시각 자질을 이용한 의미 있는 테이블 검출

이재안* 박성배** 손정우* 이상조**
*경북대학교 전자전기컴퓨터학부
**경북대학교 컴퓨터공학과
e-mail : {jalee, sbpark, jwson, sjlee}@sejong.knu.ac.kr

Discriminating Meaningful Tables Using Visual Features

Jae-An Lee*, Seong-Bae Park**, Jeong-Woo Son*, Sang-Jo Lee**
*School of Electrical Engineering and Computer Science, Kyung-pook National University
**Department of Computer Engineering, Kyung-pook National University

요 약

웹 상에서의 정보추출은 방대한 데이터를 기반으로 널리 사용되고 있다. 테이블은 웹 페이지에서 요약된 정보를 보여주는 유용한 수단이기 때문에 테이블로부터의 정보추출은 일반적인 웹 데이터의 정보추출에 비해 중요하다. 하지만 웹 페이지에 나타난 테이블은 유의미한 정보를 가지는 의미 있는 테이블과 웹 페이지의 형태의 보정을 위한 장식 테이블로 나누어진다. 따라서 웹 페이지에서 의미 있는 테이블을 구분하고 정보를 검출하는 것은 웹 상에 나타난 정보를 활용하기 위한 중요한 단계이다. 본 논문은 웹 페이지에 나타난 테이블들 중 유의미한 정보를 내포하고 있는 의미 있는 테이블을 검출할 수 있는 방법을 제안한다. 이를 위해 본 논문에서는 브라우저를 통해 보여지는 테이블의 위치적 중요도를 반영하는 새로운 자질을 정의하고, 이를 기존 자질과 결합하여 활용함으로써 시각 자질의 유용성을 평가한다. 실험을 통해 본 논문에서 제안한 방법이 기존 방법들에 비해 우수한 성능을 보임을 알 수 있었다.

1. 서론

정보추출 대상이 되는 웹 페이지 속에는 정보를 효과적으로 표현하기 위한 다양한 방법이 있다. 특히 테이블에는 유의미한 많은 정보들이 저자들에 의해 재가공된 형태로 제공된다. 예를 들어 개인의 이력이나 제품의 규격서 등은 웹 상에서 테이블의 형태로 자주 표현된다. 즉 테이블로부터의 정보 추출은 일정한 형식을 가지고 있으며 일반 문장에 비해 요약된 형태의 정보를 제공하므로 테이블로부터의 정보 추출 작업은 시맨틱 웹이나 온톨로지 인스턴스 생성 등에 효과적으로 활용될 수 있다.

웹페이지로부터 테이블을 추출하기 위해서는 <TABLE>과 </TABLE>로 이루어진 테이블들을 수집해야 한다. 하지만 수집한 테이블들이 모두 유의미한 정보를 가지고 있는 것은 아니다. 웹 페이지에서 테이블의 쓰임은 크게 유의미한 정보를 내포하고 있는 “의미 있는 테이블”과 웹 페이지의 형태를 보정하기 위해 사용된 “장식 테이블”로 나누어진다. 따라서 수집한 테이블 중에서 의미 있는 테이블만을 검출해 내는 작업은 반드시 필요하다.

의미 있는 테이블을 검출하기 위한 다양한 연구들이 계속 진행되고 있다. 최근 기계 학습 기반의 연구들은 테이블의 내용과 형태를 표현할 수 있는 자질 집합을 정의하고, 이를 바탕으로 기계학습 기법을 이용하여 의미 있는 테이블을 검출하고 있다[1, 2, 3]. 최근 Son et al. [4]은 테이블의 구조적 자질을 명시화하지 않고 파스 트리 커널(Parse Tree Kernel) 을 이용하여 테이블의 구조적 특성을 반영하였다. 이 방법은 테이블의 계층적인 특징을 트리 형태로 표현하는 특징이 있지만, 파스 트리의 모든 서브셋 트리(Subset trees)를 비교하기 때문에 속도가 느리다는 단점을 가진다.

본 논문에서는 기존 연구에서 제안한 구조 자질(Structure Features)과 내용 자질(Content Features)에 더해 시각 자질(Visual Features)을 이용하여 의미 있는 테이블을 검출하는 방법을 제안한다. 본 논문에서 제안하는 시각 자질은 브라우저 상에서 실제로 보여지는 테이블의 위치와 크기를 평가하는 자질이다. 따라서 본 연구에서는 테이블의 구조적 특성을 표현하는 구조 자질과 테이블이 포함하고 있는 콘텐츠와 관련된 내용 자질 뿐만 아니라 테이블의 위치적 중요도를 평가하는 시각 자질을 바탕으로 의미 있는 테이블 검출 작업을 수행하고자 한다. 이를 위해 의사 결정 트리 (decision tree)[5]를 사용한다.

본 논문은 다음과 같이 구성된다. 2 장에서는 의미 있는 테이블 검출을 위한 문제를 정의하고, 3 장에서는 테이블 검출을 위해 제안한 세 가지 유형의 자질 집합에 대해 설명한다. 4 장에서 제안한 시각 자질의 성능을 평가하기 위한 실험 및 그 결과들을 보여준다. 마지막으로 5 장에서 결론과 향후 연구 방향에 대해 설명한다.

2. 의미 있는 테이블 검출

의미 있는 테이블 검출은 이진 분류 문제로 살펴볼 수 있다. 테이블의 집합 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 를 가정하자. 여기서 x_i 는 $< sf_i, cf_i, vf_i >$ 로 이루어진 테이블의 자질 벡터로써, 테이블의 구조 (*sf*), 내용 (*cf*), 시각 (*vf*) 자질을 의미한다. y_i 는 i 번째 테이블이 의미 있는 테이블인 경우 *+1*, 의미 없는 테이블인 경우 *-1* 을 가진다. 기계학습 측면에서의 의미 있는 테이블 검출은 함수 $f:\chi \rightarrow y$ 를 찾는 것이며, 본 논문에서는 함수 f 를 찾기 위해 의사 결정 트리[5]를 사용한다.

의사 결정 트리는 이진 분류 문제를 다루는 대표적인 기계학습 방법 중 하나로, 주어진 문제 공간을 수집된 각 데이터의 속성 값에 따라 몇 개의 공간으로 분류하고 그 결과를 바탕으로 특정 도메인에 대한 분류와 예측을 하는 방법이다. 의사 결정 트리는 목표값을 얻기 위한 분류 함수를 트리 구조로 도표화하기 때문에 모델을 평가하고 이해하기가 쉬워 분류 문제에서 널리 사용되고 있는 방법 중 하나이다. 본 논문에서는 세 가지 유형의 자질 벡터들을 의사 결정 트리의 속성값으로 사용하며, 의미 있는 테이블인지 아닌지의 여부를 목표값으로 두고 분류 함수를 생성한다.

3. 의미 있는 테이블 검출을 위한 자질들

본 논문에서는 의미 있는 테이블을 검출하기 위해 세 가지 유형의 자질 정보를 이용한다. 각각은 테이블의 구조 정보를 보는 구조 자질과 테이블에 포함된 콘텐츠를 보는 내용 자질, 브라우저 상에 나타나는 테이블의 위치적 중요도를 보는 시각 자질이다. 표 1 은 각 유형별 자질들과 그에 대한 간략한 설명을 보여준다. 본 논문에서 사용한 자질은 각 유형별로 구조 자질 7 개, 내용 자질 7 개, 시각 자질 4 개로 구성된다.

3.1 구조 및 내용 자질

테이블의 구조와 그 내용을 보는 구조 자질과 내용 자질은 의미 있는 테이블 검출을 위한 여러 연구에서 자주 사용되고 있다[1,2]. Wang [1]은 구조 자질과 내용 자질을 다음과 같이 정의하여 사용하고 있다.

먼저 구조 자질(Structure Features)은 HTML 태그에 기반하여 추출되는 테이블의 구조와 관련된 정보이다. 웹 페이지에서 테이블은 <TR>, <TD> 등과 같은 태그들의 셋으로 구성된다. 여기서 <TR> 태그 개수는 테이블의 행의 개수를 나타내며, <TD> 태그의 개수는 각 행의 열의 개수를 나타낸다. 따라서 <TR> 태그와 <TD> 태그에 의해 만들어지는 셀들은 테이블의 전체적인 구조에 직접적인 영향을 줄 수 있다. 대부분의 의미 있는 테이블은 행과 열에 대해 비슷한 개수의 셀을 가지며 비슷한 구조를 가질 것이다. 즉 의미 있는 테이블일수록 행과 열에 대한 셀의 개수와 길이의 평균값이 비슷하며 그 표준편차는 작다. 표 1 의 자질 1 에서 4 는 행과 열에 대한 셀 개수의 평균과 표준편차를 의미하고, 자질 5 와 6 은 셀 길이의 평균과 표준편차를 평가한다. 그리고 대부분의 의미 있는 테이블은 행 또는 열에 대해서 일정한 셀 길이를 가지는 반면, 의미 없는 테이블은 셀 길이에 일관성이 없는 특성을 자질 7 에서 평가한다.

표 1. 의미 있는 테이블을 검출하기 위한 자질들

구조 자질(Structure Features)	
1	각 행에 대한 셀의 평균 개수
2	각 행에 대한 셀의 표준편차
3	각 열에 대한 셀의 평균 개수
4	각 열에 대한 셀의 표준편차
5	모든 셀 길이의 평균
6	모든 셀 길이의 표준편차
7	주어진 테이블의 셀 길이의 일관성
내용 자질(Content Features)	
8	모든 셀 안에 포함된 <IMAGE> 태그의 개수
9	모든 셀 안에 포함된 <FORM> 태그의 개수
10	모든 셀 안에 포함된 <HYPERLINK> 태그의 개수
11	모든 셀 안에 포함된 글자 개수
12	모든 셀 안에 포함된 숫자 개수
13	모든 셀 중 빈 셀의 개수
14	주어진 테이블의 내용 자질의 일관성
시각 자질(Visual Features)	
15	테이블의 왼쪽 상단의 가로축 좌표
16	테이블의 왼쪽 상단의 세로축 좌표
17	테이블의 너비
18	테이블의 높이

다음으로 내용 자질(Content Features)은 셀 안에 포함된 콘텐츠들을 반영하는 것이다. 테이블은 정보를 보다 효과적으로 전달하기 위해 문자뿐만 아니라 숫자, 이미지, 하이퍼링크 등 다양한 종류의 콘텐츠를 포함하고 있다. 예를 들어 IT 관련 제품들의 규격, 사양 등을 요약할 때는 텍스트, 숫자, 이미지 등을 일관성있게 사용하여 필요한 정보를 표현하게 된다. 이에 반해, 의미 없는 테이블은 빈 셀이 많거나 하이퍼링크만이 나열된 경우가 빈번하다. 자질 8 에서 13 은 각각 <IMAGE> 태그, <FORM> 태그, <HYPERLINK> 태그, 글자 수, 숫자 수, 빈 셀의 개수를 평가한다. 또한 의미 있는 테이블은 행 또는 열에 대해서 포함하고 있는 콘텐츠들이 일관성을 가지는 특성이 있으므로 이를 자질 14 에서 평가한다.

3.2 시각 자질

본 논문은 의미있는 테이블을 검출하기 위해 브라우저 상에서의 테이블의 위치적 중요도를 살피는 시각 자질(Visual Features)을 제안한다.

정보를 전달하기 위해 만들어진 테이블은 웹 페이지에서 가운데 넓게 위치하는 경향이 있는 반면, 리스트, 메뉴와 같이 정보를 담고 있지 않는 테이블은 페이지의 가장자리에 위치하는 경향이 있다. 또한 웹

페이지의 형태를 보정하기 위한 테이블은 다양한 위치에서 나타날 수 있으며, 의미 있는 테이블보다 크기가 큰 경향이 있다.

실제로, 실험에 사용한 데이터 셋에서 의미 있는 테이블들의 중첩되는 범위가 장식 테이블의 중첩되는 범위보다 좁고, 브라우저의 가운데에 위치하는 경향을 볼 수 있었다. 또한, 의미 있는 테이블과 장식 테이블이 너비면에서는 비슷하지만 높이면에서는 의미 있는 테이블이 장식 테이블보다 2 ~ 3 배 정도 큰 경향을 살펴볼 수 있었다. 여기서, 의미 테이블이 높이면에서 장식 테이블보다 높다는 관측은 실험에 사용한 Wang[1]의 데이터 셋이 리프 테이블(Leaf Table)[1]로만 이루어져 있기 때문이다.

결과적으로 의미 있는 테이블은 장식 테이블에 비해 효과적인 정보 전달을 위해 웹 브라우저의 중심부에 위치하며, 넓이가 큰 경향을 보인다. 따라서 브라우저 상에서의 테이블의 위치와 크기에 관련된 이러한 시각 자질은 의미 있는 테이블 검출에 유용하게 사용될 수 있다.

본 논문에서는 브라우저 상에 나타난 테이블의 좌표와 테이블의 너비, 높이 자질을 시각 자질 유형으로 정의하며, 자질 15 에서 18 이 이를 평가한다.

4. 실험 및 평가

실험은 Wang[1]이 사용한 데이터 셋을 이용하여 수행하였다. 데이터 셋은 1,393 개의 웹 페이지로 구성되어 있으며, 이들은 11,477 개의 리프 테이블을 가진다. 그 중 1,740 개는 의미 있는 테이블이며, 나머지 9,737 개는 의미 없는 장식 테이블이다.

의미 있는 테이블 검출을 위해 실험은 크게 네 가지 관점에서 수행되었다. 모든 실험은 9-fold cross validation 방식으로 정확률(P)과 재현율(R), F-measure(F)를 척도로 평가하였다. 여기서 F-measure 는 정확률과 재현율의 평균을 의미한다. 첫 번째 실험은 유용한 시각 자질을 선별하기 위해 수행하며, 두 번째 실험은 Wang[1]의 자질들과 본 논문에서 제안한 시각 자질의 성능을 비교 평가하기 위해 수행한다. 그리고 본 논문에서 제안한 시각 자질과 구조, 내용 자질들과 결합 성능을 비교 평가하며, 마지막으로 제안한 방법과 기존 연구들의 성능을 비교한다.

먼저, 본 연구에서는 테이블 검출에 사용될 시각적 자질을 선별하기 위해 6 개(Sx, Sy, Ex, Ey, W, H)의 후보 자질을 선정하고 각 후보 자질을 결합하여 실험을 수행하였다. 그 결과는 표 2 와 같다. 표 2 에서 Sx, Sy 는 테이블 왼쪽 상단의 가로축 좌표와 세로축 좌표를 평가하는 자질이며, Ex, Ey 는 테이블 오른쪽 하단의 가로축 좌표와 세로축 좌표를 평가하는 자질이다. 자질 W 와 H 는 테이블 가로 길이와 테이블의 세로 길이를 평가한다. O 는 해당 자질이 각 실험에 사용되었음을 의미한다.

표 2 에서 볼 수 있듯이, 테이블의 왼쪽 상단 좌표(Sx, Sy)만을 이용하는 (a)와 테이블의 오른쪽 하단 좌표(Ex, Ey)만을 이용하는 (b)는 비슷한 성능을 보여주고 있다. 테이블의 가로, 세로 길이(W, H) 만을 이용한 (c)는 다른 자질들에 비해 낮은 성능을 보이지만 상단 또는 하단 좌표 자질과 결합한 (e), (f), (g)에서 높은 성능을 보여준다. (e), (f), (g)는 성능면에서 크게 차이가 나지 않으므로 본 논문에서는 테이블의 왼쪽 상단 좌표(Sx, Sy) 자질과 테이블의 가로, 세로 길이(W, H) 자질을 의미 있는 테이블 검출을 위한 자질로써 선별하였다.

표 2. 시각 자질 선별을 위한 성능 평가

	시각 자질						성능(%)		
	Sx	Sy	Ex	Ey	W	H	P	R	F
(a)	O	O					84.4	86.6	84.0
(b)			O	O			83.6	86.0	83.2
(c)					O	O	71.8	84.7	77.7
(d)	O	O	O	O			87.6	88.6	88.1
(e)	O	O			O	O	**90.8**	**91.0**	**90.9**
(f)			O	O	O	O	90.8	90.8	90.8
(g)	O	O	O	O	O	O	90.8	90.9	90.9

다음으로 본 논문이 제안한 시각 자질의 성능을 기존 자질의 성능과 비교하기 위한 실험을 수행하였다. Wang[1]의 연구에서 보여진 구조, 내용 자질의 성능과 본 논문이 제안한 시각 자질의 성능에 대한 결과는 표 3 과 같다.

표 3. 세가지 유형별 자질에 대한 성능 평가

		정확률	재현율	F-Measure
Wang[1]	구조 자질	88.2%	87.2%	87.7%
	내용 자질	**95.7%**	90.8%	**93.3%**
시각 자질		90.8%	**91.0%**	90.9%

표 3 과 같이, 본 논문에서 제안한 시각 자질만을 사용하더라도 대략 91% 의 f-measure 를 보여주며, 이는 구조 자질만을 사용하는 것에 비해 좋은 성능을 보인다. 또한 시각 자질은 구조 자질 및 내용 자질과 결합하여 성능을 평가하였을 때 우수한 성능을 보여준다. 이는 세 가지 유형의 자질들을 서로 결합하여 성능을 평가한 다음 실험에서 살펴볼 수 있다. 표 4 는 그 결과를 보여준다.

앞에서 언급한 바와 같이, 본 논문에서 제안한 시각 자질과 기존 자질들을 결합한 성능이 Wang[1]의 구조와 내용 자질을 결합한 성능보다 우수함을 볼 수 있었다. 다시 말해, 시각 자질과 기존 자질들을 결합한 성능이 구조와 내용 자질만을 결합한 성능보다 0.7% 이상의 성능 향상을 보여주었다. 특히 세 가지 유형의 자질을 모두 결합한 경우에는 약 2%의 성능

[1] 리프 테이블(Leaf Table)은 내부에 또 다른 테이블은 포함하지 않는 테이블을 말한다.

향상을 보인다. 따라서, 시각 자질은 구조 혹은 내용 자질과는 다른 관점에서 테이블을 평가하고 있으므로 의미 있는 테이블 검출에 중요한 자질이다.

표 4. 결합 자질들의 성능 평가

결합한 자질	정확률	재현율	F-Measure
구조+내용	94.2%	97.3%	95.7%
구조+시각	96.5%	96.4%	96.4%
내용+시각	97.6%	97.6%	97.6%
구조+내용+시각	**97.7%**	**97.7%**	**97.7%**

마지막으로, 본 논문에서 제안한 방법의 성능을 평가하기 위해 Wang 의 방법[1]과 Jung et al.이 제안한 방법[2], Penn 이 제안한 방법[3] 들과 비교 평가하였다. Penn[3]은 전문가들로부터 의미 있는 테이블을 선별하기 위한 몇 가지 규칙을 정의하고 이를 통해 실험을 수행하였다. Wang[1] 은 세가지 유형의 자질 집합을 제안하고 의사 결정 트리와 SVMs[6]을 사용하여 실험을 수행하였으며, SVMs 의 경우 두 가지 커널 함수를 사용하였다. 표 5 는 본 논문에서 제안된 방법과 기존 연구들의 실험결과를 보여준다.

표 5. 성능 비교 실험 결과

		정확률	재현율	F-measure
Penn[3]		86.3%	89.8%	88.1%
Wang[1]	의사결정트리	97.5%	94.3%	95.9%
	SVMs (선형커널)	91.4%	93.9%	92.7%
	SVMs (RBF 커널)	95.8%	96.0%	95.9%
Jung [2]		93.4%	94.8%	94.1%
제안한 방법		**97.7%**	**97.7%**	**97.7%**

표 5 에서 볼 수 있듯이, Penn[3]의 방법이 가장 낮은 결과를 보이는 것은 테이블 검출이 전문가에 의해 정의된 간단한 규칙만으로는 높은 성능을 낼 수 없음을 의미한다. Wang[1] 이 제안한 의사 결정 트리를 활용한 방법과 SVMs 기반의 방법은 서로 비슷한 성능을 보였다. Jung[2]의 방법은 Wang 의 방법보다 조금 낮은 성능을 보이나, 이 방법은 도메인에 의존적이 않는 특징을 보인다. 본 논문에서 제안한 방법이 기존 연구들과 비교할 때 97.7%의 가장 높은 성능을 보여주었다. 따라서 본 논문에서 제안한 시각 자질은 의미 있는 테이블 검출 작업에 효과적으로 사용될 수 있으며, 구조, 내용 자질만을 사용할 때보다 높은 성능을 보여준다.

5. 결론

웹 페이지의 테이블은 일반적인 문장에 비해 사용자에 의해 요약된 정보를 간결하고 효과적으로 전달하고 있다. 따라서 웹 페이지에 나타난 많은 테이블 중에서 유의미한 정보를 내포하고 있는 테이블들을 구분해 내는 작업은 테이블로부터의 정보 추출에서 반드시 필요한 작업이다. 이를 위해, 주어진 테이블이 정보를 가진 의미 있는 테이블인지 아닌지의 여부를 판단할 수 있어야 한다.

본 논문은 웹 페이지 상에 나타난 테이블이 정보를 가진 의미 있는 테이블인지 혹은 웹 페이지의 형태를 보정하기 위한 장식 테이블인지를 판단하는 테이블 검출 방법을 제안하였다. 즉, 본 논문에서는 브라우저 상에서의 테이블의 위치적 중요도를 평가하는 새로운 자질을 제안하고, 기존 연구에서 제안된 구조 및 내용 자질들과 결합하여 활용함으로써 의미 있는 테이블 검출을 위한 성능 향상에 기여함을 보였다.

향후 연구에서는 다양한 데이터 셋을 이용하여 자질의 성능을 비교하고, 의사 결정 트리 뿐만 아니라 다양한 기계 학습 방법을 적용한 의미 있는 테이블 검출을 시도하고자 한다. 또한 의미 있는 테이블로부터 추출된 정보들을 이용하여 시맨틱 웹, 온톨로지 인스턴스 생성 등의 작업에 활용하고자 한다.

참고문헌

[1] Y. Wang and J. Hu. A Machine Learning based Approach for Table Detection on the Web. In *Proceedings of the 11th International World Wide Web Conference*, pages 242–250, 2002.

[2] S. Jung and H. Kwon. A Scalable Hybrid Approach for Extracting Head Components from Web Tables. In *Proceedings of Institute of Electrical and Electronics Engineers Trans Actions on Knowledge and Data Engineering*, pages 174–187, 2006.

[3] G. Penn, J. Hu, H. Luo, and R. McDonald. Flexible Web Document Analysis for Delivery to Narrow-bandwidth Devices. In *Proceedings of the 6th International Conference on Document Analysis and Recognition*, pages 119–130, 2004.

[4] J. Son, J. Lee, and S. Park. Discriminating Meaningful Web Tables from Decorative Tables Using a Composite Kernel. In *Proceedings of Web Intelligence 2008*, pages 368–971, 2008.

[5] Y. Yuan and M. Shaw, Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, 69, pages 125–139, 1995.

[6] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and other Kernel-based Learning Methods. *Cambridge University Press*, 2000.