삼성전자사장상

SGML / XML 문서검색 시스템

1. 소프트웨어 명

SGML/XML 문서검색 시스템

2. 제작자

충남대학교 컴퓨터공학과 신동욱 교수
주소: 대전광역시 유성구 궁동 220
전화: 042-821-6657　　전자우편: shin@comeng.chungnam.ac.kr

3. 소프트웨어 전체 요약 설명

본 소프트웨어는 SGML (Standard Generalized Markup Language) 및 XML (eXtensible Markup Language)문서들을 대상으로 효율적으로 색인하고 사용자가 Web 브라우저나 전용 브라우저를 통하여 검색할 수 있도록 하는 시스템이다. SGML 및 XML은 문서의 구조를 표현하는 Markup 언어로서 디지털 도서관 및 WWW (World Wide Web) 에서 표준으로 정착하고 있다. 본 시스템의 구성 및 기능은 아래에 설명되어 있다.

3.1 구성 및 사용자 인터페이스

본 시스템은 그림 1 에서와 같이 서버와 클라이언트로 구성되어 있으며 서버는 다시 SGML/XML 자동 색인기와 검색기로 구성되어 있다.

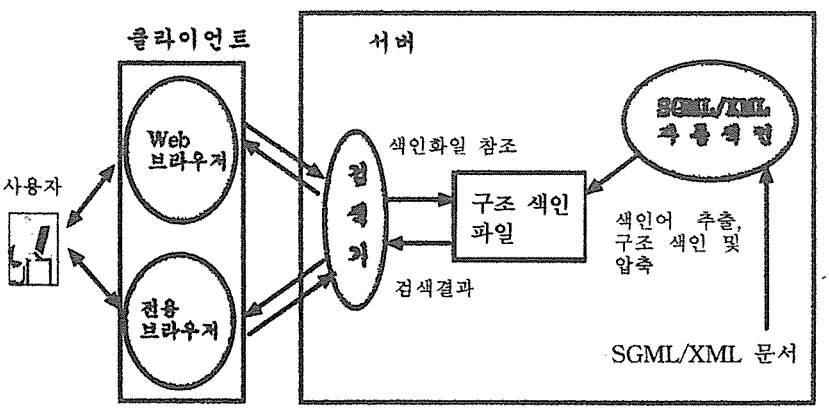

그림 1. 전체 시스템 구성도

먼저 서버에서 SGML/XML 자동색인기는 SGML/XML로 작성된 문서들로 부터 색인어를 자동으로 추출하고 문서들의 구조들을 인식하여 구조 색인 화일을 만드는데 이 과정에서 압축 (compression) 기법을 이용하여 효율적으로 색인 화일을 만든다. 구조 색인 화일을 만드는 데에는 본인이 고안한 BUS (Bottom Up Scheme) 를 이용하는데 이 기법을 이용하면 문서 구조가 아무리 복잡할 지라도 문서의 최하위 노드에서만 색인하면 되므로 매우 효과적으로 색인할 수 있다.

두번째로 검색기는 사용자가 접속할 때마다 서버에 하나씩 생기는데 이 검색기는 사용자 질의어를 파싱하고 구조 색인 화일을 접근하여 해당되는 문서 엘리먼트 (element (SGML/XML에서 사용되는 용어로 문서 구조의 각 노드를 말함)) 들을 결과로 제시한다. 이때 검색기는 사용자 질의어와 엘리먼트들 사이에 유사도 (similarity)를 측정하여 이 것을 기준으로 엘리먼트들을 랭킹 (ranking)하여 사용자에게 제시한다.

마지막으로 클라이언트는 Web 환경에서 동작할 수 있는 사용자 인터페이스와 일반 PC에서 동작되는 전용 브라우저로 구성되어 있다. 다음은 Web 환경에서 동작할 수 있는 사용자 인터페이스를 보여주고 있는데 주 윈도우는 다음과 같은 형태로 되어 있다.

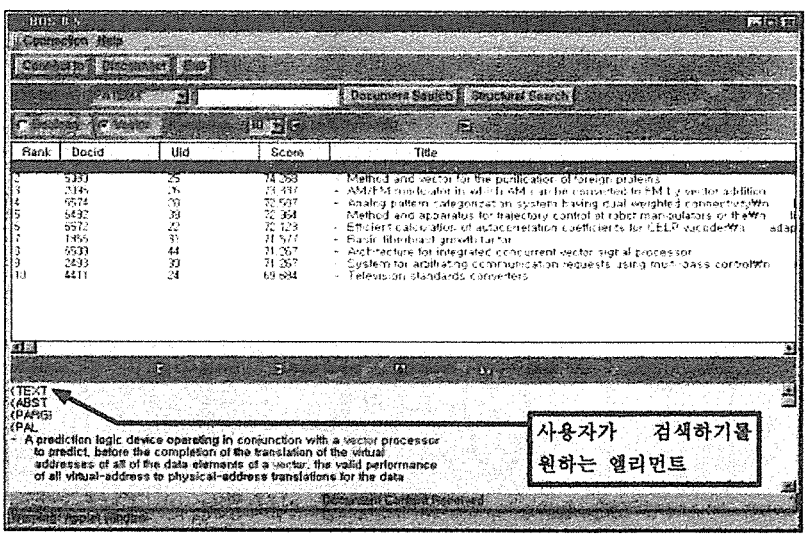

그림 2. Web 브라우저로 호출된 주 윈도우

우선 사용자가 Web 브라우저에서 다음 주소를 접근하여 BUS를 클릭하면

그림 2와 같은 화면이 나타나게 된다. 이때 사용자는 JDK 1.1.5 가 가능한 Web 브라우저를 사용하여야 한다.

http://savage.comeng.chungnam.ac.kr/~sgml

사용자가 "connect to"라는 버튼을 누르면 서버로 접속하여 검색기 프로세스가 하나씩 생기게 된다. 이때 서버에서 서비스하는 데이타베이스 이름들이 "DB select" 칸에 보여지게 된다. 그 다음에 사용자는 질의어를 입력할 수 있는데 원하는 질의어를 "Document Search" 혹은 "Structural Search" 중에 한가지를 선택하여 입력할 수 있다. 이중에서 Document Search는 기존의 검색 엔진과 같이 문서 단위로 검색할 수 있는 것이고 Structural Search는 임의의 엘리먼트 (예 초록, 장, 절, 문단) 에 조건을 주고 이 조건을 만족하는 엘리먼트들을 검색하도록 하는 것이다. Document search를 원할 경우에는 옆에 있는 창에 질의어를 입력하면 되고 Structural Search 를 원할 경우에는 "Structural Search" 버튼을 누르면 그림 3과 같은 윈도우가 나타난다.

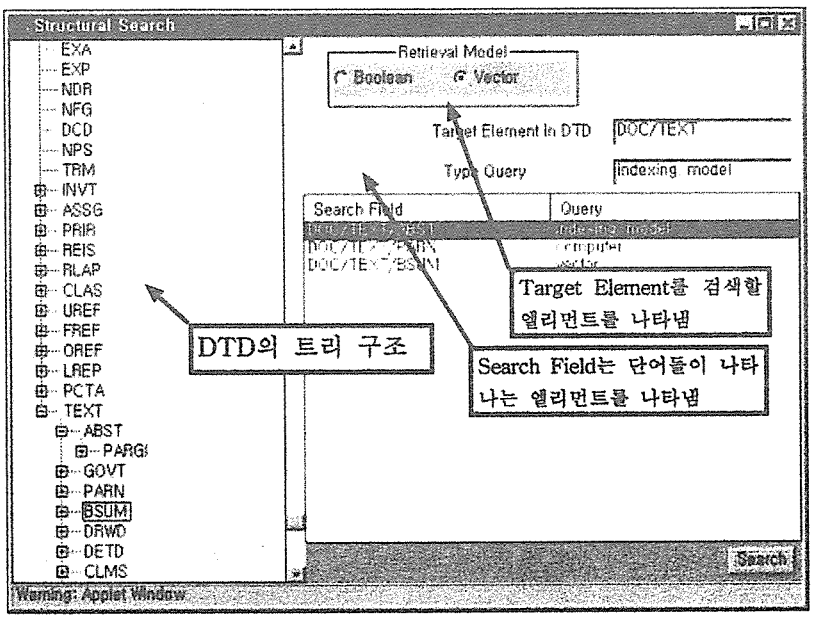

그림 3. Structural Search 윈도우

위의 윈도우에서 왼쪽 화면은 DTD (Document Type Definition) 에 정의된 문서 구조를 트리 형태로 보여주고 있다. 사용자는 이 트리를

Microsoft 탐색기와 동일한 형태로 탐색할 수 있으며 원하는 엘리먼트를 마우스의 오른편 버튼을 클릭하여 선택할 수 있다. 위에서 사용자가 입력한 질의어는 "DOC/TEXT/ABST 필드에 indexing과 model 이 나오고 DOC/TEXT/PARN에 computer가 나오며 DOC/TEXT/BSUM 에 vector가 나오는 TEXT를 찾아라" 이다. 이때 Target Element 는 사용자가 검색하기를 원하는 TEXT 이며 Search Field는 단어들이 들어 있는 ABST, PARN 과 BSUM 이 된다. 사용자는 검색어를 입력할때 Boolean 모델 혹은 벡터 공간 (Vector Space) 모델을 이용할 수 있는데 Boolean 모델을 이용하면 AND, OR, NOT 연산자를 이용할 수 있고 벡터 공간 모델을 이용하면 단어들을 "," 로 연결하여 입력할 수 있다.

위와 같이 질의어를 주면 검색기가 색인 화일을 참조하여 질의어에 적합한 (relevant) 엘리먼트들을 검색하는데 그 결과가 그림 2의 가운데 창에 나타나게 된다. 이 창에는 검색된 엘리먼트의 고유 번호 (DID(Document Identifier) 와 UID (Unique Element Identifier)) 와 유사도 (Score로 표현됨) 및 축약된 정보가 나타나게 된다.

사용자가 검색된 엘리먼트를 더블 클릭 (double click) 하면 해당하는 엘리먼트의 전문 (full text)이 맨 아래 창에 나타나는데 이때 사용자가 입력한 단어들은 빨간 색으로 하일라이트되어 나타난다. 또한 사용자가 검색된 엘리먼트의 부모 (parent), 형제 (sibling) 및 자식 (child) 노드를 탐색하고 싶으면 "Left Sibling", "Right Sibling", "Parent" 와 "First Child" 버튼을 이용하여 볼 수 있다.

일반 PC에서 동작되는 전용 브라우저는 위와 비슷한 형태를 가지고 있는데 엘리먼트의 구조를 좀더 알기 쉽게 보여 준다.

3.2 기능및 성능

본 시스템은 SGML 혹은 XML로 작성된 문서들을 다양하게 검색할 수 있도록 하는 기능을 가지고 있다. 기존의 검색 엔진들은 단순히 원하는 단어들이 포함된 문서들을 검색할 수 있는데 본 시스템은 다양한 구조 질의들을 지원한다. 예를 들어 기존의 검색 시스템들에서는 "SGML 과 검색이 포함된 문서들을 찾아라" 라는 질의어들을 지원하는데 비하여 본 시스템에서는 "SGML과 검색이 장의 제목 (chapter heading)에 나타나고 색인과 압축이 임의의 문단 (paragraph) 에 나타나는 장 (chapter)을 찾아라" 와 같은 질의어를 지원하고 있다. 이들을 포함하여 본 시스템이 지원하

는 기능을 나열하면 다음과 같다.

1. Boolean 및 벡터 공간 검색 기능
2. 다양한 구조 질의어 지원
3. 절단 검색 (*) 지원 (전방, 후방, 중위)
4. 한글 처리 기능
5. 압축 기능
6. 엘리먼트 사이 항해 (navigation) 기능

본 시스템의 색인 오버헤드는 한글 및 영문 데이터를 압축을 한 후의 결과가 20% - 35% 정도이다. 이 정도의 성능은 대표적인 외국 제품인 Inso (http://www.inso.com)의 Dynatext와 비교할 때 비슷한 오버헤드를 가지나 본 시스템에는 Dynatext에서 고려하지 않은 몇가지 색인 정보들이 더 들어가 있고 또 새로운 엘리먼트가 추가되거나 삭제될 때에 고쳐야 하는 부분이 적어 Dynatext보다도 우수하다. 검색 시간도 빨라 보통 1초 이내이다. 본 시스템이 제공하는 기능과 성능을 현재 상품화된 외국 제품들과 비교해 볼 때 전혀 떨어지지 않고 오히려 부분적으로 우수한 측면을 가지고 있다.

4. 개발 기간 및 공수

개발기간: 2년, 공수: 약 80 man-month

5. 사용 또는 개발 언어, TOOL

C/C++ (서버), Java (클라이언트)

6. 사용 시스템

서버: Sun Ultra Sparc/Solaris, 클라이언트: IBM PC/Windows 95

7. 직접 효과

본 시스템은 SGML과 XML로 작성된 문서들을 효과적으로 색인하고 검색하는 엔진이다. 최근의 추세를 보면 전세계적으로 전자 도서관을 구축하는데 있어 문서 표현은 SGML로 하는 것이 표준으로 되어 가고 있으며 W3C consortium에서 SGML을 WWW에서 사용할 목적으로 XML을 HTML과 같이 문서 표준으로 추천한데에 힘입어 상당한 주목을 받고 있다. 이에 따라 SGML/XML은 회사 등에서 인트라넷 (Intranet)을 구축하는 유력한 언어로 떠오르고 있으며 전자문서교환 (Electronic Data Interchange) 등에 중요한 수단으로 각광받고 있다. 특히 가까운 장래에 XML이 CALS/EC 분야

에 핵심이 되리라는데에는 대부분의 사람들이 공감하고 있다. 따라서 SGML 혹은 XML로 작성된 문서들을 어떻게 효과적으로 관리할 것인가가 상당한 이슈로 등장하고 있다.

그러나 이분야의 기술 수준은 가장 앞선 미국에서조차 아직 초보 단계에 있다. 우리나라에서도 일부 업체가 SGML 에디터나 브라우저 등을 상품화한 상태이지만 본 연구에서 개발한 기술과 유사한 기술을 보유하거나 상품화하지 못하고 있다. 따라서 이러한 기술을 조기에 상품화할 경우에 우리 업체가 세계 시장에서 외국 제품보다 우수한 성능으로 경쟁하고 또 세계 시장의 일부를 석권할 수 있는 계기가 될 것이다.

또한 현재 우리나라에서도 법률, 특허, 의학 정보들이 CD-ROM 으로 제작되어 공급되고 있는 실정인데 아직까지는 이들 정보들을 효과적으로 서비스하고 있지 못하다. 그런데 본 시스템을 이용할 경우에는 대량의 데이터를 작은 공간에 효과적으로 저장할 수 있을뿐만 아니라 다양한 검색 기능을 사용자에게 제공할 수 있어 매우 경쟁력이 높을 것이다.

8. 간접효과

최근에 국내 도서관들이 대부분 전산화를 추진하고 있으나 그 전산화 수준은 도서관에서 보유한 장서들의 서지 정보 (bibliographic information) 들을 저장하고 사용자에게 서비스하는 수준에 그치고 있다. 따라서 본 시스템이 상품화되어 서비스되면 국내 또는 국외의 도서관에 저장된 문서들의 구조를 자유롭게 검색할 수 있어 사용자들이 보다 편리하게 문서들을 검색할 수 있다.

9. 기타

본 시스템은 현재 충남대학교 소프트웨어 연구센터 (http://sorec.chungnam.ac.kr) 의 지원으로 상품화를 추진하고 있다.