형식개념분석(FCA)을 이용한 RDF 자동생성

윤여창, 손종수, 정인정

고려대학교 과학기술대학 컴퓨터정보학과
충청남도 연기군 조치원읍, 339-700
Tel: +82-41-860-1342, Fax: +82-41-860-0014, E-mail:{kai83, mis026, chung}@korea.ac.kr

Abstract

인터넷의 눈부신 발전으로 사용자가 이용할 수 있는 정보와 자료는 지속적으로 증가하고 있지만, 키워드 기반의 정보검색엔진은 사용자가 검색하고자 하는 정보와 연관성을 가지고 있는 다른 정보들을 보여주지 못하고 있다. 특히 자신이 원하는 논문을 찾는 학술연구자들은 해당 분야와 관련 있는 다른 개념이나 논문을 찾는데 어려움을 겪고 있다. 따라서, 본 논문에서는 여러 학술논문들에서 사용한 키워드 혹은 태그를 이용하여 해당 분야의 개념계층관계를 RDF 포맷으로 생성한다. 본 논문에서 제안하는 방법을 사용하면 학술논문검색에서 검색효율을 높일 수 있을 뿐만 아니라 검색 키워드와 관련 있는 상하위 개념을 알고 있지 못한 사용자에게 더 많은 정보를 제공해 줄 수 있다. 또한 개념계층관계를 RDF 포맷으로 생성하므로 이기종간의 호환성 문제를 해결하고 정보의 활용도를 극대화 했다.

Keyword:

Formal Concept Analysis; Field Association Term; Concept Hierarchy; RDF; Ontology;

1. 서론

인터넷의 눈부신 발전으로 사용자가 이용할 수 있는 정보와 자료는 폭발적으로 증가하고 있다[1]. 셀 수 없을 만큼 많은 정보들 속에서 사용자가 원하는 정보를 신속하고 정확하게 찾아주기 위하여 검색엔진이 등장하였고, 정보검색 기술은 점차 발전하였다. 하지만 기존 키워드 기반의 정보검색 엔진은 단지 키워드를 포함하고 있는 문서들만 결과로 보여주기 때문에 불필요한 정보들이 다수 포함되어 있으며 찾고자 하는 정보와 연관성을 가지고 있는 다른 정보들을 보여주지 못하고 있다. 특히 자신이 연구하는 분야의 논문을 찾는 학술연구자들은 해당 분야와 관련 있는 다른 개념이나 논문을 찾는데 어려움을 겪고 있다[2]. 키워드 기반의 검색엔진을 통해 논문을 검색하면 사용자가 입력한 키워드를 포함하고 있는 논문만 검색 될 뿐이며, 그 키워드와 상하위 계층관계를 갖는 관련된 개념들을 포함하고 있는 논문들을 찾아내기는 어렵다. 더욱이 검색 키워드와 관련된 개념을 알고 있지 못한 학술연구자에게 연관성 있는 논문들을 검색엔진을 통해 검색하는 것은 상당히 힘든 일이다. 따라서 학술연구자들이 논문을 검색할 때 키워드를 포함하고 있는 논문뿐만 아니라, 입력된 키워드와 관련된 상하위 계층개념까지 포함하고 있는 논문까지 검색할 수 있다면 논문검색에 들이는 노력은 크게 줄어들 수 있다. 본 논문에서는 학술연구자들의 논문검색 어려움을 해결하기 위하여 여러 학술논문들에서 사용한 키워드 혹은 태그를 이용하여 해당 분야의 개념계층관계(Concept Hierarchy)를 RDF(Resource Description Framework) 포맷으로 생성한다. 특정 분야의 개념계층관계를 만들기 위하여 본 논문에서는 형식개념분석(Formal Concept Analysis : FCA)[3]을 사용한다. 형식개념분석은 데이터 분석을 위한 클러스터링 기법으로서 객체집합(Object Set)과 속성집합(Attribute Set), 그리고 객체와 속성간의 이항관계(Binary Relation)를 이용하여 개념망(Concept Lattice)을 구축할 수 있는 방법을 제공한다[4]. 그러나 형식개념분석은 객체집합과 속성집합 및 그 이항관계를 사용한 배경도(Formal Context)를 입력으로 주어야 하기 때문에 문서의 키워드나 태그 정보를 그대로 활용하기는 어렵다. 따라서 본 논문에서는 형식개념분석 기법을 사용하기 위하여 논문의 키워드들로부터 분야연상어(Field Association Term)[5]를 추출하여 배경도에 적용하고 개념계층관계를 RDF 포맷으로 변환한다. 결과물은 키워드를 포함하고 있는 문서들로부터 개념과 개념 사이의 상하위 관계를 나타내는 개념계층관계를 표현하는 RDF 포맷 파일로 생성된다.

본 논문에서 제안하는 방법을 사용하면 논문검색에서 검색효율을 높일 수 있을 뿐만 아니라, 검색하고자 하는 키워드와 관련된 상하위 개념을 알고 있지 못한 사용자에게 더 많은 정보를 제공해 줄 수 있다. 또한, 개념계층관계를 RDF 포맷으로

생성하여 이기종간의 호환성 문제를 해결하고 정보의 활용도를 극대화 했다. 제안한 방법으로 생성된 RDF 포맷의 개념계층관계는 개념과 개념간의 상하위 관계를 나타내므로 구체적인 온톨로지[6]를 만들기 위한 기반을 제공한다.

2. 배경

2. 1. 분야연상어(Field Association Term)

분야연상어는 문서의 분야를 식별할 수 있는 한정된 단어들의 집합이다[5]. 즉, 단어를 보고 해당 분야를 인지할 수가 있다. 예를 들면, '홈런' 이라는 단어를 보면 야구의 분야연상어라는 것을 알 수가 있다. 이러한 분야연상어는 문서분류[5], 문서유사도 측정[7], 정보검색[8], 연관검색어 등에 활용될 수 있다.

분야연상어를 추출하는 방법은 크게 세 가지로 나눌 수 있다. 수동으로 추출하는 방법은 사람이 직접 분야연상어를 선택한다. 자동으로 추출하는 방법은 다시 두 가지로 나눌 수 있는데, 문서 전체에서 추출하는 방법[9], 문서 일부에서 추출하는 방법[10]이 있다.

첫째, 사람이 직접 추출하는 방법이 있다. 사람이 직접 선택하는 방법은 정확성은 어느 정도 높아질 수 있지만 시간적인 측면에서 다른 두 가지 방법에 비하여 너무 많은 비용을 지불해야 한다. 또한, 분야연상어의 중복 선택 문제를 가지고 있으며 사람이 수동적으로 하는 방법이기 때문에 추출하는 사람의 주관이 개입될 가능성이 크다.

둘째, 관련된 문서 전체에서 추출하는 방법이 있다. 문서 전체에서 추출하는 방법은 해당 분야의 문서에서 단어의 빈도수를 측정하여 추출하는 방법으로서 사람이 직접 추출하는 방법보다는 시간비용이 크게 줄어든다. 하지만, 잘못된 분야연상어의 추출 가능성을 가지고 있으며 문서 일부에서 추출하는 방법보다 효율성이 떨어진다.

마지막으로, 관련된 문서 일부분에서 추출하는 방법이 있다. 문서 일부에서 추출하는 방법은 문서에서 해당 분야와 가장 유사한 부분을 찾아서 추출하는 방법으로 기존의 방법들 보다 효율성이 뛰어나다. 그러나 해당 분야와 유사한 부분을 찾기 위한 추가적인 시간비용을 지불해야 하며, 여전히 문서에서 불필요한 단어들을 제거하는 등의 자연어처리를 필요로 한다.

본 논문에서는 위에서 언급한 방법들의 자연어처리 시간비용을 크게 줄이고 더 정확한 분야연상어를 선택하기 위하여 문서가 가지고 있는 키워드를 이용하여 분야연상어를 추출한다. 추출된 분야연상어는 형식개념분석의 배경도를 구성하는데 사용된다.

2. 2. 형식개념분석(Formal Concept Analysis)

형식개념분석은 Rudolf Wille에 의해 제안된 것으로서, 데이터 분석을 위한 클러스터링 기법의 하나이다[3][4]. 최근, 형식개념분석은 정보검색, 생태학, 데이터베이스, 인공지능, 의학, 심리학, 소프트웨어 엔지니어링, 언어학 등의 많은 분야에서 연구가 이루어지고 있다[11][12]. 특히, 정보과학 분야에서 전문가를 통해 많은 시간을 투자하여 구축했던 온톨로지 생성을 자동 혹은 반자동으로 구축하기 위한 도구로 활용되고 있다. 최근에는 개념유사도(Concept Similarity) 측정을 통한 온톨로지 병합 및 정렬에 관한 연구도 활발히 이루어 지고 있다[12][13].

형식개념분석은 배경도, 형식개념(Formal Foncept), 개념망의 세 개의 구성요소로 이루어진다. 배경도는 (O, A, R)의 세가지 요소로 구성된다. O는 객체집합, A는 속성집합, R은 O와 A의 이항관계를 나타낸다. $o \in O$, $a \in A$ 인 oRa는 객체 o가 속성 a를 포함하는 관계를 가지고 있다고 말한다. 배경도는 표 1과 같이 행렬로 표현하며, 행은 객체들을 나타내며 열은 속성들을 나타낸다. 표 1에 나타낸 배경도를 O={1, 2, 3, 4, 5}, A={a, b, c, d, e}, R={(1, a), (1, b), (1, c), (1, d), (1, e), (2, a), ... , (5, b), (5, e)} 로 표현할 수 있다.

표 1 – 배경도 예시

	a	b	c	d	e
1	X	X	X	X	X
2	X	X			
3	X	X	X	X	
4	X		X		X
5	X	X	X	X	X

형식개념은 객체들과 속성들 사이의 관계를 반영하며, $X \subseteq O$, $Y \subseteq A$를 만족하는 (X, Y) 쌍으로 표현된다. 또한, 개념을 추출하기 위해서는 (1)과 같은 수식이 정의된다.

$$X' = \{a \in A \mid oRa, \forall o \in X\}$$
$$Y' = \{o \in O \mid oRa, \forall a \in Y\} \quad (1)$$

(1)수식을 바탕으로 형식개념을 추출하기 위해서는 $X \subseteq O$, $Y \subseteq A$, X'=Y, Y'=X를 만족하는 (X, Y) 쌍을 찾아내야 한다. 또한, 객체집합 X는 외연(Extent), 속성집합 Y는 내연(Intent)으로 부른다. 예를 들어, 표 1에서 X={1, 4, 5}, Y={a, c, e}라고 하면 X'={a, c, e}, Y'={1, 4, 5}이므로 ({1, 4, 5}, {a, c, e})는 하나의 형식개념이 된다.

위와 같은 형식개념들로부터 개념망을 구성하기 위해서는 배경도에 존재하는 모든 형식개념들을 찾아내야 한다. 모든 형식개념들로부터 형식개념들

사이의 상속관계를 찾아 개념망을 구성한다.

$$(X_1, Y_1) \le (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow Y_2 \subseteq Y_1) \quad (2)$$

주어진 두 형식개념 (X_1, Y_1), (X_2, Y_2)에 대해 수식(2)를 만족하면 두 형식개념은 상속관계를 가지고 있다고 한다. 즉, 형식개념 (X_1, Y_1)은 (X_2, Y_2)에 대하여 하위개념(Subconcept), 형식개념 (X_2, Y_2)은 (X_1, Y_1)에 대한 상위개념(Superconcept)이라 부른다. 개념망은 이러한 모든 상속관계를 찾아서 그래프로 나타낸 것이다. 그림 1은 표 2로부터 상속관계를 찾아내고, 찾아낸 형식개념들의 상속관계를 이용하여 개념망으로 나타낸 것이다.

표 2 – 표1로 부터 추출된 형식개념

	형식개념(Formal Concept)
1	({a},{1, 2, 3, 4, 5})
2	({a, b},{1, 2, 3, 5})
3	({a, c},{1, 3, 4, 5})
4	({a, c, e},{1, 4, 5})
5	({a, b, c, d},{1, 3, 5})
6	({a, b, c, d, e},{1, 5})

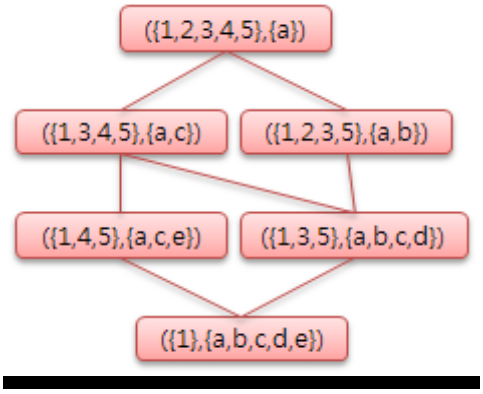

그림 1 – 개념망 예시

하지만, 생성된 개념망의 각 형식개념은 객체집합과 속성집합의 쌍으로 표현되기 때문에 각 형식개념의 단일 이름을 선택하기 위한 방법을 [16]에서 제시하였다. [16]은 개념의 단일 이름을 붙여주기는 하지만 이름이 없는 개념에 수동으로 이름을 붙여주는 작업을 필요로 한다. 따라서 본 논문에서는 이름 없는 개념을 제거하여 각 개념의 이름을 명확하게 해주는 방법을 사용한다. 그림 2의 왼쪽 그림은 [16]에서 제안한 방법으로 만들어진 개념망으로서 제일 마지막 개념의 이름이 존재하지 않는다. 오른쪽 그림은 본 논문에서 사용한 방법으로, 형식개념 목록으로부터 개념망을 생성한 뒤에 이름을 붙이는 것이 아니라 형식개념 목록으로부터 바로 개념계층관계를 생성한다.

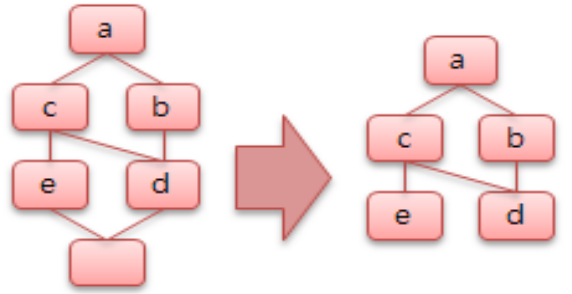

그림 2 – [15]에서 생성한 개념망(왼쪽), 이름 없는 노드를 제거한 개념계층(오른쪽)

3. 관련연구

형식개념분석을 이용하면 그래프 형식의 개념망을 생성할 수 있기 때문에 온톨로지, 개념망, 개념계층관계 등을 만들기 위한 많은 연구가 이루어지고 있다. 형식개념분석을 이용하여 온톨로지를 생성하는 대다수의 연구들은 복잡한 자연어처리 과정을 거쳐 형식개념분석의 배경도를 구성한다[2][11][14][15][16]. [11]은 특정 단어를 포함하고 있는 문서로부터 그 단어의 빈도수를 측정하고 생성된 개념망의 개념들 사이의 유사도를 측정하여 온톨로지를 구성하였다. 하지만 문서를 파싱(Parsing)하는 자연어처리 과정이 존재하여 많은 시간비용을 필요로 한다. [14]는 PFODM(Product Family Ontology Development Methodology)를 소개하여 제품 개발 디자인에 도움을 줄 수 있는 온톨로지를 개발하였다. PFODM에서는 배경도를 구성하는 새로운 방법으로 Multi-context cross table를 제시하였으나, 구성하는 과정 중에 사람의 손을 필요로 하는 단점을 가지고 있다. [15]는 개념망을 구성하고 그래프 형식의 개념망을 트리 형식의 계층구조로 변환하지만, 배경도를 구성하기 위하여 자연어처리 과정을 거친다. [2]는 학술연구자들을 위한 개념계층관계 구성방법을 제안하였다. 퍼지논리를 형식개념분석에 적용시켜 개념망을 구성하고 생성된 개념들의 유사도를 측정하여 개념들을 클러스터링하는 방법을 제안하였다. [16]은 형식분석개념으로 생성된 개념망의 개념 이름을 명시화 해주는 방법을 제안하고 있으나, 이름 없는 개념에 수동으로 이름을 붙여주는 작업이 필요하기 때문에 완전한 자동 생성이 아니다.

배경도를 구성하는 과정에서 자연어처리에 소요되는 시간을 줄이기 위하여 본 논문에서는 문서의 키워드를 이용하여 배경도를 구성한다. 또한, [16]의 반자동 생성을 자동화 하기 위한 가지치기(pruning) 방법을 제안한다.

4. 개념계층관계

4. 1. 개관

학술연구자의 논문검색의 효율성을 높이기 위한 개념계층관계를 RDF 포맷으로 생성하기 위해서, 본 논문은 5단계의 과정을 거친다.

단계 1. 서지정보 획득 및 키워드 추출

단계 2. 키워드를 이용한 분야연상어 추출

단계 3. 배경도 구성 및 개념 추출

단계 4. 개념계층관계 생성

단계 5. 개념계층관계의 RDF 변환

위의 과정 중에서 서지정보 획득을 위한 분야 선택을 제외한 모든 단계는 사람의 개입이 필요하지 않다.

4. 2. 서지정보 획득 및 키워드 추출

개념계층관계를 RDF 포맷으로 생성하는 첫 단계인 서지정보 획득을 위하여 본 논문에서는 사이언스 다이렉트(http://www.sciencedirect.com) 사이트를 선택하였다. 이 사이트에서는 연구논문뿐만 아니라 논문의 서지정보를 처리하기 쉬운 형태로 제공한다. 또한 본 논문에서는 데이터마이닝 분야를 선택하여 '데이터마이닝(Data Mining)' 키워드를 포함하는 논문 1000편의 서지정보를 사이트로부터 수집하였다. 수집된 1000편의 서지정보에서 키워드들만 추출하여 모든 키워드들의 빈도수를 측정하였다. 표 3은 '데이터마이닝' 키워드를 포함하고 있는 논문 1000편의 키워드들과 그 빈도수의 일부를 나타낸 것이다. 키워드 중, '데이터마이닝'을 제외한 데이터마이닝 분야의 총 키워드 개수는 2842개이다.

표 3 – 키워드 목록 및 빈도수 일부

키워드	빈도수
Association Rule	108
Classification	75
Clustering	71
Knowledge Discovery	53
Decision Tree	49
Neural Network	43
Machine Learning	41
Genetic Algorithm	35
Fuzzy Set	26
Feature Selection	25
Support Vector Machine	22

4. 3. 분야연상어 추출

분야연상어는 문서의 분야를 식별할 수 있는 한정된 단어들의 집합으로서 분야연상어를 통해 해당 분야를 식별할 수 있다. 본 논문에서는 문서의 키워드를 이용하여 분야연상어를 추출한다. 키워드를 이용하면 다음과 같은 장점을 가지고 있다. 첫째, 키워드 혹은 태그는 문서작성자에 의해서 만들어진 것이므로 기계적인 처리에 의해 인덱싱을 하는 것 보다 더 정확하게 문서를 요약하고 있다. 둘째, 하나의 문서에 키워드들은 두 개 이상으로 구성되어 있으며, 관련성을 가지고 있으므로 서로의 분야연상어가 될 수 있다. 셋째, 키워드의 대부분은 명사형으로 구성되어 있기 때문에 힘들고 어려운 자연어처리를 거의 필요로 하지 않는다. 넷째, 많은 웹 사이트에서 키워드나 태그를 포함한 문서를 사용하기 편리한 형태로 제공해 주고 있다.

키워드 목록으로부터 분야연상어를 추출하는 방법은 다음과 같다. 각 키워드의 가중치를 계산하고 임계값(Threshold)을 설정하여 임계값 이상의 가중치를 갖는 키워드들을 해당 분야의 분야연상어로 추출한다. 가중치는 빈도수, 전체 키워드에서의 비율, 문서의 길이를 고려한 정규화 방법 등을 사용할 수 있으며, 본 논문에서는 임계값을 2로 설정하고 가중치 값을 빈도수로 하여 2이하의 가중치 값을 갖는 키워드를 제거하여 분야연상어를 추출하였다. 데이터마이닝 분야에서 2842개의 키워드들 중 분야연상어로 추출된 키워드의 개수는 225개이다.

4. 4. 형식개념분석 배경도 구성 및 개념 추출

형식개념분석은 데이터 분석을 위한 클러스터링 기법의 하나로서, 개념들의 상하위 계층관계를 표현하는 개념망 생성방법을 제공한다. 형식개념분석을 이용하여 개념망을 생성하기 위해서는 먼저 배경도를 구성해야 한다. 배경도는 객체집합과 속성집합, 그리고 객체와 속성간의 이항관계를 행렬로 나타낸다. 본 논문에서는 하나의 분야를 객체로 보고, 분야연상어를 객체의 속성으로 본다.. 즉, 데이터마이닝 분야를 하나의 객체로 보고 데이터마이닝의 분야연상어인 연관규칙(Association Rule), 분류(Classification), 클러스터링(Clustering), 지식발견(Knowledge Discovery) 등이 데이터마이닝의 속성들이 된다. 그리고 각 속성들을 다시 하나의 분야로 보고, 각각의 분야연상어를 추출한다. 다시 말해 연관규칙, 분류, 클러스터링 등을 하나의 분야로 선택하여 단계 2, 3을 수행한다. 표 4는 각 분야의 분야연상어 목록의 일부를 나타낸 것이다. 본 논문에서는 데이터마이닝 분야를 포함하여 데이터마이닝의 분야연상어 상위 18개, 총 19개 분야의 분야연상어를 추출하였다.

연상규칙 (Association Rule)	분류 (Classification)	클러스터링 (Clustering)	데이터마이닝 (Data Mining)	의사결정나무 (Decision Tree)
Data mining	Support Vector Machine	Fuzzy clustering	Association rule	Data mining
Association rule mining	Data mining	Data mining	Classification	Classification
Knowledge discovery	Pattern classification	Hierarchical clustering	Clustering	Neural network
Fuzzy set	Machine learning	Clustering algorithm	Knowledge discovery	Genetic algorithm
Classification	Neural network	Genetic algorithm	Decision tree	Feature selection
Clustering	Feature selection	K means clustering	neural network	Pattern recognition
Fuzzy association rule	Feature extraction	classification	Machine learning	Fuzzy decision tree
Decision tree	Genetic algorithm	Data clustering	Genetic algorithm	Decision tree classifier
algorithm	Text classification	Fuzzy c mean	Fuzzy set	Computational complexity
Incremental mining	Decision tree	Neural network	Feature selection	Lower bound

표 4 – 분야별 분야연상어 목록의 일부

표 5 – 구성된 배경도의 일부 (AR : Association Rule, CA : Classification, CL : Clustering, DM : Data Mining, DS : Data Stream, DW : Datawarehouse, DT : Decision Tree, FS : Feature Selection)

	AR	CA	CL	DM	DS	DW	DT	FS
AR	X	X	X	X	X	X	X	
CA	X	X	X	X			X	X
CL	X	X	X	X	X		X	X
DM	X	X	X	X	X	X	X	X
DS	X		X	X	X			
DW	X			X		X		
DT	X	X	X	X			X	X
FS		X	X	X			X	X

표 5는 표 4의 분야별 분야연상어 목록으로부터 구성된 배경도이다. 배경도는 객체가 해당 속성을 가지고 있으면 그 속성과 이항관계를 갖는다. 예를 들어, 표 4에서 '분류(Classification)'객체는 '데이터마이닝(Data Mining)'을 분야연상어(속성)로 가지고 있으므로 이항관계가 성립하지만, '데이터스트림(Data Stream)'은 분야연상어로 가지고 있지 않으므로 이항관계가 성립하지 않기 때문에 표 5의 배경도에서 이항관계 표시가 되어있지 않다. 형식개념분석은 배경도로부터 형식개념들을 추출하고, 형식개념으로부터 개념망을 생성한다. 본 논문에서 생성된 배경도로부터 만들어진 형식개념은 총 124개이다.

4. 5. 개념계층관계 생성

형식개념분석은 형식개념들의 상속관계를 이용하여 개념망을 생성한다. 하지만, 생성된 개념망의 개념은 객체집합과 속성집합의 쌍으로 표현되기 때문에 이름의 명시화 작업이 필요하다. 기존의 이름 명시화를 위한 방법은 이름없는 개념(노드)을 생성하고 그 개념의 이름을 사람이 수동으로 붙여야하는 문제가 있다. 따라서 형식개념들로부터 이름없는 개념(노드)의 생성 없이 개념계층관계를 만들기 위한 알고리즘이 필요하며, 알고리즘은 표 6은 같다. 표 7은 알고리즘의 'AscendingSort(A);'까지 적용한 결과의 일부로서 124개의 형식개념으로부터 속성집합만 추출한 목록을 오름차순정렬로 나타낸 것이다.

표 6 – 개념계층관계 생성 알고리즘

Input : C (Set of Concepts)

Output : CH (Concept Hierarhcy)

```
BuildConceptHierarchy(C)
{
  // A : 속성집합목록
  A = ExtractAttribute(C);

  // A를 오름차순 정렬
  AscendingSort(A);

  // attr : 속성집합목록의 원소인 속성집합
  for (all attr in A)
  {
        // A에서 attr의 최대부분집합들을 찾음
        M = MaximumSubset(attr, A);

        if (attr - ∪m_i = ∅ )
              continue;
        else
        {
              // 새로운 노드 생성
              node = new node(attr);

              // attr의 부모를 최대부분집합으로
                생성된 노드로 설정
              node.parent = FindNode(M, CH);

              // 개념계층관계에 노드 추가
              CH.AddNode(node);
        }
  }
  return CH;
}
```

표 7 – 형식개념으로부터 추출한 속성집합 일부 (DM : Data Mining, AR : Association Rule, CL : Clustering, FZ : Fuzzy Logic, KD : Knowledge Discovery, DW : Datawarehouse, DS : Data Stream)

	속성집합		속성집합
1	{DM}	6	{DM, AR, DW}
2	{DM, AR}	7	{DM, CL, DS}
3	{DM, CL}	8	{DM, AR, CL}
4	{DM, FZ}	9	{DM, KD, FZ}
5	{DM, KD}	10	{DM, AR, FZ}

예를 들어 표 7의 7번 속성집합이 알고리즘에 적용될 때는 다음과 같다. 7번 속성집합 attr = {DM, CL, DS}의 최대부분집합을 찾으면 M = {{DM, CL}}이고 attr - $\cup m_i$ = {DM, CL, DS} – {DM, CL} = {DS}이다. 따라서 7번 속성집합으로 생성된 노드 DS는 {DM, CL}로 생성된 노드 CL을 부모노드로 갖게 된다. 또한, 8번 속성집합 attr = {DM, AR, CL}의 최대부분집합을 찾으면 M = {{DM, AR}, {DM, CL}}이다. attr - $\cup m_i$ = {DM, CL, DS} – {DM, CL} = ⌀ 이므로 8번 속성집합은 노드의 생성이 없다. 그림 3은 표 6의 알고리즘을 적용하여 데이터마이닝(Data Mining) 분야를 포함한 총 19개 분야의 분야연상어를 이용하여 생성된 개념계층관계이다.

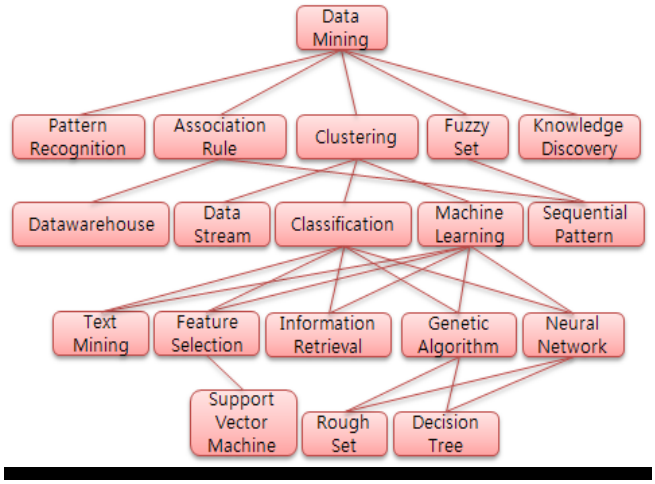

그림 3 – 데이터마이닝 분야의 개념계층관계

4. 6. RDF 변환

생성된 개념계층관계는 이기종간의 호환성 문제를 해결하기 위하여 본 논문에서는 RDF를 선택하였다. RDF는 지식표현을 위한 도구로서, 개념을 정형화하여 표현하고 개념과 개념 사이의 관계를 나타낼 수 있을 뿐만 아니라, 이기종간의 상호 호환성을 보장한다. 본 논문에서는 개념 계층관계에서 각 노드를 표현하기 위하여 rdfs:Class 태그를 사용하였으며, 상하위 상속관계를 표현하기 위하여 rdfs:subClassOf 태그를 사용하였다. 표 8은 개념계층관계를 RDF로 변환하기 위한 알고리즘이다. 알고리즘은 뿌리노드부터 자식노드를 재귀적으로 순회하면서 각 노드를 위한 RDF 표현을 추가한다. 그림 4는 그림 3의 개념계층관계를 표 8의 알고리즘을 적용하여 생성된 RDF의 일부이다.

표 8 – RDF 변환 알고리즘

Input : N(Node in Concept Hierarchy)
Output : RDF(RDF Description)

MakeRDF(N, out RDF)
{
// 노드가 RDF에 표현되지 않은 경우
if (!IsExist(N, RDF))
{
// 노드의 부모노드를 검색
parent = FindParentNode(N);
 	
// RDF 표현에 노드 N을 기술
RDF += RDFClass(N, subClassOf(parent));
 	
// 노드 N이 자식노드를 가지고 있는 경우
if (hasChild(N))
{
// 각 자식노드에 대해 재귀호출
for (each ChildNode)
{
MakeRDF(ChildNode, RDF);
}
}
}
}

```
<?xml version="1.0" encoding="euc-kr" ?>
<!DOCTYPE rdf:RDF (View Source for full doctype...)>
- <rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xml:base="http://www.semanticweb.org/ontologies/2009/9/DataMining">
    <rdfs:Class rdf:about="Data_Mining" />
  - <rdfs:Class rdf:about="Association_Rule">
      <rdfs:subClassOf rdf:resource="#Data_Mining" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Pattern_Recognition">
      <rdfs:subClassOf rdf:resource="#Data_Mining" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Clustering">
      <rdfs:subClassOf rdf:resource="#Data_Mining" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Fuzzy_Set">
      <rdfs:subClassOf rdf:resource="#Data_Mining" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Knowledge Discovery">
      <rdfs:subClassOf rdf:resource="#Data_Mining" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Datawarehouse">
      <rdfs:subClassOf rdf:resource="#Association_Rule" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Data_Stream">
      <rdfs:subClassOf rdf:resource="#Clustering" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Classification">
      <rdfs:subClassOf rdf:resource="#Clustering" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Machine_Learning">
      <rdfs:subClassOf rdf:resource="#Clustering" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Sequential_Pattern">
      <rdfs:subClassOf rdf:resource="#Association_Rule" />
      <rdfs:subClassOf rdf:resource="#Fuzzy_Set" />
    </rdfs:Class>
  - <rdfs:Class rdf:about="Text_Mining">
      <rdfs:subClassOf rdf:resource="#Classification" />
      <rdfs:subClassOf rdf:resource="#Machine_Learning" />
```

그림 4 – RDF로 표현된 개념계층관계

5. 결론

기존의 키워드 기반 검색은 키워드와 관련된 정보를 보여주지 못하고, 특히 학술연구자들에게 연구분야와 관련된 정보를 보여주지 못하고 있다. 따라서, 학술연구자들의 효율적인 논문검색 및 연구를 위해 본 논문에서는 형식개념분석을 이용하여 개념들의 상하위 관계를 나타내는 개념계층관계를 생성하였으며, 형식개념분석의 배경도를 생성하기 위해 키워드를 이용한

분야연상어를 추출하였다. 또한 생성된 개념계층관계를 RDF로 변환하여 이기종간의 상호 호환성 문제를 해결하고 지식의 활용도를 극대화하였다. 본 논문에서 제시한 방법은 모든 과정을 자동화함으로써 개념계층관계를 만들기 위한 시간비용을 줄일 수 있을 뿐만 아니라, RDF를 이용하여 표현하였기 때문에 개념과 개념들 사이의 관계를 기술하는 온톨로지를 생성하는 기반을 제공해 줄 수 있다. 또한 키워드와 태그를 기반으로 하기 때문에 자연어처리를 위한 복잡하고 어려운 과정을 제거할 수 있다. 본 논문은 학술논문을 대상으로 하였지만, 키워드를 가지고 있는 다른 모든 문서들에도 적용 가능하며, 더욱이 블로그와 같이 키워드나 태그를 사용하는 웹 2.0 어플리케이션에 크게 활용될 수 있을 것으로 기대된다. 향후 개념계층관계를 표현한 RDF를 이용하여 학술연구자들을 위한 구체적인 활용 방법에 대하여 연구하고자 한다.

6. 참고문헌

[1] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). "The Semantic Web," *Scientific American*, pp. 28-37.

[2] Quan, T.T., Hui, S.C., Fong, A.C.M., and Cao, T.H. (2004). "Automatic generation of ontology for scholarly semantic web," *In S. A. McIiraith, International Conference on Semantic Web, ISWC 2004, LNCS 3298*, pp. 726-740.

[3] Wille, R. (1982). "Restructuring lattice theory: An approach based on hierarchies of concepts," In I. Rival (Ed.), pp. 445-470.

[4] Ganter, B., and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*, Springer, Berlin

[5] Fuketa, M., Sangkon, L., Takako, T., and Makoto, O. (2000). "A document classification method by using field association words," *Information Science Journal*, Vol. 126(1-4), pp. 57-70.

[6] Gruber, T.R. (1993). "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition*, Vol. 5(2), pp. 199-220.

[7] Atlam, El-S., Fuketa, M., Morita, K., and Aoe, J.I. (2003). "Documents similarity measurement using field association terms," *Information Processing and Management*, Vol. 39, pp. 809-824.

[8] Rokaya, M., Atlam, El-S., Fuketa, M., Dorji, T.C., and Aoe, J.I. (2007). "Ranking of field association terms using co-word analysis," *Information Processing and Management*, Vol. 44(2), pp.738-755.

[9] Atlam, El-S., Elmarhomy, G., Morita, K., Fuketa, M., Aoe, J.I. (2006). "Automatic building of new field association word candidates using search engine," *Information Processing and Management Journal*, Vol. 42(4), pp. 951-962.

[10] Sharif, U., Elmarhomy, G., Atlam, El-S., Fuketa, M., Morita, K., and Aoe, J.I. (2007). "Improvement of building field association term dictionary using passage retrieval," *Information Processing and Management Journal*, Vol. 43(8), pp. 1793-1807.

[11] Weng, S.S. (2006). "Ontology construction for information classification," *Expert Systems with Applications*, Vol. 31(1), pp. 1-12.

[12] Formica, A. (2007). "Concept similarity in Formal Concept Analysis: An information content approach," *Knowledge-Based Systems*, Vol. 21, pp. 80-87.

[13] Stumme, G., and Mädche, A. (2001). "FCA-Merge: Bottom-up merging of ontologies," *In 7th International Conference on Artificial Intelligence*, pp. 225-230.

[14] Juotirmaya, N., Timothy, S., Kumara, S., Shooter, S. (2006). "A methodology for product family ontology development using formal concept analysis and web ontology language," *Journal of Computing and Information Science in Engineering*, Vol. 6(2), pp. 103-113.

[15] Cimiano, P., Hotho, A., Staab, S. (2005). "Learning concept hierarchies from text corpora using formal concept analysis," *Journal of Artificial Intelligence Research*, Vol. 24, pp. 305-339.

[16] Haav, H. (2004). "A semi-automatic method to ontology design by using FCA," Proceedings of Concept Lattices and their Applications, pp. 13-24.