제 26 회 한국정보처리학회 추계학술발표대회 논문집 제 13 권 제 2 호 (2006. 11)

의미적으로 확장된 태그들을 이용한 XML 문서들의 유사성 계산.

송인상*, 백주련*, 김응모*
*성균관대학교 정보통신공학부 컴퓨터공학과

e-mail : { insang, wise96, umkim }@ece.skku.ac.kr

Similarity Computation for XML Document with Semantically Extended Tags

In-Sang Song*, Ju-ryun Paik*, Ung-Mo Kim*
*Dept. of Computer Engineering, SungKyunKwan University

요 약

XML(eXtensible Markup language) 사용의 급속한 증가는 웹에 존재하는 많은 양의 정보들을 XML 기반 데이터로 생성하게 했으며 저장과 교환에 있어서 표준이 되도록 했다. 이는 사용자에 의한 임의의 태그정의를 가능하게 하는 XML 사용의 용이성에 기반한다. 그러나 이러한 장점은 비슷한 내용을 갖는 XML 문서에 대해서 사람들마다 개개의 태그이름과 구조를 사용한다는 문제점을 만든다. 따라서 유사한 의미를 가지고 있지만 서로 다른 문서로 분류된다. 이러한 점을 개선하기 위해 XML 문서 태그들 간의 벡터 스페이스 모델과 XML 데이터를 이용하여 시소러스를 구축하는 방법 등이 연구되고 제안되어 왔지만 아직 초보적인 단계이다. 본 논문에서는 XML 문서를 구성하는 태그들을 동의어로 확장하여 벡터를 생성하고 생성된 벡터를 가지고 태그들 간의 유사성을 체크하여 서로 다른 XML 문서들의 유사성을 수치적으로 계산한다.

1. 서론

최근 월드 와이드 웹(World Wide Web) 기술은 괄목할 만한 성장을 하면서 그 영역을 넓혀가고 있다. 1996 년 W3C 에서 제안한 XML (extensible Markup Language)[1]도 이중 역시 하나이다. XML 은 HTML 로 표현할 수 없었던 문서의 구조를 임의의 DTD 를 선언하고 태그(tag)를 정의함으로써 문서의 내용을 표현한다. 해당 문서를 즉각적인 인터넷에서의 사용이 가능하며 SGML(Standard Generalized Markup Language)보다 구현이 쉽다. 또한, [1]문서를 구성하는 태그들에 대해 정해진 규칙이 없기 때문에 사용자 임의의 명명이 가능하다. 이는 문서 내용의 자유로운 표현을 가져왔으며 그로 인해 수많은 정보가 웹으로 유입되었다[3].

하지만 XML 의 가장 큰 장점인 표현의 유연성은 XML 문서를 작성하는 사람들 마다 서로 다른 태그 이름과 구조를 사용하는 문제점을 양산했다[7]. 같은 의미를 가지는 문서들이지만 상이한 태그로 정의 되었다면 애플리케이션에서는 동일한 문서로 파악하는 것이 불가능하다. 이러 문제점 해결을 위해 시소러스를 구축하는 방법[8]과 XML 문서 태그들간의 벡터 스페이스를 생성하여 유사성을 검사하는 연구들이 있었지만, 전자의 경우 유사한 태그 정의가 미리 되어 있는 경우로 한정되고 XML 문서저장 및 검색 방법에 문제가 있었다. 하자는 정확한 유사성을 검사하지 못했으며, 두 XML 문서의 태그들만 검사하는 등의 단점이 있었다.

본 논문에서는 기존연구들의 문제점을 보완하기 위해 사용자가 정의한 용어 사전을 이용하여 시소러스

[1] 본 연구는 한국 과학재단 특정기초 (R01-2004-000-10755-0)지원으로 수행되었음

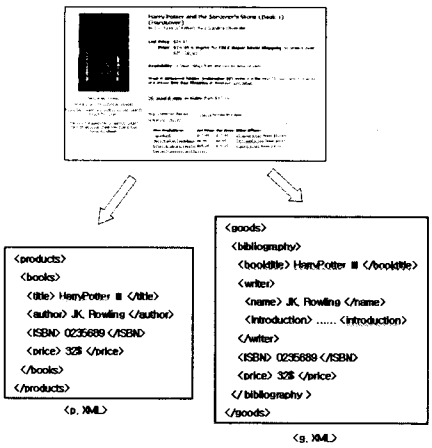
에 정의되지 않은 단어들까지 고려하여 XML 문서들 간의 유사성을 정확하게 계산하기 위해 동의어 벡터를 생성하는 방법을 제안한다.

본 논문은 2 장에서 문서 유사성과 시소러스를 살펴보고, 3 장에서 XML 문서들의 유사성 측정하기 위해 태그를 추출하는 과정과, 동의어 벡터를 생성하는 방법, XML 문서간의 유사성 계산과정을 설명하며, 4 장에서는 결론과 앞으로 나아가야 될 연구 방향을 제시한다.

2. 관련연구

2.1 문서 유사성

비슷한 내용을 갖고 있지만 문서를 작성하는 사용자에 따라 서로 다른 태그명으로 구성되기 때문에 해당 XML 문서들은 다른 범주의 문서들로 분류될 수 있다.

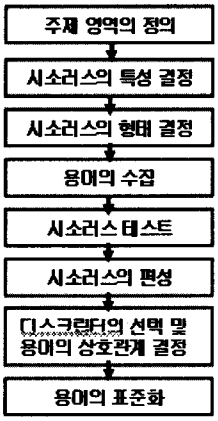


<그림 1>

예를 들어 <그림 1>에서와 같이 'harry potter'라는 책을 팔기 위해 XML 문서를 작성하는 사용자가 '책 제목'이라는 의미를 표현할 때, <title>, <subject>, <booktitle>등의 태그를 사용할 수 있다. <title>, <booktitle>등의 표현은 인간이 보면 동일한 의미로 인식이 가능하지만 애플리케이션에서는 불가능하다. 따라서 문서를 분류하는 애플리케이션은 위의 두 문서를 서로 다른 그룹으로 분류한다. 그러나 태그의 의미를 고려하여 유사성을 파악한다면 이러 잘못된 분류를 감소 시킬 수 있다.

2.2 시소러스 (Thesaurus)

시소러스는 단어를 의미에 따라 분류하고 각 단어에 대한 동의어, 유의어, 상위어, 하위어, 반의어 등을 나타낸 사전이다. 정보 검색 시스템에서 시소러스를 이용하여 사용자 질의를 검색에 적합한 형태로 변형하거나 확장함으로써 검색 시스템의 정확성과 재현성을 향상시킬 수 있다. 또한 시소러스는 용어간의 계층적인 관계를 이용하여 보다 넓은 의미의 검색어를 선정하여 광범위한 검색을 가능하게 한다. 일반적인 시소러스 형성과정은 <그림 2>와 같다[2].



<그림 2> 시소러스 형성 과정

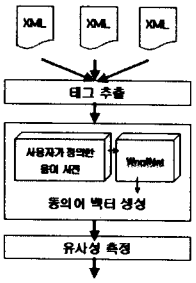
최근 프린스턴(Princeton) 대학의 Miller 가 고안한 WordNet[4]은 단어 형이 아닌 단어의 의미를 구성요소로 하는 특징을 가지고 있어서 최대한 의미를 정확히 표현하고 있다. 또한 망 형식으로 단어들을 저장하고 있기 때문에 알파벳 식 사전보다 단어들간의 연관성을 파악하기가 용이하다. 본 논문에서 사용할 WordNet 버전 2.1[5]은 명사 145,104 개와 동사 24,890 개, 형용사 31,302 개, 부사 5,720 개 총 207,016 개의 단어로 구성되어 있다.

3. XML 문서들의 유사성 측정

본 장에서는 2 장에서 언급한 기존 연구들의 문제점을 보완하여 XML 문서들의 유사성을 좀 더 정확하게 측정하기 위해 제안된 방법을 설명한다. 우선 XML 문서에서 태그들을 정제한다. 그 후, 동의어 벡터를 생성하여 태그들의 의미를 쉽게 파악할 수 있게 하는데 기존의 연구보다 좀더 확장된 의미를 가지는 동의어를 생성하고 이를 기반으로 다수의 XML 문서들의 유사성을 측정하게 된다.

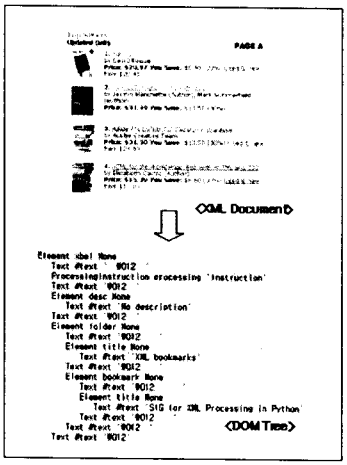
3.1 유사성 검사 시스템의 구조

본 논문에서 제안하는 유사성 검사 시스템은 크게 세 과정으로 나뉜다. 가장 먼저 행하는 과정은 XML 문서를 파싱하여 DOM 트리를 만들어 준 다음, 태그를 추출하는 과정으로 태그를 정제하는 일을 담당한다. 두 번째 과정은 WordNet 과 사용자가 정의한 용어사전을 이용하여 벡터를 생성하는 과정으로 동의 어 벡터 생성 과정을 담당한다. 최종 과정은 추출된 태그들로 만들어진 동의어 벡터를 가지고 문서들간의 유사성을 측정하는 과정으로 XML 문서들의 유사성을 측정하는 일로 나누어 진다. <그림 3>는 유사성 검사 시스템의 간략한 블록도 이다.



<그림 3> 유사성 검사 시스템의 블록도

3.2 태그 정제



<그림 4> 예제: XML 문서를 파싱하여 DOM Tree 를 생성

<그림 4>처럼 생성된 DOM 트리[9]에서 태그와 컨테츠를 분리한 후, 태그트리를 만든다. 형성된 태그 트리에서 대문자를 소문자로 바꿔주고 공백, 하이픈('-'), 언더스코어('_')등의 불필요한 문자를 제거한다. XML 문서의 유사성 측정에 전혀 영향을 끼치지 않는 모든 불필요한 단어들을 제거한 후, 사용자가 정의한 용어 사전과 WordNet 을 이용하여 각 태그에 해당하는 동의어를 찾는다.

3.3 동의어 벡터 생성

동의어 벡터 생성을 위해서 3.2 절에서 설명한 태그를 사용한다. 이때 시소러스는 단어와 단어 사이의 유사성을 검사하는데 있어서 중요한 역할을 수행한다. 지금까지 가장 널리 사용되고 있는 방법은 1990 년에 프린스턴 대학의 Miller 가 고안한 WordNet 이다. 하지만 WordNet 만 사용할 경우 다음과 같은 문제점이 발생한다.

- 첫 번째, 합성어와 생략어 등을 지원하지 못한다. 예를 들어 책 제목과 같은 <booktitle>의 합성어, <section>을 <sec>, <publication>을 <pub>로 줄여 사용하는 생략어등이 있다.
- 두 번째, 두문자를 지원하지 않는다. 예를 들어, 기독교 청년회를 나타내는 <YMCA>, 자동차, 주택, 가구 따위를 스스로 제작하거나 수리하여 쓰는 것을 뜻하는 <DIY>와 같은 것들이 있다.
- 세 번째, WordNet 은 많은 단어를 가지고 있지만 공학용어, 의학용어 등의 전문적인 용어가 부족하다.

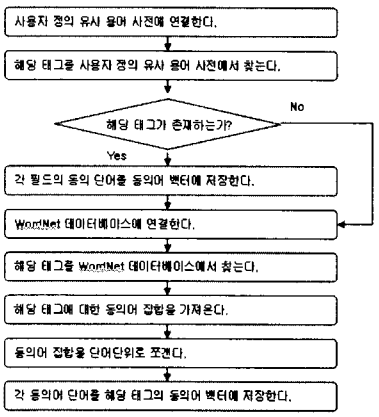
위에 나열한 문제들을 해결하기 위해 사용자가 정의해 놓은 용어 사전을 먼저 적용한다. 용어 사전의 예로 <표 1>과 같이 모델 별로 동일한 의미의 태그를 정의할 수 있다.

<표 1> 사용자 용어 사전

DocBook	TEI-List	MIL	ISO12083	HTML
p	p	Para	P	P
emphasis	emph	Emphasis	emph	EM
listItem	Item	Item	Item	LT

이렇게 정의된 용어 사전을 이용할 경우, <LT>라는 태그가 들어왔을 때, 이에 대한 동의어로써 <Item>으로 확장하여 동의어 집합에 저장할 수 있다.

태그에 대한 동의 벡터를 생성하기 위해 사용자가 정의해 놓은 용어 사전을 먼저 살펴 보고, 그 다음에 WordNet 을 사용하는 순서이다. <그림 5>은 동의 벡터를 생성하는 과정[6]이다.



<그림 5> 동의어 벡터 생성 과정

동의 벡터 생성의 예를 들면, 'author'이란 단어에 대한 동의 벡터는 title → (subject, titlebook, theme, heading, headline, caption, booktitle)와 같이 표현할 수 있다.

3.4 유사성 측정

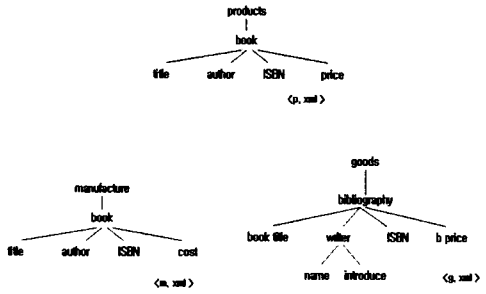
유사성을 측정하는 단계에서 세가지 XML 문서의 태그를 입력 받아 세 문서의 태그간의 유사성 매칭을 한다. 태그들간의 유사성을 비교하기 위해 다음과 같

이 레벨에 따라 유사도를 정의한다.

	유사도	내용
Level 1	1	두 태그들간의 완전 일치
Level 2	0.8	태그와 동의어 벡터의 용어 일치
Level 3	0.6	두 태그들간의 부분 일치
Level 4	0.4	태그와 동의어 벡터의 용어 부분 일치
Level 5	0.2	동의어 벡터와 동의 벡터간의 부분 일치
Level 6	0	두 태그가 완전 불일치

<표 2> 유사도 레벨

이렇게 정의된 각각의 레벨을 가지고 두 태그들의 유사성을 체크한다. 먼저 두 XML 문서를 비교해 본다. XML 문서에서 DOM 트리로 파싱되어 정제된 태그트리를 가지고 위에서부터 아래로 비교하면서 위에서 정의한 유사도를 가지고 두 문서의 유사성을 판단한다.



<그림 6> 태그 트리

예를 들어, 다음과 같은 세 개의 트리에 <표 2>를 적용한다. 세 문서 중 가장 보편화된 하나를 주된 XML 로 본다. 그리고 그 다음 태그들의 유사도를 알아본다. <m. xml>에서 <manufactures>는 유사도 0.8 를 갖는다. 또한 <book>은 1 를 갖는다. 위에 두 문서를 <표 3>에 나타내었다.

<표 3> 그림 6 의 유사도 체크

manufacture	book	title	author	ISBN	cost
0.8	1	1	1	1	0.8

goods	bibliography	Book title	writer	ISBN	bPrice
0.8	0.8	0.6	0.8	1	0.4

이렇게 나온 유사도를 가지고 <m.xml>은 93%의 유사성을 가진다. <p.xml>은 73%의 유사성을 보인다. 이러한 방법으로 각 단어가 각 문서에서 가지는 유사성을 계산하여 XML 문서 각각을 표현할 수 있다. 이렇게 표현된 XML 문서를 다수의 문서에 적용하여 유사성을 체크할 수 있게 되었다. 본 논문에서 제시한 모델을 사용할 때 유사성은 인간이 휴리스틱하게 두 문서를 비교 하였을 때와 같이 정확하게 측정되었다.

4. 결론 및 향후 연구

본 논문에서 XML 문서들의 유사성 계산을 위하여 XML 태그를 확장한 후 분석하는 기법을 제시하였다. 기존의 유사성 검사는 많은 XML 문서들 중에서 문서 두 개씩 비교해야 한다. 문서를 두 개씩 비교함으로써 시간적 소모가 많고 동의어 벡터의 양이 많아지게 된다. 우리는 다수의 XML 문서의 유사성 측정으로 인한 시간적 손실과 데이터의 양을 줄일 수가 있고 의미를 반영한 태그를 적용함으로써 XML 문서의 분류화가 정확해 진다.

그러나 우리가 제시한 사용자 정의 용어 사전은 특정 사이트나 정보를 대상으로 이루어졌다. 방대하고 다양한 XML 문서에 사용자가 정의하는 용어 사전을 적용하는 방법을 연구해야 한다. 그리고 두 문서를 비교했을 때 나타나는 유사성과 같은 유사성을 측정할 수 있어야 하는 것도 연구 과제로 남아있다.

Acknowledgement

본 연구는 21 세기 프론티어 연구개발사업의 일환으로 추진되고있는 정보통신부의 유비쿼터스컴퓨팅 및 네트워크원천기반기술개발사업의 지원에 의한 것임

참고문헌

[1] "XML (eXtensibel Markup Language)", http:/www.w3.org/xml

[2] J. Aitichison and A. Gilchrist, KINITI, "Thesaurus framing method"

[3] Minnos V, Garofalakis, Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim, "Of Crawlers, Portals, Mice and Men : Is there more to mining the web?", In Proc of the ACM SIGMOD Int. Conf. Management of data, pages 504, Philadelphia, PA, USA, 1999

[4] Miller G.A, Beckwith R., Fellbaum C., Gross D. and Miller K., "Introduction to WordNet : An On-line Lexical Database." In Five papers on WordNet, CSL report, Cognitive Science Laboratory, Princeton University, 1993

[5] "" WordNet 2.1", http://wordnet.princeton.edu/obtain

[6] Jung-Won Lee, Kiho Lee, Won Kim, "preparation for semantics-based XML Mining" , IEEE, 2001.

[7] T. Bray, J. Paoil and C. M. Sperbers-McQueen, Extensible Markup Language <XML> 1.0 W3C Recommendation, Word Wide Web Consortium, Feb, 1998.

[8] Seung-won yang, hi-youn Roh, "A Similar Tag Searching System in XML text", Journal of Telecommunication and information , Vol, 5, 2001

[9] "Document Object Model (DOM)", "http://www. microsoft.com/xml