유사 구조를 갖는 XML 문서 생성기의 설계 및 구현

이범석° 이재민 황병연
가톨릭대학교 컴퓨터공학과
{bslee°, likedawn, byhwang}@catholic.ac.kr

Design and Implementation of XML Document Generator with Similar Structure

Bum-Suk Lee° Jae-Min Lee Byung-Yeon Hwang
Dept. of Computer Engineering, The Catholic University of Korea

요 약

여러 장점을 가지고 점차 그 사용이 증가하고 있는 XML은 내용뿐만 아니라 그 구조적 정보까지 포함하고 있는 특징을 가지고 있는데, 이러한 XML 문서를 효율적으로 검색하기 위해 구조 유사성을 기반으로 하는 검색 기법이 개발되고 있다. 새롭게 개발되는 유사한 구조의 XML 문서를 검색하는 시스템의 성능 평가를 위해서는 구조적으로 유사한 다량의 XML 데이터가 필요하다.

본 논문에서는 지금까지 개발되었던 유사 구조 문서 생성기를 바탕으로 사용자가 원하는 데이터 구조를 생성하는데 보다 효과적인, 유사 구조를 갖는 XML 문서 생성기인 xTrans를 설계 및 구현한다. xTrans는 원본 XML 문서에 삽입, 삭제, 치환의 세 가지 연산을 이용하여 사용자가 원하는 일정한 비율만큼의 구조적 변화를 일으키는데, 그러한 연산은 불규칙한 위치에서 생성되므로, 같은 비율의 변화가 일어난 여러 개의 유사 구조 문서를 생성할 수 있다. 사용자는 각 연산의 변형 비율을 지정해주어 원하는 만큼 변형시킨 문서를 생성하고, 이 문서들을 이용하여 새롭게 개발되는 유사 구조 문서 검색 시스템의 성능평가에 활용할 수 있다.

1. 서 론

1996년 W3C에 의해 제안된 XML[1]은, 최근 몇 년간 데이터 교환을 위한 공통의 표준으로서 각광받아왔고, 앞으로는 컴퓨터가 사용되는 모든 분야로의 사용 증가가 기대되고 있다. XML의 가장 큰 특징은 데이터의 내용뿐만 아니라 구조까지도 포함하고 있기 때문에 데이터를 표현하고 저장하고 교환하는 분야에서 다양하게 활용될 수 있다.

XML의 사용이 증가하면서 XML 문서를 검색하는 다양한 기법[2,3]들이 연구되고 있는데, 그 중에서 구조적으로 유사한 문서를 검색하는 기법이 있다. 이러한 유사 구조 문서 검색 시스템을 성능 평가하려는 경우 구조적으로 유사한 여러 개의 XML 문서가 사용된다. 그러나 웹에서 성능 평가에 사용할 만큼 다양하고 충분한 양의 XML 문서를 구하는 것은 쉬운 일이 아니기 때문에, 유사 구조 문서 생성기가 필요하다. 이러한 필요성 때문에, DTD를 이용한 유사 구조 문서 생성기나, 혹은 XML 자체를 이용한 유사 구조 문서 생성기 등이 개발되었으나, 그 사용 목적과 기능이 제한적이었다. 따라서, 사용자는 필요에 따라 기존의 유사 구조 문서 생성기를 수정하거나, 별도의 생성기를 개발해야 할 필요성이 대두되었다.

본 논문에서는 기존의 XML 문서를 이용하거나 새로운 XML 문서를 사용하여, 원본 문서와 유사한 구조를 갖는 사용자가 원하는 만큼의 새로운 문서를 생성시킬 수 있는 유사 구조 XML 문서 생성기의 설계 및 구현 내용을 기술한다.

본 논문의 구성은 다음과 같다. 2장에서는 지금까지 개발되거나 연구된 유사 구조 XML 문서 생성기를 기술한다. 3장에서는 xTrans의 조건과 연산에 대해 제안하고, 4장에서는 xTrans의 설계 및 구현에 대해 논의한다. 마지막으로 5장에서는 결론 및 향후 계획을 기술한다.

2. 관련연구

제노바 대학교에서 특정한 DTD에 유효한 XML 문서들에 대해 유사한 구조를 갖는 다른 3개의 DTD와 대응시켜 그 구조 유사도에 따라 분류하고, 그 유사도를 측정한 실험이 있다. 이 실험에 사용된 유사한 구조를 갖는 XML 문서 생성기는 입력된 원본 DTD에 유효한 10,000개의 XML 문서를 무작위적으로 생성시키는 역할을 수행했다. 그러나 이 생성기는 구조 유사도 측정이라는 특정 실험을 위한 것이었으므로, 생성되는 XML 문서는 내용을 포함하지 않고 단순히 요소만으로 이루어진다. 또한, 원본 데이터로 XML 문서를 사용하는 것이 아니라, DTD만 사용이 가능하다는 단점을 가진다[4].

미시간 대학교의 유사 구조 문서 검색을 위해 새롭게 개발된 시스템의 성능평가를 위해 제작한 유사 구조 XML 문서 생성기도 있다. 이 생성기는 삽입, 삭제, 치환, 트리 삽입, 트리 삭제 등의 5가지 연산을 사용하여, 보다 규칙적인 알고리즘을 개발하였다. 이것을 사용하면 구조적으로 유사한 XML 문서의 생성이 간단하게 이루어지지만, 각 연산이 적용되는 비율이 무작위적으로 결정되기 때문에, 사용자가 원하는 만큼의 유사 구조 문서를 생성시키는 것은 아니다[5].

3. xTrans의 조건과 연산

3.1 xTrans의 조건

구조적으로 유사한 XML 문서를 검색할 때, 단순하게 요소들의 구조를 중요시하는 방법과 내용을 포함한 요소들에 더 비중을 두는 방법이 있는데, 본 논문의 유사구조 문서 생성기 xTrans는 후자의 것에 더 큰 비중을 두고 개발되었다. 따라서, 기존의 XML 문서를 이용해서 유사 문서를 생성할 때는 상관없지만, 새로운 XML 문서를 생성시켜 그것과 유사한 구조를 가진 문서를 만들고자 할 때에는 새 XML 문서에 텍스트 데이터를 포함한 요소를 삽입한다.

XML 문서에서 각 요소는 속성을 가질 수 있는데, 이 속성들은 DTD를 결정하는데 중요한 역할을 할 수 있다. 또한 이들 속성은 해당 요소의 자식 요소로서 표현될 수 있으나[6], 본 논문에서는 우선적으로 요소의 구조만을 중요시하기 때문에 구조적 특성으로서의 속성은 무시하였다.

3.2 xTrans의 연산 정의

XML 문서의 구조 변형을 위해 본 논문에서는 3가지의 연산을 설정하였다. 그림 1은 xTrans에서 정의하는 연산을 보여준다. 이러한 연산은 XML 문서의 각 요소에 대해 불규칙한 위치에서 생성할 수 있으며, 각 연산의 정의는 다음과 같다.

정의 3.1 [삽입] 어떤 부모 요소와 그 자식 요소 사이에 새로운 요소를 추가하거나, 혹은 어떤 요소의 자식요소에 새로운 요소를 추가하는 것을 의미한다.

정의 3.2 [삭제] 불규칙적으로 해당 요소를 삭제하는 것을 의미한다. 이때의 삭제는 특정 요소에만 작용하고, 그 하위 요소들은 그 깊이가 한 단계씩 낮아지게 된다.

정의 3.3 [치환] 이것은 어떤 한 요소를 새로운 이름의 요소로 바꾸어주는 것을 의미한다. 이것은 외형적인 구조를 바꾸지는 않지만, 요소를 바꾸어주므로 원본과 다른 구조를 가지게 된다고 볼 수 있다.

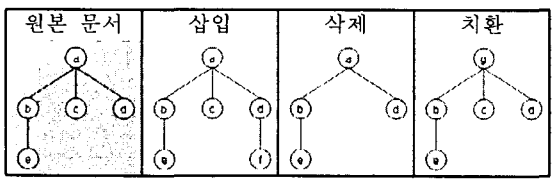

그림 1 연산의 예시

이 밖에도 트리 삽입과 트리 삭제 등의 연산을 이용할 수도 있으나, 이것은 어떤 깊이의 요소에서 적용되느냐에 따라 큰 차이가 생길 수 있다. 그렇기 때문에, 사용자가 원하는 만큼의 변형 비율이 적용된 유사 구조 문서를 생성시키려는 본래의 목적에 맞지 않아 xTrans에서는 사용하지 않았다.

4. xTrans의 설계 및 구현

xTrans는 MS Visual C#.NET으로 구현되었으며, 운영체제는 MS Windows 2000 Server가 사용되었다. 전체적인 구조는 입력 파일 정의, 변환 비율 정의, 출력 파일의 정의, 마지막으로 유사 구조 문서 생성의 네 가지 단계로 구성된다.

첫 번째 단계에서는 입력 파일을 정의한다. 우선 입력 파일로 기존 문서를 활용할 것인지, 아니면 새로운 XML 문서를 생성할 것인지 결정하고, 새로운 문서를 생성할 것이라면 새 문서의 구조를 결정하고 그에 적합한 문서를 생성시킨다. 새로운 문서를 생성시킬 때 사용자는 깊이, 요소, 속성, 그리고 요소의 이름 등을 원하는 대로 설정할 수 있다. 그림 2는 입력 파일 정의 단계를 보여준다.

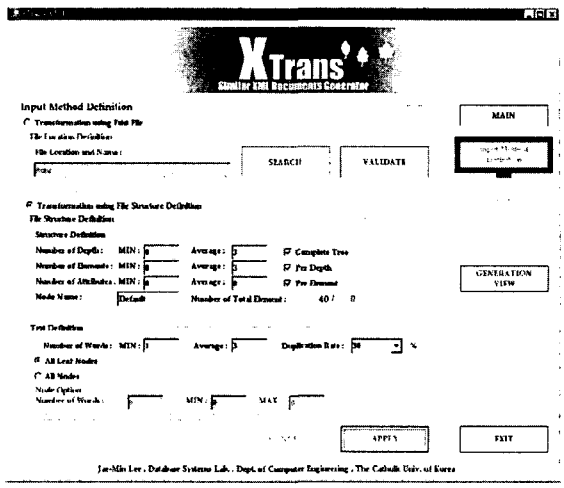

그림 2 입력 파일의 정의

두 번째 단계에서는 앞서 정의한 변형 연산의 비율을 설정해 줄 수 있다. 이때, 각 연산의 각 비율을 서로 분리하여 지정해 줄 수도 있고, 모든 변형의 전체 비율을 지정해 줄 수도 있다. 사용자가 원하는 연산의 각 비율을 지정하면, 원본 문서의 요소의 전체 개수에 대한 변형 요소의 개수인 변형률을 확인 할 수 있다. 이 때의 변형률에 대한 정의는 정의 4.1과 같다. 그림 3은 변형 연산 비율의 설정 단계를 보여준다.

세 번째의 출력 파일 정의 단계에서는 원하는 출력 파일의 개수와 파일 이름을 지정할 수 있다. 일정한 변형률 아래에서 구조적으로 유사한 문서가 여러 개 출력될 수 있기 때문에 출력 파일의 개수를 따로 지정해야 하는 것이다. 이 단계의 설정으로 같은 변형률을 가지면서 서로 다른 구조를 가지는 여러 개의 파일을 생성할 수 있

다.

정의 4.1 [변형률] 변형률($TransRate(d_1, d_2)$)은 두 개의 문서가 갖는 구조적인 차이를 나타내는 기준이다. 기준 문서 d_1을 변형하여 d_2를 생성하는 경우 기준 문서 d_1에 대한 새로운 문서 d_2의 변형률은 다음과 같이 정의한다.

$$TransRate(d_1, d_2) = \frac{TransNodeNum(d_1, d_2)}{NodeNum(d_1)}$$

$NodeNum(d_1)$: 문서 d_1에 속하는 모든 요소의 개수
$TransNodeNum(d_1, d_2)$: 문서 d_1에 속하는 요소들 중에서 d_2에 존재하지 않거나 d_2에만 존재하는 요소의 개수

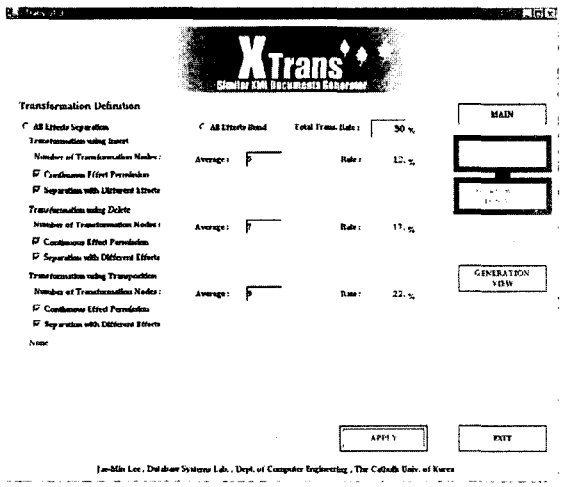

그림 3 변형률의 설정

마지막으로 유사 구조 문서 생성 단계에서는 유사 구조 문서가 생성되는 양을 보여주며, 총 몇 개의 요소에서 삽입, 삭제, 치환 연산의 각 비율을 표시하고, 그 출력 파일이 몇 개인지 메시지를 나타낸다. 또한 부가적으로 몇 개의 단어가 사용되었는지도 함께 표시한다. 그림 4는 실제 문서가 생성되는 단계를 보여준다.

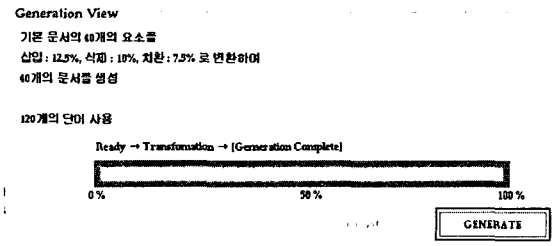

그림 4 문서의 생성

5. 결론 및 향후 계획

XML의 사용 범위는 점차 넓어지고 있고, 사용이 증가하는 만큼 문서 검색의 방법도 다양해지고 있다. 본 논문에서는 유사 구조 문서 검색 시스템의 성능평가를 위한 유사 구조 문서 생성기 xTrans를 설계 및 구현하였다.

앞서 관련 연구에 소개했던 제노바 대학교와 미시간 대학교의 유사 구조 문서 생성기는 공통적으로 특정 성능 평가를 위해 제작되었기 때문에, 범용적으로 사용할 수 있는 것은 아니었다. 그러나 본 논문에서 소개한 유사 구조 문서 생성기 xTrans는 사용자가 자신의 필요에 따라 원하는 구조를 가진 충분한 양의 유사 구조 문서를 생성해 낼 수 있기 때문에 보다 범용적으로 사용될 수 있는 장점을 가진다. xTrans는 XML 문서에 사용자가 원하는 비율만큼의 구조적 변형을 일으킨 유사 구조 문서를 다량으로 생성할 수 있으며, 기존의 문서뿐만 아니라, 사용자가 정의한 구조를 가지는 새로운 문서를 이용하여, 그와 유사한 구조를 만들 수도 있다. 원하는 만큼의 데이터를 얻을 수 있으므로 xTrans를 이용한다면, 새로운 유사 구조 문서 검색 시스템의 성능평가를 보다 정확하고 간단하게 수행할 수 있다.

향후에는 XML 문서의 요소와 더불어 속성의 구조적 특성을 이용하여 좀 더 다양한 유사 구조 문서를 생성할 수 있도록 xTrans를 확장할 것이다. 또한 DTD를 이용하여 보다 융통성있는 문서의 변환 및 생성이 가능하도록 연구를 지속할 것이다.

참고문헌

[1] W3C, "Extensible Markup Language(XML) Version 1.0 (Second Edition)," http://www.w3c.org/TR/REC-xml, October 2000.

[2] J. P. Yoon, V. Raghavan, V. Chakilam, and L. Kerschberg, "BitCube: A Three-Dimentional Bitmap Indexing for XML Documents," Journal of Intelligent Information System, Vol.17, pp.241-254, 2001.

[3] B. Cooper, N. Sample, M. Franklin, and M. Shadmon, "A Fast Index for Semistructured Data," Proceedings of the 27th VLDB Conference, Roma, Italy, 2001.

[4] E. Bertino, G. Guerrini, and M. Mesiti, "Measuring the Structural Similarity among XML Documents and DTDs," DISI Technical Report, 2002.

[5] A. Nierman and H. V. Jagadish, "Evaluating Structural Similarity in XML Documents," Proceedings of the 5th Workshop on the Web and Databases, Madison, Wisconsin, USA, 2002.

[6] S. Chawathe, A. Rajaraman, H. Graciamolina, and J. Widom, "Change Detection in Hierarchically Structured Information," Proceedings of ACM SIGMOD, pp.493-504, 1996.