http://dx.doi.org/10.5392/JKCA.2012.12.12.025

uLAMP: 자연어 처리를 위한 자원 통합 관리 플랫폼

uLAMP: Unified Linguistic Asset Management Platform for Natural Language Processing

엄정호, 신성호, 최성필, 정한민
한국과학기술정보연구원

Jung-Ho Um(jhum@kisti.re.kr), Sung-Ho Shin(maximus74@kisti.re.kr),
Sung-Pil Choi(spchoi@kisti.re.kr), Hanmin Jung(jhm@kisti.re.kr)

요약

최근 인터넷과 스마트폰 등과 같은 무선기기의 발달로 각 전문 분야별로 많은 언어 자원들이 인터넷 등에 활발히 공개되고 있다. 또한 이러한 정보들이 유용한지를 판별하기 위해 다양한 시스템이 개발되고 있다. 이러한 시스템을 구축하기 위해서는 데이터의 수집, 자연어 처리 등의 과정이 필요하다. 그렇지만 이러한 과정에 필요한 소프트웨어 및 데이터를 통합적으로 관리하는 시스템은 현재 미미한 상태이다. 이를 위해 본 논문에서는 이러한 과정에서 필요한 소프트웨어 및 데이터를 통합 관리하는 시스템인 uLAMP를 제안한다. 이 시스템은 경제적인 측면에서는 소프트웨어 및 데이터 자원의 중복 개발 또는 수집을 방지하여 비용을 절감할 수 있으며, 관리적인 측면에서는 소프트웨어 및 데이터 자원의 재활용성을 높일 수 있다. 아울러, 제안하는 uLAMP의 사용성 및 효용성 평가를 위해 사용자 설문을 진행하였으며, 이를 통해 데이터의 최신성과 사용자 편의성 측면에서 장점을 지니고 있음을 알 수 있었다.

■ 중심어 : | 자원통합관리 | 자연어 처리 | 콘텐츠 관리 시스템 |

Abstract

Due to the development of wireless devices such as smart-phone and internet, a lot of linguistic resources actively are opened in each area of expertise. Also, various systems using semantic web technologies are developing for determining whether such information are useful or not. In order to build these systems, the processes of data collection and natural language processing are necessary. But, there is few systems to use of integrating software and data required those processes. In this paper, we propose the system, uLAMP, integrating software and data related to natural language processing. In terms of economics, the cost can be reduced by preventing duplicated implementation and data collection. On the other hand, data and software usability are increasing in terms of management aspects. In addition, for the evaluation of uLAMP usability and effectiveness, user survey was conducted. Through this evaluation, the advantages of the currentness of data and the ease of use are found.

■ keyword : | Linguistic Asset Management | Natural Language Processing | Contents Management System |

접수번호 : #121030-003
접수일자 : 2012년 10월 30일
심사완료일 : 2012년 11월 27일
교신저자 : 신성호, e-mail : maximus74@kisti.re.kr

Ⅰ. 서 론

최근 인터넷 상에 게시되는 정보들이 유용한지를 판별하기 위해 시맨틱 웹 등의 기술을 이용한 다양한 소프트웨어가 개발되고 있다[1-3]. 이러한 소프트웨어는 개별적으로 우수하게 지식을 판단할 수 있는 수단을 제공한다. 이와 같은 시맨틱 웹 응용 시스템을 구축하기 위해서는 문서로부터 자연어 처리를 통한 개체 및 관계 인식, 그리고 온톨로지 스키마에 의한 맵핑 과정 등이 필요하다. 그렇지만, 이러한 시스템을 구축하기 위하여 전 과정을 하나로 통합, 관리하는 시스템은 그 필요성에 비해 구축된 사례가 미미하다.

현재, 이와 유사한 시스템으로는 야후의 Yahoo pipe[4]와 바이오 데이터 워크플로우를 관리하는 U-compare[5]가 있다. Yahoo pipe는 웹 데이터와 기존 어플리케이션을 조합하여 사용자가 새로운 웹 어플리케이션을 제작할 수 있는 환경을 제공하며, U-compare는 바이오 문헌의 지식 추출 관련 소프트웨어를 조합하여 워크플로우를 작성 및 관리할 수 있는 환경을 제공한다. 그러나 이와 같은 기 구축 시스템은 언어 처리에 관련된 소프트웨어 및 데이터를 전반적으로 통합하지 않는다.

이러한 과정에 필요한 소프트웨어 및 데이터 자원을 통합적으로 관리할 수 있는 시스템을 구축한다면 다양하고 방대한 정보의 유지 관리를 통해 사용자에게 보다 만족도 높은 서비스 제공이 가능하다. 이러한 시스템은 경제적인 측면에서 통합된 소프트웨어 및 데이터 자원 중복 개발을 방지하여 비용을 절감할 수 있고, 관리 기능 및 정책적인 측면에서, 소프트웨어 유지 보수비용이나 데이터 자원을 재활용할 수 있는 장점을 지닌다[6]. 따라서 본 논문에서는 언어 처리와 관련된 소프트웨어 및 데이터 자원을 통합 관리하기 위한 시스템인 uLAMP(unified Linguistic Asset Management Platform)을 제안한다. 이를 위해 수집되는 데이터 자원의 저장 구조 및 사용자의 소프트웨어와 데이터 셋 관리를 위한 메타 데이터 저장 구조를 설계한다. 아울러, 이러한 저장 구조에 데이터를 등록, 저장, 사용, 조회할 수 있는 기능을 설계하고, 이를 구현한 시스템에 대해서 논한다.

Ⅱ. 관련연구

기존 자원 통합 관리 시스템은 사용자가 등록하는 소프트웨어, 하드웨어, 소프트웨어의 라이선스 등을 주로 관리하면서 다수의 사람들이 이용할 수 있도록 콘텐츠 통합 관리 환경을 제공해준다. 이러한 콘텐츠 통합 관리 시스템의 현황을 살펴보면 다음과 같다. 첫째, 마이크로 소프트에서는 Software Asset Management (SAM) 시스템[7] 을 제공하며, 아울러, 이를 위해 필요한 소프트웨어 관리 프로세스 지침을 제공하고 있다. SAM 관리 프로세스는 소프트웨어 현황을 조사하고, 소프트웨어와 라이선스가 일치하는지를 파악하며, 소프트웨어 구입, 재해 복구, 사용 정책 및 절차에 대해서 검토한다. 아울러 SAM 계획 수립을 위한 설문지 조사 양식 및 소프트웨어와 하드웨어 맵 기능을 제공한다. 그렇지만 이 시스템은 소프트웨어와 하드웨어 관리만을 대상으로 하고 있으며, 사용자가 실제로 적용해볼 수 있는 환경을 제공하기 어려운 문제점을 지닌다.

둘째, M. Jakubička 의 연구에서는 SAM 구축에 필요한 요구사항 및 장점에 대해서 기술하였다[6]. 여기에서는 크게 법률, 정책 및 기능, 경제적 측면에서 SAM에 대해서 기술하였다. 먼저 법률적으로는 소프트웨어의 무단 사용 등에 대해서 적발하고, 시스템이 잘 구축될 수 있도록 환경을 조성해야 한다. 정책 및 기능적으로 SAM 시스템은 이를 사용하는 기관의 룰을 기반으로 구축되어야 한다. 경제적 측면으로는 소프트웨어 관리를 통해 중복 개발을 피함으로써 경제적 이득을 취할 수 있다. 여기서는 SAM에 대한 요구사항과 기능을 정의했지만, 시스템을 구현하지는 않았다.

셋째, Yahoo pipe에서는 주어진 소프트웨어 자원들을 파이프라인으로 연결하여 사용자가 새로운 웹 어플리케이션 형태의 원하는 콘텐츠로 개발할 수 있는 환경을 제공한다[그림 1]. Yahoo pipe는 사이트로부터 데이터를 수집하는 어플리케이션, 데이터 처리를 위한 연산, 그리고 다양한 데이터를 제공한다. 그렇지만, 이에 대한 사용 방법이 직관적이지 않으며, 데이터 처리에 대한 기능을 중심으로 구현되어 데이터 관리에 대한 기능이 미비하다. 또한, Yahoo pipe는 사용자가 원하는 소프

트웨어 등을 등록하는 기능이 없기 때문에 사용자의 다양한 요구를 반영하기가 어렵다.

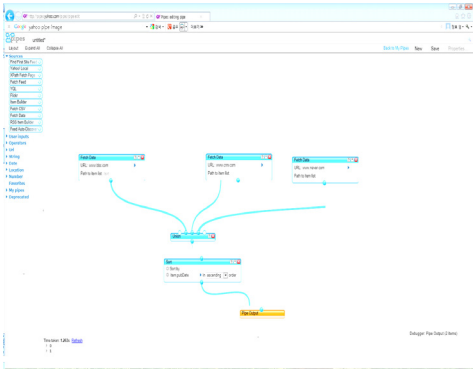


그림 1. Yahoo Pipeline 시스템 사용 화면

마지막으로 U-Compare[8]는 여러 시스템에 분산되어 있는 바이오 인포매틱스 지식 추출 관련 소프트웨어를 하나로 통합하여 바이오 텍스트 마이닝 서비스 형태로 사용자에게 제공하고 있다[그림 2]. 그러나 이는 바이오 인포매틱스 분야에만 제한되어 서비스를 제공한다.

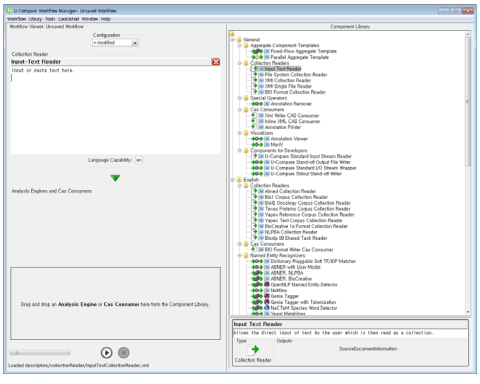


그림 2. U-Compare 시스템 사용 화면

Ⅲ. uLAMP 시스템 구조

본 논문에서 설계하는 시스템은 크게 두 가지 역할을 지니고 있다. 먼저, 웹 기사, 논문, 특허 등 다양한 원천으로부터 수집되는 원문 데이터와 이로부터 자연어 처리를 수행하여 얻어지는 개체 및 트리플 데이터의 저장 구조를 설계 및 구현한다. 이를 통해, 시맨틱 웹 등 다양한 응용에서 이를 활용할 수 있는 체제를 구축한다. 다음으로, 소프트웨어 및 데이터를 통합적으로 사용자에게 제공하고, 사용자가 소프트웨어의 등록 및 활용할 수 있는 체제를 구축함으로써 현재 다양하게 생산되는 지식을 원활히 처리할 수 있는 시스템을 구축하고자 한다.

이를 통해 얻을 수 있는 시스템의 장점은 다음과 같다. 첫째, 언어 처리를 위한 원문 및 이로부터 추출된 개체 및 트리플 자원을 조회 검색 할 수 있기 때문에, 데이터 검증 및 시맨틱 웹 응용 등에 활용할 수 있다. 둘째, 사용자가 원하는 데이터 또는 소프트웨어를 등록할 수 있는 환경을 마련함으로써, 소프트웨어 및 데이터의 공유가 가능하다. 마지막으로, 지식 획득 전반에 걸친 소프트웨어와 데이터 콘텐츠를 통합 관리함으로써 데이터의 수집과 소프트웨어의 개발에 소요되는 비용을 경감할 수 있다.

개발하고자 하는 시스템에서 관리하는 자원은 추출된 개체 및 트리플 자원, 사전, 코퍼스, 온톨로지 등의 언어 자원, 그리고 언어 및 지식 처리 관련 소프트웨어가 해당된다. 이러한 자원을 통합 관리하기 위한 시스템 전체 구조는 [그림 3]과 같다.

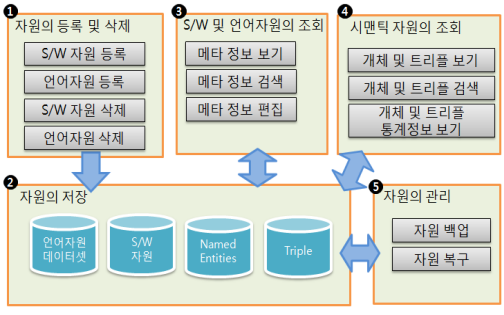


그림 3. uLAMP 시스템 구조

1. 자원의 수집

설계하는 uLAMP 시스템에 적재하기 위해 크게 두 가지 관점에서 데이터를 수집하였다. 먼저, 시맨틱 자원을 저장하기 위해, 웹 사이트, 논문, 특허로부터 데이터를 수집하였다. 또한, 언어 처리를 연구하는 전문가 집

단으로부터 관련 소프트웨어 및 코퍼스, 사전 등의 언어 자원을 제공받았다. [표 1][표 2]는 웹 기사 사이트, Wiki, 논문, 특허로부터 수집한 데이터로부터 인식된 개체 및 트리플 건수를 나타낸다. 개체의 타입은 표와 같이 기술, 제품, 인명, 기관, 위치로 분류하였다.

표 1. 시맨틱 자원(개체) 수집 건수

	기술	제품	인명	기관	위치
웹기사	293,742	544,167	265,845	1,255,730	4,333,830
Wiki	1,183,919	910,228	2,031,990	11,936,827	6,368,304
논문	2,018,990	3,943,222	1,031,365	5,725,623	1,919,186
특허	1,922,472	2,536,421	134,195	2,619,301	1,089,637
Total	5,419,123	7,934,038	3,463,395	21,537,481	13,710,957

표 2. 시맨틱 자원(트리플) 수집 건수

웹 기사	Wiki	Paper	Patent	Total
1,396,681	12,841,106	7,366,030	4,155,124	25,758,941

2. 자원 저장 구조 설계

자원의 저장은 자원의 수집 부분에서 언급했던 바와 같이 시맨틱 자원과, 수집된 소프트웨어와 언어 자원의 메타 데이터를 저장한다. 먼저, 시맨틱 자원은 원문 저장 구조와 원문으로부터 추출한 개체 및 트리플 저장 구조로 분류된다. 시맨틱 자원의 경우, 원문과 추출되는 개체 및 트리플에 대하여 식별을 하고, 원문과 추출되는 데이터 간 연결성을 유지하기 위하여 자원 고유 번호를 할당한다. 자원 고유번호의 할당 체계는 수집되는 데이터 소스의 종류, 현재 데이터베이스의 상태, 그리고 시스템에서 부여되는 레코드 번호로 구성된다. 각 소스 및 데이터베이스의 상태에 따라 부여되는 자원 고유 번호를 위한 코드표는 [표 3]과 같다.

원문의 저장 구조는 수집되는 원천에 따라 웹 기사(News), 논문(Paper), 특허(Patent)에 대한 저장 구조를 설계하였으며, 스키마 구조는 [그림 4]와 같다. 웹 기사는 원문의 내용을 보존하면서 저장하기 위해, 본문, 저자, 작성일자 등을 필드로 가진다. 또한, 카테고리, 섹션 정보 저장을 통해 기사 종류의 분류가 가능하도록 하였다.

표 3. 자원 고유 번호를 위한 코드표

자원의 형태	데이터베이스 소스명	코드명
웹 기사	IDC	IDC
	Wikipedia	WKP
	InformationWeek	INW
	Gizmag	GIZ
	TechnologyReview	TER
	Ieee spectrum	IEE
	TechNewsWorld	TNW
	DiscoverMagazine	DCM
	NewYork Times	NYT
	BBC	BBC
	Fox News	FOX
	CNN	CNN
	Thomson Reuters	REU
	USA Today	USA
	EtnTws.com	ETN
논문	프로시딩 (Proceeding)	PRO
	저널(Journal)	JNL
특허	국제공개 (International Open)	ITO
	미국공개 (USA Open)	USO
	미국등록 (USA Registration)	USR
	유럽공개 (Europe OPEN)	EUO

논문은 출판사, 저널명, ISSN 또는 ISBN, 저자 정보, 키워드 논문 주제 등 논문의 정보를 파악할 수 있는 메타 데이터와 본문으로 구성된다. 특허는 특허번호, 발명자, 명칭, 청구항, 출원국 등의 특허를 알 수 있는 메타 데이터와 특허 내용으로 구성된다. 이와 같이 저장된 원문으로부터 추출된 개체 및 트리플 그리고, 각 데이터 소스마다의 빈도수에 대한 저장 구조를 설계하였으며, 이에 대한 스키마 구조는 [그림 5]와 같다.

웹 기사

자원 고유 번호	**ResourceID**
카테고리	Category
섹션	Section
키워드	Keyword
출판 일자	PubDate
업데이트 일자	UpdateDate
작성자	Author
제목	Title
URL	URL
본문	Body
사이트명	Site
분석 일자	AnalysisDate
입수 일자	CollectionDate

특허

자원고유번호	**ResourceID**
문헌번호	CN
출원번호	AN
출원일자	AD
발명의 명칭	TIE
주발명자	MainINE
공동발명자	SubINE
출원인	PAE
국제특허분류	IC
대표 IPC	ID
우선권	PR
특허일	PD
특허번호	PN
초록	AB
청구항	CM
분석일자	AnalysisDate
입수일자	CollectionDate

논문

자원고유번호	**ResourceID**
통제번호	CN
ISBN	ISBN
ISSN	ISSN
출판사	Publisher
저널명	JournalName
시작페이지	PubstartPage
종료페이지	PubendPage
참고문헌수	RefCount
언어	Lang
주저자	MainAuthor
저자 기관	Affiliation
저자 이메일	Email
논문제목	Title
논문주제	Subject
논문 발행연도	PubDate
논문 타입	Type
초록	Abstract
분석일자	AnalysisDate
입수일자	CollectionDate
원문 위치	FullTextLocation

그림 4. 원문 저장 스키마 구조

트리플

자원 고유 번호	**ResourceID**
주제	Subject
주제 타입	SubjectType
관계	Predicate
관계 타입	PredicateType
객체	Object
객체 타입	ObjectType
추출 일자	ExtractedTime
작성 일자	PublishDate
입수 일자	CollectedDate
출처	ExtractField
부모 자원 고유 번호	ParentResourceID

개체

자원 고유 번호	ResourceID
개체명	Entity
개체 형식	EntityType
작성 일자	PublishDate
입수 일자	CollectedDate
출처	ExtractField
부모 자원 고유 번호	ParentResourceID

빈도수

소스 실제 명칭	FullSource
소스	Source
주제	Subject
주제 타입	SubjectType
관계	Predicate
관계 타입	PredicateType
객체	Object
객체 타입	ObjectType
소스별 빈도 수	CountBySource
전체에 대한 빈도 수	Count

그림 5. 시맨틱 자원 스키마 구조

개체는 개체명, 타입 등으로 구성되며, 트리플은 주체명, 주체타입, 관계명, 관계타입, 객체명, 객체타입 등 트리플을 구성하는 데이터와 타입을 저장한다. 아울러, 빈도수 테이블과 원문 테이블 간의 조인을 통해 년도별/출처별로 추출된 개체 및 트리플이 어느 데이터에서 비중이 높게 나타나는지 그리고 몇 년도부터 나타나게 되었는지를 알 수 있는 척도를 제공한다.

사용자 소프트웨어 및 언어 자원 메타 데이터는 자원의 명칭, 유형, 소개, 건 수, 설치 위치 등의 자료를 포함한다. 자원의 메타 데이터를 통해 사용자는 등록된 소프트웨어 및 언어 자원을 이해할 수 있으며, 실제 소프트웨어 또는 데이터를 업로드 또는 다운로드 할 수 있다. 이러한 기능을 지원할 수 있도록 스키마를 설계하였으며, 이는 [그림 6]과 같다.

소프트웨어 메타 데이터

소프트웨어_일련번호	SW_SEQ
소프트웨어유형	SW_TYPE
소프트웨어명	SW_NAME
이름	M_NAME
전화번호	M_TEL
이메일	M_EMAIL
설치위치정보	SAVE_ADDR
사용자매뉴얼보유여부	MANUAL_YN
프로그램등록여부	PROGRAM_YN
통합관리필요성	IM_YN
개발언어	DEV_LAN
사용방식	DEV_CATE
I/O포맷	IO_FORMAT
기능설명	FUN_TEXT
개발년도_시작	DEV_SDATE
개발년도_끝	DEV_EDATE
이력정보	HISTORY_INFO
소스코드보유여부	SOURCE_YN
소스코드_원본_파일명	SOURCE_ORG_FILE
소스코드_변경_파일명	SOURCE_CHG_FILE
소스코드_파일크기	SOURCE_SIZE
데이터셋참조번호	DS_SEQ
등록일	CREATED_TIME
등록자	CREATOR
수정일	UPDATED_TIME
수정자	UPDATOR

언어자원 메타 데이터

언어자원_일련번호	DS_SEQ
언어자원명칭	DS_NAME
담당자	M_NAME
DB저장소	DB_CATE
건수	DB_COUNT
용도	PURPOSE
데이터위치	DB_ADDR
테이블스페이스명	DB_TABLESPACE
스키마정보	DB_SCHEMA
통합관리필요성	IM_YN
설명	DS_TEXT
최종수정일	LAST_UPDATED
원시데이터 대상년도_시작	ORGDATA_SDATE
원시데이터 대상년도_끝	ORGDATA_EDATE
구축데이터 연도_시작	BUILDDATA_SDATE
구축데이터 연도_끝	BUILDDATA_EDATE
소프트웨어참조번호	SW_SEQ
샘플데이터_원본_파일명	SAMPLE_ORG_FILE
샘플데이터_변경_파일명	SAMPLE_CHG_FILE
샘플데이터_파일크기	SAMPLE_SIZE
등록일	CREATED_TIME
등록자	CREATOR
수정일	UPDATED_TIME
수정자	UPDATOR
데이터적재_분류	DATA_CATE

그림 6. 소프트웨어 및 언어 자원 메타 데이터 스키마 구조

Ⅳ. uLAMP의 구현

본 장에서는 3장에서 설계한 uLAMP를 구현하여 각 기능별로 uLAMP의 사용 절차 및 구현 모습을 기술한다. uLAMP는 소프트웨어 및 언어 자원의 등록, 조회, 수정, 삭제의 기능을 제공하며, 시맨틱 자원의 조회 및 빈도수별 통계 현황을 볼 수 있는 기능을 제공한다. 아울러, 승인된 사용자만이 자원을 등록 및 사용할 수 있도록 한다.

시스템은 JSP와 톰캣[9]을 이용하여 웹 서비스를 구현하였으며, 데이터의 저장을 위해 MS-SQL을 사용하였다. 그 이유는 시스템에서 백업 및 복구, 그리고 대용량의 데이터 처리가 가능하며, 많이 사용되는 Window 서버 플랫폼에 설치가 용이하기 때문이다. 하지만, 시스템에서 사용하는 MS-SQL 내부 구현 함수로 인하여, Linux와 같은 다른 운영체제에서 DB 서버를 운영하지 못하는 단점을 지닌다. 따라서 향후에는 이를 다른

DBMS에서도 사용할 수 있도록 변환할 계획이다.

1. 소프트웨어/언어 자원 등록

소프트웨어 및 언어 자원의 등록은 해당하는 메타 데이터의 등록 화면을 통해 이루어지며, 등록 화면은 소프트웨어 또는 언어 자원에 따라 다음과 같이 저장한다. 먼저, 소프트웨어의 등록은 소프트웨어 유형, 소프트웨어 명칭, 기능 설명, 설치 위치 정보 및 담당자 정보를 필수로 입력해야 하며 이는 사용자가 소프트웨어를 이해하기 위한 최소 정보를 의미한다. 아울러, 소프트웨어의 기능을 부연 설명하기 위해, 설명참조 이미지, 설명 참조 URL, 개발 언어, 개발 방식, IO 포맷, 최종 수정년도, 소스 코드, 매뉴얼의 업로드에 관한 정보를 등록할 수 있다[그림 7].

그림 7. 소프트웨어의 등록 화면

다음으로, 언어 자원은 언어 자원 명칭 및 유형, 설명, 데이터 업로드, 데이터 위치 정보, 담당자 부분이 필수 항목으로써 등록 시 반드시 기입해야 한다. 이는 데이터에 대한 등록이기 때문에 데이터 업로드를 필수항목으로 지정하였다. 아울러, 언어 자원 정보를 파악할 수 있도록 구축년도, 건수, 용도 등을 등록할 수 있다[그림 8].

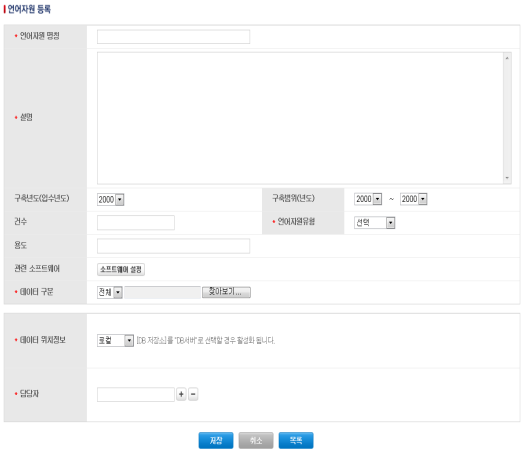


그림 8. 언어 자원 등록 화면

2. 소프트웨어/언어 자원 조회

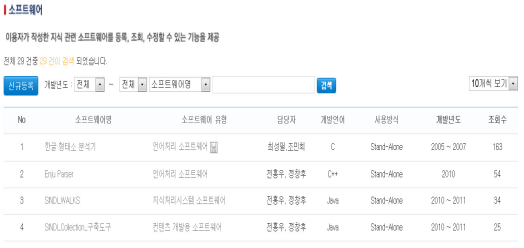


그림 9. 소프트웨어 리스트 보기 화면

소프트웨어 및 언어 자원의 조회는 등록된 전체 자원을 볼 수 있는 리스트 보기 화면 기능, 각 필드 별로 검색할 수 있는 기능 그리고 각 자원 별 메타 데이터 내용을 확인 할 수 있는 상세 보기 기능을 제공한다. 먼저, 소프트웨어 리스트 보기 화면 기능은 소프트웨어명, 소프트웨어 유형, 담당자, 개발언어 사용방식, 개발년도 등을 보여줌으로써 사용자가 필요한 소프트웨어를 빠르게 파악할 수 있다. 아울러 사용자가 원하는 소프트웨어를 검색하기 위해 소프트웨어 명칭 및 유형, 담당자 필드를 선택 검색하는 기능을 제공하며, 개발년도 별로 범위 검색 또한 지원한다[그림 9].

그림 10. 소프트웨어 상세 보기 화면

리스트 보기 화면에서 사용자가 원하는 소프트웨어 또는 언어 자원을 선택하면 이에 대한 상세 보기를 할 수 있다. 상세 보기의 구성은 등록 화면과 동일하다[그림 10].

3. 소프트웨어/언어 자원 수정

소프트웨어 및 언어 자원의 수정은 상세 보기 화면에서 수정 버튼을 클릭하면 [그림 11]과 같이 수정할 수 있다. 자원의 수정은 관리자와 자원을 등록한 사용자만이 수정할 수 있도록 권한 설정이 되어 있어 데이터의 품질을 보장하고자 하였다.

4. 시맨틱 자원 조회

시맨틱 자원 조회는 원문 데이터로부터 추출된 개체 및 트리플 정보를 보여준다. 먼저, 개체는 원문 출처, 개체명, 타입, 빈도수를 보여주며, 데이터 소스, 개체명, 개체 타입으로 검색할 수 있다. 이를 통해 사용자는 원하는 개체에 대하여 원문 출처 및 빈도수 등을 파악할 수 있다. 데이터 소스를 'IDC' 로 검색한 결과는 [그림 12]와 같으며, '3G' 개체에 대한 년도별/출처별 빈도수는 [그림 13]과 같다. 이러한 빈도수를 통해 사용자는 개체가 주로 언급되었던 시기 및 빈도수가 높은 데이터 소스 등을 파악 가능하다.

그림 11. 소프트웨어 자원 수정 화면

그림 12. 시맨틱 자원(개체) 리스트 보기 화면

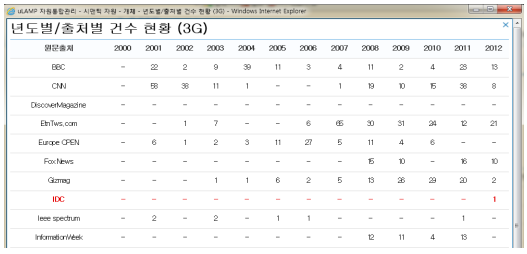


그림 13. 시맨틱 자원(개체) 빈도수 보기 화면

트리플 자원 보기 화면 또한 개체 보기와 동일한 기능을 포함하며, 트리플은 Subject, Predicate, Object의 명칭 및 타입에 따른 검색을 수행할 수 있다. 데이터 소스 'IDC'에서 트리플 명칭 'apple'로 검색했을 때 화면은 [그림 14]와 같으며, 'apple'과 'iTunes'가 'own'관계로 구성되는 트리플에 대한 빈도수 화면은 [그림 15]와 같다. 개체와 마찬가지로 트리플의 빈도수를 통해 개체 간 관계가 주로 언급된 시기 또는 주로 언급했던 데이터 소스 등에 대해서 알 수 있다.

그림 14. 시맨틱 자원(트리플) 리스트 보기 화면

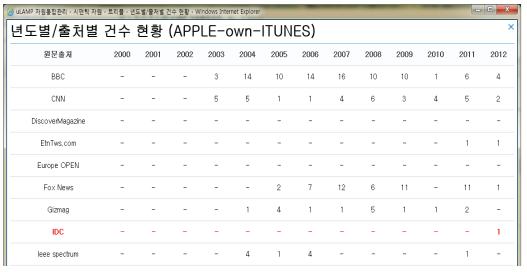


그림 15. 시맨틱 자원(트리플) 빈도수 보기 화면

V. 시스템 사용성 조사

구현한 시스템의 평가를 위해 시스템이 사용자에 게 편의성을 제공하는 정도(사용성)와 시스템의 정보가 어느 정도 유용한지를(효용성) 언어 처리 전문가 집단 11명에게 설문 조사를 통해 분석하였다. 언어 처리 전문가는 KISTI 소속의 언어 처리 전공자를 대상으로 하였으며, 전문가의 연령대는 30대~40대 초반이다. 설문 문항은 KISTI의 DCMS 시스템[10][11]에서 실시했던 설문 조사를 참고하였다[12-14]. 데이터베이스 측면에서의 설문 문항은 데이터의 정확성, 완전성, 최신성, 일관성과 같은 데이터 품질을 고려하여 작성하였다. 또한 시스템 측면에서는 검색성, 사용 용이성, 사용자 지원성, 비용과 같은 서비스 품질을 고려하였다. 시스템의 설문 조사 문항 및 결과는 [표 4]와 같다. 문항 수는 총 18문항을 설문하였으며, 답변은 Likert의 5점 척도를 사용하였다. 문항이 총 18문항이기 때문에 배점을 5.5점부터 1점씩 차등하여 계산하였다. 표에서 나타내듯, 사용자의 설문 평균 점수가 84점으로 사용자 만족도가 약 80% 정도 인 것으로 나타났다. 데이터베이스 측면에서는 데이터의 일관성 및 최신성 측면에서 높은 점수를, 시스템 측면에서는 일관된 네비게이션 구조등 사용 용이성 측면에서 높은 점수를 받았다. 또한, 데이터베이스 측면에서는 데이터의 완전성, 시스템 측면에서는 시각적 명확성에서 낮은 점수를 받았다.

제안하는 uLAMP 시스템은 언어 자원을 지속적으로 수집하기 때문에 최신성 측면에서는 장점을 지니고 있다. 하지만, 빈번한 업데이트로 인하여 완전성 측면에서는 다른 항목에 비해 낮은 점수를 지니고 있다. 따라서 이를 해결하기 위해서는 빈번한 업데이트에 대응 가능하도록 DB 스키마 구조를 최적화할 필요가 있다. 아울러, 시스템 부분에서는 사용자가 이용하기에 용이하기 때문에, 향후 시스템 배포에 장점을 지닌다. 또한, 시각적 명확성은 메인 페이지UI의 수정 등을 통해 사용자가 보기에 적당한 정보만을 선별하여 게시할 필요가 있음을 확인하였다. 관련 연구에서 소개한 기존 Yahoo pipe와 U-compare와의 시스템을 비교하면, Yahoo pipe는 uLAMP와 같이 웹 사이트에서 데이터를 수집하기 때문에 최신성 측면에서는 장점을 보이나, 입력 및 출력 데이터, 데이터의 종류 등의 설정이 필요하기 때문에 사용이 용이하지 않는 단점을 지닌다. 한편, U-compare는 바이오 데이터의 언어처리에만 한정되어 있기 때문에, 이에 대한 지식이 없으면 사용이 용이하지 않다.

VI. 결 론

본 논문에서는 시맨틱 웹 응용에 필요한 원문으로부터 추출한 자원 및 언어 처리 전반에 걸친 소프트웨어 및 언어 자원을 통합 관리하는 uLAMP 시스템을 제안하였다. 이러한 uLAMP 시스템은 사용자 설문에서 보이듯이, 서비스 사용의 용이성 및 데이터의 최신성 등에대한 장점을 지닌다.

uLAMP를 구축함으로써, 자연어 처리의 각 단계에서 필요한 소프트웨어나 언어 자원을 통합 관리하기 때문에 원하는 것을 필요에 따라 조회할 수 있으며, 원문으로부터 정제된 트리플, 개체 등을 제공함에 따라 트리플 생성 과정이 필요 없기 때문에, 이를 필요로 하는 자연어처리 소프트웨어에서 입력 데이터로 유용하게 사용될 수 있다. 이는 소프트웨어의 개발 및 데이터 유지에 필요한 비용을 감소시킨다.

향후에는 시맨틱 자원을 추출하기 위해 사용되는 원문 정보를 uLAMP에서 표현할 수 있는 방안을 설계할 것이며, 사용자의 소프트웨어 및 언어 자원을 지속적으로 수집하여 공유함으로써 언어 처리 및 콘텐츠를 다루는 전반적인 연구에 기여하고자 한다.

표 4. 설문 조사 결과

	설 문 문 항	평균 점수
데이터베이스/콘텐츠	철자오류(Spelling error)나 잘못된 데이터 값이 있습니까?	4.04
	자원(소프트웨어, 언어 자원, 시맨틱 자원)에 관한 중요한 속성들을 모두 담고 있습니까?	5.13
	레코드의 주요 필드 값이 비어있는 부분이 있습니까?	3.59
	리스트 화면에 보여주는 데이터의 값과 상세보기 화면의 데이터 값에 일관성이 있습니까?	5.04
	정보의 발생, 수집, 그리고 갱신 시기가 명료하게 표현되고 있습니까?	5.13
	불필요하게 중복되는 속성이 있습니까(이름은 틀리더라도 의미상)?	4.59
	속성에 담기는 값들이 일관성이 있습니까?	4.95
	동일한 사실을 표현하는 두 데이터 값이 불일치하는 경우가 있습니까?	4.04
	검색결과는 보기 편리하게 나열되고, 결과의 수를 조절할 수 있습니까?	5.13
시스템/콘텐츠	메뉴 항목 명칭이 일관성을 가지고 체계적이며, 기능을 명확히 나타내고 있습니까?	4.95
	모든 페이지에서 일관성 있는 네비게이션 구조를 사용합니까?	5.13
	사이트 디자인과 레이아웃, 메뉴구조가 간단명료하고 이해하기 쉽습니까?	4.95
	메인 페이지에 너무 많은 정보를 담고 있어 보기에 불편합니까?	4.04
	화면에 나타나는 데이터를 쉽게 이해할 수 있으며, 화면 당 출력량, 배열, 색상, 하이라이트 사용이 전반적으로 적절합니까?	4.77
	데이터 입력 화면과 대화상자 내의 필드는 적절한 기본 값을 가지고 있습니까?	4.95
	다양한 정렬방법(sort)을 지원합니까?	4.59
	선택된 메뉴가 선택되지 않은 메뉴에 비해 뚜렷하게 구별되어집니까?	4.68
	사이트의 용어사용에 있어서 일관성을 유지하며 문법적으로 정확합니까?	4.95
Total		84.65

참 고 문 헌

[1] 이미경, 정한민, 김평, 성원경, “연구개발 전략 수립 지원을 위한 테크놀로지 인텔리전스 서비스”, 정보과학회논문지, 제17권, 제5호, pp.337-341, 2011.

[2] W. Sung, H. Jung, P. Kim, I. Kang, S. Lee, M. Lee, D. Park, and S. Hahn, “A Semantic Portal for Researchers Using OntoFrame,” 6th International Semantic Web Conference, 2007.

[3] S. Song, H. Oh, S. Myaeng, S. Choi, H. Chun, Y. Choi, and C. Jeong, “Procedural knowledge extraction on MEDLINE abstracts,” Active Media Technology, pp.345-354, 2011.

[4] http://pipes.yahoo.com/pipes/

[5] http://u-compare.org/index.en.html

[6] M. Jakubička, "Software asset management," IEEE International Conference on Software Maintenance (ICSM), pp.1-2, 2010.

[7] http://www.microsoft.com/korea/resources/sam/

[8] Y. Kano, J. Björne, F. Ginter, T. Salakoski, E. Buyko, U. Hahn, K. B. Cohen, K. Verspoor, C.

Roeder, L. Hunter, H. Kilicoglu, S. Bergler, S. Van Landeghem, T. Van Parys, Y. Van de Peer, M. Miwa, S. Ananiadou, M. Neves, A. Pascual-Montano, A. Ozgur, D. R. Radev, S. Riedel, R. Sætre, H. Chun, J. Kim, S. Pyysalo, T. Ohta, and Tsujii, "U-Compare bio-event meta-service: compatible BioNLP event extraction services," BMC BIOINFORMATICS Vol.12, pp.481-489, 2011.

[9] http://tomcat.apache.org/

[10] W. Lee, M. Lee, K. Kim, S. Shin, H. Yoon, and W. Sung, "An Management of Digital Contents on Science and Technology," Communications in Computer and Information Science, Vol.264, pp.227-229, 2011.

[11] 이민호, 이원구, 윤화묵, 신성호, 류재철, " 해외 과학기술 학술논문 메타데이터의 비교 분석", 한국콘텐츠학회논문지, 제11권, 제9호, pp.515-523, 2011.

[12] 이주현, 이응봉, 김환민, "NDSL 웹사이트 분석 및 서비스 품질평가", 정보관리연구, 제37권, 제4호, pp.69-91, 2006.

[13] 한국데이터베이스진흥센터, "데이터베이스 품질평가 항목", 데이터베이스 표준화 연구, 2006.

[14] 한국정보통신기술협회, "데이터베이스 품질평가 항목", 정보통신 단체표준, 2000.

저 자 소 개

엄 정 호(Jung-Ho Um) 정회원

- 2006년 : 전북대학교 컴퓨터공학과(석사)
- 2011년 : 전북대학교 컴퓨터공학과(박사)
- 2011년 ~ 현재 : 한국과학기술정보연구원 선임연구원

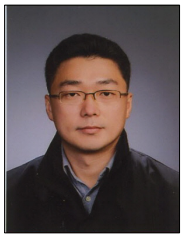
<관심분야> : 데이터베이스, 정보추출, 분산 병렬 처리

신 성 호(Sung-Ho Shin) 정회원

- 2000년 : 경북대학교 경영학과(학사)
- 2002년 : 경북대학교 경영학과-경영정보시스템(석사)
- 2002년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 정보추출, 지식공학, 시맨틱웹, MIS

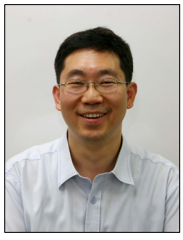
최 성 필(Sung-Pil Choi) 정회원



- 1998년 : 부산대학교 전자계산학과(석사)
- 2012년 : 한국과학기술원 정보통신공학과(박사)
- 1998년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 기계학습, 정보검색, 자연어처리, 정보추출, 텍스트마이닝

정 한 민(Han-Min Jung) 정회원



- 2003년 : 포항공과대학교 컴퓨터공학과(박사)
- 2004년 ~ 현재 : 한국과학기술정보연구원 책임연구원
- 2004년 ~ 현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 시맨틱웹, 정보검색, 자연어처리, HCI