특정 주제 웹문서의 논리적 구조 분석

이민형° 이경호
연세대학교 컴퓨터과학과
mhlee°@icl.yonsei.ac.kr khlee@cs.yonsei.ac.kr

Logical Structure Analysis of Topic-specific Web Documents

Min-Hyung Lee°, Kyong-Ho Lee
Dept. of Computer Science, Yonsei University

요 약

본 논문에서는 웹 문서를 XML 문서로 변환하기 위한 논리적 구조분석 방법을 제안한다. 제안된 방법은 비주얼 그룹화, 요소 식별, 그리고 논리적 그룹화의 세 단계로 구성된다. 특히 정교한 수준의 논리적 구조분석을 지원하기 위하여 특정 주제에 속하는 문서 유형의 논리적 계층 구조를 효과적으로 기술할 수 있는 문서 모델을 정의한다. 제안된 방법은 비주얼 그룹화를 통해서 추출된 시각적 계층구조와 문서 유형에 대한 논리적 구조 정보를 기술한 문서 모델에 기반하기 때문에 보다 정교한 수준의 구조 분석을 지원한다. 제안된 방법의 성능을 평가하기 위하여 웹으로부터 추출한 다수의 HTML 문서를 대상으로 실험한 결과, 기존 연구와 비교하여 논리적 구조분석을 성공적으로 수행하였다. 제안된 방법은 논리적 구조분석의 최종 결과로서 XML 문서를 생성하기 때문에 문서의 재 사용성을 높인다.

1. 서 론

XML (eXtensible Markup Language) [1] 은 논리적 구조를 표현할 수 있다는 장점 때문에 차세대 웹 문서 표준으로 그 중요성이 널리 인식되고 있다. 특히, 웹의 이질적인 정보를 논리적으로 조직하고 분류할 수 있는 방법이 정보화 지식 사회의 주요한 문제로 떠오르면서 XML이 이에 대한 해결책으로 대두되고 있다. 따라서, HTML 문서로부터 유용한 정보를 추출하여 XML 문서로 변환하는 방법이 요구된다.

일반적으로, 인간은 문서로부터 텍스트 영역의 기하적 특성이나 어휘 정보를 이용하여 제목 또는 요약 등의 논리적 구성 요소를 식별하고, 이를 병합하여 절 구조와 같은 복합적인 구성 요소를 식별함으로써 문서의 논리적 계층 구조를 인식한다. 이와 같이 텍스트 영역의 기하적 및 어휘적 특성으로부터 직접적인 식별이 가능한 논리적 요소를 주 구조(primary structure)라고 하며 식별된 다수의 구성 요소들을 병합함으로써 추출 가능한 구성 요소를 부 구조(secondary structure)라고 한다 [2].

그러나, HTML 문서의 논리적 구조분석에 관한 기존 연구의 대부분은 주 구조에 해당하는 구성 요소만을 추출하거나 단순한 수준의 구조분석을 지원한다. 한편, 웹 문서로부터 논리적인 계층 구조의 효과적인 추출을 위해서는 문서 유형의 논리적 계층 구조에 대한 다양한 정보를 표현할 수 있는 문서 모델이 요구된다. 기존 연구의 대부분은 단순한 수준의 문서 모델을 제공한다.[3][4][5][6][7][8]

본 논문에서는 웹 문서로부터 논리적인 구조 정보를 추출할 수 있는 효율적인 방법을 제안한다. 제안된 방법은 비주얼 그룹화(visual grouping), 요소 식별(element identification), 그리고 논리적 그룹화(logical grouping)의 세 단계로 구성된다. 특히, 제안된 방법은 정교한 수준의 구조분석을 위하여 문서 모델을 효율적으로 표현할 수 있는 언어인 MEDL(multi-level element description language)을 제안한다. MEDL은 논리적 계층 구조를 기술하기 위하여 특정 주제에 속하는 문서 집합이 포함할 수 있는 논리적 구성 요소의 종류, 포함관계, 그리고 빈도수 등에 대한 다양한 정보를 기술한다.

제안된 방법의 성능을 평가하기 위하여 웹으로부터 추출한 HTML 문서 집합을 대상으로 실험한 결과, 기존 연구와 비교하여 논리적인 구조분석을 성공적으로 수행하였다.

· 이 논문은 2003년도 LG-연세대학교 산학 프로그램의 지원에 의하여 연구되었음.

2. 문서 모델

본 논문에서는 특정 주제에 속하는 문서 집합에 대한 문서 모델을 기술할 수 있는 언어인 MEDL 정의한다. MEDL은 논리적 구성 요소가 포함할 수 있는 요소의 종류, 순서, 그리고 반복 횟수 등에 대한 정보를 정규 수식으로 표현한다. 특히 주 구조에 대하여 이를 식별하는데 필요한 키워드 및 문자열 패턴과 같은 어휘 정보를 기술한다.

한편, XML은 문서의 논리적 구조 정보를 DTD로 기술한다. 제안된 방법은 논리적 구조 정보와 더불어 주 구조의 어휘적 특성을 기술하기 위하여 XML DTD의 엘리먼트(element) 선언 방법을 확장하여 MEDL을 정의한다. 이에 대한 자세한 기술은 제한된 지면 관계상 생략한다.

3. 논리적 구조분석 방법

본 논문은 HTML 문서로부터 논리적 구조 정보를 추출하여 XML 문서를 생성하는 것을 목적으로 한다. 제안된 방법은 <그림 1>과 같이 전처리, 논리적 구조분석, 그리고 후처리의 세 단계로 구성된다.

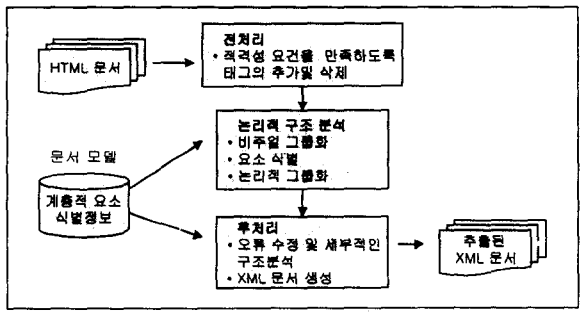

그림 1 제안된 구조분석 과정

3.1 전처

제안된 방법은 HTML 문서를 구조화하기 위하여 DOM (Document Object Model) [9] 트리에 기반한다. DOM은 HTML 문서를 구성하는 태그, 속성(attribute), 그리고 텍스트 정보를 노드로 트리를 구성한다.

한편, HTML 문서로부터 DOM 트리를 구성하기 위해서는 HTML 문서가 XML 적격성 요건을 만족하여야 한다. 제안된 방법은 HTML Tidy[10]를 적용하여 HTML 문서를 XML의 적격성 요건을 만족하는 형태로 변환한다.

3.2 논리적 구조분석

논리적 구조분석 방법은 비주얼 그룹화, 요소 식별, 그리고 논리적 그룹화의 세 단계로 구성된다.

3.2.1 비주얼 그룹화 (Visual Grouping)

일반적으로, HTML 태그 중에는 절 또는 독립적인 영역을 시각적으로 구분하는데 사용되는 태그들이 존재한다. 또한, 사용자마다 서로 다른 태그를 사용할 수 있지만 일반적으로 동일한 문서 내에서는 일관되게 사용된다. 제안된 비주얼 그룹화는 문서의 영역을 구분하는데 사용되는 HTML 태그 집합에 기반하여 HTML DOM 트리를 계층 구조를 갖는 비주얼 그룹 트리로 재구성한다.

제안된 방법은 비주얼 그룹화를 위하여 두 가지 종류의 태그 집합에 기반한다. 먼저 흔히 절 제목을 나타내기 위하여 사용되는 헤딩 태그 (heading tag) 인 태그 <H*n*>을 이용한다. 본 논문에서는 <H*n*> 태그와 같이 해당 태그만으로도 비주얼 그룹의 식별이 가능한 태그를 단일 비주얼 태그 (single visual tag) 라고 정의한다.

한편, 문서의 작성자에 따라 절 또는 독립적인 영역의 제목을 표현하기 위하여 복수의 태그를 결합하여 사용할 수 있다. 예를 들어, 단락을 구분하기 위하여 사용되는 <P> 태그와 글자의 크기 정보를 나타내는 <FONT> 태그의 "size" 속성을 결합하여 절 제목을 구분할 수 있다. 이와 같이 절 제목을 나타내기 위하여 사용되는 다수의 태그 집합을 복합 비주얼 태그 (complex visual tag)라고 정의한다.

1. DOM 트리를 하향식 너비 우선 탐색하면서 비주얼 태그를 검색한다.
2. 검색된 비주얼 태그에 대하여 <GROUP> 요소를 생성하고 해당 텍스트 내용을 "title" 속성 값으로 변환한다.
3. 너비 우선 탐색을 계속하여 현재 비주얼 태그보다 우선순위가 높거나 같은 비주얼 태그 또는 문서의 끝에 도달하면 현재 <GROUP>요소 이후에 위치하는 노드들을 <GROUP> 요소의 자식으로 추가한다.
4. 각각의 <GROUP> 요소에 대해서 1-3번 과정을 반복 적용한다.

그림 2 제안된 비주얼 그룹화 알고리즘

제안된 방법은 DOM 트리에 하향식 너비 우선 탐색(breadth first search) 과정을 적용하면서 비주얼 태그를 찾는다. 비주얼 태그 사이에는 우선순위가 존재한다. 예를 들어, 태그 <H2>가 태그 <H3>보다 우선순위가 높으며 먼저 그룹화 된다. 복합 비주얼 태그의 경우, 글자의 크기 정보를 나타내는 "size" 속성을 포함할 경우, 이와 동일한 글자 크기를 갖는 헤딩 태그의 우선순위를 갖는다. 비주얼 태그에 의하여 구분되는 영역을 <GROUP> 태그로 둘러싸며, 특히 비주얼 태그가 둘러싸고 있는 텍스트를 <GROUP>의 "title" 속성 값으로 할당한다. 제안된 비주얼 그룹화 알고리즘에 대한 자세한 기술은 <그림 2>와 같다.

3.2.2 요소 식별(Element Identification)

요소 식별은 비주얼 그룹 트리에 문서모델을 적용하여 논리적 구성 요소를 식별한다. 이를 위하여 먼저 문서에 포함된 텍스트를 구분자(delimiter)를 기준으로 토큰화(tokenization)한다. 본 논문에서는 구분자로 ';', ',', 그리고 ':'을 사용한다. 또한, 문서모델에 정의된 키워드 패턴을 포함하는 토큰(token)을 해당 논리적 요소로 식별한다.

제안된 요소 식별 방법은 논리적 구성 요소 사이의 포함 관계 및 계층 구조 정보를 포함하는 문서모델에 기반한다. 따라서 임의의 요소를 식별하는 과정에서 이미 식별된 상위 요소와의 포함관계를 고려하기 때문에 논리적 요소를 정확하게 식별한다. <GROUP> 태그의 속성 TITLE의 값에 해당하는 요소를 식별한 후, <GROUP>가 둘러싸는 영역에 대하여 해당 요소의 자손으로 올 수 있는 요소만을 식별한다.

제안된 방법은 문서모델에 정의된 논리적 계층구조에 기반하여 단계적으로 요소를 식별한다. 만일 단순히 키워드만으로 요소를 식별한다면 다수의 잘못된 요소를 식별하여 정확한 구조의 XML 문서를 생성할 수 없을 것이다. 제안된 요소 식별 방법에 대한 자세한 기술은 <그림 3>과 같다.

한편, 한 개의 토큰 내에서 두 개 이상의 요소가 식별될 때, 두 요소 중에서 하나의 요소가 다른 하나의 요소를 수식하는 관계가 발생할 수 있다. 이러한 경우, 두 개의 요소로 식별하지 않고 하나의 요소로 간주한다.

1. 비주얼 그룹 트리를 너비 우선 탐색하면서 2번 또는 3번 과정을 적용하면서 문서모델에 정의된 요소를 식별한다.
2. 비주얼 그룹화 단계에서 그룹화 된 <GROUP> 태그의 경우에는 다음을 수행한다.
   2.1 <GROUP>의 TITLE 속성 값에 해당하는 논리적 요소를 식별한다. 즉, 식별된 요소 이름에 해당하는 태그를 생성하며 속성 TITLE의 값을 생성된 요소의 속성 VAL의 값으로 변환한다.
   2.2 <GROUP> 태그로 둘러싸인 영역에 대하여 식별된 요소의 자손으로 올 수 있는 요소를 식별한다.
3. 그룹화 되지 않은 영역에 대하여 다음의 요소식별 과정을 적용한다.
   3.1 토큰이 포함하는 키워드를 기반으로 논리적 요소를 식별한다. 이때 해당 토큰을 식별된 논리적 요소를 이름으로 하는 태그의 속성 VAL의 값으로 넣어준다.
   3.2 만일 한 개의 토큰으로부터 두 개 이상의 요소가 식별되면 수식관계를 고려하여 단일의 요소로 식별한다.
   3.3 요소가 식별이 되지 않으면 해당 텍스트 값은 태그 <TEXT>의 속성 VAL의 값으로 변환한다.

그림 3 제안된 요소 식별 알고리즘

3.2.3 논리적 그룹화(Logical Grouping)

논리적 그룹화는 전 단계에서 식별된 논리적 요소를 포함하는 비주얼 그룹 트리를 대상으로 보다 정교한 계층구조를 생성한다. 특히 제안된 논리적 그룹화는 여러 요소들이 반복이 될 때 처음에 반복이 되는 요소가 반복되는 내용들을 대표할 수 있다는 점에 기인한다.

1. 비주얼 그룹 트리를 하향식 너비 우선 탐색하면서 반복되는 자식요소를 포함하는 노드 또는 HTML 리스트 아이템 태그를 검색하여 2번 또는 3번 과정을 적용한다.
2. HTML 리스트 아이템 태그인 경우 첫 번째 자식 요소를 부모로 하는 그룹을 형성한다.
3. 노드의 자식 중에 반복되는 요소가 존재할 경우, 문서 모델에 부합한다면 반복되는 요소를 부모로 하고, 반복되는 요소 사이에 있는 노드들을 자식으로 하는 계층 구조를 생성한다.

그림 4 제안된 논리적 그룹화 알고리즘

또한 <LI>, <DD> 등의 HTML 리스트 아이템 태그에 포함되어있는 내용은 논리적으로 독립된 단위로 간주하여 첫 번째 요소를 부모로 하는 단일의 그룹을 형성한다. 또한, 반복되는 논리적 요소를 기준으로 계층 구조를 생성한다. 논리적 계층 구조를 정확히 생성하기 위하여 반복되는 요소나 리스트 아이템 태그를 기준으로 그룹화할 때 그룹이 제안된 문서 모델에 부합하는지의 여부를 검사한다. 제안된 방법에 대한 자세한 기술은 <그림 4>와 같다.

4.3 후처리

후처리 과정은 논리적 구조분석 과정에서 처리하지 못한 요소를 추가로 식별하며 최종적으로 XML 문서를 생성한다. 문서

모델에 필수 요소로 기술되어 있지만 논리적 구조분석 과정에서 식별하지 못한 부 구조를 추가하거나 이전 단계에서 식별하지 못한 요소를 식별한다.

논리적 구조분석 과정에서도 이렇게 자세하게 처리해 줄 수도 있지만, 모든 텍스트를 대상으로 정규 표현을 인식하여 추출하도록 하면 시스템에 부하가 많이 걸리고, 모든 부분에서 같은 정규표현식이 적용되는 것도 아니기 때문에 후처리 과정을 두어 처리한다. 우선 정규 표현식에 대해 상위 노드부터 검색하여 정규 표현식이 있을 수 있는 노드만을 검색하여 정규 표현식을 적용할 텍스트를 찾는다. 정규 표현식을 적용할 수 있는 노드가 검색이 되면 각 노드들에 정규 표현식을 적용하여 노드를 식별한다. 이러한 정규 표현식 식별 과정이 끝나면 전술한 논리적 그룹화를 다시 한 번 적용한다.

5. 실험 결과

제안된 방법의 성능을 평가하기 위하여 Chung[5] 등의 연구에서 사용한 50개의 이력서 HTML 문서를 대상으로 실험하였다.

표 1 성능 평가 기준

기준		정의
비주얼 그룹화의 정확성		정확하게 추출된 비주얼 그룹의 수/추출된 비주얼 그룹의 수
요소의 식별률	정확률	정확하게 식별된 요소의 수/식별된 요소의 수
	재현률	정확하게 식별된 요소의 수/검증용 데이터의 요소 수
계층 구조의 정확성		$\left(1-\frac{\text{편집스크립트}}{\text{추출된 계층 구조 + 정확한 계층 구조}}\right) \times 100$

본 논문에서는 구조분석의 정확성 측면에서 제안된 방법의 성능을 분석하였다. 본 논문에서는 제안된 방법의 성능을 평가하기 위하여 <표 1>과 같이 비주얼 그룹화의 정확성, 논리적 구성 요소의 식별률, 그리고 논리적 계층 구조의 정확성의 세 가지 평가 기준을 정의한다.

먼저 비주얼 그룹화의 정확성 면에서, 제안된 방법은 94.10%의 정확률을 보였다. 비주얼 그룹화에서는 시각적으로 보기에는 그룹화 할 수 있지만, 제안된 비주얼 태그를 적용하여 비주얼 그룹화할 수 없는 경우와 HTML 문서를 작성자가 임의의 불규칙한 태그를 사용하여 비주얼 그룹을 표현한 경우에 대부분의 오류가 발생하였다. 그 밖에 비주얼 태그가 그룹화 이외의 용도로 사용되어 적절히 그룹화되지 못한 경우가 존재하였다.

요소의 식별률 면에 있어서 제안된 방법은 93.21%의 정확률과 91.62%의 재현률을 보였다. 요소 식별에서는 비주얼 그룹화가 되지 않아 요소 식별 정보를 이용하지 못한 경우와 문서에 띄어쓰기 등의 편집오류가 있는 경우에 오류가 발생했다.

마지막으로, 논리적 계층 구조의 정확성을 실험하기 위하여 구조분석의 결과로 생성된 계층 구조와 정확한 구조 사이의 구조적 차이를 계산하였다. 이를 위하여 트리 간의 편집 스크립트(edit script)를 계산하는 기존 연구인 Zhang과 Shasha의 방법[11]을 적용하였다.

실험 결과, 계층 구조의 정확성은 평균적으로 92.02%로 나타났다. 오류의 대부분은 제안된 방법의 비주얼 그룹화와 요소 식별 과정에서 찾을 수 있었다.

6. 결론 및 향후 연구 방향

XML은 논리적인 구조 정보를 표현할 수 있으며 이 기종간의 호환이 가능하다는 장점 때문에 전자 문서의 표준 포맷으로 널리 사용되고 있다. 따라서, 본 논문에서는 웹 문서로부터 XML 문서를 생성하기 위한 논리적 구조분석 방법을 제안한다.

제안된 방법은 정교한 수준의 구조분석을 위하여 문서 모델을 효과적으로 기술할 수 있는 언어인 MEDL을 제안한다. MEDL은 논리적인 계층 구조를 기술하기 위하여 특정 주제에 속하는 문서 집합이 포함할 수 있는 논리적 구성 요소의 종류, 포함관계, 그리고 빈도수 등에 대한 다양한 정보를 기술한다.

제안된 방법은 논리적 계층구조의 분석을 위해 비주얼 그룹화, 요소식별, 그리고 논리적 그룹화의 세 단계로 구성된다. 우선 비주얼 그룹화에서는 시각적으로 그룹화할 수 있는 정보를 갖는 태그들을 비주얼 태그라 정의하고 전처리를 거친 DOM 트리를 하향식 너비 우선 탐색을 적용하여 비주얼 태그들 사이의 내용을 그룹화하여 비주얼 트리를 생성한다. 요소 식별은 비주얼 트리에 문서 모델에 정의된 요소 식별 정보를 적용하여 키워드를 기반으로 하여 요소를 식별한다. 또한, 식별된 요소들의 반복과 리스트 아이템 태그들의 특징을 이용하여 논리적 계층구조로 그룹화한다. 제안된 방법의 성능을 분석하기 위해 비주얼 그룹화의 정확성, 논리적 구성 요소의 식별률, 그리고 논리적 계층 구조의 정확성의 세 가지 평가 기준으로 성능을 평가한 결과, 논리적인 구조분석을 성공적으로 수행하였다.

한편, 제안된 방법의 오류를 분석한 결과, 비주얼 그룹화가 성공적으로 이루어지지 않을 경우, 잘못된 계층 구조를 생성함을 볼 수 있었다. 또한, 정확한 요소 식별을 위해서 보다 정교한 수준의 논리적 요소 기술 방법이 요구된다. 따라서, 향후 본 연구에서는 보다 정교한 수준의 비주얼 그룹화 및 요소 식별 방법을 연구할 계획이다.

참고문헌

[1] World Wide Web Consortium, Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, http://www.w3c.org/TR/REC-xml, 2000.

[2] K. M. Summers, "Toward a Taxonomy of Logical Document Structures," Proc. Dartmouth Inst. for Advanced Graduate Studies (DAGS '95) pp. 124-133, May 1995.

[3] Seung Jin Lim and Yiu-Kai Ng, "A Heuristic Approach for Converting HTML Documents to XML Documents," Computational Logic, pp. 1181-1196, 2000.

[4] David Taniar, Y. Jiang, J. Wenny Rahayu, and L. Bishay, "Structured Web Pages Management for Efficient Data Retrieval," Proc. Int'l Conf. Web Information Systems Engineering, pp. 97-104 2000.

[5] Christina Yip Chung, Michael Gertz, and Neel Sundaresan, "Quixote: Building XML Repositories from Topic Specific Web Documents," Proc. Int'l Conf. WebDB, pp. 103-108, 2001.

[6] Wei Han, David Buttler, and Calton Pu, "Wrapping Web Data into XML," SIGMOD Record, vol. 30, no. 3, pp. 33-38, 2001.

[7] 김승원, 민준기, 정진완, "사용자와의 상호작용을 통한 HTML 문서의 XML 문서로의 변환", 정보과학회추계학술발표대회 논문집,pp. 103-105, 2002.

[8] 오금용, 황인준, "유사 패턴을 갖는 HTML 문서의 XML 자동 변환", 한국정보처리학회논문지, 제9-D권, 제3호, pp. 355-364, 2002.

[9] World Wide Web Consortium, XML DOM Level 3 Core, http://www.w3.org/TR/2003/WD-DOM-Level-3-Core, W3C Candidate Recommendation, Feb. 2003.

[10] Dave Raggett, "Clean up Your Web Pages with HP's HTML Tidy," Computer Networks and ISDN Systems, vol. 30, pp. 730-732, Apr. 1998.

[11] K. Zhang and D. Shasha, "Simple Fast Algorithms for the Editing Distance between Trees and Related Problems," SIAM Journal on Computing, vol. 18, no. 6, pp. 1245-1262, 1989.