기계학습과 규칙 기반 접근 방법을 결합한 의미 있는 표 구분과 헤드 영역 추출[1)]

정성원, 박대원, 권혁철
부산대학교 컴퓨터공학과
{swjung, bluepepe, hckwon}@pusan.ac.kr

Extracting Web-Table Information Using Decision Tree and Rule Based Approach

Sung-Won Jung, Dae-Won Park, Hyuk-Chul Kwon
Dept. of Computer Science, Pusan National University

요 약

일반적으로 HTML문서는 크게 내용과 구조로 이루어져 있다. HTML은 일반 문서와 달리 태그라는 것으로 문서에 추가 정보를 주며, 문서의 내용을 더욱 명확하게 한다. 따라서 태그를 이용하면 일반 문서보다 정보를 쉽게 구별 하고 추출할 수 있다. 이러한 여러 가지 태그들 중에서 본 연구는 표를 중점적으로 연구한다. 표는 행과 열을 이용하여 어떤 사실을 조직하여 전달하는 것으로, 다른 구조적 특성들 보다 정보를 조직하는데 매우 유용하며, 글로 기술할 많은 분량을 간단히 줄이는 역할을 한다. 이와 같은 표의 특성에 주목하여 표에서 정보를 추출하는 분야를 기존 연구자들은 Web Table Mining 명명하였다. 본 연구는 기존 연구자들이 간과한 표의 구조적인 특성을 이용하여 전체 인터넷 문서에 적용할 수 있는 방법과 함께, 표에서 의미 있는 정보 추출을 위한 단계적인 모형을 제시한다.

1. 서 론

HTML문서는 일반 문서와 달리 태그가 있으며 HTML문서에 추가적인 정보를 부여한다. 이 태그는 크게 두 가지 역할을 한다. 즉, 문서에 추가의 정보를 제공하는 〈image〉, 〈font〉와 같은 태그가 있으며, 문서의 구조를 만들 수 있는 〈p〉, 〈title〉과 같은 태그가 있다. 이와 같은 태그들은 문서의 내용을 보다 명확히 하는데 도움을 준다. 본 논문은 이러한 여러 가지 태그 중에서 표에 초점을 둔다. 다른 태그와 달리 표는 다음과 같은 이점을 가진다.

첫째, HTML문서 중에서 표 부분을 분리하기 용이하다. 일반 텍스트 문서와 달리 HTML에서는 표를 나타내기 위한 태그가 따로 존재한다. 따라서 다른 HTML 문서 영역과 표를 구별할 수 있는 기준이 명확하며, 이는 HTML에서 표에 대한 의미 추출을 시도하는 출발점이기도 하다.

둘째, 표는 그 자체가 의미 정보를 구조적으로 표현하는 수단이다. 표는 행과 열로 이루어져 있으며, 행과 열의 조합을 통해 어떤 정보를 가지고 있는 내용을 가리키게 된다. 이런 구조적인 특성으로 인하여, 많은 저자들이 어떤 정보를 구조화하여 전달할 목적으로 표를 흔히 쓰고 있다.

셋째, 표는 문서의 입장에서 보았을 때, 그 문서가 가치 있는 문서인지 아닌지에 대한 판단 기준이 될 수 있다. 일반적으로 표를 포함하고 있는 문서는 의미 있는 정보를 담고 있다. 이는 저자가 자신이 작성하는 문서를 표로 표현하고자 할 때는 고도의 지적 판단을 행하기 때문이다. 일반적으로 정보 검색 시스템이 검색하는 대상은 문서이므로, 유용한 정보를 포함한 페이지를 구별하는 것은 정보 검색에서 중요한 요소이다.

이러한 여러 가지 장점에도 불구하고 웹에서 표 추출은 다음과 같은 어려운 점이 존재한다.

첫째, HTML의 특성상 의미 있는 표와 의미 없는 테이블의 분리가 쉽지 않다. HTML은 구조와 표현이 분리되어 있지 않다. 이런 HTML의 특성은 표에도 고스란히 반영되어 있는데, 표 본래의 목적인 의미의 구조

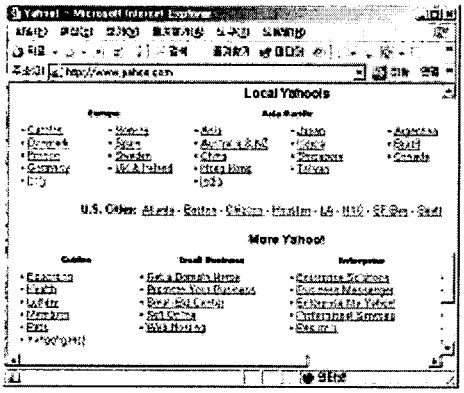
1) 이 논문은 과학 기술부(한국과학기술기획평가원)의 국가지정연구실 사업지원으로 이루어진 것임(Contract Number : M1-0412-00-0028-04-J00-00-014-00).

화와 효과적인 전달보다는 문서의 내용을 가지런히 하여, 화면을 정리하는 목적으로 쓰이는 경우가 많다. 다음 표 1은 인터넷상의 표에 대해서 수작업으로 통계를 내어 본 결과이다. 표 1에서 보는 바와 같이 전체 인터넷 문서 중 대다수(77.78%)가 표를 포함하고 있지만 거의 대부분의 표가 꾸미기 위한 목적으로 쓰이고 있다는 것을 알 수 있다.

항 목	개수
전체 문서 (A)	86475(개)
〈table〉 태그 포함 문서 (B)	67259(개)
의미 있는 표 포함 문서 (C)	1009(개)
문서 당 표의 평균 개수	15.13(개)
전체 문서 중 〈table〉 태그를 포함한 문서 비율 (B/A)	77.78(%)
전체 문서 중 의미 있는 표를 포함한 문서 비율 (C/A)	1.167(%)
〈table〉 태그를 포함한 문서 중 의미 있는 표를 포함한 문서 비율 (C/B)	1.500(%)

[표 1] 웹 문서 중 표에 대한 통계

둘째, 의미 있는 표와 의미 없는 표의 구분 기준이 모호하다. 그림 1과 같은 링크 정보를 나열한 표를 보자. 유용한 웹 페이지를 찾기 위해서는 링크 정보를 모은 페이지가 유용하겠지만, 특정한 내용을 찾기 위한 목적에서는 이 페이지는 유용하지 않다. 이렇게 보는 관점에 따라서 표의 유용성이 달라질 수 있으므로, 연구의 대상이 되는 표 중에서 의미 있는 표와 의미 없는 표에 대한 명확한 기준을 수립해야 한다.



[그림 1] 구분이 모호한 형태의 표

셋째, 표는 그 형태가 매우 다양하다. 우리가 생각하는 기본적인 형태의 표는 행 색인어와 열 색인어가 존재하는 형태이며, 이런 형태의 표를 일반적으로 의미 있는 표라고 인식하며, HTML의 내부적인 구성은 고려하지 않는다. 하지만, 실제 HTML 문서에서 표가 작성되어 있는 형태를 보면, 외형적으로 보았을 때는 여러 개의 표이나 구조적으로는 한 개의 표인 경우도 있으며, 반대로 외형적으로는 한 개의 표이지만, 구조적으로는 여러 개의 표로 이루어진 경우도 종종 볼 수 있다. 이 경우는 특히 웹 문서에서는 시각적인 효과를 주기 위하여 테이블을 여러 개 겹쳐서 사용하는 경우에 빈번히 나타난다.

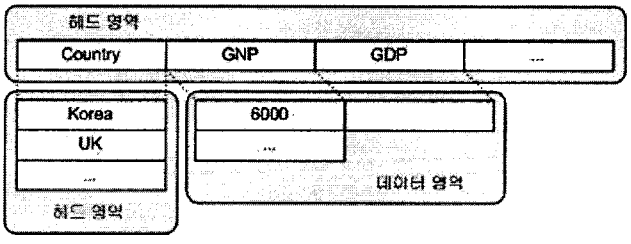
2. 관련 연구

기존 인터넷 상의 표의 정보 추출은 특정 웹 문서 형식에 국한되어 있다. [1,2,4,7]과 우리의 이전 연구[3]는 특정 영역에 한정된 표에서 정보를 추출하며, 추출 규칙을 설정한다. 따라서, 이와 같은 연구들은 일반적인 인터넷 문서에 적용하기에 적합하지 않다. 또 다른 연구[5]는 기계 학습을 사용하여 꾸미기 위한 표와 의미 있는 표를 구분하는데, 본 연구와 유사하지만, 구별 특성 중 일부가 특정 영역에 국한되는 문제점을 가진다. 본 연구는 이와 같은 이전 연구의 문제점을 해결하고 일반적인 인터넷 문서에 적용하기 위하여 특정 영역에 국한된 정보는 배제하고 표의 구조적 특징을 이용하여 의미 있는 표와 꾸미기 위한 표를 구분하고 정보를 추출하는데 그 목적이 있다.

3. 구 현

3.1 HTML상의 표의 특징

우리가 일반적으로 알고 있듯이 표는 헤드 영역과 데이터 영역으로 이루어져 있다. 헤드 부분은 보통 가장 윗행과 가장 왼쪽 열에 나타나며, 나머지는 대부분 데이터 영역이다. 이는 다음 그림 2와 같다.



[그림 2] 표의 구조

하지만, HTML에서 표는 조금 다르다. HTML에서 표를 편집하기 위해서는 〈table〉, 〈tr〉, 〈td〉 태그 등을 주로 사용한다. HTML문서에서 이 태그들은 표의 본래 목적뿐만 아니라 HTML문서에 내용들을 배치하기 위해서도 자주 사용된다. 이 두 가지 표 중에서 우리가 관심을 가지는 대상, 즉 정보를 뽑아낼 수 있는 대상은 표의 본래 목적으로 쓰인 표이다. 따라서 이 두 가지 표를 구분할 기준을 세워야 한다.

본 논문은 HTML문서 상의 표를 다음 두 가지로 구분한다.

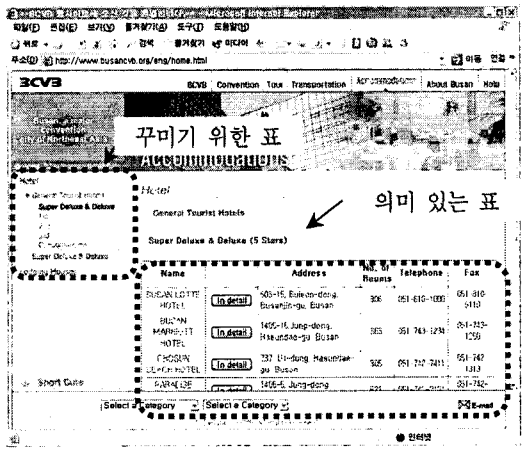
1. 의미 있는 표 : 표의 본래 목적으로 사용된 표
2. 꾸미기 위한 표 : HTML문서의 내용을 배치하기 위한 표

그림 2는 위의 정의에 대한 예제로 동일한 웹 페이지에 두 가지 형태의 표가 공존할 수 있다는 것을 보여준다. 다음은 이 두 가지 표를 구분하는 기준이다.

-표가 헤드 부분을 가지고 있으면 의미 있는 표이며 그렇지 않으면 꾸미기 위한 표이다.

표의 내용도 문자형 내용과 태그형 내용의 두 가지로 나뉠 수 있다. 이 분류는 다음 장에서 제시할 표의 구별 특성을 설정하는데 효과적인 단서가 된다. 이는 다음과 같이 정의된다.

1. 문자형 내용 : 태그를 사용하지 않고 편집된 내용 : 숫자, 문자, 단어, 특수문자 등
2. 태그형 내용 : 태그를 사용하여 편집된 내용 : 그림, 링크, 소리, 표 등 그림 3 의미있는 표와 꾸미기 위한 표의 예



[그림 3] 의미있는 표와 꾸미기 위한 표의 예

3.2 표 구별 특성 설정

앞장에서 제시한 여러 정의를 바탕으로 의미 있는 표와 꾸미기 위한 표를 구별하기 위하여 표 구별 특징을 설정해야 한다. 이를 위하여, 표의 외형적인 특징뿐만 아니라 내용의 통계적인 특징도 고려하여 구별 특징을 설정하였다. '외형적 구별 특징'은 표가 브라우저에 표현될 때의 특성과 HTML 소스에서 나타나는 특성을 이용하여 설정할 수 있는데, 표의 외곽선이나 표 제목, 행과 열의 수, 글자체의 속성 등을 고려할 수 있다. '연속적 구별 특성'은 표 내부의 내용들의 연속성을 통계적인 방법을 이용하여 유추해 내는 것으로 한 셀에 포함된 단어의 표준편차, 행에 포함된 셀의 개수에 대한 표준편차 등을 고려할 수 있다. 이렇게 설정된 구별 특징은 기계 학습 알고리즘에 적용하기 위한 학습 데이터를 만들기 위해 사용된다.

3.3. 외형적 구별 특성

우리는 이미 앞장에서 의미 있는 표의 가장 큰 특징은 헤드 부분이 있다는 것을 언급했다. 일반적으로 HTML에서 의미 있는 표를 작성할 때 헤드 부분을 다른 표의 영역과 구별하고 강조하기 위하여 〈td〉 태그에 'bgcolor'속성을 사용하여 다른 영역과 배경색을 다르게 표현하기도 하며, 〈font〉 태그를 사용하여 글자 색깔이나 크기, 글씨체를 다르게 하기도하고, 〈th〉 태그를 사용하여 명시적으로 헤드 부분임을 나타내기도 한다. 따라서 이러한 태그들은 헤드 영역과 데이터 영역을 구별하는데 좋은 단서가 될 수 있다. 본 논문에서는 이중 가장 흔히 쓰이는 박 'bgcolor'와 〈font〉 태그의 의미 있는 표에서 쓰이는 특징을 우선 관찰하였으며 그 결과는 다음과 같다.

1. 이 태그들은 적어도 2가지 값을 가진다.
2. 표는 이 태그들의 값을 기준으로 여러 구역으로 나뉠 수 있다.
3. 이 태그들이 2가지 값을 가질 때 의미 있는 태그로써 가장 효과적이다.
4. 영역의 모양은 대부분 사각형이다.

위의 관찰을 토대로 우리는 표 2와 수식 1에서 4까지의 표 구별 특성을 설정하였다.

번호	구별 특성
1	배경색에 따른 헤드의 존재 가능 정도 (DB)
2	글자 속성에 따른 헤드의 존재 가능 정도 (DF)

[표 2] 헤드의 존재 가능 정도에 따른 외형적 구별 특성

$$DB = \begin{cases} 0 & if\ bn = 1 \\ \dfrac{1}{bn-1} + (bn - bc) & otherwise \end{cases} \quad (1)$$

$$bc = \sum_{i=1}^{bn} \sum_{j=1}^{ba} baa_j - brn \quad (2)$$

$$DF = \begin{cases} 0 & if\ fn = 1 \\ \dfrac{1}{fn-1} + (fn - fc) & otherwise \end{cases} \quad (3)$$

$$fc = \sum_{i=1}^{fn} \sum_{j=1}^{fa} faa_j - frn \quad (4)$$

bn : 표 내에서 배경색의 종류
ba : 표 내에서 배경색에 따라 결정되는 영역의 개수
baaj : 배경색에 따라 결정되는 j번째 영역의 값 2 –j번째 영역이 한 개의 셀로 구성되거나 사각형이 아닐 경우 1 –그 이외
brn : 같은 배경색으로 반복되는 영역의 개수
bc : 배경색에 의해 결정되는 영역의 판정지수
fn : 표 내에서 글자 속성의 종류
fa : 표 내에서 글자 속성에 따라 결정되는 영역의 개수
faa : 글자 속성에 따라 결정되는 j번째 영역 값 2 –j번째 영역이 한 개의 셀로 구성되거나 사각형이 아닐 경우 1 –그 이외
frn : 같은 글자 속성으로 반복되는 영역의 개수
fc : 글자 속성에 의해 결정되는 영역의 판정지수

위 수식에서 bc와 fc는 각 속성에 따른 영역의 판정지수로 이 수와 속성의 종류가 같은 경우가 일반적이다. 즉 표를 작성할 때 쓰이는 속성의 개수가 곧 영역의 개수개 된다. 따라서, bn과 bc가 2이며 fn과 fc가 2일 때 결과가 1이 되며, 이 경우가 수식에서 가장 높은 값이다. 이 경우는 'bgcolor'나 〈font〉 태그에 연관된 속성이 2개 나타나고, 각 속성에 따라 헤드 영역과 데이터 영역으로 나뉘는 것으로 의미 있는 표일 확률이 가장 높다. 즉 위의 관찰 중 3번째 사항을 반영한 것이다. 그리고 DB와 DF가 0값을 가질 때는 속성값이 1개인 경우이다(속성 값이 표기되어 있지 않으면 기본적으로 1개이다). DB와 DF 값이 음수가 나오는 경우는 표에 영역의 개수가 태그의 속성의 개수보다 많은 경우이다. 이 경우는 꾸미는 표일 확률이 매우 높다. faa와 baa는 표 내부의 영역의 모양에 대한 값인데, 영역이 한 개의 셀로 이루어져 있거나 사각형이 아니면 바라는 영역의 모양이 아니므로 2를 주어 페널티를 부여하고 아닌 경우에는 1을 준다. 하지만 헤드가 가장 윗 행과 가장 왼쪽 열에 동시에 위치할 수 있으므로 이 경우는 페널티를 부여하지 않는다. 이는 4번째 관찰을 반영한 것이다.

표 2에서 제시한 구별 특성 이외에 표는 독자에게 보다 명확하게 뜻을 전달하기 위하여 다른 다양한 특성을 가진다. 꾸미기 위한 표와 의미 있는 표는 작성의 목적이 궁극적으로 다르기 때문에 형태적으로 다른 특징을 보이는데, 예를 들어 꾸미기 위한 표는 또 다른 표를 정렬하기 위하여 내용으로 가지기도 하며 시각적인 효과를 위하여 경계선도 가지지 않는 경우가 흔한 반면에, 의미 있는 표는 표 내부에 표를 가지는 경우를 거의 볼 수 없으며, 표 내부의 내용의 구별을 명확하게 보여주기 위하여 경계선을 가지기도 한다. 이러한 여러 가지 외형적인 특성은 이미 본 연구의 선행 연구에서 제시하였으며, 본 연구에서도 그 구별특성을 사용한다. 이는 다음 표3와 표4에 나타내었다.

번호	구별 특성
3	〈caption〉 태그의 유무
4	〈th〉 태그의 유무
5	border 속성의 유무
6	표 안에 표가 있는지 유무
7	전체 셀에서 수치데이터만으로 이루어진 행이나 열의 유무
8	전체 셀에서 내용이 없는 셀의 비율
9	전체 셀에서 image를 가진 셀의 비율
10	전체 셀에서 링크를 가진 셀의 비율
11	전체 셀에서 〈input〉 태그를 가진 셀의 비율
12	전체 셀에서 문자형 내용으로만 된 셀의 비율
13	전체 셀에서 특수문자를 포함하는 셀의 비율
14	전체 셀에서 숫자를 포함하는 셀의 비율
15	셀 안에 40문자가 넘는 절의 개수
16	표의 형태

[표 3] 외형적 구별 특성

번호	구별 특성 값
3	0 : 없음 1 : 존재
4	0 : 없음 1 : 존재
5	0 : 없음 1 : 존재
6	0 : 없음 1 : 존재
7	0 : 없음 1 : 존재
8	차지하는 비율
9	차지하는 비율
10	차지하는 비율
11	차지하는 비율
12	차지하는 비율
13	차지하는 비율
14	차지하는 비율
15	절의 개수(n) (만약 n 〉 10 이면 n = 10)
16	0 : span속성 없음 1 : 헤더에 행 span 속성 존재 2 : 헤더에 열 span 속성 존재 3 : 헤더에 행과 열 span속성 모두 존재 4 : 데이터 영역에 span 속성 존재

[표 4] 표 3에 대한 구별 특성 값

3.4. 연속적 구별 특성

'연속적 구별 특성'은 표의 내용의 분포가 얼마나 유사하며 반복적인가를 판단한다. 일반적으로 의미 있는 표는 그 내용이 반복적이고 유사하다. 한 예로 그림 4의 a를 보면 헤드는 표의 가장 윗 행이며 이 행은 3개의 셀을 가지고 있다. 이 행의 각 셀은 그 셀에 따르는 데이터 영역의 열의 데이터 부류를 결정한다. 즉, 그림 2의 1행 1열에 있는 'UNCode'에 의해 1열에 있는 데이터의 형태가 숫자로 통일되는 것이다. 이와 같이 내용이 얼마나 통일되어 있느냐를 알아냄으로써 의미 있는 표와 꾸미기 위한 표를 구별할 수 있다. 이는 우리가 흔히 사용하는 표준 편차 공식과 이웃 셀과의 내용 차이를 이용하여 계산한다. 이는 다음 표 5에 나타내었다. 이 표에서 모든 값은 0에서 1사이의 값으로 나타나며, 0에 가까울수록 분포가 고른 것이며, 의미 있는 표에 가깝다는 의미이다.

Table 2. Total population by country

UNCode	Place Name	Pop1990
004	Afghanistan	16556000
008	Albania	3250001
012	Algeria	24960006
020	Andorra	55300
024	Angola	9194019
660	Anguilla	6900

[그림 4] 의미 있는 표의 예

번호	구별 특성
17	행의 개수에 대한 표준 편차 (dR)
18	열의 개수에 대한 표준 편차 (dC)
19	행을 기준으로 한 셀 내 문자열 길이의 표준편차의 평균 (dCLR)
20	열을 기준으로 한 셀 내 문자열 길이의 표준편차의 평균 (dCLC)
21	행 내용 속성의 반복성 (CCR)
22	열 내용 속성의 반복성 (CCC)
23	행의 태그 반복성 (TCR)
24	열의 태그 반복성 (TCC)
25	행을 기준으로한 헤드 내용 속성의 반복성 (dHRR)
26	열을 기준으로 한 헤드 내용 속성의 반복성 (dHRC)

[표 5] 연속적 구별 특성

$$dC = \sqrt{\frac{1}{rn}\sum_{i=1}^{rn}(c_i - c)\times(c_i - c)} \quad (5)$$

$$dR = \sqrt{\frac{1}{cn}\sum_{j=1}^{cn}(r_j - r)\times(r_j - r)} \quad (6)$$

$$dCLR = \frac{1}{rn}\sum_{i=1}^{rn}\sqrt{\frac{1}{c_i}\sum_{j=1}^{c_i}(cl_{ij} - clr_i)\times(cl_{ij} - clr_i)} \quad (7)$$

$$dCLC = \frac{1}{cn}\sum_{j=1}^{cn}\sqrt{\frac{1}{r_j}\sum_{i=1}^{r_j}(cl_{ij} - clc_j)\times(cl_{ij} - clc_j)} \quad (8)$$

$$CCR = \frac{1}{rn}\sum_{i=1}^{rn}\left\{\frac{1}{c_i - 1}\sum_{j=2}^{c_i} diff(cc_{i,j-1}, cc_{i,j})\right\} \quad (9)$$

$$CCC = \frac{1}{cn}\sum_{j=1}^{cn}\left\{\frac{1}{r_j - 1}\sum_{i=2}^{r_j} diff(cc_{i,j-1}, cc_{i,j})\right\} \quad (10)$$

$$TCR = \frac{1}{rn}\sum_{i=1}^{rn}\left\{\frac{1}{c_i - 1}\sum_{j=2}^{c_i} diff(ct_{i,j-1}, ct_{i,j})\right\} \quad (11)$$

$$TCC = \frac{1}{cn}\sum_{j=1}^{cn}\left\{\frac{1}{r_j - 1}\sum_{i=2}^{r_j} diff(ct_{i,j-1}, ct_{i,j})\right\} \quad (12)$$

$$dHRR = \sqrt{\frac{1}{rhr}\sum_{i=1}^{rhr}(rhc_i - rhc)\times(rhc_i - rhc)} \quad (13)$$

$$dHRC = \sqrt{\frac{1}{chr}\sum_{j=1}^{chr}(chc_j - chc)\times(chc_j - chc)} \quad (14)$$

c : 열 개수의 평균 $\frac{1}{rn}\sum_{i=1}^{rn} c_i$

c_i : i 행에 속하는 셀의 개수

rn : 행의 개수

r : 행 개수의 평균, $\frac{1}{cn}\sum_{j=1}^{cn} r_j$

r_j : the number of cells in column j

cn : 열의 개수

cl_{ij} : i행 j열에 위치한 셀 내의 문자 수

clr_i : i행에 속하는 셀 내의 문자 수의 평균

cc_{ij} : i행 j열에 속하는 셀 내의 내용 속성

clc_j : j열에 속하는 셀 내의 문자 수의 평균

ct_{ij} : i행 j열에 속하는 셀 내의 tag 속성

$diff(c_1,c_2)$: 0 : 만약 c1과 c2가 같다면, 1 : 그 이외

rhc : 행을 기준으로 반복되는 헤드 내용 속성의 거리의 평균 $\frac{1}{rhr}\sum_{i=1}^{rhr} rhc_i$

rhc_i : 행을 기준으로 반복되는 헤드 내용 속성이 있을 때 이전 내용과의 행의 차

rhr : 행을 기준으로 반복되는 헤드 내용 속성의 개수

chc : 열을 기준으로 반복되는 헤드 내용 속성의 거리의 평균
chc_j : 열을 기준으로 반복되는 헤드 내용 속성이 있을 때 이전 내용과의 열의 차
chr : 열을 기준으로 반복되는 헤드 내용 속성의 개수

3.5. 의미 있는 표 추출

앞 절에서 우리는 의미 있는 표와 꾸미기 위한 표를 구별하기 위한 26개의 구별특성을 설정하였다. 우리는 의미 있는 표를 추출하기 위하여 여과 단계와 기계 학습 단계의 2단계로 나눠서 이 구별 특성을 적용한다. 먼저 여과 단계에서는 꾸미기 위한 표라는 확신이 있는 표를 거른다. 이는 다음과 같은 규칙으로 판단할 수 있다.

1. 1차원으로 구성된 표
2. 표의 내용이 없거나 태그형 내용만 있는 표

이 두 규칙 중 첫 번째 규칙은 행과 열의 수로 설정할 수 있으며, 두 번째 규칙은 구별 특징 8, 9, 10, 11로 얻을 수 있다. 여과 단계는 2가지 목적에서 필요하다. 첫째, 개요 부분에서 언급했듯이 꾸미기 위한 표가 의미 있는 표보다 월등히 수가 많으므로 기계학습 단계에서 꾸미기 위한 표에 기계학습 결과가 편향되는 효과를 줄일 수 있다. 둘째, 표 추출 단계에서 꾸미기 위한 표를 미리 걸러 낼 수 있으므로 표 추출 단계의 속도를 높일 수 있다.

두 번째 단계는 기계학습 단계로 분류를 위한 모델을 설정해야 한다. 이를 위하여 우리는 먼저 인터넷 상의 표를 수집하고 수집된 표를 대상으로 앞서 설정한 구별 특성을 기준으로 구별 특성 값을 설정하여 학습 데이터 집합을 만든다. 즉 하나의 표는 26차원의 벡터로 표현되며, 이 벡터에 정답을 더해서 학습 데이터를 만든 후 기계학습 알고리즘에 적용한다. 기계학습 알고리즘은 C4.5 알고리즘을 사용하였으며, WEKA 학습 도구를 이용하여 성능을 측정하였다.

3.6. 헤드 영역과 데이터 영역의 구분

앞 절까지의 기술로 HTML문서에서 의미 있는 표를 추출해 낼 수 있다. 다음 단계는 의미 있는 표에서 헤드 영역과 데이터 영역을 분리해 내는 작업이다. 이 작업은 텍스트 마이닝이나 정보 추출을 위한 중요한 선행 작업이다. 이 단계는 표 2,3,4,5에서 기술한 구별 특징 중 헤드 영역과 데이터 영역을 분리하는 단서가 되는 구별 특징을 이용한다. 이는 다음과 같은 단계를 거친다.

1. 구별 영역 후보를 설정한다.
   (1) 〈th〉 태그로 구별되는 영역
   (2) 숫자가 다수를 차지하는 표에서 숫자와 숫자가 아닌 부분으로 구분되는 영역
   (3) 〈font〉 태그와 표 바탕색 속성으로 구분되는 영역
   (4) 〈span〉 태그로 구분되는 표 형태
   (5) 표 5의 내용 반복 속성
   (6) 표 5의 태그 반복 속성
2. 구별 영역 후보 중에서 가장 가능성이 높은 후보 영역을 선택한다.
3. header 영역의 반복을 검사한다.

위와 같은 단계를 거치면 최종 후보 영역이 선택될 것이며 이를 헤드 영역과 데이터 영역의 구분으로 정한다. 본 연구에서는 두 번째 단계인 가장 가능성이 높은 후보 영역의 설정을 단순히 첫 번째 단계의 우선 순위로 봤다. 즉, 후보(1)을 가장 우선 순위가 높은 후보영역으로 보았으며, 후보 (6)을 가장 낮은 우선 순위로 보았다. 따라서, 후보 (2)와 후보 (4)로 두 개의 영역이 결정되었을 경우에는 후보 (2)의 영역을 선택한다.

4. 실 험

본 논문에서 쓰이는 데이터 집합은 크게 두 종류 이다. 실험의 객관성을 확보하기 위하여 wang[5]의 논문에서 사용한 실험데이터를 사용하였다. wang의 데이터는 전체 1,393개의 인터넷 문서에서 1,739개의 의미 있는 표를 포함하고 있으며, 13,177개의 꾸미기 위한 표를 포함하고 있다.

4.1 의미 있는 표의 추출

3.5절에서 설명한 대로 의미 있는 표의 추출 여과 단계와 기계 학습 단계의 두 단계를 거친다. 여과 학습 단계에서 전체 14,570의 표 중에서 8,280개의 표가 여과 되었다. 다음으로 남은 6,290개의 표를 이용하여 학습 데이터를 만들었으며, C4.5알고리즘에 적용하였다. 평가는 학습 데이터를 9개의 집합으로 나눠서 8개의 집합을 이용하여 학습하고 나머지 한 개의 집합을 이용하여 평가하는 방법으로 9번 반복하였다. 그 결과는 아래와 같다.

실제 부류	학습 결과 부류	
	꾸미기 위한 표	의미 있는 표
꾸미기 위한 표	4444	107
의미 있는 표	103	1636

a. confusion matrix

	정확도	재현율	F-Measure
꾸미기 위한 표	0.977	0.976	0.977
의미 있는 표	0.939	0.941	0.940

b. 실험 결과

[표 6] 표 분류 기계학습 실험 결과

4.3. 헤드 영역과 데이터 영역 구분 단계

다음 표 7은 헤드 영역과 데이터 영역 구분 결과이다. 이는 3.6절에서 기술한 방법을 사용하였으며 73.78%의 정확도를 얻었다. 결과가 좋지 않은 이유는 여러 후보 집합에 대해서 선형적인 판단 방법을 사용했기 때문이다.

전체 의미 있는 표	1739
정확하게 구분 된 표	1283
정확도	73.8%

[표 7] 헤드 영역과 데이터 영역 구분 결과

5. 결 론

본 논문은 인터넷 문서에서 표에 포함된 정보를 추출하기 위한 방법을 제시했다. 이를 위해, 본 논문에서는 먼저 인터넷 상의 표를 관찰하여 의미 있는 표와 꾸미기 위한 표의 특성을 파악하였으며, 그에 따라 구별 특성을 설정하였다. 두 번째 단계로 추출된 구별 특성을 적용 하기 위한 3단계 방법을 제시하였으며, 각각 여과 단계, 기계 학습 단계, 헤드 영역과 데이터 영역 구분 단계로 나누었다. 그 결과 HTML 문서에서 의미 있는 표의 추출 정확도는 약 94%이며, 헤드 영역과 데이터 영역 구분은 약 73.8%의 f-measure값을 얻을 수 있었다. 본 논문의 방법은 HTML문서 중 특정 영역의 정보나 패턴을 사용한 것이 아닌, 인터넷 전체 문서에서 관찰된 범용적인 구별 특성을 사용하였으므로, 이전 연구에 비하여 모든 인터넷 문서에 적용 가능하며, 인터넷 문서 중 표에서 정보를 추출하는3단계 방법론은 제시하였다.

6. 향후 과제

현재 헤드 영역과 데이터 영역을 구분하는 규칙이 너무 단순하다. 여러 후보들을 종합하여 더욱 정확하게 판단할 수 있는 방법이 필요하다. 또한, 헤드 영역과 데이터 영역의 관계를 추출하는 규칙이 필요하며, 헤드 영역에서도 색인어의 상, 하위 관계를 추출할 수 있는 유용한 정보를 설정하여 개념간의 관계 설정을 위한 효과적인 모형을 설정해야 한다.

참고 문헌

[1] H.H. Chen, S.C. Tsai, and J.H. Tsai "Mining Tables from Large Scale HTML Texts", *In Proceedings of 18th International Conference on Computational linguistics*, Saabrucken, Germmany, July 2000.

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

[3] S.W. Jung, K.H. Sung, T.W. Park, and H.C. Kwon, "Effective Retrieval of Information in Tables on the Internet", *IEA/AIE (LNAI 2358)*, pp.493-501, June 2002

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

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

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

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