□신기술해설□

공간 데이터 마이닝

황 병 연† 김 의 찬††

◈ 목 차 ◈

1. 서 론

현재 우리가 살고 있는 공간 속에는 수많은 데이터들이 있다. 이런 방대한 양의 데이터들은 데이터베이스나 데이터웨어하우스, 지리정보 시스템 등에 저장되어 있다. 이러한 데이터들은 환경이 빠르게 변하면서 자연적으로 그 양도 계속적으로 빠르게 증가하고 있다. 이렇게 방대하고 빠르게 증가하는 데이터들을 분석하여 우리가 원하는 데이터들을 쉽고 정확하게 찾기란 매우 힘들다. 최근 연구되는 분야 중에 데이터 마이닝(Data Mining) 혹은 지식발견(KDD: Knowledge Discovery in Database) 분야가 이러한 문제를 풀기 위해 나온 분야이다. 데이터 마이닝을 간단하게 정의한다면 방대한 데이터베이스 상에서 흥미 있고, 암시적인, 이전에 알려지지 않은 지식을 추출하는 것이라 할 수 있다[1, 2]. 이런 연구를 통해 많은 데이터들 중에 원하는 데이터들만을 쉽고 정확하게 찾을 수 있는 것이다.

기존의 데이터베이스로는 공간 데이터를 다루기에는 문제가 있어서 공간 데이터베이스가 생겼듯이 기존의 관계형 데이터베이스 내의 데이터 마이닝으로는 공간 데이터베이스 내의 데이터들을 다루기는 매우 부족한 점이 많다. 그래서 앞에서 데이터 마이닝을 정의했던 것과는 조금은 다르게 공간 데이터 마이닝을 정의하게 된다.

공간 데이터 마이닝 혹은 공간 데이터베이스 상의 지식 발견은 함축적인 지식 추출, 공간 관계 혹은 공간 데이터베이스 상에 명백히 저장되어 있지 않은 다른 패턴들을 추출하는 것이라 말할 수 있다[3]. 그런 지식 발견은 공간 데이터를 이해하고, 공간 데이터와 비공간 데이터간의 본질적인 관계를 찾아내며, 축약된 방법으로 데이터 규칙성을 나타내는 중요한 역할을 하게 된다. 또한 데이터 의미론을 조정하고 좋은 수행을 하도록 공간 데이터베이스를 재구성하는 역할을 한다.

공간 데이터 마이닝에서 중요한 것은 방대한 양의 공간 데이터와 공간 데이터 타입, 그리고 공간 접근 방법들을 다루는 공간 데이터 마이닝 알고리즘의 효율성이다. 공간 데이터 마이닝 방법들은 기존의 데이터 마이닝의 방법들을 확장하여 이용하면 기존의 데이터베이스와 공간 데이터베이스와의 조화를 한 층 더 높일 수 있고 효율적으로 지식의 새로운 형태를 찾아낼 수도 있을 것이다.

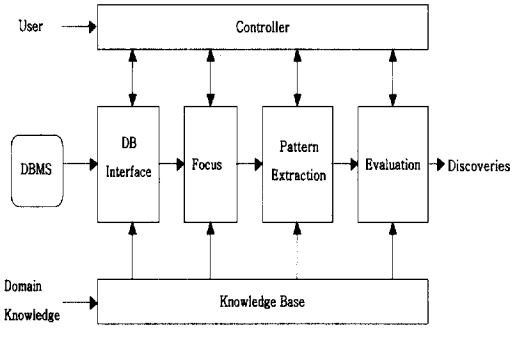
최근 대용량 데이터베이스, 그 중 하나인 공간 데이터베이스 상에서의 지식 발견 즉 공간 데이터 마이닝에 관한 연구가 많이 진행되고 있다. 이

† 정 회 원 : 가톨릭대학교 컴퓨터전자공학부 부교수
†† 준 회 원 : 가톨릭대학교 컴퓨터공학과 석사과정

에 바탕으로 본 연구에서는 공간 데이터 마이닝에서 사용되고 있는 기법들을 살펴보고 앞으로의 방향을 제시해 보고자 한다.

2. 공간 데이터 마이닝 구조

데이터 마이닝에서 다양한 모델들이 제안되었다. 그 모델들 중에는 일반적인 데이터 마이닝 프로토타입(prototype)인 DBLEARN/DBMINER에 관한 Han의 구조, Holsheimer의 병렬 구조, 그리고 Matheus의 다중컴포넌트 구조들이 포함되어 있다. 이들 구조의 거의 대부분이 공간 데이터 마이닝을 다루기 위해 사용되거나 확장되어 왔다. Matheus의 구조는 일반적인 구조로 Easter를 포함해서 공간 데이터 마이닝 상의 여러 연구가들에 의해 사용되고 있다. 이 구조는 (그림 1)과 같다[4, 5].



(그림 1) KDD 시스템 구조

이 구조에서 사용자는 데이터 마이닝 처리의 모든 단계를 조절할 수 있다. 공간과 비공간 개념 계층 같은 배경 지식은 지식 베이스(Knowledge Base)에 저장되어 있다. 데이터는 질의 최적화를 가능하게 하는 DB 인터페이스(Interface)를 사용한 저장장치로부터 가져온다. R-tree 같은 공간 인덱스 구조는 효율적인 처리를 위해 사용된다. 집중 컴포넌트(Focusing Component)는 데이터의 일부를 패턴인식에 사용하도록 한다. 예를 들어 몇몇 속성(attribute)들만이 지식 발견 작업에 적절하다고 결정하거나 혹은 좋은 결과물을 내는 것을 보장하는 객체의 사용법을 추출할 수 있다. 규칙과 패턴들은 패턴 추출(Pattern Extraction) 모듈에 의해 찾는다. 이 모듈은 통계학, 기계 학습, 그리고 규칙과 관계를 찾는 작업을 수행할 수 있는 계산 기하학적 알고리즘과 접목시킨 데이터 마이닝 기법 등에서 사용할 수 있다. 이렇게 나온 패턴들의 흥미성과 중요성은 가능한 한 명백하고 여분의 지식을 제거하기 위해 평가(Evaluation) 모듈에 의해 처리된다. 이들 네 개의 컴포넌트들은 제어기(Controller) 부분을 통해서 서로들 간의 상호 작용을 하게 된다.

3. 공간 데이터 마이닝 방법

3.1 통계학적 분석

통계학적 분석(Statistical Analysis)은 가장 일반적인 공간 데이터 분석 기법이다. 이 통계학적 분석 방법은 수치 데이터(numerical data)를 잘 다룰 수 있고 많은 알고리즘을 가지고 있다. 또한 공간 현상에 관한 모델을 얻기에 좋고 최적화를 가능하게 한다. 그러나 통계학적 분석은 공간상에 분산된 데이터들이 독립성을 갖는다고 가정해야만 한다. 이러한 가정에 따른 결과가 좋지 않은 경우는 하나의 영역에서 이웃하는 영역의 영향이 큰 경우이다[3, 4].

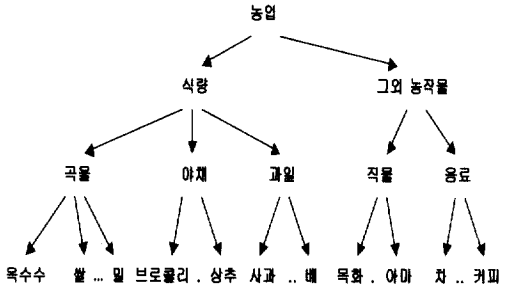
이러한 문제를 다루기 위해 공간 모델은 어떠한 변수를 포함할 수 있다. 의존 변수라는 공간적 지연 형태를 갖는 회귀모델은 이런 문제를 조금이나마 해결할 수 있다. 하지만 전체를 모델링하는 과정이 복잡하고 통계학적인 전문지식을 갖고 있는 사람만이 할 수 있다. 게다가 비선형적인 규칙들은 잘 다룰 수 없으며 이름 같은 기호 값들도 통계학적인 분석 방법에서는 잘 다룰 수 있는 것들이 아니다.

통계학적 분석 방법은 불완전하거나 비결정적인 데이터들을 잘 다룰 수 없으며 결과를 계산해 내기에 비용이 많이 든다.

3.2 일반화 기반 지식 발견

데이터베이스 상의 데이터와 객체들은 각 개념 단계에 상세한 정보를 가지고 있다. 상위 개념 단계상에서는 하위 개념 단계에서보다 데이터와 객체들의 집합들이 더 일반적인 정보를 가지고 있으며 더 요약해서 나타내게 된다. 이것이 일반화 기반 데이터 마이닝이다. 하위 개념 단계로부터 상위 개념 단계로 올라가면서 적절하게 데이터 집합을 추상화하는 것이다. 그리고 그렇게 일반화된 데이터에서 지식을 추출해 낸다.

일반화 기반 지식 발견(Generalization-Based Knowledge Discovery)[6]에서 개념 계층의 형태에는 어떤 배경 지식이 있다고 가정을 한다. 이러한 개념 계층은 전문가에 의해 만들어질 수도 있고, 자동적으로 데이터 분석에 의해 만들어질 수도 있다. 개념 계층에 대한 예를 들면 (그림 2)와 같다. 그림에서 농업보다 더 상세한 정보는 식량과 그 외 농작물로 나누어진다. 즉, 농업이 식량과 그 외 농작물보다 더 일반적인 개념인 것이다. 공간 데이터베이스 상에서는 두 종류 즉, 비공간과 공간 개념 계층을 정의할 수 있다. 개념 계층이 올라가면 올라갈수록 더 일반적인 정보를 가지고 있다. 지식 발견은 함축적이고 일반적인 정보를 가지고



(그림 2) 농업 계층도

있으며 적은 수의 분리를 갖는 상위 개념 단계에 의해 표현되어져야 한다.

데이터 일반화 기법에서 효율적인 기법으로 애트리뷰트 지향 유도(attributed-oriented induction) 방법이 있다. SQL 같은 데이터 마이닝 질의 언어를 이용하여 데이터 마이닝 질의를 해서 공간 데이터베이스 내에 있는 데이터들 중에 적절한 데이터의 집합을 추출한다. 그리고 일반화된 계층에 따라서 데이터 일반화를 수행한다. 또한, 더 높은 개념 단계상에서 공간 데이터와 비공간 데이터간의 일반화된 관계를 요약하며 수행된다.

튜플의 애트리뷰트 값이 일반화된 값으로 대치될 때 개념 계층에서 올라가고, 더 일반화하기가 불가능하고 애트리뷰트에 대한 값들이 너무 많게 되면 애트리뷰트를 삭제하고, 동일한 튜플들은 병합함으로써 비공간 데이터로 수행될 수 있다. 유도는 모든 애트리뷰트가 요구되는 단계로 일반화될 때까지 계속된다.

일반적으로 학습 요청은 지식 발견을 위해 적절한 데이터들의 집합, 개념 계층, 요구되는 규칙 형태와 같은 프리미티브(primitive)들을 제공한다. 이러한 프리미티브를 가지고 공간 데이터베이스로부터 일반 지식을 추출하기 위해 일반화는 항상 공간 데이터와 비공간 데이터 둘 다 수행할 필요가 있다. 여기서 공간 데이터와 연관된 비공간 데이터는 SAND(Spatial-and-Nonspatial Database) 구조[7]에서와 같이 연결되어 있다. 그래서 두 컴포넌트 중에서 어느 하나를 일반화했을 때 다른 컴포넌트는 그에 따라 적용하게 되는 것이다. 공간 데이터베이스에 대한 애트리뷰트 지향 유도에는 두 가지 알고리즘이 있는데 하나는 비공간 데이터 우위 일반화(Non-Spatial-Data-Dominant Generalization)이고 또 하나는 공간 데이터 우위 일반화(Spatial-Data-Dominant Generalization)이다.

일반화 기반 지식 발견의 예를 들면, British Columbia(B.C.)에서 나타나는 500가지 이상의 기

후 상태를 포함하는 기후 데이터 집합이 주어졌다고 하고, 그 데이터 집합을 이용하여 몇 연도, 어느 계절상에서 B.C. 내의 지역들에 나타나는 일반적인 기후 패턴을 찾는 것이다.

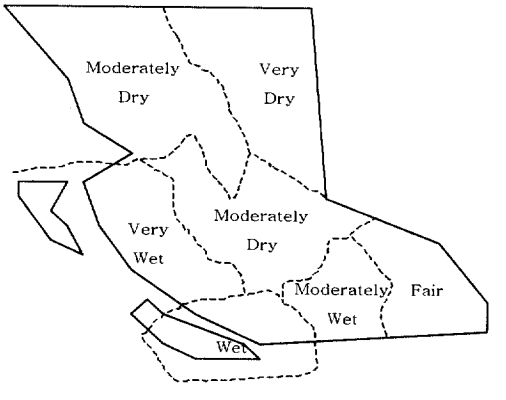
3.2.1 비공간 데이터 우위 일반화

비공간 데이터 우위 일반화 알고리즘은 똑같이 상위 단계 비공간 기술들을 분할하여 적은 수의 영역들로 이루어진 지도를 만드는 것이다. 이 알고리즘은 더 상위 개념 단계에서 비공간 애트리뷰트들을 일반화하면서 비공간 애트리뷰트로 애트리뷰트 지향 유도를 한다. 그러면 똑같은 일반화된 애트리뷰트 값을 갖는 이웃 영역 모두가 병합되어진다. 근사치를 이용하여 다른 비공간 설명을 갖는 작은 영역들은 무시할 수도 있다. 예를 들면, 우리가 찾고자 하는 것은 1990년 봄에 나타난 강우량 패턴별로 영역들을 분류하여 기록하는 것이다. 이것을 하기 위해 우리가 가지고 있는 데이터들은 상위 단계 강우량 개념과 년-계절-달 계층도, 기상 관측소 위치가 나타난 B.C.지도, 1990년 강우량 데이터 샘플 도표 등이다. 질의문은 다음과 같다.

extract region
from precipitation-map
where province = "B.C." and period = "spring"
and year = 1990
in relevance to precipitation and region.

질의문을 바탕으로 검색을 하게 되는데 봄이라고 했으므로 년-계절-달 계층도 에서 "spring"을 찾아 Mar., Apr., May인 것을 확인하고 1990년도의 강우량 데이터 샘플 도표에서 각 기상 관측소의 Mar., Apr., May 때의 데이터들을 검색한다. 3달의 평균 강우량을 구하고 평균 강우량 값에 따라 상위 단계 강우량 개념 표에 맞게 일반화한다 (e. g. Wet은 (2.0, 5.0]인데 Nanaimo 도시의 평균 강우량은 2.85, 그러므로 Nanamio 도시의 강우량을 일반화시키면 "Wet"이 된다.).

이렇게 하여 일반화된 비공간 데이터를 이용하여 공간 일반화를 하게 된다. 똑같은 상위 단계 애트리뷰트 값(Wet, Dry,...)을 갖는 영역들이 이웃하고 있으면 공간 함수 adjacent_to를 기반으로 해서 병합시킨다. 이렇게 공간 객체들은 적은 수의 영역들로 일반화시키고 병합시키기 위해서 근사치를 이용하게 된다. 예를 들면, 어느 영역에서 대부분의 위치가 Wet이고, 일부가 Very Wet이라면 이를 무시하고 그 영역을 Wet으로 일반화시키는 것이다. 이러한 방법으로 일반화된 형태가 (그림 3)에 나타나 있다.



(그림 3) 영역별 강우량 일반화 예

3.2.2 공간 데이터 우위 일반화

공간 데이터 우위 일반화 알고리즘은 상위 단계 프레디킷을 사용하는 공간 영역을 기술할 수 있다. 먼저 알고리즘은 공간 계층을 따르는 공간 영역을 병합한다. 그러면 각 영역에 대한 비공간 기술은 애트리뷰트 지향 유도 방법을 이용하여 만들어낼 수 있다. 질의에 대한 답은 각각의 일반화된 영역들을 특성화한 소수의 프레디킷의 분리를 이용한 모든 지역에 대한 기술이다.

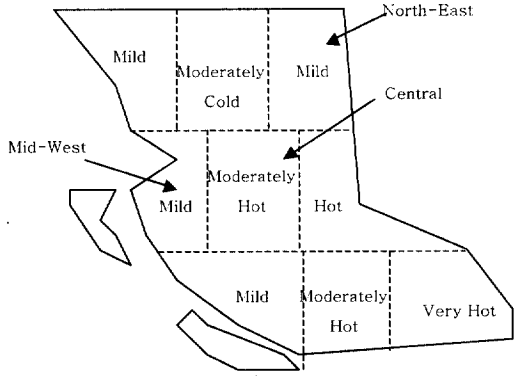
예를 들어 설명하면 영역별 온도 데이터들과 기후에 대한 상위 단계 개념 표가 주어졌다고 하

자. 우리가 찾으려고 하는 것이 1990년 여름에 미리 규정되어 있는 행정 지역의 일반적 온도 정보를 찾는 것이다. 질의문을 살펴보면 다음과 같다.

extract characteristic rule
from temperature-map
where province = "B.C." and period = "summer"
and year = 1990
in relevance to region and temperature.

이 질의문을 통해 적절한 데이터들을 모으고 이들 영역과 일치하는 공간 데이터 객체들을 클러스터링 하여 공간 데이터베이스를 일반화하고, 요구되는 개념 단계에 도달할 때까지 혹은 일반화된 공간 객체의 수가 임계값(threshold)에 이를 때까지 일치하는 비공간 포인터들을 병합한다.

공간 일반화에서 각 사분면의 객체들은 공간 계층의 첫 단계에 일치하도록 먼저 병합한다. 이들 영역 상의 온도들에 대한 평균값을 계산하고 상위 단계 개념 표를 이용하여 일반화시킨다(e. g. 사분면의 Mid-West 온도 평균이 36이면 Mild ([32, 50))로 일반화시킨다.). 이러한 방법으로 일반화된 지도가 (그림 4)에 나타나 있다.



(그림 4) 영역별 온도 일반화 예

3.3 클러스터링

클러스터링(Clustering)[8] 분석은 통계학의 한 부분으로서 수년동안 연구되어왔다. 이 클러스터링의 장점은 데이터의 특징에 대한 배경 지식 없이 데이터의 클러스터들을 찾을 수 있다는 점이다. 클러스터링 알고리즘으로 PAM(Partitioning Around Medoids)과 CLARA(Clustering LARge Applications)가 있는데 이들은 계산 복잡도의 견해로 보았을 때 조금은 비효율적이다. 그래서 PAM과 CLARA의 보완으로 CLARANS(Clustering Large Application based on RANdomized)라는 알고리즘이 나왔고 이 알고리즘을 이용하여 공간 데이터 마이닝에서 데이터 클러스터 처리를 하게 되는 것이다.

먼저 PAM과 CLARA에 대해 간단히 살펴보고 CLARANS에 대해 살펴보도록 하겠다. PAM 알고리즘은 먼저 n개의 객체가 있다고 가정한다. 그리고 k개의 클러스터를 찾는다. 이들 각 클러스터에서 대표 객체를 찾는다. 이들 대표 객체들은 메도이드(medoid)라고 부르는데 클러스터에서 중간에 위치한 점을 말한다. k 메도이드를 선택하고 나서도 가능한 모든 객체들의 쌍들을 분석하여 더 좋은 메도이드를 찾기 위해 알고리즘을 반복적으로 시도한다. 한번의 반복에서 선택된 최상의 점들은 다음 반복을 위한 메도이드로 쓰여지게 된다.

CLARA는 PAM과 유사하지만 샘플링(sampling)을 기반으로 하고 있다. 데이터의 적은 양만으로 대표값을 선택하고 PAM을 이용해서 이 샘플들로부터 메도이드를 찾는다. 샘플들이 공정한 무작위 추출방법으로 선택되었다면 이들 샘플들은 전체 데이터 집합을 비슷하게 나타낼 수 있다. 그렇게 선택된 메도이드들은 전체 데이터 집합에서 선택된 메도이드들과 유사하게 되는 것이다. CLARA는 다수의 샘플들을 추출해내고 이들 샘플들 중에서 최선의 클러스터링을 산출하게 되는 것이다.

CLARANS는 어떤 샘플에만 제한을 두지 않고 데이터 집합의 부분 집합만을 검색하여 PAM과 CLARA를 혼합하는 것이다. CLARA가 검색의 모

든 단계에서 고정된 샘플들을 갖는 반면에 CLARANS는 메도이드들 중의 한 샘플을 끌어내어 클러스터링 값과 메도이드들 중의 하나를 대치하여 생산되는 무작위 추출 이웃 클러스터를 비교한다. 더 좋은 이웃을 발견한다면 이웃 클러스터는 같은 단계를 다시 반복할 것이다. 이 과정은 M개의 이웃들보다 더 좋은 클러스터가 발견될 때까지 계속될 것이다. 그래서 위치 최적(local optimum)을 찾아낸다. CLARANS는 위치 최적들을 찾을 때까지 다른 위치 최적을 찾으려 할 것이다. CLARANS에서 모든 반복에 대한 계산 복잡도는 기본적으로 객체의 수에 비례한다.

집중 방법(Focusing Methods)을 이용한 CLARANS의 변형은 계산 비용을 감소시키는 것이 목적이다[9]. CLARANS에서 계산적으로 가장 비싼 단계가 두 클러스터들 간의 총 거리를 계산하는 것이다. 그래서 이런 단계의 비용을 감소시키기 위한 두 가지 접근 방법이 제안되었다.

첫째로 고려되는 객체의 수를 감소시키는 것이다. 중심 질의(centroid query)는 이웃 포인트가 저장되어있는 R* 트리의 리프 노드 중의 가장 중심에 있는 객체를 반환한다. 이들 객체만으로 클러스터의 메도이드를 계산하는데 사용한다. 다른 방법으로는 실제로 계산하는데 쓰이지 않는 어떤 객체들에 대한 접근을 제한하는 것이다. R* 트리 구조를 사용하여 객체들의 모든 쌍을 검사하는 대신에 클러스터링의 질을 향상시킬 수 있는 객체의 쌍으로만 계산을 수행하는 것이다.

CLARANS 알고리즘을 기반으로 하여 두 개의 공간 데이터 마이닝 알고리즘이 개발되었는데 공간 우위(Spatial Dominant) 접근인 SD(CLARANS) 알고리즘과 비공간 우위(Non-Spatial Dominant) 접근인 NSD(CLARANS) 알고리즘이다. 두 개의 알고리즘 모두 찾아낸 규칙의 형태와 적절한 데이터들은 학습 요구를 통한 사용자에 의해 규정되어진다고 가정한다.

SD(CLARANS) 알고리즘은 공간 우위 접근으로서 적절한 데이터들의 공간 컴포넌트들은 CLARANS를 이용하여 클러스터링을 한다. 그런다음 각 클러스터 상의 객체들의 비공간 컴포넌트들에 속성 지향 유도를 수행한다. 질의의 결과는 모든 클러스터 상에서 객체들의 상위 단계 비공간 기술을 나타낸다. NSD(CLARANS) 알고리즘은 비공간 우위 접근으로서 먼저 비공간 일반화를 적용하는 것이다. 속성 지향 일반화(Attribute-oriented generalization)를 비공간 애트리뷰트에 수행하고 많은 일반화된 튜플들을 만들어 낸다. 그렇게 각 일반화된 튜플에 대해 모든 공간 컴포넌트들을 모으고 k 클러스터들을 찾기 위해 CLARANS를 이용한 클러스터링을 하게 된다. 마지막 단계에서 클러스터들은 다른 타입을 기술하는 객체들의 클러스터들과 중복이 되는지를 확인하고 그렇다면 클러스터들을 병합한다. 그리고 튜플들 중에 일치하는 일반화된 비공간 기술들 또한 병합한다.

그 외 클러스터링을 이용한 방법으로는 CF(Clustering Feature)와 CF 트리를 사용한 BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)가 있고, 통계적 정보 그리드 기반 접근 방법인 STING(STatistical INformation Grid)과 STING+가 있다.

3.4 공간 연관 규칙

공간 데이터 마이닝은 기본적인 규칙들의 종류를 분류할 수 있다. 공간 특징 규칙(spatial characteristic rule)은 공간 관련 데이터의 집합에 대한 일반적 기술이다. 예를 들면, 지도상에 나타나는 지역들 중에 어떤 지역의 일반적인 기후 패턴에 대한 기술 같은 것이다. 공간 분별 규칙(spatial discriminant rule)은 공간 관련 데이터 부류의 특징들을 다른 부류와 대조하는 혹은 분별하는 것에 대한 일반적인 기술이다. 예를 들면, 두 지역 간에 기후 패턴에 대한 비교를 들 수 있다. 공간

연관 규칙(spatial association rule)[3]은 공간 데이터베이스에서 다른 집합의 특징을 통해 하나 혹은 어떤 집합의 특징에 대한 관련성 혹은 관계를 기술하는 것이다. 예를 들면, 캐나다의 큰 도시들은 대부분 캐나다와 U.S 경계에 가까이 있다라는 규칙을 들 수 있다.

일반화 기반은 공간과 비공간 관계를 일반적 개념 단계에서 찾는다. 여기서 공간 객체들은 공간 영역들을 합병하거나 공간 점을 클러스터링하여 나타낸다. 그러나, adjacent_to, near_by, inside, close_to, intersect와 같은 공간 관계들과 공간 객체들의 구조를 반영하는 규칙들은 찾을 수가 없다. 공간 연관 규칙은 공간 프레디킷을 포함하는 객체 관계를 표현한다. 예를 들면,

is_a(x, house) ∧ close_to(x, beach) → is_expensive(x). (90 %)

is_a(x, gas_station) → close_to(x, highway). (75 %)

다양한 종류의 공간 프레디킷들은 위상적인 관계를 표현한다. 또한 왼쪽, 오른쪽, 북쪽 등과 같은 공간 방위나 순서를 표현하기도 한다. 공간 연관 규칙에 대한 마이닝 연구에서는 몇 가지 예비 개념이 필요한데 다음과 같다.

[정의 1] 공간 연관 규칙은 아래와 같은 형태를 갖는 규칙이다.

P1 ∧ ... ∧ Pm → Q1 ∧ ... ∧ Qn. (c %)

여기서, 프레디킷 P1, ... , Pm, Q1, ... Qn 들 중 적어도 하나는 공간 프레디킷이고 c %는 규칙의 신뢰도를 나타낸다. 이 정의에 따라서 많은 공간 규칙들을 유도할 수 있다. 그러나 대부분의 사람들은 종종 발생되는 패턴이나 강한 관련성이 있는 규칙들에만 관심 있어한다. 큰 지지도와 높은 신뢰도를 갖는 규칙들이 강한 규칙들이다.

[정의 2] 프레디킷들의 결합인 지지도는 집합 S의 카디널리티에 대해 P를 만족하는 S내의 객체들의 수이다. S상에서 규칙 P → Q의 신뢰도 φ (P → Q/S)은 σ(P/S)에 대한 σ(P∧Q/S)의 비율이다. 다시 말하면, P가 S의 똑같은 멤버들에 의해 만족되어질 때 Q가 S의 멤버에 의해 만족되어지는 가능성이다. 하나의 단일 프레디킷은 1-프레디킷이라 부르고, k개의 단일 프레디킷들의 결합은 k-프레디킷이라 부른다.

[정의 3] 프레디킷 P의 지지도가 k 단계에 대한 최소지지 임계값 σ_k과 다름없다면 프레디킷 P의 집합은 k 단계에 집합 S상에서 크다고 한다. 그리고 개념 계층으로부터 P의 모든 조상들은 일치하는 단계들에서 크다고 한다. 신뢰도가 최소신뢰 임계값 ϕ_k과 다름없다면 규칙 "P → Q/S"의 신뢰도는 k 단계에서 높다고 한다.

[정의 4] 프레디킷 "P∧Q"는 집합 S에서 크고 규칙 "P→Q/S"의 신뢰도는 높다고 하면 규칙 "P →Q/S"는 강한 관계라고 한다.

공간 연관 규칙 마이닝을 하는 예를 살펴보자. British Columbia(B.C.) 지도를 이용하여 설명하면, 먼저 지도상의 공간 객체들을 크게 대도시, 도로, 물(호수, 강, ...), 광산, 경계로 나눈다. 다음에, 대도시와 나머지 네 개의 클래스들 사이의 일반화된 close_to(g_close_to) 관계를 플레인 스위핑(plane sweeping) 알고리즘이나 R* 트리 등으로 계산한다. 그렇게 해서 각각의 대도시별로 g_close_to 관계를 표로 정리할 수가 있다. 이 표를 다시 intersect, adjacent_to, close_to, inside 등의 프레디킷을 이용하여 재정리할 수 있다(e. g. Victoria 도시(town)인 경우 물(water)과의 관계는 <adjacent_to, J.Fuca_Strait>이고, 도로(road)와의 관계는 <intersects, highway_1>, <intersects, highway_17>이고, 경계(boundary)와의 관계는 <close_to, US>이다.).

재정리된 표를 통해서 상위 개념 단계의 k-프레디킷 집합을 표로 구성할 수 있는데 그 표는 (표 1)과 같다. (표 1)에서 count는 각 도시에 대해서 프레디킷 집합의 관계를 갖는 개수이다.

〈표 1〉 상위 개념 단계에 대한 large k-프레디킷 집합(40개 도시에 대한)

k	large k-프레디킷 집합	count
1	<adjacent_to, water>	32
1	<intersects, highway>	29
1	<close_to, highway>	29
1	<close_to, us_boundary>	28
2	<adjacent_to, water>, <intersects, highway>	25
2	<adjacent_to, water>, <close_to, us_boundary>	23
2	<close_to, us_boundary>, <intersects, highway>	26
3	<adjacent_to, water>, <close_to, us_boundary>, <intersects, highway>	22

이제 (표 1)을 통해서 다음과 같은 공간 연관 규칙을 추출할 수가 있다.

is_a(X, large_town) ∧ intersects(X, highway) → adjacent_to(X, water). (86%)

<intersects, highway>의 수가 29이고 <adjacent_to, water>, <intersects, highway>의 수가 25이므로 25/29 = 86 %가 나오게 된다. 해석을 하자면 고속도로와 교차하는 대도시가 물과 인접해 있을 확률이 86%가 된다는 것이다. 이와 같이 각 개념 단계별로 표를 만들고 그 표를 이용하여 공간 연관 규칙을 추출할 수 있다.

4. 공간 데이터 마이닝 관련 사례

4.1 GeoMiner

GeoMiner[10] 시스템은 관계형 데이터 마이닝 시스템인 DBMiner 시스템을 공간으로 확장시킨 시스템이다. 비공간 데이터들은 DBMiner 시스템에서 비공간 데이터 마이닝에 대한 함수들을 통해 처리되고, 공간 데이터들과 비공간 데이터와 공간 데이터 사이의 관계에 대한 것들은 GeoMiner 전용 함수들에 의해서 처리된다. 처리된 출력 결과는 그래픽 사용자 인터페이스를 통해 테이블, 차트, 지도 등의 형태로 제공된다.

GeoMiner 시스템은 현재 5가지의 기능 모듈이 제공되고 있다. Geo-characterizer 모듈은 공간 데이터베이스 상의 적절한 데이터들로부터 특징적 규칙들의 집합을 찾고, Geo-comparator 모듈은 적절한 데이터 집합에서 다른 클래스들 간의 일반적인 특징들을 대조하여 비교 규칙들의 집합을 찾고, Geo-associator 모듈은 적절한 데이터 집합들로부터 공간 관련 연관 규칙들의 집합을 찾는다. Geo-cluster analyzer 모듈은 적절한 비공간 기술을 갖는 포인터들의 클러스터를 찾는데, 공간 클러스터링에는 CLARANS 알고리즘을 사용하고, 클러스터들의 비공간 기술을 찾는 것은 애트리뷰트 지향 유도를 사용한다. Geo-classifier 모듈은 비공간 애트리뷰트들 중의 하나에 따라서 적절한 데이터의 집합을 클래스화하는 결정 트리를 만들기 위해 일반화 기반 결정 트리 유도 방법을 적용한다. Geo-Miner 시스템에서 사용되는 질의어로는 Spatial SQL을 확장시킨 GMQL(Geo-Mining Query Language)을 사용하고 있다.

4.2 GeoVRML Visualization

GeoVRML[11]은 미국 환경보호청(EPA: Environmental Protection Agency)에서 지리공간 데이터 집합의 상호 작용적인 탐색을 위하여 VRML(Virtual Reality Modeling Language)을 이용하여 만든 공간 데이터 마이닝을 위한 도구이다. 자바 사용자 인터페이스를 제공하며, DEM(Digital Elevation Model) 데이터 집합을 ARC/INFO 형태로 변환시키도록 하였고, ARC/INFO 포맷으로부터 웹상에서 볼 수 있고 상호작용을 위한 VRML 포맷으로 변환시킬 수 있도록 하였다. 파라미터 구성요소 선택 페이지를 시작으로 해서 ARC/INFO 데이터베이스를 검색하여 요구되는 DEM 데이터를 찾아 색상이

들어간 지도를 선택하게 된다. 다음으로 자바 애플릿을 통해 VRML 파일로 만들어지고 사용자들의 VRML 브라우저 플러그 인을 통해 VRML 파일이 로드 된다. 다수의 사용자들의 동시 접속을 위해 VRTP(Virtual Reality Transfer Protocol)를 포함하는 협동적 VRML을 지원하도록 하였다. 협동적 VRML은 공간 데이터베이스로부터 생성되는 GeoVRML 파일들의 상호 작용 실시간 공유를 가능하게 하고 인터넷상에서의 탐색을 지원해 준다. 현재 3차원 시각화 탐색을 할 수 있도록 개발 중에 있다.

5. 결론 및 향후 연구과제

공간 데이터 마이닝은 함축적인 지식 추출, 공간 관계 혹은 공간 데이터베이스 상에 명백히 저장되어 있지 않은 다른 패턴들을 추출하는 것이다. 여기서 사용되는 구조로는 Matheus의 KDD 시스템 구조가 많이 사용되고 있다.

본 연구에서는 공간 데이터 마이닝에서 사용되는 여러 방법들을 분류해서 각각의 기법들을 설명하고 공간 데이터 마이닝 관련 사례들을 살펴보았다. 공간 데이터 마이닝을 위한 방법들은 현재에도 계속 나오고 있지만 그 바탕이 되는 기본적인 알고리즘으로 통계학적 분석 방법, 공간 우위 혹은 비공간 우위의 일반화 기반 발견 방법, 클러스터링 방법, 공간 연관 규칙 방법 등이 있다. 공간 데이터 마이닝 관련 사례로는 GeoMiner와 GeoVRML 들이 있다.

향후 연구과제로는 공간 데이터 마이닝을 위한 실시간 그래픽 인터페이스 제공이 요구되며, 3차원적 시각화 데이터를 처리할 수 있는 시스템 및 애플리케이션들의 개발이 요구된다. 또한, 기존의 데이터 마이닝 방법들을 확장시키거나 현재 발표된 공간 데이터 마이닝 알고리즘을 응용하되 가정에 따른 제약을 줄일 수 있는 좀 더 효율적이고 안정적인 알고리즘을 개발하는 것이 요구된다.

참고문헌

[1] 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, 1991, pp. 1-27.

[2] M. Chen, J. Han, and P. Yu, "Data Mining: An Overview from Database Perspective," IEEE Transactions on Knowledge and Data Eng., December 1996, pp. 866-883.

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

[4] K. Koperski, J. Adhikary, and J. Han, "Spatial Data Mining: Progress and Challenges," Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 1996.

[5] C. Matheus, P. Chan, and G. Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases," IEEE Transactions on Knowledge and Data Engineering, vol. 5, 1993, pp. 903-913.

[6] W. Lu, J. Han, and B. C. Ooi, "Discovery of General Knowledge in Large Spatial Databases," Proc. Far East Workshop on Geographic Information Systems, Singapore, June 1993, pp. 275-289.

[7] W. G. Aref and H. Samet, "Optimization Strategies for Spatial Query Processing," Proc. of 17th Int. Conf. on VLDB, Barcelona, Spain, Sept. 1991, pp 81-90.

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

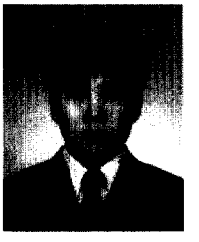
[9] M. Ester, H. P. Kriegel, and X. Xu, "Knowledge Discovery in Large Spatial Databases : Focusing Techniques for Efficient Class Identification," Proc. of the 4th Int. Symp. on Large Spatial Databases, 1995, pp. 67-82.

[10] http://db.cs.sfu.ca/GeoMiner/ "GeoMiner: A Knowledge Discovery System for Spatial Databases and Geographic Information Systems."

[11] http://www.geog.psu.edu/ica/icavis/rhyne98.html "Geo-VRML Visualization: A Tool for Spatial Data Mining"

황 병 연

1986년 서울대학교 전자계산기 공학과 졸업(공학사)
1989년 한국과학기술원 전산학과 졸업(공학석사)
1994년 한국과학기술원 전산학과 졸업(공학박사)
1999년-2000년 Univ. of Minnesota Visiting Scholar
1994년-현재 가톨릭대학교 컴퓨터전자공학부 부교수
관심분야 : 공간 데이터베이스(GIS), WWW 데이터베이스, 전자상거래
e-mail:byhwang@www.cuk.ac.kr



김 의 찬

1999년 가톨릭대학교 전산학과 졸업(이학사)
1999년-현재 가톨릭대학교 컴퓨터 공학과 석사과정 재학중
관심분야 : 공간 데이터베이스, 공간 데이터마이닝, 전자상거래
e-mail:eckim@songsim.cuk.ac.kr