HTML 테이블의 논리적 구조분석을 위한 효율적인 방법

김연석° 이경호
연세대학교 컴퓨터과학과
yskim°@icl.yonsei.ac.kr, khlee@cs.yonsei.ac.kr

An Efficient Method for Logical Structure Ananlysis of HTML Tables

Yeon-Seok Kim° Kyong-Ho Lee
Dept. of Computer Science, Yonsei University

요 약

본 논문에서는 웹 문서로부터 정보를 추출하기 위한 목적의 일환으로 HTML 테이블의 논리적인 구조를 추출하여 XML 문서로 변환하는 효율적인 방법을 제안한다. 제안된 방법은 영역구분과 구조분석의 두 단계로 구성된다. 영역구분 단계에서는 테이블의 잡음영역을 제거하고 정규화한 후 시각적 및 의미적 일관성 검사를 통하여 테이블에 존재하는 속성 및 값 영역을 구분한다. 또한 구조분석 단계에서는 구분된 영역에 제안된 테이블 모델을 적용하여 계층구조를 추출하며, 이로부터 XML 문서를 생성한다. 제안된 영역구분 방법의 성능을 평가하기 위하여 1,180개의 테이블을 대상으로 실험한 결과, 평균적으로 86.7%의 정확률을 보여 기존 연구보다 우수하였다.

1. 서론

웹 문서 표준인 HTML(Hypertext Markup Language)은 웹 문서를 시각적으로 렌더링하기 위한 포맷이기 때문에 컴퓨터로 하여금 정보를 처리하게 한다는 측면에서 한계를 갖는다. 반면 XML(eXtensible Markup Language)은 논리적 구조정보를 표현할 수 있으며 플랫폼에 독립적이라는 장점 때문에 다양한 분야에서 정보의 공유 및 교환을 위한 표준으로 널리 사용되고 있다. 따라서 HTML로부터 유용한 정보를 추출하고 이를 XML 형태로 변환하는 방법이 필요하다. 특히 HTML은 연관된 정보(relational information)를 간결하게 표현하기 위하여 테이블을 정의하는데, 본 논문에서는 웹으로부터 유용한 정보를 추출하기 위한 목적의 일환으로 HTML 테이블의 논리적 구조를 분석한 후 이를 XML 문서로 변환하는 효율적인 방법을 제안한다.

본 논문에서는 테이블을 연관성을 갖는 데이터의 배열이라고 정의하며 속성(attribute)과 값(value)의 연관관계를 포함하는 테이블을 진짜 테이블(genuine table)로 간주한다. 특히 속성을 나타내는 셀의 집합을 속성 영역(attribute area), 값을 나타내는 셀의 집합을 값 영역(value area) 이라고 정의한다. 한편, 제안된 방법은 테이블을 단순 테이블(simple table)과 합성 테이블(complex table)로 분류한다. 단순 테이블은 다시 속성 영역의 위치에 따라 열 방향 테이블(column-wise table), 행 방향 테이블(row-wise table), 그리고 타임 테이블(time table)로 분류되며, 합성 테이블은 하나 이상의 단순 테이블을 포함하는 복합 테이블(composite table)[6]과 단일 셀에 속성과 값이 함께 존재하는 혼합-셀 테이블(mix-cell table)[9]로 분류된다.

일반적으로 기존에 HTML 테이블로부터 정보를 추출하기 위한 연구가 다수 진행되어 왔다. 그러나 기존 연구의 대부분은 특정 온톨로지에 의존적이거나 단순한 규칙에 기반하기 때문에 다양한 종류의 테이블에 적용하기에 한계를 가진다[1-10] [16-18]. 또한 잡음 영역을 제거하지 않기 때문에 논리적 연관관계를 정확히 추출하는데 있어 제한적이다.

본 논문에서는 제안된 영역 구분 방법의 성능평가를 위하여 1,180개의 HTML 테이블을 대상으로 실험한 결과, 86.7%의 정확률을 보여 기존 연구보다 우수하였다.

2. 관련연구

HTML 테이블에 관한 기존 연구는 크게 웹으로부터 진짜 테이블을 식별하는 연구와 식별된 테이블로부터 논리적 구조를 분석하여 속성-값 연관관계를 추출하는 연구로 나뉘어 진다[1,7,9,11,13]. 특히 HTML 테이블의 논리적 구조분석은 포맷팅, 레이아웃, 그리고 온톨로지의 세 가지 정보를 사용한다. 포맷팅 및 레이아웃 정보는 각각 데이터와 테이블의 모양을 나타내는 정보로서 시각적 차이에 의해 속성과 값 영역을 구분하기 위하여 사용된다. 또한 속성과 값을 정의한 온톨로지는 테이블의 의미적 특징을 사용하여 영역을 구분하는데 사용된다. 표 1은 HTML 테이블의 구조분석에 관한 기존 연구의 특징을 요약한 것이다.

표 1. HTML 테이블의 구조분석 방법

관련 논문	특 징	XML 생성
[1]	7개의 휴리스틱한 규칙을 사용하여 영역 구분	×
[2]	레이아웃 및 컨텐츠 타입을 사용하여 계층구조 생성	×
[3-6]	포맷팅 정보에 기반하여 속성 및 값 영역 추출	○
[7]	107개의 규칙을 사용하여 행과 열 간의 유사도 계산	×
[8]	SVM과 HMM 기법 적용	○
[9]	3개의 휴리스틱한 규칙을 사용하여 영역 구분	×
[10]	온톨로지와 HMM 기법 적용	×
[11-14]	레이아웃 및 구문적 일관성의 차이를 이용하여 영역 구분	×
[15]	컨텐츠 인식과 구조 인식의 두 단계를 통하여 테이블의 구조인식	×
[16-17]	태그 TH와 TD를 사용하여 영역 구분	○
[18]	온톨로지를 사용한 영역 구분	×
[19]	태그속성 span에 기반에 기반하여 속성과 값의 쌍 추출	×

3. 제안된 구조분석 및 XML 변환 방법

제안된 방법은 그림 1과 같이 영역구분과 구조분석의 두 단계로 구성된다. 특히 영역구분 단계는 전처리, 시각적 일관성 검사, 의미적 일관성 검사, 그리고 후처리의 네 부분으로, 구조분석 단계는 테이블 모델에 기반한 구조분석과 XML 변환의 두

부분으로 이루어진다. 각 단계에 대한 자세한 설명은 다음과 같다.

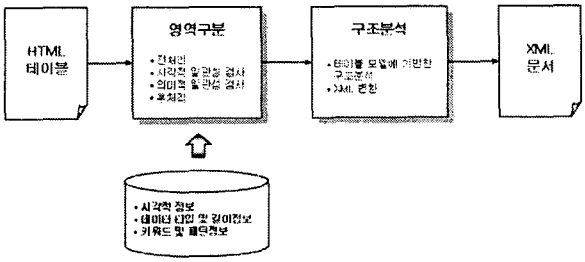

그림 1. 테이블의 구조분석 및 XML 변환 방법

3.1 영역구분

영역구분 단계에서는 전처리를 통하여 테이블의 잡음영역을 제거하고 정규화한 후 시각적 및 의미적 일관성 검사를 통하여 영역을 구분한다. 또한 합성 테이블의 영역구분을 위하여 후처리 과정을 거친다.

3.1.1 전처리

전처리 과정은 잡음영역을 제거하는 한편, 다수의 셀이 통합되어 표현된 테이블을 정규화 한다. 이를 위하여 아래와 같은 4개의 규칙을 적용한다.

① IF 행(또는 열)의 데이터가 존재하지 않음
THEN 해당 행(또는 열) 제거
② IF 첫 번째 행(또는 열)의 태그속성 스팬 값이 열(혹은 행)의 수와 동일
THEN 해당 행(또는 열)을 캡션으로 간주
③ IF 태그속성 스팬의 값이 1이 아님
THEN 해당 셀 데이터를 복사하여 셀 추가
④ IF 셀들의 데이터가 일정한 패턴을 가지고 반복됨
THEN 패턴에 따라 셀 분리

3.1.2 시각적 일관성 검사

일반적으로 <THEAD>, <TBODY>, 그리고 <TFOOT> 등과 같은 태그들은 그 자체로 영역을 구분할 수 있기 때문에 제안된 방법은 우선 위 태그들을 식별한다. 태그 식별에 의한 영역 구분이 불가능할 경우에는 포맷팅 및 구문적 일관성 검사 [20] 를 통하여 영역을 구분한다. 포맷팅 및 구문적 일관성 검사란 테이블의 속성과 값을 나타내는 셀이 각각 다른 포맷팅 및 타입과 길이를 가진다는 사실에 기반한다. 특히 포맷팅 일관성은 포맷을 결정하는 태그가 행(혹은 열)을 구성하는 셀에 사용된 빈도수를 나타낸 값으로 제안된 방법은 행(또는 열)간의 일관성 차이를 이용하여 영역을 구분한다. 한편 구문적 일관성은 셀들의 타입 및 길이 일관성을 나타낸 값으로, 테이블의 우하단 셀을 기준으로 아래에서 위(bottom-up)로 그리고

표 2. 포맷팅 일관성 검사를 위한 주요 태그

종류	의미
<i>, , <var>, <cite>	글자를 이탤릭체로 표현함
, <b>	글자를 굵게 나타냄
<u>	밑줄 그은 글자체로 표현
<h1>~<h6>	글자 크기를 변경
<big>	기본 글꼴보다 한 단계 큼
<font color="">	글자 색을 설정
<font size="">	글자 크기를 설정
<font face="">	글꼴을 변경
<tr bgcolor="">, <td bgcolor="">	행 혹은 셀의 배경색을 설정

오른쪽에서 왼쪽(right-left)으로 이웃 셀과 일관성을 계산한 후 그 차이를 이용하여 영역을 구분한다. 포맷팅 일관성을 위하여 사용된 태그는 표 2와 같다.

3.1.3 의미적 일관성 검사

시각적 일관성 검사에 의해 영역이 구분되지 않은 테이블과 2×n 열 방향 및 n×2 행 방향, n≥1, 테이블은 셀 데이터의 의미적 일관성을 사용하여 영역을 구분한다. 의미적 일관성 검사란 임의의 속성 값으로 올 수 있는 키워드 및 패턴 정보를 이용하여 대응하는 속성과 값이 의미적으로 부합하는지를 판단하는 검사로서 부합하는 속성과 값의 위치에 따라 속성과 값 영역을 구분한다.

3.1.4 후처리

후처리 단계에서는 속성과 값이 단일 셀에 존재하는 혼합-셀 테이블의 영역 구분과 함께 합성 테이블의 여부 판단 및 단순 테이블로의 분리 역할을 담당한다. 제안된 방법은 우선 혼합-셀 테이블의 여부를 판단하기 위하여 ':' 과 '=' 등의 구분자를 식별한다. 만약 테이블 내의 모든 셀에 구분자가 존재하고, 모든 셀의 구분자 좌우 데이터 타입이 각각 동일한 경우 제안된 방법은 해당 테이블을 혼합-셀 테이블로 간주하여 구분자를 중심으로 속성 및 값 쌍을 추출한다.

한편 제안된 방법은 혼합-셀 테이블을 제외한 모든 테이블에 대하여 합성 테이블 여부를 판단한다. 이를 위하여 테이블의 추출된 속성 영역의 개수를 사용하는데, 속성 영역이 2개 이상인 경우 합성 테이블로 판단하여 이를 기준으로 다수의 단순 테이블로 분리한다. 만약 추출된 속성 영역이 1개인 테이블은 태그속성 스팬을 사용하여 복합 테이블의 여부를 판단하고 스팬을 기준으로 다수의 단순 테이블로 분리한다. 이때, 단순 테이블의 속성 영역이 존재하지 않으면 합성 테이블의 속성 영역을 복사한다. 입력된 테이블이 1개의 속성 영역을 가지며, 복합 테이블이 아니라면 제안된 방법은 이를 단순 테이블로 간주한다.

3.2 구조분석

구조분석 단계에서는 영역이 구분된 테이블에 제안된 테이블 모델을 적용하여 적절한 계층구조를 생성하고, 이후 XML 변환 단계에서 깊이우선 탐색 및 병합규칙을 적용하여 XML 문서를 생성한다.

3.2.1 테이블 모델에 기반한 구조분석

테이블 모델에 기반한 구조분석 단계에서는 테이블의 구조 및 속성의 위치에 따라 열 방향, 2×n 열 방향, 행 방향, n×2 행 방향, 그리고 타임 테이블의 5가지로 분류한 후 각각 제안된 테이블 모델을 적용하여 계층구조를 추출한다. 테이블 모델(table model)이란 셀들이 갖는 계층구조를 도식화한 것으로서, 속성 $H_1,\cdots,H_n$ 과 값 $(D_{1,1},\cdots,D_{1,n}),\cdots,(D_{m,1},\cdots,D_{m,n})$을 가진다.

3.2.2 XML 변환

XML 문서는 적격성(well-formed) 요건, 즉, XML 선언 및 모든 태그가 중첩되지 않아야 하는 등의 조건,을 만족해야 하므로, 이를 위하여 제안된 방법은 깊이우선 탐색(depth first search) 기법을 사용한다. 우선, HTML 테이블을 XML 문서로 변환하기 위해서 XML 선언부를 정의한다. XML 문서의 선언부는 "<?xml ?>" 형태를 가지는 문장으로 해당 문서의 버전 및 인코딩 방식 등을 설명하며, 항상 문두에 위치한다. XML 문서를 선언한 다음에는 생성된 계층구조를 깊이우선 탐색하여 적

절한 XML 요소를 생성한다. 이때 스택을 사용하여 적절한 시작 태그와 끝 태그를 생성함으로서 적격성 요건을 만족하도록 한다. 또한 생성된 XML 문서의 불필요한 요소의 중복 제거를 위하여 Li 등 [3-6] 이 제안한 병합규칙을 사용한다.

4. 실험결과

제안된 방법의 성능을 평가하기 위하여 Hu [11-14] 의 연구에서 사용한 11,477개의 테이블 중 1,180개의 진짜 테이블을 대상으로 실험하였다. 제안된 방법의 성능은 표 3과 같다.

표 3. 성능 평가

구분	테이블 개수	정답 개수	오류 개수	정확률
열 방향	857	788	69	91.59%
행 방향	143	117	26	81.82%
타임	18	18	0	100%
합성	162	100	62	61.73%
합계	1180	1023	157	86.69%

제안된 방법은 평균 86.69%의 정확률을 보여 우수한 결과를 보였다. 이는 본 논문이 테이블의 구조를 분석하기 위하여 보다 체계적이며 정교한 방법에 기반하기 때문이다. 제안된 방법은 먼저 전처리 단계로서 잡음 영역을 제거하고 테이블을 정규화 한다. 또한 시각적 일관성은 물론이고 일관성 검사가 어려운 테이블에 대하여 의미적 일관성을 검사함으로서 보다 정교한 영역 구분이 가능하다. 표 4는 제안된 방법의 오류분석의 결과이다.

표 4. 오류분석 결과

구분	오류내용	개수
속성 → 값	포맷팅 일관성이 없음	3
값 → 속성	포맷팅 일관성이 없음	125
	구문적 일관성이 없음	3
구분불가	합성 테이블에서 각 테이블의 속성이 모두 다름	8
	2×n, n×2, 2×2 테이블의 시각적, 의미적 일관성이 없음	8
	합성 테이블로서 다양한 형태를 가진 테이블이 섞여있음	8
	비정상적인 테이블 편집으로 인한 오류	2
합계		157

5. 결론 및 향후연구

최근 들어 웹을 통하여 새롭게 생성되는 정보의 양이 급속도로 증가하면서 웹으로부터 유용한 정보를 추출하는데 많은 관심이 모아지고 있다. 특히 테이블은 연관된 정보를 효과적으로 표현할 수 있는 대표적인 방법으로서 폭넓게 사용된다. 한편, HTML은 문서를 시각적으로 랜더링 하기위한 용도로 제안된 포맷이기 때문에 컴퓨터로 하여금 유용한 정보를 추출 및 재가공 하기에는 부적합하다. 이를 위하여 논리적인 구조정보를 표현할 수 있는 XML 문서로의 변환이 필수적이다. 본 논문에서는 HTML 문서에 포함된 진짜 테이블의 영역을 구분하고 구조를 분석하여 XML 문서로 변환하는 효과적인 방법을 제안하였다. 향후 본 연구에서는 기 구축된 온톨로지 정보나 WordNet 등을 활용하여 보다 정교한 영역을 구분함과 동시에 정보의 재사용 및 자동 온톨로지 구축, 정보 검색 시스템 등의 다양한 응용분야에의 활용에 대하여 연구를 진행 할 예정이다.

참고문헌

[1] S.-W. Jung and H.-C. Kwon, "A Scalable Hybrid Approach for Extracting Head Components from Web Tables," IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 2, pp. 174-187, 2006.

[2] A. Pivk, P. Cimiano, and Y. Sure, "From Tables to Frames," Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 3, Issue 2-3, pp. 132-146, 2005.

[3] S. Li, M. Liu, G. Wang, and Z. Peng, "Capturing Semantic Hierarchies to Perform Meaningful Integration in HTML Tables," Proc. The Asia Pacific Web Conference, pp. 899-902, 2004.

[4] S. Li, Z. Peng, and M. Liu, "Extraction and Integration Information in HTML Tables," Proc. Fourth Int'l Conf. Computer and Information Technology, pp. 315-320, 2004.

[5] S. Li, M. Liu, T.-W. Ling, and Z. Peng, "Automatic HTML to XML Conversion," Proc. Fifth Int'l Conf. Web-Age Information Management, pp. 714-719, 2004.

[6] S. Li, M. Liu, and Z. Peng, "Wrapping HTML Tables into XML," Proc. Fifth Int'l Conf. Web Information Systems Engineering, pp. 147-152, 2004.

[7] H. Masuda, S. Tsukamoto, and H. Nakagawa, "Recognition of HTML Table Structure," Proc. First Int'l *Joint Conf. Natural Language Processing*, pp. 183-188, 2004.

[8] K. Itai, A. Takasu, and J. Adachi, "Information Extraction from HTML Pages and Its Integration," Proc. Int'l Symposium on Applications and the Internet, pp. 276-281, 2003.

[9] Y. Yang and W.-S. Luk, "A Framework for Web Table Mining," *Proc. Fourth Int'l Workshop on Web Information* and Data Management, pp. 36-42, 2002.

[10] M. Yoshida, "Extracting Attributes and Their Values from Web Pages," Proc. The ACL-02 Student Research Workshop, pp. 72-77, 2002.

[11] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Evaluating the Performance of Table Processing Algorithms," *Int'l Journal on Document Analysis and* Recognition, Vol. 4, No. 3, pp. 140-153, 2002.

[12] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Experiments in Table Recognition," Second Int'l Workshop on Document Layout Interpretation and its Applications, 2001.

[13] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Table Structure Recognition and Its Evaluation," *Proc. SPIE* Document recognition and Retrieval Ⅷ, Vol. 4307, pp. 44-55, 2001.

[14] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "A System for Understanding and Reformulating Tables," Proc. Fourth IAPR Int'l Workshop on Document Analysis Systems, pp. 361-372, 2000.

[15] H. Masuda, D. Yasutomi, and Hiroshi Nakagawa, "How to Transform Tables in HTML for Displaying on Mobile Terminals," Proc. 6th NLPRS Workshop of Automatic Paraphrasing: Theories and Applications, pp. 29-36, 2001.

[16] S.-J. Lim and Y.-K. Ng, "A Heuristic Approach for Converting HTML Documents to XML Documents," Proc. First Int'l Conf. Computational Logic, pp. 1182-1196, 2000.

[17] S.-J. Lim and Y.-K. Ng, "An Automated Approach for Retrieving Hierarchical Data from HTML Tables," Proc. 8th Int'l Conf. Information and Knowledge Management, pp. 466-474, 1999.

[18] H.-L. Wang, S.-H. Wu, I.-C. Wang, C.-L. Sung, W.-L. Hsu, and W.-K. Shih, "Semantic Search on Internet Tabular Information Extraction for Answering Queries," Proc. *Ninth Int'l Conf. Information and Knowledge Management*, pp. 243-249, 2000.

[19] H.-H. Chen, S.-C. Tsai, and J.-H. Tsai, "Mining Tables from Large Scale HTML Texts," Proc. 18th Int'l Conf. Computational Linguistics, pp. 166-172, 2000.

[20] Y.-S. Kim and K.-H. Lee, "Detecting Tables in Web Documents," Engineering Applications of Artificial *Intelligence*, Vol. 18, No. 6, pp. 745-757, 2005.