Research Article

pISSN : 2287-8920
https://doi.org/10.22716/sckt.2020.8.4.038

데이터 자동 추출을 위한 HTML5 테이블 구조 분석

구흥서[1]
[1]청주대학교 인공지능소프트웨어 전공 교수

A Study on Analyzing the HTML5 Table Structure for Automatic Data Extraction

HeungSeo Koo[1]
[1]Professor, Cheongju University Major in Artificial Intelligence Software

[1]**Corresponding author:** hskoo@cju.ac.kr

Received December 6, 2020; **Accepted** December 14, 2020

ABSTRACT

월드와이드웹이 정보를 출판하는 매체로 보편화되면서 웹 문서량이 폭발적으로 증가하고 있다. 웹 문서에서 HTML 테이블은 연관된 정보를 2차원 논리적 구조로 다양하게 표현할 수 있기 때문에 중요한 정보를 표현하는 수단으로 많이 사용되고 있다. 이러한 이유 때문에 HTML 테이블에서 데이터 자동 추출 방법에 대한 관심이 증가하고 있다. 본 연구는 HTML 테이블의 데이터 자동추출 문제의 해결방법으로 웹문서를 작성할 때 HTML 테이블의 코딩규약을 준수하는 방식을 제안한다. 통상적으로 사용되는 HTML 코딩규약의 목적은 웹문서 작성시 개발자들 간에 HTML 코드의 가독성을 높이고 쉽게 이해할 수 있도록 하는 것이다. 본 연구에서 제안하는 HTML 코딩규약의 목적은 웹문서의 HTML 테이블에서 데이터를 자동추출하기 용이하도록 하기 위한 것이다. 제안된 HTML 코딩규약이 활용되어 HTML 테이블에서 데이터 자동추출이 용이해질 것으로 기대된다.

As the World Wide Web becomes more common as a medium for publishing information, the volume of web documents is exploding. In web documents, HTML tables are widely used as a means of expressing important information because related information can be expressed in various ways in a two-dimensional logical structure. For this reason, interest in the method of automatically extracting data from HTML tables is increasing. This study proposes a method of complying with the coding conventions of HTML tables when creating web documents as a solution to the problem of automatic data extraction of HTML tables. The purpose of commonly used HTML coding conventions is to improve the readability of HTML code and make it easier to understand among developers when creating web documents. The purpose of the HTML coding convention proposed in this study is to facilitate automatic extraction of data from HTML tables in web documents. It is expected that the proposed HTML coding convention will be utilized to facilitate automatic data extraction from HTML tables.

Keywords: HTML table, Automatic data extraction, HTML coding convention, HTML table structure, Information extraction

1. 서 론

HTML5는 웹문서를 만들기 위한 표준 마크업 언어로 하나의 언어(JavaScript), 하나의 데이터 모델(XML, DOM), 하나의 레이아웃(CSS)을 통일적으로 제공하여 텍스트, 오디오, 비디오, 그래픽 등에 대해 통합적인 웹 플랫폼 환경을 제공한다[1]. 월드와이드웹(WWW)이 정보를 출판하는 매체로 보편화되면서 웹 문서량이 폭발적으로 증가하고 있다. 웹 문서는 표준 마크업 언어인 HTML5를 사용하여 작성하며, HTML5는 콘텐트를 브라우징 도구에 독립적으로 출판하기 위한 설계된 언어이다. 그러므로 시각적 표현을 위한 마크업 언어의 특성 때문에 정보자동화에는 한계를 갖는다[2]. 특히 웹문서에서 HTML 테이블은 연관된 정보를 2차원 논리적 구조로 다양하게 표현할 수 있기 때문에 중요한 정보를 제공하는 수단으로 많이 사용되고 있다. 이러한 이유 때문에 HTML 테이블을 자동으로 해석해서 필요한 정보를 추출하는 방법에 대한 관심이 증가하고 있다.

테이블은 연관성을 갖는 데이터의 배열이라고 정의하며, 속성과 값의 관계를 포함하는 테이블을 진짜 테이블(real table)로 간주한다[2]. HTML 문서에서 테이블의 정보를 자동 추출하는 방법에 관한 연구는 HTML 문서에서 진짜 테이블과 더미 테이블(dummy table)을 구별하는 연구와 테이블의 데이터를 정확하게 추출하기 위해 테이블 구조를 효율적으로 분석하는 방법에 관한 연구가 주류를 이루고 있다. HTML 문서에서 진짜 테이블과 더미 테이블의 판별은 유용한 정보를 2차원 구조의 논리적 형태로 표현된 진짜 테이블인지 HTML 문서의 내용을 시각적으로 보기 좋게 표현하기 위해 HTML의 table 태그가 사용된 더미 테이블인지를 판별하는 방법에 관한 연구이다. 테이블 데이터의 자동추출에 관한 연구는 HTML 테이블의 영역을 식별하고 이들 간의 관계를 분석하여 테이블의 논리적 구조를 정확하게 해석하는 방법이 핵심과제이다.

앞에서 언급한 HTML 테이블에서의 데이터 자동추출 관련 두 가지 연구주제는 다음과 같은 문제점이 존재한다. 첫째, HTML table 태그 분석에 기반한 데이터 자동추출 방식은 HTML 테이블 구조의 복잡도에 따라 테이블 구조의 정확한 분석에 한계가 있을 수 있다. 둘째, 자동추출 대상 테이블의 table 태그 구조가 변경되는 경우 데이터 자동추출 프로그램에 오류가 발생할 가능성이 높다. 이러한 경우 변경될 때마다 HTML 테이블 구조에 대해 자동추출 규칙을 새로 설계하여 프로그램을 수정해야 한다는 어려움이 있다.

본 연구는 HTML 테이블에서 데이터 자동추출 문제의 해결방안으로 웹문서를 작성할 때 HTML 테이블의 코딩규약을 준수하는 방식을 제안한다. 이를 위해 HTML5 테이블 구조와 table 태그를 분석하여 HTML 테이블의 코딩규약을 제안한다. 통상적으로 사용되는 HTML 코딩규약의 목적은 웹문서 작성시 개발자들 간에 HTML 코드의 가독성을 높이고 쉽게 이해할 수 있도록 하는 것이다. 본 연구에서 제안하는 HTML 코딩규약의 목적은 웹문서에 포함된 HTML 테이블에서 데이터를 자동추출하기 용이하도록 하기 위한 것이다.

제안된 방식은 진짜 테이블 식별문제도 기존 방식에 비해 쉽게 해결할 수 있고, 테이블에서 데이터 자동추출 프로그램 설계도 기존 방식에 비해 용이하며 복잡한 테이블 구조의 경우에도 해결 가능할 것이다. 또한 자동추출 대상 테이블의 table 태그 구조가 변경되는 경우에도, 테이블 구조의 변경을 감지하고 이에 따른 변경된 테이블 구조에서 데이터를 자동추출할 수 있도록 규칙화하는 것이 가능할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2장에서 HTML 테이블의 데이터 자동추출을 위한 관련 연구를 살펴본다. 3장에서는 웹문서에 나타낼 수 있는 HTML 테이블의 표준구조를 설명한다. 4장에서는 HTML5의 테이블 관련 요소들을 기반으로 HTML 테이블의 코딩규약을 제안한다. 그리고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

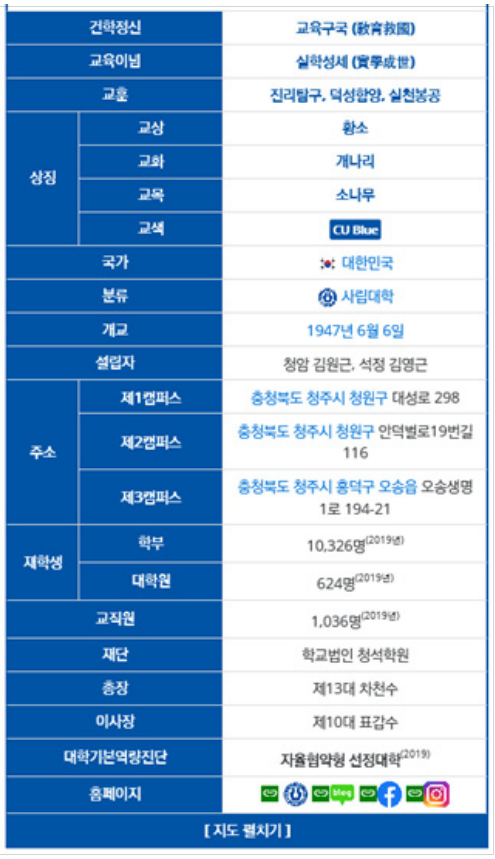
2.1 테이블 식별

상수도 HTML 테이블에 관한 연구는 크게 웹문서로부터 진짜 테이블을 식별하는 연구와 식별된 테이블로부터 논리적 구조를 분석하여 속성-값 연관관계를 추출하는 연구로 분류된다[3]. 웹문서에서 HTML의 진짜 테이블과 더미 테이블의 식별방법은 테이블 데이터의 자동추출 프로그램을 개발하는데 요구되는 핵심기술 중 하나이다. Fig. 1은 기상예보정보를 제공하는 웹사이트의 HTML 문서에 포함된 더미 테이블의 사례이다. Fig. 1의 웹페이지는 table 태그를 사용했지만 시각적인 효과를 위한 것이며 2차원 형태의 유용한 정보를 제공하기 위한 진짜 테이블이 아니다. 좀 더 자세히 살펴보면, 사용자 로그인 폼을 구성하기 위해 table 태그를 사용한 사례이며, 그림에 표시된 점선이 HTML 테이블의 셀들을 구분하여 표시된 것이다.

웹문서에서 진짜 테이블과 더미 테이블을 식별하기 위한 연구는 대부분 특정 도메인에 한정적이거나 다수의 학습 데이터를 기반으로 한다. [2]는 효율적인 테이블 식별을 위해 웹문서에 사용된 table 태그의 일반적인 특징을 추출하여 8개의 테이블 식별 규칙을 제안하였다. 대부분의 더미 테이블은 <th> 태그를 사용하지 않기 때문에 제안된 8가지 규칙으로 식별 가능하지만, <th> 태그의 존재유무로 더미 테이블을 식별하는 방법은 한계가 있다. Fig. 2는 교육기관의 정보를 제공하는 웹문서에서 <th> 태그를 사용하지 않은 진짜 테이블의 사례이다. 실제 인터넷에는 <th> 태그를 사용하지 않은 진짜 테이블을 포함하는 웹문서가 상당히 존재한다.

(source: https://www.weatheri.co.kr/)

Fig. 1. An example of the dummy table



(source: https://namu.wiki/w/청주대학교)

Fig. 2. An example of the Real table without using <th> element

2.2 테이블의 논리적 구조 분석

웹문서에서 HTML 테이블의 논리적인 구조의 정확한 분석은 테이블 데이터의 자동추출 프로그램을 개발하는데 요구되는 핵심기술 중 하나이다. HTML 테이블의 구조분석은 일반적으로 포맷팅, 레이아웃, 그리고 온톨로지의 세 가지 정보를 사용한다. 포맷팅 및 레이아웃 정보는 각각 데이터와 테이블의 정보를 나타내는 정보로서 시각적 차이에 의해 속성과 값 영역을 구분하기 위해 사용된다. 또한 속성과 값을 정의한 온톨로지는 테이블의 의미적 특징을 사용하여 영역을 구분하는데 사용된다[3]. 여러 연구[2-5]에서 제안된 방법들이 의미있는 결과를 생성했으나 복잡도가 높고 특정 도메인에 한정된다는 제약을 갖는다. 좀 더 정확한 테이블의 구조분석을 위해 다수의 연구에서 제안된 방법들이 문서모델을 표현하기 위해 DTD와 유사한 정보에 의존적이어서 범용성과 활용성이 떨어진다. 웹문서의 HTML 테이블은 다양한 구조로 출판되므로 테이블의 논리적 구조에 대한 정확한 분석은 여전히 도전과제로 남아 있다.

2.3 코딩규약

HTML 코딩 컨벤션(convention)은 컴퓨터 프로그램을 코딩을 할 때, 작성자 이외의 다른 프로그래머들도 작성자가 작성한 코드를 보고 쉽고 빠르게 이해할 수 있도록 하나의 작성표준 즉, 코딩 스타일 규약을 정한 것이다. 표준 코딩 스타일은 존재하지 않지만, 같은 기업 혹은 같은 팀 내부에서는 표준 코딩 스타일을 규정해 관리하고 사용하는 것이 소프트웨어 개발작업의 효율을 향상시키고 유지보수 비용을 감소시킬 수 있어 많이 사용된다. HTML 코딩규약의 목적은 웹문서 작성시 개발자들 간에 HTML 코드를 가독성을 높이고 쉽게 이해할 수 있도록 하는 것이다. 이와 달리 본 연구에서 제안하는 HTML 코딩규약의 목적은 웹문서에 포함된 HTML 테이블의 데이터를 자동추출하기 용이하도록 하기 위한 것이다. 본 연구에서는 웹문서 개발 측면뿐만 아니라 웹문서의 데이터 공유 측면을 고려하여 HTML의 테이블 관련 태그들의 코딩규약을 제안한다.

3. HTML 테이블 분석

[5]는 웹 문서에 나타나는 다양한 테이블의 논리적 구조분석을 위해, 테이블 유형을 다섯 가지 유형을 분류하였다. 단순 테이블과 합성 테이블로 크게 분류하고, 단순 테이블은 속성 영역의 위치에 따라 “열방향 테이블“, ”행방향 테이블“, 그리고 ”타임 테이블“로 분류하였다. 합성 테이블은 하나 이상의 단순 테이블을 포함하는 ”복합 테이블“과 단일 셀에 속성과 값이 함께 존재하는 ”혼합-셀 테이블“로 분류하였다. 이 분류 방법의 특징은 단순 테이블 유형은 테이블 제목부의 위치를 기준으로 하고, 합성 테이블 유형은 혼합-셀의 존재유무를 기준으로 하여 분류하였다. HTML5의 테이블의 구조는 다양하게 표현될 수 있지만, Fig. 3과 같은 구조의 테이블을 표준형태 테이블로 규정하였다. 이 테이블은 [5]의 “열방향 테이블“, ”행방향 테이블“, ”타임 테이블“을 모두 포함한다. 또한 Fig. 3과 같이 여러 줄의 박스제목은 <thead> 요소 내에 여러 줄의 <tr> 요소로 사용하여 표현한다. 또한 [5]의 합성 테이블은 HTML의 Fig. 3의 테이블 구조를 중첩 테이블 구조로 표현하면 구현 가능하다. 그리고 이를 바탕으로 4.1절에서 제안한 HTML 코딩규약에 따라 4.2절에서 HTML 코드의 작성 예를 제시하였다.

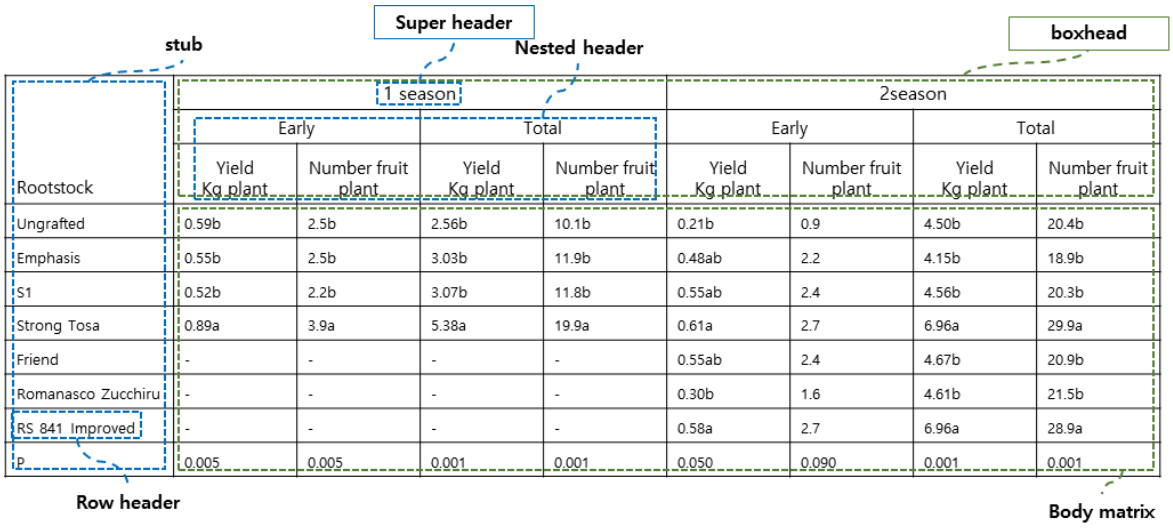


Fig. 3. A standard structure of HTML table

4. 구 현

4.1 제안된 HTML 테이블의 코딩규약

이번 절에서는 본 연구에서 제안한 HTML 코딩규약을 설명한다. Table 1은 제안한 코딩규약을 나타낸 것이다. HTML 테이블과 관련된 요소들만 포함하며 9가지 요소규약을 정의하였다. 그리고 HTML 테이블의 구조에 관련된 코딩규약을 다음 4가지 정의하였다.

- 더미 테이블(dummy table)은 테이블 제목을 나타내는 <caption> 요소를 사용하지 않는다.
- 여러 줄의 박스제목은 <thead> 요소 내에 여러 줄의 <tr> 요소로 표현한다.
- 테이블의 스텁 (Stub) 영역인 행제목은 <tbody> 요소 안에 <th> 요소로 표현한다.
- 복합 테이블은 셀제목과 값을 포함하는 하나의 셀을 하나의 단순테이블로 표현하여 테이블 전체를 중첩 테이블로 구성한다.

제안된 HTML 코딩규약의 목적은 웹문서에 포함된 HTML 테이블에서 데이터를 자동추출하기 용이하도록 하기 위한 것으로, 웹문서 개발 측면뿐만 아니라 웹문서의 데이터 공유 측면을 고려하였다.

4.2 HTML 테이블의 구현

이번 절은 4.1절에서 제안한 HTML 테이블 코딩규약의 적용 사례를 제시한다. Table 2는 3.1절의 Fig. 3의 HTML 테이블의 표준구조를 본 연구에서 제안한 코딩규약을 준수하여 구현한 예를 나타낸 것이다. Table 2의 6번~16번 줄 사이의 코드가 테이블에 포함된 각 열의 데이터 타입을 선언한 것이다. HTML 테이블의 자동추출 프로그램은 이 부분의 코드를 해석하여 테이블의 각 열에 포함된 값을 데이터 타입에 적합한 처리할 수 있을 것이다. 19번~41번 줄의 코드가 Fig. 3의 박스제목 세 줄을 구현한 것이다. 이 부분은 <thead> 요소 안에 포함되어 있기 때문에 자동추출 프로그램에서 박스제목이 세 줄로 구성됐다는 점을 쉽게 분석할 수 있을 것이다.

Table 1. HTML table coding convention

요소명	코딩 규약
<table>	2차원 테이블 구조로 데이터를 표현하기 위해 사용한다.
<caption>	테이블의 제목을 표현하기 위해 사용한다. caption 요소는 **<반드시>** 선언한다.
<colgroup>	<col> 요소를 그룹화하여 디자인을 제어할 때 사용하며, class 속성을 선언한다. class 속성에 각 열에 저장된 값의 데이터 타입을 **<반드시>** 명시한다. ☞ 사용 예 : <col class="datatype-integer"> ☞ 데이터 타입 <일반> : integer, float, string, date, time, datetime, hour, minute, second, ms ☞ 데이터 타입 <단위> : SI 국제표준단위를 사용한다. ton, kg, g, mg 등
<col>	테이블의 각 열의 너비를 지정하기 위해 선언한다.
<thead>	테이블의 머리글을 그룹화할 때 사용하는 요소로 **<반드시>** 선언한다.
<tr>	테이블의 한 행을 표현할 때 사용하는 요소로 **<반드시>** 선언한다.
<th>	scope, abbr, class, id 속성을 선언한다. scope, class 속성은 **<반드시>** 선언한다. • 테이블의 셀 머리글이 명시되지 않은 경우에도 <th> 요소를 선언하여, 열의 값의 의미를 알 수 있도록 한다. 셀 머리글을 숨김 처리하려면 CSS를 활용한다.
<tfoot>	테이블의 바닥글을 표현할 때 사용하는 요소로 **<반드시>** 선언한다. tfoot 요소는 thead와 tbody 요소 사이에 위치하도록 작성한다. • <tfoot> 요소는 열의 합계뿐만 아니라 테이블에 대한 바닥 설명글을 나타내는 용도로도 사용한다. • 바닥 설명글로 사용하는 경우는 "class" 속성에 바닥글의 유형을 **<반드시>** 선언한다. ☞ 바닥글의 유형 : date, source, desciption, legend, usage
<tbody>	테이블의 본문을 그룹화하기 위해 사용한다. 테이블에 본문이 하나인 경우도 **<반드시>** 선언한다.

Table 2. An example of implementation using the proposed HTML coding convention

```
<table id="standard-table">
<caption style="font-size:20px;font-weight:bold;">
    table Jiyong version
</caption>

<colgroup>
  <col class="datatype-stub" width="20%">
  <col class="datatype-kg" width="10%">
  <col class="datatype-float" width="10%">
  <col class="datatype-kg" width="10%">
  <col class="datatype-float" width="10%">
  <col class="datatype-kg" width="10%">
  <col class="datatype-float" width="10%">
  <col class="datatype-kg" width="10%">
  <col class="datatype-float" width="10%">
</colgroup>

<thead>
  <tr>
    <th rowspan=3>Rootstock</th>
    <th colspan=4>1 season</th>
    <th colspan=4>2 season</th>
  </tr>

  <tr>
    <th colspan=2>Early</th>
    <th colspan=2>Total</th>
    <th colspan=2>Early</th>
  </tr>

  <tr style="display: none">
    <td colspan=9 class="column-source">
        Source</td>
  </tr>

  <tr style="display: none">
    <td colspan=9 class="column-description">
        description</td>
  </tr>

  <tr style="display: none">
    <td colspan=9 class="column-legend">
        Legend</td>
  </tr>

  <tr style="display: none">
    <td colspan=9 class="column-usage">
        Usage</td>
  </tr>
</tfoot>

<tbody>
  <tr>
      <th>Ungrafted</th>
      <td>0.59b</td>
      <td>2.5b</td>
```

```
    <th colspan=2>Total</th>
  </tr>

  <tr>
    <th>Yield<br>Kg plant</th>
    <th>Number fruit<br>plant</th>
    <th>Yield<br>Kg plant</th>
    <th>Number fruit<br>plant</th>
    <th>Yield<br>Kg plant</th>
    <th>Number fruit<br>plant</th>
    <th>Yield<br>Kg plant</th>
    <th>Number fruit<br>plant</th>
  </tr>
</thead>

<tfoot>
  <tr>
    <th>P</th>
    <td>0.005</td>
    <td>0.005</td>
    <td>0.001</td>
    <td>0.001</td>
    <td>0.050</td>
    <td>0.090</td>
    <td>0.001</td>
    <td>0.001</td>
  </tr>

  <tr style="display: none">
    <td colspan=9 class="column-date">
        Date</td>
    <td>2.56b</td>
    <td>10.1b</td>
    <td>0.21b</td>
    <td>0.9</td>
    <td>4.50b</td>
    <td>20.4b</td>
  </tr>
  <tr>
    <th>Emphasis</th>
    <td>0.55b</td>
    <td>2.5b</td>
    <td>3.03b</td>
    <td>11.9b</td>
    <td>0.48ab</td>
    <td>2.2</td>
    <td>4.15b</td>
    <td>18.9b</td>
  </tr>
  <tr>
    <th>S1</th>
    <td>0.52b</td>
    <td>2.2b</td>
    <td>3.07b</td>
    <td>11.8b</td>
    <td>0.55ab</td>
    <td>2.4</td>
    <td>4.56b</td>
    <td>20.3b</td>
  </tr>
</tbody>
</table>
```

5. 결 론

웹문서에서 HTML 테이블은 연관된 정보를 2차원 논리적 구조로 다양하게 표현할 수 있기 때문에 중요한 정보를 제공하는 수단으로 많이 사용되고 있다. 이러한 이유 때문에 HTML 테이블을 자동으로 해석해서 필요한 정보를 추출하는 방법에 대한 관심이 증가하고 있다. 여러 연구]에서 HTML 테이블의 데이터 자동추출 방법을 제안하고 의미있는 결과를 제시했으나 복잡도가 높고 특정 도메인에 한정된다는 제약을 포함한다. 본 연구는 HTML 테이블의 데이터 자동추출 문제의 해결방안으로 웹문서를 작성할 때 HTML 테이블의 코딩규약을 준수하는 방식을 제안한다. 이를 위해 HTML5 테이블 구조와 table 태그를 분석하여 HTML 테이블의 코딩규약을 제안한다. 통상적으로 사용되는 HTML 코딩규약의 목적은 웹문서 작성시 개발자들 간에 HTML 코드를 가독성을 높이고 쉽게 이해할 수 있도록 하는 것이다. 본 연구에서는 웹문서 개발 측면뿐 아니라 HTML 테이블에서 데이터를 자동추출하기 용이하도록 고려하여 HTML의 테이블 관련 태그들의 코딩규약을 제안하였다. 제안된 방식은 테이블에서 데이터 자동추출 프로그램 설계가 기존 방식에 비해 용이하며 복잡한 테이블 구조의 경우에도 테이블 구조 해석의 난도가 낮아질 것으로 기대된다.

Acknowledgement

※ 이 논문은 (2019~2020)학년도 청주대학교 산업과학연구소가 지원한 학술연구조성비(특별연구과제)에 의해 연구되었음.

References

1. 이은미 “HTML5가 웹 환경에 미치는 영향”, 한국정보과학회지, 제29권 제6호, pp.55-60, 2011.
2. 김연석, 이경호 “HTML 문서의 테이블 식별을 위한 효율적인 알고리즘”, 멀티미디어논문지, 제7권 제10호, pp.1339-1353, 2006.
3. J. Hu et al., “Table Structure Recognition and Its Evaluation”, Proc. SPIE Document recognition and Retrieval VIII, Issue 1, pp.44-55, 2001.
4. 김연석, 이경호 “HTML 문서의 논리적 구조 분석을 위한 효율적인 방법”, 멀티미디어논문지, 제9권 제9호, pp.1231-1246, 2006.
5. 이민형, 이경호 “웹 문서로부터 논리적 구조 추출”, 멀티미디어논문지, 제7권 제10호, pp.1354-1369, 2004.
6. Ahmed Ktob et al, “Extracting Linked Data from HTML Tables”, Proceedings of 2017 3rd Int. Conf. on Collaboration and Internet Computing (CIC), pp.48-53, 2017.
7. 장익, 최성욱 “인터프리터 언어 코딩을 위한 효과적인 학습 모형에 대한 연구”, 융복합지식학회논문지, 제7권 제1호, pp.73-78, 2019.