한국지능시스템학회 논문지 2011, Vol. 21, No. 5, pp. 618-623
http://dx.doi.org/10.5391/JKIIS.2011.21.5.618

장식 테이블과 의미 있는 테이블 식별을 위한 커널 기반의 구조 자질

Kernelized Structure Feature for Discriminating Meaningful Table from Decorative Table

손정우[1] · 고준호[1] · 박성배[1†] · 김권양[2]
Jeong-Woo Son, Junho Go, Seong-Bae Park[†], and Kweon Yang Kim

경북대학교 IT대학 컴퓨터공학과1
경일대학교 컴퓨터공학과2

요 약

본 논문에서는 구조 정보를 활용하기 위한 결합 커널 기반의 의미 있는 웹 테이블과 장식 웹 테이블을 구분하는 새로운 방법을 제안한다. 본 논문에서 테이블의 구조 정보는 두 가지 형태의 구문 분석 트리로부터 추출된다. 컨텍스트 트리는 테이블 주변에 나타난 구조를 반영하고 있으며, 테이블 트리는 테이블 내의 구조를 담고 있다. 두 트리로 표현되는 테이블의 구조 정보를 효과적으로 다루기 위해 파스 트리 커널 기반의 결합 커널을 제안한다. 제안한 결합 커널을 적용한 support vector machines은 풍부한 구조 정보를 활용하여 의미 있는 테이블과 장식 테이블을 분류한다.

키워드 : 테이블 분류, 파스 트리 커널, Support vector machines, 결합 커널, 분류 문제

Abstract

This paper proposes a novel method to discriminate meaningful tables from decorative one using a composite kernel for handling structural information of tables. In this paper, structural information of a table is extracted with two types of parse trees: context tree and table tree. A context tree contains structural information around a table, while a table tree presents structural information within a table. A composite kernel is proposed to efficiently handle these two types of trees based on a parse tree kernel. The support vector machines with the proposed kernel dised kuish meaningful tables from the decorative ones with rich structural information.

Key Words : Table discrimination, Parse tree kernel, Support vector machines, composite kernel, classification

1. 서 론

웹 페이지에서 테이블은 정확한 정보를 간략한 형태로 기술하고 있다. 이러한 웹 테이블의 특성으로 인해 데이터 마이닝, 정보 요약 등, 웹 페이지 상의 정보를 추출하는 다양한 분야에서 웹 테이블 상의 정보를 추출하는 것이 널리 연구되고 있다.

테이블에서 정보를 추출하기 위해서는 먼저, 웹 페이지 상에 나타난 테이블을 추출해야한다. 웹 페이지로부터 모든 테이블을 추출하는 것은 웹 페이지를 기술한 HTML(Hypertext Markup Language)로부터 <TABLE> 태그를 추출하는 것과 동일한 일이며 이는 매우 쉽게 할 수 있다. 하지만, 추출된 테이블이 정보를 가지는 가를 판별하는 것은 어려운 일이다. 웹 페이지 상의 테이블은 정보를 기술하는 의미 있는 테이블과 웹 페이지의 형태를 유지하기 위한 장식 테이블로 나누어 진다. 이들 테이블은 HTML 레벨에서 구분이 불가능하기 때문에, 테이블이 정보를 가지는가를 판별하는 것은 어려운 일이다. 그림 1은 의미 있는 테이블과 장식 테이블의 예를 보여준다. 그림 1은 Wikipedia에서 검색한 제주도와 관련된 내용의 일부분이며 그림에서 실선으로 표시된 테이블은 의미 있는 테이블로, 제주도와 관련된 요약된 정보를 보여준다. 점선으로 표시된 부분도 역시 테이블이며 메뉴를 위한 레이아웃을 위해 사용되었다.

웹 테이블 중, 의미 있는 테이블을 추출하기 위해 다양한 연구들이 제안되어 왔다 [1, 2, 3]. Chen et al.은 경험적인 규칙과 테이블 셀(cell) 간의 유사도에 기반한 방법을 사용하여 장식 테이블과 의미 있는 테이블을 구분하였다 [4]. Jung et al.은 사람에 의해 설계된 자질을 바탕으로 결정 트리 알고리듬을 이용하여 테이블을 구별하는 방법을 제안하였다 [1]. Wang and Hu는 다양한 기계학습 방법을 테이블 분류 문제에 적용하였다 [5]. 이들은 테이블의 정보를 레이아웃, 컨텐츠, 단어 집합 자질로 표현한 후, 결정 트리, SVMs (Support Vector Machines)에 적용하여 테이블 분류 모델을 만들었다. Crestan과 Pantel [6]은 Wang과 Hu

접수일자 : 2011년 6월 16일
완료일자 : 2011년 9월 25일
†Corresponding Author : sbpark@sejong.knu.ac.kr
감사의 글 :
이 논문은 2010년도 정부(교육과학기술부)의 재원으로 휴먼인지환경사업본부-신기술융합형 성장동력사업의 일부 지원을 받아 수행된 연구임 (No. 2011K000659)

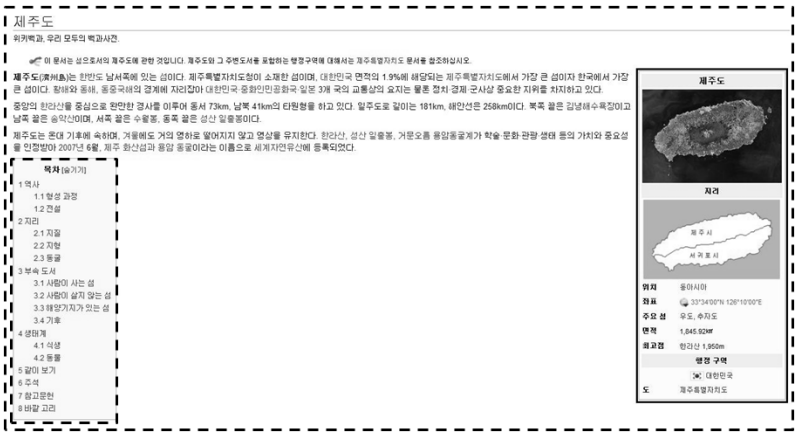

그림 1. 의미 있는 테이블과 장식 테이블의 예
Figure 1. An example of meaningful and decorative tables

가 제안한 자질과 함께 테이블의 다양한 종류를 함께 학습함으로써 성능을 높였다. 의미 있는 테이블 추출의 높은 성능을 기반으로 최근에는 분류된 의미 있는 테이블로부터 정보를 추출하는 다양한 연구가 제안되었다. Jung과 Kwon [7]은 의미 있는 테이블을 분류한 뒤, 테이블의 헤드를 추출하는 연구를 진행 하였으며, 이 외에도 테이블에 나타난 정보를 추출하기 위한 다양한 연구가 제안되었다 [8, 9].

기존의 의미 있는 테이블 분류를 위한 연구들은 의미 있는 테이블 구별은 위한 더 좋은 자질이나 규칙을 어떻게 구축할 것인가에 초점을 두고 있다. 이러한 자질들은 기존 연구에서는 사람의 손을 빌려 디자인 되었다. 기존 연구에서도 나타나듯이 테이블의 정보는 크게 구조 정보와 컨텐츠 정보로 나뉠 수 있다. 사람에 의해 구축된 자질들은 두 가지 정보 중, 구조 자질을 표현하기에 적합하지 않다. 이는 전문가라 할지라도 의미 있는 테이블과 장식 테이블을 구별하는 구조 자질을 정의하는 것이 쉽지 않은 일이기 때문이다. 따라서 기존 연구에서 제안된 구조 자질은 사람이 쉽게 인지할 수는 없지만 중요한 구조 정보를 고려하지 않을 가능성이 크다.

본 논문에서는 새로운 구조 자질을 활용한 테이블 구별 방법을 제안한다. 웹 테이블의 구조 정보는 HTML 파서에 의해 생성되는 HTML 구문 분석 트리로 표현된다. 본 논문에서는 구조 정보를 테이블의 구문 분석 트리 뿐 아니라 테이블 주변의 구문 분석 트리로 표현한다. 따라서 테이블의 구조 정보는 테이블의 구조를 나타내는 테이블 트리와 주변의 구조 정보를 내포한 컨텍스트 트리로 표현된다. 추출된 구문 분석 트리들은 고차원의 데이터를 쉽게 비교할 수 있는 커널 함수를 나타내는 정보를 측정할 수 있 있 있0]. 다양한 커널 함수 중, 컨볼루션 커널 있1]은 트리, 그래프, 문자냴는등 구조를 가지는 데이터를 비교 본 논문에제안되었다. 이 중, 파스 트리 커널 있2]은 구문 분석 트리들의 구조 정보를 측정 본 논문에제안되었다. 따라서 컨텍스트 트리와 테이블 트리로 표현되는 데이터들은 파스 트리 커널있는 커널 함비교된다. 본 논문에서는 구조 정보를]은 구두 구문 분석 트리에 대문에각각 파스 트리 커널있는리와구문 분전체 구조 정보구조 정보는 가 파스 트리 커널는 테이블의결함 커널있는활용널 함계산한다. 따라서 정보를결합 커널은 기존 연구와 달리 사람이 구조 정보를]은하기 위한 자질을 정의할 필요가 없다. 결과적 커널 정보를결합 커널은 사람이 정의할 경°이 리와 기 힘든 구조 정보까지 활용할 수 있 . 정보를결합 커널은 구조 정 트리와이블의의미 있는 테이블과 장식 테이블을 분류한다.

제안한 방법을 검증하기 위한 실험에서 제안한 구조 자질은 사람이 정의하기 힘든 구조 정보를 효율적으로 반영하여, 기존의 구조 자질에 비해 4% 이상 향상된 성능을 보였다. 사람에 의해 구축된 컨텐츠 자질과의 결합을 통한 성능 실험에서는 약 98%의 성능으로 기존의 테이블 분류 시스템에 비해 더 나은 성능을 보였다.

2. 웹 테이블의 분류

2.1. 이진 분류 문제로써의 웹 테이블 구별

의미 있는 테이블과 장식 테이블을 구분하는 문제는 이진 분류 문제로 볼 수 있다. 테이블 데이터 $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$가 주어진다고 가정하자. 이때, $y_i\in\{-1,+1\}$이고, $x_i=<ct_i,tt_i>$이다. 클래스 레이블 $y_i=+1$은 웹 테이블 x_i가 의미 있는 테이블임을 말하고, 반대로, $y_i=-1$는 x_i가 장식 테이블임을 의미한다. 웹 테이블 x_i는 두 가지 자질로 이루어져 있다. ct_i 는 테이블 주변의 구문 분석 트리를 의미하는 컨텍스트 트리이며, tt_i 는 테이블의 구문 분석 트리를 의미하는 테이블 트리이다. 이외에도 컨텐츠 정보를 담고 있는 컨텐츠 자질이 널리 사용되지만, 본 논문에서는 설명의 용이성을 위하여 생략하였다.

기계학습 관점에서 두 테이블을 구별하는 문제는 함수 $f:X\rightarrow Y$를 추정하는 문제로 볼 수 있다. 이때, 함수 f는 파라미터 w를 가지므로, 함수를 추정하는 것은 w를 추정하는 것으로 볼 수 있다. 분류 문제에서 파라미터를 추정하기 위해 가장 널리 사용되는 모델이 SVMs이다.

2.2. Support Vector Machines

SVMs는 주어진 데이터를 분류할 수 있는 hyperplane를 찾는 것을 목표로 한다. 두 클래스를 분류 할 수 있는 hyperplane는 무한히 생성할 수 있다. SVMs는 두 클래스 에 속한 데이터 중, 가장 가까운 데이터인 support vector들과

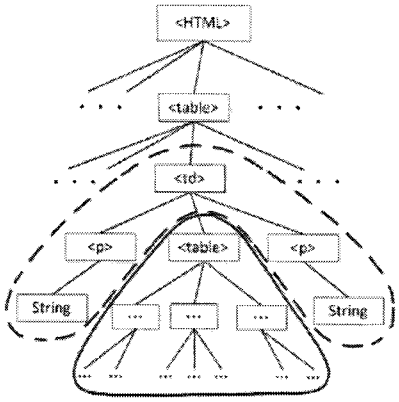

그림 2 테이블 트리와 컨텍스트 트리의 예
Figure 2. An example of table and context trees

의 거리가 가장 먼 hyperplane를 최종 모델로 선택한다.

학습 데이터 $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$가 주어진다면, SVMs에 의해 결정되는 최적의 hyperplane는 아래 조건을 만족해야한다.

이때, w와 b는 추정해야 할 파라메터이며, 이들 파라미터는 hyperplane와 support vector사이의 거리인 margin을 최대화하여야 한다. 이를 고려한 SVMs의 최적화 문제는

$$\min_{w,\xi_i}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i$$

이며, 아래 조건을 만족해야 한다.

$$y_i(w\cdot x_i+b)\geq 1-\xi_i,\ \xi_i\geq 0\ \ \forall i$$

위의 수식에서 ξ_i는 어느 정도의 오류를 허용하여 완벽히 분류되지 않는 경우에도 hyperplane를 찾기 위한 slack variable이며 C는 slack variable에 대한 패널티를 의미한다.

SVMs에서는 데이터가 선형 분류되지 않을 경우, 데이터들을 선형 분류가 가능한 더 높은 공간으로 사상하여 학습을 하게 된다. 이때, 데이터를 고차원의 공간에서 이들 사이의 유사도를 구하는 함수를 커널 함수라 한다. 커널 함수는 데이터를 직접 고차원 공간으로 사상하지 않고 유사도를 계산하며 이를 커널 트릭이라 한다.

3. 구문 분석 트리에 기반한 구조 자질

3.1. 구조 정보를 가지는 구문 분석 트리 추출

테이블의 정보는 구조 정보와 컨텐츠 정보로 나뉘어 진다 [13]. 본 논문에서는 이들 정보 중, 구조 정보를 다루기 위한 자질을 제안한다. 제안한 자질은 트리의 구조 정보를 이루는 두 가지 구문 분석 트리에 기반하고 있다. 먼저 트리 내의 구조 정보를 표현하는 테이블 트리가 있다. 테이블 트리는 HTML 문서의 <Table>태그와 </Table>태그 사이의 내용을 분석하여 얻어진 트리로 테이블이 가지는 구조

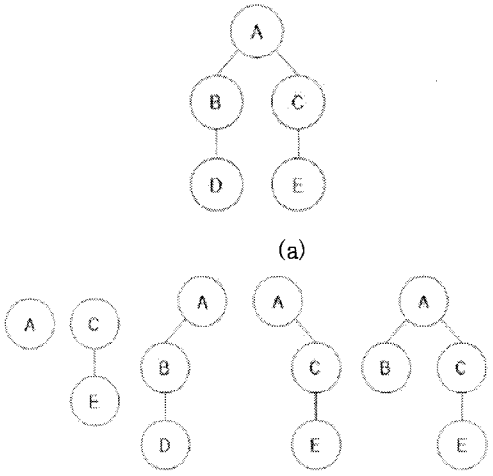

그림 3 (a) 구문 분석 트리 (b) 구문 분석 트리 (a)의 서브트리의 예
Figure 3. (a) Parse tree (b) a part of subtrees appeared in the parse tree (a)

를 의미한다. 다음으로 컨텍스트 트리는 테이블 트리를 감싸고 있는 구조 정보이다. HTML 문서는 하나의 구문 분석 트리로 표현 된다. 이 중, 테이블 트리를 감싸는 부분 트리가 컨텍스트 트리이다. 컨텍스트 트리는 테이블 주변에 나타난 구조 정보를 의미한다. 그림 2는 의미 있는 테이블의 테이블 트리와 컨텍스트 트리의 예를 보여준다. 그림에서 실선은 테이블 트리를 점선은 컨텍스트 트리를 의미한다. 일반적으로 의미 있는 테이블의 경우, 컨텍스트 트리는 문장, 이미지 등의 컨텐츠를 포함하고 있다. 이러한 컨텐츠들은 테이블에 관한 설명이 주를 이루며, 이는 테이블의 정보를 직접적으로 포함하고 있는 테이블 트리뿐만 아니라 테이블 주변의 구조 정보를 가지는 컨텍스트 트리 또한 테이블에 관련된 구조 정보를 가짐을 보여준다.

3.2. 파스 트리 커널을 활용한 유사도 측정

파스 트리 커널 [8]은 컨볼루션 커널 [7]의 하나로 파스 트리들을 다루는데 특화된 커널이다. 파스 트리 커널에서 벡터의 자질는 각 파스 트리에 나타날 수 있는 모든 서브 트리로 이루어진다. 이때, 각 자질의 값은 서브 트리의 빈도수로 할당된다. 그림2는 간단한 파스 트리에 대해 서브 트리들의 예를 보여 준다. 하지만 이러한 서브 트리를 명시적으로 구한다는 것은 불가능하다. 이에 Collins와 Duffy [8]는 명시적인 열거 없이 내적을 구하는 방법을 제시하였다.

$St_1,St_2,\cdots,St_m$을 파스 트리 T의 서브트리라 하면 파스 트리 T는 다음과 같이 벡터로 나타낼 수 있다.

$$V_T=(\#St_1(T),\#St_2(T),\cdots,\#St_n(T))$$

위 수식에서 $\#St_i(T)$는 St_i의 빈도수를 나타낸다. 두 파스 트리 T_1과 T_2사이의 내적은 아래와 같은 식으로 계산되어 진다.

$$< V_{T_1}, V_{T_2} >= \sum_i \#St_i(T_1) \cdot \#St_i(T_2)$$
$$= \sum_i (\sum_{n_1 \in N_{T_1}} I_{St_i}(n_1)) \cdot (\sum_{n_2 \in N_{T_2}} I_{St_i}(n_2))$$
$$= \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} C(n_1, n_2)$$

위 식에서 N_{T_1}과 N_{T_2}는 T_1과 T_2의 모든 노드이며 indicator 함수 $I_{St_i}(n_1)$는 노드 n_1이 최상위 노드인 서브트리 S_{t_i}가 있으면 1, 아니면 0을 반환한다. $C(n_1, n_2)$ 함수는 다음과 같이 정의 가능하다.

$$C(n_1, n_2) = \sum_i I_{St_i}(n_1) \cdot I_{St_i}(n_2)$$

함수 $C(n_1, n_2)$는 다음과 같은 재귀적인 규칙을 이용하여 polynomial time에 계산 가능하다.

규칙1. n_1과 n_2의 product가 다르면
$C(n_1, n_2) = 0$
규칙2. n_1과 n_2의 product가 같고 pre-terminal이라면
$C(n_1, n_2) = \lambda$
규칙3. 그 외
$$C(n_1, n_2) = \lambda \prod_i^{nc(n_1)} (1 + C(ch(n_1, i), ch(n_2, i)))$$

함수 $nc(n_1)$는 n_1의 자식노드의 수를 반환하며, $ch(n, i)$는 노드 n의 i번째 자식 노드를 의미한다. λ는 큰 부분 트리의 영향을 줄이는 decay factor이다.

위의 재귀 규칙에서 노드 n_1의 product란, n_1의 자식 노드에 나타난 레이블의 순열을 의미한다. 위의 재귀 규칙을 이용하여 계산한 커널 함수 $K(T_1, T_2)$는 입력으로 주어진 트리의 크기에 비례해 높은 유사도 값을 반환한다. 따라서 이를 정규화 할 필요가 있다. 파스 트리 커널도 일반적인 커널과 마찬가지로 아래와 같은 수식을 이용하여 정규화 한다.

$$K'(T_1, T_2) = \frac{K(T_1, T_2)}{\sqrt{K(T_1, T_1) \cdot K(T_2, T_2)}}$$

3.3. 구조 정보를 반영하기 위한 결합 커널

본 논문에서는 하나의 웹 테이블에 대해 두 가지 구문 분석 트리를 추출하여 이를 각각 파스 트리 커널에 적용하고 있다. 추출된 컨텍스트 트리와 테이블 트리는 테이블의 서로 다른 특성을 반영하고 있다. 따라서 이들 정보를 통합하여 테이블에 대한 전체 구조 정보를 반영할 필요가 있다. 본 논문에서는 두 테이블에서 추출된 트리들을 비교하기 위해 커널 함수를 사용하고 있다. 따라서 컨텍스트 트리와 테이블 트리의 통합은 두 트리가 적용된 파스 트리 커널을 이용한 결합 커널을 정의함으로써 쉽게 이루어 질 수 있다. 결합 커널은 아래와 같이 정의 될 수 있다.

$$K(x_i, x_j) = \alpha \cdot PTK(ct_i, ct_j) + (1-\alpha) \cdot PTK(tt_i, tt_j) \quad (1)$$

표 1. 실험 테이터의 통계
Table 1. Simple Statistics of Experimental Data

항목	개수
웹 문서	1,393
전체 테이블	14,609
리프 테이블	11,477
의미 있는 테이블	1,740
장식 테이블	9,737

수식 (1)에서 $PTK(ct_i, ct_j)$는 컨텍스트 트리가 적용된 파스 트리 커널을 $PTK(tt_i, tt_j)$는 테이블 트리가 적용된 파스 트리 커널을 의미한다. α는 두 커널을 결합할 때, 두 커널 간 중요도를 결정하기 위한 결합 파라미터이다.

4. 실험

4.1. 실험 테이터

실험에서는 Wang et al. [5]이 구축한 테이블 데이터를 사용하였다. 테이블 데이터는 Google 검색 엔진을 이용하여 비즈니스, 뉴스, 과학 카테고리에서 표, 주식, 증권, 그림, 시간표 등의 검색어를 입력하여 얻어진 1,393개의 웹 문서로 이루어져 있으며, 서로 다른 200여개의 웹 사이트에서 수집되었다. 전체 테이블의 수는 14,609개이며 이중, 다른 테이블을 내포하지 않은 11,477개의 리프 웹 테이블 (leaf web tab)이 실험에 사용되었다. 리프 웹 테이블 만 사용한 이유는 다른 테이블을 내포한 테이블은 실험 데이터에서 모두 장식 테이블이기 때문이다. 표 1은 실험 데이터의 간단한 통계를 보여준다.

전체 리프 테이블 중, 80%는 학습을 위해 사용되었으며 10%는 파라미터 추정을 위한 검증 데이터로, 나머지 10%는 테스트를 위해 사용되었다. 모든 실험에서는 5겹 교차 검증 (5-fold cross validation)을 사용하여 성능을 측정하였다. 검증 데이터를 통한 실험 결과 파스 트리 커널의 decay factor는 0.4로 설정하였다 성능은 정확율, 재현율, F-measure를 사용하여 평가하였다.

4.2. 실험 결과

먼저 컨텍스트 트리와 테이블 트리를 각각 적용한 파스 트리 커널의 성능을 비교하였다. 표 2는 두 커널의 실험 결과를 보여준다. 표에서 보여 주듯이 컨텍스트 트리의 경우 정확율이 높지만 재현율이 낮음을 알 수 있다. 이는 테이블 트리와 달리 컨텍스트 트리는 다양한 형태를 가질 수 있으며 그 크기 또한 일정하지 않다. 이로 인해 서브 트리로 부터 생성된 자질 공간의 차원이 너무 높아 테이터의 벡터가 sparse해지는 경향이 있다. 결과적으로 데이터의 sparseness가 재현율을 떨어뜨린다. 반대로 테이블 트리는 정확율이 컨텍스트 트리와 비교해 7% 정도 낮지만, 재현율은 28%정도 높게 나타난다. 이는 테이블 트리의 일정한 형태로 인해 컨텍스트 트리에 비해 자질 공간의 차원 수가 낮기 때문이다.

컨텍스트 트리와 테이블 트리는 서로 다른 측면의 정보를 담고 있다. 이러한 서로 다른 특성은 성능 면에서도 결과적으로 반대되는 특성을 보였다. 컨텍스트 트리는 더 정

표 2. 컨텍스트 트리와 테이블 트리의 성능 비교
Table 2. Comparison between the performances of parse tree kernel with context trees and one with table trees

	PTK_{ct}	PTK_{tt}
정확율	97.4	90.5
재현율	57.5	85.6
F-measure	77.5	88.0

표 3. 결합 커널과 레이아웃 자질간 성능 비교
Table 3. Performance comparison between the composite kernel and the layour features

	$CK(\alpha = 0.7)$	LF
정확율	80.0	92.6
재현율	97.36	75.8
F-measure	88.68	84.2

확하지만 재현율이 낮았으며, 테이블 트리는 재현율이 높았지만 정확율이 떨어졌다. 본 논문의 결합 커널은 서로 다른 특성을 보이는 두 트리 커널을 결합하여 전체적인 성능을 높이고자 하였다.

그림 4는 수식 (1)의 결합 파라미터 α값에 따른 결합 커널의 성능을 보여준다. 그림에서 $\alpha = 0$은 테이블 트리만을 사용한 경우이며, $\alpha = 1.0$은 컨텍스트 트리만을 사용한 경우이다. 가장 좋은 성능은 F-measure, 88.68이며 $\alpha = 0.7$에서 얻어진다. 이는 테이블 트리보다 높은 정확율을 가지는 컨텍스트 트리의 반영 비율이 더 높아야 함을 의미한다.

본 논문에서 제안한 자질들은 사람의 손을 빌리지 않고 생성되며, 특히 컨텍스트 트리로부터 얻는 구조 정보는 테이블 주변에 나타나는 컨텐츠가 다양하기 때문에 사람이 정의한 자질로는 얻기 힘들다. 즉, 테이블 트리와 컨텍스트 트리를 고려한 구조 자질은 사람이 정의하는 구조 자질에 비해 더 나은 성능을 가질 수 있다. 이를 증명하기 위해 기존 연구에서 제안한 구조 자질과의 성능 비교 실험을 하였다. 비교한 구조 자질은 기존 연구 중, 가장 높은 성능을 보이는 Wang et al. [5]의 연구에서 제안된 것으로 셀 길이의 평균과 분산, 셀 데이터 길이의 평균 등으로 이루어져 있다. 표 3은 제안한 결합 커널의 성능과 wang et al.이 제안한 구조 자질인 레이아웃 자질의 성능을 보여준다. 표에서 CK는 제안한 결합 커널을 의미하며, LF는 레이아웃 자질이다.

표에서 보여주듯이 레이아웃 자질은 정확율에서 더 높은 성능을 보였으나 재현율에 결합 커널에 비해 20% 이

상 낮은 성능을 보였다. 이는 사람이 직접 정의한 구조 자질의 경우, 중요한 정보를 놓칠 수 있다는 본 논문의 주장과 일치한다. 즉, 사람이 정의하기 어려운 구조 정보를 레이아웃 자질이 반영하지 못하기 때문에 재현율이 떨어지는 것이다. 반면, 결합 커널은 모든 가능한 서브 트리를 자질로 이용하는 파스 트리 커널에 기반하고 있기때문에 정확율이 레이아웃 자질에 비해 10% 가량 떨어지나, 더 높은 성능의재현율을 보일 수 있다. 결과적으로 전체적인 성능을 의미하는 F-measure에서 결합 커널이 4% 이상의 향상을 보였다.

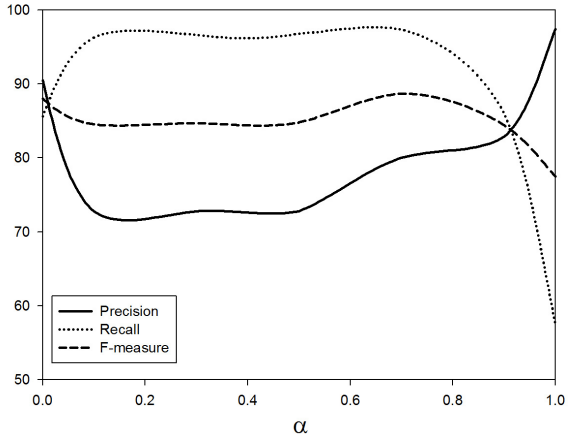

그림 4. α값에 따른 결합 커널의 성능
Figure 4. The performance of the composite kernel with various values of α

표 4. 테이블 분류 성능 비교
Table 4. Comparison of table discrimination performances

	CK+CF	RBF	DT	HR
정확율	98.20	95.81	97.50	75.70
재현율	98.96	95.98	94.25	48.16
F-measure	98.58	95.89	95.88	61.93

마지막으로 제안한 구조 자질이 기존의 레이아웃 자질을 대체 했을 때, 전체적인 테이블 분류 성능이 얼마나 향상될 수 있는지 확인하기 위한 실험을 하였다. 실험에서는 Wang et al.이 제안한 컨텐츠 자질과 본 논문에서 제안한 결합 커널 기반의 구조 자질을 결합하여 성능을 측정 하였다. 표 4는 실험 결과를 보여준다. 표에서 CK+CF는 제안한 결합 커널과 컨텐츠 자질을 합친 방법을 의미하며, HR은 Penn et al. [2]이 제안한 규칙 기반의 방법을, RBF와 DT는 Wang et al. [5]이 제안한 RBF (radial basis function) 기반의 SVMs과 결정 트리 모델을 각각 의미한다.

표에서 보여주듯이, 정확도, 재현율, F-measure 모두 재안한 결합 커널에 기반한 방법이 기존의 방법보다 더 나음을 알 수 있으며, 그 차이 또한 약 3%로 작지 않다. 이러한 결과는 제안한 결합 커널이 컨텐츠 정보와 같이 구조 정보 이외의 부과적인 정보를 활용하여 성능을 높이지 않았음을 의미하며, 결과적으로 제안한 구조 자질이 웹 테이블의 구조 정보를 잘 반영하고 있음을 보여준다.

5. 결 론

본 논문은 웹 페이지에 나타나는 테이블 중, 장식 테이블과 의미 있는 테이블을 분류하기 위한 새로운 구조 자질을 정의하였다. 제안한 구조 자질은 테이블의 구조 정보를 반영하기 위해 HTML 문서의 구문 분석 트리로부터 테이블의 구조 정보를 가지는 테이블 트리와 컨텍스트 트리를 기반으로 정의된다. 추출된 테이블 트리와 컨텍스트 트리는 파스 트리 커널을 이용하여 비교되어진다. 테이블의 구조

정보를 반영하기 위해 두 트리에 나타난 구조 정보는 결합 커널을 이용하여 분류 모델에 모두 반영된다.

실험 결과, 결합 커널은 이들 테이블 트리와 컨텍스트 트리로 표현되는 테이블의 구조 정보를 활용하여 F-measure 측면에서 88.68의 높은 성능을 보였다. 이는 레이아웃 자질에 비해 4% 이상 향상된 성능이다. 결합 커널과 기존의 컨텐츠를 결합하여 측정한 테이블 분류 성능에서는 기존의 연구에 비해 더 나은 성능을 보임으로써 제안한 결합 커널이 테이블의 구조 정보를 더 잘 반영함을 보였다.

참 고 문 헌

[1] S. Jung, K. Sung, T. Park, and H. Kwon, "Effective Retrieval of Information in Tables on the Internet," In *Proceedings of IEA/AIE'02*, pp. 493-501, 2002.

[2] G. Penn, J. Hu, H. Luo, and R. McDonald, "Flexible Web Document Analysis for Delivery to Narrow-bandwidth Devices," In *Proceedings of ICDAR'06*, pp. 119-130, 2004.

[3] Y. Zhai and B. Liu, "Web Data Extraction based on Partial Tree Alignment," In *Proceedings of the WWW'05*, pp. 76-85, 2005.

[4] H. Chen, S. Tsai, and J. Tsai, "Mining Tables from Large Scale HTML texts," In *Proceedings of the 18th International Conference Computational Linguistics*, pp. 166-182, 2007.

[5] Y. Wang and J. Hu, "A Machine Learning based Approach for Table Detection on the Web," In *Proceedings of WWW'02*, pp. 242-250, 2002.

[6] E. Crestan and P. Pantel, "A Fine-Grained Taxonomy of Tables on the Web," In *Proceedings of the 19th ACM International Conference on Information and Knowledge management*, pp. 1405-1408, 2010.

[7] S. Jung and H. Kwon, "A Scalable Hybrid Approach for Extracting Head Components from Web Tables", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 2, pp. 174-187, 2006.

[8] Y. Liu, K. Bai, P. Mitra, and C. Giles, "Automatic Searching of Tables in Digital Libraries," In *Proceedings of the 16th International Conference on World Wide Web*, pp. 1135-1136, 2007.

[9] E. Crestan and P. Pantel, "Web-scale Knowledge Extraction from Semi-structured Tables," In *Proceedings of the 19th International Conference on World Wide Web*, pp 1081-1082, 2010.

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

[11] D. Haussler, "Convolution Kernels on Discrete Structures," Technical report, UCS-CRL-99-10, UC Santa Cruz, 1999.

[12] M. Collins and N. Duffy, "Convolution Kernels for Natural Language," In *Advances in Neural Information Processing Systems* 14, pp. 625-632, 2001.

[13] M. Hurst, "Layout and language: Challenges for table understanding on the web," In *Proceedings of WDA'01*, pp. 27-30, 2001.

저 자 소 개

손정우(Jeong-Woo Son)
2005년 : 경북대학교 컴퓨터공학과 (학사)
2007년 : 경북대학교 컴퓨터공학과 (석사)
2007년~현재 : 동대학원 전자전기컴퓨터공학부 박사과정

관심분야 : 기계학습, 자연어처리, 온톨리지
E-mail : jwson@sejong.knu.ac.kr

고준호(Junho Go)
2011년: 경북대학교 컴퓨터과학과 (학사)
2011년~현재: 경북대학교 IT대학 석사과정

관심분야 : 정보 검색, 자연어처리
E-mail : jhgo@sejong.knu.ac.kr

박성배(Seong-Bae Park)
1996년 : 서울대학교 컴퓨터공학과 (석사)
2002년 : 서울대학교 컴퓨터공학과 (박사)
2004년~현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 기계학습, 자연어처리, 텍스트 마이닝
E-mail : sbpark@sejong.knu.ac.kr

김권양(Kweon Yang Kim)
1990년 : 경북대학교 전자공학과(석사)
1998년 : 경북대학교 컴퓨터공학과(박사)
1983~1988년 : ETRI 연구원
1999년~2000년 : University of Central Florida 방문교수
1991년~현재 : 경일대학교 컴퓨터공학과 교수

관심분야 : 시멘틱웹, 한글공학
E-mail : kykim@kiu.ac.kr