확장된 질의를 갖는 공간 연관 규칙의 설계 및 구현

이윤석*, 안찬민, 박상호, 이주홍
*인하대학교 컴퓨터정보공학과
e-mail:{aprilia*, ahnch1, parksangho}@datamining.inha.ac.kr,
juhong@inha.ac.kr

Design and Implementation of Spatial Association rule with Extended query

Yun Seok Lee*, Chan Min Ahn, Sang Ho Park, Ju Hong Lee
*Department of Computer Science & Engineering, Inha University

요 약

공간 연관 규칙은 공간적 특성을 바탕으로 발견되는 연관 정도를 나타낸다. 그러나 일반적인 공간 데이터베이스에서는 공간 연관 규칙을 발견하기 어렵다. 따라서 본 논문에서는 공간 데이터베이스에서 사용되는 질의를 확장하여 공간 연관 규칙을 찾는 방법을 제안한다. 본 논문에서 제안하는 시스템은 위상 정보에 따른 데이터를 구성한 후 비공간 객체 속성간의 연관규칙을 발견한다.

1. 서론

최근 공간 정보들을 효과적으로 이용할 수 있는 기술에 대한 연구가 활발하게 이루어지고 있다. 이러한 기술들은 지리 정보 시스템이나 지식 베이스 시스템에 유용하게 사용될 수 있다. 그러나 방대한 양의 공간 데이터로부터 사용자가 원하는 정보를 추출하기 위해서는 많은 비용을 필요로 한다. 이러한 문제를 해결하기 위해서 공간 데이터마이닝 기법들을 이용한다. 공간 데이터마이닝은 공간 데이터베이스에 저장된 자료들로부터 공간적 속성이 고려된 유용한 지식들을 발견하는 것이다[2].

우리가 흔히 이용하고 있는 공간 데이터베이스 시스템은 기존의 관계형 데이터베이스를 확장한 객체 지향적인 자료 구조를 가지고 있다. 그러나 객체의 위상 정보 및 그 속성과 같은 비공간 정보를 가지고 있을 뿐 공간 데이터마이닝을 효율적으로 지원하지 못한다. 따라서 본 논문에서는 공간 데이터베이스에서 사용되는 질의를 확장하여 공간 데이터마이닝 방법 중 공간 연관 규칙을 찾는 방법을 제안한다. 특히 지리 정보 시스템에 적용 가능한 모델을 구현하였다.

본 논문의 구성은 다음과 같다. 2절에서 기존 공간 연관 규칙 시스템들의 종류를 소개하고 3절에서 우리가 제안하는 방법에 필요한 요소와 질의 처리 과정 및 결과를 보인다. 마지막으로 4절에서 결론을 맺는다.

2. 관련연구

J.Han이 제안한 GeoMiner[2]에서 공간 연관 규칙을 적용한 Koperski의 방법[1] 이후 공간 연관규칙 탐사 방법에 대한 많은 연구가 이루어지고 있고, 일부는 실용화되고 있다. D.Malerba는 지리 정보 시스템과 데이터마이닝 서버를 접목한 INGENS 시스템을 발표하였다[4]. Osmar R. Zaiane은 비쥬얼 컨텐츠 데이터베이스에서 정보를 찾기 위한 MultiMediaMiner 시스템을 제안하였다[5]. 위 시스템들은 거리에 대한 Hierarchy를 구성하여 Attribute의 하나로 간주하여 공간에 대한 연관 규칙을 발견한다. 그러나 각 시스템의 질의 언어 및 최종 결과를 보면 공간 속성에 대한 분류는 건물,

1)본 연구는 대학 IT연구센터 육성・지원사업의 연구결과로 수행되었음

도로, 산, 강과 같은 큰 범주의 분류에 그칠 뿐 각 건물에 대한 속성이나 도로의 교통량과 같은 세부 속성은 나타나지 않는다.

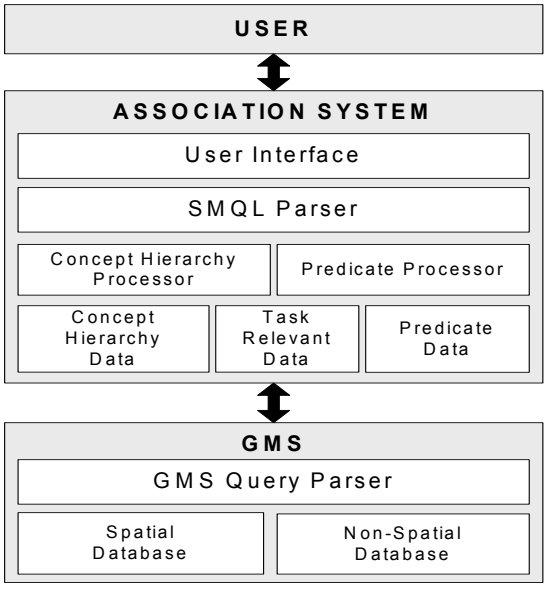
공간 속성과 비공간 속성을 동시에 고려한 시스템으로 M. May가 제안한 SPIN! System이 있다[6]. 그러나 공간 연관 규칙을 탐사 할 때는 앞서 제시된 다른 연구들과 큰 차이점을 보이지 않는다. 이러한 공간 연관 규칙의 발견은 의료 분야나 영상 처리 분야에 응용할 수 있지만 우리가 실생활에서 발견하고자 하는 공간 연관 규칙은 객체들 간의 단순한 지리적 위상 관계보다 각 객체의 속성들도 고려한 종합적인 연관 규칙탐사 방법이 요구된다.

따라서 본 논문에서는 공간 데이터베이스의 질의를 확장하여 공간 데이터마이닝, 특히 공간 연관 규칙을 찾는 방법을 제안한다. 본 논문에서 제안하는 시스템은 위상 정보에 따른 데이터를 구성한 후 비공간 객체 속성간의 연관규칙을 발견한다.

3. 설계 및 구현

본 논문에서 제안하는 시스템은 GMS를 기반으로 SMQL[3]을 사용하여 연관 규칙을 찾아내는 시스템이다. 구성은 그림 1과 같다.

GMS는 공간 DBMS로서 다중 사용자를 지원하며 공간 데이터 및 비공간 데이터를 효율적으로 저장, 관리할 수 있다.



(그림 1) 공간 연관 규칙 시스템

3.1 SMQL

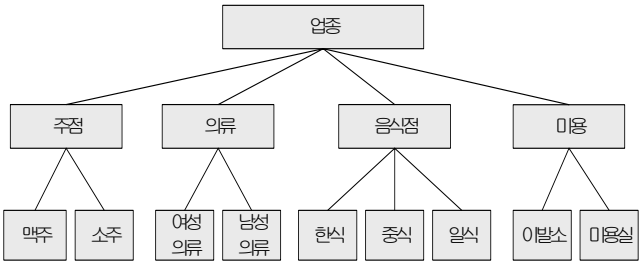
SMQL은 SIMS(Spatial Information Management System)에서 사용하기 위한 공간 데이터마이닝 질의 언어이다[3]. SMQL은 공간 데이터 마이닝 기법 중에서 Association, Classification, Clustering, Trend Analysis를 지원하는 질의 언어이며, Association의 BNF는 그림 2와 같다.

```
SMQL_query ::= ASSOCIATION [AS pattern_literal] |
        RETRIEVE TASK RELEVANT DATA AS task_name
        [ USING HIERARCHY  hierarchy_description ]
        [ USING TASK RELEVANT DATA task_name ]
        [ FOR analysis_standards ]
        FROM table_list
        [ WHERE conditions ]
        [SET threshold_specification ]
task_name ::= literal
hierarchy_description ::=
     {hierarchy_level OF} hierarchy_name
                                FOR spatial_term
threshold_specification ::=
      threshold_description THRESHOLD number
threshold_description ::=
      DISTINCT_VALUE | DESIREABLE_RULE
```

(그림 2) SMQL에서의 Association BNF

3.2 계층 구조 정의

공간 속성을 연관 규칙에 효율적으로 이용하기 위해 계층적 구조로 정리할 필요가 있으며, 이를 공간 hierarchy라고 한다. 소득이나 나이와 같은 비공간 속성도 공간 속성과 마찬가지로 계층적 구조로 표현할 수 있다. 그림 3은 계층 구조의 예를 보여준다.



(그림 3) 계층 구조의 예

이러한 공간, 비공간 hierarchy를 공간 연관 규칙 시스템 상에서 정의, 이용하기 위해서는 질의 언어를 이용한다. SMQL에서 hierarchy를 정의, 이용하는 구문의 BNF는 그림 4와 같다. hierarchy_name은

생성할 hierarchy의 이름을 말한다. apply_attribute는 어떠한 속성을 대상으로 hierarchy를 생성할 것인지를 정의한다. apply_table은 apply_attribute가 속한 테이블을 말하고, AS 이하 구문은 하위 hierarchy의 값 OF 상위 hierarchy의 값 형식으로 나타낸다.

```
Hierarchy_Generation ::=
        GENERATE HIERARCHY hierarchy_name
        FOR apply_attribute
        IN apply_table
        [AS level_number = (hierarchy_value)
                OF (hierarchy_parent | ROOT)]
hierarchy_name ::= literal      apply_attribute ::= literal
level_number ::= literal        apply_table ::= literal
hierarchy_value ::= literal | range
range ::= value-value
hierarchy_parent ::= hierarchy_value
```

(그림 4) Hierarchy 생성을 위한 BNF

3.3 질의 처리 과정 및 결과

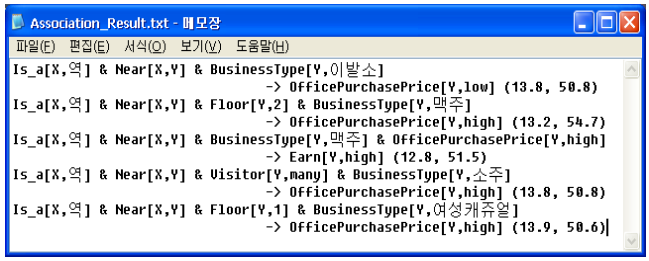
GMS에서의 공간 연관 규칙 탐사 시스템에서는 GMS에서 얻어낸 데이터를 마이닝하기 위한 SMQL과 기존 GMS의 질의를 조합하여 연관 규칙을 찾아낸다. 예제를 통해 질의 처리 과정을 살펴보면 다음과 같다. 질의는 역을 중심으로 500미터 이내의 범위에서 공간 연관 규칙을 찾도록 하였다. 이를 SMQL로 작성하면 그림 5와 같다.

```
MINE ASSOCIATION AS Station
FOR T2.*
FROM Building_T1, Building_T2
WHERE T1.BusinessType = "역"
                AND DISTANCE(T1,T2,500)
SET SUPPORT THRESHOLD 0.3 AND
CONFIDENCE THRESHOLD 0.7;
```

(그림 5) 질의 예제

우선 기준이 될 OID를 찾는다. 질의에서 Building_T1 테이블의 BusinessType="역"인 모든 OID가 검색된다. 두 번째 단계에서는 "어떻게 연관된" 정보를 찾을 것인지를 구한다. 질의에서 보면 역에서 거리가 500 이내의 범위에 속한 모든 객체들의 OID가 검색된다. 검색된 OID에서 중복된 값을 제외시키고 저장한다. 세 번째 단계에서는 이렇게 얻은 OID의 나머지 속성 데이터들을 구한다. 만약 질의에 Using hierarchy 구문이 명시되어 있으면, 구해진 데이터 집합은 개념 계층을 이용하여 구문에 지정된 level의 값으로 대체된다. 네 번째 단계에서 Apriori 알고리즘을 적용하여 연관 규칙을 발견한다.

질의 수행 결과는 그림 6과 같다. 그러나 찾아낸 규칙에서 support의 비율이 낮다. 좀 더 정확한 정보를 찾기 위해 BusinessTree의 hierarchy를 이용하여 연관 규칙을 찾는다.

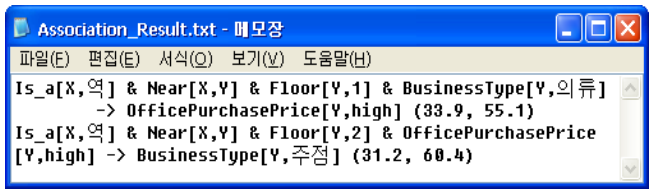


(그림 6) 질의 결과

Hierarchy를 이용한 질의는 그림 7과 같다. 질의 처리 과정 중 세 번째 단계에서 BusinessTree 라는 hierarchy 테이블을 이용하여 BuniessType의 값을 level1의 값으로 대체시킨다. 그림 8은 그 결과를 나타낸다.

```
MINE ASSOCIATION AS Station
USING HIERARCHY lv1 of BusinessTree
                        FOR T2.BusinessType
FOR T2.*    FROM Building_T1, Building_T2
WHERE T1.BusinessType = "역"
                        AND DISTANCE(T1,T2,500)
SET SUPPORT THRESHOLD 0.3 AND
CONFIDENCE THRESHOLD 0.5;
```

(그림 7) Hierarchy를 이용한 질의 예제



(그림 8) 질의 결과

발견된 규칙을 풀이하면 다음과 같다. 역 부근의 1층에 있는 의류점은 평당 가격이 높고 support는 33.9%, confidence는 55.1% 이다. 역 부근의 2층에 위치한 업종 중 평당 가격이 높은 곳은 주점인 경우가 많고 support는 31.2%, confidence는 60.4% 이다.

이를 종합하여 볼 때, 대체로 역 부근에서는 주점과 의류점이 호황이라는 것을 알 수 있다.

동일한 조건절을 가진 질의를 반복해서 수행해야 하는 경우 작업관련 데이터를 사용하여 처리시간을 단축시킬 수 있다. 그림 9는 인구가 많은 건물 근처의 건물들에 대한 연관 규칙을 찾는 질의 예제이다.

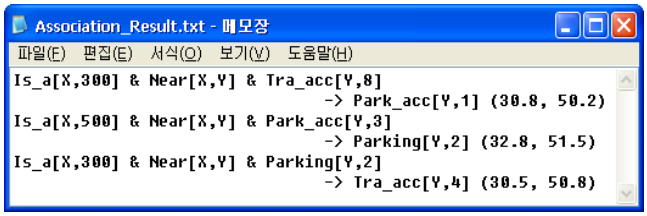
```
RETRIEVE TASK RELEVANT DATA AS pop
FOR T2.attr, T2.age, T2.price, T2.pop
FROM Building_T1, Building_T2
WHERE T1.pop > 200 AND T2.oid = T1.oid
                    AND DISTANCE(T1,T2,100)
SET SUPPORT THRESHOLD 0.3 AND
CONFIDENCE THRESHOLD 0.5;
```

(그림 9) 질의 예제

이 경우 질의 처리 과정 중 두 번째 단계까지 검색된 OID는 pop 파일로 저장된다. 질의 결과 규칙이 발견되지 않아 그림 9의 질의를 인구가 많은 건물 근처의 도로들에 대한 연관 규칙을 찾는 것으로 목표를 변경하였다. 수정된 질의는 그림 10과 같다. 이전 질의에서 처리된 pop파일을 사용하여 연관 규칙을 찾기 때문에 질의 처리 시간을 상당히 줄일 수 있다. 질의 결과는 그림 11과 같다.

```
MINE ASSOCIATION
USING TASK RELEVANT DATA pop
FOR T2.traffic, T2.parking, T2.tra_acc, T2.park_acc
FROM Building_T1, Building_T2
SET SUPPORT THRESHOLD 0.3 AND
CONFIDENCE THRESHOLD 0.5;
```

(그림 10) 작업 관련 데이터를 사용한 질의 예제



(그림 11) 질의 결과

4. 결론

최근 방대한 양의 공간 데이터로부터 흥미로운 지식을 추출하기 위한 연구들이 활발하게 진행되고 있다. 그러나 우리가 흔히 이용하고 있는 공간 데이터베이스 시스템은 공간 데이터마이닝을 효율적으로 지원하지 못한다. 본 논문에서는 공간 데이터베이스에서 사용되는 질의를 확장하여 공간 연관 규칙을 발견하는 시스템을 제안하였다. 그리고 GMS를 기반으로 단순한 지리적 위상관계보다 각 객체의 속성들이 고려된 질의 처리 결과를 보였다.

참고문헌

[1] K.Koperski, and J.Han., Discovery of Spatial Association Rules in Geographic Information Databases In Advances in Spatial Databases (Proc. 4th Symp. SSD '95), pp 47-66, Portland, ME, August, 1995.

[2] J. Han, K. Koperski, N. Stefanovic, GeoMiner : A system prototype for spatial data mining, In Proc. ACM SIGMOD Int. Conf. On Management of Data, pp 560-563, Tucson, Arizona, 1997.

[3] 박선, 박상호, 안찬민, 이윤석, 이주홍, SIMS를 위한 공간 데이터 마이닝 질의 언어, 한국정보과학회 춘계학술발표논문집 제 31권 제1호, p70-72, 2003.

[4] D. Malerba, F. Esposito, A. Lanza and F.A. Lisi, Discovering Geographic Knowledge: The INGENS System, Lecture Notes In Computer Science; Vol. 1932, Proceedings of the 12th International Symposium on Foundations of Intelligent Systems table of contents, 40 48, 2000.

[5] Osmar R. Zaïane, Jiawei Han, Ze-Nian Li, Sonny H. Chee, Jenny Y. Chiang, MultimediaMiner: A System Prototype for Multimedia Data Mining. Proceeding of International Conference on management of Data SIGMOD'98, Seattle, WA USA 1998.

[6] May, M., SPIN! an Integrated Spatial Knowledge Discovery Platform. Leopold, E. (ed.) Fachgruppentreffen Maschinelles Lernen der Gesellschaft für Informatik, Sankt Augustin 18-20.9.2000, GMD Report.