시소러스를 이용한 XML 태그 검색 시스템

양승원∪ 노희영
강 원 대 학 교 컴 퓨 터 과 학 과
ggoma405@mirae.kangwon.ac.kr rohhy@cc.kangwon.ac.kr

An XML Tag Search System By Using Thesaurus

Seung-Won Yang∪ Hi-Young Roh
Department of Computer Science Graduate School, Kangwon National University

요 약

현재 웹 기술은 HTML에서 정보를 표현하는 외형과 내용을 분리하여 정보를 구조화 할 수 있는 XML을 사용하고 있다. 구조적으로 구성된 XML은 새로운 정보 검색의 방법을 제시하였다. 즉, 태그를 이용한 정보검색으로 검색어에 의미를 부여함으로써 정보 검색자에게 좀더 효율적인 검색을 가능하게 하였다. 그러나 이러한 구조화 문서 검색 기법은 정확한 태그를 입력하였을 경우에만 결과를 얻을 수 있는 단점을 지니고 있다. 따라서 본 논문에서는 XML문서의 태그 검색에 있어 정확한 태그 검색을 확장한 유사 태그 검색 기법을 설계하고 구현하였다. 유사한 태그를 검색하기 위하여 시소러스를 구성하였으며, 작성된 시소러스를 이용하여 유사한 태그에 대한 검색을 수행하였다. 기존의 XML 문서 검색 시스템은 정확한 태그에 대한 검색만을 수행할 수 있는 반면, 본 시스템은 태그 검색에 있어 시소러스를 활용함으로써 질의에 입력되어진 태그와 유사한 태그에 대한 검색 결과를 보여줌으로서 사용자에게 질의어 선정의 불편함을 감소시켰다.

1. 서론

웹이 개발된 지난 10년간 웹 기술 자체는 괄목할 만큼 성장하여 그 응용영역을 넓혀가고 있다. 이러한 웹 환경을 대중화시키는데 큰 공헌을 한 것은 HTML(hypertext markup language)에 의한 것이라고 해도 과언이 아니다[1]. 그러나, HTML은 태그가 한정되어 있고, 표현만을 위한 기술로써 네트워크의 체증을 야기 시키고, 정보를 재사용[2]하기 위해 추가비용이 드는 등 여러 가지 단점을 가지고 있다. 이러한 HTML의 단점을 극복하기 위해 제안된 것이 XML이다[3].

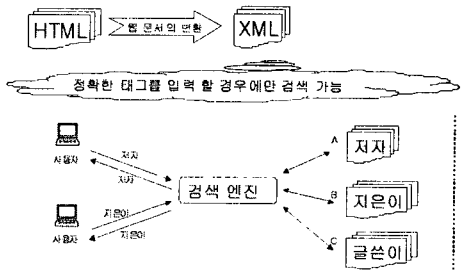
XML은 정보를 표현하는 외형과 그 내용을 분리하여 정보를 구조화함으로써 정보의 전달/재사용을 극대화하기 위한 언어이다. XML는 정보 접근의 다양성, 동적인 문서 제시, 검색정보의 일관된 관리 등에서 많은 장점을 가지고 있다. 이러한 장점과 더불어 XML은 태그를 이용한 정보검색에 있어서 검색어에 의미를 부여함으로써 정보 검색자에게 좀더 효율적인 검색을 가능하게 한다. 이러한 정보검색 방법은 계속적으로 연구·개발되고 있는 추세이며 정보 재사용의 관점에서 하나의 중요한 검색기법으로 자리 매김 할 것이다[2]. 그러나 기존의 구조화 문서 검색기법은 정확한 태그를 입력하였을 경우에만 결과를 얻을 수 있기 때문에 의미상 유사한 태그를 사용하는 경우에는 문서의 검색이 불가능하다. 즉, 정보를 제공하는 정보 제공자들은 기존의 HTML문서를 XML로 변환하는 과정에서 같은 의미를 가지고 있는 태그라도 다른 형태로 표기될 수 있다는 것이다. 따라서 이러한 혼란을 방지하고 문서의 재사용과 효율적인 검색을 위하여 정확한 태그에 대한 검색은 물론 그와 유사한 태그에 대한 검색도 가능한 시스템이 요구된다. 따라서 본 논문에서는 구조화 정보검색의 효율성을 높이기 위해 시소러스를 이용하여 유사 태그에 대한 XML문서 검색 시스템을 설계하고 구현하였다. 설계한 시스템은 시소러스 구축시 XML문서 검색이 용이하도록 하기 위해 XML Data를 이용하였다. 따라서 본 시스템은 정확한 질의 명칭을 알 수 없다고 하여도 검색을 수행할 수 있는 편의성과 관련 문서를 폭 넓게 검색할 수 있는 확장성을 제공하는 효과를 얻게 된다.

본 논문은 2장에서 시스템의 기본이 되는 시소러스를 XML Data를 이용하여 설계하는 방법에 대해 서술하였고, 3장에서는 작성된 시소러스를 이용한 XML 문서검색 시스템을 설계하고, 실험한 결과에 대해 서술하였다.

2. 본론

시소러스를 이용한 XML 문서 검색의 가장 기본적인 생각은 인터넷 서점에서 찾아 볼 수 있다. 인터넷의 확산과 함께 무수히 많은 인터넷 서점이 등장하고, 이를 이용하는 사용자도 점차로 크게 증가하고 있다. 그러나 서점마다 서로 다른 문서 형태를 가지고 있다면, 사용자들은 서점에 대한 검색을 일관되게 수행하기 어렵다. 이에 관련 정보의 재사용과 전송, 공유가 용이한 XML형태의 문서로 변환하는 작업이 이루어지고 있다. 그러나, 정보 제공자들은 기존의 HTML문서를 XML로 변환하는 과정에서 태그에 대한 제약이 없기 때문에 같은 의미를 가지고 있는 태그들이 다른 형태를 취하게 된다.

[그림2-1]에서와 같이 A라는 서점에서는 "저자"로 , B라는 서점에서는 "지은이"로, C라는 서점에서는 "글쓴이"로 각각 태그를 표현함으로써, 이런 현상은 사용자에게 적지 않은 혼란을 야기할 뿐더러 검색 시에 "저자"라는 질의를 사용한 사용자는 "저자"라는 태그를 사용한 A 서점의 결과만을 얻을 수 있을 뿐 다른 서점에서 동일한 태그로 동일한 결과의 검색이 불가능하다.



[그림2-1] XML태그 검색의 문제점

따라서 본 논문에서는 정보 제공자들이 현 HTML문서를 XML로 변환

하였을 때 발생할 수 있는 문제점을 해결하기 위해 시소러스를 이용한 XML문서 검색기법을 제안하고 설계하였다. 우선 XML 태그의 유사성을 검색하기 위하여 본 논문은 시소러스를 제작하고 이를 바탕으로 기존의 XML 태그 검색기법을 유사 태그 검색으로 확장하였으며, 시소러스의 제작과정은 보편적인 시소러스 제작과정[4]을 사용하였다.

2.1 시소러스의 생성

정보 검색 시스템에서 시소러스는 일반적으로 질의 처리 과정에서 검색 성능을 향상시키기 위해 사용될 수 있다. 즉, 사용자 질의를 시소러스를 이용하여 검색에 적합한 형태로 변형하거나 확장함으로써 검색 시스템의 정확률과 재현률을 향상시킬 수 있다.

시소러스는 색인어휘와 도입어휘(entry vocabulary)로 구성되는 통제 어휘집으로서 디스크립터(descriptor) 상호간의 관계 및 도입어와 디스크립터간의 관계가 나타나 있다. 본 논문은 이러한 관계를 이용하여 유사 태그 검색 시스템을 설계하였다.

디스크립터 상호간의 관계는 시소러스 제정과 개발을 위한 표준규격(ISO 2788)에 나타나 있으며 크게 3가지로 분류할 수 있다[5].

1)등가 관계: 등가 관계는 색인 작업 시 복수의 용어가 동일개념을 나타낸다고 인정되는 경우에 우선어 및 비우선어의 관계이다. 이는 XML의 엘리먼트의 형제 관계로 표현하였다. 따라서 형제 노드의 관계에 있는 용어들은 유사어로 정의한다. 그리고 우선어(USE)와 비우선어(UF)는 엘리먼트의 에트리뷰트 값에 의하여 결정한다.

2)계층 관계: 계층 관계는 상위 및 하위 개념을 나타낸다. 상위 개념은 클래스 또는 전체를 나타내며, 하위 개념은 한 요소 또는 일부분을 나타낸다. 이는 XML의 엘리먼트 부자 관계로 표현한다. 부자 관계에 있는 용어들은 계층 관계로 정의한다. 상위어(BT)와 하위어(NT)는 엘리먼트의 레벨에 의해 결정된다.

3)관련 관계: 관련 관계란 계층적이 아니라, 개념적으로 밀접하게 관련되어 있으나, 등가 집합에는 포함되지 않는 용어관계이다. 이는 XML의 데이터와 관련하여 에트리뷰트로 결정할 수 있다.

이와 같이 시소러스의 3가지 관계는 XML의 엘리먼트와 에트리뷰트를 사용하면 자연스럽게 매핑 시킬 수 있다. 이 관계를 사용하여 시소러스를 생성하기 위해 3개의 에트리뷰트를 사용한다.

첫째 ID 값은 디스크립터에 해당하는 엘리먼트의 깊이를 결정하기 위하여 쓰인다. 즉, ID=0 인 경우에는 가장 상위에 해당하는 용어임을 뜻한다. 그러나 이 경우에 동등한 위치의 형제 엘리먼트라도 유사어관계가 있는 것으로 해석하지 않는다.

둘째, USE 값은 우선어에 해당하는 에트리뷰트 값이다. 처음 입력되는 용어의 경우 우선어와 비우선어의 관계가 성립하지 않으므로 USE=NULL로 한다. 차후 우선어/비우선어 관계가 성립하는 용어가 입력되면 USE 값을 우선어에 해당하는 용어로 저장한다.

셋째, SN 값은 그 단어에 대한 조합레벨을 사용하거나 특정 단어와 관련 관계가 있는 경우에 입력하도록 하며 필요없을 때는 SN=NULL로 한다.

```
<Thesaurus>
        :
  <자>
    <직업 ID = "0", USE = "NULL", SN = "책과 관련된 직업">
          <지은이 ID = "1", USE = "NULL", SN = "NULL"/>
          <저자 ID = "1", USE = "지은이", SN = "NULL"/>
          <작자 ID= "1", USE = "지은이", SN = "NULL"/>
    </직업>
    <지은이 ID = "0", USE = "NULL", SN = "NULL"/>
    <저자 ID = "0", USE = "지은이", SN = "NULL"/>
    <작자 ID= "0", USE = "지은이", SN = "NULL"/>
  </자>
        :
</Thesaurus>
```

[그림 2-2] 시소러스의 예

디스크립터의 상호관계가 결정된 후 XML문서로 시소러스를 생성하기 위한 알고리즘은 다음과 같다.

step_1: 용어가 입력되면 그 해당위치(자모순 시소러스에 해당하는 위치)에 저장한다. 그리고 ID=0, USE=NULL로 하고, SN의 값은 필요시 추가한다.

step_2: n번째 용어가 들어오면 우선 step_1의 방법으로 저장한 후, n-1개의 용어와 비교하여 유사어 관계, 상하 관계를 판별한다.

step_3: 유사어 관계, 상하 관계가 판별되면, 각각의 경우에 따라 다음 과정을 적용한다.

step_3.1) 유사어 관계의 경우: n번째 용어가 어떠한 용어 x와 유사한 관계에 있으면, 우선 x용어와 똑같은 현재 엘리먼트의 내용을 복제하여 x용어 엘리먼트의 부모 엘리먼트로 생성한다. 엘리먼트를 생성한 후 ID의 값을 0으로 하고 기존의 용어 x의 ID만 1로 한다. 그리고 기존의 x용어 엘리먼트와 동등한 형제 엘리먼트로 n번째 용어를 저장한다.

step_3.2) 입력된 용어가 상위어일 경우: n번째 입력된 용어가 어떠한 용어 x 의 상위어일 경우이면, x의 용어를 n번째 입력된 용어 하위 엘리먼트로 이동시킨다. 이때 만약 유사어 관계에 있는 다른 엘리먼트가 존재할 경우 유사 엘리먼트도 모두 이동시킨다.

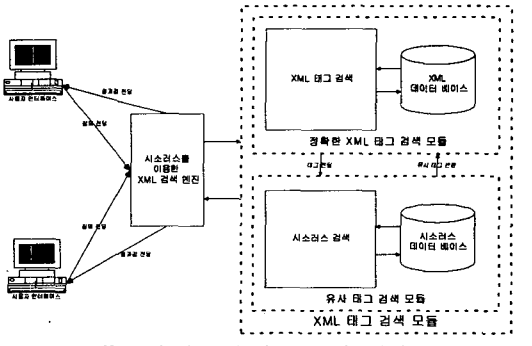
step_3.3)입력된 용어가 하위어일 경우: n번째 입력된 용어의 상위어가 있는 x용어 엘리먼트의 위치에 ID 값을 1로 하는 자식 엘리먼트로 n용어를 복제하여 넣는다. 이 때 복제하는 단어가 x용어 엘리먼트의 자식 용어와 유사한 단어가 있는지 확인하여 그 선행어를 USE 값으로 한다.

이와 같은 방법으로 인터넷 서점에서 사용 가능한 간단한 시소러스를 XML 구문을 사용하여 생성하였다. 따라서 시소러스의 제작/변경/재사용이 용이하며, 다른 시소러스와의 병합도 간단하게 해결할 수 있다.

2.2 XML 문서 검색

이 장에서는 설계된 시소러스를 이용하여 유사 태그를 검색할 수 있는 검색 엔진을 설계하고 실험한 후에 그 결과에 대해 살펴본다.

시소러스를 이용한 XML 문서 검색 시스템은 [그림2-3]과 같다.



그림[2-3] 시소러스를 이용한 XML 문서 검색 시스템

본 시스템은 기존의 XML 문서 검색 방법과 유사하나 태그에 대한 검색 수행 시 시소러스를 이용함으로써 유사한 단어의 검색을 가능하게 한다. 시소러스를 이용한 XML 문서 검색 엔진은 두 개 모듈로 구성된다.

첫째는 기존의 태그 검색 시스템과 동일한 태그 검색모듈이다. 이 모듈에서는 사용자의 질의가 입력되면 그 태그와 정확히 일치하는 XML의 태그를 검색한다.

둘째는 본 논문에서 설계하고 구현한 유사한 태그 검색 모듈이다. 이 모듈은 XML Data를 이용하여 생성된 시소러스를 데이터 베이스로 하여 사용자 질의가 입력되면 질의와 유사한 용어가 존재하는지의 여부를 시소러스의 참조로 판단하고, 유사한 태그가 존재하면 다시 첫 번째 검색 모듈로 이동하여 유사한 태그에 대한 XML 문서 검색을 실시한다.

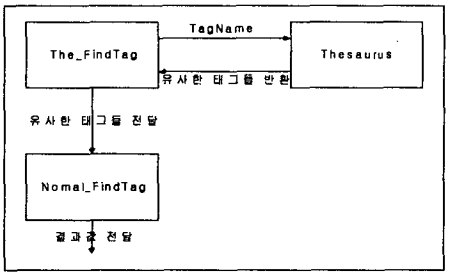
2.2.1 질의 과정

본 시스템을 테스트하기 위한 질의방법은 세 개의 입력 창을 통하여 첫 번째 입력 창에 '지은이' 또는 '저자' 와 같은 태그명을 입력하고, 두 번째 창에 '이문열'과 같은 태그에 해당하는 값을 입력하며, 세 번째 창은 찾고자 하는 정보의 태그 '도서명'을 입력하도록 한다. 즉, '지은이',

'이문열', '도서명'을 입력태그 창에 차례로 작성하면, 시스템은 "이문열이 쓴 책을 모두 검색하라"는 질의로 인식하여 검색을 시작하도록 구현했다.

2.2.2 XML 태그 검색 모듈과 개발환경

앞에서 언급했듯이 XML 문서을 검색하는 엔진은 두 개의 모듈로 구성되어 있다. 하나는 정확한 입력태그에 대해 검사하는 모듈(Normal_FindTag)이고, 다른 하나는 유사한 태그입력에 대해 검색하는 모듈(The_FindTag)이다.



[그림 2-4] 유사 태그 검색 모듈

검색 엔진은 XML 스타일 언어인 XSL[6]을 사용하였으며, XML 파일들의 탐색을 위하여 DOM(document object model)[7][8]을 사용하였다. 언어는 VBScript를 사용하였으며, 환경은 Windows 2000에서 웹 페이지를 구축하여 테스트하였다.

2.2.2.1 정확한 태그 검색 모듈

모듈 Nomal_FindTag는 태그의 명칭을 입력받아 태그를 포함하고 있는 문서의 주소와 태그가 가지고 있는 텍스트를 화면으로 보여준다.

```
Function Nomal_FindTag(tagname)
    Dim xmlDoc
    Dim ElemList
    Set xmlDoc = CreateObject("Msxml.DOMDocument")
    xmlDoc.async = False
    xmlDoc.load(URL)
    Set ElemList = xmlDoc.getElementsByTagName(tagname)
    For i = 0 To (ElemList.length-1)
            RESULTS = RESULTS + ElemList.item(i).xml
            :
    Next
    :
End Function
```

[그림 2-5] Nomal_FindTag 함수

2.2.2.2 유사한 태그 검색 모듈

유사한 태그 검색은 XML Data로 작성한 시소러스를 이용하여 질의된 태그를 시소러스에서 유사관계 및 계층관계에 따라 검색하여 그 결과를 정확한 태그 검색모듈로 넘겨주는 역할을 한다.

1) 유사한 태그에 대한 검색

질의에 사용된 용어를 Q라고 할 때, 시소러스에서 Q를 탐색한다. 만약 질의어 Q와 관계있는 단어가 없으면, 정확한 태그 검색에 대한 결과로 출력한다. 만약 질의어 Q와 관계있는 Q'가 시소러스에 존재하면, Q'의 ID값이 0이 아닌 Q'의 형제 엘리먼트가 존재하는지 검사하여, 존재하면 그 형제 엘리먼트를 반환하여 정확한 검색 모듈로 이동한다.

2) 계층관계 태그 검색

질의에 사용된 용어를 Q라고 할 때, 시소러스에서 Q를 탐색한다. 만약 질의어 Q와 관계있는 단어가 없으면, 정확한 태그검색을 최종 결과로 출력한다. 만약 질의어 Q와 관계 있는 Q'가 시소러스에 존재하면, 첫째 ID=0인 경우만 존재하면 Q'는 어떠한 용어의 상위어인 경우이므로 그 자식 엘리먼트를 Q'의 하위어로 반환하고, ID의 값이 1인 경우도 존재하면 Q'의 상위어가 존재하기 때문에 그 상위 엘리먼트를 Q'의 상위어로 반환한다.

2.2.3 테스트

현재 XML문서로 저장된 웹 문서가 극히 드물기 때문에 XML 문서수집이 불가능하다. 때문에 본 연구에서는 시스템의 테스트를 위하여 인터넷 서점에 대해 사용할 만한 시소러스와 XML문서를 생성하고 유사한 태그 검색을 위하여 [표 4-1]과 같은 입력상태를 사용하였다. 전체 문서의 수는 20개이며 이중 '지은이'가 '이문열'인 경우는 9개이다. '지은이'가 '이문열'인 9개의 문서는 같은 의미를 지니고 있지만 다른 태그를 사용하도록 하였다. 이때 유사 태그 검색이 가능한 모듈과 그렇지 않은 경우를 나누어 테스트하였다. 유사 검색모듈을 사용하지 않은 경우는 문서의 태그와 정확히 일치하는 질의를 입력하였을 경우에만 검색되어졌고, 유사 태그 검색모듈을 사용한 경우에는 시소러스에 내장되어 있는 유사단어에 해당하는 모든 태그를 결과 값으로 보여주었다.

[표 2-1] 유사 태그 검색 결과

	태그의 조합								
문서번호	1	2	3	4	5	6	7	8	9
태그이름	지은이	지은이	지은이	저자	저자	저자	작자	작자	작자
태그이름	책명	도서명	책제목	책명	도서명	책제목	책명	도서명	책제목
검색결과	×	○	○	×	○	○	×	×	×

3. 결론

본 논문에서는 XML문서에서 유사한 태그를 검색하기 위하여 시소러스를 구성하고, 이를 이용하여 유사한 태그에 대한 검색을 수행하였다. 시소러스를 구성하는 단계에서 XML Data를 사용함으로써 XML 구문을 사용하는 장점을 모두 활용할 수 있었다. 따라서 개발 시스템은 유사 태그에 대한 검색을 할 수 있게 함으로써 사용자에게 검색 질의어 선정으로 인한 어려움을 줄일 수 있다.

그러나 이 시스템은 시소러스에 유사한 태그가 정의가 되어 있는 경우에 한정되어 있기 때문에 시스템의 효율적인 사용은 시소러스의 확장이 필수적이기 때문에 앞으로 보다 효과적으로 시소러스를 생성하고 수정할 수 있는 방법이 연구되어야 한다.

4. 참 고 문 헌

[1] 이강찬, 이원석 "XML(eXtensible Markup Language)", http://dblab.comeng.chungnam.ac.kr/~dolphin/xml/
[2] "TheoSEARCH", Http://www.k4m.com
[3] "XML(eXtensible Markup Language)", Http://www.w3.org/XML
[4] 김광해 편, 유의어·반의어 사전, 한샘, 1993.
[5] "ISO2778-1986 : 단일 언어 시소러스 제정 및 개발 지침",
[6] "Extensible Style-Sheet Language(XSL)", Http://www.w3.org/XSL
[7] "Document Object Model(DOM)", Http://www.w3.org/DOM
[8] "Microsoft XML SDK Technology Preview Release-March-2000", Http://www.microsoft.com/xml
[9] J. Aitichison/A. Gilchrist, KINITI, 시소러스 작성법, 1991.7.
[10] "Extensible Markup Language(XML) 1.0" 10-February-1998, Http://www.w3.org/TR/1998/REC-xml-19980210, W3C Recommendation, 1998