규칙베이스 기반의 일반화를 확장한 공간 데이터 마이닝 시스템

최 성 민† · 김 응 모††

요 약

대용량의 공간(spatial) 데이터베이스에서 사용자에게 관심있고 일반화된 지식을 추출하는 것은 지형 정보 시스템이나 지식 베이스 시스템의 개발에 중요한 기법중의 하나이다. 본 논문은 공간 데이터 마이닝에 널리 사용되는 일반화(generalization) 방법을 확장한 공간 데이터 마이닝 모듈에 공간 데이터를 추론할 수 있도록 구축된 규칙베이스(rulebase)를 통합한 공간데이터 마이닝 시스템을 제안한다. 이를 위한 전위기로서 공간 데이터 우선(spatial data dominated)과 비공간 데이터 우선(nonspatial data dominated) 마이닝을 병합한 방식과 다중 주제도(multiple thematic map)가 주어졌을 때의 공간 지식을 추출해 낼 수 있는 방식을 제안한다. 또한 후위기로서 공간 객체들간의 위상 관계(topological relationship)를 추론하기 위한 공간 규칙 베이스를 구축한다.

A Spatial Data Mining System Extending Generalization based on Rulebase

Seong-Min Choi† · Ung-Mo Kim††

ABSTRACT

Extraction of interesting and general knowledge from large spatial databases is an important task in the development of geographical information systems and knowledge-base systems. In this paper, we propose a spatial data mining system using generalization method; In this system, we extend an existing generalization mining and design a rulebase to support deriving new spatial knowledge. For this purpose, we propose an interleaved method which integrates spatial data dominated and nonspatial data dominated mining and construct a rulebase to extract topological relationship between spatial objects.

1. 서 론

대용량의 공간(spatial) 데이터베이스에서 사용자에게 관심있고 일반화된 지식을 추출하는 것은 지형 정보 시스템이나 지식 베이스 시스템을 개발하는데 있어서 중요하다. 인공위성이나 비디오 장비 등을 사용하는 응용 분야의 발달로 인해 데이터 양이 방대해짐으로, 공간 데이터베이스로부터 사용자가 원하는 공간적인 지식을 분석해내는 것은 비현실적이고 많은 비용이 드는 과정이다. 공간 데이터 마이닝은 이러한 문제를 해결하기 위한 방법으로, 기존의 데이터 마이닝 기법을 공간 데이터베이스에 병합한 기술이다 [1,5,6,9]. 이는 공간 데이터베이스로부터 공간과 비공간 데이터의 관계, 일반화된 공간 데이터의 특성 등과 같은 사용자

※ 본 연구는 성균학술연구에 의해 수행된 것임
† 정 회 원 : 데이콤 종합연구소
†† 정 회 원 : 성균관대학교 전기 · 전자 · 컴퓨터공학부 교수
논문접수 : 1998년 5월 7일, 심사완료 : 1998년 8월 28일

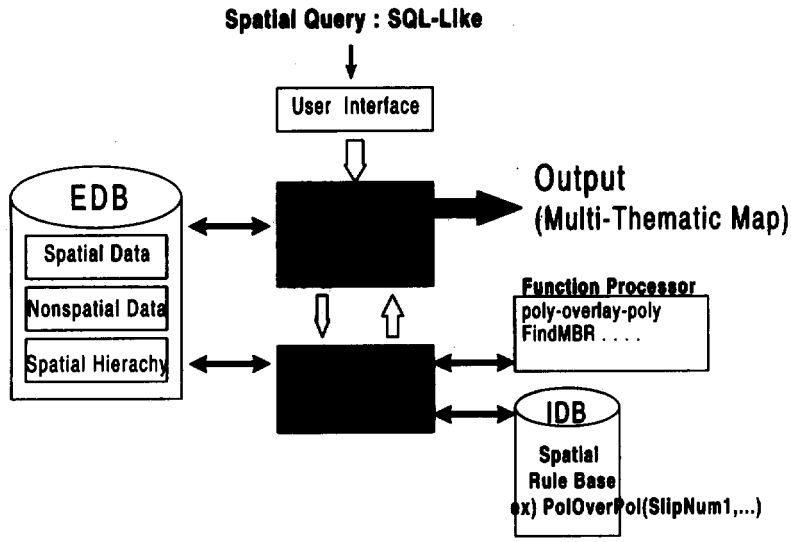
에 관심 있고, 내포적인(implicit) 공간 지식을 추출하는 과정으로 정의되고 있다.

최근까지 공간 데이터 마이닝의 기법에 관한 연구는, 공간 데이터의 일반화된 특성을 추출하는 일반화(generalization) 기법 [14], 공간 데이터간의 연관 관계를 추출하는 연관(association) 기법[11], 공간 객체들간의 유사성에 의해 그룹화하는 클러스터링(clustering) 기법[12], 통계적 방식에 근거한 기법 등으로 크게 분류할 수 있다. 이 중에서도 일반화는 대용량의 공간 객체들의 다양한 특성들을 하나의 개념으로 일반화하여 요약할 수 있는 기법으로 공간 지식 발견에 매우 중요한 기법이라 할 수 있다. 일반화는 어트리뷰트 지향 귀납법(attribute oriented induction)의 일종으로서 [9,10], 지식 발견을 위해 사용자가 사전에 설계한 개념 계층(concept hierarchy) 형태의 배경 지식(background knowledge)을 이용한다.

현재까지 제안된 일반화 기반의 공간 데이터 마이닝 시스템은 다음과 같은 문제점을 가지고 있다.
첫째, 공간 지식을 추출하는 방법에 있어서 마이닝 질의(query)의 유형이나 어플리케이션에 따라, 공간 데이터 혹은 비공간 데이터를 우선으로 일반화를 한 후에 이를 다른 데이터 형식으로 일반화시키는 방법을 사용하였다. 그러나 공간 데이터를 우선으로 하는 경우에는, 비록 공간 데이터에 정의된 한계치(threshold)까지 일반화 과정을 수행하지만 이를 처리하기 위해서는 공간 조인(join)이나 공간 합병(merge)과 같은 연산을 해야만 하므로 처리 과정이 비효율적이다[3, 14].

둘째, 일반적으로 마이닝 질의에 정의되는 주제도는 하나의 주제에 관해서만 나타내는 단일 주제도인 경우가 많이 있다. 그러나 어떤 응용에서는 마이닝 질의가 하나 이상의 주제도에 관해 일반화를 요구하는 경우가 있다. 예를 들면, "강수량과 온도에 따라 지역의 패턴을 일반화하라" 라는 질의는 강수량과 온도에 대한 두 개의 주제도가 필요하게 된다. 따라서 이러한 유형의 공간 마이닝은 사용자에게 좀 더 풍부하고 세분화된 공간 지식을 제공할 수 있게 된다[14].

셋째, 현재까지의 대부분의 지형 정보 시스템이나 공간 마이닝 시스템은 공간 데이터를 추론할 수 있는 기능이 부족하지만, 이러한 시스템에 있어서 공간 객체들간의 위상 관계를 추론해야 할 필요성이 많이 있다. 특히 공간 데이터 마이닝에서는 공간 객체 중첩(overlay) 혹은 공간 객체 합병(merging) 같은 공간 연산들이 많이 필요하다. 그런데, 이런 연산들은 기존의 관계형 데이터베이스에서 해결하기에는 많은 한계점들을 가지고 있기 때문에, 이를 규칙베이스(rulebase) 형태로 해결하면 많은 장점이 있다는 것이 지적되어 왔다.



(그림 1) 일반화 기법의 공간 마이닝 시스템의 구조
(Fig. 1) Generalization based Spatial Mining System Architecture

본 논문에서는 위의 세 가지 요소들을 통합한 공간 데이터 마이닝 시스템을 제안하고자 한다. 즉 공간 데이터 마이닝에 널리 사용되는 일반화(generalization) 기법을 확장한 공간 데이터 마이닝 모듈과 공간 데이터를 추론할 수 있도록 구축된 규칙베이스(rulebase)를 통합한 공간데이터 마이닝 시스템을 제안한다. 이를 위해 공간 데이터 우선(spatial data dominated)과 비공간 데이터 우선(nonspatial data dominated) 마이닝을 병합한 방식을 제안함으로서, 공간 데이터 우선 마이닝시에 수행되는 불필요한 공간 연산을 피할 수 있도록 하고자 한다. 또한 이를 이용하여 다중 주제도(multiple thematic map)가 주어졌을 때, 공간 지식을 추출해 낼 수 있는 방식을 제안함으로서, 다양하고 세분화된 일반화된 공간 지식을 제공할 수 있는 도구를 마련하고자 한다. 마지막으로 공간 마이닝 과정에서 필요시 되는 공간 연산을 위해, 여러 유형의 공간 객체들간의 위상 관계(topological relationship)를 추론하기 위한 공간 규칙 베이스를 구축한다. 본 논문의 구성은 다음과 같다. 제 2 장에서는 일반화 기반의 공간 마이닝 시스템의 구조 및 처리 과정에 대해서 설명하고, 제 3 장에서는 일반화를 확장하여 병합된 일반화(interleaved generalization)와 다중 주제도에 의한 일반화를 설명하고, 제 4 장에서는 공간 중첩과 같은 공간 연산을 규칙베이스로 구축한 결과를 보이고자 한다.

2. 확장된 일반화 기법의 공간 마이닝 시스템

제안하는 공간 데이터 마이닝 시스템의 구조는 (그림 1)과 같다.

위의 시스템을 구성하는 요소들은 크게 공간 마이닝의 처리 과정을 수행하는 엔진 역할을 하는 모듈과 공간 마이닝의 데이터 부분과 배경 지식을 정의하는 부분으로 정의된다. 여기서 전자는 일반화 공간 마인닝 모듈, 공간 규칙 추론 모듈, 함수 처리기 모듈로 구성되며, 후자는 외포 데이터 베이스, 공간 규칙베이스로 구성된다. 각 구성 요소에 대한 설명은 다음과 같다.

2.1 처리 과정 모듈

◆ 일반화 공간 마이닝 모듈

사용자 인터페이스로부터 제시된 SQL 형태의 마이닝 질의를 요청 받으면, 이 모듈은 질의의 유형에 따라 일반화 기반의 공간 데이터 혹은 비공간 데이터 우선의 마이닝뿐만 아니라 확장된 병합된 공간 데이터 마이닝과 다중 주제도에 의한 데이터 마이닝을 수행한다. 이때 외포 데이터베이스에 저장된 공간 개념 계층 및 관계형 테이블이나 공간 데이터 구조를 참조하면서 일반화를 하는데, 필요한 공간 연산이 있으면 공간 규칙 추론 모듈에 요청을 하게 된다.

◆ 공간 규칙 추론 모듈

공간 데이터 마이닝에서 사용되는 공간 연산들을 위한 규칙들을 처리하는 부분이다. 일반화 공간 마이닝 모듈로부터 공간 연산을 요청 받으면 공간 규칙 베이스(spatial rulebase)와 외포 데이터베이스(EDB)를 참조하면서 추론 과정을 수행하고, 공간 연산시 필요한 함수 계산은 함수 처리기에 요청을 한다. 공간 객체들간의 위상 관계를 추론한 결과는 일반화 공간 모듈에 전달되며 규칙베이스 형태로 저장한다.

◆ 함수 처리기(Function Processor)

함수 처리기는 공간 규칙 추론 모듈에서 처리하기 힘든 복잡한 일련의 공간 연산들을 함수로 구현한 것이다. 이는 어떤 함수의 계산이 요청된다면, 공간 추론 엔진으로부터 인자(argument)를 받아 해당 함수를 처리하고 참 또는 거짓의 값을 반환한다.

2.2 데이터 정의 모듈

◆ 외포 데이터베이스(Extensional Database: EDB)

외포 데이터베이스는 크게 공간 데이터, 비공간 데이터 그리고 데이터 마이닝의 실제 처리 과정에 사용되는 공간 개념 계층(spatial concept hierarchy)같은 데이터가 저장되어 있는 곳이다. 공간 데이터와 비공간 데이터의 구조는 SAND 구조[2]를 가지며, 공간 개념 계층은 사용자가 정의한 트리 형태를 갖는 배경 지식이다.

◆ 공간 규칙 베이스(Spatial Rulebase)

이는 내포 데이터베이스(intensional database)라고도 하며, 공간 중첩(spatial overlay)과 공간 합병(spatial merge)등과 같은 공간 객체들간의 위상 관계에 대한 여러 유형의 질의들을 일차 논리(first-order logic) 형태의 논리 절로 구축한 것이다.

3. 확장된 일반화 기법의 공간 마이닝 모듈

3.1 일반화의 기본 개념

공간 데이터 마이닝에서의 일반화 기법은 어트리뷰트 지향 귀납법[9,10]을 공간 데이터베이스에서의 지식 발견에 확장한 것으로서, 개념 계층과 같은 배경 지식이 사용된다. 특히 공간 데이터 마이닝에서는 기존의 데이터 마이닝에서 사용되는 개념 계층외에 공간상의 개념들을 계층적으로 표현한 공간 개념 계층이 필요시된다[8]. 이 개념 계층은 배경 지식에 사용되는 개념들을 계층적으로 표현한 것으로서 트리(tree) 형태를 가지고 있다. 여기서 상위 레벨의 개념은 하위 레벨의 개념보다 더 일반적인 개념을 나타내며 하위 레벨의 개념과 일관성을 유지해야 한다. 일반화하는 과정에서 이 계층의 경로를 따라 올라 가면서 하위 레벨의 개념을 상위 레벨의 개념으로 대치하는 귀납 과정을 하게 된다. 여기서 어느 정도의 레벨까지 일반화를 하는가를 한계치(threshold)라고 한다.

마이닝의 대상인 공간 데이터베이스의 구조는 비공간 데이터와 공간 데이터로 따로 구성되는 [2]에서 제안한 SAND 구조를 기반으로 한다. 여기서 비공간 데이터는 관계형 데이터베이스에, 공간 데이터는 공간 데이터 구조에 저장된다. 공간 객체들과 이와 관련된 비공간적인 정보는 서로 링크를 통해 연결되어있다. 일반화된 지식을 공간 데이터베이스로부터 추출하기 위해서는 공간데이터와 비공간 데이터에 대한 일반화 모두가 요구되는데, 질의 유형에 따라 어느 데이터를 먼저 일반화하는지를 결정하게 된다. 따라서 비공간 데이터와 공간 데이터 중에서 어느 것을 먼저 우선으로 일반화하는가에 따라 비공간 데이터 우선의 일반화와 공간 데이터 우선의 일반화로 분류할 수 있는데 이들 기법에서 공통적으로 적용되는 기본 과정은 다음과 같다[14].

(1) 공간 마이닝 질의에 명시된 조건을 만족하는 데이터의 수집
(2) 공간 개념 계층의 한계치까지의 일반화
(3) 공간 객체들간의 일반화된 법칙 발견

3.2 병합된 일반화

마이닝 질의의 유형이 비공간 데이터보다 공간 데이터를 우선 일반화하게 되면 이의 처리 과정이 비효율적이 된다. 예를 들면 '어떤 지역의 기후 특성을 근거로 10개 지역의 패턴으로 일반화하라' 의 경우에는 비록 이 한계치가 공간 데이터에 정의되어 있지만, 이에 대한 처리 과정은 일반적으로 공간 조인(join)이나 공간 합병(merge)같은 공간 연산들에 의존해서 공간 일반화를 해야하기 때문에 관계형 데이터베이스에서의 처리보다 비용이 더 많이 든다[3]. 이러한 이유로 비공간과 공간 데이터를 병합시켜서 일반화하는 방법이 고려되어야 하며 이에 대한 알고리즘은 다음과 같다.

알고리즘 1. 병합된 일반화

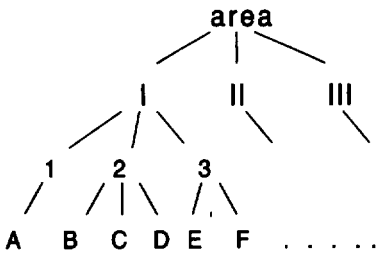
입력: (i) 비공간 데이터와 공간 데이터로 구성된 공간 데이터베이스
(ii) 공간 마이닝 질의
(iii) 공간 개념 계층과 한계치
출력: 공간 객체들간의 관계를 나타내는 일반화된 규칙
방법:
(1) 질의에 관련된 비공간 데이터와 그에 상응하는 공간 데이터를 찾는다.
(2) 공간 데이터에 대한 일반화를 하는 대신에, 공간 계층을 참조해서 일반화시킬 공간 객체들의 상위 레벨 공간 객체를 찾는다.
(2-1) 찾아서 존재하면,
1. 비공간 데이터들의 전위 링크(forward link)를 상위 레벨의 공간 객체에 연결한다.
2. 비공간 데이터의 지역을 나타내는 데이터를 상위 레벨로 일반화한다.
(2-2) 존재하지 않으면,
단지, 비공간 데이터의 지역을 나타내는 데이터를 상위 레벨로 일반화한다.
(3) (2)의 과정을 비공간 객체가 한계치가 될때까지 반복한다.
(4) 비공간 데이터에 대한 일반화를 한계치까지 한 후 전위 링크에 의해 각각의 비공간 데이터가 가리키고 있는 모든 공간 객체들을 공간 연산을 사용하여 일반화시킨다.
(5) 일반화된 법칙이나 관계성을 찾는다.

위의 알고리즘의 수행과정을 예제를 통해 설명하고자 한다.

적용 예제 1.

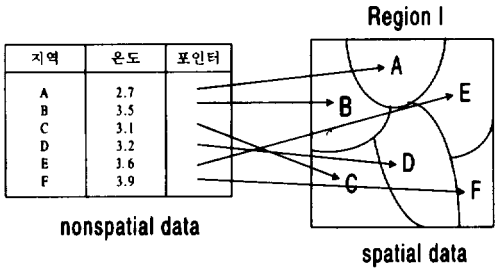
공간 마이닝 질의의 예가 "어떤 지역을 온도 패턴에 의거해서 특성을 일반화하라" 라고 하자. 이를 SQL 형태의 문장으로 표현하면 다음과 같다. 또한 이 지역에 대한 배경 지식인 공간 개념 계층은 (그림 2)와 같다.

SELECT 일반화된 규칙
FROM 온도 주제도
WHERE 지역 = 'somewhere' AND 조건
IN RELEVANCE TO 지역 AND 온도



(그림 2) 지역에 대한 공간 개념 계층
(Fig. 2) Spatial Concept hierachy of Region

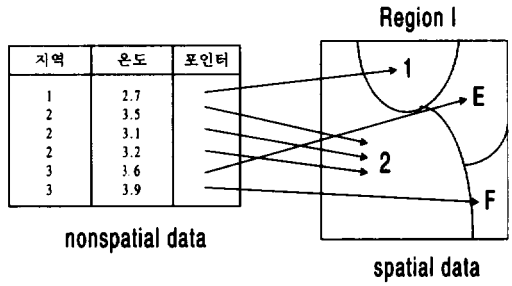
(1) 우선, 질의에 관련된 비공간 데이터를 찾고, 그에 상응하는 공간 데이터들을 찾는다. 주어진 지역에 관한 공간 계층도가 (그림 2)와 같고 찾아진 데이터들이 (그림 3)과 같다고 하자. 물론 (그림 3)의 데이터는 지역 I의 데이터만 표시한 것이지만 그 외에도 II, III 등의 여러 지역의 데이터를 찾을 수 있다.



(그림 3) 찾아진 비공간, 공간 데이터
(Fig. 3) Derived Spatial, non spatial Data

(2) (그림 2)의 I, II, III 레벨이 한계치라고 할 때 공간 데이터 우선의 일반화 경우에는 공간 데이터들을 한계치 I 까지 일반화시킨 후에 모아진 비공간 데이터들을 일반화시키는 방법을 사용하였다. 그러나, 병합된 일반화 경우에서는 공간 계층을 참조해서 그 상위 레벨 객체가 존재하는지를 우선 찾게 된다.

예를 들어, {B, C, D}에 관한 상위 레벨 공간 객체인 '2'를 공간 데이터베이스에서 찾을 수 있다면 {B,C, D}의 비공간 데이터의 지역 어트리뷰트를 '2'로 고치고, B,C,D의 포인터인 전위 링크(forward link)는 상위 레벨 개념의 공간 객체인 '2'가 존재하면 가르치게 한다. A의 경우엔 일반화된 것이 그대로이므로, 비공간 지역 어트리뷰트만 '1'로 바꿔준다. E, F의 경우 상위 레벨 개념인 '3'이 존재하지 않다고 가정하면, 비공간 데이터의 지역만 '3'으로 바꿔주며, 공간 객체는 그대로 가리킨다. 이 단계를 한 번 거친 결과의 그림은 (그림 4)와 같다.



(그림 4) 병합된 일반화 과정의 첫번째 결과
(Fig. 4) First Result of enterleaved Generalization

(3) 위의 과정을 한계치인 I까지 반복한다. 반복하면 (그림 4)의 지역 어트리뷰트는 I로 모두 바뀔 것이다.

(4) 한계치 I까지 반복했을 때, I라는 상위레벨 공간 객체가 공간 데이터베이스에 있으면 모든 전위 링크(forward link)가 이것을 가르치게 되는데, 없다면, 그대로 위와 같이 변하지 않으면서 지역 어트리뷰트만 I로 모두 바뀔 것이다. 이제, 한계치까지 반복했으므로 비공간 지역 어트리뷰트가 I인 데이터가 가르치는 모든 공간 데이터들을 전위 링크에 의해 모으고, 이것들을 공간 조인, 혹은 합병을 하면 I와 같은 영역을 얻을 수 있을 것이다.

(5) 또한, 마지막으로 비공간 지역 어트리뷰트가 I인 튜플들의 평균을 구하면, 3.33의 값이 나오며, 이것을 상위레벨 개념으로 바꾸면 "wet"과 같은 일반화된 지식을 얻을 수가 있다. 지역 I 뿐만아니라 II, III 역시 이런 과정을 거치면 일반화된 지식을 찾을 수 있게 된다.

위의 방법에 있어서 공간 계층을 참조해 상위 레벨의 공간 객체를 찾는데 많이 찾아지면 찾아질수록 공간 조인이 적게 든다.

3.3 다중 주제도를 위한 일반화

일반적으로 마이닝 질의에 정의되는 주제도는 하나 주제에 관해서만 나타내는 단일 주제도인 경우가 많이 있다. 그러나 어떤 어플리케이션들에서는 마이닝 질의가 하나 이상의 주제도에 관해 일반화를 요구하는 경우가 있다. 예를 들어, 공간 마이닝 질의가 "어떤 지역에 대해서 강수량과 온도에 따른 패턴을 찾아라" 라고 주어진다면, 이는 강수와 온도에 대한 각각의 주제도가 필요하게 된다.

두 개의 주제도를 하나의 주제도에 표현해야하는데 일반화 기반의 다중 주제도를 만들기 위해 다음과 같은 방법을 제안한다.

3.3.1 병합된 일반화를 사용하지 않은 다중 주제도

기존의 비공간, 공간 우선의 일반화를 적용시켜 아래의 방법으로 다중 주제도를 간단하게 만들 수 있다.

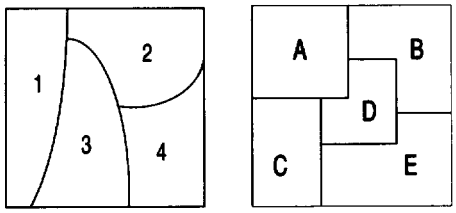
(1) 각각의 비공간 데이터들을 찾은 후에 그에 상응하는 공간 데이터들을 찾는다.
(2) 우선, 각각의 주제도를 만든다.
(3) 각 주제도를 일반화시키기 위하여 공간, 비공간 우선의 일반화를 사용하여 한계치까지 일반화시킨다.
(4) 각각의 일반화된 주제도에 있는 공간 객체들이 겹치는 지역을 찾고, 두 공간 객체의 비공간 데이터 값을 겹치는 지역의 비공간 데이터의 값으로 넣는다.

적용 예제 2

"어떤 지역의 강수량과 온도에 따른 패턴을 찾아라" 라는 다음의 질의에 있어서 병합된 일반화를 사용하지 않고 다중 주제도를 만드는 방법은 다음과 같이 적용될 수 있다.

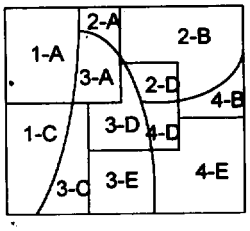
SELECT 일반화된 규칙
FROM 강우량 주제도, 온도 주제도
WHERE 지역 = 'somewhere' AND 조건
IN RELEVANCE TO 지역 AND 강수량 AND 온도

우선 강수량과 온도에 관련된 각각의 비공간 데이터를 찾고, 이에 상응하는 공간 데이터를 수집한다. 여기서 기존의 공간 혹은 비공간 데이터 우선의 방법으로 생성된, 온도와 강수량에 대한 각 주제도는 (그림 5)와 같다.



(그림 5) 온도와 강수량에 대한 일반화된 주제도
(Fig. 5) Generalized thematic map about temperature and precipitation

다음과 같이 각 주제도의 두 공간 객체가 중첩(overlay)되는 지역을 찾고, 중첩되는 지역의 비공간 어트리뷰트를 두 공간 객체의 후위 링크(reverse link)에 의해 찾은 비공간 데이터 값으로 준다. 아래의 과정을 거치면 (그림 6)와 같은 다중 주제도를 얻게 된다.



(그림 6) 온도와 강수량에 의한 다중 주제도
(Fig. 6) Multiple Thematic map about temperature and precipitation

```
for each polygygon {A, B, C, D, E} do
  for each polygon {1, 2, 3, 4} do
     call poly-overlay-poly(object i, j)
     if(there is overlay part)
     find nonspatial attribute of object i, j by
     reverse link
     make overlay part's nonspatial attribute as i-j
  end;
end.
```

위의 과정은 각각의 일반화된 주제도를 생성한 후에 다시 공간 연산을 통해서 다중 주제도를 만드는 과

정이다. 이 경우에 있어서 다음과 같은 결점이 있다.

(1) 각 주제도를 생성하기 위해 각 주제도에 대해서 비공간 혹은 공간 일반화를 수행하기 때문에 여러 번의 공간 연산들이 필요하게 된다.

(2) 각 주제도에 있는 공간 객체들의 겹치는 부분에 대한 공간 연산을 해야 한다.

위의 이유 때문에 병합된 일반화를 사용하지 않을 경우 비효율적이 된다. 그러므로, 다음과 같이 병합된 일반화를 사용하여 다중 주제도를 만드는 방법을 제안한다.

3.3.2 병합된 일반화를 사용한 다중 주제도.

병합된 일반화를 사용하지 않았을 때 생기는 비효율성을 보안하고자 병합된 일반화를 사용한다.

알고리즘 2. 병합된 일반화를 사용한 다중 주제도

입력: (i) 다중 개의 비공간 데이터와 공간 주제도로 구성된 공간 데이터베이스

(ii) 공간 마이닝 질의

(iii) 각 주제도에 대한 공간 개념 계층과 한계치

출력: 공간 객체들간의 관계를 나타내는 일반화된 규칙

방법:

(1) 공간 마이닝 질의에 명시된 조건을 만족하는 비공간 데이터와 이에 상응하는 공간 데이터를 수집한다.

(2) 각 비공간 데이터의 지역 어트리뷰트의 개념 레벨이 낮은 테이블을 높은 테이블과 같게 만든다.

/* 이 단계에서 여러 번의 공간 합병 혹은 조인 연산이 필요하게 되는데 이때, 병합된 일반화를 사용하여 보다 효율적인 공간 일반화를 할 수 있다. */

(3) 공간 일반화를 하면서 비공간 데이터에 관한 일반화도 같이 한다.

(4) 두 개의 비공간 데이터의 지역 어트리뷰트가 같은 개념 레벨로 만들었으면 두 개의 테이블 조인을 한다.

(5) 테이블 조인을 한 다음에 공간 일반화, 비공간 일반화 기법, 혹은 병합된 일반화를 사용해 원하는 한계치까지 일반화를 한다.

(6) 공간 객체들간의 일반화된 법칙이나 관계성을 출력한다.

위의 알고리즘의 수행과정을 예를 통해 설명하고자 한다.

적용 예제 3

우선 수집된 온도와 강수량에 관한 비공간 데이터에 대한 테이블이 <표 1>과 같다고 하자.

공간 개념 계층이 (그림 2)와 같다고 한다면, 온도 테이블보다 강수량 테이블이 상위 레벨 공간 객체들을 표현한다. 그러면, 온도 테이블을 강수량 테이블의 공간 객체 레벨까지 일반화시킨다. (이 예에서는 한 번만 하면 된다.) 하위 레벨의 공간 객체를 상위 레벨로 일반화시켜 표현하기 위해서는, 공간 우선 일반화를 해야하는데, 이 과정에서 병합된 일반화를 사용하면 보다 효율적이다. 공간 일반화를 하면서 비공간 데이터에 관해서도 일반화시킨다.

〈표 1〉 온도와 강수량에 대한 비공간 데이터

지역	온도	포인터
A	48.7	
B	62.5	
C	57.3	
D	59.1	
E	61.9	
F	60.8	

지역	강수량	포인터
1	1.2	
2	2.4	
3	3.5	

〈표 2〉 온도에 대한 일반화와 비공간 데이터에 대한 일반화

지역	온도 1	온도 2	포인터
1	48.7	mild	
2	59.6	moderately hot	
3	61.4	moderately hot	

지역	강수량1	강수량2	포인터
1	1.2	fair	
2	2.4	wet	
3	3.5	wet	

여기서 각 지역 어트리뷰트가 <표 2>와 같이 같은 레벨이 되었을 때, 두 개의 테이블을 조인한다. 조인한 테이블은 <표 3>과 같다. <표 3>을 보면 지역 2와 3은 같은 값을 가지면서 근접한 지역이므로 하나로 일반화될 수 있다.

〈표 3〉 온도와 강수량 테이블 조인

지역	온도	강수량	포인터
1	mild	fair	
2	moderately hot	wet	
3	moderately hot	wet	

위와 같이 2개의 주제도가 하나의 다항 주제도로 만들어 진 후에 다시 원하는 한계치까지 공간 우선의 일반화, 비공간 우선의 일반화, 병합된 일반화를 사용하여 일반화시키면 된다.

4. 공간 규칙베이스 모듈

본 공간 데이터 마이닝 시스템에서의 일반화 과정에 필요한 공간 연산들에 대한 효율적인 처리를 위해 규칙베이스를 구축하고자 한다. 이에 대한 시스템에 대한 구성 요소와 처리 과정은 2.1에서 설명했다. 이 장에서는 공간 질의에 필요한 규칙베이스를 구축하고 구축된 규칙베이스를 가지고 공간 데이터 마이닝 [illegible] 한 일반화 과정에서 질의를 처리하는 [illegible]

일반적으로 점(point), 선(line), 도형(polygon)등으로 구성되는 공간 객체들에 대해서 여러 유형의 질의가 정의될 수 있다. 예를 들면, 점과 도형간의 포함(point-containment), 도형간의 중첩(polygon-overlay-polygon), 도형간의 합병(poly[illegible] 같은 다양한 위상적 관계가 정의될 수 있다. 이 유형의 관계는 지형 정보 시스템이나 공간 마이닝 시스템에서 광범위하게 사용되는 기본적인 [illegible] 이미 이를 처리하기 위한 함수나 프로그램이 개발되어 왔다. 그러나 이들 유형의 질의를 규칙베이스로 구축 [illegible] 술한 공간 객체들 사이의 다양한 [illegible] 를 규칙베이스로 구축하였다. 본 논문에서는 지면상, 이들 중에서 (2.1에서 설명한 바와 같이) 일반화 공간 마이닝 모듈에서 많이 사용되는 공간 연산의 하나인 도형 중첩에 대한 데이터 구조 및 규칙베이스를 설명하고자 한다.

4.1 도형간의 중첩에 대한 규칙 베이스

두 공간 도형의 중첩되는 부분을 찾[illegible] (그림 7)과 같이 설명될 수 있다.

일반화 공간 마이닝 모듈에서 [illegible] 터(vector) 정보로 표현된다고 가정한다. 여기서 각 도형은 백터 정보는 동적 리스트로 저장되어 있으며, 각 도형은 노드(node)들의 연결인 체인(chain)으로 구성된다. 각 노드들은 [노드의 순서번호, X좌표, Y좌표]로 표현되며, 백터 정보 안의 점들은 시작점과 끝점을 찾

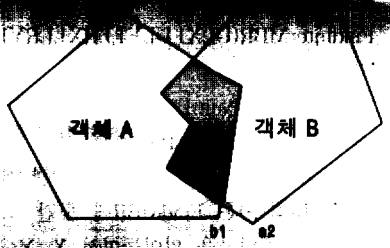


그림 7 공간 도형의 중첩되는 부분
Fig. 7) Overlay part of spatial poligon

아 놓고 시작하며, 점들은 시작점에서 시계방향으로 저장되어 있다. 발견된 중첩 영역은 InterNodeList에 [illegible] 간 마이닝 모듈에 전달된다. 이에 [illegible] 과 같다.

[illegible]gon)

입력: 두 공간 도형

출력: 두 공간 도형의 중첩 부분

방법:

[illegible] Bounded Rec-

(2) [illegible] 교차(Intersection)되는 부분을 찾

(3) 두 공간 도형에 대한 체인의 정보를 공간 데이터 구조로부터 읽어온다.

(4) 각 체인별로 MBR의 교차 부분에 들어가는

(5) 단계 (4) 에서 찾은 교차 [illegible] 보로 공간 도형의 교차 부분을 찾아, InterNodeList에 정보를 저장한다.

여기서 단계 (2)의 수행중 두 도형간의 MBR이 서로 교차하지 않으면, 두 MBR내의 공간 도형은 중첩되 [illegible] 처리 [illegible] 수행 후 [illegible] a2, b2 [illegible] 차 술어 논리(first order logic) 형태의 규칙베이스로 구축한 결과이다.

PolOverPol(SlipNum1,SlipNum2) ←
FindMBR(SlipNum1),

```
        FindMBR(SlipNum2),
        FindInterMBR(ULX1,ULY1,LRX1,LRY1,ULX2,
        ULY2,LRX2,LRY2),
        InterMBR(SlipNum1,SlipNum2,InterULX,Inter
        ULY,InterLRX,InterLRY,
      [InterNodeList]).
```

/* FindMBR(함수)은 FindMBR에 공간 도형 ID를 주면, 그 도형의 체인들을 읽어들여서 X, Y 좌표들의 최대, 최소를 구해 MBR을 만든다. */

```
findInterMBR(ULX1,ULY1,LRX1,LRY1,ULX2,ULY2,
LRX2,LRY2) ←
      findInterX(ULX1,LRX1,ULX2,LRX2,&InterULX,
      &InterLRX),
      findInterY(ULY1,LRY1,ULY2,LRY2,&Inter
      ULY,&InterLRY)
```

/* 두 MBR이 교차할 때 그 교차하는 MBR의 좌표를 찾아낸다. 여기서 findInterX(함수)는 MBR의 X좌표 부분이 교차하는 지의 여부와 교차하는 X좌표(InterULX,InterLRX)를 찾는다. findInterY(함수)는 findInterX와 같은 역할을 하나 Y좌표에 관한 함수이다. */

```
InterMBR(SlipNum1,SlipNum2,InterULX,InterULY,
InterLRX,InterLRY) ←
      findChain(SlipNum1,InterULX,InterULY,Inter
      LRX,InterLRY),
      findChain(SlipNum2,InterULX,InterULY,Inter
      LRX,InterLRY),
      InterChain(SNX1,SNX2,SNY1,SNY2,NodeCount1,
      NodeCount2,
      [NodeList1],[NodeList2],stSN2,stEN1,stEN2,[Inter
      NodeList]).
```

/* 교차하는 MBR내에서 교차하는 도형을 찾아낸다 */

```
findChain(SlipNum,InterULX,InterULY,InterLRX,Inter
LRY) ←
      chain(ChainNum,SNX,SNY,NodeCount,[Node
      List],SlipNum),
      storeChain,InterULX,InterULY,InterLRX,Inter
      LRY,SNX,SNY,
      NodeCount,[NodeList],&tSN,&tEN).
```

/* 각 Polygon의 chain중 MBR에 들어가는 부분만 찾아낸다. 여기서 storeChain(함수)에서 이 부분을 수행하면 현재의 SlipNum의 chain들 중 교차되는 MBR내에 들어있는 chain들의 첫 번째 node의 순서번호(stSN에)와 마지막 node의 순서번호(stEN에)가 저장된다. 또한 InterChain(함수)에서 이 부분을 수행하면 도형의 겹쳐진 부분의 좌표가 InterNodeList에 저장된다. */

5. 결 론

대용량의 공간(spatial) 데이터베이스에서 사용자에게 관심있고 일반화된 지식을 추출하는 것은 지형 정보 시스템이나 지식 베이스 시스템의 개발에 중요한 기법이다. 그러나 공간 데이터가 갖는 데이터 양의 방대함으로 인해, 공간 데이터베이스로부터 사용자가 원하는 공간적인 지식을 분석해내는 것은 비현실적이고 비용이 많이 드는 과정이다. 공간 데이터 마이닝은 이러한 문제를 해결하는 것으로서, 최근까지 공간 데이터 마이닝의 기법에 관한 다양한 연구가 활발히 진행되고 있다. 이 중에서도 일반화는 기법은 대용량의 공간 객체들의 다양한 특성들을 하나의 개념으로 일반화하여 요약할 수 있는 기법으로 공간 지식 발견에 매우 중요한 기법이라 할 수 있다.

이 일반화 기반의 공간 데이터 마이닝 시스템은 다음과 같은 문제점을 가지고 있다. 첫째, 공간 데이터 혹은 비공간 데이터를 우선으로 일반화를 한 후에 이를 다른 데이터 형식으로 일반화시키는 방법을 사용하였다. 그러나 공간 데이터를 우선으로 하는 경우에는, 이를 처리하기 위해서 공간 조인(join)이나 공간 합병(merge)과 같은 연산을 해야만 하므로 처리 과정이 비효율적이 된다. 둘째, 일반적으로 마이닝 질의에 정의되는 주제도는 하나의 주제에 관해서만 나타내는 단일 주제도인 경우가 많이 있다. 그러나 어떤 어플리케이션들에서는 마이닝 질의가 하나 이상의 주제도에 관해 일반화를 요구하는 경우가 있다. 따라서 이러한 유형의 공간 마이닝은 사용자에게 좀 더 풍부하고 세분화된 공간 지식을 제공할 수 있게 된다. 셋째, 현재까지의 대부분의 지형 정보 시스템이나 공간 마이닝 시스템은 공간 데이터를 추론할 수 있는 기능이 결여되어 있다. 이들 시스템들은 공간 객체들간의 위상 관계를 추론해야 할 필요성이 많이 있다. 특히 공간 데이터 마이닝에서는 공간 객체 중첩(overlay) 혹은 공간 객체 합병(merging) 같은 공간 연산들이 많이 필요시 된다.

본 논문에서는 이러한 문제점을 해결하기 위해, 공간 데이터 마이닝에 널리 사용되는 일반화(generalization) 기법을 확장한 공간 데이터마이닝 모듈과 공간 데이터를 추론할 수 있도록 구축된 규칙베이스(rule-base)를 통합한 공간 데이터마이닝 시스템을 제안했다. 이를 위해 공간 데이터 우선(spatial data dominated)과 비공간 데이터 우선(nonspatial data dominated) 마이닝을 병합한 인터리브(interleaved)된 방식을 제안함으로서, 공간 데이터 우선 마이닝시에 수행되는 불필요한 공간 연산을 피할 수 있도록 했다. 또한 이를 이용하여 다중 주제도(multiple thematic map)가 주어졌을 때의 공간 지식을 추출해 낼 수 있는 방식을 제안함으로서, 다양하고 세분화된 일반화된 공간 지식을 제공할 수 있는 도구를 마련했다. 마지막으로 공간 마이닝 과정에서 필요시 되는 공간 연산을 위해, 여러 유형의 공간 객체들간의 위상 관계(topological relationship)를 추론하기 위한 공간 규칙 베이스를 구축했다.

참 고 문 헌

[1] R. Agrawal and R. Srikant, "Fast algorithms for Mining Association Rules," In Proc. Int'l. Conf. VLDB., pp.487-499, Santiago, Chile, Sept. 1994b.

[2] W. G. Aref and H. Samet, "Optimization Strategies for Spatial Query Processing," In Proc. Int'l Conf. VLDB, pp.81-90, Barcelona, Spain, Sept. 1991.

[3] N. Beckmann, and H.-P. Kriegel, R. Schneider, and B. Seeger. "The R*-tree:An Efficient and Robust Access Method for Point and Rectangles," In Proc. of 1990 ACM-SIGMOD, pp.322-331, Atlantic City, USA, May 1990.

[4] T. Brinkhoff, H.-P. Kriegel, and B. Seeger. "Efficient Processing of Spatial Joins using R-trees," In Proc. ACM-SIGMOD, pp.237-246, Washington, D.C., May 1993.

[5] M. Ester, H. -P. Kriegel, and X. Xu. "Knowledge discovery in large spatial databases: Focusing Techniques for Efficient Class Identification," In Proc. 4th Int'l. Symp. on Large Spatial Databases (SSD'95), pp.67-82, Portland, Maine, August 1995.

[6] U. M. Fayyad and G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, editors. 'Advances in knowledge Discovery and Data Mining,' AAAI/MIT Press, Menlo Park, CA, 1996.

[7] R. Guttman, "A dynamic index structure for spatial searching," In Proc. ACM SIGMOD, pp. 47-57, Boston, MA, 1984.

[8] J. Han and Y. Fu, "Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases," In Proc. AAAI 94 Workshop on Knowledge in Databases, pp.420-431, Zurich, Swizterland, Sep., 1995.

[9] J. Han, and Y. Fu. "Exploration of the Power of Attribute-Oriented Induction in Data mining," In [6], pp.421-440, 1996.

[10] J. Han, Y. Cai, and Nick Cercone, "Knowledge Discovery in Databases : An Attribute-Oriented Approach," Proceedings of the 18th VLDB Conference. Vancouver, British Columbia, Canada 1992.

[11] K. Koperski, J. Han, "Discovery of Spatial Association Rules in Geographic Information Databases," In Proc. 4th Int'l Symp. on Large Spatial Databases, pp.47-66, Portland, Maine, August 1995.

[12] R, Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Mining," In Proc. 1994 Int'l. Conf. Very Large Data Bases, pp.144-155, Santiago, Chile, Sep. 1994.

[13] B. C. Ooi. "Efficient Query processing for Geographic Information Systems," PhD thesis, Monash University, Victoria, Australia, (1988). (Lecture Notes in Computer Science 471, Springer-Verlag, Berlin, 1990).

[14] W. Lu, J. Han, and B.C. Ooi, "Discovery of General Knowledge in Large Spatial Databases," In Proc. Far East Workshop on Geographic Information Systems, pp.271-289, Singapore, June 1993.

최 성 민

cherish@halla.dacom.co.kr

1996년 성균관대학교 정보공학과 졸업(학사)

1998년 성균관대학교 대학원 정보공학과 졸업(공학석사)

1998년~현재 데이콤 종합연구소 근무

관심분야 : 데이터마이닝, 지형정보시스템, 객체지향데이터베이스, 네트웍 관리 시스템

김 응 모

umkim@yurim.skku.ac.kr

1981년 성균관대학교 수학과(학사)

1986년 Old Dominion University 컴퓨터과학과(석사)

1990년 Northwestern University 컴퓨터과학과(박사)

1997년 8월~1998년 7월 Univ. of Cal., Irvine, 방문연구원

현재 성균관대학교 전기·전자·컴퓨터공학부 부교수

관심분야 : 지형정보시스템, 데이터마이닝, 객체지향 데이터베이스