Journal of The Korea Society of Computer and Information
Vol. 19, No. 10, October 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.10.023

비정형 문서의 정보추출을 통한 OWL 온톨로지 구축 시스템의 설계 및 구현

조 대 웅*, 최 지 웅**, 김 명 호**

The Design and Implementation of OWL Ontology Construction System through Information Extraction of Unstructured Documents

Dae Woong Jo *, Ji Woong Choi **, Myung Ho Kim**

요 약

정보검색 분야의 발전은 많은 양의 정보를 빠르게 찾아주는 것에서 사람이 원하는 정보를 정확하게 찾아주는 연구 분야로 넓혀가고 있다. 핵심 기술로는 개인화 및 시맨틱 웹 기술을 활용하고 있다. 웹 문서에 대한 자동색인 기술과 처리능력은 연구단계를 넘어 실용 서비스로 나타나고 있다. 하지만 웹 문서 이외의 첨부된 문서 형태에 대한 문서정보검색에 관한 연구는 미진한 상황이다. 본 논문에서는 텍스트, 워드, 한글과 같은 형식으로 작성된 비정형 문서의 본문 내용을 분석하여 OWL 온톨로지로 구축하는 방법에 대해 설명한다. 문서 온톨로지의 TBox를 구축하고, 문서로부터 얻을 수 있는 자원을 선정하여, 구축된 문서 온톨로지의 인스턴스로 활용할 수 있도록 시스템으로 구현한다. 이와 같은 비정형 문서의 온톨로지 자동 구축으로 해당 문서의 시맨틱 기술을 이용한 정보검색 및 문서관리 시스템에서 효과적으로 활용 가능하다.

▸ Keywords : OWL, 문서정보검색, 온톨로지, 시맨틱 웹, 링크드 데이터

Abstract

The development of the information retrieval field is evolving to the research field searching accurately for the information from thing finding rapidly a large amount of information. Personalization and the semantic web technology is a key technology. The automatic indexing technology about the web document and throughput go beyond the research stage and show up as the practical service. However, there is a lack of research on the document information retrieval field about the attached document type of except the web document. In this paper, we illustrate about the method in which it analyzed the text content of

•제1저자 : 조대웅 •교신저자 : 김명호
•투고일 : 2014. 8. 5, 심사일 : 2014. 8. 18, 게재확정일 : 2014. 9. 3.
* 숭실대학교 대학원 컴퓨터학과 (Deptartment of Computer Science and Engineering, Soongsil University)
** 숭실대학교 컴퓨터학부 (School of Computer Science and Engineering, Soongsil University)

the unstructured documents prepared in the text, word, hwp form and it how to construction OWL ontology. To build TBox of the document ontology and the resources which can be obtained from the document is selected, and we implement with the system in order to utilize as the instant of the constructed document ontology. It is effectually usable in the information retrieval and document management system using the semantic technology of the correspondence document as the ontology automatic construction of this kind of the unstructured documents.

▸ Keywords : OWL, Document Information Retrieval, Ontology, Semantic Web, Linked Data

I. 서 론

정보검색 분야의 발전은 많은 양의 정보를 빠르게 찾아주는 것에서 사람이 원하는 정보를 정확하게 찾아주는 연구 분야로 넓혀가고 있다. 정확한 정보를 찾는 것은 이용자에 대한 축적된 데이터를 활용한 개인화, 추천 서비스 등을 이용하거나 시맨틱 웹 기술과 같은 기계의 정보처리능력의 향상을 이용한 정보 추론, 연관관계 등의 서비스가 가능한 기술의 활용으로 정보검색의 결과의 정확성을 높일 수 있다.

현재 구글과 같은 웹 검색 서비스에서는 웹 자원에 대한 자동색인 및 처리능력의 향상으로 효과적인 정보검색 서비스를 지원하고 있다. 페이지랭크 알고리즘[1]의 적용으로 해당 키워드가 들어간 웹페이지의 참조 수에 순위를 매겨서 가장 높은 순위부터 이용자에게 보여주는 식이다. 이러한 웹 페이지 외에 첨부된 문서 형태에 대한 문서정보검색에 관한 방법도 키워드가 들어간 문서의 제목이나 본문에 해당 키워드가 들어가 있으면 문서를 검색 결과에 보여주고 있다. 이러한 방식은 이용자들이 가장 많이 찾는 정보를 보여준다는 이점이 있지만 이용자에 따라서는 원하는 정보를 찾기 위해 더 많은 시간을 할애해야 할지도 모른다.

시맨틱 웹 기술을 활용한 검색은 웹페이지 검색에 한정되어 부분적으로 서비스가 되고 있다. 웹페이지 안의 하이퍼텍스트 처리 외에 페이지 안에 첨부된 문서파일에 대한 시맨틱 웹 기술을 이용한 연구는 현재 미진하다. 물론, 일반 웹 페이지로부터 상당량의 정보를 얻고는 있지만 이미 구축되어 있는 논문, 기술백서, 사례연구, 보고서 등의 문서들에서는 웹 페이지에서는 볼 수 없는 고품질의 정보를 가지고 있는 경우가 많다. 따라서 기 구축된 문서를 활용하고, 정보검색의 다양성과 정확성을 높이기 위해서 문서정보검색에 시맨틱 웹 기술을 적용해서 처리하여야 한다.

본 논문에서는 텍스트, 워드, 한글과 같은 문서의 형식으로 작성된 파일의 본문 내용을 분석하여 OWL 온톨로지로 구축하는 방법에 대해 설명한다. 문서 온톨로지의 TBox를 구축하며, 본 논문에서는 D-Ontology(Document Ontology)로 명명한다. D-Ontology는 문서의 기본 정보들인 저자 정보, 문서 종류, 생성 날짜, 크기 등과 같은 문서의 메타데이터에 대한 속성 정보와 문서의 중요 키워드 및 계층 구조 간의 관계를 온톨로지로 기술하며, 추가로 D-Ontology가 해당 도메인 검색에 그치는 것이 아니라 이미 구축된 다른 온톨로지와의 연계 및 연결된 웹 데이터를 위해 링크드 데이터(Linked Data)[2] 형태로 확장, 구축될 수 있는 방법을 제안한다.

본 논문에서는 TBox로 구축된 D-Ontology에 인스턴스 매칭을 위한 시스템을 구현한다. 시스템 요소로는 비정형 문서를 구조화하기 위한 항목 정보 추출 처리기, 문서의 메타데이터 정보 추출을 위한 파일처리기, 문서 내의 중요 키워드 추출을 위한 키워드 추출기까지 문서 분석 엔진에서 처리하며 추출된 각각의 정보들은 온톨로지 구축 엔진에 의해 D-Ontology의 인스턴스로 매칭 한다. 인스턴스 매칭이 완료된 D-Ontology는 확장 검색을 위해 관련된 다른 웹 문서 페이지와 연결한다. 본 논문에서는 DBpedia[3]의 SPARQL Endpoint[4]에 비정형 문서로부터 추출한 키워드와 관련된 페이지의 링크 정보를 SPARQL 질의를 통해 얻어오고, 얻어온 링크 정보와 키워드와의 연관을 맺어서 링크드 데이터 형태로 확장 검색할 수 있는 발판을 마련한다. 이와 같은 비정형 문서를 관리할 수 있는 D-Ontology의 구축은 해당 문서의 시맨틱 기술을 이용한 정보검색 및 문서관리 시스템에서 효과적으로 활용 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서

비정형 문서를 구조화시키기 위한 연구사례와 온톨로지 구축 방법 및 도서 온톨로지 구축 현황에 대해 설명을 한다. 3장 본론에서는 본 논문에서 제안하는 비정형 문서의 온톨로지 구축 방법에 대한 설명과 정보 추출 범위 및 알고리즘 설명을 하고 추출된 데이터를 온톨로지로 자동 구축하기 위한 계획과 방법에 대해 설명한다. 4장에서는 온톨로지 구축 시스템의 설계 및 구현, 구축된 온톨로지를 이용한 활용 및 질의 문 예시에 대해 설명한다. 마지막으로 5장에서 결론을 내린다.

II. 관련 연구

1. 비정형 문서 구조화

비정형 문서를 구조화시키기 위한 연구는 엔지니어링 문서의 자동 추출 연구[5]가 존재한다. [5]에서는 엔지니어링 문서에서 각 제목의 머리 기호가 그 문서의 논리적 계층 구조를 표현한다는 점을 이용하여 문서 내 각 제목의 계층을 자동으로 분류하는 방법론을 제시하였다. 제시한 방법론은 비정형 텍스트 문서에서 세부 제목을 추출하는 방법과 추출된 제목의 계층을 정의하는 방법으로 구성된다. 문서의 세부 제목은 문장의 맨 앞에 위치한 머리기호의 형태를 미리 정의된 머리 기호 그룹과 비교하여 추출하며, 추출된 제목의 계층은 머리 기호 형태의 변화에 따라 각 제목 간의 상대적 위치를 파악함으로써 정한다. 제시된 방법론을 이용하여 비정형 텍스트 문서를 세부 제목에 따라 구조화된 XML 문서로 변환하는 시범 모듈을 개발하였다. 하지만 머리 기호의 상대적인 위치를 기반으로 항목의 레벨을 정하고 있으며, 이와 같은 방식은 레벨이 다르지만 같은 위치로 작성이 된 문서는 레벨의 깊이를 명확하게 판별할 수 없다는 문제도 존재한다.

본 논문에서의 항목 추출방식은 정규 표현식 패턴에 의해 항목들을 선별하고, 선별된 항목별로 레벨 값과 깊이 우선 방식의 트리 기반의 자료구조를 이용해서 항목 간의 관계 깊이를 결정할 수 있다는 장점을 가지고 있다.

2. 온톨로지 구축

온톨로지 구축은 특정 도메인을 선정하고, 개념을 자동 혹은 반자동으로 추출하여 추출된 개념들의 관계를 설정하는 것을 말한다. 이러한 온톨로지 구축 과정에서 어려운 점은 온톨로지 구축의 이론적인 체계와 원리가 아직 미흡하다는 것이다. 하지만 온톨로지 구축에 필요한 기본적인 설계 규칙이나 패턴정의에 대한 방법론이 있으며, 본 논문에서는 Ontology Development 101 방법론[6]을 통해 온톨로지를 구축한다. Ontology Development 101에서는 7단계로 온톨로지 구축 방법에 대해 설명하고 있다. (1) 온톨로지 도메인과 범위를 결정 (2) 기존 온톨로지의 재사용 고려 (3) 온톨로지에 있어서 중요한 용어들을 열거 (4) 클래스 간의 계층을 정의 (5) 클래스의 속성들을 정의 (6) 속성을 정제 (7) 인스턴스를 생성한다로 되어 있으며, 본 논문에서는 이와 같은 7단계의 규칙을 준수하여 D-Ontology를 구축한다.

3. 문서 온톨로지 구축 현황

문서 처리와 관련된 온톨로지는 도서관 분야의 링크드 데이터(LLD, Library Linked Data)[7] 프로젝트가 있다. 서지 데이터인 서명, 저자, 날짜와 전거 데이터인 분류, 주제명, 저자명, 시소러스 등의 데이터를 가지고 링크드 데이터로 구축하는 것을 뜻한다.

링크드 데이터는 데이터에 링크정보를 이용해서 각 데이터들 간의 연결을 통한 통합적 정보 처리 능력을 활용하는 것이다. 링크드 데이터 발행 규칙은 다음과 같다[2]. (1) URI로 해당 리소스를 판별하고, (2) HTTP URI를 통해 사람들이 해당 리소스를 볼 수 있어야 한다. (3) RDF 형태의 트리플로 기술되어야 하며, SPARQL과 같은 표준 질의를 이용해서 유용한 정보를 제공할 수 있어야 한다. (4) 또 다른 URI를 포함하고 있어서 더 많은 것 즉, 개념들을 탐색할 수 있어야 한다. 본 논문에서도 링크드 데이터 발행 규칙에 따라 문서 온톨로지를 구축한다.

LLD와 관련한 도서관 분야의 온톨로지 구축 및 변환 프로젝트는 도서관에서 기존에 저장, 관리하던 정형적 데이터인 DB를 중심으로 트리플 형태로 변환 구축하는 작업이다. 이미 영국, 스웨덴, 독일, 미국 등 많은 곳에서 LLD 데이터를 구축하고 있다. 이러한 링크드 데이터 구축은 지식베이스의 설명을 위해 통일된 용어 어휘집이 필요한데 기존의 RDFS/OWL 기반 외에 SKOS(Simple Knowledge Organization System)[8], DC(Dublin Core)[9], FOAF(Friend of a Friend)[10] 같은 수많은 용어 어휘집을 링크드 데이터 구축에 사용하고 있으며, 이러한 용어 어휘집은 사실상 표준으로 자리 잡아가고 있다. 본 논문에서도 SKOS, DC, FOAF 어휘집을 활용하여 온톨로지를 구축한다.

III. 본 론

1. 비정형 문서의 정보 추출 및 자료 구조

본 장에서는 특정한 형식이 없는 비정형 문서를 구조화된 정보로 변환하기 위한 방법과 중요 키워드 선별 방법에 대해 설명한다.

비정형 문서 내에서 문서를 구조화하기 위한 방법은 문서가 가지는 항목들을 추출해서 컴퓨터가 이해할 수 있는 계층화된 정보로 변환하는 것이다. 또한, 계층적 항목 정보와 함께 항목아래 존재하는 본문 내용들도 계층적인 정보로 저장, 관리가 가능하다면 문서 전체에 대한 관계를 구조화 시킬 수 있다.

하지만 문서마다 성격과 목적이 다르고, 통일된 규칙이 존재하지 않아 항목 유형과 계층관계가 모두 다르다. 따라서 본 논문에서는 다양한 형식의 기본 항목 유형을 선별하고, 그에 따라 추출 가능한 범위를 산정한다. 또한, 항목 추출 알고리즘에 의해 추출된 항목은 트리형식의 자료구조로 자동 변환, 저장되어 구조화된 문서로 변환 된다.

비정형 문서로부터 중요 키워드를 선별하기 위한 방법은 선별된 항목 정보에 중요 키워드가 있다는 전제하에 항목 정보로부터 키워드를 선별, 저장하도록 한다. 선별된 키워드는 레벨 값을 부여하여 추후 검색 시 레벨 값에 따라 키워드의 중요도를 선택하여 나타낼 수 있다.

1.1 항목 추출 대상 유형 정의

비정형 문서의 항목 추출 유형 대상을 정의하기 위해선 기본적으로 많이 쓰이는 유형들을 조사할 필요가 있다.

모든 항목들에는 대체로 그 항차(項次)와 항순(項順)을 나타내기 위한 항번(項番)을 붙이는데, 그 형식에는 장절식(章節式), 숫자-문자식(number-letter), 십진법식(decimal system) 등이 있다[11]. 항목 추출 대상 유형을 정의한다는 것은 항번의 유형을 정의하는 것과 같다. 장절식은 법령문서에서 쓰는 형태로 본 논문에서는 숫자-문자식, 십진법식 형태의 유형들을 선정하고, 그에 따른 형식 정의를 통해 추출 가능한 범위를 계산한다.

표 1은 항목 추출을 위한 대상들을 정리해서 나타낸 것이다. 가장 많이 쓰는 유형은 유형1, 2, 3, 4와 같은 형태이며, 유형1혹은 유형 4가 보통 가장 최상위 체계로 자리하고 있다. 하지만 이와 같은 상-하위 체계는 문서마다 상이 할 수 있으므로 큰 의미는 없다. 본 논문에서는 각 항목 유형별 추출을 상-하위 체계 상관없이 구분 짓기 위해 트리구조의 깊이우선

표 1. 항목 추출 대상 유형
Table 1. Items Extraction Target Type

유형 1	1., 2., 3., ...(하위형식 1.1, 1.1.1, 1.1.1.1,..)
유형 2	1), 2), 3), ...
유형 3	(1), (2), (3), ...
유형 4	I., II., III., ...
유형 5	A., B., C., ...
유형 6	A), B), C), ...
유형 7	(A), (B), (C), ...
유형 8	가, 나, 다, ...
유형 9	가)., 나)., 다),. ...
유형 10	(가), (나), (다), ...

(depth-first)방식으로 자료들을 저장하면서 해당 항목들의 우선순위를 정한다.

1.2 항목추출을 위한 알고리즘

그림 1의 알고리즘 동작 예를 보면 좌측 칸에는 입력되는 문서의 구조를 나타낸다. 본 알고리즘은 항목을 추출하고자 하는 문서를 라인 단위로 1회 스캔하며 처리를 완료한다. 그림 1의 가운데 칸의 항목 레벨 배열은 본 알고리즘이 처리를 종료하였을 때의 모습이며 최초에는 비어있다. 우측 칸의 항목 트리는 본 알고리즘이 처리를 종료하였을 때의 모습이며, 최초에는 root만을 가지고 있다. 참고로, root의 레벨은 0으로 한다. 알고리즘 동작 예는 다음과 같다.

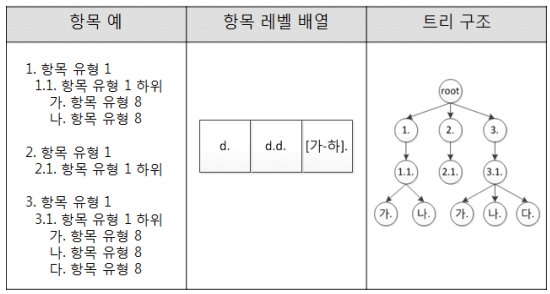


그림 1. 알고리즘 동작 예
Fig. 1. Example of Algorithm Behavior

'1.'이 읽혀졌을 때, '1.'의 정규 표현식 패턴인 'd.'가 항목 유형 우선순위 배열의 첫째 칸에 저장된다. 이것은 패턴 'd.'가 레벨 1임을 의미한다. 따라서 항목 트리의 레벨 1에 '1.'에 해당하는 노드를 추가한다.

'1.1.'이 읽혀졌을 때, '1.1.'의 정규 표현식 패턴인 'd.d.'가 항목 유형 우선순위 배열의 두 번째 칸에 저장된다. 이것은 패턴 'd.d.'가 레벨 2임을 의미한다. 따라서 항목 트리의

레벨 2에 '1.1.'에 해당하는 노드를 추가한다. 이 때, 가장 마지막으로 추가된 레벨 1 즉, '1.'의 자식으로 추가된다.

'가.'이 읽혀졌을 때, '가.'의 정규 표현식 패턴인 '(가-하).'가 항목 유형 우선순위 배열의 세 번째 칸에 저장된다. 이것은 패턴 '(가-하).'가 레벨 3임을 의미한다. 따라서 항목 트리의 레벨 3에 '가.'에 해당하는 노드를 추가한다. 이 때, 가장 마지막으로 추가된 레벨 2 즉, '1.1.'의 자식으로 추가된다.

'나.'이 읽혀졌을 때, '나.'의 정규 표현식 패턴인 '(가-하).'가 항목 유형 우선순위 배열의 세 번째 칸에 이미 저장되어 있다. 이것은 패턴 '나.'가 레벨 3임을 의미한다. 따라서 항목 트리의 레벨 3에 '나.'에 해당하는 노드를 추가한다. 이 때, 가장 마지막으로 추가된 레벨 2 즉, '1.1.'의 마지막 자식으로 추가된다. '2.'이 읽혀졌을 때, '2.'의 정규 표현식 패턴인 'd.'가 항목 유형 우선순위 배열의 첫 번째 칸에 이미 저장되어 있다. 이것은 패턴 '2.'가 레벨 1임을 의미한다. 따라서 항목 트리의 레벨 1에 '2.'에 해당하는 노드를 추가한다. 이 때, root의 마지막 자식으로 추가된다. 이와 같은 방식으로 항목들을 유형에 맞게 추출하고, 트리구조로 추출된 항목을 저장하게 된다.

1.3 키워드 추출방법

기술문서를 검색하는 것은 문서 안의 키워드의 정의를 찾아보는 경우가 많다. 문서 검색의 핵심은 중요 키워드를 선별하고 해당 키워드와 연관된 부가적인 정보들을 온톨로지로 보여주는 것이다. 본 논문에서는 해당 문서에 대한 중요 키워드는 문서의 항목에 나타난다는 전제하에 추출된 항목을 기반으로 항목 정보에서 키워드를 추출하는 방식을 택한다. 키워드 추출은 텍스트 마이닝의 한 분야로 자동 용어 인식, 자동 색인, 자동 키워드 추출이라고도 한다. 자동 키워드 추출을 위한 기법은 통계기반, 언어학적 기반, 기계학습 기반 등의 방법이 있다[12]. 본 논문에서는 키워드의 빈도수 계산을 통한 통계기반 접근법을 응용하여 키워드를 추출한다.

키워드 추출방법은 다음과 같다. (1) 키워드는 해당 문서의 항목으로부터 추출한다. (2) 형태소 분석을 통해 항목으로부터 명사를 찾아내고, 명사 중에서 고유명사, 보통명사, 외국어를 선별한다. (3) 고유명사와 외국어는 보통명사보다 높은 레벨에 위치한다. (4) 보통명사는 나오는 빈도수를 책정하여 중요도 레벨을 정한다. 텍스트 마이닝 분야에서 빈도수는 가장 많이 나오는 고빈도어와 가장 적게 나오는 저빈도어의 키워드는 중요 키워드가 될 확률이 낮다는 통계 결과[12]에 의해 고빈도어, 저빈도어의 키워드는 가장 낮은 레벨로 책정을 하며, 중간 빈도어의 키워드를 보통 명사 내에선 가장 높은 레벨로 정한다. (5) 책정된 레벨과 명사는 키, 벨류 형식의 〈레벨, 명사〉 형태로 키워드 저장소에 저장된다. (6) 키워드 저장소에 저장된 레벨 값을 통해 해당 명사의 중요도로 이용하여 온톨로지 구축 후 질의 시 레벨이 높은 키워드가 먼저 질의 결과에 나오도록 한다. 질의 시 레벨 값 조정을 통해 다른 키워드도 함께 도출은 가능하도록 한다.

키워드 추출과 관련된 예는 다음과 같다. "2. 온톨로지 구축 시스템의 설계", "2.1 시맨틱웹(SemanticWeb) 에서의 활용 구축"의 항목이 있다면 온톨로지, 시맨틱웹, SemanticWeb, 시스템은 고유명사 및 외국어로서 가장 높은 단계의 레벨을 부여받게 된다. 그 외의 보통명사들인 구축, 설계, 활용이 나올 수 있으며 빈도수 계산을 통해 각각의 레벨 값을 부여한다.

2. 온톨로지 구축

2.1 비정형 문서의 온톨로지 구축 목표

온톨로지는 일종의 지식베이스(Knowledge Base)로 TBox(terminological component), ABox(assertion component)를 합쳐놓은 개념을 뜻한다. 온톨로지의 구축은 지식베이스의 구축을 의미하고, 각 개념에 해당하는 TBox와 ABox의 구분을 명확하면서도 의미적으로 구분하는 것이 중요하다. TBox는 스키마, 공리(axiom)의 집합이고, ABox는 데이터와 공리의 집합으로 정의된다. TBox의 스키마 정의는 OWL DL 수준의 FOL(First-Order Logic)의 클래스, 오브젝트, 프로퍼티의 생성규칙, 추론 규칙, 분류 법칙 등을 이용하여 온톨로지를 구축하고, ABox는 인스턴스, 속성값, 일관성 체크 등을 확인하여 온톨로지로 구축한다.

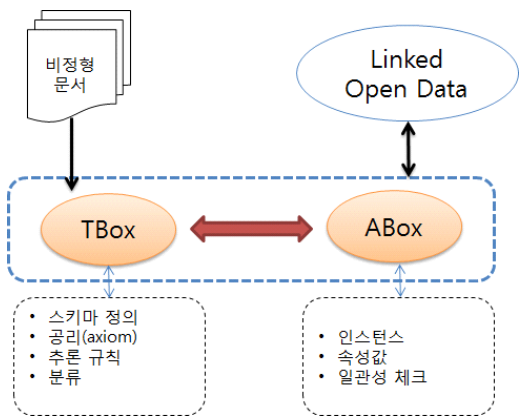


그림 2. 비정형 문서의 온톨로지 변환 형태
Fig. 2. Ontology Transformation Form of The Unstructured Documents

본 논문은 기술문서와 관련된 정보를 효과적으로 관리 및 문서 정보 검색을 하기 위한 D-Ontology를 구축하며, 이는 TBox와 ABox를 모두 구축한 문서 지식베이스로서의 역할을 한다. 그림 2는 본 논문에서 구축하는 D-Ontology의 변환 형태 및 응용에 대한 부분을 설명하고 있다. TBox에 해당하는 스키마는 문서로부터 얻을 수 있는 정보를 바탕으로 구축 툴을 이용하여 구축하고, ABox에 해당하는 인스턴스는 비정형 문서로부터 자동 추출하여 TBox와 매핑할 수 있도록 한다.

본 논문에서 구축하는 D-Ontology는 단순 문서들의 온톨로지를 만드는 것이 아니라, LOD(Linked Open Data)[13] 형태의 공유된 지식베이스를 구축하는 것을 목표로 한다. 비정형 문서로부터 얻어낸 키워드 및 문서의 종류, 타이틀, 저자의 이름과 소속 등의 정보들은 하나의 문서에서 얻을 수 있는 정보지만 LOD 형태로 구축하면 해당 정보의 확장성 있는 의미 검색을 가능하게 한다. LOD는 전 세계적으로 데이터를 오픈 연결하는 프로젝트로 시맨틱 웹의 궁극적인 목적이 될 수 있다. 키워드는 문서로부터 얻지만 키워드와 관련된 정의는 문서 하나에서 정의한 것 외에 웹에서는 다른 정의가 있을 수 있고, 그와 관련된 다른 생각을 가진 사람들도 존재한다. 이러한 부분을 DBpedia로 연결 확장 시키도록 하는 것이 D-Ontology 구축의 실제적인 목표가 된다.

2.2 온톨로지 구축 설계

D-Ontology의 구축은 그림 3에서와 같이 16개의 클래스를 구축하고, 그에 따른 각각의 18개의 프로퍼티 관계와 공리, 인스턴스들까지 구축된다. 프로퍼티로 사용되는 어휘는 D-Ontology에서 정의한 오브젝트 프로퍼티 관계, OWL DL 수준의 프로퍼티들, 키워드 개념에 대한 설명을 위한 SKOS, 문서의 메타정보 기술을 위한 DC, 저자 정보 관계 기술을 위한 FOAF 어휘를 활용하여 구축한다.

표 2는 각각의 오브젝트 프로퍼티 관계들을 기술한 것이고, 그림 3을 참조하면 Metadata 클래스는 dc:title, dc:format, dc:language, dc:size 데이터 프로퍼티 관계로 DC 어휘를 사용해서 문서의 메타데이터 속성을 나타내고 있다. Person은 foaf:name, foaf:mbox 관계로 데이터 프로퍼티 관계를 기술하고, DocType 클래스는 dc:type 관계로 해당 문서의 타입 정보를 기술한다.

Contents 클래스는 Keyword, Section 클래스를 isA 관계의 하위 클래스로 두고 있으며, Section 클래스는 Head, Body, Level 클래스를 하위 클래스로 두고 항목 정보의 항번과 제목, 레벨 값을 추출하여 인스턴스로 매핑 한다. Keyword 클래스는 항목의 제목으로부터 형태소 분석을 통해 얻은 명사의 종류와 빈도수 계산을 통해 해당 문서의 키워드를 선별하고, 키워드의 가중치를 위한 Value 클래스를 하위

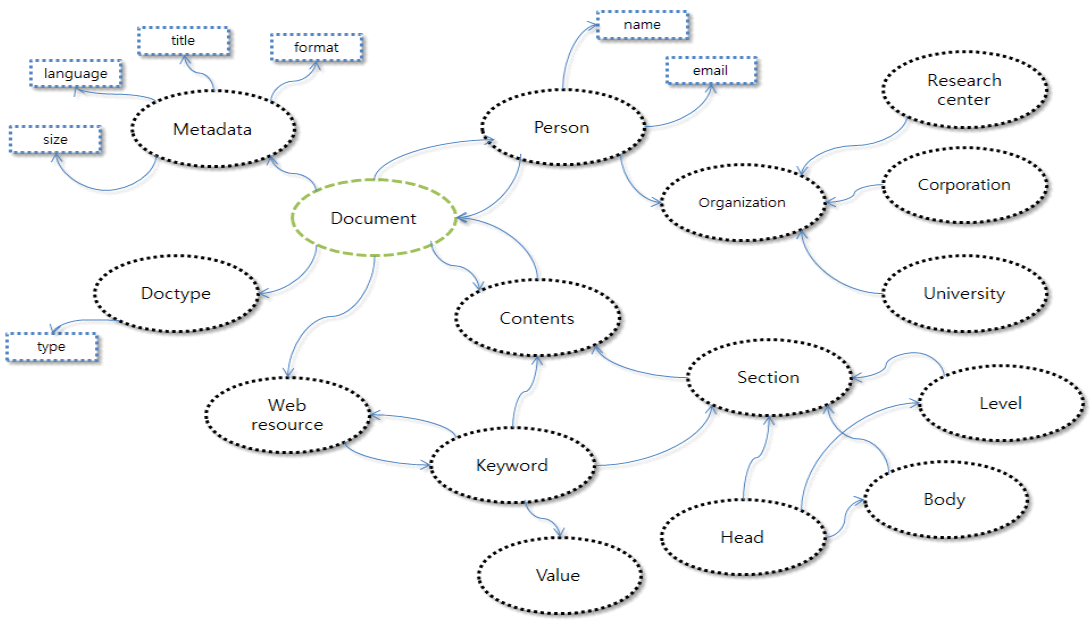


그림 3. D-Ontology 클래스 구조
Fig. 3. Class Structure of The D-Ontology

표 2. 오브젝트 프로퍼티 관계
Table 2. Object Property Relations

#	Domain	Property	Range
1	Document	hasAuthor	Person
2	Document	hasContents	Contents
3	Document	hasWebResource	Webresource
4	Document	hasDocType	DocType
5	Document	hasMetaData	MetaData
6	Person	isDocumentOf	Document
7	Person	foaf:organization	Organization
8	Contents	isBelongTo	Documents
9	Head	hasBody	Body
10	Head	hasLevel	Level
11	Keyword	owl:sameAs	Webresource
12	Keyword	hasValue	Value
13	Keyword	hasSection	Section

클래스로 둔다. 또한, 키워드와 관련된 항목 정보 연결을 위해 hasSection 관계로 연결한다. Webresource 클래스는 해당 문서로부터 얻어낸 키워드와 관련된 DBPedia 페이지의 주소를 인스턴스로 가지는 클래스다. Webresource 클래스의 인스터스 추출은 DBPedia에서 SPARQL Endpoint를 운영하고 있으며, DBPedia에 있는 정보를 SPARQL 질의를 통해 접근할 수 있다. 질의를 통해 얻은 페이지 정보와 해당 키워드를 owl:sameAs 관계를 이용해서 기술하고, 서로간의 참조를 위해 owl:SymmetricProperty 설정을 한다.

본 논문에서 구축된 D-Ontology를 활용하여 해당 문서와 관련된 다양한 질의 연산이 가능하며, 문서 자체의 문서 탐색 뿐이 아니라 LOD 형태의 다른 추가적인 정보도 확장 질의가 가능한 구조를 가지게 된다.

IV. 실 험

1. 온톨로지 구축 시스템의 설계

본 논문에서 구축하는 시스템은 문서 분석 엔진과 온톨로지 구축 엔진으로 나눌 수 있다. 문서 분석 엔진에서는 구조화되지 않은 비정형 문서의 형태 구조화 및 파일처리, 키워드 추출과 관련된 일을 담당하고, 온톨로지 구축 엔진에서는 문서 분석 엔진으로부터 추출된 정보와 도메인 온톨로지와의 연계를 통한 ABox의 자동 구축 일을 담당한다.

그림 4는 본 논문에서 제안하는 비정형 문서의 온톨로지 구축 시스템의 주요 흐름도를 나타낸 것이다. 비정형 문서는 doc, hwp, txt 타입의 문서 포맷을 대상으로 문서 텍스트 추출기에 의해 텍스트 부분만을 뽑아서 문서 분석 엔진으로 전달된다. 문서 분석 엔진에서는 비정형 문서의 항목 추출 대상 저장소와의 비교 연산을 통해 항목 정보를 뽑아내서 계층화 시킨다. 계층화된 문서의 정보는 내부 트리 구조에 의해 저장되어 관리 및 추적이 가능하도록 한다. 추출된 항목 정보를 바탕으로 형태소 분석기에 의해 명사를 추출하고, 추출된 명사의 품사 태깅을 통해 고유명사, 외국어, 보통명사형태로 나눈다. 명사의 종류와 빈도수를 계산해서 가중치를 매기고, 중요 키워드를 선정한다. 선정된 중요 키워드는 내부 해시맵 자료구조로 저장한다. 그리고 문자열 처리 및 파일 처리를 통해 문서로부터 얻을 수 있는 메타정보들을 추출해서 내부 객체로 저장 관리한다.

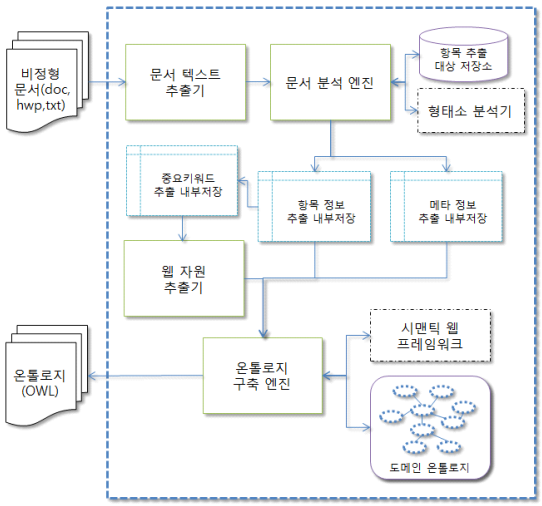


그림 4. 비정형 문서의 온톨로지 구축 시스템 흐름도
Fig. 4. Ontology Construction System Flow Chart of The Unstructured Documents

또한, 웹 자원 추출기에서는 문서로부터 얻은 중요 키워드와 관련된 DBpedia의 정보를 추출하는 역할을 한다. 중요 키워드와 관련된 정보를 얻기 위해 DBpedia의 SPARQL Endpoint에 질의를 던져 DBpedia에 저장된 검색결과를 얻어온다. 질의 방식은 SPARQL의 트리플 패턴 매칭 방식으로 중요 키워드와 관련된 위키페이지 링크 주소와 그와 관련된 외부링크 페이지 주소를 얻어온다. 트리플 패턴 매칭 방식은 주어, 술어, 목적어 형태의 트리플과 매칭 되는 결과를 보여주는 SPARQL의 기본적인 질의 방식이다.

그림 5는 웹 자원 추출기에서 사용하는 SPARQL 질의 표현 방식을 나타낸 것이다. 빨간색 동그라미로 표시된 부분이

문서로부터 추출된 중요 키워드가 들어갈 자리이며 트리플 패턴에서 주어로서의 역할을 한다. 즉, 'Ontology' 키워드와 관련된 위키페이지 링크 주소와 외부 페이지 링크 주소를 알려달라는 질의 처리를 하고 있다.

```
PREFIX db:<http://dbpedia.org/resource/>
PREFIX dbpedia-owl:  <http://dbpedia.org/ontology/>

SELECT ?wikilink ?externLink
WHERE {
{ db:Ontology  foaf:isPrimaryTopicOf  ?wikilink.
  db:Ontology  dbpedia-owl:wikiPageExternalLink  ?externLink}
}
```

그림 5. 웹 링크 추출 질의처리
Fig. 5. Web Link Extraction Query Process

온톨로지 구축 엔진에서는 문서로부터 추출된 계층화된 트리 구조 정보 및 키워드 정보, 웹 자원 주소, 메타정보를 바탕으로 OWL 온톨로지로 구축 한다. 온톨로지 구축 엔진은 오픈소스 기반의 시맨틱 웹 프레임워크를 이용해서 OWL DL 수준의 디스크립션 로직에 따라 구축한다. 구축된 온톨로지는 기본적으로 RDF/XML로 직렬화(serialization)되어 파일로 저장된다. OWL/XML, Turtle 파일 형식 또한 변환 구축 저장되도록 시스템을 구성한다.

2 온톨로지 구축 및 시스템 구현

2.1 도메인 온톨로지 구축

D-Ontology의 TBox 구축은 온톨로지 구축 툴을 이용한 관계 매핑 작업을 거쳐 구축한다. 본 논문에서 사용한 온톨로지 구축 툴은 protege[14]를 이용해서 구축을 한다. 클래스 집합(owl:Class)을 선정하고, 클래스의 계층 관계 설정(rdfs:subClassOf), 제한사항(owl:allValuesFrom, owl:someValuesFrom, owl:hasValue), 정의역(rdfs: domain), 공역(rdfs:range), 오브젝트 프로퍼티(owl: objectProperty), 데이터타입 프로퍼티(owl: datatypeProperty) 등의 여러 관계들을 설정한다. 그림 6은 protege 4.3으로 구축된 D-Ontology의 클래스 및 오브젝트 프로퍼티의 계층 구조를 나타내고 있다.

본 논문에서는 101 Onotology Development 방법론의 7단계에 의해 온톨로지를 구축한다. 다음은 각 단계에 따른 온톨로지 구축 절차에 대해 설명한 것이다. (1) 온톨로지 도메인의 범위는 비정형 문서에 해당하는 기술문서들을 대상으로 하며 (2) 기존 온톨로지 재사용을 위해 링크드 데이터를 활용한 DBpedia와의 연계를 하고, (3) 문서에서 필요한 중요 용어들을 선별 클래스로 구성한다. (4),(5),(6)에 해당하는 클래스 및 속성간의 관계 정의는 해당 용어간의 성격에 따라 공리를 취해 구축한다. 마지막 (7) 인스턴스 생성은 본 논문에서 제안하는 구축 시스템에 의해 문서로부터 추출하여 데이터 공리로 자동 매핑 시킨다.

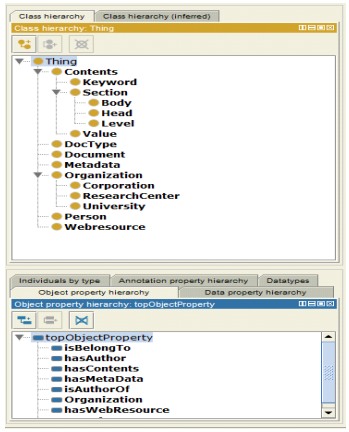


그림 6. 비정형 문서의 온톨로지 변환 형태
Fig. 6. Ontology Transformation Form of The Unstructured Documents

2.2 온톨로지 구축 시스템 구현사항

본 절에서는 시스템 설계를 바탕으로 비정형 문서의 온톨로지 구축 시스템 구현에 대한 부분을 설명한다. 비정형 문서의 온톨로지 구축 시스템은 Java 언어를 사용하여 파일 I/O 기능과, 오픈소스 기반의 시맨틱 웹 프레임워크인 OWL API[15]를 활용하여 구현한다. 시스템은 최종적으로 비정형 문서를 구조화 시키고, 문서의 중요 키워드 및 메타정보를 추출하여 OWL DL 수준의 온톨로지로 구축된다.

OWL API는 OWL2 수준의 문법을 지원하는 인-메모리 방식의 싱글노드에 적합한 오픈소스 프레임워크 이고, Jena[16]와 함께 현재 가장 많이 쓰는 시맨틱 웹 프레임워크로 Jena에 비해 트리플 로딩 속도에서 더 나은 성능으로 측정되었으며[15] RDF/XML, OWL/XML, Turtle, OWL Functional Syntax, OBO Flat 파일 형식까지 파싱, 읽기, 쓰기가 가능하다. 또한, FaCT++, Hermit, Pellet, Racer 등의 추론엔진들도 사용 가능하게 되어 있다. 본 논문에서는 OWL API 기반으로 시스템을 설계 구축하였으며, 비정형 문서를 파싱 추출된 정보를 OWL 온톨로지로 구축하기 위한 로딩, 공리, 매핑 역할을 수행한다. 비정형 문서의 파일 처리는 Java 6의 파일 I/O 기능을 이용해서 수행되며, 항목 추출과 관련해서는 정규 표현식 클래스를 사용해서 구현되었다. 또한, 텍스트 파일 이외의 hwp 파일 변환은 java-hwp[17] 라

이브러리를 이용하였으며, doc 파일에 대한 변환작업은 아파치제단의 POI 3.10[18] 라이브러리를 사용하여 구현하였다. 형태소 분석기는 jhannanum 0.8.3[19]을 이용하여 개발하였다.

2.3 온톨로지 구축 결과

온톨로지 구축은 본 논문에서 구현한 온톨로지 구축 시스템에 의해 비정형 문서로부터 자동으로 정보들을 추출하여 도메인 온톨로지의 인스턴스로 매핑 된다. 표 3은 비정형 문서[20]로부터 자동 추출되어 완성된 온톨로지 구축 현황을 숫자로 나타낸 것이고, 표 4는 구축된 온토롤지의 일부를 RDF/XML형식으로 나타내고 있다. 표 3의 구축 현황을 살펴보면 프로퍼티의 숫자는 오브젝트, 데이터타입 프로퍼티의 숫자를 합한 것이며 71개의 인스턴스가 실험에 사용된 문서로부터 추출되었다. 197개의 공리는 클래스, 프로퍼티, 어노테이션의 관계와 추출된 인스턴스와 자동 매핑 처리되었다. 공리의 숫자는 추론 전의 논리적 공리(logical axiom)의 개수를 나타내고 있다.

표 5는 SPARQL 질의어를 이용해서 온톨로지에 트리플 패턴 매칭 방식으로 정보를 찾는 예를 나타내고 있다. 표 5에서 제시한 질의어는 해당 문서 온톨로지로부터 얻을 수 있는 결과를 질의어로 표현한 것이다.

표 3. 온톨로지 구축 실험 결과
Table 3. Experiment of Ontology Construction Results

Class	Property	Individual	Axiom
16	18	71	197

표 4. 온톨로지 구축 결과 (RDF/XML)
Table 4. Ontology Construction Results (RDF/XML)

```
〈rdf:RDF xmlns="http://dss.ssu.ac.kr/document#"
xmlns:dc="http://purl.org/dc/elements/1.1#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:document="http://dss.ssu.ac.kr/document#"
xmlns:foaf="http://xmlns.com/foaf/0.1#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xml:base="http://dss.ssu.ac.kr/document"〉
〈owl:Ontology
rdf:about="http://dss.ssu.ac.kr/document"〉
〈owl:versionIRI
rdf:resource="http://dss.ssu.ac.kr/document/1.0"/〉
〈/owl:Ontology〉

--중략--

〈!-- http://dss.ssu.ac.kr/document#1 --〉
〈owl:NamedIndividual
rdf:about="http://dss.ssu.ac.kr/document#1"〉
〈rdf:type
df:resource="http://dss.ssu.ac.kr/document#Level"/〉
〈/owl:NamedIndividual〉

〈!-- http://dss.ssu.ac.kr/document#HTTP --〉
〈owl:NamedIndividual
rdf:about="http://dss.ssu.ac.kr/document#HTTP"〉
〈rdf:type
rdf:resource="http://dss.ssu.ac.kr/document#Concept"/〉
〈hasSection
rdf:resource="http://dss.ssu.ac.kr/document#HTTP_프로토콜을_이용한_OWL_온톨로지_접근의_필요성"/〉
〈hasValue
rdf:resource="http://dss.ssu.ac.kr/document#high"/〉
〈owl:sameAs
rdf:resource="http://en.wikipedia.org/wiki/HTTP"/〉
〈/owl:NamedIndividual〉

〈!-- http://dss.ssu.ac.kr/document#OWL --〉
〈owl:NamedIndividual
rdf:about="http://dss.ssu.ac.kr/document#OWL"〉
〈rdf:type
rdf:resource="http://dss.ssu.ac.kr/document#Concept"/〉
〈hasSection
rdf:resource="http://dss.ssu.ac.kr/document#HTTP_프로토콜을_이용한_OWL_온톨로지_접근의_필요성"/〉
〈hasValue
rdf:resource="http://dss.ssu.ac.kr/document#high"/〉
〈owl:sameAs
rdf:resource="http://en.wikipedia.org/wiki/OWL"/〉
〈/owl:NamedIndividual〉

〈!-- http://dss.ssu.ac.kr/document#SOAP --〉
〈owl:NamedIndividual
rdf:about="http://dss.ssu.ac.kr/document#SOAP"〉
〈rdf:type
rdf:resource="http://dss.ssu.ac.kr/document#Concept"/〉
〈hasSection
rdf:resource="http://dss.ssu.ac.kr/document#SOAP메시지_변환_필요성"/〉
〈hasValue
rdf:resource="http://dss.ssu.ac.kr/document#high"/〉
〈owl:sameAs
rdf:resource="http://en.wikipedia.org/wiki/SOAP"/〉
```

```
〈owl:sameAs
rdf:resource="http://shivasoft.in/blog/java/create-soap-message-using-java/"/〉
〈owl:sameAs
rdf:resource="http://www.ibm.com/developerworks/edu/x-dw-cosoap-i.html"/〉
〈owl:sameAs
rdf:resource="http://www.w3.org/2000/xp/Group/"/〉
〈owl:sameAs
rdf:resource="http://www.w3.org/TR/soap/"/〉
〈owl:sameAs
rdf:resource="http://www.w3.org/TR/soap12/"/〉
〈/owl:NamedIndividual〉

--중략--
```

표 5. 온톨로지 질의 예
Table 5. Example of Ontology Query

소속이 대학교인 사람들이 쓴 문서의 주요 키워드와 그와 관련된 웹 페이지의 주소 목록을 출력하는 예

SELECT ?person ?university ?keyword ?resource
WHERE
{
?x document:hasAuthor ?person.
?person foaf:Organization ?university.
?university rdf:type document:University.
?x document:hasContents ?keyword.
?keyword owl:sameAs ?resource
}

키워드가 SOAP이 들어간 문서의 제목과 그 문서의 다른 주요 키워드 및 해당 저자의 소속을 출력하는 예

SELECT ?person ?keyword ?title ?or
WHERE
{
?x document:hasAuthor ?person.
?x document:hasContents document:SOAP.
?x document:hasContents ?keyword.
?x document:hasMetaData ?meta.
?meta dc:title ?title.
?person foaf:Organization ?or.
}

V. 결 론

본 논문은 문서정보검색 분야에 시맨틱 검색이 가능하도록 시맨틱 웹 기술을 활용하여 문서 온톨로지를 구축하였으며, 비정형 문서로부터 얻을 수 있는 정보들을 자동 추출 매핑 시키는 온톨로지 구축 시스템을 설계 및 구현하였다. 본 논문에서 구축한 D-Ontology는 구조화되지 않는 문서로부터 구조화시키기 위한 항목 정보의 추출, 텍스트 마이닝 기술을 활용한 키워드 추출, 문서의 메타정보 추출이 하나의 시스템에서 자동적으로 이루어지며 추출된 데이터는 D-Ontology의 인스턴스로 자동 매핑 처리된다. 또한, 모든 데이터 처리는 OWL DL 수준의 디스크립션 로직 레벨로 작성되며, 링크드 데이터 발행 규칙에 따라 기술된다. 이와 같은 링크드 데이터 형태의 데이터 발행 방식을 이용하여 향후, LOD 형태로 데이터를 공개, 공유할 수 있으며 문서로부터 얻을 수 있는 한정된 정보에서 확장, 추론 질의가 가능한 문서 온톨로지로 발전할 수 있다. 이와 같은 비정형 문서의 자동 온톨로지 구축으로 시맨틱 웹 기술을 이용하여 문서정보검색의 효과성을 증대할 수 있을 것으로 기대된다.

향후, D-Ontology의 TBox 구축에 필요한 추가 스키마 구축이 필요하고, 키워드 추출과 관련된 텍스트 마이닝 기법의 정확도 비교 실험 및 LOD 형태의 오픈 데이터로의 확장, 추론 실험연구가 필요하다.

참고문헌

[1] A. N. Langville, and C. D. Meyer, "Google's PageRank and beyond: The science of search engine rankings," Princeton University Press, 2011.

[2] T. Heath, and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan and Claypool Publishers, 2011.

[3] J. Lehmann, et al., "DBpedia-A Large-scale, Multilingual Knowledge Base Exracted from Wikipedia," Semantic Web Journal, Jun. 2013.

[4] SPARQL Endpoint, http://dbpedia.org/sparql

[5] S. I. Park et al., "A Methodology for Automatic Hierarchy Definition of Sentences in Engineering Documents," Journal of the computational structural engineering institute of Korea, Vol. 22, No. 4, pp. 323-330, Aug. 2009.

[6] N. F. Noy, and D. L. McGuinness. "Ontology development 101:A Guide to Creating Your First Ontology," Stanford Knowledge Systems

Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.
〔7〕 W3C Library Linked Data Incubator Group http://www.w3.org/2005/Incubator/lld/
〔8〕 Simple Knowledge Organization System Reference, http://www.w3.org/TR/skos-reference
〔9〕 Doublin Core, http://dublincore.org
〔10〕 Friend of a Friend, http://xmlns.com/foaf/spec
〔11〕 Naver encyclopedia of knowledge, http://terms.naver.com/entry.nhn?docId=64937&cid=544&categoryId=544
〔12〕 S. H. Han, "A Study on Keyword Extraction From a Single Document Using Term Clustering," Journal of the Korean Society for Library and Information Science, Vol. 44, No. 3, pp. 155-173, Aug. 2010.
〔13〕 State of the LOD Cloud, http://lod-cloud.net/state/
〔14〕 M. Horridge, "A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Editon 1.3," The University of Manchester, 2011.
〔15〕 M. Horridge and S. Bechhofer, "The OWL API_A Java API for OWL Ontologies," Semantic Web Journal, Vol. 2, No. 1, pp. 11-21, 2011.
〔16〕 Apache Jena, https://jena.apache.org/
〔17〕 Java-hwp, https://github.com/ddoleye/java-hwp
〔18〕 The Apache POI Project, https://poi.apache.org/
〔19〕 HanNanum, http://kldp.net/projects/hannanum/
〔20〕 D. W. Jo, J. W. Choi and M. H. Kim, "SPARQL Query Tool for Using OWL Ontology," Journal of The Korea Society of Computer and Information, Vol. 14, No. 11, pp. 21-30, Nov. 2009.
〔21〕 S. S. Cho, D. W. Jo and M. H. Kim, "The Design and Implementation of The Amendment Statement Automatic Generated System for Attached Tables in Legislation," Journal of the computational structural engineering institute of Korea, Vol. 19, No. 4, pp. 111-122, Apr. 2014.

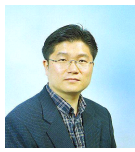
저 자 소 개

조 대 웅

2008: 한림대학교 컴퓨터공학과 공학사
2010: 숭실대학교 컴퓨터학과 공학석사
현재 : 숭실대학교 컴퓨터학과 박사과정
관심분야: 시스템소프트웨어, 분산/병렬 컴퓨팅, 시맨틱 웹, 정보검색, BI
Email : jodw@ssu.ac.kr

최 지 웅

2001: 숭실대학교 컴퓨터학부 학사
2003: 숭실대학교 컴퓨터학과 공학석사
2007 - 2008: 고등기술연구원 연구원
2011: 숭실대학교 컴퓨터학과 공학박사
2013 - 현재: 숭실대학교 컴퓨터학부 교수
관심분야: 시스템소프트웨어, 분산/병렬 컴퓨팅, 시맨틱 웹, BI, 보안
Email : iamjwchoi@gmail.com



김 명 호

1989: 숭실대학교 컴퓨터공학부 학사
1991: 포항공과대학교 전자계산학과 공학석사
1995: 포항공과대학교 전자계산학과 공학석사
1995: 한국전자통신연구소 선임연구원
1998, 2006: 미국 테네시주립대 교환교수
1995 - 현재: 숭실대학교 컴퓨터학부 교수
관심분야: 분산/병렬 컴퓨팅, 시스템소프트웨어, 정보보안, 시맨틱 웹, BI
Email : kmh@ssu.ac.kr