HTML 문서의 테이블 식별

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

Detecting Tables in HTML Documents

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

요 약

HTML의 <TABLE> 태그는 연관된 정보를 기술하기 위한 테이블은 물론이고 웹 문서의 레이아웃을 표현하기 위하여 사용된다. 본 논문에서는 웹으로부터 유용한 정보를 추출하기 위한 목적의 일환으로 HTML 문서로부터 테이블을 식별하는 효율적인 방법을 제안한다. 제안된 방법은 전처리와 속성-값 연관관계 추출의 두 단계로 구성된다. 전처리 단계에서는 진짜 테이블 또는 레이아웃용으로 사용된 <TABLE> 태그의 일반적인 특징을 반영한 규칙을 적용하여 진짜 또는 가짜로 명확히 식별이 가능한 <TABLE> 태그를 추출한다. 속성-값 연관관계 추출 단계에서는 테이블 영역을 속성 및 값 영역으로 구분한 후, 값 영역에 대하여 구문적 일관성 검사를 수행한다. 또한 값 영역의 크기가 작아서 구문적 일관성 검사를 수행할 수 없는 경우, 속성-값 영역의 의미적 일관성을 검사한다. 제안된 방법의 성능을 평가하기 위하여 1,393개의 HTML 문서로부터 추출한 11,477개의 <TABLE> 태그를 대상으로 실험한 결과, 평균적으로 97.54%의 정확률과 99.22%의 재현률을 보여 기존 연구보다 우수하였다.

1. 서 론

최근 들어 웹이 일상생활에서 보편적으로 사용됨에 따라 웹 문서의 양이 급증하고 있다. 그러나 HTML (Hypertext Markup Language)은 웹 문서를 사용자에게 보이기 위한, 즉 시각적으로 렌더링하기 위한 포맷이기 때문에 컴퓨터로 하여금 정보를 처리하게 한다는 측면에서 한계를 갖는다. 따라서 HTML 문서로부터 유용한 정보를 추출하는 방법은 많은 분야에서 주요 관심사가 되고 있다[1].

일반적으로 테이블은 연관된 정보 (relational information)를 구조적이며 간결하게 표현할 수 있는 방법이다. 본 논문에서는 테이블을 연관성을 갖는 데이터의 배열이라고 정의하며, [2]와 마찬가지로 속성 (attribute)과 값 (value)의 관계를 포함하는 테이블을 진짜 테이블 (genuine table)로 간주한다.

HTML 문서의 테이블 식별에 관한 연구는 크게 특정 도메인에 의존적인 방법과 도메인에 독립적인 방법의 두 가지로 나누어진다[2]. 도메인에 의존적인 방법은 특정 도메인에 대한 정보를 이용하여 테이블을 식별하는 방법이며, 도메인 독립적인 방법은 임의의 HTML 문서에 포함된 테이블을 도메인 정보에 상관없이 식별하는 방법이다. 그러나 이러한 연구들은 특정 어휘정보 혹은 패턴에 의존적이거나 많은 학습을 필요로 한다는 점에서 개선의 여지가 있다[3~7].

본 논문에서는 제안된 방법의 성능을 평가하기 위하여 1,393개의 HTML 문서로부터 추출한 11,477개의 <TABLE> 태그를 대상으로 실험하였으며 그 결과 평균 97.54%의 정확률과 99.22%의 재현률을 보여 기존 연구보다 우수하였다.

2. 테이블 식별방법

제안된 방법은 그림 1과 같이 전처리와 속성-값 연관관계 추출의 두 단계로 구성된다.

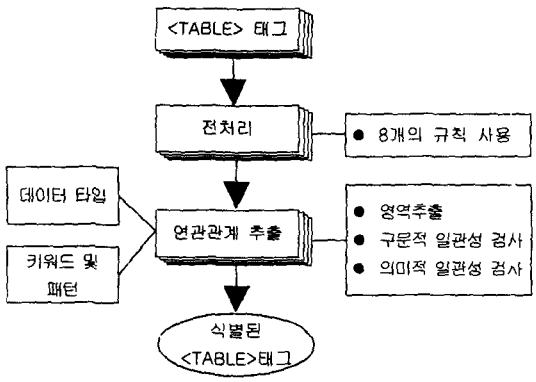

그림 1. 제안된 테이블 식별 방법

HTML 문서로부터 추출된 <TABLE> 태그에 8개의 규칙을 적용하여 일차적으로 테이블을 식별한 다음, 식별하지 못한 테이블에 대하여 영역 추출 단계에서 테이블을 속성 영역과 값 영역으로 구분한다. 이렇게 구분된 값 영역에 대하여 세로 또는 가로 방향으로 구문적 일관성이 존재하는지의 여부를 검사한다. 특히 제안된 방법은 값 영역의 데이터 타입 (data type)과 길이 (length) 정보를 기반으로 일관성을 검사한다. 구문적 일관성 검사를 통하여 식별이 어려운 경우 의미적 일관성 검사를 통하여 값 영역에 포함된 내용이 해당 속성과 일관성을 갖는지를 판별한

다. 이를 위하여 값 영역의 데이터 타입과 특정 속성의 값으로 올 수 있는 키워드 또는 패턴 등의 부가정보를 이용한다. 각 단계에 대한 자세한 설명은 다음과 같다.

2.1 전처리

제안된 전처리 과정은 진짜 또는 가짜 테이블이 갖는 일반적인 특징을 기준으로 테이블을 식별한다. 이를 위하여 아래와 같은 8개의 규칙을 적용한다.

① <CAPTION> 태그가 존재하면 진짜 테이블이다.
② 1×1테이블은 box를 표현하기 위한 가짜 테이블이다.
③ 문자가 존재하지 않는 테이블은 가짜 테이블이다.
④ 대부분의 셀이 하이퍼링크로 이루어진 테이블은 레이아웃을 위한 가짜 테이블이다.
⑤ 대부분의 셀이 이미지로 이루어진 테이블은 레이아웃을 위한 가짜 테이블이다.
⑥ 대부분의 셀이 공백으로 이루어진 테이블은 레이아웃을 위한 가짜 테이블이다.
⑦ <th> 태그와 이에 상응하는 <td> 태그가 존재하는 테이블은 진짜 테이블이다.
⑧ 테이블을 포함하는 테이블은 가짜 테이블이다.

2.2 연관관계 추출

연관관계 추출 단계는 전처리 단계에서 진짜 혹은 가짜 테이블로 판별되지 않은 <TABLE> 태그를 대상으로 한다. 제안된 연관관계 추출 방법은 영역 추출, 값 영역의 구문적 일관성 검사, 그리고 속성-값 영역의 의미적 일관성 검사의 세 부분으로 이루어지며, 제안된 방법은 연관관계 추출에 앞서 테이블을 속성과 값 영역으로 구분한다. 연관관계 추출 알고리즘은 그림 2와 같다.

2.2.1 영역 추출

제안된 방법은 연관관계의 추출을 위하여 먼저 테이블 영역을 속성과 값 영역으로 구분한다. 즉, 주어진 테이블에 대하여 속성을 가질 수 있는 셀 영역을 속성 영역으로, 나머지 부분을 값 영역으로 구분한다.

2.2.2 구문적 일관성 검사

구문적 일관성 검사는 속성에 대응하는 값이 가로 방향 (또는 세로 방향)으로 두개 이상 존재할 경우, 추출된 값 영역에 대하여 가로 방향 (또는 세로 방향) 구문적 일관성을 검사한다. 가로 (또는 세로) 방향 구문적 일관성이란 값 영역을 구성하는 행 (또는 열) 일관성의 평균값으로 정의하며 "행 (또는 열) 방향 일관성"이라한다. 특히 각각의 행 (또는 열) 일관성을 계산하기 위하여 데이터 타입 일관성 (data type coherency)과 길이 일관성 (length coherency)을 사용한다 (식 (1)-(4) 참조). 한편 임의의 테이블에 대하여 가로와 세로의 양 방향으로 일관성 검사가 가능할 경우, 양 방향으로 일관성 검사를 수행한 후 크기가 큰 값을 테이블 일관성 (table coherency)으로 간주한다.

$$\text{행 (또는 열)방향 일관성} = \frac{\sum \text{행 (또는 열) 일관성}}{\text{행 (또는 열)의 수}} \quad (1)$$

$$\text{행 (또는 열) 일관성} = W_1 \times \text{데이터 타입 일관성} + W_2 \times \text{길이 일관성} \quad (2)$$

$$\text{데이터 타입 일관성} = \frac{\text{주요 데이터 타입을 갖는 셀의 수}}{\text{행 (또는 열)의 전체 셀 수}} \quad (3)$$

$$\text{길이 일관성} = \frac{\text{길이가 범위 } \alpha \text{ 이내에 포함되는 셀의 수}}{\text{행 (또는 열)의 전체 셀 수}} \quad (4)$$

```
입력: HTML <TABLE> 태그
출력: 테이블의 진위 여부 (IsGenuineTable)
함수 및 변수 정의.
 Boolean IsGenuineTable  ::= 진짜 또는 가짜인지 여부 저장
 Boolean IsThereSemanticCoherency()
    ::= 의미적 일관성이 존재하면 TRUE, 그렇지 않으면 FALSE
 Boolean IsThereSyntacticCoherency()
    ::= 구문적 일관성이 존재하면 TRUE, 그렇지 않으면 FALSE

방법:
1: If (테이블의 크기가 1×2, 2×1, 또는 2×2인 경우)
2:   // 속성이 단일의 값을 가지기 때문에 의미적 일관성을 검사
3:   IsGenuineTable = IsThereSemanticCoherency();
4: // 속성이 2개 이상의 값을 갖는 경우, 먼저 값 사이의 구문적
      일관성을 검사
5: // 또한 구문적 일관성 검사를 통하여 판별할 수 없는 경우를
      위해서 의미적 일관성을 검사
6: else if (테이블의 크기가 1×n 또는 n×1인 경우) {  // n>2
7:   if(IsThereSyntacticCoherency())
8:      IsGenuineTable=True;
9:   else
10:     IsGenuineTable = IsThereSemanticCoherency();
11: }
12: else if (2×n || n×2) { // n≥3
13:  if (테이블이 값 영역을 포함하지 않는 경우)
14:     IsGenuineTable = False;
15:  else {
16:     if(IsThereSyntacticCoherency()) IsGenuineTable=True;
17:     else IsGenuineTable = IsThereSemanticCoherency();
18:  }
19: }
20: else {   // 크기가 3× 3 이상인 경우
21:  if (테이블이 값 영역을 포함하지 않는 경우)
22:     IsGenuineTable = False;
23:  if (값 영역이 2개 이상의 row/column으로 이루어진 경우) {
24:     if(IsThereSyntacticCoherency())
25:        IsGenuineTable=True;
26:     else
27:        IsGenuineTable = IsThereSemanticCoherency();
28:   }
29:   else // 값 영역이 1개의 row/column으로 이루어진 경우
30:   IsGenuineTable = IsThereSemanticCoherency();
31: }
```

그림 2. 제안된 연관관계 추출 알고리즘

제안된 방법은 일관성 검사를 위하여 15개의 데이터 타입을 정의하였으며, 계산된 테이블 일관성이 임계값보다 클 경우 진짜 테이블로 식별한다. 만일 계산된 테이블 일관성이 임계값보다 적을 경우에는 해당 <TABLE> 태그에 의미적 일관성 검사를 적용하여 다시 한번 테이블의 진위 여부를 검사한다.

2.2.3 의미적 일관성 검사

추출된 값 영역에 대하여 구문적 일관성 검사를 적용할 수 없거나, 계산된 테이블 일관성이 임계값보다 적은 경우, 의미적 일관성을 추가로 검사한다. 제안된 방법은 서로 대응하는 속성과 값이 의미적으로 부합하는지의 여부를 검사하는데 이를 위하여 임의의 속성의 값으로 올 수 있는 키워드 및 패턴 정보를 정의한다. 즉, 임의의 속성-값 영역에 대하여 의미적 일관성이 존재하면 <TABLE> 태그를 진짜 테이블로 식별한다.

3. 실험결과

제안된 방법의 성능을 평가하기 위하여 Wang과 Hu[8]의 연구에서 사용한 11,477개의 <TABLE> 태그를 대상으로 실험하였다. 실험에 사용된 <TABLE> 태그는 1,675개의 진짜 테이블과 9,802개의 가짜 테이블을 포함한다. 특히 실험을 위하여 테이블 일관성의 임계값, 데이터 타입 일관성의 가중치 (W_1), 그리고 길이 일관성의 가중치 (W_2)를 각각 0.54, 0.6, 그리고 0.4로 설정하였다. 제안된 방법의 성능을 정량적으로 평가한 결과는 표 1과 같다.

표 1. 성능 평가 (%)

실험 데이터	정확률	재현률	F-measure
1,393개의 웹 페이지에서 추출한 11,477개의 <TABLE>태그	97.54	99.22	98.38

본 논문에서는 Wang과 Hu의 연구와 마찬가지로 테이블 식별을 정확률 (precision), 재현률 (recall), 그리고 F-measure의 세 가지 측면에서 제안된 방법의 성능을 분석하였다 그 결과 제안된 방법은 97.54%의 정확률과 99.22%의 재현률을 보여 97.50%의 정확률과 94.25%의 재현률을 보인 Wang과 Hu의 방법보다 우수하였다. 이는 본 논문이 테이블을 식별하기 위해서 보다 체계적이며 정교한 방법에 기반하기 때문이다. 표 2는 제안된 방법의 오류분석의 결과이다.

4. 결론 및 향후연구

최근 들어 웹을 통하여 새롭게 생성되는 정보의 양이 급속도로 증가하면서 웹으로부터 유용한 정보를 추출하는데 관심이 모아지고 있다. 특히 테이블은 연관된 정보를 효과적으로 표현하며 웹 문서표준인 HTML은 테이블의 표현을 위해서 <TABLE> 태그를 정의한다. 그러나 HTML은 본래 문서의 내용을 시각적으로 렌더링하기 위한 용도로 제안된 포맷이기 때문에 컴퓨터로 하여금 유용한 정보를 추출 및 재가공 하기에는 부적합하다. 반면, 논리적 구조 정보를 표현할 수 있는 XML (Extensible Markup Language)은 기존에 데이터로써의 가치가 떨어지며 이질적인 형

표 2. 오류분석 결과

구분	오류내용	개수
가짜 → 진짜	값 영역이 길이가 비슷한 문자열로 구성되어 구문적 일관성 존재	35
	구문적 일관성 존재	3
	의미적 일관성 존재	2
진짜 → 가짜	대부분의 셀이 하이퍼링크로 구성	4
	대부분의 셀이 이미지로 구성	1
	대부분의 셀이 공백 셀로 구성	3
	2×2 테이블이며 의미적 일관성 부재	2
	비정상적인 테이블 편집으로 인한 식별오류	3
합 계		53

태의 웹 컨텐츠를 컴퓨터 처리 및 가공이 가능한 형태로 급속하게 변화시키고 있다. 향후 본 연구에서는 식별된 테이블 정보의 효과적인 재사용을 지원하기 위하여 테이블의 논리적인 구조를 인식하여 이를 XML 형태로 변환하는 연구를 진행할 계획이다.

참고문헌

[1] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo, "Extracting Semistructured Information from the Web," Proc. PODS/SIGMOD, pp. 18~25, Tucson, Arizona, May 1997.

[2] Y. Yang, Web Table Mining and Database Discovery. MSc Thesis, Simon Fraser University, Aug. 2002.

[3] M. Hurst, "Layout and language: Challenges for Table Understanding on the Web," Proc. First Int'l Workshop on Web Document Analysis, pp. 27~30, Seattle, USA, Sep. 2001.

[4] H.-H. Chen, S.-C. Tsai and J.-H. Tsai, "Mining Tables from Large scale HTML Texts," Proc. 18th Int'l Conf. Computational Linguistics, Vol. 1. pp. 166~172, 2000.

[5] G. Penn, J. Hu, H. Luo, and R. McDonald, "Flexible Web Document Analysis for Delivery to Narrow-Bandwidth Devices," Proc. Fifth Int'l Conf. Document Analysis and Recognition(ICDAR01), pp. 1074~1078, Seattle, USA, Sep. 2001.

[6] M. Yoshida, K. Torisawa, and J. Tsujii, "A Method to Integrate Tables of the World Wide Web," Proc. First Int'l Workshop on Web Document Analysis(WDA 2001), pp. 31~34, Seattle, USA, Sep. 2001.

[7] M. Hurst, "Classifying TABLE Elements in HTML," Proc. 11th International World Wide Web Conference, Honolulu, HI, May 2002. http://www2002.org/CDROM/poster/115/index.html

[8] Y. Wang and J. Hu, "Detecting Tables in HTML Documents," Proc. 5th IAPR Int'l Workshop on Document Analysis System (DAS'02), pp. 249~260, Princeton, USA, Aug. 2002.