GMS 에서의 공간 연관 규칙 탐사 시스템의 설계 및 구현

안찬민*, 이주홍*, 전석주**
*인하대학교 컴퓨터정보공학과, **서울교육대학교 컴퓨터교육과
e-mail : ahnch1@datamining.inha.ac.kr*, juhong@inha.ac.kr*, chunsj@snue.ac.kr**

Design and Implementation of Spatial Association Rule in GMS

Chan Min Ahn*, Ju Hong Lee*, and Seok Ju Chun**
*Department of Computer Science & Engineering, Inha University
**Department of Computer Education, Seoul National University of Education.

요 약

본 논문에서는 지리정보 시스템인 GMS 를 기반으로 한 공간 연관 규칙의 구현과 설계 방법을 제안한다. GMS 에는 비공간 데이터와 공간 데이터가 테이블로 구분되어 저장되어 있다. 이를 이용하여 비공간 데이터 집합에서 관련된 데이터 집합을 추출한 후 그에 해당되는 공간 데이터를 이용하여 공간 연관 정보를 찾아내서 연관 규칙을 발견하는 방법에 대입하여 공간 연관 규칙을 발견한다.

1. 서론

정보화 사회에서 지식 사회로 발전하면서 과거에는 주로 정보의 저장 역할을 해내던 DB 시스템에도 많은 변화가 요구되고 있다. 통신 기술과 인터넷의 발달로 정보 Source 의 절대량이 기하급수적으로 증가하고 있는 현재, 단순한 정보의 저장은 이미 그 한계에 다다르고 있다. 미래 사회에서는 정보 저장 자체보다도 수많은 정보 속에서 사용자에게 필요한 유용한 지식을 찾아내는 일의 가치가 증대되고 있다. 이러한 유용한 지식을 찾아내는 기술이 데이터 마이닝이다.

본 논문에서는 연관 규칙 탐사 시스템의 구현을 목표로 하고 있으며 특히 지리정보 시스템에서의 연관 규칙을 찾는 방법에 초점을 맞춰 실제 지리정보 시스템에 적용 가능한 시스템의 모델을 제안하고 구현하는 것을 목표로 한다. 방대한 데이터에서 서로 연관된 특징을 갖는 유용한 정보를 찾는 것을 연관 규칙 마이닝이라고 한다.

공간 연관 규칙은 공간적 특성을 기준으로 발견되는 연관 정도를 나타낸다. 즉 주어진 데이터 집합에서 출현하는 빈도가 높은 값을 기준으로 연관 규칙을 찾는 것이 데이터 마이닝의 연관 규칙이라면 거리의 멀고 가까운 정도나 인접한 정도, 특정 지리적 특성을 갖는 곳에서 발견되는 공간적인 특성이 포함된 데이터 집합에서의 출현 빈도를 찾아내서 연관 정도를 찾는 것이 공간 연관 규칙이다.

이러한 데이터 마이닝 기술을 기반으로 한 지리 정보 시스템에서의 데이터 마이닝 시스템으로 GeoMiner 가 있다.[3] 이 GeoMiner 에서 연관 규칙을 사용하기 위해 Koperski 는 기존의 연관 규칙 방법을 5 단계로 나누어 공간 연관 규칙을 발견하는 방법을 제안하였다.[2]

본 논문에서는 GMS 시스템을 이용하여 찾고자 하는 목표를 설정한 후 범위에 관계 없이 존재하는 데이터 집합에서 발견되는 모든 연관 규칙을 찾아내는 방법을 적용하고 지도 데이터를 이용하여 비공간 및 공간 데이터 집합에서 공간 연관 규칙을 찾는 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 관련 연구를 보이고 3 절에서는 알고리즘과 함께 실제 구현하여 실행하는 모습을 보인 뒤 4 절에서 결과를 분석한다. 마지막으로 5 절에서 결론을 맺고 향후 개선해야

할 과제를 정리한다.

2. 관련 연구

2.1 연관 규칙

데이터 마이닝의 방법은 Characterization, Discrimination, Association, Classification, Prediction, Clustering, Outlier Analysis 등의 6 가지로 분류된다[1].

연관 규칙 탐사는 주어진 데이터 집합에서 흥미로운 관계를 발견하는 작업으로 동시에 발생되는 빈도를 확률적으로 분석하여 일정 확률 이상이 발생되면 높은 연관 관계를 갖는다고 보고 이러한 연관 관계를 생성하는 규칙을 찾아내는 작업이다. 예를 들어 “기저귀를 구입하는 사람은 맥주를 구입할 확률이 높다” 와 같이 자료들 간에 직접적인 연관은 없지만 동시에 발생되는 빈도가 확률적으로 높은 경우를 발견할 수 있다.

연관 관계의 유용성을 판단하는 기준은 Confidence 와 Support 로 정의된다. Itemset A 와 B 가 있다고 가정할 때 A 를 포함하는 Tuple 에서 A 와 B 모두를 포함하는 Tuple 의 비율을 Confidence 라고 한다. 그리고 전체 Tuple 에서 A->B 를 만족하는 Tuple 의 비율을 Support 라고 한다. 사용자가 설정한 Support 와 Confidence 보다 많은 출현 빈도를 가지면 유용한 정보라고 할 수 있다.

동시에 발생하는 사건을 분석하는 대표적인 방법으로 Apriori 알고리즘이 있다.[4] 이는 후보 집합을 설정해 줌으로서 불필요한 계산량을 줄이고 성능을 개선시키는 방법이다.

2.2 공간 연관 규칙

공간 연관 규칙은 기존의 데이터 연관 규칙과는 달리 공간적인 속성이 포함된 연관 규칙을 말한다. 예를 들어 “역과 주거지역간의 거리와 집값의 연관규칙을 찾아라”, “집의 가격과 주변 도로와 교통량의 연관 규칙을 찾아라”와 같이 공간적인 속성이 포함되어야 한다. 공간적인 속성은 거리의 가까움과 먼 정도 등을 나타낸다.

공간 연관 규칙을 찾아내는 방법 중 대표적인 것으로 Koperski 의 알고리즘이 있다.[2] 이는 GeoMiner 를 이용하여 다섯 단계의 과정을 거쳐 연관 규칙을 찾아내는 방법을 보이고 있다..Malerba, D.는 ILP 시스템을 이용한 공간 연관 규칙 탐사 방법을 제안하였다[5]. 이는 공간 데이터 중에서 빈번하게 발생하는 데이터를 찾아내어 연고나 규칙을 발견함을 보이고 있다.

3. 설계 및 구현

본 논문에서 제안하는 시스템은 GMS 를 기반으로 하여 SMQL[6]을 사용, 연관 규칙을 찾아내는 시스템이다. 그 구성은 그림 1 과 같다.

3.1 GMS

GMS 는 공간 DBMS 로서 다중 사용자를 지원하며 공간 데이터 및 비공간 데이터를 효율적으로 저장, 관리할 수 있다.

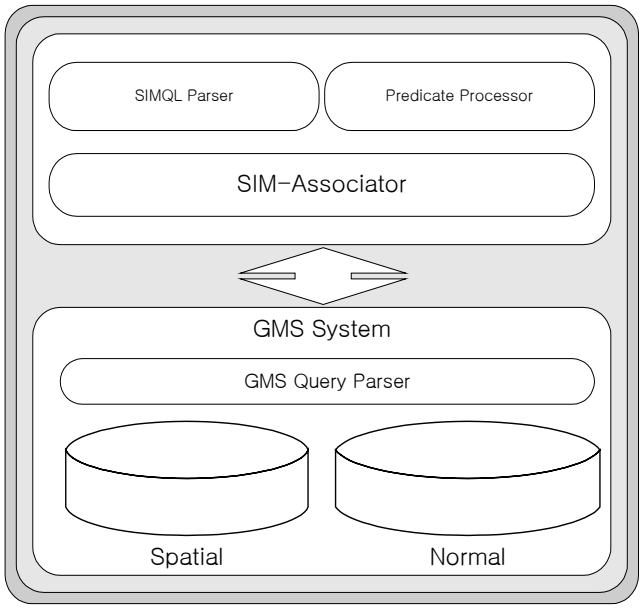

그림 1. GMS 를 이용한 연관규칙 탐사 시스템

GMS 의 자료 구조는 크게 맵 테이블과 노멀 테이블로 나눌 수 있다. 맵 테이블은 건물, 도로, 강, 산 등이 각각의 테이블로 저장되며 각 객체의 위치와 객체 형태 등의 공간 정보가 저장된다. 노멀 테이블에는 비공간 속성 등의 비공간 정보가 저장된다. 맵 테이블과 노멀 테이블은 서로 조인 연산이 가능하여 비공간 속성을 기준으로 분류한 객체끼리의 공간 연산을 적용할 수 있다. 자료 검색을 위해 질의와 함께 Open GIS 의 표준을 따르는 공간 Predicate 를 적용하여 필요한 연산 결과를 확인할 수 있다.

GMS 의 질의는 SQL92 의 표준을 기반으로 하며 공간 데이터 처리를 위한 공간 SQL 을 포함한 확장된 SQL 을 사용한다. 공간 SQL 은 OGC 에서 표준으로 제안하는 7 개의 기본 공간 데이터 타입과 9 개의 공간 관계, 확장된 공간 데이터 타입 및 공간 관계 연산자 및 공간 함수를 지원한다.

3.2 SMQL

SMQL 은 SIMS(Spatial Information Management System)에서 사용하기 위한 공간 데이터 마이닝 질의 언어이다. [6] 공간 데이터 마이닝 기법 중에서 Assoc-iation, Classification, Clustering, Trend Analysis 를 지원하는 질의 언어이며 Association 에 해당되는 스펙은 그림 2 와 같다.

3.3 GMS 에서의 공간 연관 규칙 탐사 시스템

본 논문에서 제안하는 공간 연관 규칙 탐사 시스템은 크게 자료가 저장되는 공간 데이터베이스 부분과 정보를 추출하는 데이터 마이닝 부분으로 구성되어 있다. 자료는 공간 정보는 맵 테이블, 비공간 정보는 노멀 테이블과 같이 서로 다른 테이블에 저장되어 있으며 모든 객체와 테이블에 부여된 OID(Object ID)를 통해 해당 객체의 공간 정보와 비공간 정보를 열람할 수 있다.

```
SMQL_query ::= MINE rule_header
        [ USING HIERARCHY   hierarchy_description ]
                [ FOR analysis_standards ]
                FROM table_list
                [ WHERE conditions ]
                [SET threshold_specification ]

rule_header ::= association_header | characteristic_header

association_header ::= ASSOCIATION [AS pattern_literal]

threshold_specification ::=
        threshold_description THRESHOLD number
```

그림 2. SMQL 에서의 Association 부분의 스펙(일부)

GMS 에서의 공간 연관 규칙 탐사 시스템에서는 GMS 에서 얻어낸 데이터를 마이닝하기 위한 SMQL 과 기존 GMS 의 질의를 조합하여 연관 규칙을 찾아내며 그 방법은 그림 3 같다.

그림 3 GMS 에서의 공간 연관 규칙 탐사 시스템의 흐름도

우선 기준이 될 OID A 를 찾는다. 공간 연관 규칙을 찾고자 하는 기준의 역할은 검색 목표를 분명히 하는 효과를 갖는다. 두번째 단계에서는 기준이 될 A 를 기점으로 "어떻게 연관된" 정보를 찾을 것인지 전체적인 집합 B 를 구한다. B 는 비공간적인 연관 데이터와 공간적인 연관 데이터들을 포함한다. 이때 비공간적인 연관 데이터는 검색이 용이하나 공간 연관 데이터의 검색은 데이터 마이닝을 고려하지 않은 GMS 시스템 성능의 한계 때문에 많은 처리 시간이 요구된다. 따라서 비공간 연관 데이터를 미리 찾은 후 그 결과들에 대한 공간 연관 데이터를 검색하여 검색 효율을 높인다. 세번째 단계에서는 이렇게 얻은 데이터집합 B 중을 기준으로 규칙을 발견하고자 목표했던 데이터 집합을 구하고 네번째 단계에서 각각의 경우에 대한 가장 많은 출현 빈도를 갖는 데이터 값을 찾는다. 데이터 집합의 개수는 종류에 관계없이 사용자가 지정하게 되며 얻어진 값들은 처음에 찾고자 했던 목표와 관련 정도가 높다고 할 수 있다. 다섯번째 단계에서 Apriori 알고리즘을 적용하여 연관 규칙을 발견한다.

4. 실험 및 결과

본 논문에서는 결과를 보이기 위해 GMS 의 강남지역교통도를 사용했다. 맵 테이블에는 강, 강명, 건물, 건물명, 녹지, 도로, 도로명, 지하철, 행정구계, 행정구명, 행정동계, 행정동명, 산, BACKGR, 기타 TEXT 라는 테이블이 존재한다. 노멀 테이블에는 맵 테이블에 저장된 객체들의 OID 와 일치하는 비공간 데이터가 저장되어 있다.

GMS 에서는 공간 데이터 마이닝 기술을 지원하는 연산이 없으므로 공간 연관 규칙을 위해 몇 가지 기준을 정의하여야 한다. Predicate 은 데이터 베이스에 저장된 정보들 중에서 비슷한 의미를 갖거나 수치 데이터의 묶음으로 표현하여 연산량을 줄이고 객관적인 지표로 정보를 얻는 효과를 갖는다. 본 논문에서는 수치 데이터의 묶음으로 표현하기 위해 소득 수준 값의 범위를 설정하여 표 1 과 같은 소득 수준 Predicate 을 설정하였다. 또한 공간적으로 비슷한 경향을 갖는 거리 지표에 대해서 표 2 와 같이 거리 구분 Predicate 을 정의하였다.

소득수준	소득값
저소득	0~1500
평균 소득	1500~5000
고소득	5000 이상

표 1 연관 규칙을 찾기 위한 소득 수준 Predicate

거리구분	거리값
이웃	0~120
인접	120~300
주변	300~500
근처	500 이상

표 2 연관 규칙을 찾기 위한 거리 구분 Predicate

이들은 SMQL 질의에서 사용자가 범위를 지정해 줄 수 있다.

이 기준으로 " 소득이 높은 집 주변 건물과의 연관 규칙" 에 해당되는 질의는 그림 4 와 같다.

```
MINE Association Rule AS "소득수준"
FOR T2.속성, T2.시공연도
FROM 건물 T1, 건물 T2
WHERE T1.소득 = 고소득
        AND DISTANCE(T1,T2,주변);
SET SUPPORT THRESHOLD 0.9
SET CONFIDNCE THRESHOLD 0.5
```

그림 4 공간 연관 규칙 탐사를 위한 질의의 예

5. 결론 및 향후 과제

본 논문에서는 GMS 에서 공간 연관 규칙을 발견하는 시스템을 제안하였다. 그러나 공간 정보는 우리 실생활과 밀접한 관계를 갖고 있으므로 정적인 자료가 아닌 동적인 자료의 집합이다. 주기적으로 갱신되는 정보들을 분석하고 예측할 수 있는 시스템이 보다 우리 생활에 밀접하게 연관되고 도움을 줄 것이다.

또한 본 시스템에서 사용한 도로의 거리 부분은 지도상의 직선 거리를 기준으로 개발된 내용으로 실제 생활에서 느껴지는 체감 거리와는 크게 다르다. 따라서 공간 데이터베이스의 도로 정보에 위상 정보가 추가되어 직선 거리 뿐 아니라 실제 생활에서 적용되는 거리에 대한 정보를 분석할 수 있도록 개발되어야 할 것이다.

Acknowledgement

본 연구는 대학 IT 연구센터 육성 지원사업의 연구결과로 수행되었음

참고문헌

[1] Jiawei Han, Micheline Kamber, Data Mining, Concepts and Techniques, Morgan Kaufmann Publishers, 2001

[2] 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

[3] 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

[4] Rakesh Agrawal, Tomasz Ilielinski, and Arun Swami, "Datamining : A performance perspective", IEEE Transactions on Knowledge and Data Engineering, 5(6), 12, 1993.

[5] Malerba, D., Lisi, F.A., An ILP method for spatial association rule mining. In A. Knobbe and D.van der Wallen(Eds.), Notes of the ECML/PKDD 2001 Workshop on Multi-Relational Data Mining, 18,29, Freiburg, Germany, 2001

[6] 박선,박상호,안찬민,이윤석,이주홍, SIMS 를 위한 공간 데이터 마이닝 질의 언어, 한국정보과학회 춘계학술발표논문집 제 31 권 제 1 호, p70-72, 2003