2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

확장된 공간 연관 규칙 탐사기법

하 단 심* 황 부 현*
*전남대학교 전산학과
E-mail : {dsha, bhhwang}@sunny.chonnam.ac.kr

Extended Method of Discovery of Spatial Association Rules

Danshim Ha* Buhyun Hwang*
*Dept. of Computer Science, Chonnam National University

요 약

공간 데이터가 증가함에 따라 이를 효율적으로 저장하고 분석할 수 있는 기술이 필요하게 되었다. 공간 데이터 마이닝은 데이터베이스에서 유용한 지식을 추출하는 기술로, 기존의 데이터 마이닝 방법에 공간의 개념을 추가하여 확장함으로써 공간 패턴, 공간 객체들의 연관 관계 등을 얻을 수 있다.

본 논문에서는 공간 데이터 마이닝의 기법 중의 하나인 공간 연관 규칙 탐사 기법을 제안 한다. 제안하는 방법은 공간 관계를 포함한 공간 연관 규칙 뿐만 아니라 공간 객체의 비공간 속성도 함께 고려함으로써 보다 확장되고 다양한 공간 연관 규칙을 탐사할 수 있다.

1. 서 론

데이터 마이닝은 대용량의 데이터베이스 구축이 일반화되고 컴퓨팅 능력의 향상으로 인하여 발달하고 있다. 데이터 마이닝이란 데이터베이스에 존재하는 데이터를 분석하여 드러나지 않은 유용한 지식을 발견하는 과정이다[1]. 공간 데이터 마이닝은 기존의 데이터 마이닝에 공간의 개념을 추가하여 확장한 것으로, 공간 데이터에 대한 지식 탐사 과정으로 정의할 수 있다. 그리고 공간 데이터 마이닝을 통해 공간 패턴, 공간 객체간의 연관 관계 등을 얻을 수 있다[2,3]. 공간 데이터 마이닝은 GIS(Geographical Information System), 의료 영상 기기 등의 응용에 사용된다[2,3,8].

본 논문에서는 기존의 공간 관계만을 포함한 연관 규칙 탐사 방법을 확장하여 데이터의 비공간 속성이 추가된 확장된 공간 연관 규칙 탐사 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 연관 규칙과 공간 데이터 마이닝에 대하여 기술하고, 3장에서는 본 논문에서 제안하는 비공간 속성을 함께 고려한 확장된 공간 연관 규칙 탐사 방법을 기술한다. 끝으로 4장에서 결론과 향후 연구방향을 기술한다.

※ 이 논문은 한국과학재단 1999년도 특정기초연구비 (1999-2-303-006-3) 지원에 의하여 연구되었음.

2. 관련 연구

2.1 연관 규칙

연관 규칙은 사건의 동시 발생에 대한 규칙으로써 한 항목들의 그룹과 다른 항목들의 그룹 사이의 연관성에 대한 정보를 담고 있다[4,5]. $X \cap Y = \emptyset$의 관계에 있는 두 데이터 항목 집합 X, Y에 대하여 연관 규칙이 존재한다면 $X \rightarrow Y(c\%)$로 표현하며, 이는 X를 포함하는 트랜잭션이 Y도 동시에 포함할 확률이 c%인 규칙임을 의미한다.

데이터 항목 집합 X, Y에 대해 X의 지지도 sup(X)는 전체 트랜잭션에서 X가 차지하는 비율이며, 규칙 $X \rightarrow Y$의 신뢰도 conf($X \rightarrow Y$)는 X를 포함하는 트랜잭션 중에서 Y가 함께 발생하는 트랜잭션의 비율을 의미한다[4,5,6].

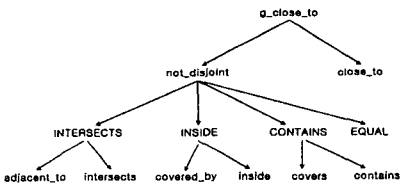
2.2 공간 데이터 마이닝

공간 데이터베이스는 공간 데이터와 비공간 데이터로 구성이 되어 있다. 공간 데이터는 특정한 공간에 위치하는 객체와 관련된 데이터로써 거리, 위상에 대한 정보를 담고 있다[2,3,8]. 비공간 데이터는 공간 데이터를 설명해 주는 데이터이다. 실세계의 공간 객체는 "위치"와 같은 공간 데이터와 "이름","전화번호"와 같은 비공간 데이터가 결합하여 표현된다[2].

2.3 공간 연관 규칙 탐사

공간 연관 규칙은 기존의 연관 규칙을 공간 데이터 마이닝에서도 사용할 수 있도록 확장한 것이다[7]. 공간 연관 규칙은 X→Y로 표현되며, X,Y는 adjacent_to(인접), intersects(교차)와 같은 공간 관계를 설명하는 공간 관계 술어가 포함된 조건식의 집합으로 정의된다[7]. 공간 연관 규칙은 기존의 연관 규칙 탐사 방법과 유사하게 지지도와 신뢰도를 이용하여 탐사되며 탐사 결과로 공간 데이터와 공간 데이터 간의 공간 관계에 대한 공간 연관 규칙을 얻을 수 있다. 예를 들어 「is_a(x, gas-station) →close_to(x, highway) (75%)」라는 규칙은 gas-station이 highway와 근접할 확률이 75%라는 공간 연관 규칙이다.

공간 연관 규칙 탐사 기법은 공간 데이터의 계층도, 공간 관계 술어들의 계층도와 MBR(Minimum Bounding Rectangle)과 같은 근사치 연산을 이용하여 지지도와 신뢰도를 적용하여 탐사된다[2,3,7]. 공간 관계 술어에 대한 계층도의 예는 그림 1과 같으며 공간 데이터 Town, Road, Water에 대한 계층도의 예는 표 1과 같다.



(그림 1) 공간 관계 술어(g_close_to)의 계층도

- town의 계층도

  (town(large_town(big_city,medium_sized_city), small_town(...), ...) ...)

  water의 계층도

  (water(sea(Strait(...), river(large_river(Fraser_river, ...), ...), lake(large_lake(Okanagan_Lake, ...), ...), ...), ...)

- road의 계층도

  (road(national_highway(route1, ...),
  provincial_highway(highway7,...),
  city_drive(Hasting ST., Kingsway, ...),
  city_street(E_1st Ave, ...), ...), ...), ...)

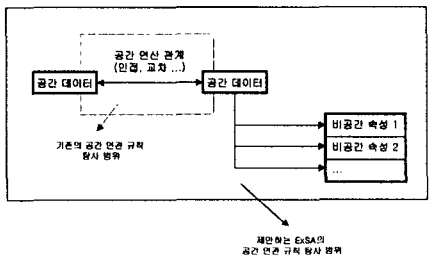
<표 1> 공간 데이터 관계의 개념 계층도

그러나 기존의 공간 연관 규칙 탐사 방법은 데이터의 공간 속성만을 고려하여 공간 데이터와 공간 데이터 사이에 존재하는 공간 관계간의 공간 연관 규칙만을 탐사한다. 따라서 데이터의 비공간 속성은 고려하지 않음으로써 효율적인 공간 연관 규칙을 탐사할 수 없다는 문제점이 있다.

3. 확장된 공간 연관 규칙 탐사 기법

3.1 ExSA(EXtended Spatial Association Rule)

기존의 공간 연관 규칙 탐사 방법은 공간 데이터와 공간 데이터 사이의 공간 관계의 연관성을 탐사할 수 있다. 그러나 기존의 공간 연관 규칙 탐사 방법으로는 비공간 데이터가 포함된 공간 연관 규칙을 탐사할 수 없다. 본 논문에서는 기존의 공간 연관 규칙 탐사 방법을 보다 확장하여 공간 데이터들의 비공간 속성과 공간 데이터 사이의 공간 관계간의 연관성을 함께 탐사할 수 있는 확장된 공간 연관 규칙 탐사 기법인 ExSA를 제안한다. 그림 2는 ExSA가 탐사하는 범위이다.



(그림 2) ExSA의 탐사 범위

ExSA에서는 비공간 속성에 대한 계층도가 존재하는 경우 그들의 계층도에 따라 확장시키면서 규칙을 탐사한다. 예를 들어 강수량과 같은 비공간 속성에 대한 계층도가 존재한다면 표 2와 같이 표현될 수 있다.

very dry	dry	moderately dry	fair
[0, 0.1]	(0.1, 0.3]	(0.3, 1.0]	(1.0, 1.2]
moderately wet	wet	very wet	
(1.2, 2.0]	(2.0, 5.0]	5.0 & up	

<표 2> 강수량 데이터에 대한 계층도

3.2 ExSA의 전개 예

예 1은 제안하는 ExSA를 이용하여 각 단계별로 k개의 항목으로 구성된 k-빈발항목집합을 구하는 과정이다.

사용자는 Region_Name이라는 지역안에서 Town과 g_close_to 관계를 갖는 Road, Water에 대해서 Town의 강수량에 관계된 공간 연관 규칙을 탐사하고자 한다. 탐사할 Road, Water의 유형은 type 속성에서 지정한다. 질의는 <질의 1> 다음과 같다.

<질의 1> 확장된 공간 연관 규칙 탐사 질의 예
discover extended spatial association rules
inside Region_Name
from Road R, water W
in relevance to Town T, Town.precipitation Town_p

where g_close_to(T.geo, X.geo) and X in {R,W}
and T.type = large_town and R.type = national_highway
and W.type in {Sea, River, Lake}
and Town_p in {dry, fair, wet}

<예 1>

사용될 Road, Water의 type과 Town과의 관계는 그림 3과 같다.

- W1은 T1, T2, T5와 close_to, T3와는 adjacent_to 한다
- W2는 T2와 close_to 하고 T3,T4,T5와는 adjacent_to 한다
- W3는 T3, T4와 close_to 하다
- R1은 T1, T2, T3와 close_to 하고 T5와 intersect 한다
- R2는 T1, T2, T4와 intersect 하며 T3, T5와는 close_to 한다
- W1, W2는 River, W3은 lake이고 R1, R2는 national highway이다

(그림 3) Road, Water의 type과 Town과의 공간 관계

질의를 통해 최초에 수집되는 데이터는 표 3과 같다. 최소 지지도를 40%로 지정하고 표 3에서 그림 3의 Town과의 공간 관계에 따라 adjacent_to, close_to, intersect 등의 하위 단계로 확장한 결과는 표 4와 같다.

Town	Water	Road	Precipitation
T1	W1	R1, R2	0.5
T2	W1, W2	R1, R2	1.1
T3	W1, W2, W3	R2	0.6
T4	W2, W3	R2	0.7
T5	W1, W2	R1, R2	1.1

<표 3> 수집된 작업 관련 데이터

Town	Water	Road	Precipitation
T1	<close_to, water>	<close_to, road> <intersect, road>	wet
T2	<close_to, water>	<close_to, road> <intersect, road>	fair
T3	<adjacent_to,water> <close_to, water>	<intersect, road>	wet
T4	<adjacent_to,water> <close_to, water>	<intersect, road>	wet
T5	<close_to, water> <adjacent_to,water>	<close_to, road>	fair

<표 4> g_close_to 연산의 확장

첫 번째 단계는 Town의 비공간 속성인 강수량에 따라 <공간 관계 술어, 공간 데이터>를 묶어 그들의 지지도를 구한다. 예 1에서 생길 수 있는 1-후보항목집합은 표 5와 같으며, 표 5에서 지지도가 2이상인 데이터들로 구성된 집합이 최소 지지도를 만족하는 1- 빈발항목집합이 된다.

precipitation	1-후보항목집합	support
Town_p=wet	<close_to, water>	3
Town_p=fair	<close_to, water>	2
Town_p=wet	<adjacent_to, water>	2
Town_p=fair	<adjacent_to, water>	1
Town_p=wet	<close_to, road>	1
Town_p=fair	<close_to, road>	2
Town_p=wet	<intersect, road>	3
Town_p=fair	<intersect, road>	1

<표 5> 1단계에서의 1-후보항목집합

다음 단계에서는 표 5에서 생성된 빈발항목집합들의 join을 통해 원소가 2개인 2-후보항목집합을 구성한다. 구성 방법은 Town의 강수량에 대해 동일한 값을 갖는 것 끼리 join을 통해 구성하며 2-후보항목집합은 표 6과 같다.

precipitation	2-후보항목집합	support
Town_p=wet	<close_to,water> <adjacent_to,water>	2
Town_p=wet	<close_to, water> <intersect, road>	3
Town_p=wet	<adjacent_to, water> <intersect, road>	2
Town_p=fair	<close_to, water> <close_to, road>	2

<표 6> 1단계에서의 2-후보항목집합

표 6에서 3-후보항목집합을 생성하여 최소 지지도를 만족하는 3-빈발항목집합을 생성한다. 이러한 과정은 더 이상의 후보항목집합을 생성할 수 없을 때 까지 반복하며, 표 7은 첫 번째 단계에서 생성된 최종의 빈발항목집합이다.

precipitation	3-후보항목집합	support
Town_p=wet	<close_to,water> <adjacent_to,water> <intersect, road>	2

<표 7> 1단계에서의 최종 빈발항목집합

공간 관계에 대한 하위 단계로의 확장을 통해 빈발항목집합을 구성하는 첫 번째 단계가 끝나면 다음 단계로 공간 데이터의 하위 단계로의 확장을 통해 빈발항목집합을 탐사해 나간다. 생성될 수 있는 1- 후보 항목 집합은 표 8과 같다. 후보 항목 구성은 앞단계와 동일한 방법으로 구성되며, 같은 비공간 속성을 갖는 항목들끼리 묶어 구성한다. 표 9는 2단계에서 생긴 최종의 빈발항목집합이다.

precipitation	1-빈발항목집합	support
Town_p=wet	<close_to, river>	2
Town_p=wet	<close_to, lake>	2
Town_p=fair	<close_to, river>	2
Town_p=wet	<adjacent_to, river>	2
Town_p=fair	<close_to, national highway>	2
Town_p=wet	<intersect, national highway>	3

<표 8> 2단계에서의 1-빈발항목집합

precipitation	3-빈발항목집합	support
Town_p=wet	<adjacent_to, river> <close_to,lake> <intersect, national highway>	2

<표 9> 2단계에서의 최종 빈발항목집합

마지막으로 각 단계마다 생성된 최종 빈발항목집합들을 이용하여 규칙을 구성하며, 최소 신뢰도를 이용하여 비공간 속성을 포함한 확장된 공간 연관 규칙을 얻게 된다. 예 1에서는 「강에 근접하거나 호수에 가깝다면 Town의 강수량은 많다(wet) (100%)」와 같은 비공간 속성까지 확장된 공간 연관 규칙을 발견할 수 있다.□

제안하는 ExSA는 비공간 속성에 대하여 같은 값을 갖는 항목들로 후보항목집합을 구성하고 사용자가 정의한 최소 지지도를 이용하여 빈발항목집합을 구성한다. 공간 관계 술어나 공간 데이터, 비공간 속성에 계층도가 존재한다면 각각의 단계별로 반복하여 비공간 속성을 포함한 확장된 공간 연관 규칙을 탐사한다.

제안하는 ExSA의 알고리즘은 알고리즘1과 같다.

알고리즘 1 : ExSA Algorithm

```
for(level=1; level<max_level or EL[level,1] ≠ 0; level ++ )
  EL[1,1] ={ 최상위 단계 모든 데이터의 support 중에서 minsup를 만족하는 데이터 집합}
  for(k=2; EL[level, k-1] ≠ 0; k++)
    Ck = get_candidate_merging_nonspatial_attr(EL[level, k-1]);
    forall spatial object s ∈ SDB do begin
      Ct = subset(Ck , s);
      forall candidates c ∈ Ct do begin
        c.count++;
      end
    end
    EL[level, k] = {c∈Ck | c.count ≥ minsup}
  end
  LL[level] = UkEL[level,k];
  result = generate_association_rule(LL[level], minconf);
end
```

※ C_k : 후보 항목 집합 저장소
※ get_candidate_merging_nonspatial_attr : 같은 비공간 속성을 갖는 값들끼리 합병하여 후보항목 생성하는 함수
※ subset : 각 후보항목이 트랜잭션에 존재하는지 조사
※ k-itemset : 항목수가 k 개인 집합
※ EL[level, k] : level 단계에서의 빈발 k-itemset
※ LL[level, k] : 각 level마다의 빈발항목집합을 저장

4. 결 론

본 논문에서는 공간 데이터간의 공간 관계에 대한 연관성만을 탐사했던 기존의 공간 연관 규칙을 확장하여 공간 객체의 비공간 속성이 추가된 공간 관계간의 연관 규칙을 탐사할 수 있는 ExSA를 제안하였다.

향후 연구 방향으로는 공간 데이터의 속성을 보다 잘 살릴 수 있는 공간 연관 규칙 탐사 기법의 연구 등이 필요하다.

참고문헌

[1] Ian H.Witten, Eibe Frank "Data Mining: Practical Machine learning Tools and Techniques with Java implementation" Morgan Kaufuman

[2] 오병우, 박지웅, 한기준 "GIS 데이터베이스를 위한 공간 데이터 마이닝" 98 데이터베이스 연구회지 1998

[3] Krzysztof Koperski, Junas Adhikary, Jiawei Han . "Spatial Data Mining: Progress and Challenges Survey Paper" SIGMOD'96 Workshop. on Research Issues on Data Mining and Knowledge Discovery (DMKD'96) 1996.

[4] 박종수, 유원경, 홍기형 " 연관 규칙 탐사와 그 응용" 정보과학회지 1998

[5] Rakesh Agrawal, Ramakrishnan Srikant "Fast Algortihms for Mining Association Rules" Proceedings of VLDB conference. 1994

[6] Rakesh Agrawal, Tomasz Imielinski, Arun Swami. "Mining Association Rules between Sets of Items in Large Database" Proceeding of ACM SIGMOD USA. 1993

[7] Krzystof Koperski, Jiawey Han "Discovery of spatial assocation rules in Geographic Information Databases" Advances in Spatial Databases. Proceedings of 4th Symposium, SSD'95. 1995

[8] Wei Lu, Jiawey Han, Beng Chin Ooi "Discovery of General Knowlege in Large Spatial Databases" Proceedings of Far East Workshop on Geographic Information Systems. Singapore. 1993