# 지식 문서에서 도메인 온톨로지를 이용한 개념 추출 기법

문 현 정[†] · 우 용 태[††]


## 요 약

본 논문에서는 도메인 온톨로지를 이용하여 XML 형식의 지식 문서를 분류하고 대표 개념을 효과적으로 추출하기 위한 기법을 제시하였다. 먼저, 도메인 온톨로지는 텍스트마이닝 기법과 통계적 기법을 이용하여 생성하였다. 이를 위해 XML 문서의 구조적인 특징을 이용하여 도메인 대표용어 집합을 구성하였다. 그리고 XML 문서를 효과적으로 분류하기 위한 *DScore*기법과 지식 문서로부터 개념을 추출하기 위한 *TScore* 기법을 제시하였다. 본 논문에서 제안한 기법의 효율성을 검증하기 위하여 295편의 컴퓨터 관련 논문을 대상으로 실험하였다. 실험 결과, 본 연구에서 제안한 도메인 대표 용어 집합을 이용한 분류 결과가 기존의 방법보다 우수한 성능을 보였다. 특히 *TScore*기법에서는 문서에서 출현한 용어의 빈도수는 낮더라도 문서의 개념을 대표할 수 있는 용어를 효과적으로 추출할 수 있음을 보였다. 본 연구는 개념 기반의 검색 기법을 통하여 대량의 지식 문서를 효과적으로 관리하기 위한 지식 관리 모델에 적용할 수 있다.


**키워드 : 온톨로지, 지식 관리 시스템, KDD, 개념기반검색**

# Concept Extraction Technique from Documents Using Domain Ontology

Hyeon Jeong Mun[†] · Yong Tae Woo[††]


## ABSTRACT

We propose a novel technique to categorize XML documents and extract a concept efficiently using domain ontology. First, we create domain ontology that use text mining technique and statistical technique. We propose a *DScore* technique to classify XML documents by using the structural characteristic of XML document. We also present *TScore* technique to extract a concept by comparing the association term set of domain ontology and the terms in the XML document. To verify the efficiency of the proposed techniques, we perform experiment for 295 papers in the computer science area. The results of experiment show that the proposed technique using the structural information in the XML documents is more efficient than the existing technique. Especially, the *TScore* technique effectively extract the concept of documents although frequency of term is few. Hence, the proposed concept-based retrieval techniques can be expected to contribute to the development of an efficient ontology-based knowledge management system.


**Key Words : Ontology, Knowledge Management System, KDD, Concept-based Retrieval**

## 1. 서 론

기존의 지식 관리 모델은 대량의 지식 자원에 대한 편리한 접근을 위한 효율적인 문서관리에 초점을 두고 비구조적 형태의 문서를 대상으로 키워드 기반의 지식 관리 기법을 주로 사용한다. 하지만 이 기법은 여러 가지 문제점이 지적되고 있다. 첫째, 문서에서 추출된 키워드가 문서 내용과 관련이 없거나, 해당 문서의 개념을 대표하는 중요한 키워드가 누락될 수 있다. 둘째, 키워드만으로는 사용자가 원하는 정확한 지식 문서를 검색하기 어렵다. 셋째, 사용자에게 친숙하지 않은 도메인에서는 사용자가 적절한 키워드를 선정하기 어렵다. 넷째, 대량의 문서 집합에서 검색 결과에 대한 정확도와 신뢰도가 떨어질 수 있다[1].

이러한 키워드 기반의 지식 관리 모델의 문제점을 보완하기 위하여 메타 데이터를 이용한 내용 기반 지식 관리 모델에 대한 연구가 활발하게 진행되고 있다. 이 모델은 문서의 내용과 연관된 메타 데이터를 삽입하여 지식을 관리하는 모델로서 질의어와 가장 유사한 개념을 가진 문서를 검색하는 방법이다. 이 모델에서 메타 데이터로 사용하는 용어 집합은 용어에 대한 애매 모호성을 줄이기 위해 시소로스와 같은 통제된 용어 집합을 주로 사용한다. 하지만 이러한 용어 집합에서는 용어간의 의미적인 연관성을 표현하기 어렵다. 또한 지식 문서에서 용어간의 연관 관계를 이용한 개념이나 지식 추론이 어렵다.

※ 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2004-050-D00016)
† 정 회 원 : 창원대학교 컴퓨터공학과 연구교수
†† 종신회원 : 창원대학교 컴퓨터공학과 교수
논문접수 : 2006년 4월 19일, 심사완료 : 2006년 5월 15일

최근에는 메타 데이터로서 온톨로지를 이용하여 지식 관리[2-4], 정보 검색[5-7], 시멘틱 웹[1, 8] 분야에 적용하기 위한 연구가 활발하게 진행되고 있다. 온톨로지는 지식을 표현하기 위한 용어 집합과 용어간의 관계를 통하여 표준화된 용어에 대한 의미를 정의하고 구조화하기 위한 목적으로 사용한다. 또한 온톨로지는 지식 문서를 분류하기 위한 계층구조의 구성, 사용자의 관심 도메인을 지속적으로 관리하기 위한 사용자 모델링, 문서 개념에 대한 추론 등에 응용될 수 있다.

온톨로지 기반의 지식 관리 모델에서 핵심 기능의 하나는 지식 문서를 대표하는 개념을 효과적으로 추출하는 기능이다. 이러한 개념은 사용자가 문서마다 명시적으로 입력한 대표 키워드를 참조하거나 도메인 전문가의 수작업에 의해 구성할 수 있다. 하지만 대표 키워드는 지식 문서를 작성한 사용자가 직접 명시하지 않으면 구성하기 어렵다. 또한 도메인 전문가에 의한 수작업은 많은 시간과 처리 비용이 요구된다. Yahoo에서는 문서의 대표 개념 추출과 유사한 형태의 카테고리 관리를 위해 도메인 전문가 100여명이 투입되었다[9]. 또한 이 방법은 전문가의 관점에 따라 분류 기준에 대한 일관성을 유지하기 어려운 문제점이 있다.

최근에 국내에서도 온톨로지를 이용한 정보검색기법이나 문서간의 연관성을 분석하기 위한 연구가 이루어지고 있다. 최옥경 등은 시맨틱 웹을 기반으로 한 정보검색시스템을 제안하였고[7], 김명숙 등은 온톨로지를 이용한 XML 질의 확장 기법을 제안하였다[10]. 그리고 이무훈 등은 웹 문서간의 의미적 연관성을 기술하기 위한 온톨로지 에디터를 제안하였다[11]. 하지만 아직까지 국내에서 온통로지를 이용한 지식관리 모델에 대한 연구결과는 빈약한 수준이다.

본 논문에서는 도메인 온톨로지를 이용하여 대량의 XML 문서를 체계적으로 분류하고 대표 개념을 자동적으로 추출하기 위한 새로운 기법을 제시하였다. 본 연구는 크게 전처리 과정, 도메인 온톨로지 생성, 도메인 온톨로지의 개념별 연관 용어 집합 구성, 온톨로지를 이용한 지식 문서 분류 그리고 대표 개념 추출 과정으로 구성된다.

먼저, 전처리 과정에서는 XML 문서의 구조적인 특징을 가중치로 반영한 엘리먼트 가중치와 *TWIDF(Term Weight Inverse Document Frequency)* 기법을 이용하여 문서를 대표하는 용어를 추출하였다. 그리고 도메인별 대표 용어 집합은 $\chi^2$ 분포에서 일정한 유의 수준을 만족하는 용어를 대상으로 상관관계와 도메인 가중치를 이용하여 구성하였다. 도메인 대표 용어간의 관계는 용어간의 포함 관계(subsumption)를 분석하여 용어간의 상, 하위 관계를 방향성 그래프에 의해 표현하였다. 그리고 도메인 온톨로지는 연관 규칙 탐사 알고리즘을 사용하여 신뢰도(confidence)와 지지도(support)가 높은 용어 집합을 대상으로 향상도(lift)순으로 Top-N개의 연관 용어 집합으로 구성하였다.

본 연구에서 제안한 온톨로지를 이용한 지식 문서의 개념 추출 기법은 크게 도메인 온톨로지를 이용한 문서 분류, 도메인 온톨로지 개념과 지식 문서간의 매핑 과정으로 나뉘어진다. 본 연구에서는 지식 문서가 속하는 도메인을 효과적으로 분류하기 위한 *DScore(Document Score)* 기법과 온톨로지상의 개념과 지식 문서를 매핑하기 위한 *TScore(Term Score)* 기법을 제시하였다. *DScore* 기법은 지식 문서가 속하는 도메인을 결정하기 위하여 도메인 온톨로지와 지식 문서의 대표 용어간의 유사도 비교하기 위한 기법이다. 임의의 지식 문서는 가장 높은 *DScore* 값을 가지는 도메인으로 분류된다. 그리고 *TScore* 기법은 온톨로지상의 개념과 지식 문서를 매핑하기 위하여 온톨로지의 개념별 연관 용어 집합과 문서에서 추출한 대표 용어간의 유사도를 비교하기 위한 기법이다. 지식 문서에 대한 대표 개념은 해당 도메인에서 가장 높은 *TScore* 값을 가지는 용어를 추출한다.

## 2. 관련 연구

### 2.1 키워드 기반의 지식 관리 모델

키워드 기반의 지식 관리 모델에서 문서로부터 추출한 키워드는 반드시 문서와 연관 있는 용어가 아니거나 문서와 관련된 중요한 키워드가 누락될 수 있다. 또한 이 모델에서는 사용자가 원하는 정보를 정확하게 정형화하기 힘들다. 즉, 키워드만으로는 사용자가 원하는 지식의 수준을 판단하기 어렵다. 또한 사용자는 전문적인 지식 도메인에서 정확한 질의어를 구성하기가 어렵다. 이 경우 질의어와 연관된 연관 용어를 함께 제공하여 사용자가 원하는 정보를 좀 더 효과적으로 검색할 수 있다. 이러한 용어간의 연관 관계는 사용자가 원하는 문서를 검색하거나 질의의 검색 범위를 넓히는 중요한 요소로 사용될 수 있다. 하지만 이 모델은 용어간의 연관 관계를 제공하기 어렵고, 동음이의어와 같은 용어간의 애매 모호성으로 인해 검색의 정확도가 떨어질 수 있다.

### 2.2 메타 데이터를 이용한 내용 기반의 지식 관리 모델

키워드 기반 지식 관리 모델의 문제점을 개선하기 위하여 메타 데이터를 이용한 내용 기반 지식 관리 모델에 관한 연구가 진행되고 있다[2-4]. 이 모델은 지식 문서에 메타 데이터를 삽입하여 관리하는 모델이다. 메타 데이터로 사용할 수 있는 정보는 문서의 도메인, 대표 키워드, 문서의 요약, 문서에서 내포하고 있는 정보, 지식 문서의 위치 정보 등이 있다. 이러한 메타 데이터는 문서를 분류하거나 내용 기반 검색 과정에 이용할 수 있다. 메타 데이터로 주로 사용되는 기법은 시소로스나 온톨로지를 이용한 기법이 널리 사용되고 있다. 시소로스는 문서를 공유하는 사용자간에 용어의 애매 모호성을 줄이기 위해 사용되는 통제된 용어 집합이다. 하지만 시소로스를 이용한 메타 데이터는 유사 용어 집합과 용어간의 상, 하위 관계를 표현할 수 있지만, 용어간의 연관 관계 정의나 추론이 어렵다.

### 2.3 온톨로지 기반의 지식 관리 모델

온톨로지는 'Ontology is a formal, explicit specification

of a shared conceptualization'라고 정의된다[2]. 여기서 formal은 온톨로지를 기계가 이해할 수 있어야 한다는 사실을 의미한다. explicit는 개념의 형태와 개념에 대해 부여된 제약에 대한 명시적인 정의를 의미한다. shared는 온톨로지가 그룹이 공유할 수 있는 합의된 개념이어야 한다는 것을 의미한다. 최근에는 온톨로지를 지식관리를 위한 메타 데이타로 이용하기 위한 연구가 활발하게 진행되고 있다.

$(KA)^2$에서는 온톨로지 기반으로 지식 수집(know- ledge gathering), 지식 조직화 및 구조화(knowledge organization and structuring), 지식 정제(knowledge refinement) 과정으로 구성된 지식 관리 모델을 제시하였다[4]. Ontobroker에서는 온톨로지를 이용하여 웹 문서에 메타 정보를 삽입하기 위하여 HTML을 확장하는 방법과 HTML 문서를 RDF 문서로 변환하는 방법을 제시하였다[5]. FindUR에서는 온톨로지를 이용하여 카테고리의 주제 집합에 대해 클래스간의 관계를 정의하였다. 그리고 카테고리 정보를 문서에 대한 카테고리를 표현하기 위한 메타 정보로 사용하여 질의에 대한 검색 범위를 제한하는데 사용하였다[6]. Ontogator와 MuseumFinland에서는 이미지 데이터에 온톨로지 기반의 메타 데이터를 삽입하여 사용자에게 다양한 관점의 검색 기능과 연관된 이미지를 추천하는 기법을 제시하였다[1, 8].

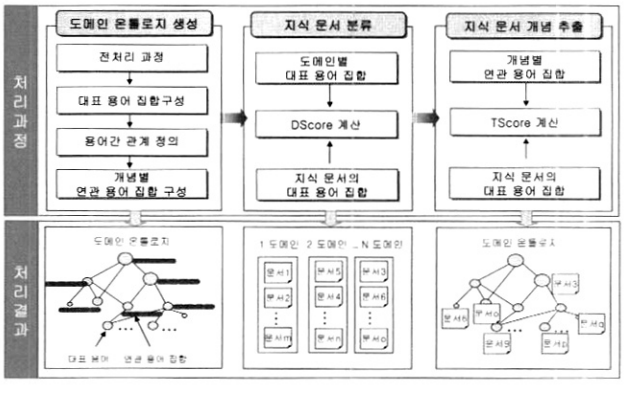
### 2.4 지식 관리를 위한 도메인 온톨로지 구성 방법

대량의 지식을 체계적으로 관리하기 위한 도메인 온톨로지는 크게 객체 지향형과 계층구조형 온톨로지가 있다. 먼저, 객체 지향형 온톨로지는 일반적인 객체 지향형 모델에서 사용하는 형태이다. 이 기법은 개념을 클래스로 정의하고, 각 클래스마다 속성 값을 가지는 property를 정의하는 기법이다. 이 기법에서는 온톨로지를 이용하여 논리적으로 의미 있는 정보 단위별로 메타 데이터를 삽입하고, 검색 결과 또한 논리적인 정보 단위로 추출한다. 하지만 $(KA)^2$의 분석 결과에 의하면 약 5페이지 정도의 문서에 대해 메타 데이터를 삽입하는데 한 시간정도의 시간이 소요되었다. 따라서 제한된 지식 도메인이라 할지라도 기하급수적으로 증가하는 지식 문서에 대해 메타 데이터를 수작업으로 구성하기 어렵다. 또한 문서 작성자가 논리적인 정보 단위별로 지식 문서에 메타 데이터를 추가해야 하는 부담이 있다. 이러한 형태의 객체 지향형 온톨로지는 Wine, Food 온톨로지 등이 있다[12].

계층구조형 온톨로지는 기존의 지식 문서 관리를 위해 주로 사용되는 기법이다. 이 기법은 Yahoo에서처럼 계층구조 형태를 기본 구조로 구성하여 개념간의 상, 하위 관계뿐만 아니라 의미적인 관계까지 연결하는 형태이다. 계층구조형 온톨로지에서 지식 문서는 문서의 대표 개념과 가장 유사한 개념의 온톨로지에 대한 인스턴스로 연결된다. 이 모델에서 검색 결과는 문서 단위가 되며, 키워드 기반의 검색 기법보다 개념적으로 유사한 문서를 효과적으로 검색할 수 있다. 또한 메타 데이터가 세분화될수록 검색의 정확도는 향상될 수 있지만, 메타 데이터 작업에 대한 레벨을 정하는 과정이 복잡해 질 수 있다. 이러한 형태의 계층구조형 온톨로지는 WordNet[13], SENSUS[14], Cyc[15] 온톨로지 등이 있다.

## 3. 도메인 온톨로지의 기반 문서 분류 및 대표 개념 추출 기법

본 논문에서는 XML 문서로 구성된 지식 정보를 체계적으로 관리하기 위한 온톨로지 기반의 지식 관리 모델을 제시하였다. 먼저, XML 문서의 구조적인 특징과 텍스트마이닝 기법을 이용하여 도메인 온톨로지를 생성하였다. 지식 문서는 도메인 온톨로지를 기반으로 도메인을 결정하여 분류하였다. 지식 문서를 대표하는 개념은 해당 도메인 온톨로지와 지식 문서의 대표 용어간의 유사도 비교를 통하여 추출하였다. 지식 문서의 개념은 도메인 온톨로지의 개념별 연관 용어 집합과 지식 문서의 대표 용어간의 유사도를 비교하여 추출하였다. 본 연구에서는 XML 문서의 구조적인 특징을 이용하여 문서를 분류하기 위한 *DScore* 기법과 문서의 개념을 추출하기 위한 *TScore* 기법을 제시하였다. *TScore* 기법에 의해 추출된 대표 개념을 메타 데이터로 사용하여 해당 도메인의 인스턴스로 매핑하였다. 다음 (그림 1)은 본 논문에서 제시한 온톨로지 기반의 지식 관리 모델의 전체적인 개념도이다.




(그림 1) 도메인 온톨로지 기반의 문서 분류 및 대표 개념 추출 기법 개념도

### 3.1 전처리 과정

전처리 과정은 크게 형태소 분석, 도메인 기본 용어 추출, 동의어 처리, 엘리먼트 가중치 계산 과정으로 구성된다. 엘리먼트 가중치 계산 과정은 문서를 대표하는 용어를 추출하기 위하여 XML 문서의 구조적 특성을 이용하여 용어별로 가중치를 부여하는 과정이다. XML 형식의 문서는 엘리먼트와 속성에 의해 문서에 대한 구조적인 정보를 포함하고 있다. 따라서 XML 문서에서는 용어가 속한 엘리먼트에 의해 평면적인 문서보다 용어에 대한 중요도를 쉽게 구별할 수 있다. 일반적으로 문서의 제목이나 요약 부분에서는 전체 문서에 대한 중요한 내용을 축약하여 표현하고 있다. 따라

서 이러한 부분에서 출현하는 용어는 빈도수는 낮아도 문서의 내용을 대표하는 개념으로 간주할 수 있다. 엘리먼트 가중치는 엘리먼트의 레벨과 레벨내의 중요도에 의해 정의된다[16].

본 연구에서는 XML 문서의 구조적인 특징을 이용하여 문서에서 용어에 대한 영향력에 따라 용어별로 가중치를 부여하기 위한 기법을 제시하였다. 이를 위하여 기존의 *TFIDF*(*Term Frequency Inverse Document Frequency*) 기법을 수정한 *TWIDF* 기법을 제안하였다. *TWIDF* 기법에서는 XML 문서의 엘리먼트별로 용어 빈도수와 엘리먼트 가중치의 곱에 대한 합으로 정의한 용어 가중치 값을 계산하여 구조적인 정보를 반영하였다. *TWIDF* 기법에서 특정 도메인에 속한 문서 집합 $C$는 $C=\{c_1,c_2,\cdots,c_L\}$라 정의하고, XML 문서에서 추출된 용어 집합 $T$는 $T=\{t_1,t_2,\cdots,t_M\}$라 정의한다. 그리고 엘리먼트 인스턴스의 집합 $E$는 $E=\{e_1,e_2,\cdots,e_N\}$라 정의한다. 임의의 문서 $c_i$ 에서 용어 $t_k$ 에 대한 *TWIDF* 계산식은 다음과 같다.

$$TWIDF(c_i,t_k)=\sum_{j=1}^{N}(ew_j\times tf_{jk}^{i})\times\log(\frac{|C|}{cf_k}) \quad (1)$$

$ew_j$ : 엘리먼트 인스턴스$e_j$의 가중치
$tf_{jk}^{i}$ : 문서 $c_i$의 엘리먼트$e_j$에서 용어$t_k$의 빈도수
$|C|$ : 문서 수
$cf_k$ : 용어$t_k$가 출현한 문서 수

### 3.2 도메인 온톨로지 생성

도메인 온톨로지 생성은 도메인을 대표하는 용어 집합에서 용어간의 방향성 관계를 구성하는 과정이다. 도메인 대표 용어 집합은 문서별로 *TWIDF* 기법에 의해 추출된 대표 용어를 기반으로 $\chi^2$ 통계량을 사용하여 일정한 유의수준을 만족하는 용어 집합으로 구성하였다. 그리고 도메인 대표 용어간의 방향성 관계는 양의 상관관계를 가지는 용어를 대상으로 포함(subsumption) 관계를 분석하여 정의하였다. 용어간의 상관관계는 연관 규칙 알고리즘의 향상도를 사용하여 용어간의 연관성 정도를 분석하였다. 본 연구에서 용어간의 관계는 포함 관계를 이용하여 상, 하위 관계만을 분석하였다. 두 용어 $t_k$, $t_s$ 간의 포함 관계는 $t_k$ 용어가 출현하는 문서중에서 $t_s$ 용어가 항상 출현하고, $t_s$ 용어가 출현하는 문서중에서 $t_k$ 용어가 출현하지 않는 경우로 정의된다. 따라서 용어간의 계층구조에서 $t_k$가 $t_s$를 포함하면 $t_k$는 $t_s$보다 일반화된 상위 용어이고, $t_s$는 $t_k$보다 상세화된 하위 용어임을 의미한다[16]. 다음 식 (2)는 $t_k$, $t_s$ 용어간의 포함 관계에 대한 정의식이다.

$$Subsumption(t_k,t_s)=\exists t_k\exists t_s[P(t_k|t_s)=1 \text{ and } P(t_s|t_k)<1] \quad (2)$$

### 3.3 도메인 온톨로지의 개념별 연관 용어 집합 구성

연관 용어 집합은 도메인 온톨로지상의 개념과 연관성이 높은 용어로 구성된 집합이다. 이러한 연관 용어 집합은 개별적인 용어만으로 문서의 내용을 파악하기 어려운 경우에 문서에 대한 정확한 개념을 분석하는데 유용하다. 예를 들어, 문서에서 출현한 '공', '방망이', '장갑'과 같은 개별 용어는 이 문서의 내용이 '야구'를 의미하는지 알기 어렵다. 하지만, '야구'와 관련된 연관 용어 집합을 함께 분석할 경우에는 이 문서의 내용이 '야구'를 설명하는 문서임을 알 수 있다. 문서 검색 과정에서 연관 용어의 중요성은 Google 온톨로지 키워드 분석 과정에서도 알 수 있다. Google에서 키워드는 평균 4.8개의 연관 용어를 가지고 있으며, 검색 결과에서 상위 레벨의 문서일수록 연관 용어를 많이 포함하고 있다. Google에서 상위 검색 결과에 포함된 연관 용어는 평균 2.17개의 연관 용어를 가진다[17].

연관 용어 집합은 도메인 온톨로지를 구성하는 개념별로 연관 규칙 탐사 알고리즘을 사용하여 신뢰도, 지지도와 향상도를 기준으로 연관성이 높은 용어 집합으로 구성하였다. 이 알고리즘은 하나의 장바구니에 담긴 상품 집합을 트랜잭션으로 정의하여 항목간의 연관 규칙을 발견하기 위한 대표적인 데이터마이닝 알고리즘이다. 본 연구에서 연관 규칙을 발견하기 위한 트랜잭션 단위는 하나의 문서에서 추출된 용어 집합이다. 그리고 용어 집합은 전처리 과정에서 형태소 분석을 통하여 추출된 명사 형태의 전문용어를 추출하여 구성하였다.

### 3.4 지식 문서에서 대표 개념 추출

#### 3.4.1 지식 문서 분류

지식 문서에서 대표 개념을 추출하기 위한 과정은 임의의 XML 문서에 대한 도메인을 결정하는 문서 분류 과정과 도메인 온톨로지의 개념과 문서를 매핑하는 과정으로 이루어진다. 먼저, 문서가 속하는 도메인을 결정하기 위하여 도메인 대표 용어 집합과 분류용 문서에서 추출한 용어 집합간의 유사도를 계산하기 위한 *DScore* 기법을 제안하였다.

*DScore* 기법에서는 유사도 계산에서 XML 문서의 구조적인 특징을 반영하기 위하여 엘리먼트 가중치와 빈도수, 그리고 도메인 가중치를 고려하였다. 여기서 도메인 가중치는 도메인 대표 용어에 대하여 전체 빈도수에 대한 도메인별 비율에 따라 결정된다. XML 문서는 가장 높은 *DScore* 값을 가지는 도메인으로 분류된다.

임의의 도메인 $D_i$ 를 대표하는 용어 집합 $DT$를 $DT_i=\{dt_1,dt_2,\cdots,dt_O\}$라 정의하면 *DScore*에 대한 정의식은 다음 식 (3)과 같다.

$$DScore=\sum_{l=1}^{O}(dtw_l\times I_j) \quad (3)$$

$$I_j=\sum_{j=1}^{N}(ew_j\times dtf_{jl}^{i})$$

$dtw_l$ : 특정 도메인에서 임의의 대표 용어 $dt_l$의 도메인 가중치
$ew_j$ : 엘리먼트 인스턴스$e_j$의 가중치
$dtf_{jl}^{i}$ : 문서$c_i$의 엘리먼트$e_j$에서 대표 용어$dt_l$의 빈도수

3.4.2 지식 문서의 개념 추출

문서를 대표하는 개념을 효과적으로 추출하는 작업은 지식 검색에서 핵심적인 역할을 수행한다. *TFIDF*와 같은 기존의 기법에서는 문서에서 출현한 용어의 단순 빈도수에 의해 문서에서 용어가 미치는 영향을 분석하였다. 본 연구에서는 도메인 온톨로지를 이용하여 지식 문서에서 대표 개념을 자동적으로 추출하기 위한 *TScore* 기법을 제시하였다. 또한 이 기법은 도메인 온톨로지의 개념별로 연결된 연관 문서를 통하여 특정 개념과 연관된 문서 집합을 효과적으로 검색하는데 사용할 수 있다.

*TScore* 기법은 지식 문서에서 대표 개념을 추출하기 위하여 도메인 온톨로지의 개념별 연관 용어 집합과 문서의 대표 용어간의 유사도를 비교하기 위한 기법이다. 문서의 대표 용어와 도메인 온톨로지의 개념별 연관 용어 집합간의 유사도 비교에서 가장 높은 *TScore* 값을 가지는 용어를 해당 문서에 대한 대표 개념으로 추출하였다. 본 연구에서 제안한 *DScore* 기법과 *TScore* 기법은 다르게 XML 문서의 구조적인 특징을 효과적으로 반영할 수 있다.

임의의 도메인 $D_i$를 대표하는 용어 $dt_l$의 연관 용어 집합 $AT$를 $AT_{dt_l}=\{(at_{k1},atw_{k1}),(at_{k2},atw_{k2}),\cdots,(at_{kP},atw_{kP})\}$라 정의하면 대표 용어 $dt_l$에 대한 $TScore$는 다음 식 (4)와 같이 정의된다.

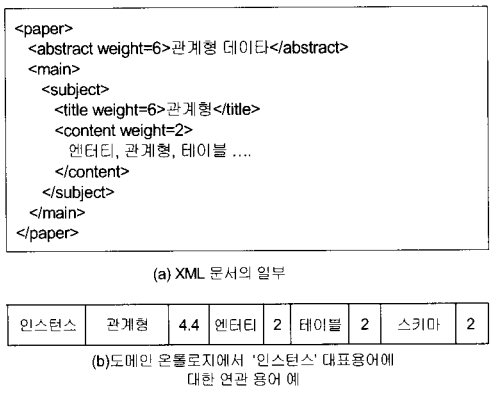
$$TScore(dt_l)=\frac{1}{N}\sum_{m=1}^{P}(atw_{km}\times I_j) \quad (4)$$

여기서, $I_j=\sum_{j=1}^{N}(ew_j\times atf^{i}_{jm})$

$atw_{km}$ : 임의의 대표 용어 $dt_l$에 대한 연관 용어 $at_m$의 가중치
$ew_j$ : 엘리먼트 인스턴스 $e_j$의 가중치
$atf^{i}_{jm}$ : 문서 $c_i$의 엘리먼트 $e_j$에서 연관 용어 $at_m$의 빈도수
$N$ : 임의의 대표 용어 $dt_l$에 대한 연관 용어 수

다음 (그림 2)는 *TScore* 기법을 이용한 유사도 계산 과정을 보여주기 위한 XML 문서의 일부와 데이터베이스 온톨로지에서 '인스턴스' 대표 개념에 대한 연관 용어 집합의 예이다. '인스턴스' 대표 개념에 대한 *TScore* 값은 다음과 같이 계산된다. 먼저, '인스턴스' 대표 개념의 각 연관 용어에 대한 가중치를 계산한다. 연관 용어에 대한 가중치 합은 연관 용어가 출현한 엘리먼트 가중치와 해당 엘리먼트에서 연관 용어의 빈도수를 곱한 값의 합으로 정의하였다. 그리고 각 대표 개념은 연관 용어의 빈도수가 다른 관계로 가중치를 연관 용어가 출현한 빈도수로 나누어 주는 정규화 과정을 거친다.

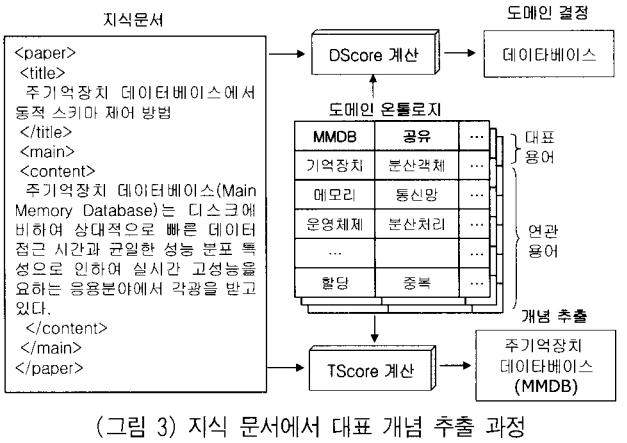
(그림 2)에서 데이터베이스 도메인 온톨로지의 '인스턴스' 대표 용어에 대한 계산 과정은 다음과 같다. '인스턴스'와 관련된 연관 용어중에서 XML 문서에서 출현한 '관계형' 연관 용어는 <abstract>, <title>, <content> 엘리먼트에서 한 번씩 출현하였다. 엘리먼트 가중치는 식(4)에서 $ew_j$에 의해 계산된다. 예제에서 <abstact>와 <title>에 배정된 엘리먼트 가중치는 6이고 <content> 엘리먼트의 가중치는 2이다. 그




(그림 2) 지식 문서에서 대표 개념 추출을 위한 예제 데이타

리고 '관계형' 연관 용어에 대한 $I_j$는 연관 용어가 출현한 엘리먼트의 가중치와 각 엘리먼트에서 연관 용어의 빈도수를 곱한 합으로 6*1+6*1+2*1로서 14가 된다. $I_j$ 값인 14와 '관계형' 연관 용어의 가중치 4.4를 곱하면 61.6이 된다. '인스턴스' 용어에 대한 *TScore* 값은 모든 연관 용어 집합에 대하여 위와 같이 계산하여 더한 값에서 '인스턴스' 용어에 대한 연관 용어 수 4로 나누는 정규화 과정을 거치면 17.4가 된다. 마지막으로 문서를 대표하는 개념은 도메인 온톨로지의 각 개념별로 *TScore* 값을 구하여 가장 높은 *TScore* 값을 가지는 개념으로 추출한다.

다음 (그림 3)은 본 논문에서 제안한 기법에 의해 XML 형식의 지식문서로부터 개념을 추출하는 과정을 보여주는 그림이다. 먼저, 지식문서는 *DScore* 기법에 의해 지식 문서에서 출현한 용어 집합과 도메인 온톨로지의 연관 용어 집합간의 유사도를 비교하여 데이타베이스 도메인으로 분류된다. 그리고 이 문서에서 출현한 용어와 데이터베이스 도메인의 대표 용어별로 구성된 연관 용어 집합간의 *TScore* 값을 계산하여 문서의 개념을 추출한다.

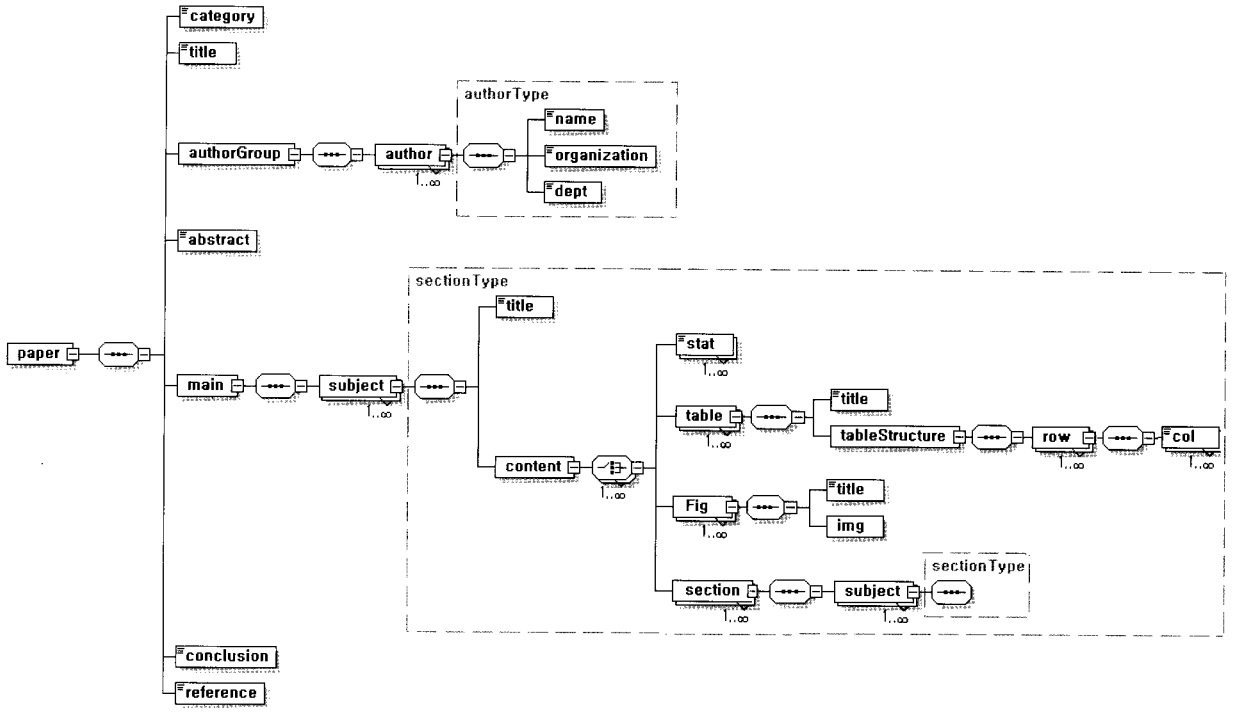



(그림 3) 지식 문서에서 대표 개념 추출 과정

## 4. 실험 및 결과

4.1 실험 데이터 및 문서 스키마

본 논문에서 제안한 기법의 효율성을 검증하기 위하여 컴




(그림 4) 실험용 XML 문서에 대한 스키마 구조

퓨터 관련 학회에서 발표된 295편의 학술 논문을 대상으로 실험하였다. 실험용 문서에 대한 구조는 XML 스키마로 정의하였다. 실험에서는 XML 문서에 대한 형태소 분석을 통해 추출된 용어에 대하여 컴퓨터 용어 사전에 수록된 전문용어를 추출하였다. 전체 논문에서 추출된 도메인 전문용어는 약 24,986개로 편당 평균 104개의 전문용어가 추출되었다. '시스템' 처럼 일반화된 전문용어를 제거하여 최종적으로 추출된 전문용어 수는 11,336개이며, 문서별로 약 86개의 전문용어가 추출되었다.

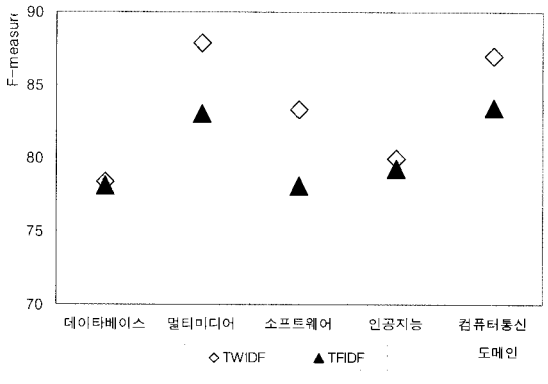
위의 (그림 4)는 실험용 XML 문서에 대한 스키마 구조이다.

### 4.2 실험 결과 및 분석

#### 4.2.1 지식 문서 분류 실험

본 연구에서 제안한 *TWIDF* 기법을 이용하여 XML 문서의 구조적인 정보가 문서 분류에 미치는 영향에 대해 실험하였다. 먼저, 문서 분류 성능은 *TWIDF* 기법과 기존의 *TFIDF* 기법을 사용하여 도메인 대표 용어를 구성하여 F-measure 척도에 의해 비교하였다. 실험용 문서는 *DScore* 값이 가장 높은 도메인으로 분류하였다. 그리고 XML 문서의 엘리먼트별로 가중치를 다양하게 부여하여 용어 가중치가 분류 결과에 미치는 영향력을 분석하였다. 엘리먼트 가중치는 XML 문서의 제목, 요약·결론, 본문 레벨1, 본문 레벨2, 본문 레벨3으로 구분하여 부여하였다. XML 문서에서 공통적으로 가지는 루트 엘리먼트에 대한 가중치는 고려하지 않았다.

다음 (그림 5)는 *TWIDF*와 *TFIDF* 기법에 의해 추출된 대표 용어 집합을 이용하여 F-measure 척도에 의해 분류 성능을 비교한 결과이다. (그림 5)에서처럼 대부분의 도메인

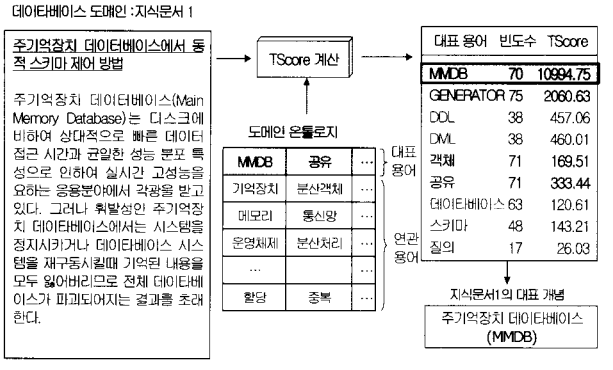



(그림 5) *TWIDF*와 *TFIDF* 기법간의 문서 분류 성능 비교

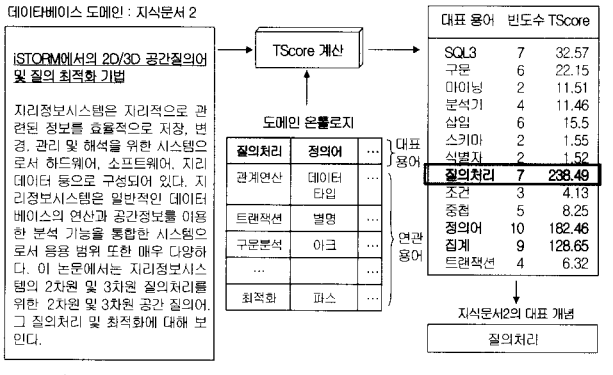
에서 기존의 *TFIDF* 기법보다 본 연구에서 제안한 *TWIDF* 기법의 분류 성능이 향상되었음을 알 수 있다.

#### 4.2.2 지식 문서 개념 추출 실험

지식 문서에서 개념을 추출하는 과정은 문서에서 출현한 용어 집합과 해당 도메인의 개념별 연관용어 집합간의 *TScore* 기법에 의해 유사도를 비교하는 과정이다. 다음 (그림 6)은 본 논문에서 제안한 *TScore* 기법에 의해 XML 문서로부터 대표 개념을 추출한 실험 결과이다. 먼저, 이 문서는 *DScore* 기법에 의해 데이타베이스 도메인으로 분류되었다. 그림 우측의 *TScore*는 이 문서에서 출현한 용어와 데이터베이스 도메인의 대표 용어별로 구성된 연관 용어 집합간의 *TScore* 값을 계산한 결과이다. 실험 결과에서 문서에서 출현한 '주기억장치 데이타베이스' 용어의 빈도수는 70으로




(그림 6) *TScore* 기법에 의한 XML 문서의 대표 개념 추출 결과(1)




(그림 7) *TScore* 기법에 의한 XML 문서의 대표 개념 추출 결과(2)

'GENERATOR', '객체', '공유'보다 적다. 하지만 '주기억장치 데이터베이스'의 *TScore* 값이 가장 높은 관계로 이 문서에 대한 대표 개념으로 추출하였다. 실제로 그림 왼편의 문서에서 설명하는 주된 개념이 '주기억장치 데이타베이스' 와 관련된 내용이라는 것을 알 수 있다.

위의 (그림 7)은 데이타베이스 도메인으로 분류된 임의의 XML 문서에서 대표 개념을 추출한 다른 실험 결과이다. '질의처리' 용어는 출현 빈도수가 3번째이지만, *TScore* 값이 가장 높은 관계로 이 문서에 대한 대표 개념으로 추출되었다. 실제로 문서에서 설명하는 주된 내용이 '질의처리'와 관련된 개념이라는 것을 알 수 있다.

문서에서 출현한 용어의 단순 빈도수를 중심으로 개념을 검색하는 기존 기법에서는 문서를 대표하는 개념을 효과적으로 추출하기 어렵다. 하지만 본 실험 결과에서처럼 온톨로지를 이용한 내용 기반의 지식 탐사 기법은 출현 빈도수는 낮더라도 용어의 중요도를 반영하여 대표 개념으로 추출할 수 있다. 결과적으로 본 기법은 사용자가 원하는 지식을 효과적으로 추출할 수 있는 새로운 형태의 지식 관리 모델이라 생각한다.

## 5. 결 론

본 논문에서는 도메인 온톨로지를 이용하여 대량의 XML 문서를 체계적으로 분류하고, 지식 문서로부터 대표 개념을 자동적으로 추출하기 위한 새로운 기법을 제시하였다. 먼저, XML 문서의 구조적인 특징을 가중치로 반영한 엘리먼트 가중치와 *TWIDF* 기법을 이용하여 문서를 대표하는 용어 집합을 구성하였다. 그리고 도메인별 대표 용어 집합은 연관 규칙 탐사알고리즘을 사용하여 향상도순으로 Top-N개의 연관 용어 집합으로 구성하였다. 도메인 온톨로지는 용어간의 포함 관계를 분석하여 도메인 대표 용어간의 상, 하위 관계를 방향성 그래프에 의해 표현하였다.

그리고 본 논문에서는 XML 문서의 구조적인 특징을 이용하여 문서를 분류하기 위한 *DScore* 기법과 지식 문서에서 대표 개념을 자동적으로 추출하기 위한 *TScore* 기법을 제시하였다. 먼저, *DScore* 기법은 문서가 속하는 도메인을 결정하기 위하여 도메인 대표 용어 집합과 분류용 문서에서 추출한 용어 집합간의 유사도를 계산하기 위한 기법이다. *DScore* 기법에서는 엘리먼트별 가중치와 도메인 가중치를 이용하여 XML 문서의 구조적인 특징을 반영하여 문서를 분류하였다.

*TScore* 기법은 지식 문서에서 대표 개념을 추출하기 위하여 도메인 온톨로지의 개념별 연관 용어 집합과 문서의 대표 용어간의 유사도를 비교하기 위한 기법이다. 문서에서 추출한 대표 용어중에서 가장 높은 *TScore* 값을 가지는 용어를 해당 문서에 대한 대표 개념으로 추출하였다. 이 기법은 도메인 온톨로지의 개념별로 연결된 연관 문서를 통하여 특정 개념과 연관된 문서 집합을 효과적으로 검색하는데 사용할 수 있다.

본 논문에서 제안한 기법의 효율성을 검증하기 위하여 295편의 학술 논문을 대상으로 문서 분류 실험과 개념 추출 실험을 하였다. 먼저, 문서 분류 실험에서는 XML 문서의 엘리먼트별로 가중치를 다양하게 부여하여 용어 가중치가 분류 결과에 미치는 영향력을 분석하였다. *TWIDF* 기법과 기존의 *TFIDF* 기법에 의해 구성된 도메인 대표 용어에 대하여 F-measure 척도에 의한 문서 분류 실험을 통하여 제안된 기법의 분류 성능이 우수함을 보였다. 그리고 본 연구에서 제안한 *TScore* 기법에 의해 문서에서 출현한 용어 집합과 해당 도메인의 개념별 연관 용어 집합간의 유사도 비교를 통하여 지식 문서에서 개념을 추출하기 위한 실험을 하였다. 실험 결과에서 출현 빈도수는 낮더라도 *TScore* 값이 높은 용어가 문서를 대표하는 개념으로 효과적으로 추출될 수 있음을 보였다.

본 논문에서 제안한 기법은 지식 관리 모델에서 개념 기반의 검색 기법을 통해 사용자가 원하는 지식을 효과적으로 검색하는데 이용될 수 있다. 또한 본 기법은 시맨틱 웹, e-Learning, KDD, 전자상거래 등과 같은 다양한 분야에서 응용될 수 있다. 앞으로 사용자 프로파일과 연계된 도메인을 분석하여 개인별로 관심 있는 지식 문서를 추천하기 위한 기법에 대한 연구가 필요하다. 그리고 사용자끼리 지식 공유를 위한 사용자 협업 모델에 대한 연구를 진행할 계획이다.

## 참 고 문 헌

[1] E. Hyvönen, S. Saarela, K. Viljanen, "Ontogator: combining view- and ontology-based search with semantic

browsing," *Proc. of the XML Finland Conference*, 2003.
[2] P. V. Benjamins, D. Fensel, and A. G. Perez, "Knowledge Management through Ontologies," *Proc. of the Practical Aspects of Knowledge Management*, 1998.
[3] Y. Sure and et al., *On-To-Knowledge: Semantic Web Enabled Knowledge Management*, J. Wiley and Sons, 2002.
[4] R. Benjamins and D. Fensel, "The Ontological Engineering Initiative(KA)$^{2}$," *Proc. of Formal Ontologies in Information Systems*, pp.287-301, 1998.
[5] S. Decker, M. Erdmann, D. Fensel, and R. Studer, *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*, In R. Meersman et al., editors, Database Semantics: Semantic Issues in Multimedia Systems, Kluwer Academic Publisher, pp.351-369, 1999.
[6] D. L. McGuinness, "Ontological Issues for Knowledge-Enhanced Search," *Proc. of the Formal Ontology in Information Systems*, pp.302-316, 1998.
[7] 최옥경, 한상용, "자동화된 통합 프레임워크를 위한 시맨틱 웹 기반의 정보 검색 시스템," *한국정보처리학회 논문지*, Vol.13, No.1, pp.129-136, 2006.
[8] E. Hyvönen and et al "Finish Museum on the Semantic Web User's Perspective," *Proc. of the Museums and the Web*, 2004.
[9] A. Maedche, "A Machine Learning Perspective for the Semantic Web," *Proc. of the Semantic Web Working Symposium*, 2001.
[10] 김명숙, 공용해, "온톨로지-DTD 정합에 의한 XML 질의 확장," *한국정보처리학회 논문지*, Vol.12, No.5 pp.773-780, 2005.
[11] 이무훈, 조현규, 조현성, 조성훈, 장창복, 최의인, "웹 문서의 의미적 연관성 기술을 위한 온톨로지 에디터," *한국정보처리학회 논문지*, Vol.12, No.5, pp.881-888, 2005.
[12] 오삼균, "Web Ontology Language와 그 활용에 관한 고찰," *데이터베이스연구학회지*, Vol.18, No.3, pp.63-79, 2002.
[13] G. A. Miller, "WordNet : A Lexical Database for English," *Communication of the ACM*, Vol.38, No.11, pp.39-41, 1995.
[14] K. Knight and S. Luk, "Building a Large-Scale Knowledge Base for Machine Translation," *Proc. of the AAAI*, 1994.
[15] Cycorp, "Cyc Knowledge Server," http://www.cyc.com, 2002.
[16] H. J. Mun, J. Y. Lee and Y. T. Woo, "A Domain Ontology Creation Method for Ontology-based Knowledge Management Model," *Int'l Journal of ACIS*, pp.99-108, 2005.
[17] goRank,.com, "Google Ontology Analysis," http://www.gorank.com/research/google_ontology_analysis.php, 2004.

**문 현 정**

e-mail : mun@changwon.ac.kr
1994년 한국방송대학교 전자계산학과 (이학사)
1996년 창원대학교 전자계산학과(이학석사)
2003년 창원대학교 컴퓨터공학과(공학박사)
2004년~현재 창원대학교 연구교수
관심분야 : 온톨로지 기반 지식관리, 시맨틱 웹, 텍스트마이닝

**우 용 태**

e-mail : ytwoo@changwon.ac.kr
1982년 경북대학교 전자공학과(공학사)
1984년 경북대학교 전자공학과(공학석사)
1995년 경북대학교 전자공학과(공학박사)
1987년~현재 창원대학교 컴퓨터공학과 교수
관심분야 : 온톨로지 기반 지식관리, 추천시스템, 시맨틱 웹, 텍스트마이닝, e-Learning