웹상의 표에서 머리와 몸체 분리 방안 연구

정성원, 권혁철
부산대학교 컴퓨터공학과 한국어정보처리 연구실
{swjung, hckwon}@pusan.ac.kr

Separating Head from Body in Web-Tables

Sung-Won Jung., Hyuk-Chul Kwon
Korean Language Processing Lab, School of Electrical & Computer Engineering,
Pusan National University

요 약

본 논문은, 웹상의 표에서 유용한 정보를 뽑기 위하여 표 머릿부분과 몸체부분을 효과적으로 분리하는 방안을 제안한다. 웹상의 표로부터 정보를 뽑기 위해서는 웹상의 표를 기계가 해석할 수 있는 형태, 즉 속성-값의 쌍으로 변환해야 한다. 이중 속성은 보통 표 머리에 해당하며, 그에 해당하는 값은 표 몸체에 해당하는데, 이는 머리가 해당 몸체 부분을 대표하여 나타내는 단어이기 때문이다. 본 연구의 선행 연구에서는 인터넷상의 표가 표 본래의 정보 전달을 위한 목적 이외에 인터넷 문서의 정렬이나 구조화를 목적으로 쓰이는 경우가 많으므로 이러한 표를 제거하고 표 본래의 의미를 전달하는 표(의미 있는 표)만 추출하는 연구를 하였다. 본 연구에서는 이를 바탕으로 의미 있는 표에서 표 머리와 몸체를 분리하기 위한 휴리스틱에 기반을 둔 모델을 제안한다. 이를 위하여, 표의 본래 특성과, 표를 작성하는 저자의 작성 습관을 관찰하여 머리와 몸체를 분리하기 위한 방안을 설정하고, 이 방안들을 결합하는 모델을 구축한다. 본 연구에서는 이 결과로 80.3%의 표 머리 추출 정확도를 얻을 수 있었다.

1. 서론

정보 추출에서 다루는 문서의 형태는 일반 문서, 구조화된 문서, 반구조화된 문서로 나뉠 수 있다. 이중 일반 문서는 자연언어 문장으로 구성되었으며 가장 쉽게 접할 수 있다. 일반 문서에 정보를 자유롭게 추출하기 위해서는 자연 언어 처리 기법을 이용하여 문장에 대한 이해가 필요하지만, 이는 현재의 기술로는 힘들므로, 특정 패턴이나, 단서를 이용하여 특정 영역에 부합하는 정보를 추출하는 수준에 머물러 있다. 이와 반대로, 구조화된 문서는 저자가 특수한 목적에 맞도록 문서를 정형화시켜놓은 것으로 데이터베이스와 같은 것이 이에 해당한다. 이는 컴퓨터가 이해하기 쉽고, 정보를 추출하기가 용이하지만, 특정 목적 이외에는 다른 용도로 사용하기 어렵다는 단점이 있다. 반 구조화된 문서는 앞의 두 가지 문서 형태의 중간적인 특성을 보이며, 대표적인 예로 표가 있다. 반 구조적인 문서는 보통 문법적으로 완전한 문장으로 표현되지 않으며, 구나 단어 더불어 반 구조문서 자체가 가지는 구조적인 특성을 이용하여 정보를 표현한다. 반 구조화된 문서는 보통 일반 문서보다 분석이 쉬우며, 자체 구조 정보로 인하여 보다 조밀하고 유용한 정보를 가지고 있다. 본 연구는 이러한 반 구조문서 중 가장 대표적인 예인 표를 다룬다. 표는 HTML 문서에서 쉽게 접할 수 있으며, HTML 태그를 이용하여 쉽게 추출할 수 있다.

표는 구조적인 측면에서 봤을 때, 행과 열로 이루어져 있으며, 의미적인 측면에서 봤을 때, 표 머리와 표 몸체로 나뉠 수 있다. 표 머리는 표 머리에 소속된 표 몸체 중의 데이터를 범주화시킨다. 즉, 행과 열은 범주화의 단위가 되며, 행과 열의 가장 윗 셀이나 왼쪽 열은 범주화된 단위의 대표가 된다. 이러한 구조적인 특성 때문에 표는 본질적으로 데이터 집합 내의 의미 연관관계를 제공할 수 있다. 따라서 많은 웹페이지들(증권, 날씨, 통계청)이 데이터 베이스 내의 유용한 정보를 표현하기 위하여 표 형태를 많이 사용한다.

이러한 표의 정보를 활용하기 위하여, 표를 기계가 읽을 수 있는 형태로 바꾸는 것이 필요하다. 즉, 반 구조적인 형태의 문서를 구조적인 형태로 변환해야 한다. 이를 위하여 표 머리와 표 몸체를 속성-값 형태로 변환할 필요가 있다. 속성-값 형태로 변환하면, 이를 바탕으로 구조화된 연관관계를 재구성할 수 있으며, 이는 우리의 연구의 궁극적인 목적이다.

이를 위하여 먼저 (1) 웹상의 HTML 문서에서 의미 있는 표만 추출하는 전처리 단계가 필요하며 (2) 추출된 의미 있는 표에서 정보를 추출한다. 전처리 단계는 본 연구의 선행 연구로 이루어졌다. HTML의 특성상 HTML 상의 표는 표 본래 목적인 대량의 정보를 구조화를 통한 효과적인 전달 이외에 HTML 문서 자체의 구조를 설정하는 목적으로도 쓰인다. 따라서 HTML 문서 자체의 구조를 설정하기 위해 사용하는 표를 제거하고, 정보 추출의 대상이 되는 의미 있는 표만 남긴다. 이를 위하여 두 가지 표를 구별하기 위한 구별 특성을 설정했으며, 기계학습 기법을 사용하여 구별을 위한 모델을 만들었다. 본 연구는 선행 연구에서 추출된 의미 있는 표를 이용하여 표 머리와 몸체를 분리하는 연구를 수행하였다. 이 연구는 의미 있는 표에서 정보를 추출하기 위한 중요한 전단계로 표에서 속성-값 추출을 위한 과정이다. 이를 위하여 선행 연구에서 추출한 구별 특성을 바탕으로 셀과 셀 사이의 관계를 행과 열을 기준으로 반영하기 위한 휴리스틱을 설정한다. 본 연구의 결과는 속성-값 쌍을 추출하는 후보가 될 것이며, 이는 단어 간 계층적 연관관계를 재구성하기 위한 중요한 단서가 된다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 살펴보고, 3장에서 본 연구의 선행연구를 간단히 살펴 본다. 4장에서는 본 논문의 주제인 표에서 머리와 몸체를 분리하는 방안을 살펴본다. 이를 바탕으로 5장에서 실험 결과를 알아보고 마지막으로 결론과 향후 연구 계획을 간략하게 서술한다.

2. 관련 연구

표 정보추출은 정보추출의 하부 영역으로 1990년대 후반부터 시작된 새로운 분야로 초기 선행 연구자들이 이 분야를 "테이블 마이닝"으로 이름 붙였으며, 많은 다른 선행연구자들이 이 용어를 사용하고 있다. 테이블 마이닝은 초기 특정 영역에 한정적인 연구로 시작하여 서서히 특정 영역에 독립적인 연구로 진행되고 있다.

특정 영역에 한정적인 연구는 특정 영역에 맞는 추출 규칙을 사용하여 특정 정보만 추출하는 래퍼(wrapper)의 한 부분으로 연구되기 시작하였다. 이러한 연구 중에서 [1, 2, 4] 는 특정한 표 형태에서만 정보를 추출하였으므로, 다양한 형태의 표를 가지는 웹 문서에 적용하는데 한계를 가졌다.

특정 영역에 독립적인 연구는 Wang[5]의 연구가 대표적이다. Wang은 기계 학습 기법을 사용하여 의미 있는 표와 꾸미기 위한 표의 구분을 시도하였다. Wang은 여러 가지 통계적 속성을 표에서 뽑아냈으며, 전통적인 정보검색에서 쓰이는 기법인 단어 출현 빈도(term frequence)와 표 출현빈도(table frequence, 정보 검색에서 document frequence에 해당)를 사용하였다. 하지만, 이러한 기법은 새로운 표에서 새로운 단어가 출현했을 때 반영할 수 없는 한계가 있다. Yang[7]은 테이블 마이닝의 궁극적인 목적이 표에서 추출되는 속성-값의 쌍으로 표의 모체가 되는 데이터베이스를 재구축하는 것이라고 보았다. Yang은 속성-값 쌍을 추출하기 위하여 언어에서 나타나는 패턴으로 추출 규칙을 설정하였다. 따라서, 언어의 패턴에 편향되는 특성 때문에 언어가 바뀌거나 고려하지 않은 패턴이 나타날 경우 적용하기 어려운 문제점이 있다.

이러한 관련 연구에서의 문제점은 영역 독립적인 연구라 하더라도 새로운 영역이나 새로운 표가 출현했을 때는 유연하게 대처하지 못한다. 또한, 관련 연구들 마다 의미 있는 표와 꾸미기 위한 표, 또는, 진짜 표(genuine table)와 거짓 표(non-genuine table), 데이터를 가진 표(data table)와 데이터를 가지지 않은 표(non-data table) 등 표에 대한 정확한 정의가 없다. 따라서 본 논문에서는 이러한 문제점을 해결하기 위하여, 의미 있는 표와 꾸미기 위한 표를 정의하고, 표의 편집시 나타나는 속성과 구조적인 정보만을 고려하여 영역 독립적인 표 정보 추출 방안을 제안한다.

3. 의미 있는 표와 꾸미기 위한 표의 구분

인터넷상의 표는 일반적인 문서 편집에서 대량의 정보를 효과적으로 구조화하여 전달하는 목적 이외에 문서의 구조를 설정하고, 문서 내 요소를 효과적으로 배치할 목적으로 사용하는 경우가 많다. 본 장에서는 이러한 표의 유형을 정의하며, 이를 구별하기 위한 본 연구의 선행 연구[3]를 간략하게 소개한다.

3.1. 표의 정의

웹상의 표는 HTML의 여러 요소 중의 하나로 HTML 문서 중 태그(tag)로 표현된다. HTML의 태그는 크게 내용을 표현하는 태그와 구조를 표현하는 태그로 나눌 수 있으며, 표는 구조를 표현하는 태그이다. HTML 문서에서는 표를 구성하는 태그는 <table>, <tr>, <td>, <th> 등이 있다. 공백문자나 <p>,
 태그 등을 혼합하여 표와 유사한 형태로 만들 수 있지만, 이는 본 논문에서 다루지 않기로 한다.

정의 1 : 웹상의 표는 HTML 소스 코드 상에서 <table>태그로 시작하여 </table>로 끝나는 영역이다.

또한, 표는 일반적으로 그림 1과 같이 네 개의 부분으로 구성되어 있으며, HTML 태그로 각각은 <caption>(표 제목), <thead>(머리), <tbody>(몸체), <tfoot>(다리)에 대응된다.

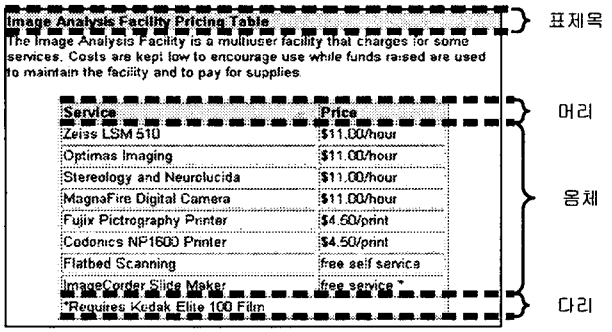

그림 1. 표의 구조

표는 HTML의 구조를 표현하는 태그이므로, 표 본래 목적인 대량의 데이터를 구조화하여 전달하는 것 이외에 HTML 문서 자체의 구조를 설정하고 내용을 정렬하는 등의 목적으로 쓰이기도 한다. 따라서 웹상의 표는 다음과 같은 두 가지 종류가 있으며 그림 2와 같이 하나의 웹페이지에 같이 출현할 수도 있다.

정의 2 :

(1) 의미 있는 표 : 대량의 정보를 효과적으로 전달하는 표 본래 목적을 위해 사용되는 표로 표의 내용은 표 구조와 밀접한 연관이 있다.

(2) 꾸미기 위한 표 : HTML 문서를 구조화할 목적으로 사용되는 표로 표의 내용은 표의 구조와 연관이 없다.

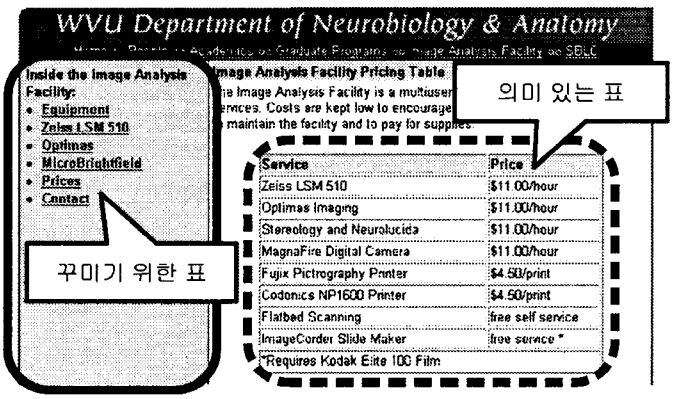

그림 2. 웹상의 표의 종류

위의 정의에서 알 수 있듯이, 의미 있는 표는 표의 구조를 흩뜨리면 표의 내용이 어떤 의미가 있는지 알 수 없다. 예를 들어 그림 2의 의미 있는 표에서 'Price'가 표의 머리이며 그에 해당하는 열이 모두 'Price'의 값임을 알 수 있다. 하지만, 표 머리인 'Price'를 삭제한다면 그에 해당하는 데이터는 어떤 의미인지 알 수 없다. 이와 같은 예에서 알 수 있듯이 의미 있는 표에서 대량의 데이터를 구조화하기 위해서는 표 머리를 이용하여 데이터를 그룹화하여 추상화한다. 따라서, 의미 있는 표와 꾸미기 위한 표의 구별은 다음 같다.

정의 3 : 웹상의 표에서 표 머리가 있으면 의미 있는 표이고, 머리가 없으면 꾸미기 위한 표이다.

3.2. 의미 있는 표와 꾸미기 위한 표의 구분

앞 장에서 살펴 본대로 표는 크게 네 개의 부분으로 구성되어 있으며 각각을 표현하기 위한 태그도 존재한다. 이러한 태그를 사용자들이 표의 구조에 맞게 잘 사용한다면, 웹상의 표에서 정보를 뽑아내는데 한결 수월하겠지만, 실제 HTML 문서에서는 앞에서 서술한 태그들이 거의 사용되지 않으며, <tr>과 <td>만 사용하여 표를 구성한다. 본 연구의 선행 연구에서는 이러한 점을 극복하고자, 의미 있는 표와 꾸미기 위한 표를 구분을 먼저 시도하였다.

웹상의 표를 작성하는 사람은 표가 쓰이는 목적에 따라 표를 편집하는 방법이 다르다. 이를 관찰하여, 우리는 표의 "외형적 구별 특성"과 표 내부 내용의 연속성을 바탕으로 한 "연속적 구별 특성"을 설정하였다.

외형적인 구별 특성은 표가 쓰이는 목적에 따른 표의 외형을 살핀다. 예를 들어, 꾸미기 위한 표는 내용의 정렬이 목적이므로 1개의 셀로 구성되어 있거나 1차원이며, 표 안에 또 다른 표를 가지고 있는 경우도 많다. 반면, 의미 있는 표는 2차원이며, 표 안의 표를 가지고 있는 경우가 별로 없다.

연속적인 구별 특성은 표의 내용의 분포가 얼마나 유사하며 반복적인가를 판단한다. 일반적으로 의미 있는 표는 그 내용이 반복적이고 유사하다. 이는 표의 머리와 그에 연관된 데이터의 관계 때문에 발생하는 현상이다. 즉, 표의 머리는 데이터를 대표하는 것이므로, 그에 연관된 데이터는 연속적인 특성이 있는 것이 당연하다. 이는 인접한 셀의 내용의 유형(그림, 링크, 숫자, 문자 등)이나, 인접한 셀 내부의 문자열의 길이 등을 이용하여 통계에서 흔히 사용하는 표준편차 공식에 적용 계산한다.

이와 같은 기준으로 본 연구의 선행 연구에서는 23개의 구분 특성을 설정하였으며, 이를 자동 추출할 수 있는 프로그램을 만들었다. 여기에 수작업으로 정답을 붙여 학습 데이터를 만들고, 이를 C4.5 알고리즘에 적용하여 구분 모델을 만들었다. 구분 특성에 대한 자세한 사항은 본 논문의 주제에서 벗어나므로 생략하였으며, 자세한 사항은 선행연구[3]에서 참조할 수 있다.

4. 표 머리와 몸체의 구분

앞 장까지 인터넷상에서 꾸미기 위한 표를 제거하고, 의미 있는 표를 추출하는 방법을 살펴 보았다. 의미 있는 표가 인터넷 문서에서 분리되면, 다음 단계로 의미 있는 표에서 머리와 몸체 부분을 분리해야 한다. 이는 속성-값 쌍을 추출하기 위한 중요한 단계이다. 선행 연구에서는 표 머리와 몸체를 구분하기 위한 간단한 휴리스틱을 소개하였다. 본 연구에서는 선행 연구를 좀 더 정밀하게 보완하고 각각의 휴리스틱을 효과적으로 결합하는 방안을 소개한다.

4.1. 표 머리와 몸체를 구분하기 위한 휴리스틱 설정

표의 머리와 몸체를 구분하기 위해서는 의미 있는 표와 꾸미기 위한 표를 구분하는 특성을 더 세밀하게 적용하여야 한다. 즉, 기존 연구에서 의미 있는 표와 꾸미기 위한 표의 구분 특성이 표 전체의 경향을 반영하는 것이라면, 표 머리와 몸체를 구분하는 특성은 셀과 셀 사이의 관계를 행과 열을 기준으로 반영하는 것이라 할 수 있다. 이를 위하여 크게 다음과 같은 두 가지 경향을 고려할 수 있다.

먼저, 의미 있는 표를 작성할 때, 저자는 머리와 몸체를 분리하여 독자에게 자신의 의도를 더욱 정확하게 전달하기 위하여 다양한 기법을 사용한다. 즉, 표 머리 부분과 몸체를 다른 배경 색깔, 혹은 다른 글자체, 글자 크기 등으로 지정한다. 두번째로, 표 머리는 머리와 연관된 데이터를 대표하므로, 표 머리와 연관된 데이터는 비슷한 형태로 나타나지만, 표 머리는 데이터와 다른 형태로 나타나는 경우가 많다. 이와 같은 경향을 고려하여, 다음과 같은 관찰을 통하여 휴리스틱으로 설정할 수 있다.

관찰 1. <th> 태그는 일반적으로 잘 쓰이지 않으며, 대신 <td> 태그가 많이 쓰인다. 따라서, 어떤 표 내에서 <th> 태그를 사용했다면, 저자가 의도적으로 표 머리를 표시하기 위하여 사용했다고 볼 수 있다.

휴리스틱 1. 만약 한 셀이 <th> tag로 표현되면, 그 셀은 머리에 속할 가능성이 크다.

관찰 2. 저자가 웹상의 표를 편집할 때, 표 머리와 몸체를 분리하기 위하여 여러 가지 꾸밈을 사용한다. 즉, 배경색을 달리한다든지, 표 머리에 나타나는 단어의 글씨체나 크기, 속성을 몸체에 나타나는 단어와 다르게 한다. 여기에 사용되는 HTML 속성이나 태그로는 'bgcolor' 속성, <b> , <i>, <u>등의 글씨 꾸밈 태그, <font> 태그에 속한 'face', 'color', 'size' 등의 글씨체 속성이 있다. 이 8가지 속성들을 '표 머리 꾸밈 속성'이라 부르며, 이러한 속성은 표에 2가지 종류가 나타나며, 각각이 표에서 두 개의 영역으로 분리될 때, 표 머리와 몸체를 분리하는데 가장 좋은 단서가 된다. 이 중 'bgcolor'는 표 머리와 몸체를 구분하는 것 이외에도 단순히 행과 열을 쉽게 구분하는 목적으로도 쓰일 수 있다. 또한, 글씨관련 속성은 보통 저자가 글씨 관련 속성 중 한 번에 한두 개의 속성만 사용하는 경향이 있으므로 묶어서 생각하는 것이 편하다. 따라서 'bgcolor'속성과 글씨관련 속성은 나눠서 각각의 휴리스틱으로 구성한다.

휴리스틱 2. 만약 표 머리 꾸밈 속성중 'bgcolor'를 이용하여 표가 두 개의 영역으로 나누어진다면, 가장 위나 가장 왼쪽에 있는 영역이 표 머리일 가능성이 크다.

휴리스틱 3. 만약 표 머리 꾸밈 속성중 글씨관련 속성을 이용하여 표가 두 개의 영역으로 나누어진다면, 장 위나 가장 왼쪽에 있는 영역이 표 머리일 가능성이 크다.

관찰 3. 의미 있는 표는 표 머리글과 연관되어 표 몸체 부분의 셀 내용의 타입이 반복되어 나타나는 경향이 있다. 예를 들어 어떤 표의 머리글에 '홈페이지'가 있다면, 그와 연관된 몸체의 셀에는 <a href> 태그가 계속해서 나타날 것이다. 이러한 셀 내용 타입에는 링크, 그림, 숫자, 문자 등이 있으며, 한 셀에 이러한 내용 타입이 섞여서 나타날 수도 있다.

휴리스틱 4. 만약 어떤 행이나 열의 가장 왼쪽 셀이나 가장 윗셀을 제외한 그 행이나 열에 속한 다른 셀에서 셀 내용 타입이 반복적으로 나타나면, 그 행이나 열의 가장 왼쪽 셀이나 윗셀을 포함한 가장 왼쪽 열이나 가장 윗행은 표 머리일 가능성이 크다.

관찰 4. 셀 내부에 문자열은 특별한 단위형태의 연속으로 이루어져 있다. 단위형태는 셀 내부 문자열의 부분으로 특별한 구별자로 구분이 될 수 있다. 본 논문에서는 이 단위 형태를 알파벳 문자, 숫자, 태그, 특수문자의 네 가지로 설정했다. 예를 들어 그림 3을 보면 표 몸체 영역은 '숫자-알파벳 문자'의 형태임을 알 수 있다. 이를 '내용 패턴'이라 부르기로 한다.

Cut in length	Gamma Energy in PbWO4	e- Energy in PbWO4
50 micron	12 keV	136 keV
100 micron	20 keV	210 keV
500 micron	60 keV	642 keV
1 mm	85 keV	1.13 MeV
5 mm	137 keV	5.34 MeV
1 cm	219 keV	11.5 MeV

그림 3. 내용 패턴을 가진 의미 있는 표의 예

휴리스틱 5. 만약 어떤 행이나 열의 가장 왼쪽 셀이나 가장 윗셀을 제외한 그 행이나 열에 속한 다른 셀에서 셀 내용 패턴이 반복적으로 나타나면, 그 행이나 열의 가장 왼쪽 셀이나 윗셀을 포함한 가장 왼쪽 열이나 가장 윗행은 표 머리일 가능성이 크다.

관찰 5. 웹상의 표는 가끔 태그로 표 내부의 셀을 합친다. 이때, 가장 왼쪽 열이나 가장 윗행에 합쳐진 형태의 셀이 나타나면, 바로 오른쪽 열이나 아래의 행에는 합쳐진 셀의 값이 나타나는 경우가 많다(그림4 참조). 즉, 합쳐진 셀의 내용이 그와 연관된 오른쪽이나 아래 인접 셀의 상위 개념이 되며, 이를 통하여 표에서 다중 계층 구조를 표현한다.

Total Receipts	Reseller Pricing Per Domain Year			
	.com/.net .org	.biz/.info .us/.name	.in	.co.in/.net.in/.org.in .gen.in/.firm.in/.ind.in
Low Volumes	$6.99	$7.99	$12.59	$6.89
> US $675	$6.75	$7.49	$12.59	$6.89
> US $1688	$6.75	$7.25	$12.39	$6.69
> US $3245	$6.49	$6.99	$12.15	$6.39
> US $6490	$6.49	$6.75	$12.15	$6.39
> US $12980	$6.49	$6.49	$11.99	$6.19

그림 4. 합친 셀을 가진 의미 있는 표의 예

휴리스틱 6. 표에서 이 가장 왼쪽 열이나 가장 윗행에 나타나면, 그에 인접한 셀을 추상화하는 것이므로, 인접한 셀까지 표 머리에 속할 가능성이 크다.

관찰 6. 일반적인 문서상에서 표의 가장 윗행과 가장 왼쪽 열이 모두 표 머리이면, 표의 1행 1열은 사선으로 나누어지고 각각 행과 열을 대표하는 개념을 서술한다. 하지만, 웹상의 표에서는 셀 내부에 사선을 표현하기가 힘들다. 따라서, 이 경우에 웹상의 표에서는 빈칸으로 비워두는 경우가 많다.

휴리스틱 7. 표의 1행, 1열이 내용이 없는 빈 셀이면, 그 셀을 포함하는 가장 윗행과, 가장 왼쪽 열은 모두 표 머리일 가능성이 크다.

4.2. 후보 행렬 생성

하나의 표는 4.1절에 기술한 휴리스틱에 바탕을 두어 행렬로 변환된다. 변환된 행렬의 각 요소의 값은 표 머리가 될 가능성이 수치로 표현되어 있다. 즉, 행렬의 요소의 값이 '1'이면 표 머리가 될 가능성이 크며, '0'이면 표 몸체가 될 가능성 크다는 뜻이다. 그림 3은 이와 같은 변환 과정을 도식화한 것이다. HTML 상의 표 그림 3의 (a)는 표 머리를 표현하기 위하여 <th> tag (휴리스틱 1), 숫자 타입(휴리스틱 3), 숫자 형태 패턴 (휴리스틱 4), 1행, 1열에 있는 빈 셀(휴리스틱 6) 등 4가지 속성이 있다. 따라서, 그림 5의 (b)와 같이 4개의 후보 행렬로 변환된다.

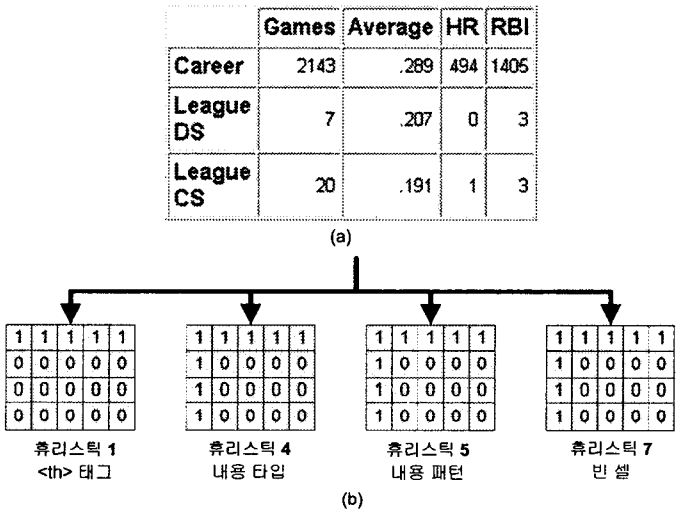

그림 5 표의 후보 행렬로의 변환

4.3. 후보 행렬의 결합을 통한 표 머리 추출

4.2절에서, 표 하나가 여러 개의 후보 행렬로 변환되는 과정을 살펴보았다. 각 후보 행렬의 요소 중 '1'을 값으로 가지는 요소는 표 머리가 위치할 가능성이 있다는 뜻이다. 이와 같은 여러 후보 행렬을 합쳐서 최종 결과 행렬을 만들기 위하여 선형 결합 방법을 사용한다. 이는 다음 수식 (1) 과 같다.

$$S = \sum_{i=1}^{6} \lambda_i H_i \qquad \text{----------} \quad (1)$$

$$\text{where } 0 \le \lambda_i \le 1 \text{ and } \sum_i \lambda_i = 1$$

여기서 S는 행렬의 합이며, λ_i는 각 후보 행렬의 가중치, Hi는 후보 행렬이다.

S는 후보 행렬의 가중치에 따른 합이다. 따라서 S의 각 요소를 평가하여 최종 결과 행렬로 만들어야 한다. S의 각 요소는 다음 수식 (2)로 평가된다.

$$Mid = \frac{\max(S_{ij}) + \min(S_{ij})}{2}$$

$$R_{ij} = \begin{cases} 1 & \text{if } S_{ij} \geq Mid \\ 0 & otherwise \end{cases} \quad \text{---------- (2)}$$

$max(S_{ij})$는 S의 요소 중 가장 큰 값이며, $min(S_{ij})$는 S의 요소 중 가장 작은 값이다. 따라서 Mid는 S의 요소 중 가장 큰 값과 작은 값의 중간 값이 된다. 이 값을 이용하여 S의 요소가 Mid 값보다 크면 1 아니면 0으로 설정하여 행렬 R을 만든다. 따라서, R의 요소 중 1인 요소는 표 머리에 속하며, 0인 요소는 표 몸체에 속한다. 다음 그림 6은 이와 같은 과정을 도식화한 것이다.

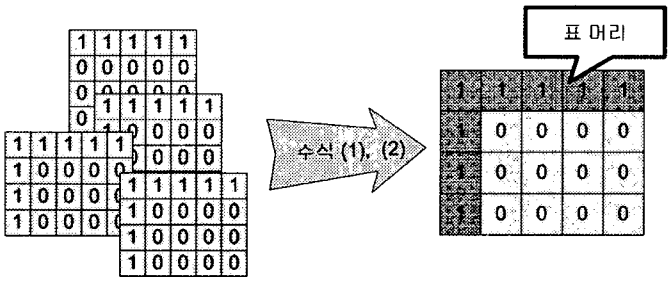

그림 6. 표 머리 선택 과정

5. 실험

5.1. 실험 데이터

실험을 위하여 본 연구실에서 보유하고 있는 웹 로봇을 이용하여 HTML 문서를 수집하였고, 그 중 의미 있는 표를 가진 인터넷 문서를 선택하였다. 또한, Wang의 실험 데이터도 사용하였다. 실험 데이터의 특성은 다음 표와 같다.

항목	자체 수집 데이터	Wang 데이터	합계
의미 있는 표 수	964	969	1,933
꾸미기 위한 표 수	2,249	2,009	4,258
합 계	3,213	2,978	6,191

표 1 데이터 특성

실험데이터는 Wang의 제안에 따라 구성하였다. Wang은 각 표에 해당하는 <table> 태그에 HTML의 표준에 없는 속성을 추가하였다. 'tabid'는 실험에 쓰이는 표의 고유번호로써 숫자를 속성값으로 가진다. 'genuinetable'은 'yes'와 'no'의 속성값을 가지며 각각 의미 있는 표와 꾸미기 위한 표를 의미한다. 본 연구에서는 표의 머리와 몸체를 분리하는 실험을 수행해야 하므로 여기에 추가의 속성인 'head'를 설정하였다. 'head'는 각 셀에 설정되어야 하므로, <td>나 <th>의 속성이며, 속성값은 'yes'와 'no'로 표 머리에 각 셀이 속하는지를 표시한다.

5.2. 각 휴리스틱의 적절성

4.1절에서 설정한 휴리스틱이 얼마나 효과적인지 알아보기 위하여 각 휴리스틱을 하나씩 적용했을 때의 성능을 조사해 보았으며, 표 2가 그 결과이다.

휴리스틱	적용 정도		셀 기준 성능	
	표의 개수	적용비율	정확도	재현율
1.	246	12.7%	75.4%	82.0%
2.	469	24.3%	65.0%	94.9%
3.	430	22.2%	76.4%	98.7%
4.	1471	76.1%	74.3%	89.6%
5.	1191	61.6%	67.8%	88.4%
6.	375	19.3%	48.5%	76.9%
7.	267	13.8%	57.1%	91.6%

표 2 휴리스틱의 적절성

표2의 결과로 보듯이 4번, 내용 타입과 5번 내용 패턴이 가장 적용 범위가 넓었다. 반대로 휴리스틱 1, <th> 태그의 출현 여부는 적용 범위가 좁으며, 이는 <th> tag를 사용하여 작성된 표가 별로 없다는 것을 말한다. 셀 기반 성능은 적용된 범위 내에서의 셀을 기준으로 한 정확도와 재현율을 나타내고 있으며, 적용된 범위 내에서는 어느 정도 유용성이 있다고 판단된다.

5.3. 표 머리와 몸체 분리 정확도

휴리스틱 적용의 정확도를 높이기 위하여 4.3의 수식 (1), (2)를 이용하여 각 휴리스틱을 최적의 방법으로 결합한다. 먼저, 휴리스틱 4를 최하 성능으로 설정한다. 휴리스틱 4는 적용 정도가 가장 높고 셀 기준 성능도 높으므로, 전체 휴리스틱 중 가장 높은 성능을 나타낸다. 다음으로, 휴리스틱에 모두 가중치 1을 준 단순 결합방법으로 결합해 보았다. 마지막으로 선형 결합시 가중치를 힐 클라이밍 방법으로 찾아내어 적용하여 보았다. 표 3은 그 결과이며, 표 4는 힐 클라이밍 방법으로 얻은 최적 가중치 값이다.

추출 모델	셀 기준 성능		표 기준 성능	
	정확도	재현율	맞는 표 수/전체 표수	정확도
최하 성능	74.3%	64.5%	1176/1933	60.8%
가중치 동일 결합	71.1%	81.3%	1409/1933	72.9%
힐 클라이밍 방법으로 가중치 조절 후 결합	78.5%	86.4%	1552/1933	80.3%

표 3 표 머리 추출 실험 결과

휴리스틱	1	2	3	4	5	6	7
가중치 (λ_i)	0.125	0.188	0.125	0.125	0.0625	0.1875	0.187

표 4 최적 가중치

6. 결론 및 향후 연구

본 논문은 웹상의 표에서 머리와 몸체를 분리하는 방안을 연구하였다. 몇몇 선행 연구들이 의미 있는 표와 꾸미기 위한 방안을 연구하였으나, 머리와 몸체를 분리하는 방안에 대한 연구는 미미하며, 대부분 언어적인 패턴을 사용하므로, 언어에 대해 편향성을 가지는 경우가 많다. 본 연구는 모든 인터넷 문서에 적용 가능하도록 언어적인 요소를 최대한 억제하였으며, 표 자체의 특성과 표를 작성하는 저자의 습관, 표 내부의 내용의 반복성 등을 고려하였다. 이를 기반으로 7개의 휴리스틱을 이용한 표 머리와 몸체 분리 규칙을 설정하였으며, 이를 효과적으로 결합하는 방안을 제시하였고, 최종 결과로 80.3%의 추출 정확도를 달성하였다.

본 연구의 최종 목적은 반 구조적인 언어자원인 표에서 표의 구조적인 정보를 이용하여 계층적인 정보를 설정하는 것이다. 향후, 본 연구에서 추출된 표 머리 몸체 정보를 활용하고, 표의 구조적인 정보, WordNet 등 기구축된 언어 자원, 단어 간 연어 정보, 클러스터링 등의 기법을 활용하여 더욱 정확한 속성-값 관계를 추출하며, 이를 정보 검색 시스템, 온톨로지 반자동 구축 등 다른 응용 분야에 활용하는 연구를 진행할 것이다.

6. 참고 문헌

[1]Chen, H.H., Tsai, S.C., Tsai, J.H.:Mining Tables from Large Scale HTML Texts. Proceedings of 18th International Conference on Computational Linguistics, Saabrucken, Germany, July 2000.

[2]Hurst, M.: Layout and Language: Beyond Simple Text for Information Interaction - Modeling the Table. Proceedings of the 2nd International Conference on Multimodal Interfaces, Hong Kong, 1999.

[3]정성원,박대원,권혁철 : 기계학습과 규칙 기반 접근 방법을 결합한 의미 있는 표 구분과 헤드 영역 추출, 제16회 한글·언어·인지 학술대회 제16권 제 1호, p.5~11,2004.10.

[4]Ning, G., Guowen, W., Xiaoyuan, W., Baile, S.: Extracting web table information in cooperative learning activities based on abstract semantic model. Computer Supported Cooperative Work in Design, The Sixth International Conference, 2001, 492-497.

[5]Wang, Y., Hu, J.: A Machine Learning Based Approach for Table Detection on The Web in Proceedings of The Eleventh International World Wide Web Conference WWW2002, Sheraton Wailili Honolulu, Hawaii, USA, 2002, 7-11.

[6]Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Pub., 2000.

[7]Yang, Y.: Web Table Mining and Database Discovery. M.Sc. thesis, Simon Fraser University, August, 2002.