웹사이트 관리를 위한 RDF 메타데이타 생성시스템

(The RDF Metadata Generation System for WebSite Management)

이 미 경 † 하 얀 †† 김 용 성 †††

(Mi Kyung Lee) (Yan Ha) (Yong Sung Kim)

요 약 웹 자원을 효율적으로 관리하기 위하여 메타데이타(metadata)의 활용이 증가하고 있으며, 이 활용 분야 중 웹사이트 관리를 위한 응용들은 표준화되지 않아서 상호호환성의 문제를 가지고 있다. 따라서 본 논문에서는 메타데이타의 표준화와 상호운영을 목적으로 하는 RDF(Resource Description Framework)를 이용하여 웹사이트 관리를 위한 메타데이타 생성 시스템을 제안한다. 제안된 시스템은 웹사이트를 기관 도메인별로 자동 분류하고, 정보를 구조적 형식으로 기술하여 효율적인 검색 환경을 제시할 수 있다. 이를 위해 더블린 코어를 기반으로 한 메타데이타 모델 및 RDF 메타데이타를 자동 생성하기 위한 사상 규칙과 알고리즘을 제시하고 시스템을 구현한다.

Abstract The practical use of metadata increases for the efficient management of web resources. But it is the problem that, among the application fields of metadata, applications for the website management are not exchanged mutually because they are not standardized. Therefore this paper suggests metadata generation system for the management of Website on the basis of Resource Description Framework(RDF) proposed with the purpose of standardization and interoperability of metadata. The suggested system automatically classifies websites into each domain and offers the efficient retrieval environment by describing the information to structural format. We propose the metadata model based on the Dublin Core, the rules and an algorithm for generating RDF metadata, and then implement a system.

1. 서 론

최근 웹의 발전으로 네트워크 상에서 접근할 수 있는 자원들이 증가하고 다양화되므로, 웹 자원을 정확하고 신속하게 검색하기 위하여 메타데이타의 필요성이 증가하고 있다[1]. 즉, 웹 자원에 대한 메타데이타는 웹 자원의 내용, 구조, 특징을 기술한 정보로 자원에 대한 이해를 높이고 유용성을 판단할 수 있는 기준을 제공한다. 따라서 자원 생성자가 메타데이타를 직접 기술하고, 웹 문서 내에 메타데이타를 포함함으로서 자원 검색 시간과 노력이 경감되고 신뢰성이 향상된다. 또한 자원 발견, 웹 문서 수집, 내용 평가, 전자상거래, 디지털 서명, 지적재산권 명시, 웹사이트 관리 등 여러 분야에서 응용되고 있다[2].

이러한 메타데이타 응용 중 웹사이트를 수집하고 관리하기 위한 영역에는 HTML 메타데이타와 XML(eXtensible Markup Language)을 이용한 CDF(Channel Definition Format), MCF(Meta Content Framework), RDF(Resource description Framework)가 있다.

HTML 메타데이타는 확장성, 구조성, 데이타 검색 기능의 단점이 드러났고 이 문제들을 해결하기 위하여 단순하고 구조 검색 및 전문 검색이 가능한 XML이라는 새로운 마크업 언어가 개발되었으며, 이를 기반으로 하는 CDF와 MCF는 푸시 기술만을 지원하고 표준화되지 않았다. 따라서 메타데이타 관리를 위한 표준안인

† 정 회 원 : 서울정수기능대학 정보통신설비과 교수
mklee@sjpc.ac.kr
†† 정 회 원 : 경인여자대학 멀티미디어정보전산학부 교수
yanha@hanmail.com
††† 종신회원 : 전북대학교 컴퓨터과학과 교수
yskim@moak.chonbuk.ac.kr
논문접수 : 2000년 8월 24일
심사완료 : 2001년 2월 2일

RDF가 제안되었고 이는 웹 상에서 컴퓨터가 이해하고 처리할 수 있는 메타데이타를 표현하기 위한 언어로 웹 자원의 자동화에 초점을 두고 있다[2].

웹사이트 관리를 위하여 기존에는 개별적으로 웹사이트를 접근하고 분석하지 않고서는 정보를 알 수가 없고, 정보를 기술하기 위한 메커니즘이 표준화되지 않았다. 이러한 문제점을 해결하기 위하여 메타데이타는 사용자가 원하는 웹사이트를 검색하는데 신속하고 정확한 의사결정을 내릴 수 있게 하고, RDF는 웹사이트 메타데이타를 기술하기 위한 표준화된 프레임워크를 제공하여 데이타베이스 구축 및 질의 검색이 용이하게 된다. 메타데이타를 자동으로 기술하는 것은 어려운 일이지만 수동으로 기술하는 것은 거의 불가능하다. 따라서 메타데이타를 자동 생성하기 위한 도구로 편집기와 생성기가 개발되어 있고 RDF 형식으로 기술하는 연구가 진행되고 있다[3][4][5].

따라서 본 논문에서는 웹사이트 관리를 위한 메타데이타를 RDF를 이용하여 자동 생성하고 메타베이스를 구축하는 시스템을 제안한다. 웹사이트를 기관 도메인별로 관리하기 위하여 URL (Uniform Resource Locator)의 서브도메인을 이용하여 기관별로 범주화하고 RDF 구문을 이용하여 웹사이트 정보를 구조적 형식으로 기술함으로써 효율적인 검색 환경의 제공과 웹 자동화 처리를 지원하는데 목적을 둔다. 이를 위해 메타데이타 모델은 표준화된 더블린 코어 요소 중에서 선택하여 제안하고, HTML로 표현된 웹사이트 정보를 RDF 데이타 모델링 기법과 제안된 메타데이타 요소를 이용하여 RDF 메타데이타로 자동 생성하기 위한 사상규칙과 알고리즘을 제시하고 시스템을 구현한다.

본 논문의 구성은 다음과 같다. 2장에서는 웹사이트를 수집하고 관리하기 위한 메타데이타 응용 영역과 메타데이타 생성기에 대한 관련 연구를 알아보고, 3장에서는 웹 자원 메타데이타 모델인 더블린코어와 RDF에 대해 기술한다. 4장에서는 웹사이트 관리를 위하여 더블린 코어를 기반으로 하는 메타데이타 모델과 RDF 메타데이타를 생성하기 위한 사상규칙과 알고리즘을 제안한다. 5장에서는 웹사이트 관리를 위한 RDF 메타데이타를 생성하는 시스템을 설계하고 구현한다. 6장에서는 결론 및 향후 연구과제를 제시한다.

2. 관련연구

웹사이트 관리를 위한 메타데이타 연구로는 HTML 메타데이타, XML을 기반으로 하는 CDF, MCF, RDF가 있다. HTML 메타데이타[6]는 <META>와 <LINK> 요소를 이용하여 기술한다. <META> 요소는 HTML 문서에 메타데이타를 포함하는 데 사용되며, 속성을 기술하고 속성 값을 할당한다. <LINK> 요소는 문서 외부의 다른 메타데이타 스키마를 참조하는데 사용한다. HTML 메타데이타는 사용자들이 확장된 메타데이타를 제공할 수 없고 정확한 검색이 불가능하며, 복잡한 문서 관계를 표현하는데 적합하지 않다.

CDF[7]는 채널을 기본 개념으로 하고 XML을 기반으로 하여 웹 페이지들의 정보를 구성하여 제공하는 것으로 마이크로소프트가 W3C(World Wide Web Consortium)에 표준으로 제안하였으며 공개되어 있다.

<CHANNEL>은 <ITEM> 요소들로 구성되어 있으며 제목(title), 설명(abstract), 저자(author), 최종수정일(last modification date) 정보에 의해 웹페이지를 기술한다. 단지 푸시를 지원하는 방식이라는 제한점을 가진다. MCF[8]는 CDF에 맞서서 넷스케이프사에서 제안하였고 웹페이지, 이미지, 웹사이트에 대한 정보를 제공하는 것이다. 웹페이지는 크기(size), URL, 저자(author)에 의해 기술한다. MCF는 다양한 목적을 가지는 메타 정보를 처리할 수 있는 파일을 만들 수 있지만, 기존의 메타 태그를 사용하는 것보다 복잡하다는 단점이 있다. 그리고 CDF와 MCF는 상호 호환이 되지 않는 문제점이 있고 표준안으로 채택되지 않았다. 이에 반해 RDF[2]는 웹 자원들을 통합하고 관리하는데 관련된 기술로 웹사이트의 정보를 조직화하고 기술하기 위하여 XML을 기반으로 하며 MCF를 기초로 하고 있다. RDF는 표준화된 메타데이타에 대한 일관된 표현 및 교환 처리가 가능한 상호운영의 특징을 가지므로 CDF, MCF 같은 응용들도 RDF를 이용하여 기술되어야 한다.

또한 메타데이타 생성도구는 메타데이타 작성자가 편집기에 의해 수작업으로 데이타 요소를 입력하고 특정 형식으로 메타데이타를 생성하여 데이타의 최신성과 일관성을 유지하는 도구와 생성기에 의하여 입력된 URL의 페이지를 분석하고 데이타를 자동으로 추출하여 메타데이타를 자동 생성하는 도구들이 연구·개발되고 있다. 먼저 국외의 생성도구에 대해 살펴보면 다음과 같다. Reggie[3]는 메타데이타의 스키마와 기술 언어, URL을 입력하여 메타데이타를 생성하고, 더블린 코어의 한정어를 데이타요소에 사용할 수 있으며 각종 도움말을 제공한다. 변환 형식은 HTML v3.2, v4.0, RDF/XML이고 WWW/Java로 구현되었으며, Netscape v4.0 또는 Internet Explorer v4.0 환경을 필요로 하는 편집기이다. Dublin Core Metadata template[4]는 더블린 코어의 모든 데이타 요소와 한정어를 사용하며, 각종 도

움말을 제공하여 데이타 변환을 지원하고, 최소 수준의 메타데이타를 생성할 수 있다. 변환 형식은 HTML v3.2와 v4.0이고 WWW/Perl로 구현되었다. DC-dot[5]은 메타데이타 생성기로 자원의 URL을 입력하면 해당 웹 페이지를 검색하여 더블린 코어의 메타태그로 자동 생성하는 도구로서, 더블린 코어의 기본 데이타요소만을 사용하고 한정어를 사용하지 않는다. 필요한 경우 데이타요소에 관련 사항을 입력하여 메타데이타를 재생성할 수 있다. 변환 형식은 HTML과 XML/RDF이고 WWW/Perl로 구현되었다. 국내 관련 연구로는 SeriCore 메타데이타 편집기[9]가 있으며, SeriCore 메타데이타 모델은 과학 기술 분야와 관련된 논문, 보고서, 기술문서와 이미지를 위한 모델이다. 이 편집기는 SeriCore 메타데이타의 정보를 수동으로 입력하여 생성하고, Visual C++ 4.2와 Visual Basic 4.0의 OCX를 이용하여 구현되었고 SGML을 기반으로 메타데이타를 생성한다. 기존 생성 도구들은 대부분 더블린 코어의 모든 요소를 HTML과 SGML형식과 상이한 메타데이타간의 표준화를 위하여 RDF 형식으로 메타데이타를 기술하고 있다. 따라서 본 연구에서는 웹사이트의 효율적인 관리를 위하여 웹사이트 메타데이타 정보를 RDF 형식으로 자동 생성하고 메타베이스 구축과 검색기능을 추가한 시스템을 구현한다.

3. 더블린 코어와 RDF

본 장에서는 웹 자원 메타데이타 모델인 더블린 코어와 웹 자원 메타데이타를 기술하기 위한 RDF에 대해 살펴본다.

3.1 더블린 코어

웹 상에서 자원의 유형과 접근방법이 다양해지고 메타데이타마다 요소를 달리하고 있다. 이러한 문제를 해결하기 위하여 1995년 3월 OCLC(Online Computer Library Center)와 NCSA (National Center for Supercomputer Applications)가 더블린에서 개최된 워크숍에서 웹 자원 메타데이타의 형식으로 더블린 코어를 합의하였다. 데이타의 호환성을 유지하고 네트워크 자원을 기술하고 접근하는데 필요한 15개의 데이타 요소를 기본요소로 결정하였다[10]. 더블린 코어의 데이타 요소는 정보 저장을 위하여 자원 내용 관련 요소, 지적재산권 관련 요소, 자원 설명 관련 요소의 3개 그룹으로 분류할 수 있다[11].

자원 내용과 관련된 요소는 Title, Subject, Description, Type, Source, Relation, Coverage등이 있고, 지적재산권과 관련된 요소는 Creator, Publisher, Contributor, Rights가 있으며, 자원의 설명과 관련된 요소로는 Date, Format, Identifier, Language가 있다. 웹 자원 검색을 용이하게 하는 메타데이타인 더블린 코어는 데이타 요소의 구조가 간단하고, 사용이 용이하며, 상호운영성을 확보할 수 있고, 유연성과 확장성을 포함하며 자원의 접근성이 높다는 특징을 가지고 있다[10].

3.2 RDF(Resource Description Framework)

RDF는 XML을 기반으로 웹 상의 분산된 다양한 자원들을 기술하기 위한 구조이고, 웹 상의 자원을 대상으로 메타데이타의 표준화 작업과 효율적이고 체계적인 관리를 위하여 W3C에서 제안하였다. RDF는 웹 자원 검색을 위한 상이한 메타데이타를 효율적으로 교환하고 공유할 수 있는 상호운영을 목적으로 한다. 또한 RDF는 웹 자원의 자동화 처리와 컴퓨터가 이해할 수 있는 정보 교환 수단을 제공한다[2].

W3C는 RDF 모델 및 구문을 1999년 2월에 권고안(Recommendation)으로 제정하였고, RDF 스키마를 2000년 4월에 후보 권고안(Candidate Recommendation)으로 제정하였다.

다음은 RDF 모델 및 구문과 스키마에 대해 개략적으로 기술한다.

3.2.1 RDF 데이타 모델과 구문

RDF 데이타 모델[2]은 메타데이타를 정의하고 사용하기 위한 추상적이고 개념적인 구조를 의미하며, 자원과 관련된 속성유형과 속성값들은 노드-아크 다이어그램(nodes-arcs diagram)과 3-튜플(tuples) 형식을 사용하여 표현한다. 자원은 URI(Uniform Resource Identifier)로 구별되는 모든 웹 객체로서 웹사이트, 웹페이지, 웹페이지의 일부분 등이 될 수 있다. 속성은 자원을 기술하기 위한 특징, 관계들을 나타내고, 각각 특정한 의미와 허용되는 값, 기술할 수 있는 자원의 유형 등을 가지고 있다. 이러한 속성의 특징들은 RDF 스키마에서 표현된다. 속성값은 다른 자원 또는 문자열, 숫자와 같은 원자값일 수 있다. 문장은 자원과 속성, 속성값으로 이루어진 특수한 자원을 의미하고, 문장은 자원을 의미하는 주어(subject), 속성명을 의미하는 술어(predicate), 속성값을 의미하는 목적어(object)로 이루어진다.

RDF 데이타 모델에서 자원은 노드로, 속성은 아크로, 속성값이 원자값인 경우는 사각형으로, 자원인 경우는 노드로 표현한다. RDF의 데이타 모델의 예는 그림 1과 같다.

RDF 구문[2]은 RDF 데이타 모델을 생성하고 교환하기 위한 목적으로 인간과 기계가 읽을 수 있고 처리할 수 있는 형태로 인코딩되는 형식을 말하며, RDF 스

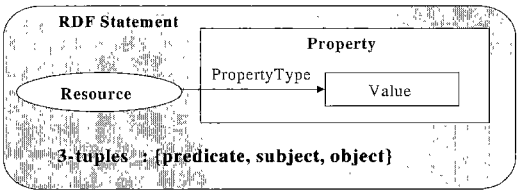

그림 1 RDF의 데이타 모델

키마에서 정의된 자원을 이용하여 생성된 RDF 데이타 모델을 XML 문법에 따라 인코딩한다. XML의 이름공간(namespace) 기능을 사용하여 속성유형과 속성유형이 정의된 스키마를 분명하게 연결할 수 있다. 이 속성의 의미를 분명히 정의함으로써 메타데이타를 일관되게 입력하고 교환할 수 있으며, 상이한 분야에서 정의된 독립된 메타데이타 패키지를 상호 교환할 수 있다.

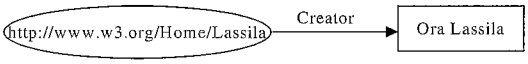

그림 2 노드-아크 다이어그램

그림 2를 3-튜플 형식으로 표현하면 {Creator, [http://www.w3.org/Home/Lassila], Ora Lassila)이고, RDF 구문으로 표현하면 다음과 같다.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:s="http://description.org/schema/">
    <rdf:Description about="http://www.w3.org/Home/Lassila">
        <s:Creator>Ora Lassila</s:Creator>
    </rdf:Description>
</rdf:RDF>
```

3.2.2 RDF 스키마(Schema)

RDF 스키마[12]는 자원들 사이의 속성과 관계를 포함한 정보 집합인 타입 시스템을 제공하고 RDF 문장들을 해석하기 위해 응용된다. 또한 특정 분야의 정보 자원에 대한 속성을 표현하는 요소들을 선언하기 위해 사용되고, 기계가 이해할 수 있는 요소들을 정형화함으로써 다른 메타데이타에서 사용된 요소들을 재사용하거나 교환할 수 있다. 다른 메타데이타 스키마를 구별하기 위하여 RDF의 이름공간 개념을 사용한다. RDF 스키마에 대한 자원과 클래스들의 집합과 요소들을 표현하면 그림 3과 같다. 그림 3에서 둥근 직사각형은 클래스를 나타내고, 큰 점들은 각 자원을, 화살표는 자원이 정의하는 클래스를 나타낸다. 그리고 서브클래스는 슈퍼클래스에 둘러싸여 있다. RDF 자원은 클래스, 프로퍼티, 제한조건, 설명부분으로 구성된다.

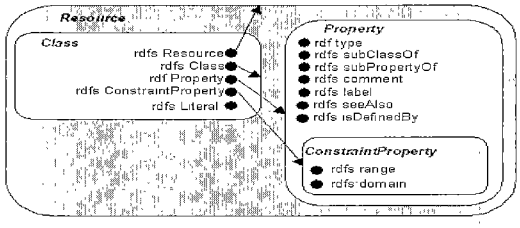

그림 3 자원과 클래스

4. RDF 메타데이타 생성

본 장에서는 더블린 코어를 기반으로 웹사이트 관리를 위한 메타데이타 모델을 제안하고, HTML로 기술된 웹사이트 정보를 RDF 메타데이타로 자동 생성하기 위한 사상규칙과 알고리즘을 기술한다.

4.1 웹사이트 관리를 위한 메타데이타 모델

웹사이트에 대한 정보를 메타데이타로 표현하면, 사용자는 메타데이타를 통하여 실제 데이타 즉, 웹사이트에 접근하게 됨으로써 메타데이타는 양질의 정보를 제공하여 사용자와 웹사이트를 연결하는 중개자 역할을 수행한다. 따라서 웹사이트를 효율적으로 관리하고, 쉽게 검색할 수 있는 메타데이타의 구축과 관리는 중요하다. 본 연구에서는 홈페이지, 논문, 기술보고서, 출판물 등 다양한 웹 자원에 대한 메타데이타 모델인 더블린 코어를 기반으로 웹사이트 관리에 필요한 요소를 선택하여 검색정보와 관리정보로 구분하고, 검색정보에 구분정보를 추가하여 제안한다. 구분정보는 URL의 도메인을 분석하여 추출한다. 제안된 메타데이타 요소는 웹사이트를 유일하게 정의할 수 있고, 대량의 웹사이트 정보를 체계적이고 구조적으로 관리하여 효율적인 검색 환경을 제공할 수 있다. 웹사이트 관리를 위한 메타데이타 모델의 요소는 다음 표 1과 같고, HTML 문서의 태그들을 분석하여 쉽게 추출할 수 있다.

표 1 웹사이트 관리를 위한 메타데이타 요소

구분	요소명	
검색정보	구분정보	Category
	제목정보	Title
	주제어정보	Keyword
	내용정보	Abstract
관리정보	위치정보	Identifier
	수정일정보	Date

(1) 검색정보

웹사이트에 관련된 내용을 기술한 정보로서, 메타데이타를 이용한 웹사이트 검색 시 유용하게 사용되는 요소이다.

- Category : 웹사이트의 범주를 할당해주는 구분정보로, 서브도메인의 기관 종류에서 추출하여 부여한다. 서브도메인을 분석하여 부속도메인의 edu와 ac는 Educational Facilities, gov와 go는 Government agencies, org와 or는 Public Institution, com과 co는 Company, net와 ne는 Network Enterprise, int는 International Organization으로 구분한다.
- Title : 웹사이트에 부여된 제목 정보로, HTML <TITLE> 태그에서 추출하고 주제어 정보의 생성에 도움을 준다.
- Keyword : 웹사이트의 내용이나 주제명으로 표현되는 키워드 정보로, Keyword <META> 태그에서 추출한다.
- Abstract : 웹사이트 내용과 관련된 요약을 기술한 정보로, Description <META> 태그에서 추출한다.

(2) 관리정보

효율적인 메타데이타 관리를 위하여 웹사이트의 존재유무, 변경 유무에 따라 메타데이타를 삭제하거나, 수정 또는 재생성하기 위해 사용되는 정보이다.

- Identifier : 웹사이트를 식별하기 위한 정보로, URL을 포함한다.
- Date : 웹사이트의 최종 수정일 정보로, HTTP Last-modified 헤더 또는 Date <META> 태그에서 추출한다.

4.2 사상 규칙

RDF 메타데이타를 기술하기 위하여 RDF 데이타 모델을 정의한 후 인코딩하며, RDF 메타데이타 모델은 노드-아크 다이어그램 또는 3-튜플 형식으로 표현한다. 따라서 HTML 문서를 RDF 데이타 모델로 모델링하기 위하여 다음과 같이 정의한다.

[정의] HTML 문서의 <HEAD> 부분을 RDF 메타데이타 생성을 위한 RDF 데이타 모델링 결과로 사상한다.

위 정의에 따라 HTML 문서의 태그들을 분석 추출하여 RDF 데이타 모델링을 위한 규칙은 다음과 같다.

[규칙 1] 웹사이트의 URL 주소는 자원(Resource)이 된다.

[규칙 2] 웹사이트 도메인의 기관 종류를 분석하여 웹사이트의 범주를 자동 부여한다. 범주는 서브도메인의 교육기관은 Educational Facilities, 정부기관은 Government agencies, 공공기관은 Public Institution, 회사는 Company, 네트워크 관련은 Network Enterprise, 국제기관은 International Organization으로 자동 부여된다.

(예 1) 메타데이타를 생성하기 위한 URL주소는 "http://www.stanford.edu/" 이다.

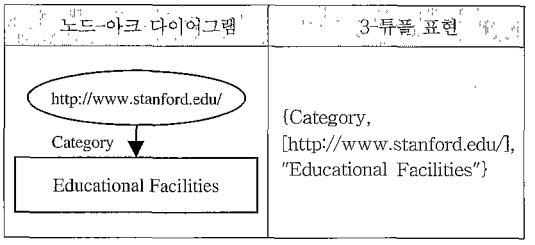

그림 4 (예 1)의 RDF 데이타 모델

4.2.1 TITLE 태그

TITLE 태그는 자원의 내용을 정의하여 자원에 대한 정보를 제공한다.

[규칙 3] TITLE 태그는 RDF 데이타 모델의 속성유형이 되고, 요소내용은 속성값이 된다.

(예 2) <title>Stanford Home Page: Welcome to Stanford University</title>

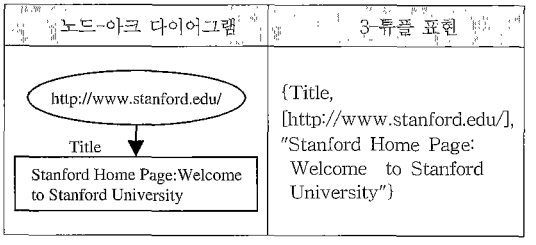

그림 5 (예 2)의 RDF 데이타 모델

4.2.2 META 태그

META 태그는 자원에 대한 정보를 정의하고 검색엔진, 로봇 등에 의해 탐색될 때 중요한 정보를 제공한다.

[규칙 4] META 태그의 name과 http-equive 속성은 RDF 데이타 모델의 속성유형이 되고, content 속성은 RDF 데이타 모델의 속성값이 된다.

(예 3) <META NAME="description" CONTENT= "The Stanford University home page is a good place to start your search of Stanford University's web resources and websites.">

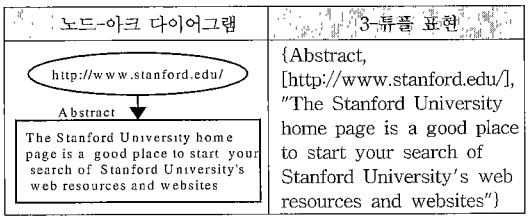

그림 6 (예 3)의 RDF 데이타 모델

[규칙5] META 태그의 content 속성값이 여러 개일 때는 RDF의 BAG 컨테이너(container) 특징을 부여한다.

(예 4) <META NAME="keywords" CONTENT= "Stanford, Stanford Websites, Stanford web, Stanford home page, university, college">

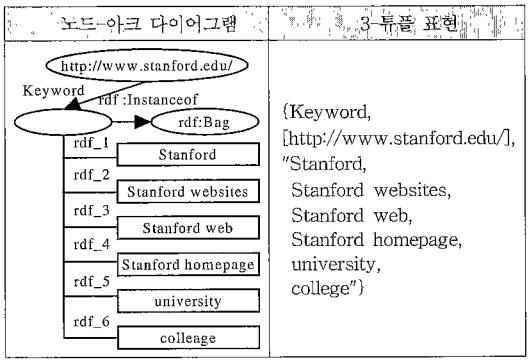

그림 7 (예 4)의 RDF 데이타 모델

다음은 제시한 HTML 문서의 태그에 대한 규칙을 적용하기 위한 예로서 스탠포드 대학교 홈페이지의 <HEAD> 부분이다.

```
<html>
 <head>
 <title>Stanford Home Page: Welcome to Stanford University</title>
 <META NAME="description" CONTENT="The Stanford University home page is a good place to start your search of Stanford University's web resources and websites.">
 <META NAME="keywords" CONTENT="Stanford, Stanford websites, Stanford web, Stanford homepage, university, college">
 <META name="provider" content="andyk@leland.stanford.edu">
 </head>
</html>
```

위 스탠포드 대학교 홈페이지에 대해 RDF 데이타 모델링 규칙을 적용한 RDF 데이타 모델은 다음 그림 8과 같다.

위 RDF 데이타 모델의 3-튜플 형식을 자원-속성 테이블 형태로 표현하면 다음 표 2와 같다.

4.3 RDF 메타데이타 생성

RDF 메타데이타는 위에서 생성된 자원-속성 테이블을 이용하여 자동 생성한다. RDF 메타데이타 생성 과정은 필수적인 선언부를 기술하고, RDF 모델링 결과를 규칙에 적용하여 자원과 속성에 대한 RDF 메타데이타를 생성한다.

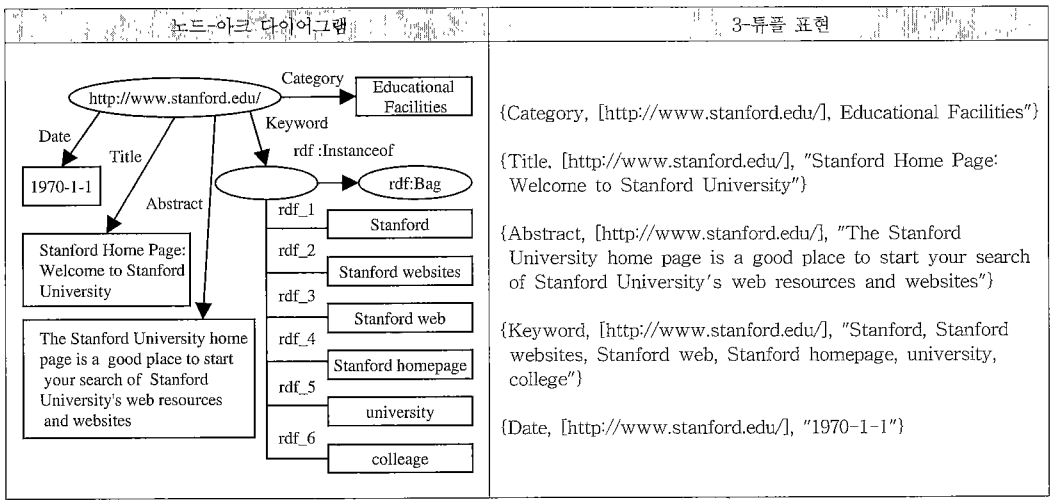

그림 8 스탠포드 대학교 홈페이지에 대한 RDF 데이타 모델

표 2 HTML 문서의 자원-속성 테이블

HTML 태그	자원	속성	
		속성유형	속성값
URL	http://www.stanford.edu/	Category	Education Facilities
TITLE		Title	Stanford Home Page:Welcome to Stanford University
META		Keyword	Stanford, Stanford websites, Stanford web, Stanford home page, university, college
		Abstract	The Stanford University homepage is a good place to start your search of Stanford University's web resources and websites.
URL		Identifier	http://www.stanford.edu/
META		Date	1970-1-1

4.3.1 RDF의 선언부 기술

선언부는 RDF 메타데이타 생성시 필요한 부분으로 다음은 RDF 메타데이타 기술에 필수적인 선언부의 표현이다. <rdf:RDF>는 RDF 기술 내용이 XML 문서 안에 정확하게 포함될 수 있는 역할을 수행하고, 이름공간과 관련된 URI는 스키마를 참조하는데 사용한다.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-RDF-syntax-ns#
    xmlns:prefix="URI-reference of metadata schema">

</rdf:RDF>
```

4.3.2 자원과 속성 기술

RDF 메타데이타 기술의 선언부가 처리된 상태에서 다음과 같은 규칙을 RDF 데이타 모델링 결과에 적용하여 자원과 속성에 대한 RDF 메타데이타를 생성한다.

[규칙 6] RDF 데이타 모델의 자원을 표시하고 설명하기 위해 <rdf:Description>를 사용한다. 다음과 같은 형식으로 자원은 URI 참조 부분에 기술된다.

```
<rdf:Description about="URI-reference">
</rdf:Description>
```

(예 5) 웹 자원의 URL이 "http://www.stanford.edu/"이고 더블린 코어를 이용한 경우 선언부와 자원을 기술하면 다음과 같다.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-RDF-syntax-ns#
    xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description about="http://www.stanford.edu">

  </rdf:Description>
</rdf:RDF>
```

[규칙 7] RDF 데이타 모델의 속성유형은 RDF 구문의 속성명으로, 속성값은 요소내용으로 대응되어 다음과 같은 형식으로 기술된다.

```
<prefix:propName> value </prefix:propName>
```

(예 6) RDF 데이타 모델 {Title, [http://www.stanford.edu/], "Stanford Home Page: Welcome to Stanford University"}을 RDF 구문으로 기술하면 다음과 같다.

```
<dc.Title > Stanford Home Page: Welcome to Stanford University </dc.Title>
```

[규칙 8] RDF 데이타 모델의 속성값이 여러 개일 때는 rdf:Bag 컨테이너 구문을 이용하여 다음과 같은 형식으로 기술한다.

```
<prefix:propName>
    <rdf:Bag>
        <rdf:li> value 1 </rdf:li>
        <rdf:li> value 2 </rdf:li>
    </rdf:Bag>
</prefix:propName>
```

(예 7) RDF 데이타 모델 {Keyword, [http://www.stanford.edu/], "stanford, stanford websites, stanford web, university"}을 RDF 구문으로 기술하면 다음과 같다.

```
<dc:Keyword>
    <rdf:Bag>
        <rdf:li> stanford </rdf:li>
        <rdf:li> stanford websites </rdf:li>
        <rdf:li> stanford web </rdf:li>
        <rdf:li> university </rdf:li>
    </rdf:Bag>
</dc:Keyword>
```

다음은 4.2절에서 생성된 스탠포드 대학교 홈페이지의 RDF 데이타 모델을 RDF 구문 생성 규칙에 적용한 결과이다.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:ws="http://210.119.188.47/ws/schema.html/">
 <rdf:Description about="http://www.stanford.edu">
  <ws:Category> Educational Facilities</ws:Category>
  <ws:Title> Stanford Home Page: Welcome to
                    Stanford University </ws:Title>
  <ws:Keyword>
    <rdf:Bag>
      <rdf:li> Stanford </rdf:li>
      <rdf:li> Stanford websites </rdf:li>
      <rdf:li> Stanford web </rdf:li>
      <rdf:li> Stanford homepage </rdf:li>
      <rdf:li> university </rdf:li>
      <rdf:li> college </rdf:li>
    </rdf:Bag>
  </ws:Keyword>
  <ws:Abstract> The Stanford University home page is a
    good place to start your search of Stanford University's
    web resources and websites. </ws:Abstract>
  <ws:Identifier>http://www.stanford.edu/</ws:Identifier>
  <ws:Date> 1970-1-1 </ws:Date>
 </rdf:Description>
</rdf:RDF>
```

또한, 생성된 RDF 메타데이타는 3-튜플 형식으로 관계형 데이타베이스에 효율적으로 저장되고, 필드검색이 가능하여 정확률이 향상된다. 메타데이타 정보를 테이블 형태로 자원은 레코드, 속성유형은 필드명, 속성값은 필드값으로 사상시켜 다음 그림 9와 같은 인덱스 구조로 저장한다. 인덱스 구조에서 Category Name은 웹사이트의 범주를 나타내고 Category ID는 웹사이트 범주 구분번호이며, Website Data는 웹사이트에 대한 정보를 나타낸다. 따라서 구조적 인덱스 관리가 가능하고 저장된 인덱스 정보는 웹사이트 검색에서 질의를 이용하여 검색되고, 검색 결과에 대한 정보를 제공한다.

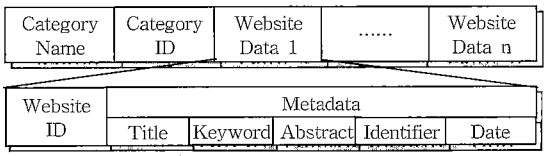

그림 9 인덱스 저장 구조

웹사이트 검색은 메타데이타에 의한 검색과 색인어에 의한 검색을 수용하여 사용자가 원하는 검색 의도를 효과적으로 표현하고, 검색의 신뢰도 및 효율을 높일 수 있다. 검색은 작성된 메타데이타 필드 중에서 사용자가 Category와 Keyword 필드의 내용을 입력하고, 이 메타데이타 필드들은 불리안 조건식으로 결합되어 검색된다.

4.4 알고리즘

다음은 4.2절에서 정의된 사상 규칙에 따라 HTML 문서의 META 태그를 RDF 데이타 모델로 사상시켜 생성된 자원-속성 테이블로부터 4.3절에 정의된 생성 규칙에 따라 RDF 메타데이타를 생성하고 저장, 검색하는 알고리즘이다.

```
public class WebSite extends JApplet {
  Connection connection = null;

  // 자바 애플릿이 웹브라우저에서 시작 시 실행
  public void init() {
    Class.forName("com.inet.tds.TdsDriver").newInstance();
    // JDBC Driver Type 4 클래스 로딩
    connection = DriverManager.getConnection(address,login,password);
    // 데이타베이스 서버 연결
  }

  // 자바 애플릿이 웹브라우저에서 종료 시 실행
  public void destroy() {
    connection.close(); // 데이타베이스 서버 연결 해제
  }

  // RDF 메타데이타 생성과 메타베이스 구축
  public void runCreateRDF() {
    URL getAdd = new URL(url.getText()); // 자원 생성
    URLConnection conn = getAdd.openConnection();
    // 자원과 연결 설정
    Categorycategory = Generating_Category(getAdd.get DoaminName()
    ); // Category 속성 생성
    Date date = conn.getLastModifiedDate(); // Date 속성 생성
    Identifier identifier = getAdd; // Identifier 속성 생성
    HtmlTitleTag title = conn.getHtmlTitleTag(); // Title 속성 생성
    HtmlMetaTag keyword = conn.getHtmlMetaTag(); // Keyword 속
    성 생성
    HtmlMetaTag abstract = conn.getHtmlMetaTag();  // Abstract 속
    성 생성
    RDF rdf = Generating_RDFSyntax(category, date, identifier, title,
    keyword, abstract); // RDF 메타 데이타 생성
    Save_DataBase(rdf); // 데이타베이스 저장
  }

  // 사용자가 입력한 내용에 대한 질의 생성과 데이타베이스 검색
  public void runMetaSearch() {
    Statement st = connection.createStatement();
    String query = "SELECT DISTINCT Identifier, Title, Abstract,
    Date FROM RDFTable ";
    query += " WHERE ("Keyword LIKE '%"+keyword.get
    Text().trim()+"%' ";
    query += " OR Abstract LIKE '%"+keyword.getText().trim()+"%' ";
    query += " OR Title LIKE '%"+keyword.getText().trim()+"%' )";
```

```
    query += " AND Category = '"+category.getSelectedItem()+"'";
    query += " ORDER BY Title";
    ResultSet rs = st.executeQuery(query); // 질의 생성
    write_searchResult(rs);  // 검색 결과 출력
  }
}
```

5. 시스템 설계 및 구현

본 장에서는 웹사이트 관리를 위한 RDF 메타데이타 생성 시스템을 구성하는 각 모듈의 구성과 기능을 알아보고, 구현 결과에 대해 살펴본다.

5.1 시스템 구성

RDF 메타데이타 생성 시스템은 메타데이타 생성 부분, 생성된 메타데이타의 데이타베이스 저장 부분, 사용자인터페이스를 통한 검색 부분으로 구성된다. RDF 메타데이타 생성 시스템 구성도는 그림 10과 같다.

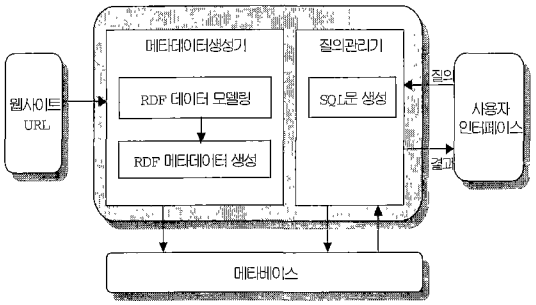

그림 10 웹사이트 관리를 위한 RDF 메타데이타 생성 시스템

5.1.1 메타데이타생성기

웹사이트 URL을 입력받아 추출된 HTML 문서 <HEAD> 부분의 태그들을 분석하고 RDF 데이타 모델링 규칙을 적용하여 3-튜플 형식으로 모델링한다. 입력된 URL은 자원으로 사상되고 META 태그의 name 속성은 속성유형으로 content 속성은 속성값으로 사상시켜 3-튜플 데이타 모델을 자원-속성 테이블로 생성한다. 필요한 경우 추가 정보를 입력하여 자원-속성 테이블을 수정할 수 있다. 그리고 RDF 메타데이타를 기술하기 위하여 RDF 구문의 선언부를 자동 생성하고, 자원-속성 테이블을 이용하여 연속구문과 축약구문으로 4.3절에서 정의한 규칙에 따라 RDF 메타데이타를 자동 생성한다.

5.1.2 메타베이스

생성된 RDF 메타데이타는 필드 구조 형식으로 속성유형은 필드명으로, 속성값은 필드값으로 사상되고, MS-SQL 7.0을 이용하여 데이타베이스에 저장하고 메타베이스를 구축한다.

5.1.3 질의관리기

사용자인터페이스에서 사용자가 입력한 범주와 키워드 값을 이용하여 적합한 질의문을 만들고 검색 결과를 보여준다.

5.2 시스템 구현

본 시스템의 구현 환경은 IBM-PC 호환기종에서 MS사의 Windows NT 기반 하에 자바 2 SDK 1.2.2로 구현하였고 웹 브라우저 Internet Explorer 5.0을 통하여 실행하였다. 구현된 RDF 메타데이타 생성 시스템은 RDF 메타데이타 생성기와 웹사이트 검색기로 구성된다.

다음은 스탠포드 대학교의 홈페이지를 시스템에 적용시킨 결과이다. RDF 메타데이타 생성기에 URL이 입력되어 GET 버튼을 클릭하면 HTML 문서를 읽어들여 RDF 모델링 결과인 3-튜플 형식이 자원-속성 테이블로 생성된다. URL의 서브도메인이 "edu"이므로 Category는 Educational Facilities로 자동 부여된다. Title은 <TITLE> 태그, Keyword와 Abstract는 <META> 태그를 분석하여 부여되며, Identifier는 URL이 부여되고 Last Modified Date는 웹페이지의 최종 수정일이 자동 부여되거나 1970-1-1로 부여된다. 생성 결과는 다음 그림 11과 같다.

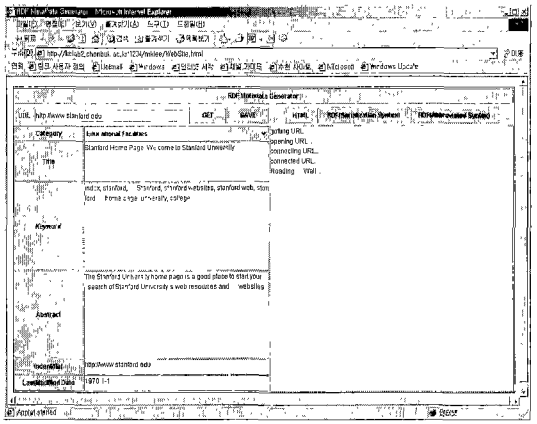

그림 11 자원-속성 테이블 생성 화면

생성된 자원-속성 테이블을 이용하여 4.3절에서 기술한 규칙에 따라 RDF 메타데이타를 연속구문과 축약구문으로 생성한다. 연속구문으로 RDF 메타데이타를 생성한 화면은 다음 그림 12와 같다.

그림 12 RDF 메타데이타 생성 화면

그림 12의 RDF 메타데이타 생성화면에서 SAVE 버튼을 클릭하면 RDF 메타데이타는 필드구조 형식으로 데이타베이스에 저장되고 결과 화면은 그림 13과 같다.

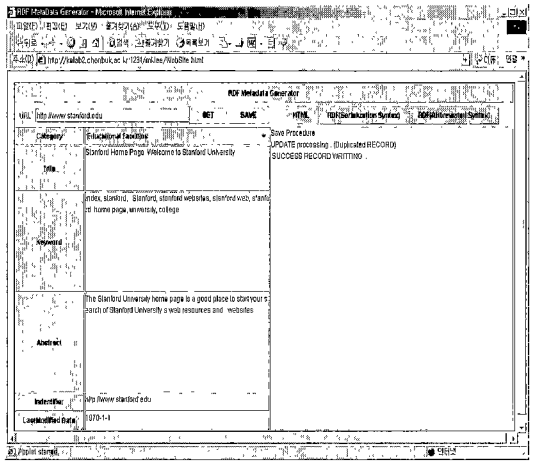

그림 13 생성된 RDF 메타데이타의 저장 화면

구축된 메타베이스는 inet사의 JDBC 드라이버를 이용하여 접근하고 사용자 인터페이스를 통하여 사용자가 질의를 입력하면 질의 관리기에서 적합한 질의문을 만들고, 질의에 적합한 결과를 사용자에게 보여준다. 사용자가 사용자인터페이스에서 구분 정보(Category)의 교육기관(Education Facilities)을 선택하고 "university"라는 질의를 입력하였을 때 사용자의 질의에 대한 검색 결과는 다음 그림 14와 같다.

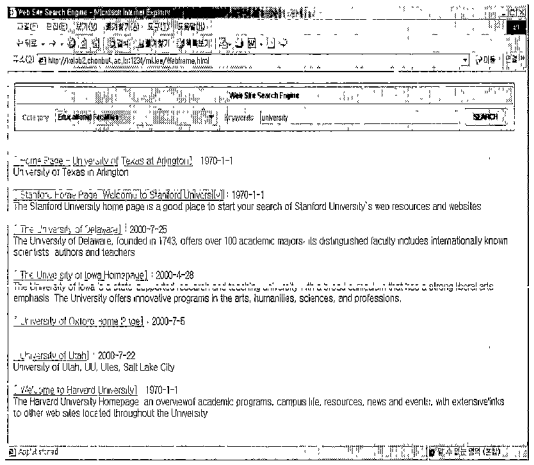

그림 14 사용자 질의와 검색 결과 화면

다음 그림 15는 그림 14의 검색 결과 중 스탠포드 대학교를 선택하여 홈페이지로 연결된 결과이다.

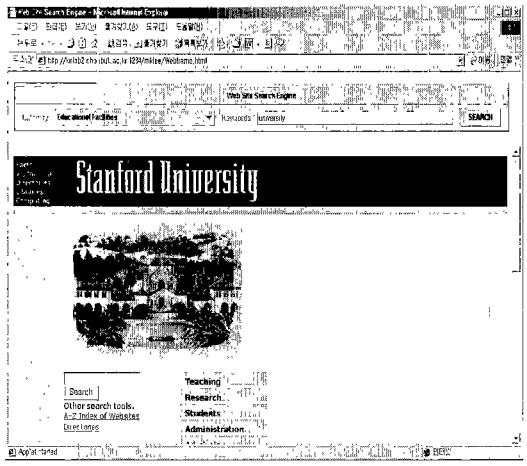

그림 15 검색 결과에서 스탠포드 대학교를 선택한 화면

5.3 비교 분석

본 연구의 RDF 메타데이타 생성 시스템의 평가를 위해 메타데이타 생성도구[3][4][5]들과 비교, 분석해 본다. 비교 결과는 다음 표 3과 같다.

표 3에서 보면, Reggie[3], Nordic Template[4]는 더블린 코어 요소와 한정어를 메타데이타 요소로 사용하며, 입력된 URL의 메타데이타 정보를 자동 변환하지 못하고 정보를 수작업으로 입력하는 편집기이다. DCdot[5]은 더블린 코어 요소만을 사용하며, 입력된 URL의

표 3 메타데이타 생성 시스템의 비교

	Reggie[3]	Nordic Template[4]	DC-dot[5]	본 연구
유 형	편집기	편집기	생성기	생성기
메타데이타 요소	더블린코어 데이타요소/ 한정어	더블린코어 데이타요소/ 한정어	더블린코어 데이타요소	제안된 데이타요소
온라인 편집	○	○	○	○
URL 입력변환	×	×	○	○
변환형식	HTML, RDF, RDF Abbr	HTML	HTML, RDF	HTML, RDF, RDF Abbr
RDF 컨테이너 속성 적용	×	×	×	○
DB구축	×	×	×	○
검색기능	×	×	×	○
프로그램	www/Java	www/Perl	www/Perl	www/Java

메타데이타 정보를 자동 변환하는 생성기이다. 이에 비해 본 연구의 시스템은 더블린 코어 요소 중 웹사이트 관리에 필요한 요소만을 선택하여 RDF 메타데이타를 자동 생성하고, 입력과 출력이 동일한 화면에서 이루어진다. 또한 데이타베이스 구축 기능과 검색기능을 추가하여 RDF의 특징이 충분히 반영된 생성기/검색기이다.

6. 결 론

본 논문에서는 HTML로 기술된 웹사이트 정보를 RDF 데이타 모델링 기법을 이용하여 RDF 메타데이타로 자동 생성하고 필요한 경우 추가 정보를 입력하여 메타데이타를 재생성할 수 있는 새로운 시스템을 제안하였다. 본 시스템에서는 더블린 코어 요소 중에서 웹사이트 관리에 필요한 요소를 선택 제안하고 범주를 기술하는 요소를 추가하였다. 또한 HTML 메타데이타를 입력받아 제안된 메타데이타 요소에 사상시켜 RDF 메타데이타를 자동 생성하고 필드 구조로 메타베이스를 구축하였다. 따라서 웹사이트를 검색할 경우 웹사이트의 내용을 직접 확인하지 않고도 정확하게 원하는 사이트를 검색할 수 있는 환경을 제공하였다.

본 연구의 수행 결과를 통하여 개발된 RDF 메타데이타 생성 시스템의 특징은 다음과 같다. 첫째, 메타데이타 모델을 표준화된 더블린 코어를 기반으로 제안함으로서 유용성, 사용자 적합성 및 접근성과 같은 메타데이타의 역할을 충족시킬 수 있다. 둘째, 도메인을 분석하여 범주를 자동으로 부여하고 웹사이트에 대한 RDF 메타데이타를 자동 생성하며, 메타베이스를 구축함으로서 대량의 웹사이트를 효율적이고 체계적으로 저장 관리한다. 셋째, RDF 메타데이타 자동 생성은 거대해지는 웹 자원을 기술하는데 유용하고 RDF의 목적인 웹 자원의 자동처리가 가능하다. 넷째, 웹사이트 검색뿐만이 아닌 내용평가, 전자상거래와 같은 RDF 응용들에 폭 넓게 전개할 수 있다. 향후 연구과제로는 웹사이트 검색과 관리에 효율적인 메타데이타 모델 요소를 추가하고 메타베이스의 수정, 삭제를 위한 메타데이타 관리기를 구현하는 것이다.

참 고 문 헌

[1] 이미경, 하 얀, 김용성, "RDF 스키마에서 UML 클래스 다이어그램으로의 변환", 정보처리학회 논문지, 제7권 1호, pp. 29-40, 2000. 1.

[2] Lassila, Ora, and Ralph R. Swick., "Resource description framework(RDF) model and syntax specification," W3C Recommendation 24 February 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[3] Distributed Systems Technology Centre., "Reggie: the metadata editor," 1998. http://metadata.net/dstc

[4] Nordic Metadata Project., "Dublin Core Metadata Template(Creator)," January 1999. http://linnea. helsinki.fi/meta/index.html

[5] Andy Powell., UKOLN, University of Bath., "DC-dot(Dublin Core Generator)," 1999. http://www.ukoln.ac.uk/metadata/dcdot/

[6] Kunze, J., "RFC 2731: Encoding Dublin Core Metadata in HTML," December 1999. http://www.ietf.org/rfc/rfc2731.txt.

[7] Castedo Ellerman, "Channel Definition Format (CDF)," W3C Submission 09 March 1997. http://www.w3.org/TR/NOTE-CDFsubmit.html.

[8] R.V. Guha, "Meta Content Framework Using XML," W3C Submission 24 June 1997. http://www.w3.org/TR/NOTE-MCF-XML/.

[9] 정효택, 양영종, 김순용, 이상덕, 최윤철, "Web상의 전자문서를 위한 메타데이타 모델의 제안 및 관리 시스템의 개발", 정보처리학회 논문지, 제5권, 제4호, pp.924-940, 1998. 4.

[10] Dublin Core Metadata Initiative."Dublin Core Metadata Element Set Reference Description. Version 1.1.," 1999-07-02. http://purl.org/dc/ elements/1.1.

[11] J. Kunze, C. Lagoze, M. Wolf OCLC Online Computer Library Center,Inc.,"Dublin Core Metadata for Resource Discovery," September 1998. http://www.ietf.org/rfc/rfc2413.txt

[12] Brickley, Dan, and R.V. Guha. 2000. "Resource

Description Framework (RDF) Schema Specification 1.0," W3C Candidate Recommendation 27 March 2000.
http://www.w3.org/TR/rdf-schema.
[13] Miller, Eric & Renato Iannella. 1998. "Dublin Core Examples in RDF," 1998-03-06
http://www.dstc.edu.au/Research/Projects/rdf/dc-in-rdf-ex.html
[14] Miller, Eric. "An introduction to the resource description framework," 1998.
http://www.dlib.org/dlib/may98/miller/05miller.html.

이 미 경
1988년 전북대학교 전산통계학과(이학사). 1991년 전북대학교 전산통계학과(이학석사). 1999년 전북대학교 전산통계학과 박사수료. 1988년 ~ 1995년 전주영생여상고 교사. 1995년 ~ 현재 서울정수기능대학 정보통신설비과 조교수. 관심분야는 정보 검색, 지식 공학, 전자 상거래, 컴퓨터 교육

하 얀
정보과학회논문지 : 소프트웨어 및 응용
제 28 권 제 3 호 참조

김 용 성
정보과학회논문지 : 소프트웨어 및 응용
제 28 권 제 1 호 참조