제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

다국어를 지원하는 XML 문서 검색 시스템: HyREX

한예지, 채종대, 김수희
호서대학교 컴퓨터공학과
e-mail:shkim@office.hoseo.ac.kr

HyREX: Universal XML Retrieval Engine for XML

Ye-Ji Han, Jong-Dae CHae, Su-Hee Kim
Department of Computer Engineering, Hoseo University

요 약

HyREX는 연구용 프로토타입 XML 하이퍼미디어 문서 검색시스템으로 다국어를 지원하고 있다. HyREX는 검색을 위한 효율적인 접근 경로들을 처리하는 물리적 계층 HyPath와 질의어를 처리하는 논리적 계층 XIRQL 그리고 사용자 인터페이스인 HyGate 계층으로 이루어져 있다. 이 연구에서는 영어와 독일어 등의 검색을 지원하는 기존의 HyREX 시스템을 한글 XML 문서 검색시스템으로 확장하기 위해 먼저 한글 데이터타입을 위한 클래스를 구현하였다. 앞으로 한글 XML 문서 검색에서 정확율과 재현율을 향상하기 위해 각 문서의 인덱스에 대해 *tf · idf* 공식을 이용하여 가중치를 부여하고 이를 개발하고자 한다.

1. 서론

인터넷 사용과 정보의 양이 기하급수적으로 증가함에 따라 자료 검색의 효율성을 위해 검색엔진이 출현하게 되었다. 웹에 분산되어 있는 정보를 표현하는 수단으로 최근까지 가장 많이 사용되어온 언어인 HTML은 제한된 태그의 사용과 문서 자체가 구조화 되어 있지 않아 효과적인 검색이 어렵다는 단점이 있다[1]. 이에 웹 발전의 주도적인 역할을 하고 있는 W3C(World Wide Web Consortium)에서는 유연한 정보 표현 능력을 가진 확장 가능한 마크업 언어인 XML을 제안하였다[2].

XML이 차세대 웹 문서의 표준으로 급부상하면서 이를 효과적으로 관리하기 위한 XML 관련 연구들도 늘어나고 있다. 그 예로는 XML 저장 관리 시스템 개발에 대한 연구(XML 문서들의 저장, 관리 및 검색), XML 관련 질의어(검색, 링크)에 관한 연구, 기존 데이터베이스 시스템에 저장된 데이터를 XML 문서로 변환하는 도구 개발에 관한 연구 등이 있다[3].

독일 University of Dortmend 컴퓨터학과의 Dr. Norbert Fuhr가 이끄는 정보검색(IR) 그룹에서는 XML 형태로 표현된 데이터를 효율적으로 검색하기 위한 포괄적인 기술을 개발하기 위해, 다음과 같은 연구 목표를 설정하여 수행 중에 있다.

- XML 문서를 대상으로 한 질의와 질의에 대한 검색 문서의 랭킹을 제공하는 질의 처리
- 북마크, 디렉토리, 온톨로지를 포함하는 XML 정보 소스에 대한 의미론적인 메타데이터 개발
- XML 데이터를 메타데이터 구조로 구성하기 위한 자동적인 분류
- 사용자, 의미론적 메타데이터, 관련 정도의 피드백을 기초로 한 질의의 확장과 개선에 대한 기술 개발

위의 연구와 병행하여, XML 문서를 대상으로 사용자들이 원하는 정보를 편리하게 검색할 수 있도록 하는 검색엔진 HyREX(Hypermedia Retrieval Engine for XML)를 개발하고 있다.

이 연구에서는 HyREX를 한글 XML 문서를 처리할 수 있는 검색엔진으로 확장하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 XML 검색 기법과 검색 질의어에 대한 기존 연구들을 살펴본다. 3장에서는 HyREX의 특징과 구조에 대해, 4장에서는 HyREX를 이용한 한글 XML 문서 검색에 대해 소개한다. 마지막으로 5장에서는 결론 및 향후 연구방향을 제시한다.

2. 관련 연구

2.1 XML 검색 기법

XML 문서는 보통 엘리먼트(태그)로 구성된다. 그러므로 정보검색 시스템에서는 문서 단위 뿐만 아니라 임의 깊이에 존재하는 엘리먼트 단위의 검색도 가능해야 하며 엘리먼트를 설명하는 속성 애트리뷰트 값에 대한 검색도 가능해야 한다.

위와 같은 조건을 만족하기 위해 XML 문서에 대한 검색은 내용 기반 검색, 구조 기반 검색, 내용/구조 기반(혼합) 검색, 애트리뷰트 기반 검색으로 구분된다[4].

내용 기반 검색은 주어진 키워드와 관련 있는 문서 또는 엘리먼트 내용을 검색하는 것으로 검색 대상은 #PCDATA(글자데이터) 값이다. 구조 기반 검색은 엘리먼트 이름(태그)을 기반으로 검색하는 것으로 내용을 배제한 문서의 논리적 구조(DTD)와 관련이 있다. 내용/구조 기반 검색은 내용과 구조 정보를 이용하여 검색하는 것으로 검색 대상은 논리적 구조를 고려한 임의의 엘리먼트의 내용이다. 마지막으로 애트리뷰트 기반 검색은 주어진 애트리뷰트 이름과 값을 검색하는 것이다.

2.2 XML 검색 질의어

현재 XML 검색 질의어는 1998년에 제안된 XQL와 XML-QL, 1999년에 제안된 XPath, 2001년에 제안된 XQuery 등이 있으며 이들은 모두 앞에서 소개한 내용/구조 기반 검색이 가능하도록 지원하고 있다. [3]에서는 XML 검색 질의어들의 주요 특징을 <표 1>과 같이 비교하고 있다.

제일 처음 제안된 XQL(XML Query Language)[5]은 SQL 데이터베이스에서 질의하는 방식으로 XML 데이터를 질의할 수 있도록 개발되었으며 XSL 내에서 사용되는 형식 구문에 기반을 두고 있어 XSL의 확장판으로서 제안되고 있는 질의어이다. XML-QL[6]은 XML 문서에서 데이터를 추출, 통합 및 다중 XML 문서간의 통합에 주로 사용되는 질의어이다. XPath[7]는 XML 문서의 계층을 탐색하는데 경로 표기법을 사용하며 XML 문서에 대해 각 엘리먼트와 애트리뷰트(노드) 지정과 문서 주소 지정을 위한 포괄적인 질의어이다. XQuery[8]는 Quilt에 기반한 질의어로 순수 XML 측면(XML이 데이타베이스, 파일, 메시지 등 어디에 있던 관계없이)에서 XML 질의 기능을 처리한다는 장점이 있다.

비교항목	XQL	XML-QL	XPath	XQuery
제안	W3C			
질의 구조 형태	경로표현, 필터 연산자, 비교연산자 등	WHERE-CONSTRUCT 구조	위치경로	경로 표현, 엘리먼트 생성자, FLWR(FOR, LET, WHERE, RETURN) 표현식
결과 형태	노드, 노드리스트, XML문서, 배열, 기타구조	XML 데이터	노드의 집합	XML 데이터
지원하는 검색 기법	내용/구조 기반 검색			
질의 가능 영역	한 문서 또는 XML 저장소 내 모든 문서에 대한 질의 지원			
문서 간의 연산	지원하지 않음	지원	지원하지 않음	지원
검색 결과의 후처리 가능 여부	불가	가능	불가	가능
링크 지원 여부	지원하지 않음			

<표 1> XML 관련 질의어 비교

3. HyREX

3.1 HyREX의 특징

HyREX[9,10]의 주된 특징은 그 이름으로 설명될 수 있다.

- hyper

HyREX는 사용자에게 명시적 링크와 묵시적인 링크를 제공한다.

- media

HyREX는텍스트를 포함하는 다양한 미디어로 표현되는 정보에 대한 검색이 용이하다.

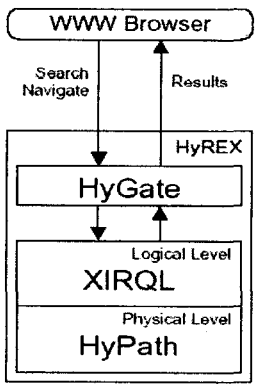
- retrieval engine

  XML을 통하여 사용할 수 있는 모든 종류의 정보구조를 검색하는 것이 가능하다.
- XML

  XML 문서의 내용과 내재하는 문서구조를 고려한 검색이 가능하다.

3.2 HyREX의 구조

HyREX의 구조는 논리적/물리적 레벨 사이의 분리가 명확하다는 점에서 데이터베이스 관리 시스템의 구조와 유사하다. HyREX는 검색을 위한 효율적인 접근 경로들을 처리하는 물리적 계층 HyPath, 질의어를 처리하는 논리적 계층 XIRQL 그리고 사용자 인터페이스인 HyGate 계층으로 이루어져 있다. <그림 1>은 HyREX의 구조를 간단히 나타내고 있다[9]. HyREX의 세 계층에 대한 특징들을 간단히 다음과 같이 요약할 수 있다.



(그림 1) HyREX 구조

(1) HyGate

가장 상위 레벨에서 사용자가 웹 브라우저를 이용하여 HyREX를 연결하고 필요한 정보를 요청하면 이 요구서가 HyGate에 의해 접수된다. HyGate는 사용자의 요청을 XIRQL 질의로 변환하고 그 처리를 하위 레벨로 위임한다. 그리고 처리결과를 사용자에게 적당하게 표현한다.

- 검색과 브라우징을 위한 사용자 인터페이스
- 사용자의 질의를 XIRQL 질의로 변환
- 검색 결과 표현

(2) XIRQL

DTD에 의하여 문서의 다양한 부분들의 데이터 타입을 명시한다. 이것은 각 문서의 클래스를 표현하기 위한 DDL에서 이루어진다.

- XML 정보 검색 질의어
- 정보 검색 기능을 가지도록 XPath를 확장
- 가중치가 부여된 문서 내용과 질의 조건
- 검색 결과를 위한 랭킹
- 어떤 타입의 정보도 검색가능한 강력한 언어
- 관련성에 기인한 검색

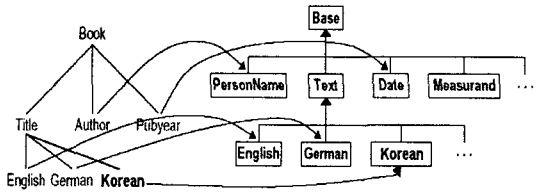
(3) HyPath

HyPath는 문서의 접근 구조를 명시한다. 이것은 DDL 내에서 이루어진다.

- 내용과 구조 기반의 효율적인 접근 경로
- 응용분야 특유의 접근 경로의 선택

(4) HyREX에서의 데이터타입

XML의 마크업은 문서의 논리적 구조를 나타낼 뿐만 아니라 추가로 문서의 의미적 정보를 제공한다. 작성된 문서가 어떤 데이터 타입에 의해 만들어진 것인지 언급될 수 있으며, 이 정보가 검색시에 이용될 수 있다. 특정한 문서 부분에 데이터 타입을 할당하는 것은 <그림 2>에 예시된 것처럼 DTD에 의해 수행된다.



(그림 2) 데이터타입과 문서와의 매핑

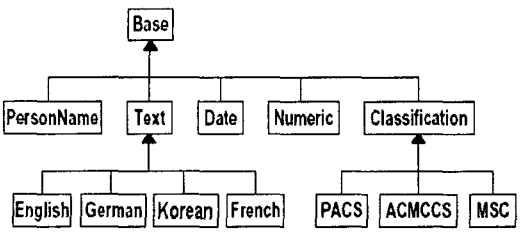
4. HyREX의 확장: 한글 XML 문서 검색 지원

이 연구에서는 한글로 작성된 XML 문서를 검색할 수 있도록 HyREX를 확장하는 데 중점을 둔다.

먼저 한글 문서를 인식할 수 있도록 한글 데이터 타입을 위한 클래스를 만들었고 단어를 분리하는 간단한 모듈을 추가하였다.

다음 단계로 HyREX에 한글의 정보검색의 기능을 지원하기 위해 구문을 분석하여 인덱스를 구축하는 기능을 추가하고자 한다. 이 인덱스를 이용하여 데이터베이스에 있는 XML 문서들과 질의어 간의 유사도를 계산하여 그 값이 높은 순서대로 검색이 되는 기능을 부여하고자 한다. 그러기 위해서는 각 문서의 인덱스에 대한 가중치를 부여하는 모듈이 필요

한 데, 이를 *tf* · *idf* 공식을 이용하여 개발하고자 한다. <그림 3>은 현재 HyREX가 지원하는 데이터타입들을 나타내고 있다.



(그림 3) HyREX가 지원하는 데이터타입들

5. 결론 및 향후 연구

HyREX는 XML 문서 검색을 위한 하이퍼미디어 검색 엔진으로 세 레벨의 스키마로 이루어져 있다. HyREX는 궁극적으로 다양한 언어를 지원하는 XML 검색 엔진을 비전으로 하고 있다. 현재 영어와 독일어를 위한 기본적인 검색 기능이 구현되었으며, 이 두 언어에 대해서도 개발한 모델에 입각하여 계속적인 구현 작업이 필요하다.

현재 HyREX에는 한글 데이터타입을 위한 클래스가 구현되었고 단어를 분리하는 기능을 지원하는 정도로 XML 한글문서의 검색을 지원하기에는 아직 초기단계에 있다. 독일에서 개발 중에 있는 시스템을 한글을 지원하는 시스템으로 확장하기 위해, 이 시스템을 이해하기 위해 많은 어려움을 겪었지만 이젠 전체 아키텍처와 구현을 위한 내부 디자인을 상세히 파악하고 있는 만큼 필요한 기능의 추가에 집중할 수 있게 되었다.

앞으로 한글 구분 분석을 통한 스테밍과 가중치가 부여된 인덱싱 모듈을 추가하는 작업을 수행하여 정확율과 재현율을 향상하는 시스템으로 HyREX를 확장하고자 한다.

참고문헌

[1] 정희경, "차세대 웹 문서 표준 XML", 정보처리 제6권 제3호(1999.5)

[2] http://www.w3.org/, W3C

[3] 문찬호, 강현철, "링크 질의를 통한 XML 문서의 검색 기법", 정보처리학회논문지D 제8-D권 제4호 (2001.8)

[4] 박종관, 손충범, 강형일, 유재수, 이병엽, "XML 문서의 효율적인 구조 검색을 위한 색인 모델", 정보처리학회논문지D 제8-D권 제5호(2001.10)

[5] J. Robie et al., "XML Query Language(XQL)", http://www.w3.org/TandS/QL/QL98/pp/xql.html, 1998

[6] A. Deutsch et al., "XML-QL : A Query Language for XML", http://www.w3.org/TR/NOTE-xml-ql/, 1998

[7] J. Clark and S. DeRose, "XML Path Language (XPath) Version 2.0", http://www.w3.org/TR/xpath20/, 2002

[8] D. Chamberlin et al., "XQuery : A Query Language for XML", http://www.w3.org/TR/xquery/, 2002

[9] Norbert Fuhr, "HyREX", http://ls6-www.cs.uni-dortmund.de/ir/projects/hyrex/, Department of Computer Science, University of Dortmund

[10] Mohammad Abolhassani, Norbert Fuhr, Norbert Gövert, Kai Großjohann, "HyREX : Hypermedia Retrieval Engine for XML", Department of Computer Science, University of Dortmund

[11] Norbert Fuhr, Kai Großjohann, "XIRQL : An XML Query Language Based on Information Retrieval Concets", University of Dortmund, Germany