# FCA 기반의 온톨로지 확장

김현식[0] 김인철
경기대학교 전자계산학과
{advance7, kic}@kyonggi.ac.kr

# FCA-Based Ontology Augmentation

Hyun-Sik Kim[0], In-Cheol Kim
Department of Computer Science, Kyonggi University

**요 약**

기존 온톨로지를 확장하는 한 가지 방법은 기존의 기초 개념들로부터 새로운 하위 개념들을 유도해내거나 개념들간의 새로운 관계들을 발견해내는 것이다. 본 논문에서는 의학분야의 기존 온톨로지를 확장하는데 정형적 개념 분석(FCA) 방법이 갖는 잠재적 역할을 분석해보자 한다. 이를 위해 우리는 영역 특정 문서들로부터 기존 개념들의 실례(instance)들을 추출할 수 있다고 가정한다. 본 논문에서는 3 단계로 이루어진 FCA기반의 온톨로지 확장 방법론을 설명하고, MeSH 온톨로지 확장에 관한 경험을 소개한다.

## 1. 서론

하나의 온톨로지는 개념들과 그들의 속성, 그리고 개념들간의 관계로 표현되는 특정 분야의 지식 모델이다. 지식공유를 위해 다양한 분야에서 많은 온톨로지들이 현재 개발되고 있다. 하지만 특정 분야에서 하나의 온톨로지를 개발하는 일은 여전히 어렵고 시간이 많이 소모되는 작업이다. 따라서 우리는 종종 처음부터 완전히 새로운 온톨로지를 구축하기 보다는 기존 온톨로지를 확장하는 것을 더 선호한다. 기존 온톨로지를 확장하고 내용을 풍부하게 만드는 방법의 하나는 기존의 기초 개념들로부터 새로운 하위 개념 또는 하이브리드 개념들을 유도해내거나 개념들간의 새로운 관계들을 발견해내는 것이다. 정형적 개념 분석(Formal Concept Analysis, FCA)은 공통의 속성을 가지는 개체들을 그룹화함으로써 개념들의 격자를 생성하는 효과적인 방법으로 알려져 있다. 본 논문에서는 의학분야의 기존 온톨로지를 확장하는데 이 정형적 개념 분석(FCA) 방법이 갖는 잠재적 역할을 분석해보자 한다.

## 2. 정형적 개념 분석

정형적 개념 분석(Formal Concept Analysis, FCA)은 격자 이론(lattice theory)에 기초한 수학적 데이터 분석 방법이다. FCA는 공통적인 속성들을 가지는 개체들의 그룹을 찾아내는 방법을 제공한다. 우리는 여기서 이 본문에 필요한 만큼만 간략히 FCA의 기초개념을 소개한다.

**정의 1.** 하나의 정형적 문맥은 하나의 튜플 $K := (G, M, I)$ 이다. 이때 $G$는 개체들의 집합(set of objects), $M$은 속성들의 집합(set of attributes), 그리고 $I$는 집합 $G$와 $M$간의 이진관계(binary relation)이다. 즉, 이것은 $I \subseteq G \times M$ 을 의미한다. $(g,m) \in I$ 는 _*개체 g가 속성 m을 가진다(object g has attribute m)*_라고 읽는다.

**정의 2.** 한 개체들의 집합 $A \subseteq G$ 에 대해, 우리는 집합 $A' := \{m \in M \mid \forall g \in A : (g,m) \in I\}$ 을 정의한다. 또, 한 속성들의 집합 $B \subseteq M$ 에 대해, 우리는 집합 $B' := \{g \in G \mid \forall m \in B : (g,m) \in I\}$ .을 정의한다

**정의 3.** 정형적 문맥 $K := (G, M, I)$ 에 기초한 한 정형적 개념은 $A \subseteq G$, $B \subseteq M$, $A' = B$, 그리고 $B' = A$를 만족하는 하나의 $(A,B)$ 쌍이다. 이때 우리는 $A$ 를 개념 $(A,B)$ 의 외연(*extent*), 그리고 $B$ 를 내연(*intent*)이라고 부른다. 또 집합 $\beta(G,M,I)$ 는 정형적 문맥 $K := (G,M,I)$ 에 기초한 모든 개념들의 집합을 나타낸다.

**정의 4.** $(A_1,B_1)$ 와 $(A_2,B_2)$ 가 하나의 정형적 문맥에 기초한 개념들이고, $A_1 \subseteq A_2$ 이거나 혹은 $B_2 \subseteq B_1$ 이면, $(A_1,B_1)$는 $(A_2,B_2)$의 한 하위 개념(subconcept)이다. 이 경우에, $(A_2,B_2)$ 는 $(A_1,B_1)$ 의 한 상위 개념(superconcept)이며, $(A_1,B_1) \leq (A_2,B_2)$ 라고 표기한다. 관계(relation) $\leq$ 는 개념들의 계층적 순서(hierarchical order)라고 부른다. 문맥 $K := (G,M,I)$ 의 가능한 모든 개념들과 이들간의 순서 관계를 개념 격자(concept lattice)라 부르며, $\underline{\beta}(K)$로 표시한다.

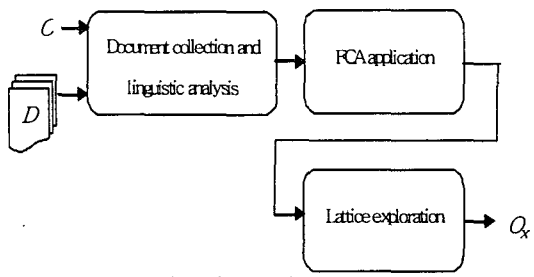



**그림 1 온톨로지 확장 단계들**

### 3. 온톨로지 확장 프로세스

이 절에서는 그림 1과 같은 3 단계로 이루어진 FCA기반의 온톨로지 확장 과정에 관해 설명한다. 온톨로지 확장의 첫 번째 단계에서는 기존 온톨로지 내의 한 개념 계층(concept hierarchy)과 연관된 텍스트 문서들을 수집하고, 자연어 처리를 통해 정형적 문맥을 생성한다. 예컨대, 그림 2와 같은 MeSH 온톨로지 내의 한 개념 계층을 선택하여 그것을 확장하려고 한다고 가정해보자. 먼저 우리는 MEDLINE과 같은 문헌데이터베이스(literature Database)로부터 "Anxiety disorder"와 그것의 하위 개념들에 관련이 있는 문서들을 검색하여야 한다. 우리는 MEDLINE 검색시스템에 약간의 제약조건과 더불어 키워드 "Anxiety disorder"를 포함하는 질의(query)를 제시함으로써 원하는 문서들을 수집할 수 있다. 수집된 문서들에 대한 자연어 분석과정을 통해 우리는 각 개체(object)는 하나의 문서(document)에 대응되고, 각 속성(attribute)은 하나의 특징 키워드(feature keyword)에 대응되는 정형적 문맥(formal context)을 생성할 수 있다. 이때 정형적 문맥의 속성들에 대응되는 특징 키워드들은 온톨로지 내의 기존 개념들의 이름들로 제한할 수 있다.

온톨로지 확장을 위한 두 번째 단계에서는 앞 단계에서 구해진 정형적 문맥을 기초로 정형 개념 분석(FCA) 과정을 실제 적용하여 정형적 개념들의 격자(lattice of formal concepts)를 유도한다. 이 단계는 기존의 FCA과정을 그대로 따라 진행된다. 끝으로, 온톨로지 확장을 위한 마지막 단계에서는 격자 탐색(lattice exploration)을 통해 새로운 유용한 개념들과 순서 관계들을 찾아내고, 이들을 기존 온톨로지에 추가하거나 이들에 맞게 기존 온톨로지를 갱신한다. 새로 생성된 개념들의 대부분 기존 개념들보다 좀더 세부적인 의미를 담고 있는 하위 개념들(subconcepts)이거나 서로 다른 두 가지 이상의 기존 개념들의 속성을 함께 가지는 하이브리드 개념들로 볼 수 있다. 유도된 개념 격자에는 개념들간의 새로운 순서 관계들도 발견될 수 있다. 영역 전문가들(domain experts)의 판단과 조언에 따라, 이들 중 유용한 의미를 지닌 것들만 선별적으로 온톨로지에 포함시킴으로써 기존 온톨로지를 확장한다. 이와 같은 FCA기반의 온톨로지 확장 방법론은 다양한 영역에 적용 가능하다. 하지만 본 논문에서는 의학분야 특히, 정신질환(mental disorder)을 다루는 정신과 영역(psychopathologic domain)에서 이러한 방법론의 가능성을 입증해보려고 한다.

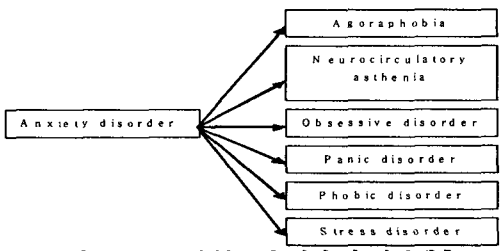



**그림 2 MeSH 온톨로지 내의 한 개념계층**

### 4. MeSH 온톨로지 확장

이 절에서는 MeSH 의학 온톨로지의 작은 예를 가지고 FCA기반의 온톨로지 확장 과정을 예시한다. 원래 본 연구의 실험은 약 43 개의 정신과 관련 개념들을 포함하는 MeSH 온톨로지의 한 부분집합과 MEDLINE으로부터 수집한 약 320 개의 연관 문서들을 기초로 수행되었다. 하지만 본 논문에서는 예시의 목적으로 그림 2와 같은 MeSH 온톨로지의 작은 한 부분집합을 대신 다루기로 한다. 이 온톨로지 부분에는 "Anxiety disorder" 개념과 그것의 하위 개념들로 이루어진 개념 계층을 포함하고 있다. 앞서 언급한 바와 같이, 관련 문서들의 수집을 위해 MEDLINE 검색시스템에 ("Anxiety disorder" AND (Agoraphobia OR "Neurocirculatory asthenia" OR ... OR "Stress disorder"))와 같은 질의(query)를 제시하였고, 그 결과 이들 개념에 연관된 약 24 개의 문서들을 입수하였다. 그런 다음, 단어들간의 경계와 "Neurocirculatory asthenia" 와 같은 두 단어 표현들(2 gram expressions)들을 찾아내기 위해 토큰 분석기(tokenizer)로 문서를 분석하였다. 어휘분석(lexical analysis)은 어휘목록의 영역 세부적인 부분에 대응되는 하나의 항목이 존재하면 단일 단어나 두 단어 표현을 MeSH 온톨로지 상의 기존 개념과 연관지어 준다. 예컨대, "Agoraphobia scale"과 같은 표현은 "Agoraphobia" 개념과 연관되어 진다. 즉, "Agoraphobia" 개념이 기존 MeSH 온톨로지에 존재하고 한 문서 $g$가 "Agoraphobia scale" 표현을 포함하고 있으면, $(g, Agoraphobia) \in I$ 관계가 만족된다. 이와 같은 방식으로, 우리는 그림 3과 같은 하나의 정형적 문맥 $K := (G, M, I)$ 을 생성할 수 있다. 문서들의 집합 $D$는 개체들의 집합으로 $(G := D)$, 개념들의 집합 $C$는 속성들의 집합으로 $(M := C)$ 대응된다. 문서 $g$가 $m$의 한 실례(instance)를 포함하고 있을 때, 관계 $(g, m) \in I$ 는 만족된다.

정형적 문맥 $K := (G, M, I)$ 을 기초로, 우리는 개념 격자 $\underline{\beta}(K)$ 를 계산할 수 있다. 그 결과 얻어진 개념 격자는 그림 4와 같다. 격자의 최상단 노드는 "Anxiety disorder" 개념을 나타낸다. 이 결과는 "Anxiety disorder" 표현이 모든 문서들에 포함되어 있기 때문인 것으로 추정된다.

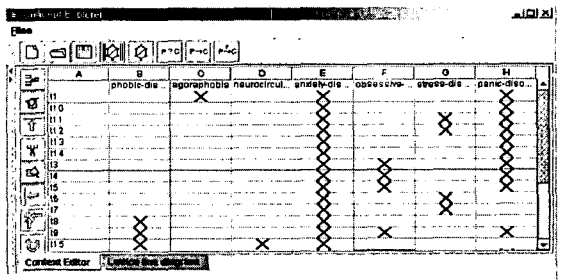



**그림 3 정형적 문맥의 일부**

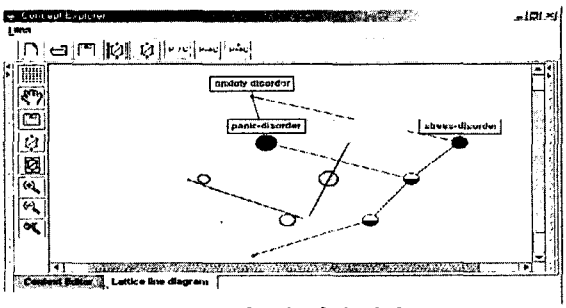
유도된 개념 격자의 상단부는 기존 MeSH 온톨로지의 개념 계층과 거의 일치한다. 하지만 격자의 중간과

하단부에서는 흥미로운 새로운 개념들과 순서 관계들을 발견할 수 있다. 예컨대, 우리는 그림 4의 (a)에서 "Panic disorder" 개념과 "Stress disorder" 개념의 공통 하위 개념(subconcept)들을 찾을 수 있다. 이 하위 개념들은 "Panic disorder" 개념의 성질과 "Stress disorder" 개념의 성질들을 함께 가지고 있는 것으로 볼 수 있다. 또, 이것은 두 질환을 동시에 앓는 환자들이 존재할 수 있다는 사실을 의미하기도 한다. "Panic-stress disorder"로 표기하는 이 새로운 개념은 기존의 개념들을 바탕으로 하는 일종의 하이브리드 개념으로도 볼 수 있다. 본 연구의 자문을 맡아 준 정신과 전문의들은 실제로 정신과 환자들 중 두 질환을 동시에 앓는 환자들이 상당수 존재한다는 사실을 확인해 주었으며, 새로운 독립적인 개념으로서 "Panic-stress disorder"의 가치를 인정해주었다.

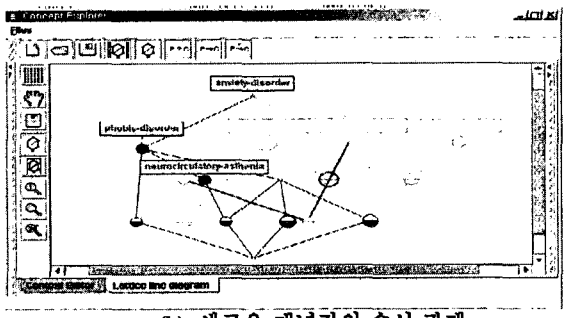
한편, 우리는 그림 4의 (b)에서 "Phobic disorder" 개념과 "Neurocirculatory asthenia" 개념사이에 새로운 순서 관계(ordering relationship)을 발견할 수 있다. 기존 MeSH 온톨로지 개념 계층에서는 이 두 개념간에는 아무런 순서 관계가 없었다. 하지만 새로 유도된 개념 격자상으로는 "Phobic disorder" 개념이 "Neurocirculatory asthenia" 개념의 상위 개념(superconcept)이 되었다. 자문을 맡은 정신과 전문의들은 이 두 개념간의 순서 관계 역시 어떤 측면에서 중요한 의미를 가진다고 인정해주었다. 따라서 온톨로지 확장을 위해 개념 격자에서 발견한 이러한 새로운 개념들과 관계들은 기존 온톨로지에 통합될 수 있다고 판단한다.

## 5. 결론

본 논문에서는 의학분야의 기존 온톨로지를 확장하는데 정형적 개념 분석(FCA) 방법의 잠재적인 역할을 분석해보았다. 우리는 자연어 처리 기술을 적용함으로써 영역 특정 문서들로부터 기존 개념들의 실례(instance)들을 추출할 수 있다고 가정하였다. 본 논문에서는 3 단계로 이루어진 FCA기반의 온톨로지 확장 방법론을 설명하였고, MeSH 온톨로지 확장에 관한 세부경험을 소개하였다. 이러한 연구결과를 토대로 우리는 FCA방법이 온톨로지 확장에 중요한 수단이 될 수 있다고 믿는다.




(a) 새로운 하위 개념들




(b) 새로운 개념간의 순서 관계

그림 4 정형적 개념 격자

## 참고 문헌

[1] B. Diaz-Agudo, P.A. Gonzalez-Calero: Formal Concept Analysis as a Support Technique for CBR, Knowledge-Based Syst. Vol.14, pp.163-171, 2001.

[2] B. Ganter, R. Willer: Formal Concept Analysis: Mathematical Foundations, Springer, Berlin, 1997.

[3] E.A. Mendonca, J.J. Cimino: Automated Knowledge Extraction from MEDLINE Citations, Proc. AMIA Symp. pp.575-579, 2000.

[4] S.L. Moigno, J. Charlet, D. Bourigault, P. Degoulet, M. Jaulent: Terminology Extraction from Text to Build an Ontology in Surgical Intensive Care, Proc. AMIA Symp. pp.430-434, 2002.

[5] G. Stumme, A. Maedche: Merging Ontologies by Means of Formal Concept Analysis, First International Workshop on Databases, Documents, and Information Fusion, Magdeburg, Germany, 2001.