韓國컴퓨터情報學會 論文誌
第15卷 第5號, 2010. 5.

2010-15-5-1-1

키워드를 활용한 온톨로지 인스턴스 생성에 관한 연구

한 광 록*, 강 현 민**, 손 석 원***

A Study on Ontology Instance Generation Using Keywords

Kwang-Rok Han*, Hyun-Min Kang**, Surgwon Sohn***

요 약

시맨틱 웹의 성공 여부는 온톨로지 구축과 생성을 위해서 지식을 체계화하는 시맨틱 어노테이션에 달려있다. 그러므로 각 분야의 많은 지식 표현을 변환하여 온톨로지 인스턴스로 생성하기 위해서 시맨틱 어노테이션의 효율성이 중요하다. 본 논문에서는 기존 웹에서 시맨틱 어노테이션 작업을 통하여 온톨로지 인스턴스를 정확하고 효율적으로 생성하는 규칙기반 온톨로지 인스턴스 생성 시스템을 제안한다. 기존연구에서는 사용자가 관련 정보를 찾아서 온톨로지와 대조하여 정보를 입력하는 수동적인 과정이 필요하였다. 그러나 제안한 방식에서는 추출할 정보들에 관한 키워드 데이터와 규칙정보를 분할해서 관리한다. 따라서 소수의 키워드와 규칙정보들을 추가함으로써 다양한 웹문서의 효율적 정보 추출이 가능하다. 이것은 여러 사이트에서 규칙과 키워드를 재사용할 수 있는 온톨로지 인스턴스 생성이 가능하다는 것을 보여준다.

Abstract

The success of semantic web depends largely on the semantic annotation which systematizes knowledge for the construction and production of ontology. Therefore, the efficiency of semantic annotation is very important in order to change many knowledge expressions and generate into ontology instances. In this paper, we presents a generation system of rule-based ontology instances which are produced accurately and efficiently via semantic annotation in conventional web sites. In conventional studies, the manual process is necessary for finding relevant information, comparing it with ontology, and entering information. We propose a new method that manages keyword data regarding extracted information and rule information separately. Thus, it is quite practical to extract information efficiently from various web documents by adding a small number of keywords and rules. The proposed method shows the possibility of ontology instance generation which reuses the rules and keywords from the various websites.

▸ Keyword : 온톨로지(Ontology), 인스턴스 생성(Instance Generation), 키워드(keyword)

• 제1저자 : 한광록 교신저자 : 손석원
• 투고일 : 2009. 12. 23, 심사일 : 2010. 01. 26, 게재확정일 : 2010. 04. 07.
* 호서대학교 컴퓨터공학부 교수 **호서대학교대학원 메카트로닉스공학과 석사과정
*** 호서대학교 뉴미디어학과 부교수
※ 이 논문은 2009년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임(2009-0016)

Ⅰ. 서 론

시맨틱 웹(Semantic Web)은 기존 웹을 확장하여 컴퓨터가 이해할 수 있는 잘 정의된 의미를 기반으로 인간과 컴퓨터 간의 효과적인 협력체계를 구축하기 위한 것이다. 일반적으로 시맨틱 웹은 온톨로지(Ontology), 주석화된 웹(Annotated Web), 그리고 사용자를 대신하여 정보를 수집, 검색하는 에이전트(Agent)로 구성된다[1]. 즉, 시맨틱 웹의 목적은 웹에 산재한 정보의 의미를 사용자인 인간뿐만 아니라 자동화된 기계 즉, 에이전트 프로그램이 그 의미 정보를 해석할 수 있는 일종의 표준 의미정보의 교환 수단을 만드는 것이다[2]. 따라서 웹 상에 방대한 양의 온톨로지가 산재하고, 이를 자동으로 해석하여 처리할 수 있는 에이전트 소프트웨어에 사람 또는 에이전트가 질의를 하면, 컴퓨터가 자동으로 분산된 온톨로지를 탐색하고 추론하여 원하는 결과를 돌려주는 것이다.

시맨틱 웹이 활성화되기 위해서는 온톨로지와 같은 컴퓨터가 이해할 수 있는 의미정보를 포함하고 있는 웹 페이지가 대량생산되어야 한다. 온톨로지로 주석화 된 웹페이지의 개발이 어려운 만큼, 기존의 웹 페이지에 컴퓨터가 이해할 수 있는 의미정보를 추가하는 것이 현실적인 방법이다[1]. 이러한 목적을 이루기 위한 기본이 되는 기술이 바로 HTML 문서로 구성된 기존의 웹에서 정보를 추출하여 온톨로지 인스턴스(Instance)를 생성하는 것이다[3]. 기존 웹 문서에서 정보를 추출할 때의 문제점은 웹 문서의 목적이 사람에게 시각적으로 정보를 제공하는 사이트 및 문서 작성자의 스타일에 따라서 사용하는 단어와 레이아웃이 서로 다르다는 것이다. 따라서 원하는 정보를 추출하기 위해서 화면 구성 관련 내용과 이미지 그리고 형식(Format) 등의 정보를 제거해야 한다. 또한 의미는 같지만 어휘를 서로 다르게 사용하는 여러 정보 소스로부터 필요한 정보를 추출해야 하는데 이것은 매우 힘든 과정이다[4].

본 논문에서는 구조화 또는 준 구조화된 웹 문서를 분석하여 데이터 규칙 정보 및 사이트 규칙 정보를 생성하고 추출할 지식분야의 키워드를 활용하여 네이버 및 다음 등 5개 사이트의 웹 문서로부터 자료가치가 있는 정보를 추출한다. 추출된 정보를 도메인 온톨로지를 통하여 온톨로지 인스턴스를 자동으로 생성하는 방법에 대하여 기술한다. 규칙과 키워드를 분할해서 관리를 하므로 다른 지식분야의 온톨로지 인스턴스 생성시 키워드 교체만으로 보다 적은 비용으로 효율적인 온톨로지 인스턴스를 생성할 수 있다.

제2장에서는 온톨로지 인스턴스 생성과 관련된 연구를 기술하고, 제3장에서는 본 연구에서 제안한 방법인 데이터 및 사이트 규칙정보와 키워드를 이용한 온톨로지 인스턴스 생성의 구조 및 과정을 설명한다. 제4장에서는 제안된 방법을 이용하여 5개의 영화 사이트에 적용하여 규칙과 키워드를 이용한 온톨로지 인스턴스를 생성실험을 하였다. 제5장에서 결론을 맺는다.

Ⅱ. 관련 연구

시맨틱 웹은 정보의 처리와 검색을 기계가 대신할 수 있는 방안을 찾아서 자동으로 정보의 처리와 검색이 이루어지게 하는 기술이다. 이를 위해서 웹 페이지에 적절한 주석(Annotation)을 달아서 이를 기계가 이해할 수 있도록 한다. 주석화된 웹 구현 방식은 시맨틱 웹을 구현하는데 있어서 가장 현실적인 방법 중 하나이다. 주석화된 웹 구현은 이미 존재하는 웹 페이지에 대하여 온톨로지를 기반으로 추가적인 설명을 덧붙이는 것으로 주로 정보 검색의 정확도를 높이는데 크게 기여할 수 있다[5].

주석방식에 있어서 핵심요소인 온톨로지에 관해서는 그것의 다양한 활용 범위와 높은 시장성 그리고 사회, 경제 및 기술적 파급 효과에 대한 인식이 커지면서 국내외적으로 온톨로지 구축과 응용에 대한 연구가 활발하게 진행되고 있다[6,7].

최근 수동이나 자동으로 웹 문서의 주석을 용이하게 하는 어노테이션 도구들을 기반으로 온톨로지를 개발하는 연구들이 진행되고 있다. 예를 들면, KIM이라는 시맨틱 어노테이션 플랫폼(Semantic Annotation Platform)은 지식 정보 관리(KIM:Knowledge Information Management)와 자동화된 시맨틱 어노테이션, 인덱싱, 그리고 비구조화 및 반자동화된 콘텐츠들의 검색을 위한 서비스를 제공한다[8]. 또 다른 도구인 MnM은 마찬가지로 시맨틱 웹 콘텐츠의 페이지를 주석처리하기 위하여 반자동 및 자동방식 모두를 지원한다[9].

국내에서 개발된 시맨틱 어노테이션 도구들은 서울대학교에서 개발한 SARM과 KISTI에서 개발한 OntoManager 등이 있다[10,11]. SARM은 특정 도메인과 언어의 주석에 적합한 도구이다. 이들은 웹 페이지로부터 개체명을 추출하고 온톨로지 인스턴스에 맵핑시키는 방식을 사용하여 시맨틱 정보를 구축하는 것을 목표로 하고 있다. 이 기법을 특정 웹 페이지에 대한 시맨틱 메타 데이터 구축에 활용한다면 해당 웹 페이지를 시맨틱 웹 응용 서비스를 통해 접근시키는 것이 가능해질 것이다. 또한 2006년부터 KAIST를 중심으로 국가 IT 온톨로지 인프라 기술 개발을 진행하고 있다. 그 중에서 IT 핵심 온톨로지 구축 과정의 하나로 의미주석화 (Semantic Annotation) 작업을 진행하고 있으며, 주석이라는 용어가

전통적으로 문서의 특별한 부분에 설명과 정보를 기록하는 것을 의미하고 있듯이 IT 핵심 온톨로지를 위한 의미 주석화 온톨로지 구축 시스템인 OntoCS를 이용하여 반자동으로 진행 되었다[12]. 이밖에도 온톨로지를 기반으로 웹 페이지의 텍스트를 블록 지정해서 온톨로지 클래스 트리에 드롭하는 방식으로 주석화하는 방식도 소개되고 있다[13].

III. 법용 온톨로지 인스턴스 생성

1. 온톨로지 인스턴스 생성을 위한 데이터

본 논문에서 제안한 시스템은 크게 원본 데이터에서 정보를 추출하는 데이터 추출기와 추출한 정보를 이용하여 온톨로지 인스턴스를 생성하는 인스턴스 생성기로 구성된다. 그림 1은 본 논문에서 제안한 시스템의 구조를 나타낸다.

데이터 추출과정은 다음과 같다. 우선 웹에 산재되어 있는 웹문서를 변환하여 HTML 형식의 원본 데이터로 저장한다. 해당 웹 문서를 해석하기 위한 사이트 규칙 정보를 만들고 원본 데이터에 적합한 규칙들을 분류한다. 분류한 규칙들을 원본 데이터에 적용시켜 콘텐츠를 추출한다. 원본 데이터에 해당하는 분야의 키워드를 저장하고 있는 키워드 데이터에서 키워드를 로드한다. 추출한 콘텐츠에 로드한 키워드와 규칙을 적용시켜 추출된 정보는 추출 정보 데이터라고 한다.

인스턴스 생성과정은 데이터 추출과정에서 생성한 추출 정보 데이터를 이용하여 추출한 정보에 해당하는 분야의 개념 온톨로지를 이용하고 추출한 정보에 해당하는 키워드에 관련된 온톨로지의 클래스와 객체속성을 검색한다. 추출한 정보에 검색한 클래스와 객체속성을 적용시켜서 개념 온톨로지에 입력하면 OWL 온톨로지를 생성할 수 있고, 이를 추론하여 결과에 따라 온톨로지 인스턴스 데이터베이스에 저장한다.

1.1 데이터 규칙 정보(Rule.xml)

데이터 규칙 정보는 웹 페이지의 구조화되어 있는 문서에서 정보를 추출하기 위한 규칙들을 정의해놓은 XML형식의 데이터이다. 규칙은 사이트별로 관리하지 않고 통합으로 관리함으로써 동일한 규칙이 필요한 사이트의 경우에 재작성 할 필요 없이 그대로 이용 가능하도록 하였다.

표 1. 규칙 명세
Table 1. Rule Specification

규칙	정의
Content- Rule	문서내의 추출할 정보가 정의되어 있는 테이블의 위치를 추출해내기 위한 규칙으로 테이블의 시작위치를 검색하기 위한 규칙과 테이블의 끝을 검색하기 위한 규칙으로 되어 있다.
Keyword-Rule	정보를 추출하기 위하여 문서 내를 키워드 검색할 때 적용시켜야하는 규칙으로서 태그 내에서 키워드를 검색해야하는지에 대한 규칙과 키워드를 검색할 범위를 나타내는 규칙으로 이루어져있다.
DataRule	추출된 정보가 하나의 정보가 아닌 여러 정보들이 복합되어 있을 때 이를 분리시키기 위한 규칙으로서 정보를 분리시킬때 적용할 분리 문자를 규칙으로 설정해 정보를 분리한다.

데이터 규칙 정보는 표 1과 같이 3가지의 규칙으로 구분되어 있다. 규칙은 해당하는 규칙헤더로 구분될 수 있게 정의되어 있으며 규칙헤더의 하위 노드엔 그에 해당하는 규칙이 정의되어 있다.

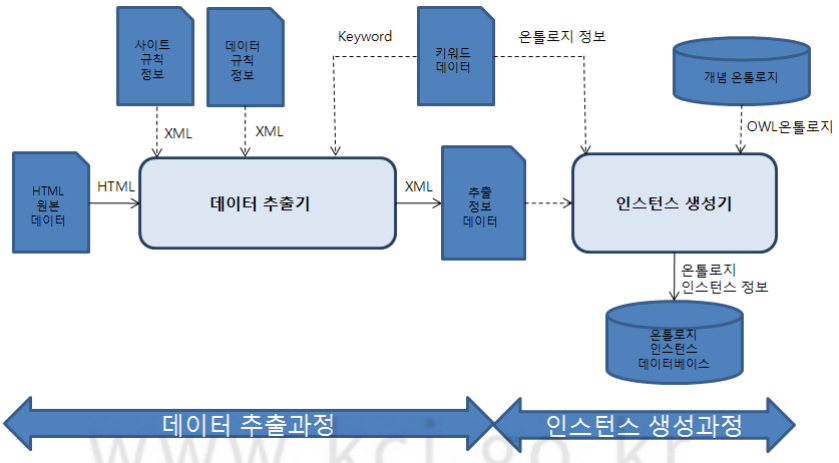

그림 1. 온톨로지 인스턴스 생성 시스템 구조
Fig. 1. System Architecture of Ontology Instance Generation

표 2는 본 논문의 데이터 규칙 정보에 정의한 규칙에 대한 설명이다. 현재 정의해 놓은 규칙들을 각 사이트에 적용한 규칙들만 사이트 규칙 정보에 규칙 헤더로서 가지고 있고, 정보 추출시 사이트 규칙 정보에 정의해놓은 규칙들을 사이트별로 적용 가능하게 하였다. 다른 새로운 사이트에서 정보를 추출할 때에도 데이터 규칙 정보에 미리 정의해 놓은 규칙을 이용하는 것이 가능하며 새로운 규칙이 필요할 때만 새롭게 정의하고 이에 따른 규칙 헤더만 사이트에 적용시키면 된다.

1.2 사이트 규칙 정보(Site.xml)

데이터 규칙 정보에서 통합으로 관리하고 있는 규칙을 참조하여 각 사이트에 적합한 규칙들을 사이트와 분야별로 관리하기 위한 데이터를 사이트 규칙 정보라고 한다. 사이트 규칙 정보는 규칙을 적용시킬 사이트 이름과, 추출 정보가 해당하는 분야, 사이트의 인코딩 방식, 적용시킬 콘텐츠 규칙(ContentRule)과 키워드 규칙(KeywordRule), 데이터 규칙(DataRule)으로 구성된 규칙헤더(Rule Header)가 저장되어 있다.

1.3 키워드 데이터(Keyword.xml)

키워드 데이터는 원본 데이터에서 추출할 정보를 검색하기 위한 키워드가 저장되어 있으며 각 키워드에 따른 개념 온톨로지의 클래스와 객체 속성을 저장하고 있다. 키워드는 각 분야별로 분류되며 추출대상인 원본 데이터의 분야에 따라 키워드를 검색할 수 있다. 키워드 데이터는 원본데이터의 분야와 이에 따른 키워드, 그리고 해당분야의 개념 온톨로지 경로이름과 키워드에 해당하는 온톨로지 클래스 및 객체속성으로 구성되어 있다.

표 2. 정의된 규칙
Table 2. Defined Rules

규칙분류	규칙	비고
Content-Rule	ContentTableStart	추출할 정보가 있는 테이블의 시작 위치를 나타내는 규칙
	ContentTableEnd	추출할 정보가 있는 테이블의 끝 위치를 나타내는 규칙
Keyword-Rule	DelectTag	키워드 검색시 HTML태그를 전부 삭제할 것인가를 나타내는 규칙
	RoopStart	콘텐츠 내에서 키워드 정보 데이터에 있는 키어드로 검색을 시작할 위치
	RoopEnd	RoopStart규칙에 의해서 키워드 검색이 되었을 시 정보 추출을 완료하는 위치를 나타내는 규칙
	RoopCount	RoopStart 규칙이 시점부터 일정범위 안에 RoopEnd가 없을 시 추출된 데이터를 무효화시키는 규칙
	TitleLoad	원본문서의 〈title〉태그에 정보가 있을 시 정보 추출을 위한 규칙
DataRule	Contentdivide	추출한 데이터에 여러 데이터가 복합적으로 구성되어 있을 시 이를 분리하기 위한 분리규칙(예 : 애니메이션|미국|94분)
	Datadivide	동일한 데이터가 일정 특수문자로 묶여있을 시 이를 분리하기 위한 규칙 (예 : 애니메이션,액션,코미디)
	Titlesplit	TitleLoad 규칙으로 추출한 정보에서 유효한 정보만 추출하기 위한 규칙

1.4 추출 정보 데이터(Output.xml)

원본 데이터에서 데이터추출기를 통하여 추출된 정보를 담고 있는 XML형식의 파일을 추출 정보 데이터라고 한다. 추출 정보 데이터에는 원본 데이터의 지식분야와 키워드, 그리고 각 키워드에 인덱싱되는 정보들이 저장되어 있다. 이 인덱스 정보 데이터를 인스턴스 생성기로 입력되어 개념 온톨로지를 이용해서 온톨로지 인스턴스를 생성한다.

1.5 개념 온톨로지

개념 온톨로지는 기본 정보를 포함하고 있는 OWL 온톨로지이다. 본 논문의 개념 온톨로지인 영화 개념 온톨로지는 영화 타이틀과 장르, 국가, 개봉 날짜 등의 영화에 관련된 클래스와 객체속성들로 이루어져있다. 정상적인 정보 추출의 판단은 키워드 데이터에 저장되어 있는 개념 온톨로지 정보에 따라서 해당되는 개념 온톨로지에 정보를 저장하고 추론을 하였을 때 오류가 발생하는지의 여부로 판단한다.

1.6 온톨로지 인스턴스 데이터베이스

온톨로지 인스턴스 데이터베이스는 정상적으로 생성된 온톨리지 인스턴스를 저장하는 데이터베이스로서 본 논문에서는 영화 온톨로지 인스턴스 데이터베이스를 구축하였다.

2. 데이터 추출 과정

그림 2는 웹에 산재되어 있는 웹문서들을 검색하여 온톨로지 인스턴스 생성에 필요한 정보를 추출하는 과정을 나타낸다.

2.1 HTML 파서

원본 데이터에 있는 HTML 문서를 파싱하고 그에 따른 관련 사이트 규칙 정보를 콘텐츠 추출기에 전달한다.

2.2 콘텐츠 추출기

HTML 파서로부터 전달받은 사이트 규칙 정보를 이용하여 데이터 규칙 정보에 있는 ContentRule을 원본 데이터에 적용하여 정보를 추출한 콘텐츠의 위치를 파악하고 이를 출력 정보 생성기에 전달한다.

2.3 추출정보 생성기

원본데이터가 해당하는 분야를 키워드 데이터에서 검색하여 해당하는 분야에 관련된 키워드를 검색결과로 받아온다. 이 키워드를 콘텐츠추출기에서 전달받은 콘텐츠에 KeywordRule과 DataRule을 적용하여 정보를 추출하고 추출정보 데이터(Output.xml)를 생성한다. 본 논문에서는 원본데이터가 영화 분야이므로 키워드 데이터에서 영화 분야에 속하게 설정해 놓은 키워드인 장르, 국가, 상영시간, 개봉일자 등의 키워드를 검색결과로 받아오게 된다. 각 키워드에 KeywordRule과 DataRule을 적용시켜서 콘텐츠를 검색하여 그 결과를 추출정보 데이터로 생성하게 된다.

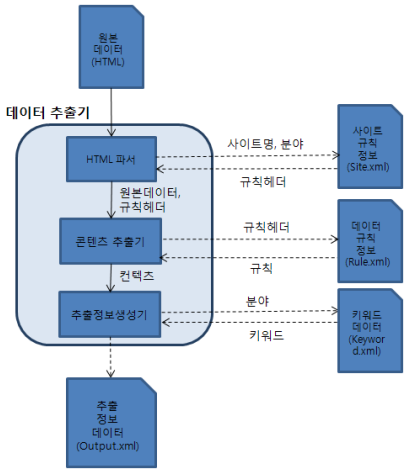

그림 2. 데이터 추출 프로세스
Fig. 2. Process of Data Extraction

3. 인스턴스 생성 과정

그림 3은 데이터 추출과정에서 생성한 추출정보 데이터를 대상으로 개념온톨로지를 참조하여 온톨로지 인스턴스를 생성하는 과정을 나타낸다.

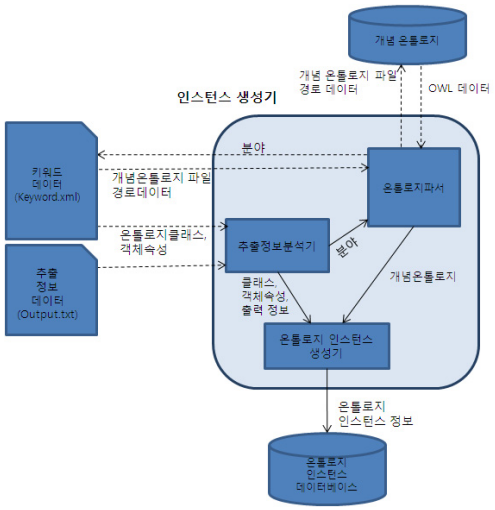

그림 3. 온톨로지 인스턴스 생성 프로세스
Fig. 3. Process of Ontology Instance Generation

3.1 추출정보 분석기

데이터 추출기에서 생성한 추출정보 데이터(Output.xml)와 그에 따른 키워드 데이터를 비교하여 개념 온톨로지 경로 이름과 온톨로지 클래스 및 객체속성을 검색한다. 개념 온톨로지 경로이름은 온톨로지 파서로 전달되고 온톨로지 클래스와 객체 속성 등의 추출한 정보는 온톨로지 인스턴스 생성기로 전달한다.

3.2 온톨로지 파서

추출한 정보에 해당하는 분야의 개념 온톨로지를 로드하기 위한 온톨로지 파서이다. 출력정보 분석기에서 전달받은 분야의 개념 온톨로지 파일 경로를 키워드 데이터에서 로드하여 해당 개념 온톨로지를 파싱하여 온토롤지 인스턴스 생성기에 전달한다.

3.3 온톨로지 인스턴스 생성기

온톨로지 파서에서 전달받은 개념 온톨로지에 출력정보 분석기에서 전달받은 온톨로지 클래스와 객체 속성을 적용하여 추출한 정보들에 대응하는 OWL 온톨로지를 생성하고 이를 추론하여 결과에 따라 온톨로지 인스턴스를 온톨로지 인스턴스 데이터베이스에 저장한다.

4. 영화 온톨로지 인스턴스 생성과정

본 절에서는 앞에서 기술한 범용 온톨로지 생성 시스템을 영화 사이트에 적용하여 온톨로지를 생성해내는 과정을 예로

들어 설명한다. 영화 정보에 필요한 개념 온톨로지를 정의하고 몇 개의 키워드와 규칙정보만을 추가함으로써 간단히 영화 사이트와 관련된 정보의 추출이 가능하다.

```
<?xml version="1.0" encoding="euc-kr" ?>
- <SiteConfig>
  - <SiteRule>
      <Site name="naver" />
      <Section value="영화" />
      <naverEncoding>EUC-KR</naverEncoding>
      <naverContentRule>ContentRule1,ContentRule2</naverContentRule>
      <naverKeywordRule>KeywordRule2,KeywordRule3,KeywordRule4,KeywordRule5,KeywordRule7</naverKeywordRule>
      <naverDataRule>DataRule1,DataRule2,DataRule5</naverDataRule>
    </SiteRule>
  - <SiteRule>
      <Site name="daum" />
      <Section value="영화" />
      <daumEncoding>UTF-8</daumEncoding>
      <daumContentRule>ContentRule3,ContentRule2</daumContentRule>
      <daumKeywordRule>KeywordRule1,KeywordRule2,KeywordRule3,KeywordRule6,KeywordRule7</daumKeywordRule>
      <daumDataRule>DataRule1,DataRule2,DataRule6</daumDataRule>
    </SiteRule>
```

그림 4. 사이트 규칙 정보
Fig. 4. Site Rule information

그림 4는 사이트 규칙 정보의 XML구조를 보여주고 있으며 네이버와 다음의 정보를 추출하기 위한 규칙들을 각각 정의하고 있다. 그림 5는 네이버와 다음의 구조화된 영화 문서를 보여주고 있다. 구조화되어 있는 부분에 영화에 대한 전반적인 정보들이 포함되어 있기 때문에 영화 온톨로지를 구성하는데 필요한 요소들을 모두 얻을 수 있다. 그림 6은 그림 5의 구조화된 영화 문서로부터 정보를 추출하기 위한 규칙들을 정의한 XML 형식의 정보 데이터(Rule.xml)이다. 그림 7은 원본 데이터에 규칙을 적용한 예로서 사이트 규칙 정보는 그림 7에 ①~⑦으로 표시되어 있듯이 사이트에 적용될 규칙에 규칙헤더를 가지고 있으며 데이터 규칙 정보는 각 규칙헤더에 대해 규칙을 정의하고 있다.

A~G는 데이터 규칙 정보의 규칙들이 원본 데이터에 적용되는 예로서 A는 CotentTableStart로 원본 데이터내의 콘텐츠의 시작 위치를 나타내는 규칙이며, B는 ContentTableEnd로서 콘텐츠가 끝나는 위치를 나타내는 규칙이다.

그림 5. 네이버와 다음의 구조화된 문서
Fig. 5. Structured Documents of Naver and Daum

```
<?xml version="1.0" encoding="euc-kr" ?>
- <Rule>
  - <ContentRule>
    - <ContentRule1>
        <ContentTableStart>dl class=\"summary\"</ContentTableStart>
      </ContentRule1>
    - <ContentRule2>
        <ContentTableEnd>></dl></ContentTableEnd>
      </ContentRule2>
    - <ContentRule3>
        <ContentTableStart>dl class="cu mainInfo"</ContentTableStart>
      </ContentRule3>
    </ContentRule>
  - <KeywordRule>
    - <KeywordRule1>
        <DelectTag>TRUE</DelectTag>
      </KeywordRule1>
    - <KeywordRule2>
        <RoopStart>dd</RoopStart>
      </KeywordRule2>
    - <KeywordRule3>
        <RoopEnd>/dd</RoopEnd>
      </KeywordRule3>
    - <KeywordRule4>
        <RoopCount>1</RoopCount>
      </KeywordRule4>
    </KeywordRule>
  - <DataRule>
    - <DataRule1>
        <Contentdivide>[|]</Contentdivide>
      </DataRule1>
    - <DataRule2>
        <Datadivide>[,]</Datadivide>
      </DataRule2>
    - <DataRule3>
        <Datadivide>-</Datadivide>
      </DataRule3>
    - <DataRule4>
        <Datadivide>[.]</Datadivide>
      </DataRule4>
    - <DataRule5>
        <Titlesplit>::</Titlesplit>
      </DataRule5>
    - <DataRule6>
        <Titlesplit>[-]</Titlesplit>
      </DataRule6>
    </DataRule>
  </Rule>
```

그림 6. 데이터 규칙 정보
Fig. 6. Data Rule Information

C는 콘텐츠 내의 추출할 정보가 위치하는 시작점을 나타내는 규칙이며 D는 추출할 정보가 위치하는 끝지점을 나타내는 규칙이다. E는 원본 데이터의 타이틀을 분석한다는 규칙

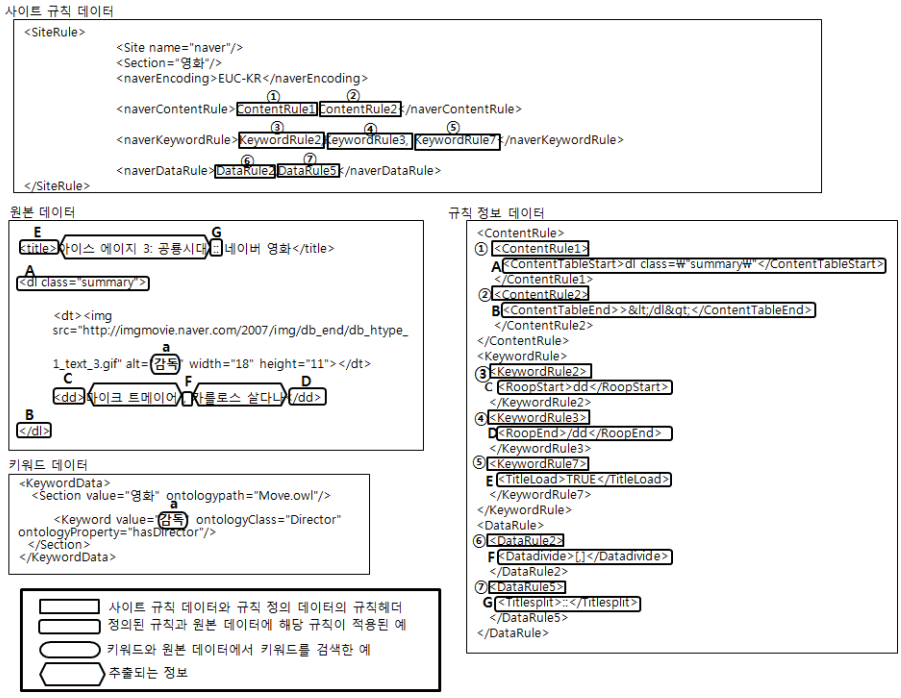

그림 7. 규칙 적용의 예
Fig. 7. Example of Rule Application

그림 8. 키워트 데이터
Fig. 8. Keyword Data

이며 F는 정보가 2개 이상이 묶여있을 시 이를 분리하기 위한 규칙이고 G는 타이틀에서 필요한 정보만 분리하기 위한 규칙이다.

정보의 추출을 위해서는 키워드 데이터에서 추출한 정보에 관련 키워드를 로드하여 콘텐츠를 검색한다. a는 콘텐츠를 감독이라는 키워드로 검색하여 그 위치를 파악하고 감독의 정보를 추출한다.

그림 8은 영화정보를 추출하기 위해서 작성한 키워드 데이터이다. Section은 추출할 정보가 해당하는 분야이며 ontologypath는 온톨로지 경로를 가지게 되어 있다. 검색할 키워드의 정의로서 Keywords는 여러 정보들의 묶음인 키워드를 나타내는 것이고, Keyword는 추출할 정보를 검색하기 위한 키워드이다. ontoloyClass는 해당 키워드의 온톨로지클래스이고 ontologyProperty는 해당 키워드의 온톨로지 속성이 된다.

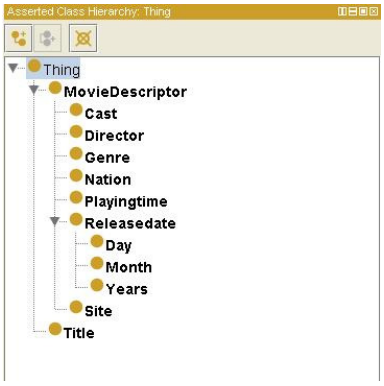

그림 9. 영화 개념 온톨로지
Fig. 8. Conceptual Ontology of Movie

그림 9는 다음과 네이버에서 추출한 정보로 온톨로지 인스턴스를 생성하기 위한 개념 온톨로지를 나타내고 있다.

Ⅳ. 실험 및 평가

본 장에서는 구현된 시스템을 평가하기 위해 다음과 같은 방법으로 실험한다. 특정 사이트의 특정 분야에 대한 웹 문서를 수집하여 본 논문에서 제안한 시스템으로 원하는 정보를 추출하여 정보 추출 신뢰도를 측정하였고 추출한 정보로 온톨로지 인스턴스를 생성하여 온톨로지 인스턴스 생성률을 측정하였다.

1. 원본 데이터 수집

본 논문의 원본 데이터 수집은 구조화가 잘 되어 있는 포털사이트인 네이버, 다음, 야후, 네이트와 영화예매 사이트인 맥스무비에서 제공하는 영화정보 페이지 중 각 사이트 별로 100 페이지를 수집하여 원본 데이터로 이용하였다.

2. 평가방법

각 사이트에 적합한 사이트 규칙 정보와 데이터 규칙 정보, 키워드 데이터를 작성하고 각각의 사이트에서 추출된 100 페이지의 웹문서 원본 데이터를 데이터 추출기에 입력한다. 그리고 추출된 추출 정보 데이터를 원본데이터와 비교하여 추출 결과를 측정하고 이를 인스턴스 생성기에 입력하여 온톨로지 인스턴스의 생성 결과를 측정한다.

3. 실험 결과

본 논문에서 제안한 시스템은 웹문서의 URL를 데이터 추출기에 입력하여 정상적으로 추출된 추출 정보 데이터인지 확인한다. 이 추출정보 데이터를 인스턴스 생성기에 입력하여 온톨로지 인스턴스를 얻는 방식으로 실험을 진행하였다.

그림 10은 데이터추출기를 통해 나온 결과물로서 맥스무비 사이트의 출력 정보 데이터들과 가필드라는 영화에 대한 출력 결과를 보여주고 있다.

그림 10. 데이터 추출 결과
Fig. 10. Result of Data Extraction

표 3. 실험 결과
Table 3. Experimental Results

(1) 실험1

	원본데이터	추출 정보 데이터	정보 추출 신뢰도
다음	100	100	100%
네이버	100	99	99%
야후	100	93	93%
네이트	100	97	97%
맥스무비	100	98	98%
합계	500	487	97%

(2) 실험2

	실험1의 추출 정보 데이터	온톨로지 인스턴스 생성	온톨로지 인스턴스 생성률
다음	100	75	75%
네이버	99	77	78%
야후	93	76	82%
네이트	97	68	70%
맥스무비	98	62	63%
합계	487	358	73%

그림 11은 출력 정보 데이터를 인스턴스 생성기가 로드하여 개념 온톨로지를 적용시켜 생성된 온톨로지를 보여 주고 있다.

그림 11. 온톨로지 인스턴스 생성 결과
Fig. 11. Generation Result of Ontology Instance

실험결과는 표 3에 나타내었다. 실험 1은 각 사이트의 원본 데이터에서 추출한 정보추출 신뢰도이며 실험 2는 실험 1의 추출정보 데이터의 온톨로지 인스턴스 생성률을 나타내고 있다.

실험 결과 본 시스템에서 제안한 온톨로지 인스턴스 생성 방법을 사용할 경우, 같은 분야의 구조화된 문서에서 정보를 추출할 경우 정상적인 정보를 추출하는 정보추출 신뢰도는 평균 97%의 신뢰도를 나타내었다. 그러나 온톨로지 인스턴스 생성은 평균 73%의 인스턴스 생성율을 나타내었다. 추출한 정보의 인스턴스 생성률이 떨어지는 이유는 사이트 내에 출연자나 상영시간 등의 정보누락으로 인하여 개념 온톨로지에서 설정해놓은 조건을 원본 데이터가 충족하지 못하는 경우였다. 본 논문에서 제안한 온톨로지 인스턴스 생성 방법에서는 사이트 규칙 정보와 데이터 규칙 정보, 키워드 데이터의 세 가지의 데이터 작성만으로도 여러 사이트의 같은 분야의 정보를 추출할 수 있다. 또한 추출한 정보를 가지고 온톨로지 인스턴스를 생성할 수 있다는 것을 확인할 수 있었으며 이것은 각각의 사이트마다 서로 다른 규칙을 만들어서 인스턴스를 생성하는 것보다 효율적임을 알 수 있었다. 이를 응용하고 보완하여 각각의 사이트에서 분야별로 규칙을 정의해주고 키워드 데이터만을 변경함으로써 동일 사이트의 다른 분야의 정보를 추출하여 온톨로지 인스턴스를 생성할 수 있을 것이다.

V. 결론

기존의 규칙기반 온톨로지 인스턴스 생성시스템에서는 규칙과 키워드의 통합된 구조로 인해서 각각의 사이트마다 규칙을 재정의해야 하는 번거로움이 있다. 본 논문에서는 규칙과 키워드를 분리하여 관리함으로써 여러 사이트에서 이를 재사용할 수 있는 범용 규칙기반 온톨로지 인스턴스 생성 시스템에 대하여 기술하였다. 이를 위하여 사이트 규칙 정보와 데이터 규칙 정보를 기술하였고 영화 분야에 대한 키워드 데이터를 기술하였다. 이것을 기반으로 데이터추출기를 생성하여 추출 정보 데이터를 생성하였다. 또한 추출 정보 데이터는 인스턴스 생성기를 통하여 온톨로지 인스턴스로 가공되어 온톨로지 인스턴스 데이터베이스로 저장되고 추후 온톨로지 기반의 검색엔진에 사용할 수 있도록 하였다.

논문의 검증을 위하여 다음, 네이버, 야후, 네이트, 맥스무비에서 제공하는 영화 정보 500건을 대상으로 정보 추출과 온톨로지 인스턴스 생성을 수행하여 성능을 평가하였다. 결과적으로 사이트 규칙 정보에 각 사이트의 규칙헤더를 추가하고 데이터 규칙 정보에 정의되어 있지 않은 규칙을 추가하여 키워드 데이터의 키워드를 추가함으로써 각 사이트에서 정보를 추출하고 온톨로지 인스턴스가 생성된다는 것을 확인하였다. 이는 여러 사이트에서 규칙정보와 키워드를 재사용함으로써 온톨로지 인스턴스 생성이 가능함을 보여주었다.

하지만 정보 추출 신뢰도에 비해 온톨로지 인스턴스 생성률이 낮았고 5개의 사이트의 규칙만 데이터 규칙 정보가 추가되어 있기에 다른 사이트의 정보 추출시 데이터 규칙 정보가 없는 규칙을 계속 추가해야하는 번거로움이 있다. 따라서 차후에 보다 많은 사이트의 데이터를 추출하고 온톨로지 인스턴스를 생성하기 위하여 다양한 사이트 분석을 통한 효율적 데이터 규칙 정보의 구축이 필요하다. 즉, IT부품을 판매하는 다나와 사이트[14]와 같은 웹 사이트를 이용하여 본 논문에서 제안한 방법을 이용하여 실험하는 것이 필요하다. 또한 온톨로지 인스턴스 생성율을 높이기 위해서 각 정보에 알맞은 정형화된 개념 온톨로지의 구축에 대한 연구가 필요할 것으로 사료된다.

참고문헌

[1] 한성국, 정영식, 유재규, "웹2.0과 시맨틱 웹, 그리고 진화의 방향," 정보과학회지, 제 25권, 제 10호, 57-66쪽, 2007년 10월.

[2] Berners-Lee, T.,Hendler, J. and Lassila, O., "The Semantic Web," Scientific American, 2001.5

[3] C. Patel, K. Supekar, and Y. Lee. "OntoGenie: Extracting Ontology Instances from WWW," Human Language Technology for the Semantic Web and Web Services, ISWC, 2003.

[4] Latifur Kahn, Feng Luo, "Ontology Construction for Information Selection," Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence, pp.122-127, 2002,

[5] J. Euzenat, "Eight questions about Semantic Web annotations," IEEE Inteligent Systems, Vol. 17, No. 2, pp.46-53, 2001.

[6] 선복근, 위다현, 한광록, "OWL 온톨로지를 기반으로 하는 논문 검색시스템에 관한 연구," 컴퓨터정보학회논문지, 제 14권, 제 2호, 169-180쪽, 2008년, 2월.

[7] 정한민, 이미경, 성원경, "시맨틱웹 2.0 기술동향," 정보통신연구진흥원 주간기술동향 통권 1344호, 15-28쪽, 2008년. 4월.

[8] Popov Borislav, A.K. Angel Kirilov, Manov Dimitar, Ognyanoff Damyan, Miroslav, "Kim-Semantic Annotation Platform," 2nd International Semantic Web Conference(ISWC2003), Vol. 2870. pp.834-839, Springer, Verlag Berlin Heidelberg, 2003.

[9] Vargas-Vera Maria, Motta Enrico, Domingue John, Lanzoni Mattia, Stutt Arthur and Ciravegna Fabio, "MnM: Ontology driven Semi-Automatic and Automatic Support for Semantic Markup," The 13th International Conference on Knowledge Engineering and mamagement(EKAW 2002), ed Gomez-Perez, A., Springer Verlag, 2002.

[10] H. Zheng, B. Kang, S. Koo, H. Choi, K. Kim, and H. Kim, "A semantic Annotation Tool to Extarct Instances from Korean Web Documents," In Proceedings of Semantic Authoring and Annotation workshop of ISWC, 2006.

[11] 이미경, 정한민, 성원경, "히스토리 기반 온톨로지 인스턴스 도구 관리," 제19회 한글 및 한국어 정보처리 학술대회, 2007.

[12] 김재호, 신지애, 최기선, "국가 IT 온톨로지 구축," 2006년도 한국정보과학회 가을 학술발표논문집, 제 33, No.2(B), 16-19쪽, 2006년.

[13] 박재훈, 유재규, 전양승, 정영식, 한성국, "온톨로지 기반의 시맨틱 어노테이터 구현," 프로그래밍언어 논문지 제 2권, 제 2호, 17-23쪽, 2006년 11월.

[14] 장문수, 강선미, "도메인지식의 계층화를 통한 온톨로지 인스턴스의 속성정보 추출," 퍼지 및 지능시스템학회 논문지, 제 17권, 제 3호, 291-296쪽, 2007년 6월.

저 자 소 개

한 광 록

1984년 : 인하대학교 전자공학과 공학사
1989년 : 인하대학교 정보공학전공공학박사
1991년 ~ 현재 : 호서대학교 컴퓨터공학부 교수
관심분야 : 정보검색, HCI, e-Health, 시맨틱웹

강 현 민

2008년 : 호서대학교 게임공학과 공학사
2010년 : 호서대학교 메카트로닉스공학과 공학석사
2010년 ~ 현재 : HITS 연구원
관심분야 : HCI, 시맨틱웹

손 석 원

1985년 : 인하대학교 전자공학과 공학사
2007년 : 인하대학교 컴퓨터정보공학과 공학박사
1999년 ~ 현재 : 호서대학교 뉴미디어학과 부교수
관심분야 : 인공지능,무선통신망, e-Health