XML 문서의 내용기반 검색을 위한 인덱싱 모델 및 색인어의 가중치 부여

한예지, 한창우, 서동혁, 김수희°
호서대학교 컴퓨터공학과
shkim°@office.hoseo.ac.kr

Indexing Model and Weight Assignment on Keywords for Contents based Retrieval in XML Documents

Yeji Han, Changwoo Han, Donghyuk Seo, Suhee Kim°
Department of Computer Engineering, Hoseo University

요 약

본 논문에서는 XML 문서의 내용을 효율적으로 검색하기 위해 필요한 메타데이터의 스키마를 개발하고, 이를 바탕으로 구축되는 내용기반 인덱싱 모델을 제안한다. 제안하는 내용기반 인덱싱 모델은 엘리먼트 타입에 따라 랭킹 검색과 불리언 검색을 지원한다. 랭킹 검색 결과의 재현도와 정확도를 높이기 위해, 검색 결과의 출력 기준 노드가 리프 노드와 내부 노드인 경우를 구별하여 색인어에 대한 가중치를 부여하고, 이를 이용하여 질의와 엘리먼트간의 유사도를 계산하는 방법을 제안한다.

1. 서 론

인터넷의 발전으로 접근할 수 있는 정보의 양이 기하급수적으로 증가하면서 이러한 정보를 보다 효율적으로 사용하고자 하는 연구가 활발히 진행되고 있다. 현재 인터넷상의 정보는 다양한 형태의 문서로 존재하고 있으며, 이러한 문서를 표현하기 위한 표준을 요구하게 되었다. 이에 웹 발전의 주도적인 역할을 하고 있는 W3C에서는, 데이터의 표현과 교환을 위한 차세대 웹 문서의 표준으로 XML을 제안하였다. XML은 HTML의 단점을 보완하고 SGML의 장점을 반영한 것으로, 문서의 다양한 구조를 표현하면서 쉽게 사용할 수 있다는 장점을 가지고 있다. 그러므로 웹 문서뿐만 아니라 전자상거래, 전자도서관, 대규모 기관의 인트라넷 등의 다양한 분야에서 폭넓게 활용되고 있는 추세이다. XML 관련 연구로는 XML 저장관리 시스템에 관한 연구[1,2], XML 질의어에 관한 연구[3] 등이 있다. 본 논문에서는 XML 저장관리 시스템의 한 부분인 (내용기반) 인덱싱 모델[4,5]에 중점을 두고 있다.

사실상, 모든 면에서 우수하고 강력한 인덱스를 구축하는 것은 상당히 어려운 문제이다. 이 논문에서는 XML 문서의 내용기반 검색을 효율적으로 하기 위한 메타데이터를 개발하고, 각 엘리먼트에서 추출한 색인어에 대해 가중치를 부여하는 방법을 제안하고자 한다.

2. 관련 연구

XML 문서를 구성하는 정보표현의 기본단위는 엘리먼트이다. 그러므로 XML 문서 검색에서는 문서단위의 검색보다는 엘리먼트 단위의 검색에 중점을 두고 있다.
내용기반 검색을 위한 인덱스는 색인어들로 구성되는 인덱스 파일과 이들이 추출된 문서 및 엘리먼트의 정보를 나타내는 포스팅 파일로 구성된다. 기존에 제안된 인덱스의 방법들을 살펴보면 다음과 같다.

*Shin et al.*은 *K*-ary 완전 트리를 이용하는 BUS (Bottom Up Scheme)[4]를 제안하였다. BUS는 XML 문서의 트리 구조를 *K*-ary 완전 트리로 가정하여, 각 문서 트리의 노드에 식별자 UID(Unique element IDentifier)를 부여하는 방법이다. 이 방법은 노드의 깊이가 깊어질수록 사용하는 노드에 비해 데이터의 양이 커지고(가상 노드가 존재하므로), *K*값이나 노드의 변경(삽입, 삭제)에 따라 UID를 재구성해야 하는 단점을 가지고 있어 동적 환경에 부적합하다. BUS의 내용기반 인덱스는 리프 노드(텍스트 엘리먼트)만을 인덱싱하여 저장 장소의 오버헤드를 줄이고, 색인어를 추출한 각 엘리먼트의 정보인 <문서번호(DID), UID, 엘리먼트 타입(ETY), 레벨(LEV), 색인어의 빈도수(FREQ)>들로 포스팅 파일을 구성하여 엘리먼트 단위의 검색과 랭킹 검색을 지원한다. 한 내부 노드에서 각 색인어의 빈도수는 그에 속한 하위 노드들의 빈도수들의 합으로 계산된다. 이러한 방법으로 내용기반 검색을 수행할 경우, 항상 루트 노드가 검색되거나 재현도가 낮게 나올 가능성이 있다.

김성완을 중심으로 한 연구[5]에서는 BUS의 단점을 개선하기 위해 XML 문서 트리에 실제로 존재하는 각 노드에만 UID를 부여하기 때문에, 가상 노드가 존재하지 않으며 노드의 변경에 영향을 받지 않아 동적환경에 적합하다. 이들이 제안한 내용기반 인덱스는 인덱스의 엔트리와 관련된 <DID, UID, ETY, FREQ, next_link>들로 포스팅 파일을 구성하고, BUS와 동일한 방법으로 랭킹 검색을 지원한다. next_link는 동일한 {DID, UID}를 갖는 포스팅들을 연결하기 위해 사용하며, 이는 노드의 삭제에 의한 처리를 쉽게 하는 역할을 한다.

3. 메타데이터의 개발

이 장에서는 효율적인 내용기반 검색을 지원하기 위하여 각 엘리먼트의 내용을 기초로 한 랭킹 검색과 불리언 검색을 위하여 필요한 메타데이터를 개발한다.

◆ 파일 정보
XML 문서의 구조를 정의하고 있는 DTD 파일과 인덱스하고자 하는 XML 파일들이 존재하는 디렉토리를 명시한다.

◆ 인덱스하지 않을 엘리먼트
일반적으로 내용기반 인덱스에서는 리프 노드(텍스트 엘리먼트)에서 추출한 색인어만으로 인덱스를 구축함으로써, 저장 장소의 오버헤드를 줄인다[4,5]. 본 논문에서는 텍스트 엘리먼트 중에서 엘리먼트 타입의 의미로 볼 때 인덱스하지 않아도 되는 엘리먼트를 선별하여 명시함으로써, 저장 장소의 오버헤드를 좀 더 줄일 수 있다.

◆ 패턴매칭할 엘리먼트
전통적인 정보검색에서는 검색에서는 빈도수를 이용하여 랭킹 검색을 지원한다. 하지만 XML 문서에서는 엘리먼트 타입에 따라 랭킹 검색보다는 패턴매칭을 하여 질의의 결과로 리턴하는 것이 더 타당한 경우가 있다. 이러한 엘리먼트들을 선별하여 불리언 검색을 지원한다. 일반적으로 엘리먼트의 내용이 사람 이름과 같은 고유명사이거나 년/월/일과 같은 날짜를 나타내는 경우에는 불리언 검색을 하는 것이 바람직하다.

◆ 엘리먼트별 중요도
텍스트 엘리먼트를 대상으로 엘리먼트 타입의 의미를 따른 중요성을 명시한다. 일반적으로, 논문 형식의 문서에서는 논문의 제목이나 요약이 본문에 있는 내용보다 그 중요도가 높다고 볼 수 있다. 엘리먼트별 중요도는 각 엘리먼트에서 추출되는 색인어의 가중치를 계산하기 위해 빈도수와 함께 사용될 수 있다.

◆ 가중치 감소 비율
기존의 연구에서 질의의 대상 엘리먼트가 내부 노드일 경우, 리프 노드의 가중치를 상위 노드로 그대로 반영(누적)하여 내부 노드의 가중치를 구한다. 이 경우, 색인어가 존재하는 리프 노드의 레벨과 상관없이 가중치가 결정되므로 바람직하지 못한 결과가 나타날 수 있다. 그러므로 한 색인어의 가중치는 적당한 비율로 감소하여 상위 노드에 반영되는 것이 타당하다[6].

◆ 문서 식별자 DID와 파일 이름 Filename
DID(Document IDentifier)는 각 XML 문서를 구별하기 위해 부여되는 식별자이다. Filename은 실제 XML 문서 파일의 이름으로, 검색 결과인 DID를 이용하여 내용을 출력할 때 실제 해당 문서 파일을 찾기 위해 사용될 수 있다.

◆ 엘리먼트들의 수(Number of elements) Nele
각 문서에 존재하는 엘리먼트들의 수를 나타내며, 이는 내용기반 검색에서 색인어의 가중치를 계산하기 위해 사용된다.

◆ 엘리먼트(노드) 식별자 UID
XML 문서 상의 각 엘리먼트를 구별하기 위해 부여되는 식별자로, 문서 상에 나타나는 순서대로 부여된다.

◆ 엘리먼트 타입(Element TYpe) ETY와 절대경로 Path
DTD에 명시되어 있는 엘리먼트의 정보를 토대로 구성될 수 있는 모든 절대 경로(Path)에 대한 식별자인 엘리먼트 타입(Element TYpe : ETY)은 엘리먼트별 중요도를 식별하여 가중치를 계산하기 위해 사용될 수 있다.

◆ 엘리먼트의 레벨 LEV
각 엘리먼트의 레벨(level)로, XML 문서 트리내의 각 노드는 해당 ETY와 동일한 레벨 값을 갖는다. LEV는 가중치 감소 비율을 적용하는데 사용될 수 있다.

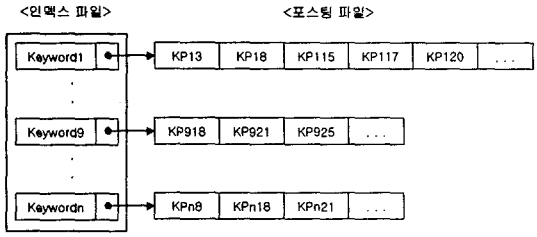
◆ 색인어 Keyword와 빈도수(frequency) FREQ
Keyword는 텍스트 엘리먼트(패턴매칭할 엘리먼트 제외)에서 추출되는 색인어로, 랭킹 검색의 키로 사용된다. FREQ는 각 엘리먼트에서 추출된 색인어별 빈도수로, 가중치를 계산하기 위해 사용된다.

◆ 패턴매칭할 색인어 PMKeyword
패턴매칭할 엘리먼트에서만 추출된 색인어로, 불리언 검색의 키로 사용된다.

4. 내용기반 인덱싱

4.1 랭킹 검색을 위한 인덱스

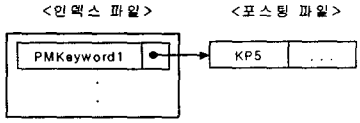
패턴매칭할 엘리먼트를 제외한 텍스트 엘리먼트에서 추출된 색인어(Keyword)로 인덱스를 구축한다. 색인어들의 포스팅 정보로는 <Keyword, ETY, DID, UID, FREQ>가 될 수 있다([그림 1] 참조).



[그림 1] 랭킹 검색용 순수 내용기반 인덱스의 구조

4.2 불리언 검색을 위한 인덱싱

패턴매칭할 엘리먼트에서 추출된 색인어(PMKeyword)만으로 인덱스를 구축한다. 포스팅 정보로는 <PMKeyword, ETY, DID, UID>가 될 수 있다([그림 2] 참조).



[그림 2] 불리언 검색용 순수 내용기반 인덱스의 구조

5. 색인어의 가중치

XML 문서에서 각 색인어에 부여되는 가중치는 XML 문서의 기본 단위인 엘리먼트별로 계산하는 것이 바람직하다. 일반적으로 색인어는 텍스트 엘리먼트(리프 노드)

에서만 추출되므로 각 색인어의 가중치는 리프 노드에서 부여될 수 있다. 그러나 문서 단위의 검색이나 혼합 검색에서는 내부 노드가 검색 결과의 출력 기준 노드로 지정될 수 있으므로, 각 색인어가 내부 노드에서 차지하는 가중치를 계산하는 방법도 필요하다.

5.1 리프 노드에서 색인어의 가중치

리프 노드에서 색인어의 가중치($ew_{i,j,k}$)는 기존의 $tf*idf$ 공식[7]을 변형한 (수식 1)을 적용하여 계산할 수 있다.

$$ew_{i,j,k} = etf_{i,j,k} * ief_i * es_l \quad \text{..... (수식 1)}$$

- $ew_{i,j,k}$: 문서 k 의 엘리먼트 j 에서 색인어 i 의 가중치
- $etf_{i,j,k}$: 문서 k 의 엘리먼트 j 에서 색인어 i 의 빈도수
- ief_i : 색인어 i 가 나타나는 엘리먼트의 수에 대한 역 엘리먼트 출현 빈도(inverse element frequency)

  $ief_i = \log_e \frac{(eN+1)}{ef_i}$

  - ef_i : 색인어 i 가 나타나는 엘리먼트의 수
  - eN : 전체 엘리먼트의 수
- es_l : 엘리먼트 타입 l 에 주어지는 중요도

5.2 내부 노드에서 색인어의 가중치

검색 기준 노드가 내부 노드(문서 단위 검색)일 경우에는 하위 노드(리프 노드)들에 있는 각 색인어의 가중치를 반영하여 내부 노드의 가중치를 구하여야 한다. 리프 노드에 있는 어떤 색인어의 가중치 값을 그 노드의 상위 노드에 그대로 반영하는 경우, 색인어가 존재하는 리프 노드의 레벨과 상관없이 가중치 값이 결정되므로, 바람직하지 못한 결과가 나타날 수 있다. 그러므로 하위 노드들에 있는 각 색인어의 가중치를 적당한 비율로 감소하여 상위 노드에 반영한다[6]. 내부 노드에서 그에 속한 하위 노드들을 고려한 확장된 가중치($xew_{i,j,k}$)는 (수식 2)를 적용하여 계산할 수 있다.

$$xew_{i,j,k} = \sum_{m=1}^{D} af^m * dew_{i,j,k} \quad \text{..... (수식 2)}$$

- $xew_{i,j,k}$: 문서 k 의 엘리먼트 j 에서 그에 속한 하위 노드들에 나타나는 색인어 i 의 가중치들을 반영한 확장된 가중치
- D : 검색 기준인 내부 노드와 그에 속한 최하위 노드의 레벨의 차
- af : 가중치 감소 비율
- $dew_{i,j,k}$: 문서 k 의 엘리먼트 j 와 레벨의 차가 m 인 하위 노드들에서 색인어 i 의 가중치들의 합

6. 유사도

질의문 내에서 검색하고자 하는 키워드가 다수일 경우, 각 키워드에 대한 검색 비중을 명시할 수 있다. 이 값들이 문서 내에서의 그 키워드의 가중치와 함께 유사도 계산에 사용될 수 있다. 질의와 XML 문서 엘리먼트 간의 유사도는 검색 결과의 출력 기준 노드가 리프 노드일 경우에는 (수식 3)을, 내부 노드일 경우에는 (수식 4)를 적용하여 계산할 수 있다.

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

$$sim(E_{j,k}, Q) = \sum_{i=1}^{n} xew_{i,j,k} * w_{qi} \quad \text{..... (수식 4)}$$

- $sim(E_{j,k}, Q)$: 질의문 Q 가 주어졌을 때, 검색 결과의 출력 기준 노드($E_{j,k}$) 별로 계산되는 유사도
  - $E_{j,k}$: 문서 k 의 엘리먼트 j
  - Q : 사용자에 의해 주어지는 질의문
- $ew_{i,j,k}$: (수식 1)을 적용하여 계산한 가중치
- $xew_{i,j,k}$: (수식 2)를 적용하여 계산한 가중치
- n : 질의문 Q 에 있는 서로 다른 색인어들의 수
- w_{qi} : 질의문 Q 에서 색인어 i 의 가중치(검색 비중)

7. 결 론

차세대 웹 문서의 표준인 XML의 활용 분야가 늘어나면서 XML 문서의 저장, 관리 및 검색을 효율적으로 지원할 수 있는 XML 데이터베이스 시스템이 필요하게 되었다. 이 논문에서는 XML 문서를 효율적으로 검색하기 위한 내용기반 인덱싱 모델을 제시하였고, 각 엘리먼트에서 추출되는 색인어에 대한 가중치 부여법을 제안하였다. 제안하는 내용기반 인덱싱 모델에서는 지정된 텍스트 엘리먼트(리프 노드)에서 정보를 추출하여 포스팅 파일을 구성하고, 엘리먼트 타입에 따라 랭킹 검색과 불리언 검색을 지원한다. 랭킹 검색은 기존의 연구들에서 취약점이었던 재현도와 정확도를 높이기 위해, 색인어의 가중치를 기반으로 하는 유사도를 이용한다. 가중치는 각 엘리먼트에서 추출된 색인어의 빈도수 정보를 통해 계산되며, 문서 단위 또는 내부 노드 단위의 검색 시의 정확도를 높이기 위해 가중치 감소 비율을 적용한다.

향후 연구 방향은 제안한 내용기반 인덱싱 모델의 성능을 정량적으로 측정하기 위해 XML 코퍼스를 대상으로 실험을 수행하고, 그 결과를 바탕으로 제안한 모델을 수정하고 개선하는 것이다.

8. 참고문헌

[1] 강형일, 최영길, 이종설, 유재수, 조기형, "RDBMS와 IRS를 이용한 XML 저장관리 시스템 설계 및 구현", 정보과학회논문지, 제7권, 제1호, pp.1-11, 2001

[2] 박충희, 이상준, "효율적 문서 검색 및 변경을 위한 XML 문서 저장 시스템 설계", 한국정보과학회 봄 학술발표논문집, 제30권, 제1호, pp.548-550, 2003

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

[4] Dongwook Shin, Hyuncheol Jang, Honglan Jin, "BUS: An Effective Indexing and Retrieval Scheme in Structured Documents", ACM DL, pp.235-243, 1998

[5] Sung Wan Kim, Jaeho Lee, Hae Chull Lim, "Indexing and Retrieval of XML-Encoded Structured Documents in Dynamic Environment", Springer-Verlag Berlin Heidelberg, pp.141-154, 2002

[6] Norbert Gövert, Norbert Fuhr, Mohammad Abolhassani, and Kai Großjohann. "Content-oriented XML retrieval with HyREX", ERCIM, pp.26-32, 2003

[7] Salton G., "Automatic Text Processing : The Transformation , Analysis, and Retrieval of Information by Computer", Addison- Wesley Publishing Company, 1989