시소러스 구축을 위한 웹 기반 워크벤치 개발

이승준[1], 정한민[2], 성원경[2], 최 광[1], 이상헌[1], 최석두[3]
㈜오롬정보[1], 한국과학기술정보연구원[2],한성대학교[3]
{dino[1], choik[1], shlee[1]}@orom.com, {jhm[2], wksung[2]}@kisti.re.kr,
sdchoi@hansung.ac.kr[3]

Development of Web-based Workbench for the Construction of Thesaurus

Seung jun Lee[1], Hanmin Jung[2], Won-Kyung Sung[2], Kwang Choi[1], Sang hun Lee[1], Sukdoo Choi[3]
Orominfo co. Ltd.[1], Korea Institute of Science and Technology Information[2], Hansung University[3]

요 약

본 연구에서는 다양한 개념 패싯과 관계 패싯들을 수용한 범용 과학기술 시소러스 구축용 웹 기반 워크벤치 개발에 대해 기술한다. 기존 국내 시소러스 구축용 워크벤치들이 제공하는 기본적인 용어 관계 구축 기능을 확장하여 개념 패싯, 범주 관계 패싯, 의미역 관계 패싯, 속성 관계 패싯 및 속성 키워드 처리 기능을 원활히 제공할 수 있는 사용자 중심적 워크벤치를 개발함으로써 시소러스 상의 개념들에 대한 효율적인 구축이 가능하도록 한다. 또한, 시멘틱 웹 상의 온톨로지 영역에 보다 근접한 고도화된 시소러스 구축을 위해 용어들을 개념화시키고, 개념간의 다양한 관계를 설정하는 프로세스 중심적 설계로 분야 적합성이 높은 정보 처리 기반을 갖춘다. 궁극적으로는 여러 마이크로 시소러스들을 통합하여 운용할 수 있는 복합 모델을 구축하는 것을 목표로 하고 있다.

이러한 목적에 부합하는 시스템 구현을 위해 CBD (Component Based Development) 개발 방법론으로 MSF/CD 를 이용하였으며, 분산 환경에서 이기종간의 데이터 교환을 용이하게 하기 위하여 웹 서비스 (XML Web Services)를 이용하였다. 또한, 시멘틱 웹 기반 연구자 간 협업 지원 서비스 구현을 위한 확장 검색용으로서도 활용할 수 있도록 하였다. 시소러스 반출은 CSV, XML 및 RDF 를 모두 지원할 수 있도록 함으로써 다양한 사용자 요구 사항에 부합할 수 있도록 하였다. 시소러스 브라우징을 시각화 기반의 3단계 구조를 가진 플래시로 구현하여 사용자가 쉽게 시소러스를 탐색하고 분석할 수 있는 기반을 제공하였다. 또한, 다양한 검색 요구를 만족시키고자 기본 검색, 고급 검색, 메타 검색을 선택할 수 있도록 하며, 개념 편집 및 시소러스 브라우징과 연동시켜 효율적인 시소러스 구축이 가능하도록 하였다.

본 연구의 워크벤치를 이용하여 구축된 시소러스는 기존 시소러스들에 비해 사용자가 보다 폭넓은 의미 기반 검색을 수행할 수 있도록 함으로써 다각적인 정보를 쉽게 획득할 수 있는 기반을 마련하고 있다는 데 의의가 있으며, 다국어 시소러스 및 다중 시소러스를 수용할 수 있는 방향으로 발전시킬 계획이다.

Keyword : Thesaurus, Web-based Workbench, Semantic Web, Ontology

1. 서론

정보의 종류와 그 활용방안이 다양화됨에 따라 수많은 종류의 정보구축 시스템과 정보검색 시스템이 대두되었고, 이를 활용하는 정보 기기 역시 다양화되었다. 이런 정보들이 정해진 일련의 규칙이나 해당 도메인에 따라 정제되어 분류되어 있고, 이들 간의 관계가 직관적인 정보를 제공해

준다면 좀더 유용하게 활용될 것이다. 시소러스는 정보의 기본 단위가 되는 용어의 정의와 용어들 간의 관계를 정의해 주는 중요한 개념으로 오늘날 다양한 분야에서 이를 활용하는 시스템이 활발히 연구 중이며 그 대표적 사례가 국립중앙도서관의 주제명표목표 시스템이다[1]. 그러나 현재 국내에는 과학기술분야 시소러스 구축을 위하여 참고할 구축 지침이 마련되어 있지 않으며, 범용 시소러스 구축을 위해 국제 표준으로 사용되고 있는 ISO-2788 을 사용하고 있는 실정이다. 또한 시소러스 구축지침에 정합되는 워크벤치 역시 찾아볼 수 없다.

본 연구에서는 범용 과학 기술 시소러스 구축용 웹 기반 워크벤치 개발에 대해 논하며, 2005 년 KISTI 에서 진행한 "10 만 범용 과학기술분야 전문용어에 대한 계층적 개념망/어휘망 구축 및 워크벤치 개발" 사업의 일환으로 작성된 시소러스 구축지침 V.11 을 준용하였다[2].

2. 워크벤치의 설계

2.1 관계지시기호 및 패싯

워크벤치에서 구축되는 시소러스는 전술한 시소러스 구축지침에서 일반 시소러스와는 다른 부분을 갖고 있다. 즉, 구축의 대상이 용어에서 개념으로 변경되었으며, RT 관계를 생각하지 않는다는 것이 그것이다. 그 차이를 표로 보면 <표 1>과 같다[2].

표 1. 관계지시기호 비교

	전통적 시소러스	목표 시소러스
대상	용어(Term)	개념(Concept)
관계	상위관계(Broader Term) : BT	상위관계(Broader Term) : BT
	하위관계(Narrower Term) : NT	하위관계(Narrower Term) : NT
	관련관계(Related Term) : RT	우선관계(Use) : USE
	우선관계(Use) : USE	비우선관계(Used For) : UF
	비우선관계(Used For) : UF	

또한 15 개의 개념패싯과 6 개의 범주관계패싯, 9 개의 의미역 관계패싯, 16 개의 속성관계패싯을 활용하여 관계정보를 확장시킨다. 사용되는 패싯의 종류는 <표 2>와 같다[2].

표 2. 패싯의 종류

개념패싯	범주관계 패싯	의미역관 계패싯	속성관계 패싯
분야・이론・방법	전체부분 -의미적	근원	분야・이론・방법
감각・감정	전체부분 -조어적	대상	감각・감정
위치・공간	사례 -의미적	도구	위치・공간
기기・장치・부속	사례 -조어적	시간	기기・장치・부속
물질・재료	일반화 -의미적	목표	물질・재료
조직	일반화 -조어적	장소	조직
생물		행위자	생물
상태・성질		수혜자	상태・성질
시간		방식	시간
언어			언어
단체			단체
질병・증상			질병・증상
내용			내용
행위			행위
현상・사건			현상・사건
			사례

2.2 시스템 구성

워크벤치는 MS 사의 .Net Framework 을 기반으로 설계/개발하였으며 CBD(Component Based Development)개발 방법론 중 MSF/CD(Microsoft Solution Framework/Components Development)를 이용하였다. .Net Framework 은 개발자의 직관적인 개발환경을 제공함으로써 생산성을 향상시켜주며 응용프로그램의 안정성과 다양한 언어간의 통합성을 지원한다. 이는 워크벤치가 MS 사의 Windows 를 플랫폼으로 이용하는데 최상의 환경을 제공한다. MSF 는 개념설계, 논리설계, 물리설계의 세

단계로 설계를 진행 하며 매우 유연하고 직관적인 모델을 제시하여 본 워크벤치처럼 Business Logic 을 지속적으로 변경하거나 다양한 환경정보를 모듈화시켜야 할 경우 매우 유용한 방법론이다[3]. 워크벤치에서는 모든 구성요소를 컴포넌트로 정의하고 이를 개발/활용하기 위한 효과적인 방법으로 MSF 의 여러 항목 중 CD 를 선택하였다. 또한 이기종간의 데이터 교환을 위해 SOAP 기반의 웹 서비스를 이용하였다. 웹 서비스는 네트워크을 통해 데이터를 서로 다른 시스템간에 공동으로 사용할 수 있도록 통신규약으로 SOAP(Simple Object Access Protocol)과 저장소인 UDDI(Universal Description, Discovery, and Integration), 기술언어로 WSDL(Web Services Description Language)을 지원한다[4]. 워크벤치에서는 웹 서비스를 통하여 시멘틱 웹 기반 연구자 간 협업 지원 서비스 구현의 확장검색 모듈로 사용하였다. <그림 4>는 웹 서비스를 통하여 제공되는 XML 코드의 예이다.

```
<?xml version="1.0" encoding="utf-8" ?>
- <DataSet xmlns="http://tempuri.org/">
  + <xs:schema id="NewDataSet" xmlns=""
      xmlns:xs="http://www.w3.org/2001/XMLSchema"
      xmlns:msdata="urn:schemas-microsoft-com:xml-msdata">
  - <diffgr:diffgram xmlns:msdata="urn:schemas-microsoft-com:xml-
      msdata" xmlns:diffgr="urn:schemas-microsoft-com:xml-diffgram-
      v1">
    - <NewDataSet xmlns="">
      - <table diffgr:id="table1" msdata:rowOrder="0">
          <Concept>예측프로그램</Concept>
          <ConceptRelation>BT</ConceptRelation>
          <RelationConcept>프로그램</RelationConcept>
        </table>
      - <table diffgr:id="table2" msdata:rowOrder="1">
          <Concept>예측프로그램</Concept>
          <ConceptRelation>NT</ConceptRelation>
          <RelationConcept>기술수요예측프로그램</RelationConcept>
        </table>
      - <table diffgr:id="table3" msdata:rowOrder="2">
          <Concept>예측프로그램</Concept>
          <ConceptRelation>NT</ConceptRelation>
          <RelationConcept>기술예측프로그램</RelationConcept>
        </table>
      </NewDataSet>
    </diffgr:diffgram>
  </DataSet>
```

그림 1. 웹 서비스상에서 출력되는 XML 코드의 예

2.3 시각화 도구

본 시스템에서는 시소러스 브라우징을 시각화 기반의 3 단계 구조를 가진 플래시로 구현하여 사용자가 쉽게 시소러스를 탐색하고 분석할 수 있는 기반을 제공하였다.

시소러스 데이터를 친숙하게 표현하기 위해서 매크로미디어의 솔루션인 플래시를 기반으로 가시화 단계를 처리하였는데, 플래시로 개발된 응용 프로그램은 백엔드 프로그램과 XML 로 데이터 통신한다. XML 이 데이터를 응용 프로그램으로 전달하는 데 사용되는 산업 표준으로 자리 매김한 후 Flash 는 XML 에 대한 뛰어난 프로그래밍 기능을 지원하기 시작하였고 비즈니스 프레젠테이션의 컨텐츠를 XML 파일로 캡슐화한 후 Flash 응용 프로그램에서 이 파일을 읽고 프레젠테이션을 제공한다. 워크벤치에서는 Flash 를 시각화 도구로 이용하기 위해 XML 을 이용한다[5]. <그림 2>는 플래시 브라우저의 예이다.

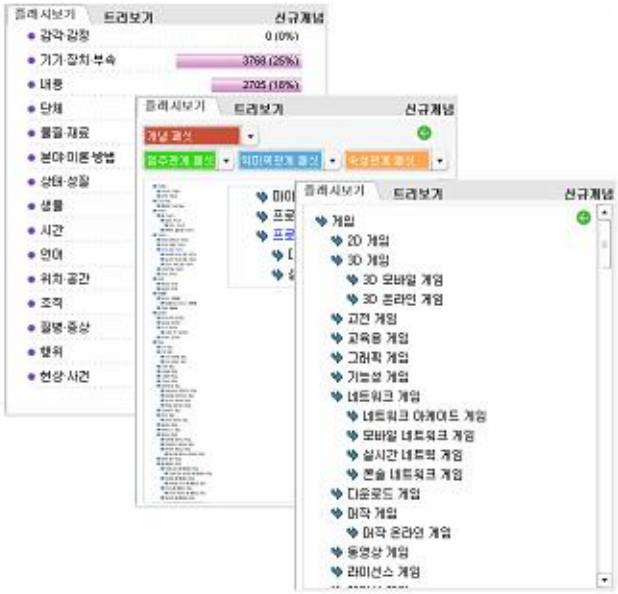

그림 2. 플래시 브라우저

2.4 워크벤치의 구조

본 시스템에서는 기존 국내 시소러스 구축용 워크벤치들이 제공하는 기본적인 용어 관계 구축 기능을 확장하여 개념 패싯, 범주 관계 패싯, 의미역 관계 패싯, 속성 관계 패싯 및 속성 키워드 처리 기능을 원활히 제공하기 위해 용어를 개념화시키고 개념의 구조를 정의하여 시소러스 상의 개념들에 대한 효율적인 구축이 가능하도록 하였는데, 그 첫 단계가 개념화이다. 개념화는 구축대상 용어에 대하여 개념패싯을 부여하는 것으로 개념화가 된 용어만이 관계설정의 대상이 되며 시소러

스 데이터를 구성하게 된다.

본 시스템을 통하여 시소러스를 구축하기 위해 <그림 3>과 같은 단계가 진행되는데, 말뭉치를 통해 추출된 용어를 개념패싯을 이용하여 개념화하고 이렇게 구성된 개념에 관계를 부여하며, 관계패싯과 속성키워드를 부여하면서 시소러스 데이터를 구축하게 되며 데이터의 보정과 오류를 최소화하기 위해 시소러스 검수 지침에 의거 데이터들의 검수작업을 진행한다[6]. 이렇게 검수단계까지 진행된 데이터들은 XML 형태로 플래시 브라우저나 트리 브라우저, 혹은 웹 서비스를 통하여 이 기종 시스템에게 서비스 할 수 있는 형태로 변형되며 CSV(Comma Separated Value), XML, RDF(Resource Description Framework) 의 파일형태로 추출되어 다양한 사용자의 요구에 부합되도록 제공된다.

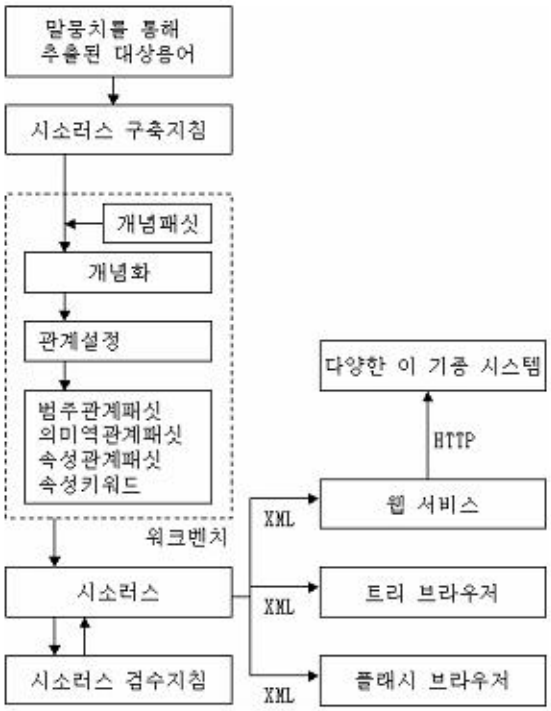

그림 3. 워크벤치를 이용한 시소러스 구축/활용

실제 <그림 3>의 과정을 거쳐 구축된 시소러스 데이터는 기계적으로 통제 가능한 데이터 형태로 <그림 4>의 우측과 같은 개념적 구조를 가지게 되며, 이는 향후 본 시스템이 확장할 다양한 마이크로 시소러스 시스템의 근간을 이룬다.

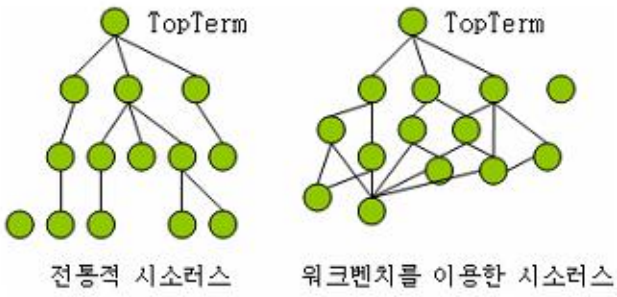

그림 4. 전통적 시소러스와 워크벤치를 이용한 시소러스의 개념적 구조

본 시스템에서는 개념의 필수 항목으로 개념명, 개념패싯, 출처, 용례로 정의하고 영문개념명과 범위주기를 부가적 정보로 정의하였으며 개념관계는 관계정보, 범주관계패싯, 속성관계패싯, 속성키워드를 필수정보로 의미역관계패싯을 부가적 정보로 정의하여 시소러스를 구축하도록 하였다. 이렇게 정의된 개념의 구조는 <그림 5>과 같으며 이렇게 정의된 개념이 지속적으로 축적되어 시소러스를 구성하게 된다.

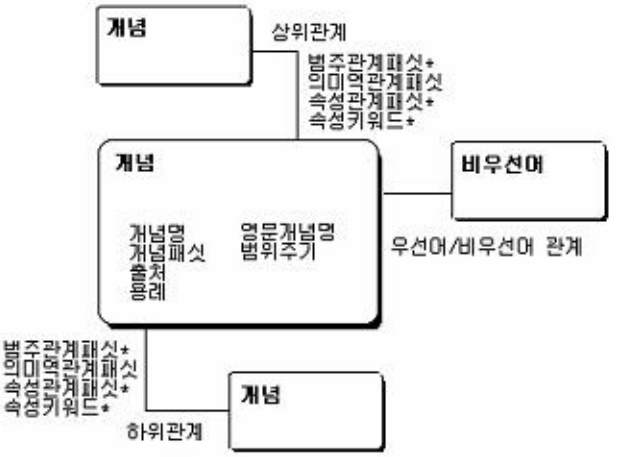

그림 5. 개념과 개념관계

구축된 시소러스는 <그림 6>과 같은 상하개념의 그룹을 이루게 되며 대문자는 용어, 소문자는 해당 용어의 동의어를 나타낸다. 하나의 대문자가 하나의 개념을 표현하는 용어라고 할 수 있다. 용어에 따라 의미역 관계패싯을 갖기도 하며, 상하개념의 구분은 조어적 기준과 의미적 기준에 따라 결정된다. <그림 6>에서 A→NO, A→M 등은 의미적 기준에 따른 것이며, A→AB(aB), A→CA(Ca) 등은

의미적 기준에 따른 상하개념 설정의 유형이다[2].

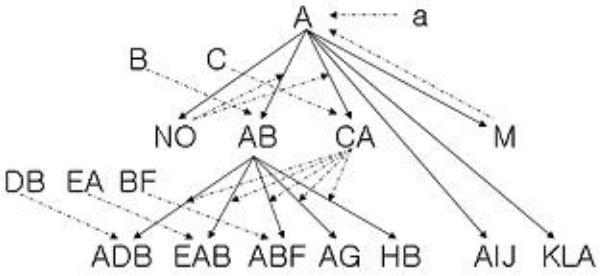

그림 6. 개념관계의 유형

2.5 워크벤치 구현

본 시스템은 위에서 언급한 개념의 구조를 바탕으로 하여 Interface Layer, Business Layer, Utility Layer, Database Layer 의 계층구조를 이룬다. Interface Layer 는 User Interface 와 Service 로 세분화 되어 시스템에서 제공하는 시소러스 구축, 검수, 검색과 같이 이용자가 직접 접근하여 사용하는 부분은 User Interface 가, 웹 서비스나 파일시스템 반출부분과 같이 연계시스템, 혹은 이기종 시스템이 이용할 부분은 Service 부분이 처리하게 되며 이들이 이용할 각종 구축, 검수, 브라우징 규칙들은 Business Logic 이 담당하게 된다. Utility 부분은 Business Logic 과 Database Agent 부분의 연결과 파일시스템의 접근 등 좀 더 물리적인 부분의 처리를 담당하며, Database Agent 는 데이터베이스 처리부분을 진행한다. 각각의 레이어는 Unicode 기반의 XML 로 통신을 하며 <그림 7>은 본 시스템의 구성도이다.

이렇게 구성된 시스템은 시소러스 구축지침, 시소러스 검수지침을 각각 컴포넌트화 하여 Business Logic 에 배치하고 이들이 변경될 때마다 각각의 컴포넌트를 재조립하게 된다. 또한 향후 발전모델인 다국어지원 다중 시소러스를 고려하여 Database Agent 는 Unicode 와 KSC 5601 을 동시에 지원하며 다양한 형태의 RDBMS(Relational Database Management System)를 지원하기 위하여 외부의 Adapter 를 이용하도록 구현하였다.

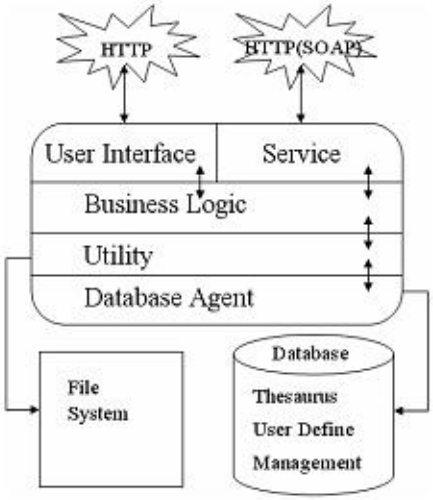

그림 7. 시스템 구성도

<그림 8>은 실제 구현된 시소러스 구축을 위한 웹 기반 워크벤치의 화면으로 A 영역은 구축된 시소러스 데이터를 관계를 중심으로 한 검색을 실행하는 검색과 전문가를 위한 고급검색, 외부 데이터 검색을 위한 메타검색 부분이며, B 영역은 플래시 브라우저와 트리 브라우저를 이용하여 개념 및 개념관계를 탐색하는 브라우저 부분이다. C 영역은 용어를 개념화 시키고 개념관계를 설정하며 각종 패싯을 부여하고 시소러스 검수를 진행하는 편집화면이다.

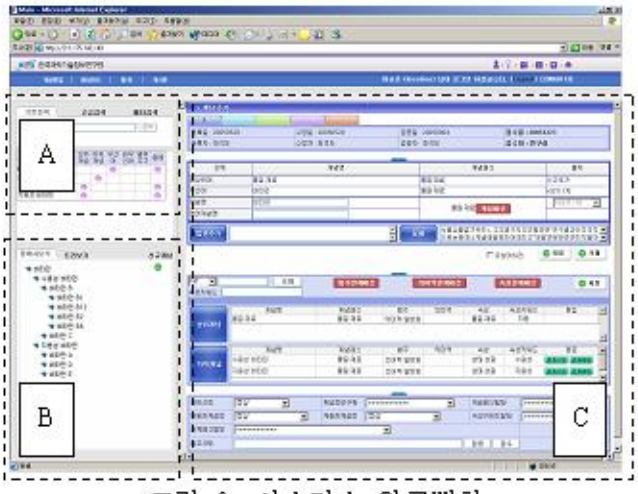

그림 8. 시소러스 워크벤치

3. 결론 및 향후 방향

본 시스템을 통하여 용어를 개념으로 확장하고 개념 패싯, 범주 관계 패싯, 의미역 관계 패싯,

속성 관계 패싯, 속성 키워드를 사용자 중심적으로 활용할 수 있도록 함으로써 온톨로지 영역에 좀더 근접한 시소러스 구축이 가능하게 되었으며 분야 적합성이 높은 정보 처리 기반을 갖추게 되었다. 또한 기존 시소러스 구축 시스템이 지니지 못한 시소러스 데이터의 기계적 통제와 다양한 각도(조어적, 의미적)의 브라우징이 가능하게 되었다. 향후 이 시스템을 활용, 확장하여 Unicode 기반의 다국어 시소러스와 다중 시소러스를 구축/활용할 수 있도록 할 예정이며, 이를 통해 여러 마이크로 시소러스 들을 통합하여 운용할 수 있는 복합 모델을 구축할 수 있을 것이다.

참고문헌

[1] 이상헌, "국립중앙도서관 주제명표목표 개발", 2002

[2] 최석두, 김이겸, 한상길, 최상기, 윤혜영, 백태현, "시소러스 구축지침 V.11", 한국과학기술정보연구원, 2005

[3] John Erik Hansen & Carsten Thomsen, "Enterprise Development with Visual Studio.NET, UML, and MSF", Apress, 2004

[4] David Booth, Hugo Haas, Francis McCabe, Eric Newcomer, Michael Champion, Chris Ferris, David Orchard , "Web Services Architecture", 2004

[5] Mike Chambers, "Flash Professional and Flex 2", Macromedia, 2005.

[6] 황순희, "시소러스 검수지침", 한국과학기술정보연구원, 2005