

지리정보 통합데이터베이스 구축을 위한 형식개념분석(FCA)의 적용 Application on Formal Concept Analysis for Constructing Integrated GIS Database

김병선, 구자용, 윤성민

Byung Sun Kim, Cha Yong Ku, Sung Min Yun

상명대학교 지리학과 GIS 연구실

gisguy@paran.com, koostar@smu.ac.kr, kkong1042@smu.ac.kr

Abstract : 국토 모니터링을 위해서는 다양한 출처와 종류의 자료를 통합하여 제공하여야 한다. 특히 국토 모니터링의 범위에는 원격탐사와 같은 공중 모니터링과 지상 모니터링 자료가 모두 포함되기 때문에 매우 다양한 종류와 속성을 가진 지리정보 자료들이 통합되어야 한다. 본 연구에서는 다양한 출처와 종류의 지리정보자료를 효과적으로 통합하는 방법으로 형식개념분석(Formal Concept Analysis, FCA)을 살펴보고 사례분석을 통해 이 기법의 적용 가능성을 파악하고자 한다. 연구결과 형식개념분석을 통하여, 다양한 종류의 자료가 가지고 있는 중복을 제거하고 이를 체계적으로 정리하여 통합할 수 있다. 본 연구에서는 형식개념분석의 개념을 파악하고 현재 활용되고 있는 지리정보자료들에 적용하여 평가함으로써 국토 모니터링 자료의 통합기법으로 적용될 수 있는 가능성을 연구하였다.

1. 연구배경 및 목적

형식개념분석기법은 도메인 내의 다양한 데이터들로부터 객체와(Object)와 속성(Attribute)들을 추출하고, 이들 사이의 포함관계를 파악하여 개념(Concept)을 생성하고 개념계층구조(Conceptual hierarchy)를 구축하기 위한 수학적 분석기법이다. 이런 형식개념분석기법은 특정 분야의 전문가나 지식 공학자들에 의해서 많은 시간과 노력을 투입하여 작성되었던 계층적 개념구조를 보다 수월하게 작성할 수 있다는 장점이 있기 때문에 의학 및 바이오인포매틱스, 사회과학, 소프트웨어공학, 온톨로지공학 등의 다양한 분야에서 폭넓게 활용되고 있다(김홍기 외, 2006).

지리정보분야에서 형식개념분석기법은 통합데이터베이스 구축에 활용될 수 있는데, 특히 파일 단위의 기초공간자료들이 의미론(semantic)적으로 중복되는 경우 이를 조직화할 수 있는 방법을 제공해 준다. 실제 구축된 파일단위의 개별 지리정보자료들은 특정 분야(교통, 환경, 도시계획 등)에 종속적으로 도형정보와 속성정보를 갖는다. 이러한 개별 정보들의 특성을 조직화하는 것이 통합데이터베이스 구축에 근본이라 할 수 있으며 동시에 가장 큰 문제점이라고 할 수 있다. 따라서 정제된 형태의 통합데이터베이스를 구축하기 위해서는, 개별 지리정보자료들의 의미론적인 특성들 즉 이들 간의 유사성이나 이질성 그리고 속성들의 계층적 위계들이 표준화된 기준을 통해 체계적으로 정립하여 데이터모델을 구축해야 한다(Kavouras and Kolka, 2002).

이러한 배경에서 본 연구에서는 지리정보 통합데이터베이스 구축에 있어 기반이 되는 기본 데이터모델 구축을 위해 현재 데이터모델링과 관련하여 다양한 분야에게 활용되고 있는 형식개념분석기법을 적용하고, 이에 대한 실용 가능성을 검토하는 것을 연구의 목적으로 한

다. 특히 개별지리정보의 의미론적인 중복성을 체계화하여 이를 통합하는데 중점을 두었다.

본 연구는 다음과 같이 진행된다. 첫째, 형식개념분석에 대해 설명하고 둘째, 기구축된 개별지리정보 자료들을 형식개념분석기법을 이용하여 체계적으로 통합하고 마지막으로 형식개념분석의 국토모니터링 통합데이터베이스 구축에 있어 활용 가능성과 보완점에 대해 논의한다.

2. 형식개념분석기법

형식개념분석기법은 개념격자(concept lattice)라는 수학적 모델을 기반으로 하는 데이터 분석기법의 일종으로, 본 장에서는 개념을 형성하고 이를 분류하는 기본적인 정의만을 중심으로 다룬다.

[정의1] Formal context $K=(G,M,I)$ 는 객체들의 집합 G 와 속성들의 집합 M , 그리고 G 와 M 사이의 이항관계 $I \subseteq G * M$ 로 구성으로 G 와 M 의 관계를 반영한다. 즉 G 와 M 의 원소들은 각각 해당 Formal context의 객체들과 각 객체들이 가질 수 있는 속성들을 나타낸다. 또한 어떤 객체 g 가 속성 m 을 가지고 있을 경우, gIm 또는 $(g,m) \in I$ 로 나타내며, g 는 m 을 갖는다는 것을 의미한다.

Formal context는 형식개념분석기법의 기본이 되는 구조로서, 행렬 형태로 나타낼 수 있으며, 해당 행렬의 행과 열의 헤더는 Formal context를 구성하는 객체들과 속성들로 구성된다. 그리고 행렬의 각 셀에 대해서는 해당 셀에 관련된 객체와 속성이 이항관계 I 를 만족할 경우에는 'X'표시를 하고, 이외의 경우는 공란으로 남겨둔다.

[정의2] 임의의 Formal context $K=(G,M,I)$ 에 대하여, $O \subseteq G$, $A \subseteq M$ 일 때, $intent(O)=A \wedge extent(A)=O$ 를 만족하는 (O,A) 를 형식개념(Formal context)라 한다(단, $intent(O):=\{a \in M \mid \forall o \in O:(o,a) \in I\}$, $extent(A):=\{o \in G \mid \forall a \in A:(o,a) \in I\}$).

임의의 $O \subseteq G$ 에 대하여, $intent(O)$ 에 의해 O 의 모든 객체들이 공통적으로 갖는 속성들의 집합을 구할 수 있다. 역으로 $A \subseteq M$ 에 대하여, $extent(A)$ 에 의해 A 의 속성들을 갖는 객체들의 집합을 구할 수 있다.

결국, 정의2에 따라 각 개념들은 (O,A) 와 같은 형태의 쌍(pair)으로 정의되며 객체집합 O 는 속성집합 A 의 extent이며, 동시에 속성집합 A 는 객체집합 O 의 intent가 된다. 이와 같은 방법으로 formal context $K=(G, M, I)$ 로부터 추출한 모든 개념들의 집합을 $B(K)$ 로 나타내며 $B(K):=\{(O,A) \in 2^G * 2^M \mid intent(O)=A \wedge extent(A)=O\}$ 와 같이 정의된다.

이렇게 정의된 intent와 extent를 토대로 하여 상위-하위개념관계를 정의할 수 있다.

[정의3] 주어진 Formal context의 임의의 형식개념 $(O_1, A_1), (O_2, A_2) \in B(K)$ 에 대하여, “상위-하위개념 관계(Superconcept-Subconcept relation)” $(O_1, A_1) \leq (O_2, A_2)$ 는 일종의 반 순서관계(partial order relation)로서 $(O_1, A_1) \leq (O_2, A_2) \Leftrightarrow O_1 \supseteq O_2 (\Leftrightarrow A_1 \supseteq A_2)$ 와 같이 정의된다.

Formal context $K=(G,M,I)$ 로부터 만들어진 모든 개념들 간의 상위-하위개념관계 \leq 는 일종의 반순서관계(partial order relation)에 해당하며, 개념들과 그들 사이의 상위-하위개념관계에 의해 만들어진 계층적 개념구조(Concept hierarchy)를 개념격자(Concept Lattice)라고 부르고 $L=(B(K), E_{\leq})$ 과 같이 표현한다. 이와 같은 개념격자를 Hasse Diagram을

사용하여 가시화할 수 있다.

이러한 수학적 정의를 시스템으로 구성하면 분석코드1과 같다.

<분석코드 1> GenerateConcepts and BuildConceptLattice

```
//INPUT : a formal context K := (G, M, I)
//OUTPUT : Concept Lattice L : (B(K), E≤)
for all g∈G(K) do
  B(K) ← B(K) ∪ (extent(intent(g)), intent(g));
end for
for all c∈B(K) do
  for all g∈(G-extent(c)) do
    X ← extent(c) ∪ {g};
    if (extent(intent(X)), intent(X)) ∉ B(K) then
      B(K) ← B(K) ∪ (extent(intent(X)), intent(X));
    end if
  end for
end for
for all c1∈B(K) do
  for all c2∈B(K)-{c1} do
    if (c1 ≤ c2) ∧ (∄ c3 ∈ B(K)-{c1, c2} [(c1 ≤ c3) ∧ (c3 ≤ c2)]) then
      E ≤ ← E≤ ∪ {(c1, c2)};
    end if
  end for
end for
```

3. 사례분석

본 연구에서는 파일단위의 상이한 지리정보자료를 통합데이터베이스로 구축하는 과정에서 형식개념분석의 활용성을 검토하기 위해 사례분석을 수행하였다. 활용자료는 토지이용자료와 중분류 토지피복자료를 사용하였으며, 분석도구로는 Opensource 기반의 Toscana-J 1.61을 사용하였다. 그림1과 같이 중분류 토지피복자료와 토지이용자료를 객체로써 열에 입력하고 이와 대응하는 기준속성으로 교통, 주거, 상업, 공업, 시설, 수자원 등 6개의 속성을 임의로 행에 할당하여 형식개념 테이블을 생성하였다. 여기서 속성은 통합 DB를 반영하는 표준이 될 수도 있으며 또는 업무나 연구분야 등 특정 분야에 사용되는 기준이 될 수도 있다. 그러나 본 연구에서 이러한 기준을 정립하는 것이 목적이 아니므로 본 연구에서는 6개의 기준 속성을 임의로 할당 하였다.

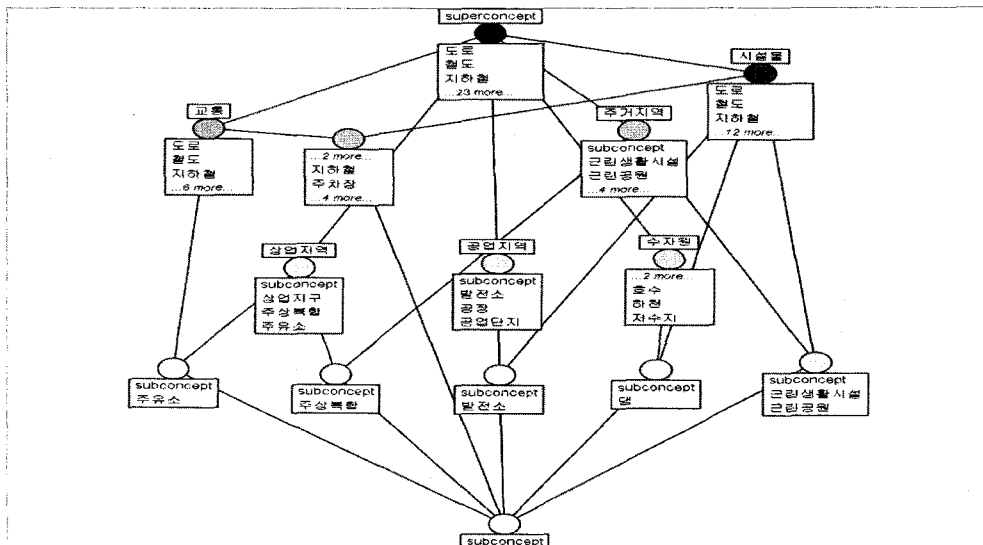
그림1과 같이 속성과 객체가 연관되는 부분에 대해 'X' 표시를 한다. 이 과정은 전적으로 자료를 사용하는 업무담당자나 연구자와 같이 자료를 사용하는 목적 또는 분야에 종속적일 수밖에 없으므로 자동으로 처리할 수 없는 부분이다.

<그림 2> 형식개념 테이블

| | superconcept | 교통 | 주거지역 | 상업지역 | 공업지역 | 시설물 | 주요물 |
|--------------|--------------|----|------|------|------|-----|-----|
| superconcept | x | | x | x | x | | x |
| 도로 | x | x | | | | x | |
| 철도 | x | x | | | | x | |
| 지하철 | x | x | | | | x | |
| 주차장 | x | x | | | | x | |
| 빌딩보도 | x | x | | | | x | |
| 주유소 | x | x | | x | | | |
| 상업지역 | x | | x | | | | |
| 상업지역 | x | | x | | | | |
| 주상복합 | x | | x | x | | | |
| 공업지역 | x | | x | | | | |
| 공업지역 | x | | x | | | x | |
| 상업지구 | x | | | x | | | |
| 주유소 | x | | | | x | | |
| 빌딩보도 | x | | | | x | x | |
| 상수도 | x | | | | | x | |
| 하수도 | x | | | | | x | |
| 가드레일 | x | x | | | | x | |
| 가스관 | x | | | | | x | |
| 호수 | x | | | | | | x |
| 하천 | x | | | | | | x |
| 계곡 | x | | | | | | x |
| 터널 | x | x | | | | x | |

그림1과 같이 행렬이 구축되면 이를 바탕으로 분석코드1을 적용하여 정형화된 개념을 추출하게 된다. 추출된 개념은 그림2와 같이 모델링으로 도식화 될 수 있다. 그림2를 포괄적으로 설명하면 상위개념으로부터 하위개념으로 속성이 상속되고, 하위개념으로부터 해당 객체들이 전파되는 관계를 나타낸다.

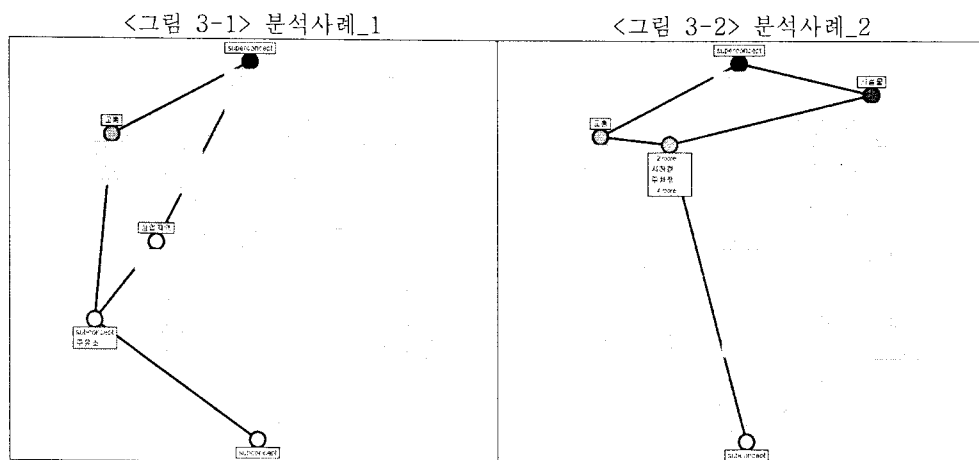
<그림 3> 형식개념을 통한 모델링 구축



한 예로 그림 3-1에서 주유소는 교통개념에 속하면서 동시에 상업지역개념에 속하는 객체이고 그림 3-2에 철도·지하철·주차장·빌딩보도 객체는 교통과 시설물개념을 동시에 상속받는 관계가 된다는 것을 파악할 수 있다.

지리정보자료에서 형식개념분석을 적용한 결과 상이한 두 개의 개별 자료들이 체계적으로

분류되고 또한 통합할 수 있는 데이터모델을 구축할 수 있다는 것을 확인할 수 있었다.



4. 결과 및 논의

각 분야별로 다양한 목적을 위해 구축된 지리정보 자료들을 통합하기 위해서는 우선 이들 간의 중복되는 개념들을 의미론적 차원에서 체계적으로 정립할 필요가 있다. 본 연구에서는 이를 위해 현재 다양한 분야에서 데이터모델 구축에 활용되고 있는 형식개념분석 기법을 지리정보 자료에 적용하였다.

본 연구에서는 토지이용자료와 중분류 토지피복분류도에 분석기법을 접목하여 실험한 결과 두 자료에 중복되는 부분을 일목요연하게 식별할 수 있었고 모델링을 통해 이를 체계적으로 분류할 수 있었다.

본 연구에서 제시한 기법을 앞으로 구축될 지능형 국토의 통합 데이터베이스에 적용할 경우 개별적으로 생산되는 다양한 자료들을 체계적으로 결합하고 관리할 수 있는 기반을 제공할 수 있다는 점에서 시사하는 바가 매우 크다. 특히 형식개념분석은 수학적 기반으로 명확하게 정의되어 있기 때문에 시스템적으로 통일성 있게 자료를 관리하는데 매우 유용하다. 또한 형식개념분석을 수행할 수 있는 다양한 분석도구들이 이미 Opensource 기반으로 개발되어 있기 때문에 불필요한 개발을 최소화 할 수 있다는 장점이 있다.

그러나 앞으로 형식개념분석이 지능형 국토 사업의 통합 데이터베이스 구축에 활용되기 위해서는 몇 가지 선행되어야 할 과제가 남아있다. 우선 공중 및 지상모니터링에서 생산되는 자료가 명확하게 정의되어야 한다. 또한 현재 구축 중에 있는 기본지리정보와 같은 데이터의 표준체계가 모니터링차원에서도 심도 있게 논의되고 구축될 필요가 있다.

따라서 표준화된 체계와 자료의 프로토타입이 국가 모니터링사업을 통해 정립되고, 형식개념분석에 이를 접목하여 응용한다면 통합 데이터베이스의 기반이 되는 데이터모델이 매우 효율적으로 개발될 수 있을 것으로 판단된다.

사사

이 연구는 국토해양부 지능형국토정보사업에서 연구비를 지원받아 수행되었습니다. 감사드립니다.

참고문헌

- 김홍기, 강유경, 황석형, 김동순, 2006, “형식개념분석을 위한 자동화 도구의 개발과 의료분야에서의 적용사례”, 정보처리학회논문지, 13(7), 997-1008.
- 토지이용현황도 및 도로망도 제작, 건설교통부, 1999.
- Kavouras, M., Kokla, M., 2002, A method for the formalization and integration of geographical categorizations, *International Journal of Geographical Information Science*, 16(5), 439-453.