내용기반 XML 문서의 검색

김수희° 조명찬 한예지
호서대학교 컴퓨터공학과
shkim°@office.hoseo.ac.kr

Information Retrieval from XML Documents based on Contents

Suhee Kim° Myoungchan Cho Yeji Han
Department of Computer Engineering, Hoseo University

요 약

이 연구에서는 XML 문서의 효율적인 검색을 위해 XML 데이터에서 색인어를 추출하고 가중치를 부여하여 내용기반 인덱스를 구축하고, 질의와 문서간의 유사도가 높은 문서들을 사용자에게 제공함으로써 기존의 경로 중심 혹은 패턴매칭 형태의 XML 문서 검색 기능을 확장하고자 한다. 내용기반 검색을 지원하는 XML 문서 검색시스템을 설계하고, 내용기반 검색과 관련한 이슈들을 논의한다. 개발 중에 있는 연구용 프로토타입 시스템을 이용하여 질의에 대한 내용기반 검색 결과를 간단히 소개한다.

1. 서 론

정보화, 지식기반 사회화, 그리고 글로벌화로 대표되는 21세기는 정보통신 기술의 혁명을 통하여 공업산업사회로부터 변천하여 지식정보를 자원화 하는 지식정보사회이다. 인터넷관련 기술의 급속한 발전에 따라 전자도서관, 대규모 기관의 인트라넷, 전자상거래를 비롯하여 다양한 분야의 응용들이 인터넷 기반으로 매우 활성화되고 있다.

1998년 W3C에서는 데이터의 표현과 교환을 위한 표준으로 XML(eXtensible Markup Language)을 채택하였다. XML은 HTML의 단점을 보완하고 SGML의 장점을 반영한 메타언어라 할 수 있다. XML은 인터넷 기반 응용시스템의 구축에 중요한 위치를 차지하고 있다. 머지않아 전자상거래, 전자도서관, 과학적인 데이터의 리포지터리 등 많은 분야의 문서들이 XML 포맷으로 작성될 것이다. 그러나 XML의 풍부한 태그 구조가 오늘날의 탐색엔진에서는 반영되어 있지 않다. 문서 검색 엔진의 다른 측면에서의 향상에도 불구하고 의미적인 레벨에서 적절한 정보 접근 방법들이 여전히 결여되어 있다[1,2,3,4]. 탐색엔진에서 XML 데이터를 효율적으로 처리하기 위해 여러 기술들이 개발되어 왔지만, 그들 각각이 본질적으로 내재한 한계를 가지고 있다.

- 데이터베이스 시스템 관점에서 볼 때, XPATH, XQL, XML-QL 혹은 Quilt와 같은 XML 질의 언어들이 가장 유망한 추세이다[5,6,7,8]. 이 검색 언어들은 XML 문서의 구조와 레이블들을 패턴 매칭과 전통적인 SQL 형태의 논리 조건과 결합하여 이용한다. 그러나 의미적인 유사도에 대한 개념이 없으므로 검색결과를 순위로 나타내는 방법이 존재하지 않는다.
- 정보검색 분야(IR)에서는 최첨단의 웹 탐색 엔진들을 포함한 대부분의 연구들이 XML 타입의 데이터 구조나 XML 문서의 엘리먼트와 애트리뷰트의 레이블, DTD, 스키마에 내재하고 있는 잠재적인 온톨로지 정보를 고려하지 않고 있다.

이 연구에서는 XML 문서의 효율적인 검색을 위해 가중치와 유사도를 이용하여 내용기반 정보검색을 수행하고 질의와 문서간의 유사도가 높은 문서들을 사용자에게 제공함으로써, 기존의 경로 중심 혹은 패턴매칭 형태의 XML 문서 검색 기능을 확장하고자 한다. 이를 위해, XML 데이터 검색시스템을 설계하고, 내용기반 검색과 관련한 이슈들을 논의하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 XML 문서 검색시스템의 설계를, 3절에서는 색인어에 가중치를 부여하는 문제를 다룬다. 4절에서는 질의문과 문서간의 유사도 계산에 대해 논의하며, 5절에서는 개발 중에 있는 시스템을 간단히 소개한다. 마지막으로 결론 및 향후 연구 방향을 제시한다.

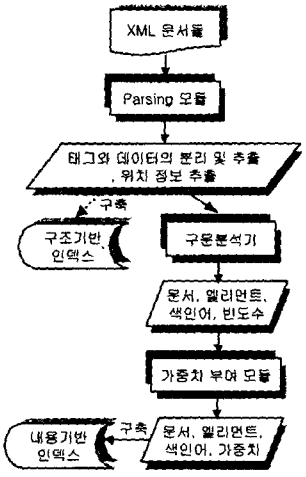
2. XML 문서 검색시스템의 설계

다양한 종류의 XML 문서의 효율적인 검색을 위해서는 문서의 태그를 중심으로 하는 구조기반 검색과 문서의 내용을 중심으로 하는 검색이 모두 지원되어야 한다. 내용기반 XML 문서의 검색은 전통적인 문서 정보검색 기법을 도입하여 응용할 수 있다. XML 문서에서 데이터 부분을 추출하고, 이를 대상으로 색인어를 추출하여 가중치를 부여하고 인덱스를 구축할 수 있다. 전통적인 문서 검색에서는 검색의 효율을 높이기 위해 일반적으로 역화일이나 시그니쳐 파일로 인덱스를 구축한다. <그림 1>과 <그림 2>는 내용기반 검색을 지원하는 검색시스템을 구축하기 위한 개요이다.

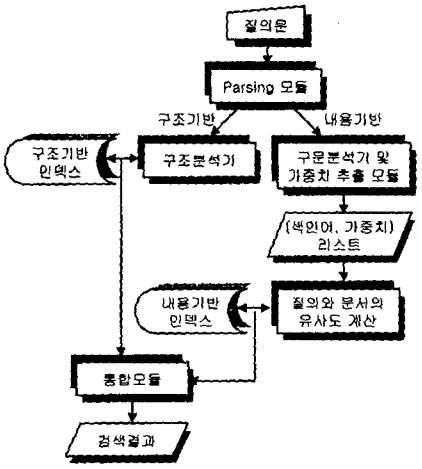
3. 가중치 부여

XML 문서는 데이터만으로 구성된 리프 노드와 데이터와 자식 노드들로 구성된 내부 노드로 이루어진 트리 구조를 가지고 있다.

각 노드에 있는 데이터를 대상으로 구문분석기를 이용하여 주요 색인어들을 추출하고 이들의 빈도수를 계산한다. 각 색인어의 빈도수를 이용하여 가중치를 계산한다.



<그림 1> 인덱스 구축 과정



<그림 2> 문서 검색 과정

3.1 문서 단위와 정보의 검색 단위

하나의 XML 파일을 하나의 XML 문서로 취급하기에는 각 XML 파일이 저장하고 있는 데이터의 양은 천차만별이다. 같은 DTD나 스키마를 가지고 있는 다양한 크기의 파일들이 여러 곳에 분포해 있을 수 있다. 하나의 XML 파일을 논리적인 관점에서 몇 개의 문서로 분할할 수 있다. 이 연구에서는 XML 파일에 있는 태그와 내용을 토대로 XML 문서 단위를 수동으로 지정한다.

리프 노드를 기본 검색 단위로 하고 내부 노드는 그 하부 노드의 정보를 반영하여 검색한다.

3.2 가중치 계산

• 색인어의 가중치

문서 k 의 j 경로에 있는 엘리먼트의 노드에서 색인어 i 의 가중치($w_{i,j,k}$)는 (수식 1)을 적용하여 계산한다.

$$w_{i,j,k} = tf_{i,j,k} * idf_i \quad \text{..... (수식 1)}$$

- $tf_{i,j,k}$: 문서 k 의 j 경로에 있는 엘리먼트의 노드에서 색인어 i 가 나타나는 빈도수
- idf_i : 색인어 i 가 나타나는 문서들의 수에 대한 역 문서 출현 빈도 (*inverse document frequency*),

$$idf_i = \log_e \frac{(N+1)}{df_i}$$

- df_i : 색인어 i 가 나타나는 문서의 빈도수
- N : 전체 문서의 수

• 색인어의 확장된 가중치

내부 노드에서는 그 노드 자체만이 아닌 하부노드에 있는 각 색인어의 가중치를 반영할 필요가 있다. 한 노드에 있는 어떤 색인어의 가중치가 그 노드의 상부 노드들에게 그대로 반영되는 경우, 항상 루트 노드에서 가장 큰 가중치와 높은 랭크를 가지게 되고 이결과는 바람직하지 못하다[9]. 그러므로 하부 노드에 있는 각 색인어의 가중치를 적당한 비율로 감소하여 반영하여야 한다.

각 노드에서 그 하부 노드를 고려한 확장된 가중치는 (수식 2)와 같이 계산될 수 있다.

$$ew_{i,j,k} = w_{i,j,k} + af * w_{i,j,k} - w_{i,j,k} * af * cw_{i,j,k} \quad \text{..... (수식 2)}$$

- $ew_{i,j,k}$: 문서 k 의 j 경로에 있는 엘리먼트의 노드에서 그 노드 자체의 색인어 i 에 대한 가중치와 더불어 그 자식노드들에서 나타나는 색인어 i 에 대한 가중치를 반영한 확장된 가중치
- $cew_{i,j,k}$: 문서 k 의 j 경로에 있는 엘리먼트의 노드에서 그 자식 노드들에서 색인어 i 에 대한 확장된 가중치들의 합
- af : 감소 비율

위 내용으로 리프노드에서 각 색인어의 확장된 가중치 $ew_{i,j,k}$는 $w_{i,j,k}$와 동일하다는 것을 쉽게 알 수 있다.

4. 질의와 문서간의 유사도 계산

3절에서 설명한 방법으로 각 색인어에 가중치를 부여하여 내용(키워드) 기반의 인덱스를 구축할 수 있다. 우리가 목표로 하는 것은 사용자의 질의문에 가장 적절한 정보를 포함하고 있는 엘리먼트나 문서들을 제공하는 것이다.

질의문(Q)과 엘리먼트나 문서와의 유사도를 계산하는 방법들은 여러 가지가 있지만, 여기에서는 두 가중치 벡터의 내적(inner product)으로 유사도를 계산한다. 유사도가 높은 순서대로 데이터베이스에 있는 정보들을 제공할 수 있다.

• 내적을 이용한 유사도의 계산

$$sim(E_{j,k}, Q) = \sum_{i=1}^{n} ew_{i,j,k} * w_{qi} \quad \text{..... (수식 3)}$$

- n : 데이터베이스에 있는 서로 다른 색인어들의 수
- $E_{j,k}$: 문서 k 의 j 경로에 있는 엘리먼트 노드
- w_{qi} : 질의문 Q 에서 색인어 i 의 가중치 비율

5. 시스템 구현

XML 문서를 대상으로 구조기반과 내용기반 검색 기능을 지원하기 위해, 연구용 프로토타입의 시스템을 구현하고 있다.

질의와 데이테베이스에 있는 XML 문서간의 유사도가 높은 문서를 검색하기 위한 주요 개발 과정을 샘플 XML 문서를 이용하여 간단히 소개한다. 내용기반인덱스를 구축하기 위해 역화일을 이용하고, 문서단위를 수동으로 지정한다.

- 샘플 XML 문서의 DTD

샘플로 사용한 문서의 내용은 한글로 된 몇 편의 동시들이다. 이 동시들은 <그림 3>과 같은 DTD로 구성되어 있다. 여기에서는 각 동시(poem)를 독립된 문서로 취급하여 가중치 계산에 이용한다.

```
<!ELEMENT poetry (poem)+>
<!ELEMENT poem (title, author, content)>
<!ELEMENT title      (#PCDATA)>
<!ELEMENT author   (#PCDATA)>
<!ELEMENT content (#PCDATA)>
```

<그림 3> 동시집의 DTD

- 경로와 색인어의 추출

문서내에 있는 각 엘리먼트를 추출하고, 거기에 속해 있는 데이터를 대상으로 색인어를 생성하고 빈도수를 계산한다.

색인어를 생성하기 위해 HAM 6.0을 이용하였다[10]. 그 다음으로 3절에서 소개한 방법으로 각 색인어에 대한 가중치를 계산한다. <그림 4>는 "잔디에 누워"라는 시에서 HAM 6.0을 이용하여 추출한 색인어, 빈도수와 가중치를 나타내고 있다.

<잔디에 누워>

잔잔한 물결처럼 보드라운
잔디에 홀로 누워 하늘을 본다

바다가 저렇겠지 끝없이 푸른
하늘엔 흰 구름이 둥둥 떠가고

외로운 나도야 흰 구름 따라
한없이 한없이 가는 것 같다

⇨

색인어	빈도수	가중치
구름	2	4.030
하늘	2	2.644
물결	1	2.708
나도	1	2.708
바다	1	2.708
잔디	1	2.708

<그림 4> 데이터 및 추출된 색인어

- 유사도 계산

(수식 3)을 이용하여 유사도를 계산하고, 질의에서의 각 키워드의 가중치를 편의상 1.0으로 간주한다.

- 질의 및 검색 결과
  - 질의의 예 : 구름&&하늘&&물결 (복수질의 가능)
  - 검색 결과

순위	파일이름	경로	유사도
1	poetry1.xml	poetry[0]/poem[0]/content[2]	9.383
2	poetry2.xml	poetry[0]/poem[3]/content[2]	6.673
3	poetry2.xml	poetry[0]/poem[3]/title[0]	3.337
4	poetry3.xml	poetry[0]/poem[2]/content[2]	1.322
5	poetry1.xml	poetry[0]/poem[2]/content[2]	1.322

6. 결론 및 향후 연구

이 연구에서는 XML 문서를 대상으로 내용기반 검색을 지원하기 위한 시스템을 설계하고, 이와 관련한 몇 가지 이슈들을 다루었다.

XML 문서를 대상으로 내용기반 검색을 효율적으로 수행하기 위해서는 문서의 단위, 검색의 기본 단위, 내부 인덱스 노드에서의 색인어 가중치 계산, 관련한 메타 데이터들의 효율적인 저장 및 관리, 효율적인 인덱스의 구축 등 해결해야 하는 많은 과제들이 있다. 그리고 경로기반 검색과 내용기반 검색을 효율적으로 접목하기 위한 연구가 수행되어야 한다. 향후 연구로서는 이러한 문제들을 해결하기 위해 여러 가지 기법들을 연구하고, 현재 개발하고 있는 시스템을 이용하여 다양한 종류의 코퍼스를 대상으로 개발한 기법들을 실험을 통하여 비교 평가하고자 한다.

7. 참고문헌

[1] E. A. Fox and G. Marchionini(Guest Editors), Special issue "Toward a worldwide digital library", *Communication of the ACM*, 41(4), 1998.

[2] H. Maurer, Web-based knowledge management, *IEEE Computer*, 31(3): 122-123, 1998.

[3] A. Paepcke, C. K. Chang, H Garcia-Molina, and T. Winograd, Interoperability for digital libraries worldwide, *Communications of the ACM*, 41(4):33-43, 1998.

[4] SemanticWeb.org, "The semantic web community portal", 2000, http://www.semanticweb.org.

[5] S. Aviteboul, S. Buneman, and D. Suciu, Data on the Web-From Relations to Semistructured Data, **Morgan Kaufman Publishers**, San Francisco, 2000.

[6] A. Deutch D. Fernandez, M. and Florescu, and D. Suciu, "A query language for XML", **WWW8/Computer Networks**, 31(11-16): 1155-1169, 1999.

[7] D. Kossmann, "Special Issue on XML", *IEEE Data Engineering Bulletin*, 22(3), 1999.

[8] D. Chamberlin, J. Robie, and D. Florescu, Quilt: "An XML query language for heterogeneous data sources", In D. Suciu and G. Vossen, editors, *3rd International Workshop on the Web and Databases*, Dallas, Texas 2000.

[9] Norbert Gövert, Norbert Fuhr, Mohammad Abolhassani, and Kai Großjohann. Content-oriented XML retrieval with HyREX. In **Proceedings of the 1st INEX Workshop**. ERCIM Workshop Proceedings, pages 26-32. ERCIM, Sophia Antipolis, France, 2003.

[10] HAM 6.0 , http://nlp.kookmin.ac.kr