적이고, 잡음이나 불일치성 데이타를 잘 처리하지 못하므로 대용량 공간 데이타베이스를 위해 그대로 적용될 수 없다. 그리하여, 기계 학습 알고리즘에서의 튜플-지향과 반대되는 애트리뷰트-지향 알고리즘이 제안되었고, 그 후에 이 기술은 공간 데이타베이스에서 사용될 수 있도록 확장되었다.

3.1.1 애트리뷰트-지향 유도[10]

일반화-기반 지식 발견은 개념 계층 형태의 배경 지식을 필요로 한다. 공간 데이타베이스의 경우에는 비공간 및 공간 두 가지 종류의 개념 계층이 존재한다. 개념 계층들은 전문가에 의해 명시적으로 주어질 수 있고, 또는 데이타 분석에 의해 자동으로 생성될 수도 있다. 예를 들면, 일반화 과정에서 "동"들이 모여 조금 더 큰 지역인 "구"를 이룰 수 있다. 애트리뷰트-지향 유도는 일반화 계층을 따라 올라가고, 또한 더 높은 개념 레벨에서 공간 및 비공간 데이타를 요약하면서 수행된다. 애트리뷰트-지향 유도는 각각의 애트리뷰트가 원하는 레벨로 일반화될 때까지 다음의 수행 과정을 반복한다.

① 개념 계층의 상위 레벨로 올라가면서 튜플의 애트리뷰트 값들을 일반화한다.

② 더이상 개념 계층을 통한 일반화가 불가능하고 특정 애트리뷰트에 매우 다양한 값이

존재할 때 그 애트리뷰트를 제거하여 일반화한다.

③동일한 튜플들끼리 통합한다. 이때 추가적으로 "count" 애트리뷰트를 두어 증가시킨다. 이러한 "count"는 얻어진 지식의 정량적인 표현에 사용된다.

3.1.2 공간-데이타-우위 일반화[17]

이 방법은 공간 데이타에 대해 먼저 일반화한 후에 비공간 데이타를 분석하는 방법으로서, 사용자 질의와 관련된 모든 데이타를 수집하는 것으로 시작한다. 그리고, 수집된 데이타 중 공간 데이타에 대해 개념 계층에 근거하여 공간 지역들을 통합한다(예를 들면, "동"들이 모여 "구"를 이룬다). 공간 데이타에 대한 일반화는 사용자가 지정한 공간 일반화 임계치에 도달할 때까지 계속되며, 그 위에 애트리뷰트-지향 유도 기법을 사용하여 비공간 데이타가 검색되어 분석된다. 질의의 결과는 일반화된 지역 각각에 대한 특성들의 "or 결합"으로 나타난다. 예를 들면, "광진구의 인구밀도는 매우 높다(40%) 또는 높다(60%)"로 나타낼 수 있다.

3.1.3 비공간-데이타-우위 일반화[17]

이 방법은 비공간 데이타에 대해 먼저 일반화한 후에 공간 데이타를 분석하는 방법으로서, 공간-데이타-우위 일반화와 같이 첫번째 단계에서 사용자 질의와 관련된 모든 데이타를 수집하는 것으로 시작한다. 두번째 단계에서는 비공간 애트리뷰트에 대해 더 높은(더 일반적인) 개념 레벨로 일반화하며 애트리뷰트-지향 유도를 수행한다. 일반화 임계치는 일반화 과정을 계속할지 또는 멈출지를 결정하는데 사용된다. 이 단계에서 공간 객체에 대한 포인터들이 집합으로 수집되어 일반화된 비공간 데이타와 합쳐진다. 알고리즘의 세번째와 마지막 단계에서 일반화된 동일한 애트리뷰트들을 가진 인접하는 영역들은 "adjacent_to" 공간 함수에 근거해 서로 통합된다. 잡음 처리를 수행하기 위해서는 근사치를 사용하여 다른 비공간 데이타를 갖는 작은 지역들을 무시한다. 예를 들어, 만약 대부분의 영역이 공장이고 서너개의 상점이 그 지역에 존재한다면 전체 영역은 공장지역으로 일반화된다. 질의의 결과는 상위 레벨 기술에 따른 서너개의 작은 지역들을 갖는 지도의 형태로써 표현될 수 있다.

이러한 일반화에 근거한 알고리즘은 모두 개념 계층에 근거를 두고 있다. 개념 계층은 제한된 분야에 관해서는 전문가가 일일이 기술하여 높은 효율성을 보장하지만 광범위한 분야에서 각 애트리뷰트마다 개념 계층을 생성하고 이를 유지·관리한다는 것은 매우 힘든 일이다. 데이타 마이닝의 결과는 개념 계층의 정의에 의해 크게 좌우된다. 또한, 개념 계층이 데이타의 분포에 따라 자동적으로 생성되기도 하는데, 이때 생성된 개념 계층들에 대한 적합성 판명은 여전히 사용자(특히, 전문가)가 해야 한다. 개념 계층에 근거한 기존의 방법들은 비공간 데이타에 적용되던 개념 계층을 공간 애트리뷰트로 확장하여 사용하므로 공간 데이타의 모델링이 미흡하고 공간 데이타만의 특성을 제대로 처리하지 못하는 단점이 있다.

3.2 클러스터링 방법

클러스터 분석은 수년동안 중점적으로 연구된 통계학의 한 지류로서 개념 계층과 같은 배경 지식을 사용하지 않고 데이타로부터 직접 클러스터를 발견할 수 있다는 장점을 갖는다.

3.2.1 PAM 및 CLARA[13]

PAM(Partition Around Medoids) 알고리즘은 n개의 객체들이 있다고 가정하고, 각 클러스터를 위한 대표 객체를 먼저 발견하여 k 클러스터들을 찾는다. 클러스터에서 가장 중앙에 위치한 점인 대표 객체는 "medoid"라 불린다. k medoid들을 선정한 후에, 이 알고리즘은 medoid와 일반 다른 객체들로 이루어진 모든 가능한 객체들의 쌍을 분석하면서 반복적으로 최선의 medoid들을 선정하기 위해 시도한다. 각 쌍의 결합에 대해 클러스터링 정도가 계산되고, 한 단계에서 최선의 점들의 선택은 다음 단계를 위한 medoid들로 선정된다. PAM 알고리즘은 n과 k의 값이 크면 상당히 비효율적이다.

PAM 알고리즘은 모든 데이타를 검색하는

반면, CLARA(Clustering LARge Applications) 알고리즘은 샘플링에 기초하여 실제 데이타의 작은 부분만을 사용하여 medoid들을 선정한다. 이 방법은 만약 샘플을 매우 적절하게 무작위로 추출할 수 있다면 무작위로 추출된 샘플들로부터 선정된 medoid들과 전체 객체들로부터 선정된 medoid들이 유사할 것이라는 생각에 근거를 두고 있다. CLARA는 다중 샘플들을 뽑고 이러한 샘플로부터 최선의 클러스터링을 산출한다. 그래서, CLARA는 PAM보다 큰 데이타 집합을 처리할 수 있다.

3.2.2 CLARANS[19] 및 Focusing[2]

CLARANS(Clustering Large Applications based upon RANdomized Search) 알고리즘은 샘플을 고정하지 않고 데이타 집합의 부분집합만을 검색하여 PAM과 CLARA를 결합한다. CLARA가 검색의 모든 단계에서 고정된 샘플만을 갖는 반면에 CLARANS는 검색의 각 단계에서 특정한 무작위성으로 샘플을 뽑는다. 클러스터링 과정은 모든 노드가 잠재적인 해(즉, k medoid들의 집합)인 그래프를 검색하는 것으로 생각할 수 있다. 하나의 medoid를 대체한 후에 얻어진 클러스터링은 현재 클러스터링의 이웃(neighbor)이라고 불린다. 무작위로 시도되는 이웃들의 개수는 매개변수 maxneighbor에 의해 제한된다. 만약 더 좋은 이웃이 발견되면 CLARANS는 그 이웃 노드로 옮겨가서 과정을 다시 시작한다. 만약 발견되지 않는다면 현재의 클러스터링이 지역적 최적(local optimum)을 산출한다. 만약 지역적 최적이 발견되면 CLARANS는 새로운 지역적 최적의 검색을 위해 새롭게 무작위로 선택된 노드로부터 출발한다. 검색된 지역적 최적들의 개수는 매개변수 numlocal에 의해 제한된다.

CLARANS 알고리즘은 PAM과 CLARA 둘보다는 효율적이라고 평가되지만, 아직 효율성이 만족할만한 수준은 아니고 모든 객체가 메모리에 적재되어야 한다는 단점을 갖고 있어서 샘플링의 질을 높이는 Focusing 방법이 제안되었다. Focusing 방법은 공간 데이타 마이닝 과정의 전처리에 해당하는 단계로서 GIS에서 공간 인덱스로 많이 사용되는 R*-tree의 리프 노드에서 제일 중심에 위치한 객체들만을 클러스터링하여 샘플링에 의한 클러스터링의 손실을 줄인다.

3.2.3 SD 및 NSD[17]

공간 데이타 마이닝에서는 CLARANS에 근거하여 공간-우위 접근 방법(SD:Spatial Dominant)인 SD(CLARANS) 알고리즘과 비공간-우위 접근 방법(NSD : Non-Spatial Dominant)인 NSD(CLARANS) 알고리즘이 개발되었다. SD(CLARANS) 알고리즘은 공간 우위 접근 방법으로서 먼저 관련있는 데이타 집합의 공간 애트리뷰트에 대해 CLARANS를 사용하여 클러스터링한 후에 각 클러스터에 속한 객체들의 비공간 애트리뷰트에 대해 애트리뷰트-지향 유도를 수행한다. 질의의 결과는 각 클러스터에 속한 객체들의 상위 레벨에서 일반화된 비공간 특성을 표현한다.

NSD(CLARANS) 알고리즘은 비공간 우위 접근 방법으로서 먼저 비공간 일반화를 수행한 후에 클러스터링하는 방법이다. 즉, 비공간 애트리뷰트들에 대한 애트리뷰트-지향 일반화를 통해 일반화된 튜플들을 생성한 후에, 각각의 일반화된 튜플의 공간 애트리뷰트에 대해 CLARANS를 사용하여 클러스터링한다.

3.2.4 BIRCH[24]

대용량의 데이타 집합을 위한 클러스터링 알고리즘인 BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)에서는 CF(Clustering Feature)와 CF tree를 사용한다. CF는 전체 점들을 모두 저장하는 대신 서브클러스터에 대한 정보를 요약한 것으로 CF=(N, $\overline{LS}$, SS)로 정의된다. N은 서브클러스터에 있는 점들의 개수이고, $\overline{LS}$는 $\sum \overline{X}$으로서 N점들의 선형 합계이다. 그리고, SS는 $\sum \overline{X}^2$으로서 데이타 점들의 제곱 합계이다. CF는 클러스터를 계산하기 위한 효율적인 정보들을 포함한다.

CF tree는 CF들을 저장하기 위해 사용되는 균형 트리로서 B(branching factor)와 T(threshold) 값을 갖는다. B는 자식 노드들의 최대 개수이고, T는 리프 노드에 저장된 서브

클러스터의 최대 반지름으로 사용 가능한 메모리에 따라 T값이 조정될 수 있다. CF tree의 중간 노드는 자식 노드의 CF들에 대한 합을 저장하고 있어서 자식 노드에 대한 정보를 요약하고 있다. CF tree는 새로운 데이타의 삽입에 대해 점진 수정 방법을 사용하여 동적으로 구축된다. BIRCH 알고리즘은 데이타 점들을 한번만 읽어서 트리를 구성할 수 있다는 장점이 있다.

3.2.5 DBSCAN[3]

공간 데이타 마이닝을 위한 밀도에 근거한 클러스터링 알고리즘인 DBSCAN은 클러스터의 밀도를 Eps와 MinPts로 조절한다. MinPts는 계산의 복잡성을 줄이기 위해 4로 고정하여 사용된다. Eps를 결정하기 위해서는 먼저 임의의 점과 k(k=4)번째 가까운 점들 사이의 거리를 계산하고, 계산된 거리에 따라 모든 점들을 정렬한다. 그리고, k-dist 그래프를 그려 사용자가 첫 "valley"를 찾으면 그 거리를 Eps로 정하는데, 이는 클러스터링의 결과에 매우 중요한 영향을 준다. DBSCAN의 특징은 저밀도 지역을 이루는 잡음을 클러스터로부터 제거할 수 있다는 것과 R*-tree를 사용해 좋은 효율성을 얻을 수 있다는 것이다.

3.2.6 STING[23]

STING(STatistical INformation Grid) 알고리즘은 공간 데이타 마이닝을 위한 통계 정보 그리드-기반 접근 방법이다. STING에서는 그리드 셀 계층구조(Grid Cell Hierarchy)를 사용해 공간을 다중 레벨의 사각형 셀들로 나눈다. 1번째 레이어에서는 오직 하나의 셀만을 가지고 있고 i번째 레이어에서는(i-1)번째 레이어의 4배에 해당하는 셀들을 가진다. 각각의 셀들은 다음과 같은 6개의 매개변수들을 갖는다. n은 셀에 속한 점들의 개수이고, m은 셀에 속한 모든 값들의 평균이며, s는 애트리뷰트 값들의 표준편차이다. min과 max는 각각 애트리뷰트의 최소값 및 최대값이며, distribution은 "normal", "uniform", "exponential", "NONE" 등의 값을 갖는다. STING은 특정 질의와는 상관없이 각각의 그리드 셀에 속한 데이타에 대한 요약 정보를 갖고 있어서 다양한 질의에 응답할 수 있고, 점진적인 수정을 지원하며, 또한 계산이 간단해 높은 효율성을 제공한다는 특징을 갖는다.

3.3 기타 알고리즘

본 절에서는 공간 연관 지식을 탐색하는 방법과 근사치를 사용하는 다중 필터링 방법에 대해서 언급한다.

3.3.1 공간 연관 지식 탐색[15]

연관 지식은 트랜잭션 데이타 마이닝에서 "basket analysis"를 위해 주로 사용되었던 지식으로서 이를 공간 데이타 마이닝에서도 사용할 수 있도록 확장된 것이 공간 연관 지식이다. 공간 연관 지식은 adjacent-to, near-by, inside, close-to, intersecting 등과 같은 프레디키트를 포함하는데, 예를 들면 "is-a(x, gas-station)→close-to(x, highway).(75%)"가 공간 연관 지식에 속한다. 또한, 빈번하게 나타나고 대부분의 경우를 만족하는 패턴에 중점을 두기 위하여 minimum support와 minimum confidence 개념이 사용된다. 공간 객체들의 집합 S에서 패턴 A의 support는 S의 멤버가 패턴 A를 만족할 확률이고, A→B의 confidence는 패턴 A가 나타날 때 패턴 B가 나타날 확률이다.

공간 연관 지식 탐색에 있어 공간 계산의 비용을 줄이기 위하여 두 단계 공간 계산 기법, 일반화된 프레디키트 기법 등을 사용하는 최적화 방법이 사용된다. 두 단계 공간 계산 기법은 MBR을 사용하여 큰 집합에 대한 근사치 계산을 수행한 후에 적은 집합에 대한 정제 계산을 수행하는 방법이고, 일반화된 프레디키트 기법은 "일반화된 close_to(g_close_to)"와 같이 다중 개념 레벨들을 사용하여 지식을 유도하는 방법이다.

3.3.2 CRH[14]

CRH(Encompassing Circle, Isothetic Rectangle, and Convex Hull) 알고리즘은 두가지 문제를 해결하기 위한 알고리즘이다. 첫번째는 주어진 점들의 클러스터에 대해 클러스터

내의 점들의 대다수와 가까운 형상을 효율적으로 찾는 것이다. 이를 해결하기 위해서는 기하학적인 근사치 방법(즉, circles, rectangles, and convex hulls)을 사용한다. 두번째는 주어진 n개의 클러스터에 대해 n개의 클러스터 대부분이 적용되는 aggregate proximity 공통성들을 추출하는 것이다. 이를 위하여 다른 클러스터에는 없는 다수의 의미있는 공통성들을 효과적으로 찾을 수 있도록 개념 일반화를 사용하는 GenCom 알고리즘을 개발하였다.

CRH 알고리즘은 점차적으로 후보 형상들을 줄여가면서 클러스터와 가까운 형상을 찾기 위한 방법으로서, 먼저 원에 의한 근사치를 사용하여 클러스터와 멀리 떨어진 형상들을 제거하고, 그 뒤에 사각형에 의한 근사치를 사용하여 추가로 형상들을 제거한다. 그리고, 클러스터 내의 점들과 각 형상의 경계와의 거리를 계산하고 총계를 계산하여 최소, 최대, 평균 거리, 그리고 명시된 임계치보다 적은 거리에 존재하는 점들의 백분율을 반환한다. CRH 알고리즘은 50,000 형상들을 처리하는 실험에서 2초 미만의 응답 시간을 갖는 것으로 보고되었다.

4. 관련 URL

본 장에서는 공간 데이타 마이닝과 관련된 인터넷상에서의 중요한 URL들을 나열한다.

• http://db.cs.sfu.ca/GeoMiner

-공간 데이타 마이닝 시스템인 GeoMiner에 관한 홈페이지로서 주요 기능 모듈들에 대한 설명, GeoMiner의 snapshot, 공간 데이타 웨어하우스 및 마이닝에 대한 튜토리얼 자료 등이 수록

• http://www-aig.jpl.nasa.gov/public/mls/skicat/skicat-home.html

-천문학 이미지 분석 시스템인 SKICAT (The Sky Image Cataloging and Analysis Tool)에 관한 홈페이지

• http://www.cs.waikato.ac.nz/~ml

-뉴질랜드 Waikato 대학의 환경 분석을 위한 WEKA 홈페이지로서 많은 논문을 포함하는 출판물 링크와 관련된 연구에 대한 링크 등을 포함

• http://dml.cs.ucla.edu/~weiwang/myresearch.shtml

-VLDB 1997에 실린 STING 알고리즘에 대한 논문과 참고 문헌들에 대한 링크도 포함

• http://www.cs.ubc.ca/nest/dbsl/spatial.html

-근사치 근접 관계를 발견하기 위한 CRH 알고리즘에 대한 간략한 설명을 포함

• http://www.spatial.maine.edu

-미국 Maine 대학의 공간정보공학과의 홈페이지로서 연구 및 출판물과 on-line 관련 문헌 등에 대한 링크를 포함

• http://www.almaden.ibm.com/cs/quest/index.html

-IBM에서 개발하는 Quest 데이타 마이닝 시스템과 관련된 홈페이지

• http://www.sgi.com/Products/software/MineSet

-실리콘 그래픽스사의 데이타 마이닝과 시각화를 위한 MineSet에 대한 홈페이지

• http://www.cs.uoregon.edu/~wolf/research/bib.html

-데이타 마이닝에 대한 관련 문헌 목록을 일반 논문, 공간 데이타베이스, 기계학습, 통계학 등으로 분류하고 포스트스크립트 포맷으로 다운로드 받을 수 있도록 하며 관련 URL도 소개

• http://www.cs.sfu.ca/research/groups/DB/sections/publication/kdd/kdd.html

-데이타 마이닝에 대해 많은 연구를 수행한 캐나다 Simon Fraser 대학의 논문을 다운로드 받을 수 있도록 나열

• http://www.cs.helsinki.fi/research/pmdm/publications

-핀랜드의 헬싱키대학의 FDK(From Data to Knowledge) 관련 출판물에 대해 연도별로 분류하고 다운로드 받을 수 있도록 제공한 목록

• http://fas.sfu.ca/cs/people/Faculty/Han

-데이타 마이닝 분야에서는 저명한 Jia-Wei Han의 개인 홈페이지로서 현재 연구중인 DBMiner, LogicBase, GeoMiner, WebMiner

에 대한 링크와 출판물 및 교재 등을 수록

• http://fas.sfu.ca/cs/people/GradStudents/koperski/personal/research/research.html

—Kris Koperski의 개인 홈페이지로서 공간 데이타 마이닝에 대한 소개와 관련된 논문들을 다운로드 받을 수 있도록 제공

5. 결론 및 향후 연구 분야

공간 데이타 마이닝은 저장된 공간 및 비공간 데이타로부터 사용자에게 유용한 지식을 자동으로 추출해 주고, 사용자는 추출된 지식을 사용하여 의사 결정에 근거 자료로서 활용할 수 있다. 기존의 공간 데이타 마이닝 시스템으로는 관계형 데이타베이스 시스템과 GIS를 결합한 GeoMiner, 객체 지향 데이타베이스를 확장하고 Open GIS의 CORBA 표준을 채택한 ASK-ME, 이미지 프로세싱과 기계 학습을 통해 지능적으로 이미지 데이타를 분석하는 SKICAT, 기계 학습 도구들을 통합하여 공통 프레임워크와 단일 사용자 인터페이스를 제공해 농업 부문의 데이타 집합을 분석하기 위한 WEKA 등이 있다.

공간 데이타를 추출하기 위해 사용되는 알고리즘은 크게 일반화-기반 지식 발견과 클러스터링 방법이 있다. 일반화-기반 지식 발견은 트랜잭션 데이타 마이닝에서 사용하는 튜플-지향 알고리즘에 대조되는 관계형 데이타 마이닝을 위한 애트리뷰트-지향 알고리즘을 사용한 것으로서 개념 계층을 사용하고 추출되는 지식이 개념 계층에 크게 의존적이어서 개념 계층의 정의가 지식을 좌우한다는 단점이 있다. 클러스터링 방법은 개념 계층과 같은 배경 지식을 사용하지 않고 데이타로부터 직접 클러스터를 효율적으로 발견할 수 있다는 장점이 있다. 그러나, 사용자의 다양한 요구를 충족시키고 정확한 지식을 추출하기 위해서는 일반화-기반 지식 발견과 클러스터링 방법을 적절히 결합하는 형태가 필요하다.

현재까지 공간 데이타 마이닝은 주로 기존의 관계형 및 트랜잭션 데이타 마이닝에 공간 데이타 처리를 조금 추가하는 수준에 머무르고 있는데, 이는 공간 데이타의 특성을 제대로 반영하지 못하므로 바람직하지 않다고 생각된다. 그러므로, 공간 데이타의 특성들을 모두 지원하는 전용의 공간 데이타 마이닝 시스템에 대한 본격적인 연구가 필요하다.

컴퓨터의 활용 범위가 확대되면서 3차원 데이타를 위한 공간 데이타 마이닝이 필요하고, 공간뿐만 아니라 시간에 따른 지식도 추출하기 위해 시공간 데이타 마이닝에 대한 연구가 요구된다. 또한, 공간 데이타의 다양한 버젼들을 효율적으로 유지 관리하기 위해 버젼 관리의 지원에 관한 연구도 요구된다.

공간 데이타 사이의 관계를 측정하는 것은 공간 데이타 마이닝에서 기본이 되므로 효율적인 공간 유사 함수의 개발이 우선 해결하여야 할 문제이다. 공간 데이타 사이의 관계는 주로 객체간의 거리 또는 위상 관계에 의해 표현되므로 효율적인 거리 계산 알고리즘이나 위상 관계에 근거한 공간 데이타 마이닝의 연구가 필요하다.

효율적인 공간 데이타 마이닝을 위한 클러스터링 알고리즘과 클러스터링에 근거한 공간 인덱스 구조의 연구도 요구된다. 특히, 읽기 전용의 특성을 반영하는 공간 인덱스 또는 위상 관계 등의 부가적인 데이타를 추가로 저장할 수 있는 공간 데이타 마이닝 전용의 공간 인덱스 구조에 대한 연구도 필요하다.

지식을 추출하기 위한 데이타를 선택하기 위하여 사용자의 요구를 모두 수용할 수 있는 사용자 인터페이스에 대한 연구가 필요하다. 그리고, 추출된 지식을 사용자가 쉽게 인식할 수 있도록 표현하는 방법에 대한 연구도 필요하다. 특히, 추출된 지식중 공간 데이타를 포함하는 지식은 그래픽 출력을 사용하여야 하므로 그래픽 사용자 인터페이스에 대한 연구가 절실히 요구된다.

공간 데이타 마이닝 시스템은 다양한 분야의 기술이 집적된 형태로서 활용 범위를 넓히기 위하여 추가적인 다른 기술들과의 연계도 가능하다. 예를 들면, 공간 데이타 마이닝과 데이타 웨어하우스 기술을 결합하거나, 인터넷을 사용하는 공간 데이타 마이닝 기술 등에 관한 연구들이 앞으로 필요하겠다.

참고문헌

[1] D.A. Bell, S.S. Anand, and C.M. Shapcott, "Database Mining in Spatial Databases," Proc. of Int. Workshop on Spatio-Temporal Databases, 1994.

[2] M. Ester, H.P. Kriegel, and X. Xu, "Knowledge Discovery in Large Spatial Databases : Focusing Techniques for Efficient Class Identification," Proc. of the 4th Int. Symp. on SSD'95, pp. 67-82, 1995.

[3] 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.

[4] U. Fayyad, et al., "Automated Analysis of a Large-Scale Sky Survey : The SKICAT System," Proc. of KDD Workshop, pp. 1-13, 1993.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, 1996.

[6] S. Ford, J.A. Blakeley, and T.J. Bannon, "Open OODB : A Modular Object-Oriented DBMS," Proc. of ACM SIGMOD Int. Conf. on Management of Data, pp. 552-553, 1993.

[7] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus, "Knowledge Discovery in Databases : An Overview," Knowledge Discovery in Databases, AAAI Press/The MIT Press, pp. 1-27, 1991.

[8] S.R. Garner, "WEKA : The Waikato Environment for Knowledge Analysis," Proc. of New Zealand Computer Science Research Students Conference, University of Waikato, pp. 7-64, 1995.

[9] J. Han, K. Koperski, and N. Stefanovic, "GeoMiner : A System Prototype for Spatial Data Mining," http://db.cs.sfu.ca/GeoMiner.

[10] J. Han and Y. Fu, "Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases," Proc. of KDD Workshop, pp. 157-168, 1994.

[11] J. Han, S. Nishio, and H. Kawano, "Knowledge Discovery in Object-Oriented and Active Databases," F. Fuchi and T. Yokoi(eds), Knowledge Building and Knowledge Sharing, Ohmsha/IOS Press, pp. 221-230, 1994.

[12] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, A. Rajan, N. Stefanovic, B. Xia, and O.R. Zaiane, "DBMiner : A System for Mining Knowledge in Large Relational Databases," Proc. of the 2nd Int. Conf. on KDD-96, pp. 250-255, 1996.

[13] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data : An Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[14] E. Knorr and R.T. Ng, Applying Computational Geometry Concepts to Discovering Spatial Aggregate Proximity Relationship, Tech. Report, University of British Columbia, 1995.

[15] K. Koperski and J. Han, "Discovery of Spatial Association Rules in Geographic Information Databases," Proc. of the 4th Int. Symp. on SSD'95, pp. 47-66, 1995.

[16] K. Koperski, J. Adhikary, and J. Han, "Spatial Data Mining : Progress and Challenges," Proc. of SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96), 1996.

[17] W. Lu, J. Han, and B.C. Ooi, "Knowledge Discovery in Large Spatial Data-

bases," Proc. of Far East Workshop on Geographic Information Systems, pp. 275-289, 1993.

[18] L. Mohan and R.L. Kashyap, "An Object-Oriented Knowledge Representation for Spatial Information," IEEE Trans. on Software Engineering, Vol.14, No.5, pp. 675-681, 1988.

[19] R. Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Data Mining," Proc. of Int. Conf. on VLDB, pp. 144-155, 1994.

[20] B.W. Oh, J.K. Yun, and K.J. Han, "A Spatial Data Mining System Based on Clustering," Proc. of the 6th Int. Conf. on RSDMGrC(Rough Set, Data Mining, and Granular Computing) in JCIS'98, 1998(accepted).

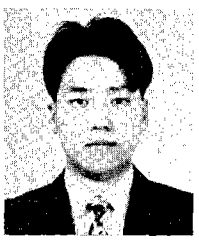
[21] Open GIS Consortium, Inc., OpenGIS Simple Features Specification for CORBA, Revision 1.0, 1998.

[22] P. Stolorz et al., "Fast Spatio-Temporal Data Mining of Large Geophysical Datasets," Proc. of the 1st Int. Conf. on KDD-95, pp. 300-305, 1995.

[23] W. Wang, J. Yang, and R. Muntz, "STING : A Statistical Information Grid Approach to Spatial Data Mining," Proc. of the 23rd VLDB Conf., pp. 186-195, 1997.

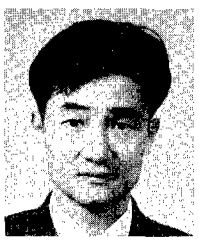
[24] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH : an Efficient Data Clustering Method for Very Large Databases," Proc. of ACM-SIGMOD Int. Conf. on Management of Data, pp. 103-114, 1996.

[25] 윤재관, 오병우, 한기준, "공간 데이타 마이닝을 위한 객체 관리 시스템," 한국정보과학회 학술발표논문집, 25권 1호, pp. 36-38, 1998.



오 병 우

1993 건국대학교 컴퓨터공학과 졸업(공학사)
1995 건국대학교 대학원 컴퓨터공학과 졸업(공학석사)
1995～현재 건국대학교 대학원 컴퓨터공학과 박사과정
관심분야 : 지리 정보 시스템, 객체 지향 데이타베이스, 공간 데이타 마이닝, 공간 데이타 웨어하우스
E-mail : bwoh@db.konkuk.ac.kr



이 강 준

1995 건국대학교 컴퓨터공학과 졸업(공학사)
1997 건국대학교 대학원 컴퓨터공학과 졸업(공학석사)
1997～현재 건국대학교 대학원 컴퓨터공학과 박사과정
관심분야 : 지리 정보시스템, 주기억-상주 데이타베이스, 객체 관계형 데이타베이스, 분산 데이타베이스
E-mail : kjlee@db.konkuk.ac.kr

한 기 준

1979 서울대학교 수학교육학과 졸업(이학사)
1981 한국과학기술원 전산학과 졸업(공학석사)
1985 한국과학기술원 전산학과 졸업(공학박사)
1985～현재 건국대학교 컴퓨터공학과 교수
1990 Stanford대학 전산학과 visiting scholar
관심분야 : 지리 정보 시스템, 객체 지향 데이타베이스, 공간 데이타 마이닝, 공간 데이타 웨어하우스, 주기억-상주 데이타베이스, 디지탈 라이브러리
E-mail : kjhan@db.konkuk.ac.kr