Journal of Korea Multimedia Society Vol. 7, No. 10, October 2004(pp. 1339-1353)

HTML 문서의 테이블 식별을 위한 효율적인 알고리즘

김연석[†], 이경호[††]

요 약

HTML의 table 태그는 연관된 정보를 기술하기 위한 테이블은 물론이고 웹 문서의 레이아웃을 표현하기 위하여 사용된다. 본 논문에서는 웹으로부터 유용한 정보를 추출하기 위한 목적의 일환으로 HTML 문서로부터 테이블을 식별하는 효율적인 방법을 제안한다. 제안된 방법은 전처리와 속성-값 연관관계 추출의 두 단계로 구성된다. 전처리 단계에서는 진짜 테이블 또는 레이아웃용으로 사용된 table 태그의 일반적인 특징을 반영한 규칙을 적용하여 진짜 또는 가짜로 명확히 식별이 가능한 table 태그를 추출한다. 속성-값 연관관계 추출 단계에서는 테이블 영역을 속성 및 값 영역으로 구분한 후, 값 영역에 대하여 구문적 일관성 검사를 수행한다. 또한 값 영역의 크기가 작아서 구문적 일관성 검사를 수행할 수 없는 경우, 속성-값 영역의 의미적 일관성을 검사한다. 제안된 방법의 성능을 평가하기 위하여 1,393개의 HTML 문서로부터 추출한 11,477개의 table 태그를 대상으로 실험한 결과, 평균적으로 97.54%의 정확률과 99.22%의 재현률을 보여 기존 연구보다 우수하였다.

An Efficient Algorithm for Detecting Tables in HTML Documents

Yeon-Seok Kim[†], Kyong-Ho Lee[††]

ABSTRACT

<TABLE> tags in HTML documents are widely used for formatting layout of Web documents as well as for describing genuine tables with relational information. As a prerequisite for information extraction from the Web, this paper presents an efficient method for sophisticated table detection. The proposed method consists of two phases: preprocessing and attribute-value relations extraction. For the preprocessing where genuine or ungenuine tables are filtered out, appropriate rules are devised based on a careful examination of general characteristics of <TABLE> tags. The remaining is detected at the attribute-value relations extraction phase. Specifically, a value area is extracted and checked out whether there is a syntactic coherency. Futhermore, the method looks for a semantic coherency between an attribute area and a value area of a table that may be inappropriate for the syntactic coherency checkup. Experimental results with 11,477 <TABLE> tags from 1,393 HTML documents show that the method has performed better compared with previous works, resulting in a precision of 97.54% and a recall of 99.22% in average.

Key words: Table Detection(테이블 식별), HTML Document(HTML 문서), Genuine Table(진짜 테이블), Attribute-Value Relations(속성-값 연관관계), Information Extraction(정보 추출)

※ 교신저자(Corresponding Author) : 김연석, 주소 : 서울시 서대문구 신촌동 134(120-749), 전화 : 02)2123-3878, FAX : 02)365-2579, E-mail : yskim@icl.yonsei.ac.kr
접수일 : 2004년 1월 19일, 완료일 : 2004년 4월 19일

[†] 연세대학교 대학원 컴퓨터과학과 석사과정
[††] 정회원, 연세대학교 컴퓨터산업공학부 조교수
(E-mail : khlee@cs.yonsei.ac.kr)

1. 서 론

최근 들어 웹이 일상생활에서 보편적으로 사용됨에 따라 웹 문서의 양이 급증하고 있다. 한편 HTML (Hypertext Markup Language)[1]은 웹 문서를 사용자에게 보이기 위한, 즉 시각적으로 렌더링하기 위한 포맷이기 때문에 컴퓨터로 하여금 정보를 처리하게 한다는 측면에서 한계를 갖는다. 따라서 HTML 문서로부터 유용한 정보를 추출하는 방법에 대한 관심이 고조되고 있다[2].

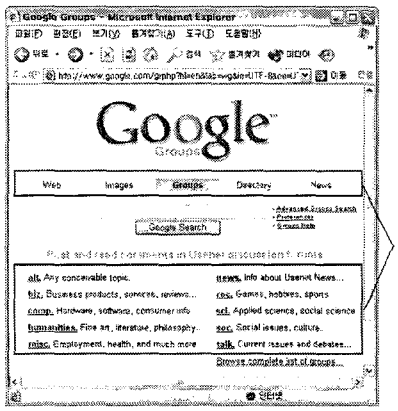
HTML은 테이블의 표현을 위하여 table 태그를 정의하는데, 그림 1(a)는 HTML 문서에 포함된 진짜 테이블의 예이다. 또한 기존 HTML 문서의 상당수는 문서의 레이아웃을 보기 좋게 표현하기 위하여 table 태그를 사용하였다(그림 1(b) 참조). 따라서 HTML 문서로부터 유용한 정보를 추출하기 위해서는 먼저 table 태그가 진짜 테이블을 표현하기 위하여 사용되었는지의 여부를 판별할 필요가 있다[3].

일반적으로 테이블은 연관된 정보(relational information)를 구조적이며 간결하게 표현할 수 있는 방법이다. 본 논문에서는 테이블을 연관성을 갖는 데이터의 배열이라고 정의한다. 따라서 제안된 방법은 [8]과 마찬가지로 속성(attribute)과 값(value)의 관계를 포함하는 테이블을 진짜 테이블(genuine table)로 간주한다. 그림 2는 테이블을 구성하는 속성과 값의 예이다. 특히 본 논문에서는 속성을 나타내는 셀들의 집합을 속성 영역(attribute area), 그리고 값을 나타내는 셀들의 집합을 값 영역(value area)이라고 정의한다.

기존에 HTML 문서로부터 테이블을 식별하기 위한 연구가 진행되었다[4-9]. 그러나 기존 연구의 대부분은 특정 도메인에 의존적이거나 테이블 식별을 위해서 다수의 학습 데이터 및 학습 시간을 요구한다. 본 논문에서는 웹 문서로부터 정보를 추출하기 위한 목적의 일환으로 HTML 문서에 포함된 테이블을 식별하는 효율적인 방법을 제안한다. 제안된 방법은 전처리 그리고 속성-값 연관관계 추출의 두 단계로 구성된다. 전처리 단계에서는 진짜 또는 가짜로

지수명	현재가	대비
KOSPI	847.95 ▲	2.29
KOSPI 200	110.63 ▲	0.29
KODI	1,594.77 ▲	0.73
KOGI	1,404.50 ▲	1.69
KOSPI 100	835.56 ▲	2.28
KOSPI 50	799.20 ▲	1.58
KOSPI IT	716.95 ▲	5.50
KOSDAQ	44.54 ▼	0.46
DOW 30	10,600.51 ▲	45.65
NASDAQ	2,140.45 ▲	31.38
S&P 500	1,139.93 ▲	7.78

(a) table태그를 사용하여 표현한 진짜 테이블



(b) table태그를 사용하여 문서의 레이아웃을 표현한 예

그림 1. HTML table 태그를 사용하여 각각 진짜 테이블과 레이아웃을 표현한 예

코 드	상 품 명	출 발 일
bdm-650	제주도 1박2일	매일
bdm-651	제주도 2박3일(3일관광)	매일
bdm-652	제주 TOUR 2박 3일(관광1급,특급호텔)	월,화,수,목,금,토
bdm-653	제주 TOUR 2박 3일(귀빈관광호텔)	월,화,수,목,토
bdm-654	색동 투어(관광1급,특급호텔)	월,화,수,목,금,토
bdm-655	모두랑 투어(크라운프라자특급호텔)	수
bdm-656	자유 여행(렌트카)	월,화,수,목,금,토
bdm-657	펜션 자유 여행	월,화,수,목,금,토

속성영역

값 영역

그림 2. 테이블을 구성하는 속성 영역과 값 영역의 예

명확히 식별이 가능한 table 태그를 추출한다. 이를 위하여 진짜 테이블 또는 레이아웃용으로 사용된 table 태그의 일반적인 특징을 반영한 8개의 규칙을 제안한다.

연관관계 추출 단계에서는 전처리 과정에서 식별되지 않은 테이블을 대상으로 테이블 영역을 속성 및 값 영역으로 구분한 후, 값 영역에 대하여 구문적 일관성(syntactic coherency) 검사를 수행한다. 한편 구문적 일관성이 존재하지 않거나 값 영역의 크기가 작아서 구문적 일관성 검사를 적용할 수 없는 경우, 속성-값 영역에 대하여 속성과 값의 의미적 일관성(semantic coherency) 검사를 수행한다. 제안된 방법의 성능을 평가하기 위하여 1,393개의 HTML 문서로부터 추출한 11,477개의 table 태그를 대상으로 실험한 결과 평균 97.54%의 정확률과 99.22%의 재현률을 보여 기존 연구보다 우수하였다.

본 논문의 구성은 다음과 같다. 2절에서는 관련연구를 통하여 HTML 문서로부터 테이블을 식별하는 기존의 연구를 간략히 기술하고, 3절에서는 제안된 테이블 식별방법을 전처리와 연관관계 추출의 두 단계로 구분한 후, 각각에 대한 자세한 설명을 기술한다. 4절에서는 실험 결과를 통하여 제안된 방법의 성능을 기존 연구와 비교 및 분석하며, 마지막으로 5절에서는 결론 및 향후 연구방향을 기술한다.

2. 관련 연구

HTML 문서의 테이블 식별에 관한 연구는 크게 특정 도메인에 의존적인 방법과 도메인에 독립적인 방법의 두 가지로 나누어진다[8]. 도메인에 의존적인 방법은 특정 도메인에 대한 정보를 이용하여 테이블을 식별하는 방법이며, 도메인 독립적인 방법은 임의의 HTML 문서에 포함된 테이블을 도메인 정보에 상관없이 식별하는 방법이다. 표 1은 HTML 문서로부터 테이블을 식별하는 기존 연구의 특징을 간략히 기술한 것이다.

Chen 등[4]은 야후 (yahoo) 웹 사이트로부터 수집된 여행관련 웹 페이지를 대상으로 테이블을 식별 및 인식하는 방법을 제안한다. 본 방법은 필터링 모듈과 테이블 식별 모듈을 적용하여 진짜 테이블을 식별하는데, 필터링 모듈은 규칙에 기반하며 식별 모듈은 문자열 유사도(string similarity), 개체 유사도(named entity similarity), 그리고 수 카테고리 유사도 (number category similarity)의 세 가지 유사도를 사용한다. 한편 제안된 방법은 특정 도메인(중국항공정보)에 의존적이므로 다양한 도메인에 적용하는데 한계를 갖는다.

Penn 등[5]은 단말 테이블(leaf table) 여부, 다중 행(multi-row) 또는 다중 열(multi-column)의 여부, 셀이 가질 수 없는 태그의 존재여부, 그리고 제한된 셀 길이 등 네 개의 규칙을 적용하여 테이블을 식별한다. 그러나 진짜 테이블의 특성인 구문적 및 의미적 일관성을 고려하지 않으며 셀 길이에 대한 제한을 두기 때문에 다양한 종류의 테이블을 식별하는데 한계를 갖는다.

Yoshida 등[6]은 사용된 어휘와 EM(Expectation-Maximization) 알고리즘[10]에 기반하여 테이블의 구조를 미리 정의된 9가지 중 하나로 식별한다. 그러나 제안된 방법은 테이블의 식별을 위하여 적절한 온톨로지 정보를 요구한다.

표 1. HTML 문서로부터 테이블 식별을 위한 방법

저자	연도	특 징
Chen 등[4]	2000	규칙에 기반한 테이블 필터링 모듈을 적용하여 가짜 테이블을 제거한 후 문자열, 개체, 그리고 수 카테고리 유사도를 이용하여 테이블을 식별한다.
Penn 등[5]	2001	네 개의 휴리스틱 규칙을 적용하여 테이블을 식별한다.
Yoshida 등[6]	2001	속성 및 값 영역에 사용된 어휘정보에 기반하여 테이블의 구조를 9가지로 분류한다.
Hurst[7]	2002	테이블 식별을 위하여 5개의 DOM 특징과 3개의 모델 특징을 다양한 분류기에 적용한 후, 가장 좋은 결과를 보인 분류기의 결과를 분석하였다.
Yang[8]	2002	두 개의 규칙을 적용하여 얻어진 테이블 후보 집합에 대하여 테이블의 속성과 값이 포함할 수 있는 문자열의 패턴 정보를 적용하여 진짜 테이블을 식별한다.
Wang과 Hu [9]	2002	테이블 식별을 위하여 7개의 레이아웃 특징, 8개의 컨텐츠 타입 특징, 그리고 1개의 워드 그룹 특징에 기반한 기계학습 기법을 제안한다.

Hurst[7]는 행(row), 열(column), 테두리(border) 등을 포함한 5개의 HTML DOM (document object model) 특징과 문자열 컨텐츠의 비율(string content ratio), 단일 셀의 비율(singular cell ratio), 그리고 행/열의 카테고리등 3개의 모델(model) 특징을 제안한다. 특히 테이블의 식별을 위하여 두 종류의 특징을 추출하고, 이를 나이브 베이스 분류기(naive bayes classifier)[11], 결정트리(decision tree)[11], Winnow 분류기[12], 그리고 최대 엔트로피 분류기(maximum entropy classifier)[13]에 적용하였으며 가장 우수한 성능을 보인 두 분류기의 결과를 비교하였다.

Yang[8]은 두 개의 규칙에 기반한 필터링 과정을 적용하여 추출된 후보 테이블 집합에 미리 정의된 AIEP(attribute indicating entity pattern)와 VIEP (value indicating entity pattern)를 적용하여 테이블을 식별한다. AIEP와 VIEP는 각각 속성(attribute)과 해당 속성의 값(value)으로 올 수 있는 문자열의 패턴을 의미한다. 따라서 제안된 방법은 AIEP와 VIEP에 포함된 정보의 양에 크게 의존적이다.

Wang과 Hu[9]는 결정트리에 기반한 테이블 식별 방법을 제안한다. 제안된 방법은 레이아웃 특징 7개, 컨텐츠타입 특징 8개, 워드그룹 특징 1개 등 총 16개의 특징을 이용하여 시스템을 학습한다. 비즈니스, 뉴스, 그리고 과학 분야에 속하는 1,393개의 HTML 파일에 포함된 11,477개의 table 태그를 대상으로 실험한 결과, 높은 정확률과 재현률을 보였다. 그러나 학습 데이터로 비교적 많은 수의 특징을 사용하기 때문에 시스템을 학습시키는데 많은 비용이 소요된다.

3. 제안된 테이블 식별 방법

본 절에서는 제안된 테이블 식별 방법을 기술한다. 제안된 방법은 그림 3과 같이 전처리와 속성-값 연관관계 추출의 두 단계로 구성된다. 특히 연관관계 추출은 속성 및 값 영역의 추출, 값 영역의 구문적 일관성 검사, 그리고 속성-값의 의미적 일관성 검사의 세 부분으로 이루어진다.

HTML 문서로부터 추출된 table 태그에 8개의 규칙을 적용하여 일차적으로 테이블을 식별한 다음, 식별하지 못한 테이블에 대하여 영역 추출 단계에서 테이블을 속성 영역과 값 영역으로 구분한다. 이렇게 구분된 값 영역에 대하여 세로 또는 가로 방향으로 구문적 일관성이 존재하는지의 여부를 검사한다. 특히 제안된 방법은 값 영역의 데이터 타입(data type)과 길이(length) 정보를 기반으로 일관성을 검사한다. 구문적 일관성 검사를 통하여 식별이 어려운 경우 의미적 일관성 검사를 통하여 값 영역에 포함된 내용이 해당 속성과 일관성을 갖는지를 판별한다. 이를 위하여 값 영역의 데이터 타입과 특정 속성의 값

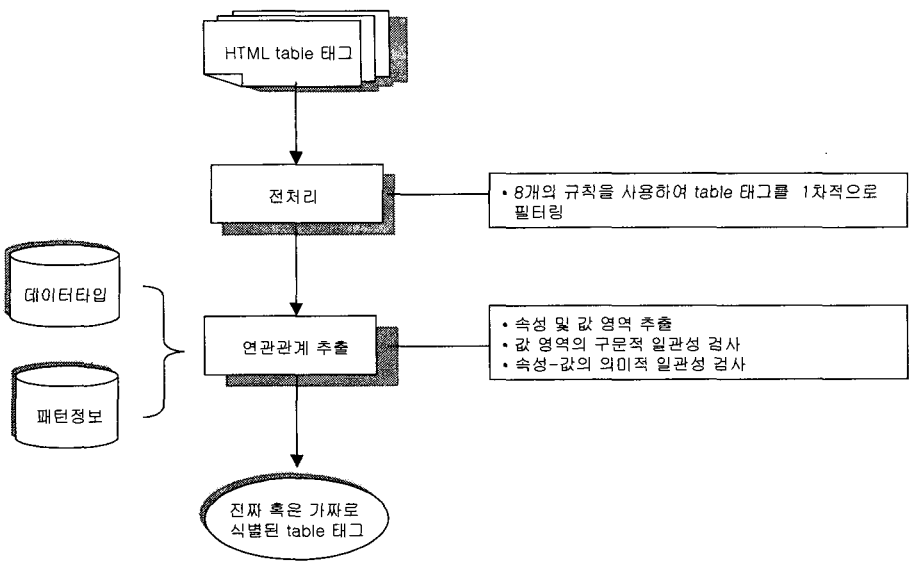


그림 3. 제안된 테이블 식별 방법

으로 올 수 있는 키워드 또는 패턴 등의 부가정보를 이용한다. 각 단계에 대한 자세한 설명은 다음과 같다.

3.1 전처리

제안된 전처리 과정은 진짜 또는 가짜 테이블이 갖는 일반적인 특징을 기준으로 테이블을 식별한다. 이를 위하여 그림 4와 같이 8개의 규칙을 적용한다.

3.2 연관관계 추출

연관관계 추출 단계는 전처리 단계에서 가짜 혹은 진짜 테이블로 판별되지 않은 table 태그를 대상으로 한다. 제안된 연관관계 추출 방법은 영역 추출, 값 영역의 구문적 일관성 검사, 그리고 속성-값 영역의 의미적 일관성 검사의 세 부분으로 이루어진다. 제안된 방법은 연관관계 추출에 앞서 테이블을 속성과 값 영역으로 구분한다. 일반적으로 속성에 대응하는 값이 2개 이상 존재할 경우, 해당 값들은 일관된 데이터 타입을 갖는다. 한편 각 속성이 갖는 값이 유일하여 구문적 일관성을 검사할 수 없는 경우, 해당 속성과 값 사이의 의미적 일관성을 검사한다. 제안된 연관관계 추출 알고리즘은 그림 5와 같으며 이에 대한 자세한 설명은 다음과 같다.

3.2.1 영역 추출

제안된 방법은 연관관계의 추출을 위하여 먼저 테이블 영역을 속성과 값 영역으로 구분한다. 그림 6과 같이 크기가 1×2(그리고 1×n) 또는 2×1(n×1)인 테이블의 경우, 각각 첫 번째 열(column)과 행(row)이 속성 영역(음영으로 표시된 부분)에 해당한다. 한편 크기가 2×2이며 스팬(span) 속성을 포함하지 않는 경우, 첫 번째 행 또는 열이 속성 영역에 해당될 수 있으며, 만일 첫 번째 행(또는 열)을 속성 영역으로 간주하면 두 번째 행(또는 열)이 값 영역에 해당한다.

특히 크기가 1×2, 2×1, 그리고 2×2인 테이블의 경

전처리 규칙 (1) : 테이블의 제목을 나타내는 <caption> 태그의 존재여부
IF: <caption> 태그가 존재한다.
THEN: table 태그를 진짜 테이블로 식별한다.

전처리 규칙 (2) : 일반적으로 박스(box)를 표현하기 위하여 사용되는 크기가 1×1인 테이블
IF: 테이블의 크기가 1×1이다.
THEN: table 태그를 가짜 테이블로 식별한다.

전처리 규칙 (3) : 문자의 존재여부
IF: 테이블에 문자가 존재하지 않는다.
THEN: table 태그를 가짜 테이블로 식별한다.

전처리 규칙 (4) : 레이아웃 관점에서 하이퍼링크를 적절히 배열하기 위하여 테이블 사용
IF: 테이블을 구성하는 대부분의 셀들이 하이퍼링크만을 포함한다.
THEN: table 태그를 가짜 테이블로 식별한다.

전처리 규칙 (5) : 대부분의 셀들이 이미지로 구성
IF: 테이블을 구성하는 셀들이 대부분 이미지로 구성된다.
THEN: table 태그를 가짜 테이블로 식별한다.

전처리 규칙 (6) : 대부분의 셀들이 공백으로 구성
IF: 테이블을 구성하는 셀들이 대부분 공백으로 구성된다.
THEN: table 태그를 가짜 테이블로 식별한다.

전처리 규칙 (7) : 일반적으로 <th> 태그는 테이블의 헤더를 표현하기 위하여 사용한다. 따라서 본 논문에서는 <th> 태그와 이에 대응하는 데이터 셀(<td>)이 존재할 경우 해당 테이블을 진짜 테이블로 간주한다.
IF: (1) 테이블이 <th> 태그를 포함한다.
(2) <th> 태그에 오른쪽 또는 아래 방향에 <td> 태그가 존재한다.
THEN: table 태그를 진짜 테이블로 식별한다.

전처리 규칙 (8) : 일반적으로 레이아웃을 표현하기 위하여 중첩된 테이블을 사용한다.
IF: 해당 table 태그가 중첩된 table 태그를 포함한다.
THEN: 해당 table 태그를 가짜 테이블로 식별한다.

그림 4. 제안된 전처리 규칙

```
입력: HTML table 태그
출력: 테이블의 진위 여부 (IsGenuineTable)
함수 및 변수 정의:
    Boolean IsGenuineTable                ::= 테이블 진짜 또는 가짜인지 여부 저장
    Boolean IsThereSemanticCoherency()    ::= 속성-값 간의 의미적 일관성이 존재하면 TRUE,
                                              그렇지 않으면 FALSE를 반환한다.
    Boolean IsThereSyntacticCoherency()   ::= 값 영역에 구문적 일관성이 존재하면 TRUE,
                                              그렇지 않으면 FALSE를 반환한다.
방법:
1:    If (테이블의 크기가 1×2, 2×1, 또는 2×2인 경우)
2: // 속성이 단일의 값을 가지기 때문에 속성-값 간의 의미적 일관성을 검사
3:     IsGenuineTable = IsThereSemanticCoherency();
4: // 속성이 2개 이상의 값을 갖는 경우, 먼저 값 사이의 구문적 일관성을 검사
5: // 또한 구문적 일관성 검사를 통하여 판별할 수 없는 경우를 위해서 의미적 일관성을 검사
6:    else if (테이블의 크기가 1×n 또는 n×1인 경우) { // 여기서 n>2
7:     If(IsThereSyntacticCoherency())
8:       IsGenuineTable=True;
9:     else
10:      IsGenuineTable = IsThereSemanticCoherency();
11:   }
12:   else if (2×n || n×2) { // 여기서 n≧3
13:     if (테이블이 값 영역을 포함하지 않는 경우)
14:       IsGenuineTable = False;
15:     else {
16:       if(IsThereSyntacticCoherency()) IsGenuineTable=True;
17:       else IsGenuineTable = IsThereSemanticCoherency();
18:     }
19:   }
20:   else { // 크기가 3×3 이상인 경우
21:     if (테이블이 값 영역을 포함하지 않는 경우)
22:       IsGenuineTable = False;
23:     if (값 영역이 2개 이상의 row 또는 column으로 이루어진 경우) {
24:       if(IsThereSyntacticCoherency())
25:           IsGenuineTable=True;
26:       else
27:           IsGenuineTable = IsThereSemanticCoherency();
28:     }
29:     else // 값 영역이 1개의 row 또는 column으로 이루어진 경우
30:       IsGenuineTable = IsThereSemanticCoherency();
31:   }
```

그림 5. 제안된 연관관계 추출 알고리즘

우, 각각의 속성에 대응하는 값 영역이 유일하여 값 영역의 구문적 일관성을 검사할 수 없기 때문에 해당 속성과 값 영역 간의 의미적 일관성을 검사한다. 물론 크기가 2×2이며 스팬 속성을 포함하지 않는 경우, 가로 방향(row-wise) 또는 세로 방향(column-wise)으로 의미적 일관성이 존재하는지의 여부를 검사한다. 한편 1×n(또는 n×1) 테이블의 경우, 가로 방향(또는 세로 방향)으로 구문적 일관성을 검사한다.

한편, 크기가 2×n 또는 n×2 (n≧2)인 테이블의 경우 값 영역을 포함하지 않을 수 있다. 예를 들어 그림 7과 같이 첫 번째 행 또는 열이 하나의 셀(cell)로 이루어진 경우 값 영역을 포함하지 않는 것으로 간주하여 가짜 테이블로 식별한다. 즉, 그림 7(a)에서 단일의 셀을 포함하는 첫 번째 행을 주 속성(main attribute) 영역으로 그리고 다음 행을 부 속성(sub attribute) 영역으로 추출한다. 결과적으로 그림 7(a)는 값 영역을 포함하지 않기 때문에 별도의 검사 과정을 수행할 필요 없이 가짜 테이블로 식별한다. 마찬가지 이유로 그림 7(b) 역시 테이블이 아닌 것으로 처리된다.

한편 그림 8과 같이 첫 번째 행의 일부 셀이 스팬

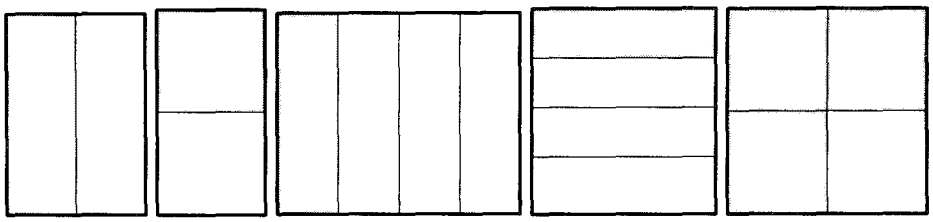
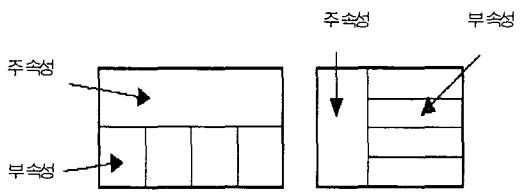


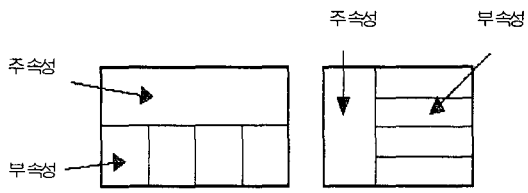
그림 6. 크기가 각각 1×2, 2×1, 1×n, n×1, 그리고 2×2인 테이블과 가능한 속성 영역 (음영으로 표시된 부분)

속성을 포함하여 2개 이상의 열에 걸쳐 있을 경우, 만일 첫 번째 행을 속성 영역이라고 가정하면 두 번째 행 역시 부 속성으로 간주된다. 따라서 이 경우 첫 번째 열을 속성 영역 그리고 나머지 열을 값 영역으로 추출한다(그림 8(a) 참조). 마찬가지로 그림 8(b)에서 첫 번째 행을 속성 영역으로 추출하고 세로 방향으로 구문적 일관성 검사 그리고(또는) 의미적 일관성 검사를 적용한다.

그 밖의 경우, 2×n(또는 n×2) 테이블은 구조적 특성상 가로 방향(또는 세로 방향)으로의 구문적 일관성과 양방향으로의 의미적 일관성을 검사할 수 있다. 예를 들어, 그림 9(a) (또는 그림 9(b))에서 첫 번째 열(또는 행)을 속성 영역으로 가정한다면, 가로(또는 세로) 방향으로의 구문적 일관성 검사와 의미적 일관성 검사를 수행할 수 있다. 반면에 첫 번째 행(또는 열)을 속성 영역으로 가정한다면, 세로(또는 가로)방향으로 의미적 일관성 여부만을 검사한다.



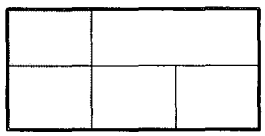
(a) 값 영역을 포함하지 않는 2×n 테이블의 예



(b) 값 영역을 포함하지 않는 n×2 테이블의 예

그림 7. 크기가 2×n(또는 n×2)이며 값 영역을 포함하지 않는 예 (n≥2)

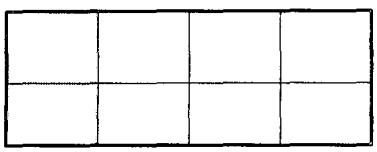
마지막으로 3×3 이상의 테이블 역시 2×n(그리고 n×2) 테이블과 마찬가지로 속성 및 값 영역을 추출한다. 추출된 영역에 대하여 구문적 및 의미적 일관성을



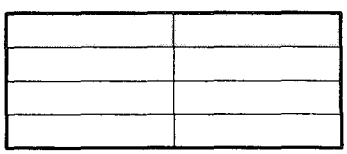
(a) 첫 번째 행에 스팬 속성을 포함하는 테이블

(b) 첫 번째 열에 스팬 속성을 포함하는 테이블

그림 8. 첫 번째 행(또는 열)에 스팬 속성을 포함하는 2×n(또는 n×2) 테이블의 예



(a) 2×4 테이블



(b) 4×2 테이블

그림 9. 2×n(그리고 n×2) 테이블의 예

검사하는 방법에 대한 자세한 기술은 다음과 같다.

3.2.2 구문적 일관성 검사

속성에 대응하는 값이 가로 방향(또는 세로 방향)으로 두개 이상 존재할 경우, 추출된 값 영역에 대하여 가로 방향 (또는 세로 방향)으로 구문적 일관성을 검사한다. 이를 위하여 가로(또는 세로) 방향으로의 일관된 정도를 "행(또는 열) 방향 일관성"이라고 부르며 값 영역을 구성하는 행(또는 열) 일관성의 평균값으로 정의한다(식 (1)과 (2) 참조).

$$\text{행(또는 열) 방향 일관성} = \frac{\sum \text{행(또는열)일관성}}{\text{행(또는열)의수}} \quad (1)$$

$$\text{행(또는 열) 일관성} = W_1 \times \text{데이터 타입 일관성} \; W_2 \times \text{길이 일관성} \quad (2)$$

특히 각각의 행(또는 열) 일관성을 계산하기 위하여 데이터 타입 일관성(data type coherency)과 길이 일관성(length coherency)을 정의한다(식 (3)과 (4) 참조). 데이터 타입 일관성은 행(또는 열)을 구성하는 주요 데이터 타입(major data type)의 비율로 정의하며, 각각의 셀에 해당하는 데이터 타입 중에서 빈도수가 가장 높은 데이터 타입을 주요 데이터 타입이라 한다. 제안된 방법은 데이터 셀에 포함된 태그의 종류에 따라 이미지 또는 폼(form) 타입으로 타입을 설정하며, 이외의 셀에 대해서는 데이터 셀의 내용에 기반하여 각각의 데이터 타입을 식별한다. 이를 위하여 제안된 방법은 표 2와 같이 대표적인 데이터 타입이 포함하는 패턴 정보를 정의한다. 본 논문은 각 셀의 데이터 타입을 그림 10과 같이 15가지로 분류한다.

$$\text{데이터 타입 일관성} = \frac{\text{주요 데이터 타입을 갖는 셀의 수}}{\text{행(또는 열)의 전체 셀 수}} \quad \ldots\ldots\ldots (3)$$

표 2. 데이터 타입 및 해당 문자열 패턴과 키워드

타입	패 턴	포함하는 키워드 정보
우편번호	ddd:ddd,ddd-ddd	
시간	d+:d+,d+:d+:d+	(am), pm(pm), hour(hr), minute(min), second(sec)
날짜	d+-d+-d+,d+/d+/d+	
월	112	January(Jan), February(Feb) … December(Dec)
일	131	Monday(Mon), Tuesday(Tue) … Sunday(Sun)
온도	정수, 실수	°
전압/전류	정수, 실수	pV, nV, mV, kV, MV, pA, nA, mA, kA, pW, nW, mW, kW, MW, Hz, kHz, MHz, GHz, THz, pF, nF
무게	정수, 실수	mg, kg, ton, grain, oz(온스), lb(파운드)
통화	정수, 실수	$, , , \, USD, CAD
백분율	정수, 실수	%

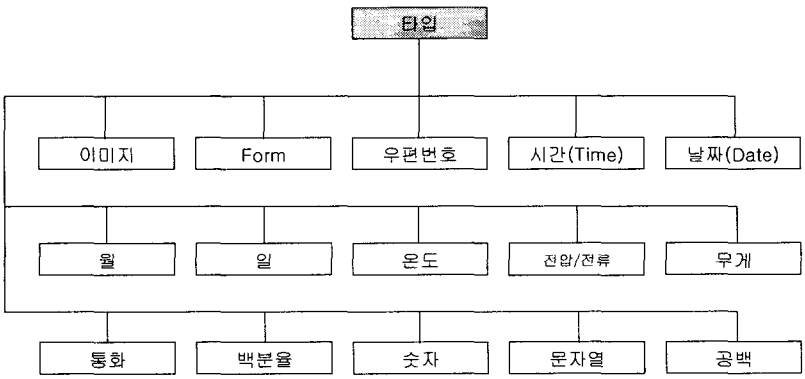


그림 10. 15가지 데이터 타입

한편 길이 일관성을 통하여 셀이 포함하는 내용의 길이가 얼마나 비슷한 지를 검사한다. 길이 일관성은 각 행(또는 열)의 평균길이를 기준으로 일정한 범위 $\propto$(셀의 평균길이×0.5≤$\propto$≤셀의 평균길이×1.5) 이내의 길이를 갖는 셀의 빈도수로 계산한다(식 (4) 참조).

$$\text{길이 일관성} = \frac{\text{길이가 범위 } \propto \text{ 이내에 포함되는 셀의 수}}{\text{행(또는 열)의 전체 셀 수}} \quad \cdots\cdots\cdots (4)$$

한편 임의의 테이블에 대하여 가로와 세로의 양 방향으로 일관성 검사가 가능할 경우, 양 방향으로 일관성 검사를 수행한 후 크기가 큰 값을 테이블 일관성(table coherency)으로 간주한다. 제안된 방법은 일관성 검사를 통하여 계산된 테이블 일관성이 임계값보다 클 경우, 진짜 테이블로 식별한다. 만일 계산된 테이블 일관성이 임계값보다 적을 경우, 해당 table 태그에 의미적 일관성 검사를 적용하여 다시 한번 테이블의 진위 여부를 검사한다.

3.2.3 의미적 일관성 검사

추출된 값 영역에 대하여 구문적 일관성 검사를 적용할 수 없거나, 계산된 테이블 일관성이 임계값보다 적은 경우, 의미적 일관성을 추가로 검사한다. 제안된 방법은 서로 대응하는 속성과 값이 의미적으로 부합하는지의 여부를 검사한다.

예를 들어, 그림 11은 속성과 값 영역 사이에 의미

E-mail	Telephone	Web Site
citizenspark@aiken.net	(803) 642-7760	www.@aiken.net

그림 11. 의미적 일관성이 존재하는 테이블의 예

표 3. 의미적 일관성 검사를 위한 키워드와 패턴 정보

분류	속성에 포함된 키워드	데이터 셀의 패턴 및 키워드
버전정보	Version	d+.d+, d+.d+.d+
이메일	E-mail, email	w+@w+.w+
운영체제	OS	Windows, NT, XP, UNIX, LINUX, Tru64
전화번호	Telephone, Phone, TEL, FAX, Contact	(d+)d+-d+, d+-d+
URI	Web Site, Online	http, www
날짜	period, Date, Create, Revise, Deadline, Start, End	d+-d+-d+, d+/d+/d+,d+Januaryd+, …, d+Decemberd+,d+Jand+, …,d+Decd+
시간	time	d+:d+:d+, (am), pm(pm), hour(hr), minute(min), second(sec)
월	Month, Date	1-12, January(Jan), February(Feb) … December(Dec)
일	Day	1-31, Monday(Mon), Tuesday(Tue) … Sunday(Sun)
높이	length	cm, km, mile, ft, inch, yd
길이	height	cm, km, mile, ft, inch, yd
너비	width	cm, km, mile, ft, inch, yd
가격	Cost, Asset, Purchases, profit, value, Price	$, , , \, USD, CAD
온도	Temperature. TEMP	°
습도	Humidity, Hum	%
속도	wind, speed	mph
거리	Visibility	km, mile, ft, inch, yd
빈도	Frequency	Hz, kHz, MHz, GHz, THz
출력	Battery, Speaker, Power	Ohm, Watts
직업	Company	Corp., Inc.

적 일관성이 존재하는 경우이다. 이 경우, 테이블의 크기가 2×n이므로 세로 방향으로 구문적 일관성을 검사할 수 없다. 첫 번째 열을 속성 영역으로, 나머지를 값 영역으로 설정한 다음 가로 방향으로의 구문적 일관성을 검사한다. 이때 값 영역에 데이터 타입과 길이 측면에서 일관성이 존재하지 않는다. 그러나 의미적 일관성 검사를 적용하여 세로 방향으로 의미적 일관성이 존재함을 알 수 있다. 즉, 속성 'E-mail', 'Telephone', 그리고 'Web Site'에 대응하는 값 영역이 모두 의미적으로 부합되는 값을 갖는다.

이를 위하여 표 3과 같이 임의의 속성의 값으로 올 수 있는 일반적인 키워드 및 패턴 정보를 기술한다. 제안된 방법은 임의의 속성-값 영역에 대하여 의미적 일관성이 존재하면 table 태그를 진짜 테이블로 식별한다.

4. 실험 결과

본 논문에서는 Wang과 Hu의 연구와 마찬가지로 테이블 식별의 정확률(precision), 재현률(recall), 그리고 F-measure의 세 가지 측면에서 제안된 방법의 성능을 분석한다. 세 가지 성능 평가 기준에 대한 정의는 표 4와 같다.

표 4. 성능 평가 기준

기준	정의
정확률	$\frac{N_{gg}}{N_{gg}N_{ng}}$
재현률	$\frac{N_{gg}}{N_{gg}N_{gn}}$
F-measure	$\frac{PR}{2}$

N_{gg} : 진짜 테이블을 진짜로 식별한 수.
N_{gn} : 진짜 테이블을 가짜로 식별한 수.
N_{ng} : 가짜 테이블을 진짜로 식별한 수.
N_{nn} : 가짜 테이블을 가짜로 식별한 수.

4.1 성능 평가

제안된 방법의 성능을 평가하기 위하여 Wang과 Hu의 연구에서 사용한 11,477개의 table 태그를 대상으로 실험하였다. 실험에 사용된 table 태그는 1,675개의 진짜 테이블과 9,802개의 가짜 테이블을 포함한다. 특히 실험을 위하여 테이블 일관성의 임계값, 데이터 타입 일관성의 가중치(W_1), 그리고 길이 일관성의 가중치(W_2)를 각각 0.54, 0.6, 그리고 0.4로 설정하였다. 이 값들은 실험을 통하여 가장 우수한 결과를 보인 가중치들이다. 제안된 방법의 성능을 정량적으로 평가한 결과는 표 5와 같다.

제안된 방법은 표 5와 같이 97.54%의 정확률과 99.22%의 재현률을 보여 97.50%의 정확률과 94.25%의 재현률을 보인 Wang과 Hu의 방법보다 우수하였다. 이는 본 논문이 테이블을 식별하기 위해서 보다 체계적이며 정교한 방법에 기반하기 때문이다. 제안된 방법은 먼저 전처리 단계로서 진짜 테이블과 레이아웃용 테이블을 구분한다. 또한 값 영역의 구문적 일관성 검사는 물론이고 일관성 검사가 어려운 table 태그에 대하여 의미적 일관성을 검사한다. 예를 들어, 일반적으로 크기가 2×2인 테이블의 경우 속성 영역과 값 영역 사이의 의미적 일관성의 유무를 검사함으로서 식별이 가능하다. 이러한 경우 Wang과 Hu의 방법과 같은 테스트 데이터에 의한 학습을 통해서는 해결하기 어렵다.

표 5. 성능 평가

실험 데이터	정확률	재현률	F-measure
1,393개의 웹 페이지에서 추출한 11,477개의 table 태그	97.54%	99.22%	98.38%

실험 결과, 제안된 방법의 테이블 식별오류는 표 6과 같으며 각각에 대한 자세한 설명은 다음과 같다. 그림 12와 그림 13은 제안된 방법이 테이블을 잘못 식별한 예이다. 그림 12는 레이아웃을 표현하기 위해서 사용된 table 태그를 진짜 테이블로 잘못 식별한 경우이다. 예를 들어 그림 12(a)와 그림 12(b)는 각각 세로 및 가로 방향으로 구문적 일관성이 존재하여 진짜 테이블로 잘못 식별되었다. 이와 같이 값 영역이 길이가 비슷한 문자열로 이루어져 구문적 일관성 검사 과정에서 참으로 판별된 경우, 의미적 일관성 검사를 추가로 적용할 필요가 있음을 알 수 있다. 실제로 의미적 일관성 검사를 적용하였더니 이들 테이블은 모두 정확하게 식별되었다.

그림 12(c)는 값 영역의 타입이 모두 날짜 타입이며 길이가 유사하기 때문에 가로 방향으로 구문적 일관성이 존재하여 진짜 테이블로 잘못 식별한 경우

표 6. 오류분석 결과

구분	오류내용	개수 (%)
가짜 → 진짜	값 영역이 길이가 비슷한 문자열로 구성되어 구문적 일관성 존재	35 (66%)
	구문적 일관성 존재	3 (6%)
	의미적 일관성 존재	2 (3%)
진짜 → 가짜	대부분의 셀이 하이퍼링크로 구성	4 (8%)
	대부분의 셀이 이미지로 구성	1 (2%)
	대부분의 셀이 공백 셀로 구성	3 (6%)
	2×2 테이블이며 의미적 일관성 부재	2 (3%)
	비정상적인 테이블 편집으로 인한 식별오류	3 (6%)
합 계		53 (100%)

이다. 그림 12(d)는 테이블이 가로방향으로 의미적 일관성 즉, 속성 "Phone"에 대응되는 값 "(281)487-0000"을 가짐으로써 진짜 테이블로 잘못 식별된 경우에 해당한다.

한편, 그림 13은 진짜 테이블을 가짜 테이블로 잘못 식별한 예이다. 그림 13(a), 그림 13(b), 그리고 그림 13(c)는 전처리 과정에서 가짜 테이블로 잘못 식

Batter (Sac Bunts included in Sac Fly in TurboStats)	Pitcher (Balks)
Batter (Hit by Pitch)	Pitcher (Home Runs)
Fielder (Double Plays)	Pitcher (Intentional Walks)
Runner (Picked Off)	Pitcher (Sac Fly and Bunts)
Catcher (Thrown runner out stealing)	Pitcher (Wild Pitch)
Catcher (Stolen base off), Catcher (Pass Ball)	Pitcher (Complete Game), Games Started, (Shut outs)

(a) 길이가 비슷한 문자열로 구성되어 세로 방향으로 구문적 일관성 존재 (문자열 타입)

Rank	Programming Service MSO with Ownership Interest

(b) 길이가 비슷한 문자열로 구성되어 가로 방향으로 구문적 일관성 존재

ASSETS	September 30, 1997	December 31, 1996 (1)

(c) 가로 방향으로 구문적 일관성 존재

	SWS COMMUNICATIONS	
	5233 Spencer Highway	
	Pasadena, Texas 77505	
Phone (281) 487-5233		Fax (281) 487-0000

(d) 가로 방향으로 의미적 일관성 존재

그림 12. 가짜 테이블을 진짜 테이블로 잘못 식별한 예

별한 경우로서, 테이블의 대부분이 각각 하이퍼링크(hyperlink), 이미지, 그리고 공백으로 구성되어 있기 때문에 전처리 규칙 4, 5, 그리고 6에 의하여 가짜로 잘못 식별되었다. 그림 13(d)는 테이블의 크기가 2×2로 의미적 일관성 검사 결과, 진짜 테이블을 잘못 식별한 경우이다. 이것은 의미적 일관성을 검사하기 위해서 구축된 정보가 해당 키워드를 포함하지 않아 테이블을 식별하지 못한 경우로서, 만약 속성과 값의 키워드로 각각 "document type"과 "interview"를 추가한다면 진짜 테이블로 식별이 가능하다. 그림 13(e)는 비정상적인 테이블 편집에 의하여 가짜로 식별된 경우로서 속성 영역(Functions, #FTEs)과 값 영역이 통합되어 있기 때문에 제안된 방법은 두 번째 줄 전체를 부 속성으로 간주하였다.

4.2 기존 연구와의 비교

전술한 바와 같이 제안된 방법은 동일한 실험데이터 및 성능평가 기준을 사용한 Wang과 Hu의 방법보다 우수한 결과를 보였다. 한편 Yang은 표 7과 같이 정확률을 본 논문과 다르게 정의하며 표 8과 같이 Chen 등이 제안한 방법과 성능을 비교하였다.

본 연구에서는 Yang이 사용한 실험 데이터를 입수할 수 없었다. 따라서 제안된 방법의 성능을 Yang의 실험 결과와 정량적으로 비교하는 것은 불가능하다. 그러나 방법론의 측면에서 제안된 방법과 Yang의 방법을 정성적으로 비교하면 다음과 같다. 먼저 두 방법 모두 임의의 규칙을 적용하여 전처리 단계에서 테이블을 식별한다. 특히 Yang은 두 개의 규칙을 적용하여 가짜 테이블을 추출하는 반면에 제안된 방법은 진짜 테이블 또는 레이아웃용으로 사용된 table 태그의 일반적인 특징을 반영한 보다 정교한 규칙을 제안한다.

또한 Yang은 본 논문의 의미적 일관성 정보와 유

Course Code	Course Name
150	Compaq Advocate Orientation Seminar
278	Compaq Products Quarterly Training (CPQ)
278C	Compaq Products Quarterly Training (CPQ) - Canada
505	Selling Compaq Business Systems In Today's Markets
509	Alpha Sales Training
510	Selling Compaq Enterprise Solutions In Today's Markets
521	ProLiant Full Line Training
544	Internet Enabling Workshop Nonstop? Himalaya Servers
559	StorageWorks Full Line Sales Training
560	Selling Compaq Enterprise Storage Solutions
562	Winning the Financial Sale: A Finance and Leasing Seminar
GOVERNMENT, EDUCATION, AND MEDICAL COURSES (GEM)	
472	Certified Education Partner (CEP) Summit

(a) 데이터 셀의 대부분이 하이퍼링크로 이루어져 가짜 테이블로 잘못 식별된 예

Index	Last	Change	% Chg
Dow	11,142.66	269.69	2.48%
NASDAQ	2,154.03	68.45	3.28%
S&P 500	1,278.96	29.52	2.36%
30 Yr	58.73	0.21	0.35%
Russell	496.24	6.61	1.34%

(b) 데이터 셀의 대부분이 이미지로 이루어져 가짜 테이블로 잘못 식별된 예

The Gallup Poll March 26-28, 2001 N=1,024 adults nationwide MoE 73

"Next, we'd like to ask you some questions about sports. What is your favorite sport to watch?"

	%
Football	28
Basketball	16
Baseball	12
Auto racing	6
Golf	4
Ice/Figure skating	4
Ice hockey	3
Soccer	2
Tennis	2
Boxing	2
Gymnastics	1
Motorcross	1
Wrestling	1
Volleyball	1
Other	4
None	12
No opinion	1

"For each of the following, please say whether you are a fan of that sport or not. . . ."

	Yes %
Professional football	54
Professional baseball	46
College football	44
Figure skating	40
College basketball	38
Professional basketball	36
Auto racing	31
Professional golf	27
Professional ice hockey	24
Professional tennis	19
Professional wrestling	12

(c) 데이터 셀의 대부분이 공백으로 이루어져 가짜 테이블로 잘못 식별된 예

Title	Document Type
Manager Interview: The View From Long-Term Corporate Fund	Interview

(d) 의미적 일관성의 부재로 인하여 가짜 테이블로 잘못 식별된 예

A-76 Studies Initiated in FY 1997	
Functions	# FTEs
Social services	2,331
General maintenance and repair	6,460
Installation support	5,868
Real property maintenance	5,168
Base multifunction services	9,223
Data processing	751
RDT&E support	743
Other nonmanufacturing	2,817
Education and training	569
Health services	360

(e) 비정상적인 테이블 편집에 의하여 가짜 테이블로 잘못 식별된 예

그림 13. 진짜 테이블을 가짜 테이블로 잘못 식별한 예

사한 AIEP와 VIEP 개념을 제안한다. 그러나 진짜 테이블의 식별을 위해서 속성과 값으로 사용될 수 있는 키워드 패턴을 모두 포함하여야 한다는 제약을 갖는다. 예를 들어, Yang의 방법은 표 8과 같이 값 영역이 숫자로 이루어진 테이블을 식별하는데 한계를 갖는 반면, 본 연구에서는 테이블의 일반적인 특성인 구문적 일관성의 유무로 테이블을 식별한 후

표 7. Yang의 성능 평가 기준

기 준	정 의
정확률	$\frac{N_{gg}N_{nn}}{N_{gg}N_{ng}N_{gn}N_{nn}}$
재현률 및 F-measure	본 논문과 동일

표 8. Yang의 방법과 Chen 등의 방법의 성능 비교[8]

방 법	테이블 수	정확률	재현률	F-measure
Chen 등	1,927	82.78%	48.32%	65.55%
Yang	1,927	94.57%	100%	97.29%

나머지 테이블에 대하여 의미적 일관성 검사를 적용하기 때문에 모든 가능한 키워드를 포함하지 않고도 대부분의 테이블을 식별할 수 있다.

한편, Penn은 진짜 테이블이 갖는 기본적인 특성을 사용하여 간단하게 테이블을 식별하였는데, 이는 제안된 방법의 전처리 부분에 해당된다. 따라서 Penn의 방법은 테이블의 주된 특성인 구조적 및 의미적 일관성을 고려하지 않기 때문에 테이블을 식별하는데 있어 한계를 갖는다. Yoshida는 테이블의 구조를 미리 정의된 9개의 유형중에 한가지로 분류한다. 따라서 테이블의 식별이라기보다는 구조 인식을 위한 방법에 해당한다.

Hurst는 비교적 적은 수의 실험 데이터에 대하여 다양한 분류기의 성능을 비교하였으며, 또한 테이블 식별을 위한 새로운 방법보다는 기존 분류기의 성능 평가에 초점을 두었다.

따라서 제안된 방법은 기존 연구들과는 달리, 데이터에 대한 학습이나 많은 온톨로지 정보(속성과 값으로 사용 될 수 있는 키워드 패턴)를 사용하지 않고도 효율적으로 테이블의 식별이 가능하다. 표 9는 HTML 문서로부터 테이블을 식별하는 방법에서 사용한 실험 데이터의 종류 및 성능 평가 결과를 정리한 것이다.

5. 결론 및 향후 연구 방향

최근 들어 웹을 통하여 새롭게 생성되는 정보의 양이 급속도로 증가하면서 웹으로부터 유용한 정보를 추출하는데 관심이 모아지고 있다. 특히 테이블은 연관된 정보를 효과적으로 표현하며 웹 문서표준인 HTML은 테이블의 표현을 위해서 table 태그를 정의한다. 한편 table 태그가 유용한 정보를 포함하는 진짜 테이블은 물론이고 문서의 레이아웃을 렌더링하기 위한 용도로도 널리 사용되고 있다. 따라서 본 논문에서는 HTML 문서에 포함된 table 태그로부터 진짜 테이블을 식별할 수 있는 효율적인 방법을 제안하였다.

제안된 방법은 보다 체계적이며 정교한 수준의 테이블 식별을 위하여 전처리와 연관관계 추출의 두 단계로 구성된다. 먼저 전처리 과정을 통하여 진짜 또는 가짜 테이블로 명확히 식별 가능한 table태그를

표 9. 관련 연구의 실험 데이터 및 실험 결과

관련 연구	실험 데이터	실험결과		
		정확률	재현률	F-measure
Chen 등[4]	야후 웹 사이트에서 추출한 여행관련 918개의 테이블	92.92%	80.07%	86.50%
Penn 등[5]	뉴스, 텔레비전, 라디오, 단체 관련 75개의 웹페이지	86.30%	89.80%	88.05%
Yoshida 등[6]	임의의 웹 페이지에서 추출한 175개의 테이블	79.44%	85.86%	82.65%
Hurst[7]	임의의 웹 페이지에서 추출한 339개의 테이블	95.00%	93.50%	94.20%
Yang[8]	주식, 날씨, 레스토랑, 여행, 옥션, 정부기관 등 150개의 웹 페이지에서 추출한 1,927개의 테이블	94.57%	100%	97.29%
Wang과 Hu[9]	비즈니스, 뉴스, 과학 관련 총 1,393개의 웹 페이지에서 추출한 11,477개의 테이블	97.50%	94.25%	95.88%
제안된 방법	Wang과 Hu와 동일한 실험 데이터	97.54%	99.22%	98.38%

추출한다. 이를 위하여 일반적으로 진짜 테이블 또는 레이아웃을 표현하기 위해서 사용되는 table 태그의 특징을 반영한 규칙을 제안하였다. 두 번째 연관관계 추출은 테이블 영역을 속성과 값 영역으로 구분한 후, 값 영역에 대하여 구문적 일관성이 존재하는지의 여부를 검사한다. 이는 테이블을 구성하는 값 영역이 해당 속성에 대하여 일관된 데이터타입 및 비슷한 길이를 갖는다는 사실에 기인한다. 한편 테이블의 크기가 작아서 구문적 일관성 검사를 수행할 수 없는 경우, 속성과 값 영역 사이의 의미적 일관성 검사를 통하여 테이블을 식별한다. 실험 결과, 제안된 방법은 기존 연구와 비교하여 우수한 성능을 보였다. 특히 기존 연구와 달리 방대한 양의 데이터에 대한 학습 시간이 불필요하다.

한편, HTML은 본래 문서의 내용을 시각적으로 렌더링하기 위한 용도로 제안된 포맷이기 때문에 컴퓨터로 하여금 유용한 정보를 추출 및 재가공 하기에는 부적합하다. 반면, 논리적 구조 정보를 표현할 수 있는 XML (Extensible Markup Language)[14]은 기존에 데이터로써의 가치가 떨어지며 이질적인 형태의 웹 컨텐츠를 컴퓨터 처리 및 가공이 가능한 형태로 급속하게 변화시키고 있다. 향후 본 연구에서는 식별된 테이블 정보의 효과적인 재사용을 지원하기 위하여 테이블의 논리적인 구조를 인식하여 이를 XML 형태로 변환하는 연구를 진행할 계획이다.

참 고 문 헌

[1] World Wide Web Consortium, Hypertext Markup Language (HTML) 4.01, W3C Recommendation, http://www.w3.org/TR/html4/, 1999.

[2] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo, "Extracting Semistructured Information from the Web," *Proc. PODS/SIGMOD*, pp. 1825, Tucson, Arizona, May 1997.

[3] M. Hurst, "Layout and language: Challenges for Table Understanding on the Web," *Proc. First Int'l Workshop on Web Document Analysis*, pp. 27~30, Seattle, USA, Sep. 2001.

[4] H.-H. Chen, S.-C. Tsai and J.-H. Tsai, "Mining Tables from Large scale HTML Texts," *Proc. 18th Int'l Conf. Computational Linguistics*, Vol. 1. pp. 166172, 2000.

[5] G. Penn, J. Hu, H. Luo, and R. McDonald, "Flexible Web Document Analysis for Delivery to Narrow-Bandwidth Devices," *Proc. Fifth Int'l Conf. Document Analysis and Recognition(ICDAR01)*, pp. 10741078, Seattle, USA, Sep. 2001.

[6] M. Yoshida, K. Torisawa, and J. Tsujii, "A Method to Integrate Tables of the World Wide Web," *Proc. First Int'l Workshop on Web Document Analysis(WDA 2001)*, pp. 3134, Seattle, USA, Sep. 2001.

[7] M. Hurst, "Classifying TABLE Elements in HTML," *Proc. 11th International World Wide Web Conference*, Honolulu, HI, May 2002. http://www2002.org/CDROM/poster/115/index.html.

[8] Y. Yang, *Web Table Mining and Database Discovery*. MSc Thesis, Simon Fraser University, Aug. 2002.

[9] Y. Wang and J. Hu, "Detecting Tables in HTML Documents," *Proc. 5th IAPR Int'l Workshop on Document Analysis System (DAS'02)*, pp. 249~260, Princeton, USA, Aug. 2002.

[10] Hartley, 1958, Biometrics, 174-194.

[11] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[12] N. Littlestone and M. K. Warmuth, "The Weighted Majority Algorithm," *Information and Computation*, Vol. 108, No. 2, pp. 212261, 1994.

[13] K. Nigam, J. Lafferty, and A. McCallum, "Using Maximum Entropy for Text Classification," *Proc. IJCAI Workshop on Information Filtering*, pp. 6167, 1999.

[14] World Wide Web Consortium, Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, http://www.w3c.org/TR/REC-xml, 2000.

김 연 석

2003년 명지대학교 전자정보통신공학부 졸업(학사)
2003년~현재 연세대학교 컴퓨터과학과 석사과정

관심분야 : 웹문서 분석, 정보 추출 및 통합, XML

이 경 호

1995년 연세대학교 전산과학과 졸업(학사)
1997년 연세대학교 컴퓨터과학과 졸업(석사)
2001년 연세대학교 컴퓨터과학과 졸업(박사)
2001년 National Institute of Standards and Technology(NIST) 객원연구원
2002년~현재 연세대학교 컴퓨터산업공학부 조교수
관심분야 : 멀티미디어 문서처리, XML, 웹 서비스