2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

XML 기반 웹 사전 프레임워크

유응구*, 구자룡*, 김고운*, 이금석*, 김병구**
*동국대학교 컴퓨터공학과, ** ㈜티엔테크
e-mail : {engus, jekal, gowoon, kslee}@dgu.ac.kr, bkim@tntech.co.kr

The XML-based Web Dictionary Framework

Eung-Gu You*, Ja-Ryong Koo*, Go-Woon Kim*, Keum-Suk Lee*, Byung-Gu Kim**
*Dept. of Computer Engineering, Dongguk University, **Transnodes Technology, Inc

요 약

인터넷이 널리 사용되면서 인터넷 사용자들의 웹 사전 사용 빈도가 크게 증가하였다. 현재 다양한 형태의 웹 사전들이 다양한 서비스를 제공하고 있지만, 대부분의 웹 사전들은 검색엔진 형태의 단순질의 검색과 인덱스를 이용한 검색 기능을 제공하고, 고정된 표현 형식만을 제공한다. 또한 사전 내용, 인덱스, 참조 및 관련 사이트 정보를 관리하기 위한 도구의 부재로 관리가 어렵고, 저장형식으로 HTML을 사용하기 때문에 사전 데이터의 재사용에 문제가 있다.

따라서 본 논문에서는 기존의 웹 사전들을 사용자-관리자 측면에서 분석하여 문제점을 제시하고, 이를 해결할 수 있도록 카테고리 검색 및 히스토리 검색 서비스를 지원하고, 데이터를 XML 문서로 저장하며, 전용의 사전 관리 도구를 제공하는 XML 기반 웹 사전 프레임워크를 제안하고 구현하였다.

1. 서론

인터넷이 널리 보급됨에 따라 사용자들은 더욱 많은 정보들을 접하게 되었다. 다양한 정보를 이해하기 위해서는 정보를 구성하는 용어에 대한 이해가 필요하고, 이를 위해서 사용자는 다양한 용어 사전을 이용한다. 최근에는 정보를 이해하기 위해 책이나 CD 형태의 사전을 사용하기 보다는 인터넷에 접속하여 사용하는 웹 사전을 이용하는 사용자가 크게 증가하였다.

웹 사전에 대한 사용자의 수요는 크게 증가하였지만 기존의 웹 사전들은 대부분 검색엔진 형태의 검색 기능과 인덱스 검색 기능만을 제공하고, 사전 내용에 스타일 정보가 포함되어 있기 때문에 고정된 표현 형식을 갖는다. 또한 사전 데이터를 HTML 형태로 저장하기 때문에 공유, 교환 등 재사용이 어렵고, 사전 내용을 수정, 추가, 삭제 등 사전 내용 관리와 참조, 인덱스 및 관련 사이트 정보 관리를 관리자가 텍스트 에디터나 웹 에디터를 이용하기 때문에 관리에 어려움이 많다. 따라서 본 논문에서는 사용자와 관리자의 요구를 수용하여 기존 웹 사전의 문제점을 해결할 수 있는 웹 사전 프레임워크를 제안하고, 구현한다.

본 논문은 2 장에서 기존의 웹 사전들을 분석하여 보고, 3 장에서 기존의 웹 사전들이 갖는 문제점을 해결하기 위한 웹 사전 프레임워크를 설명하고, 구현한다. 4 장에서는 이 논문의 결론과 향후 연구를 살펴본다.

2. 관련 연구

본 장에서는 기존의 웹 사전을 제공하는 검색 방법 및 특징, 개발언어 및 사전 내용 저장 방식, 관리도구 지원 측면으로 분석해본다.

2.1 검색방법 및 특징

기존의 웹 사전 대부분은 검색엔진 형태의 단순질의 검색과 인덱스를 이용한 인덱스 검색을 제공한다. 웹 사전들 중 일부만이 사전 내용들을 주제별로 분류하여 찾는 영역을 줄임으로써 빠르고 정확하게 검색하는 카테고리 검색이나 유의어 및 근접어를 이용한 검색을 지원하고, 최근 추가 단어와 자주 찾는 단어 리스트를 이용한 검색을 제공한다. 표 1 은 기존의 웹 사전이 제공하는 검색 방법과 각각의 특징을 나타낸다[1, 2, 3]. 현재의 웹 사전은 소규모이고 사용자의 수준에 따른 검색이 아니라 단순한 형태의 검색방법만을 제공한다.

이름	단순질의 / 인덱스	특징
		URL
네이버 영어사전	🖃 / ρ	근접어, 유사어 검색
		http://dic.naver.com
로저사전	🖃 / 🔧	콤보박스를 이용한 정확도 조정, CGI
		http://user.chollian.net/~roser/
매일경제 용어사전	🖃 / 🖃	최근 추가 단어 및, 인기 단어 리스트 제공, PHP 로 개발
		http://dic.mk.co.kr/dic_index.html
콜린스 영영사전	🖃 / 🔧	Java Applet 으로 개발
		http://springbank.linguistics.ruhr-uni-bochum.de:8099/ccsd-set.html
야후 백과사전	🖃 / 🔧	카테고리 검색 제공, 검색 결과에 관련 용어 및 카테고리 정보 제공, 확장 검색 기능 제공, PHP 로 개발
		http://kr.encycl.yahoo.com
IT 용어사전	🖃 / 🖃	최근 추가 용어 리스트 제공, PHP 로 개발
		http://www.etimesi.com/db/word/index.html
돌도끼	🖃 / 🔧	사용자 단어 추가/수정 기능 제공, PHP 로 개발
		http://dic.doldoki.org
Webopedia	🖃 / 🔧	카테고리 검색 제공, 검색 결과에 관련 용어 및 카테고리 정보 제공, 내부 링크 제공, HTML 파일 I/O
		http://www.webopedia.com
Britannica Eng	🖃 / 🖃	단순질의와 인덱스 검색 기능을 통합, ASP 로 개발
		http://www.eb.com/limited_search.html
Britannica Kor	🖃 / 🖃	검색 결과에 관련 용어 및 관련 사이트 정보 제공, ASP 로 개발
		http://preview.britannica.co.kr/
텀즈 용어사전	🖃 / 🖃	카테고리 서비스 제공, 관련 사이트 정보, 부가 정보 및 새로 추가한 용어 리스트 제공, 다른 용어 참조, 검색엔진과 연동, PHP 로 개발
		http://www.terms.co.kr

표 1 기존 웹 사전들의 검색 방법 및 특징

2.2 개발 언어 및 사전 내용 저장 방식

기존의 웹 사전은 대부분 HTML, CGI, PHP, ASP 등으로 개발되었다[1, 2, 3]. HTML 이나 CGI 로 개발된 웹 사전은 서비스 요청을 프로세스 수준에 처리하기 때문에 대규모 웹 사전에는 적합하지 않다. PHP 와 ASP 로 개발된 웹 사전은 서비스 요청을 쓰레드 수준에서 처리하기 때문에 대규모 웹 사전 구축에 사용할 수 있지만, PHP 의 경우 스크립트 언어로 다양한 서비스 개발이 어렵고, ASP 의 경우 플랫폼에 종속적이다. 최근에는 Java Servlet 또는 JSP 와 같은 언어로도 개발되고 있다[4].

사전 내용 저장 방식을 살펴보면 HTML 문서를 파일 입출력하는 방식과 관계형 데이터베이스를 이용하여 질의하는 방식이 있다. 또한 객체지향 데이터베이스를 이용한 방식이 있지만 널리 사용되지는 않고 있다. HTML 은 간단하게 웹 서비스 문서를 작성할 수 있지만 문서 구조 서술에 제약이 많고, 효과적인 문서 검색이 어렵다는 단점이 있다. 이를 해결하기 위해 인터넷 표준언어인 XML 이 제안되었다. XML 은 문서 구조 서술에 제약이 없고 데이터베이스화하기가 쉬우며 효과적으로 문서를 검색할 수 있다는 장점이 있다[5].

2.3 사전 관리 도구

기존의 웹 사전은 사전 내용을 작성, 인덱스 검색을 위한 인덱스 구축 및 용어간 참조를 생성하는데 관리자가 텍스트에디터나 웹 저작도구를 이용하기 때문에 웹 사전의 효율적인 관리에 어려움이 많다.

3. 웹 사전 프레임워크

3.1 제공 기능

기존의 웹 사전은 대부분 단순한 형태의 검색 방법만을 제공하고, 저장형식으로 HTML 을 사용하기 때문에 사전데이터의 재사용이 어려우며, 개발언어도 ASP 나 CGI 등을 사용하고 있다. 아울러 사전 내용 편집, 인덱스 생성, 참조 생성 작업이 전용의 도구가 없기 때문에 관리가 매우 어렵다.

본 논문에서는 XML 기반 웹 사전 프레임워크를 제안한다. 제안한 XML 기반 웹 사전은 단순질의 검색, 인덱스 검색을 지원하고, 사용자 편의를 위해 카테고리 검색, 히스토리 검색 및 참조를 통한 검색 등 다양한 검색방법을 제공한다. 또한 관리자를 위해 사전 내용, 인덱스 및 자동 참조 관리도구를 제공한다. 사전데이터는 재사용을 위해 카테고리 정보 및 인덱스 정보를 XML DOM[6] 객체로 데이터베이스에 저장한다. 아울러 개발언어로는 플랫폼에 독립적인 Java Servlet 을 사용하였고, 대규모 웹 사전 구축에 필요한 세션 및 분산 트랜잭션 관리를 제공하기 위해 웹 어플리케이션 서버와 연동한다[7].

3.2 시스템 구성

그림 1 은 본 논문에서 제안한 웹 사전 프레임워크의 구성도를 나타낸다.

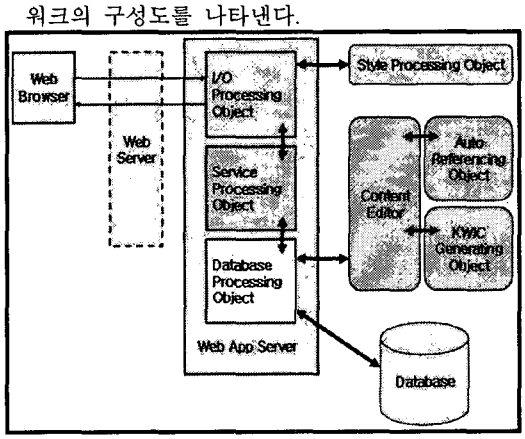

그림 1 웹 사전 프레임워크의 시스템 구성도

IOPO(I/O Processing Object)는 웹 브라우저 요청을 SPO(Service Processing Object)에게 전달하고, 처리된 결과에 StPO(Style Processing Object)와 연동하

여 스타일 적용한 후 웹 브라우저에게 반환한다.

SPO 는 HTTP 요청을 분석하고 처리한 뒤 DPO(Database Processing Object)에게 전달한다. 또한 DPO 가 반환한 결과를 이용하여 카테고리 검색을 위한 트리 인터페이스를 생성하거나 검색하고, 히스토리 객체를 이용하여 히스토리 서비스를 제공한다.

DPO 는 SPO 가 분석한 요청을 이용하여 질의문을 생성하고, 데이터베이스에 질의하여 결과를 SPO 에게 전달한다. 또한 사전 내용 편집, 자동 인덱스 구성, 참조 생성을 처리하는 CE(Content Editor)의 요청을 처리한다.

CE 는 DPO 를 이용하여 데이터베이스 내부에 존재하는 사전 내용 및 카테고리 정보를 검색, 수정, 삭제하고, ARO(Auto Referencing Object)와 연동하여 사전 내용에 참조를 생성하며, KGO(KWIC Generating Object)를 이용하여 키워드 인덱스를 구성한다.

3.3 구현

제안한 웹 사전 프레임워크의 제공 기능을 토대로 소프트웨어 용어 사전을 구현하였다. 개발환경은 다음과 같다.

- 운영체제 : Windows 95/98/2000/NT 및 Linux
- 구현언어 : Java
- 데이터베이스 : Oracle 8.1.5 이상
- 웹 어플리케이션 서버 : Enhydra 3.1 이상
- 웹 서버 : Apache 1.3.2 및 IIS4.0

그림 2 는 Content Editor 의 화면 구성을 나타낸다.

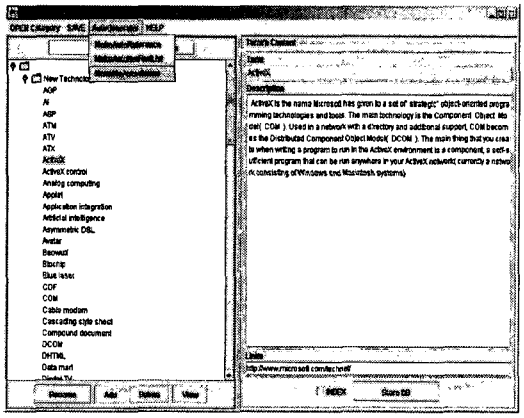

그림 2 Content Editor 화면 구성

카테고리 검색을 지원하기 위해서는 용어를 주제별로 분류하는 작업이 필수적이다. 분류에 따라 관리자는 Content Editor 의 좌측 트리 인터페이스를 이용하여 용어 노드를 추가, 삭제, 변경하고, 내용을 확인을 한다. 카테고리 정보는 내부적으로는 XML DOM 객체이다. DOM 객체는 초기 로딩 시간이 길지만 메모리에 상주하기 때문에 자바스크립트를 사용하는 것보다 매우 빠르다[6]. 구현한 시스템은 카테고리 및 인덱스 정보를 XML DOM 객체로 저장하였다. XML DTD 에 따라 설계된 우측 인터페이스를 통해서 문서를 구성하는 용어 이름, 용어 내용, 관련 사이트 정보를 쉽게 수정하고 저장할 수 있다. 저장버튼을 누르면 DPO 를 호출하여 카테고리 정보를 나타내는 DOM 객체를 저장하고, 데이터베이스 내부에 존재하는 용어에 대한 참조를 자동으로 생성한다.

또한 CE 는 KGO 를 호출하는 메뉴를 이용하여 트리 형태의 인덱스를 자동으로 구성하고, ARO 를 호출하는 메뉴를 이용하여 자동 참조를 생성한다. 자동 참조는 용어 내용에 포함된 단어 중 데이터베이스에 용어로 등록된 단어를 참조할 수 있도록 하이퍼링크로 연결한다. 자동 참조 생성 시 성능 향상을 위해 용어 내용을 분석하여 불용어 리스트를 생성한다. 관리자는 필요에 따라 인덱스 구성, 참조 생성 및 불용어 리스트 구성 작업을 수행한다.

그림 3 는 카테고리 검색을 이용하여 검색한 결과를 나타낸다. 결과 화면은 DPO, SPO 가 처리한 결과를 IOPO 가 XSL[8] 혹은 CSS[9] 형태의 스타일을 적용하여 나타난 결과이다. 즉 사전 내용과 표현을 분리함으로써 같은 내용을 다양하게 표현할 수 있다. 우측 상단에 나타난 'HTML', 'CGI'는 자동으로 생성된 참조로 선택하면 해당 단어내용이 새로운 창에 나타난다. 또한 우측 하단의 'whatis'는 관련 사이트 정보로 선택하면 관련 사이트가 화면에 나타난다.

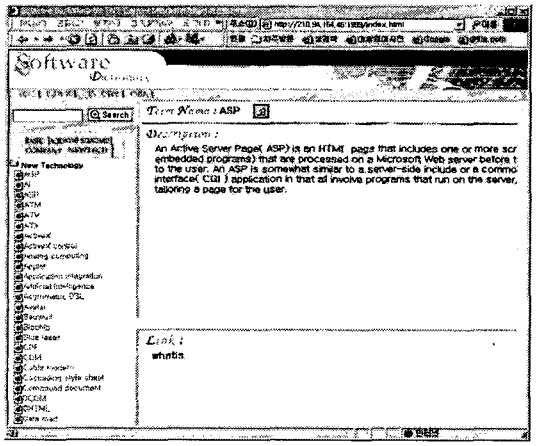

그림 3 검색 화면 구성 및 카테고리 검색 결과

그림 4 는 인덱스 검색을 이용하여 검색한 결과를 나타낸다. 기존의 웹 사전들은 나열 형태의 인덱스 검색을 제공하였지만 구현한 시스템에서는 트리 형태의 인덱스를 자동으로 구성하여 검색을 제공한다. 트리 형태의 인덱스는 유사한 단어로 시작하는 인덱스들에 연관성을 부여한다.

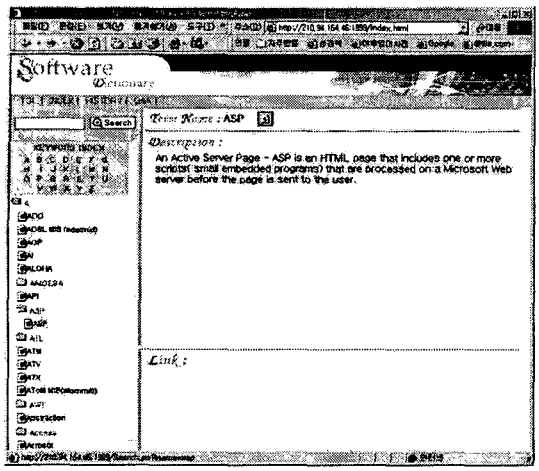

그림 4 트리 형태의 인덱스 검색

그림 5 는 단순질의 검색과 히스토리 검색 창을 나타낸다. 단순질의 검색은 텍스트 상자에 입력한 단어를 포함하는 용어를 데이터베이스로부터 검색한다. 같은 단어가 여러 카테고리에 존재할 수 있기 때문에 다중 결과인 경우 카테고리 정보를 제공하여 사용자가 올바른 선택을 할 수 있도록 하였다. 또한 히스토리 검색은 세션 정보를 이용하여 세션동안 검색한 단어의 리스트를 제공하고 검색할 수 있도록 한다. 최근에 찾은 단어가 리스트의 최상위에 위치한다.

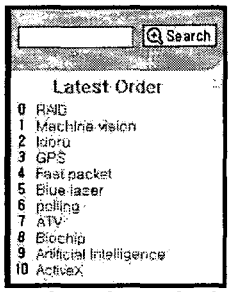

그림 5 히스토리 검색

4. 결론 및 향후 연구

웹이 널리 사용됨에 따라 웹 사전에 대한 사용자의 수요가 증가하였지만 기존의 웹 사전으로는 사용자와 관리자의 요구를 충족시킬 수 없었다.

본 논문에서 제안하고 구현한 웹 사전 프레임워크는 다양한 검색 기능과 사용자 수준에 따른 검색 기능을 제공하고, 인덱스, 참조 생성 및 사전 내용 처리를 위한 효율적인 도구를 제공하며, 사전 데이터를 XML DOM 으로 표현하고 저장함으로써 재사용을 가능하도록 하였다. 또한 사전 내용과 스타일을 분리하여 다양한 결과 표현을 제공한다.

향후 연구로는 사전내용을 XML DOM 으로 생성하여 관계형 데이터베이스에 저장할 때 발생하는 메모리 관리 문제에 대한 연구가 필요하다. 또한 질의 용어의 위치를 고려한 검색과 검색 범위를 제한하여 검색 하는 기능이 추가되어야 하고, 현재 시스템에서 지원하는 한글 질의 처리뿐 아니라 한글 인덱스나 참조를 자동으로 구성하는 부분에 대한 보완이 필요하다.

참고문헌

[1] http://www.whatis.com
[2] http://www.terms.co.kr
[3] http://www.webopedia.com
[4] HALL, "Core Servlets and Java Server Pages," pp7-12, Prantice Hall, 2000.
[5] W3C, "Extensible Markup Language (XML) 1.0 (Second Edition)," http://www.w3.org/TR/2000/REC-xml-20001006, 2000.
[6] W3C, "Document Object Model (DOM) Level 3 Core Specification Version1.0," http://www.w3.org/TR/2002/WD-DOM-Level-3-Core-20020114, 2002
[7] http://www.enhydra.org
[8] W3C, "eXensible Stylesheet Language(XSL) Version 1.0," http://www.w3.org/TR/2001/REC-xsl-20011015/, 2001.
[9] W3C, "Cascading Style Sheets(CSS) level 1.0," http://www.w3.org/TR/REC-CSS1-19990111, 1999.