GML데이터에서 지역적 연관규칙 탐색 기법
A Local Association Rule Search Method from GML Data

홍성한*, 황병연
Sung-Han Hong, Byung-Yeon Hwang
가톨릭대학교 컴퓨터공학과
{hongta, byhwang}@catholic.ac.kr

요 약

GIS분야에 대한 다양한 연구가 진행됨에 따라 그 활용에 대한 관심도 확대되고 있다. Open GIS Consortium에서는 GML(Geography Markup Language)을 개발하여 이를 GIS 응용분야에 활용하고자 하는 연구가 활발히 진행되고 있다. GML데이터에서 의미 있는 정보를 추출하기 위해서는 데이터 마이닝 기법 활용이 필수적이다. 최근에 데이터마이닝 기법 중 연관규칙을 이용한 탐색 방법이 제안되었다. 그러나 이 방법은 전체 데이터를 대상으로 의미 있는 정보를 탐색 하므로, 데이터 내에 포함되어 있는 부분 속성인 지리 공간적 연관성을 탐색하는데 한계를 가지고 있다. 따라서 본 연구에서는 GML데이터에서 부분적 속성을 고려한 지역적 연관규칙 탐색 기법을 제안한다.

1. 서 론

OGC(Open GIS Consortium)에서는 GML(Geography Markup Language)을 개발하여 다양한 응용분야에서 활용하려는 연구가 지속적으로 진행되고 있다[1].

GML언어는 W3C(W3Consortium)에서 웹 개발과 관련된 표준화 언어인 XML(eXtensible Markup Language)[1, 2]의 다양한 특징들을 포함한다. 그러나 GML에 XML에서 연구되었던 다양한 특징적 기법들을 직접 적용하는 것은 문제를 가지고 있다. GML 데이터가 가지고 있는 다양한 지리 공간적 속성 때문이다. 특히, 의미 있는 정보를 추출하는데 있어 지리 공간 속성 정보를 고려한 탐색은 필수적이다.

본 연구에서는 지리 공간적 부분 속성을 고려한 지역적 연관규칙 탐색 기법을 제안하려 한다.

XML/GML 문서에서 의미 있는 정보를 탐색하기 위해서는 데이터 마이닝 기법의 접목이 필수적이다[3, 4]. 기존의 XML 문서에서는 의미 있는 정보를 추출하기 위해 데이터 마이닝 기법의 순차패턴 마이닝, 연관규칙, 분류화, 일반화 기법 등을 변형하여 의미 있는 정보를 추출하였다[3, 5].

최근에 GML문서 데이터로부터 의미 있는 정보를 추출하기 위한 기법이 제안되었다[5]. 이 기법에서는 GML 데이터에서 태그를 제외한 내용을 추출하고, 데이터 마이닝 기법 중 연관규칙을 적용하여 연관성을 탐색하였다.

일반적인 연관규칙 탐색 프로세스를 살펴보면, 전체 데이터베이스에서 후보 항목집합을 찾고, 이미 설정된 최소지지도 임계값을 넘는 빈발 항목집합을 찾는다.

GML문서에서 의미 있는 정보를 추출하기 위해서는 속성 정보를 고려한 연관 규칙 탐색이 필요하다.

예를 들어, A도시에 거주하는 사람들의 세대, 주택, 기반시설 등에 따라 형성되는 시설물들의 연관성도 달라진다. 즉, 아파트와 단독주택 주변 시설물간의 지역적 연관성이 다르다. 따라서 이를 일반적인 연관규칙만을 고려해서 탐색한다면, GML 데이터에서 의미 있는 지역적 연관성 탐사가 불가능하다.

본 연구에서는 GML데이터로부터 부분적 속성을 고려한 연관성 탐색 방법을 제안하려고 한다.

2장에서는 기존 관련연구에 대한 기술을 하고, 3장에서는 본 논문에서 제시하는 기법을 설명하며, 4장에서는 실험 예제를 통해 본 연구의 내용을 살펴 볼 것이다. 5장에서는 결론 및 향후 연구에 대해서 기술한다.

2. 관련 연구

데이터 마이닝이란 대량의 데이터로부터 감춰진 유용한 정보를 추출하는 방법을 말한다[6]. 데이터 마이닝은 흔히 KDD(Knowledge Discovery in Database)라고도 불리며, 데이터 마이닝 기법에는 대표적으로 연관규칙, 클러스터링, 분류화 기법들이 있다[4, 5].

그 중에 연관규칙이란 동시에 발생하는 사건들을 규칙의 형태로 표현한 것으로 특정 사건이 발생하면 동시에 혹은 일정한 시간 간격 사이에 다른 사건과의 관련성을 탐색하는 기법을 말한다[5]. 즉, 트랜잭션들을 대상으로 항목 또는 속성간의 연관관계를 발견하는 방법을 말한다.

연관규칙을 찾는 방법을 간단히 살펴보면, 전체 데이터베이스에서 먼저 후보 항목집합을 찾고, 미리 설정된 최소 지지도(Minimum Support) 임계값을 넘는 빈발 항목을 찾아낸다. 빈발 항목 집합 탐색 시 전체 트랜잭션을 반복적으로 검색하면서 조인연산을 지속적으로 실행하게 되는 구조이다[7]. 이 때 규칙의 타당성을 검증하기 위한 기준으로서 지지도(Support)와 신뢰도(Confidence)가 적용된다. 지지도는 전체 항목 중에서 연관 규칙 R : X→Y를 지지하는 비율을 의미한다. 또한 신뢰도는 X의 모든 항목을 포함하고 있는 트랜잭션의 개수에 대하여 Y 또한 포함하는 비율을 의미한다[2, 3, 4].

연관규칙은 다양한 방법으로 응용되고 있다[2, 5]. 최근에 연관 규칙 탐색은 복수의 지지도를 사용하는 연관 규칙의 탐사[2], 일정 주기 상에서 나타나는 순환적인 연관 규칙 탐사[4], 공간 연관 규칙 탐사[3], 일정한 시간을 간격으로 동시에 나타나는 순차 패턴의 탐사[3] 등으로 응용되어 활발히 연구되고 있다.

연관규칙 탐사와 관련한 대표적인 알고리즘은 Apriori 알고리즘이다[4]. 이 알고리즘의 특징을 살펴보면, 빈발 항목집합의 최대 길이가 k일 때 k+1만큼 반복적으로 스캔하여 후보항목 집합을 생성하는 구조를 가진다[5].

기존의 XML 데이터를 가지고 연관 규칙을 적용한 연구들이 있다. [4]에서는 XML문서 데이터에서 빈발한 경로를 찾는 연구를 하였고, [3, 4]에서는 연관 규칙 탐색을 위한 확장된 XQuery를 제안하였다.

[4]에서는 FP-growth 알고리즘을 기반

으로 XSD-AR(XML Structural Delta Association Rule)이라는 연관 규칙 기반의 타입을 제안하였다.

최근에 GML데이터에서 의미 있는 정보를 추출하기 위한 기법이 제안되었다[5]. 이 기법은 GML데이터로부터 내용을 추출하고, 데이터 마이닝 기법 중 연관규칙을 적용하여 의미 있는 정보를 탐색했다. 그러나 GML문서 데이터 내부에 포함되어 있는 지리 공간 정보의 속성을 고려한 지역적 연관성은 찾을 수 없다. 예를 들어, 서울시 전체 구에 공동주택 주변 서비스 업종의 연관성과 강남구 내에 공동주택 중 아파트 주변 서비스업종 연관성은 다르다. 따라서 GML데이터처럼 다양한 지리 공간적 속성을 가진 문서는 지역적 연관성 탐색이 필요하다.

3. 지역적 연관규칙 탐색

본 장에서는 GML문서에서 부분적 속성을 고려한 지역적 연관규칙 탐색 기법을 제안한다.

3.1 지역적 연관성 탐색 정의

지역적 연관성이란 부분 속성 고려 없이 의미 있는 정보를 탐색하는 전역적 연관 탐색에 대비되는 개념이다.

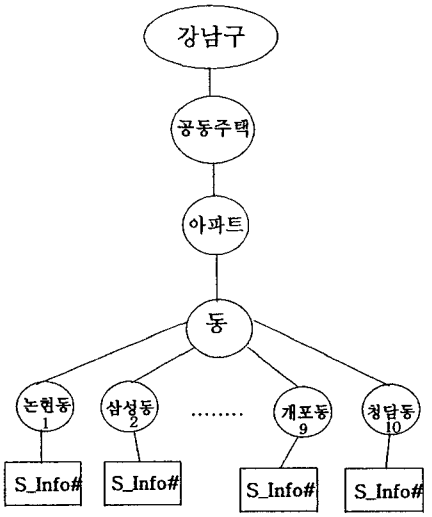
예를 들어, 서울시 각 구별 공동주택 주변에 서비스업종에 대한 연관성을 탐색한다고 가정하자. 이 때는 서울시 전체를 대상으로 탐색이 이뤄지기 때문에 전역적 탐색이라 할 수 있다. 이에 반에 지역적 연관성 탐색은 특정 구 내에 공동주택에 관한 세부 속성을 고려한다. 즉, 아파트, 연립주택, 다세대 주택 주변에 서비스 업종에 대한 연관성을 탐색하게 된다.

따라서 지역적 연관성 탐색은 GML문서처럼 지리 공간적 정보를 기반으로 한 문서에서 필요한 탐색 방법이 된다.

3.2 GML데이터로부터의 항목 추출

GML데이터로부터 의미 있는 정보를 추출하기 위해서는 태그를 제외한 내용을 추출해야 한다. 일반적으로 의미 있는 정보를 탐색 할 때는 데이터 자체의 값을 이용하여 규칙을 찾기 때문이다.

또한, GML문서는 지리 공간 정보의 다양한 속성을 포함한다. 따라서 특정 속성 내에 존재하는 항목들 간에 의미 있는 정보를 추출하기 위해서는 문서 구조가 동일해야 한다.



<그림 1> GML 스키마의 예

<그림 1>은 동일한 구조를 가진 GML 문서 스키마를 나타낸 예이다. 즉, <그림 1>에서 강남구 내에 있는 아파트를 '동'별로 분류하여 주변 서비스 업종 환경정보

를 나타낸 GML문서 구조이다.

<그림 1>에서 타원들은 태그를 나타내며, 타원 내부는 내용을 의미한다. 또한 각 사각형들 내부는 타원 속 내용에 대한 속성을 의미한다. Info#는 각 속성들에 대한 Value(값)를 나타낸다.

먼저 연관규칙을 탐색하기 위해서는 트랜잭션 테이블을 구성해야 한다. 따라서 추출된 내용을 기반으로 데이터베이스를 구성해 보면, <그림 1>에서처럼 강남구 내에 공동주택 중 아파트를 분류한 '동'들이 트랜잭션(D_TID)이 된다. 또한 S_Info#에 해당되는 Value(값)들은 연관규칙을 탐색하기 위한 서비스 업종(항목)들이 된다. 다시 말해, 강남구에 아파트(공동주택)를 '동'별로 분류하여 그 주변 서비스 업종의 연관성을 탐색하는 것이다. 추출된 트랜잭션 테이블의 예는 <표 1>과 같다.

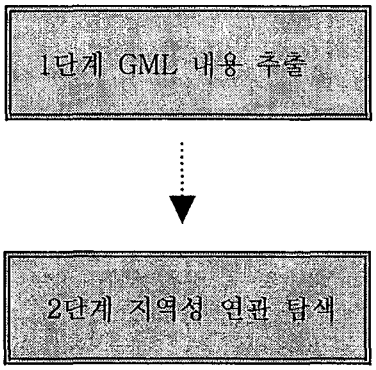
<표 1> 추출된 트랜잭션 테이블

D_TID	서비스 업종(항목)
1	C, F, H
2	F, G, H
3	F, H, N
4	F, H, L, M
5	C, E, F, J
6	B, F, H, L
7	B, F, H
8	A, B, C, F, H
9	A, B, C, F
10	A, B, C, D, G

3.3 지역적 연관규칙 탐색

3.3.1 지역적 연관규칙 탐색 단계

지역적 연관규칙 탐색은 <그림 2>과 같이 2단계 과정으로 구성된다.



<그림 2> 지역적 연관규칙 탐색단계

1단계는 GML데이터로부터 태그를 제외한 내용을 추출하는 단계이다. 1단계에서 추출된 내용들을 트랜잭션 ID와 아이템 셋(Item Set)으로 구분하여 전체 트랜잭션 테이블을 구성한다.

2단계는 지리 공간적 속성을 고려한 트랜잭션을 중심으로 Apriori 알고리즘을 적용하여 지역적 연관규칙을 탐색한다.

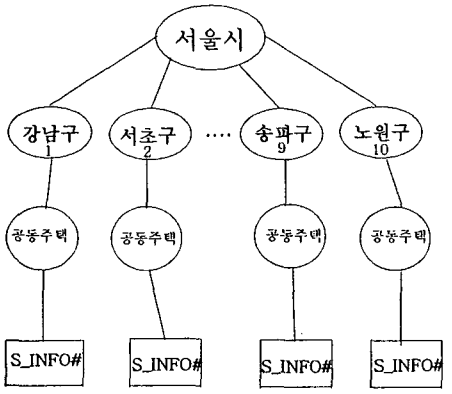
4. 실험 예

이 장에서는 지역적 연관 규칙 탐색과 전역적 연관 규칙 탐색 방법을 비교 실험한다.

실험 순서는 3.3절에서 제시한 2단계의 과정을 거쳐 실험한다.

먼저, 비교 탐색을 위해서 GML문서가 동일한 구조라고 가정한다. 문서구조가 상이할 경우 부분속성 고려의 의미가 없기 때문이다.

<그림 3>은 전역적 탐색을 위한 GML 문서 구조이다.



<그림 3> 전역적 GML문서 구조

<그림 3>은 서울시 각 구별 공동주택 주변 서비스 업종 정보를 GML문서 구조로 나타낸 것이다. 여기서 공동주택은 아파트, 연립주택, 다세대 주택을 포함한다.

<그림 3>에서 전역적 연관성 탐색을 하기 위해서는 내용을 추출하여 트랜잭션 테이블을 구성해야 한다. 즉, 서울시 각 구가 트랜잭션(G_TID)이 되고, 서비스 업종 정보가 항목이 된다. 트랜잭션 테이블을 구성하면 <표 2>와 같다.

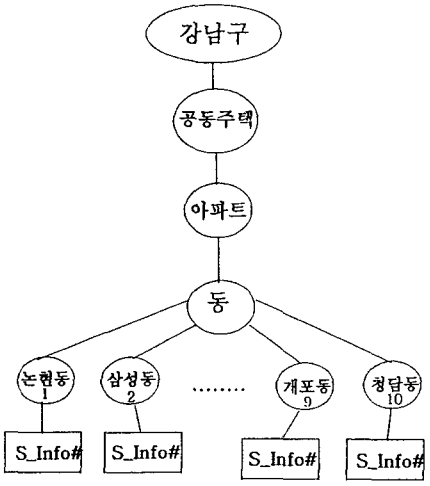
<표2> 추출된 트랜잭션 테이블

G_TID	서비스 업종(항목)
1	A, B, D, E
2	B, C, D
3	A, B, D, E
4	A, C, D, E
5	B, C, D, E
6	B, D, E
7	C, D
8	A, B, C
9	A, D, E
10	B, D

<표 2>로부터 연관성을 탐색하기 위해서는 대표적 탐색 알고리즘인 Apriori알고리즘을 이용한다. 최소 지지도 임계값은 50%로 가정한다.

탐색 결과, 최소 지지도 50%를 만족하는 2-빈발 항목 집합 {D, E}가 탐색되었다.

다음은 비교 실험을 위해 지역적 연관규칙 탐색을 한다.



<그림 4> 지역적 GML문서 구조

<그림 4>는 강남구 내에 공동 주택 중 아파트를 '동'을 기준으로 분류하여 주변 서비스 업종을 나타낸 GML문서 구조이다. <그림 4>로부터 지역적 연관성을 탐색하기 위해서는 추출된 내용을 기반으로 트랜잭션 테이블을 구성해야 한다. 즉, 강남구 내에 각 '동'들이 트랜잭션(D_TID)이 되고, 서비스 업종 정보들이 항목이 된다.

지역적 연관성 탐색을 위해 추출된 트랜잭션 테이블은 <표 3>과 같다.

<표 3> 추출된 트랜잭션 테이블

D_TID	서비스 업종(항목)
1	C, F, H
2	F, G, H
3	F, H, N
4	F, H, L, N
5	C, E, F, J
6	B, F, H, L
7	B, F, H
8	A, B, C, F, H
9	A, B, C, F
10	A, B, C, D, G

실험조건은 전역적 연관성 탐색 방법과 동일하게 Apriori 알고리즘을 이용한다.

또한, 최소 지지도 임계값은 50%로 설정한다.

탐색 결과, 전역적 방법으로 탐색되지 않았던 최소지지도 50%를 만족하는 2-빈발 항목집합 {F, H}가 탐색 되었다.

5. 결과 및 향후 연구

본 연구에서는 GML데이터로부터 부분적 속성을 고려한 연관규칙 탐색 방법을 제안하였다.

기존의 전역적 연관규칙 탐색에서는 GML데이터의 지리 공간적 부분 속성이 고려되지 않았다. 그러나 제안된 지역적 연관규칙 탐색에서는 부분적 공통속성을 고려했다. 따라서 GML문서로부터 의미있는 다양한 정보를 추출할 수 있다.

향후 연구로는 GML데이터에서 다양한 이동객체의 속성들을 고려한 연관 규칙 탐색 방법에 대한 연구가 필요하다.

< 참 고 문 헌 >

[1] Open GIS Consortium, Inc, Geography Markup Language Specification (GML) v3.1.1, 2004.

[2] D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P.L. Lanzi, "A Tool for Extracting XML Association Rules," proc. of the 14th IEEE International Conf. On Tools with Artificial Intelligence, Washington DC, USA, Nov. 2002, pp. 57-64.

[3] A. Meisels, M. Orlov and T.Maor, "Discovering Association in XML Data," proc. of the 3rd International Conf. on Web Information Systems Engineering, Singapore. Dec. 2002 pp. 178-183.

[4] L. Chen, S.S Bhowmick, and L.T. Chia, "Mining Association Rules from Structural Deltas of Historical XML Documents," Proc. of the 8th Pacific-Asia Conf. Sydney, Australia, May. 2004, pp. 452-457.

[5] 김의찬, 황병연, "GML 데이터에서 연관규칙 추출," 한국공간정보시스템학회 추계학술대회, 2005, pp. 55-60.