Journal of Korea Multimedia Society Vol. 9, No. 9, September 2006(pp. 1231-1246)

HTML 테이블의 논리적 구조분석을 위한 효율적인 방법

김연석†, 이경호††

요 약

시각적 렌더링 포맷인 HTML은 연관된 정보를 간결하게 표현하기 위하여 테이블을 사용한다. 그러나 HTML은 컴퓨터로 하여금 정보를 처리 및 가공하게 한다는 측면에서 한계를 갖기 때문에 논리적 구조정보의 표현이 가능한 XML 문서로의 변환이 필요하다. 본 논문에서는 웹으로부터 정보를 추출하기 위한 목적의 일환으로 HTML 테이블의 논리적인 구조를 추출하여 XML 문서로 변환하는 효율적인 방법을 제안한다. 제안된 방법은 영역구분과 구조분석의 두 단계로 구성된다. 영역구분 단계에서는 잡음영역을 제거하며 시각적 및 의미적 일관성 검사를 통하여 테이블에 존재하는 속성과 값 영역을 구분한다. 이후 구조분석 단계에서는 구분된 영역에 제안된 테이블 모델을 적용하여 계층구조를 추출하며, 이로부터 XML 문서를 생성한다. 제안된 영역구분 방법의 성능을 평가하기 위하여 1,180개의 테이블을 대상으로 실험한 결과, 평균적으로 86.7%의 정확도를 보여 기존 연구보다 우수하였다.

An Efficient Method for Logical Structure Analysis of HTML Tables

Yeon-Seok Kim†, Kyong-Ho Lee††

ABSTRACT

HTML is a format for rendering Web documents visually and uses tables to present a relational information. Since HTML has limits in terms of information processing and management by a computer, it is important to transform HTML tables into XML documents, which is able to represent logical structure information. As a prerequisite for extracting information from the Web, this paper presents an efficient method for extracting logical structures from HTML tables and transforming them into XML documents. The proposed method consists of two phases: Area segmentation and structure analysis. The area segmentation step removes noisy areas and extracts attribute and value areas through visual and semantic coherency checkup. The hierarchical structure between attribute and value areas are analyzed and transformed into XML representations using a proposed table model. Experimental results with 1,180 HTML tables show that the proposed method performs better than the conventional method, resulting in an average precision of 86.7%.

Key words: HTML Table(HTML 테이블), Structure Analysis(구조분석), Attribute-Value Relations(속성-값 연관관계), Information Extraction(정보추출), XML

1. 서 론

웹 문서 표준인 HTML(Hypertext Markup Language)은 웹 문서를 시각적으로 렌더링하기 위한 포맷이기 때문에 컴퓨터로 하여금 정보를 처리하게 한다는 측면에서 한계를 갖는다. 반면 XML (eXtensible Markup Language)은 논리적 구조정보를 표현할 수 있으며 플랫폼에 독립적이라는 장점

※ 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 : 2006년 4월 4일, 완료일 : 2006년 7월 20일

† 연세대학교 컴퓨터과학과 박사과정
†† 정회원, 연세대학교 컴퓨터산업공학부 부교수
(E-mail: khlee@cs.yonsei.ac.kr)
※ 본 논문은 교육부 BK21 사업의 연구비 지원을 받았음

때문에 다양한 분야에서 정보의 공유 및 교환을 위한 표준으로 널리 사용되고 있다. 따라서 HTML로부터 유용한 정보를 추출하고 이를 XML 형식으로 변환하는 방법이 필수적이다. 특히 HTML은 연관된 정보(relational information)를 간결하게 표현하기 위하여 테이블을 사용하는데, 본 논문에서는 웹으로부터 유용한 정보를 추출하기 위한 목적의 일환으로 <그림 1>과 같이 HTML 테이블의 논리적 구조를 분석한 후 이를 XML 문서로 변환하는 효율적인 방법을 제안한다.

본 논문에서는 테이블을 연관성을 갖는 데이터의 배열이라고 정의하며 속성(attribute)과 값(value)의 연관관계를 포함하는 테이블을 진짜 테이블(genuine table)로 간주한다. 특히 <그림 1(a)>와 같이 속성을 나타내는 셀의 집합을 속성 영역(attribute area), 값을 나타내는 셀의 집합을 값 영역(value area) 이라고 정의한다. 한편, 제안된 방법은 테이블을 단순 테이블(simple table)과 합성 테이블(complex table)로 분류한다. 단순 테이블은 다시 속성 영역의 위치에 따라 열 방향 테이블(column-wise table), 행 방향 테이블(row-wise table), 그리고 타임 테이블(time table)로 분류되며, 합성 테이블은 하나 이상의 단순 테이블을 포함하는 복합 테이블(composite table)[6]과 단일 셀에 속성과 값이 함께 존재하는 혼합-셀 테이블(mix-cell table)[9]로 분류한다. <그림 2>는 테이블 분류의 예이다.

HTML 테이블에 관한 연구는 크게 웹으로부터 진짜 테이블을 식별하는 연구와 식별된 테이블로부터 논리적 구조를 분석하여 속성-값 연관관계를 추출하는 두 가지 연구로 나누어진다[13]. 그러나 논리적 구조 추출을 위한 기존 연구의 대부분은 특정 온톨로지에 의존적이거나 단순한 규칙에 기반하기 때문에 다양한 종류의 테이블에 적용하는데 있어 한계를 가진다[1-10][16-18]. 또한 잡음 영역을 제거하지 않기 때문에 논리적 연관관계를 정확히 추출하는데 제한적이다.

본 논문에서는 기존 연구들의 문제점을 개선하기 위하여 테이블의 논리적 구조를 분석하고 이를 XML 문서로 변환하는 효율적인 방법을 제안한다. 제안된 방법은 테이블을 속성과 값 영역으로 구분하는 영역 구분과 테이블의 논리적 계층구조를 추출하여 XML 문서를 생성하는 구조분석의 두 단계로 구성된다. 영역구분 단계에서는 먼저 잡음 영역을 제거한 후 시각적 및 의미적 일관성 검사를 적용하여 테이블을 속성과 값 영역으로 구분한다. 또한 합성 테이블의 영역 구분을 위하여 휴리스틱한 규칙을 적용한다. 한편 구조분석 단계에서는 제안된 테이블 모델을 기반으로 속성 및 값 영역으로부터 논리적 계층구조를 추출하

속성 영역

값 영역

Company Name	Last	Change	%Change
CSCO	24.74	+0.92	0.01%
INTC	30.88	+0.96	0.01%
MSFT	27.08	+0.12	0.00%
JDSU	4.89	+0.15	0.01%
ORCL	13.42	-0.14	-0.01%
SIRI	2.80	+0.10	0.01%
SUNW	5.41	+0.10	0.00%
LVLT	5.14	+0.23	0.01%
AMAT	22.38	+1.02	0.01%
DELL	32.95	+0.84	0.01%

(a)

```
<?xml version="1.0" encoding="UTF-8" ?>
<table>
<CompanyName>
  <CSCO>
    <Last>24.74</Last>
    <Change>+0.92</Change>
    <Change>0.01%</Change>
  </CSCO>
  <INTC>
    <Last>30.88</Last>
    <Change>+0.96</Change>
    <Change>0.01%</Change>
  </INTC>
  <MSFT>
    <Last>27.08</Last>
    <Change>+0.12</Change>
    <Change>0.00%</Change>
  </MSFT>
  <JDSU>
    <Last>4.89</Last>
    <Change>+0.15</Change>
    <Change>0.01%</Change>
  </JDSU>
  <ORCL>
    <Last>13.42</Last>
    <Change>-0.14</Change>
    <Change>-0.01%</Change>
  </ORCL>
  <SIRI>
    <Last>2.80</Last>
    <Change>+0.10</Change>
    <Change>0.01%</Change>
  </SIRI>
  <SUNW>
    <Last>5.41</Last>
    <Change>+0.10</Change>
    <Change>0.00%</Change>
  </SUNW>
  <LVLT>
    <Last>5.14</Last>
    <Change>+0.23</Change>
    <Change>0.01%</Change>
  </LVLT>
  <AMAT>
    <Last>22.38</Last>
    <Change>+1.02</Change>
    <Change>0.01%</Change>
  </AMAT>
  <DELL>
    <Last>32.95</Last>
    <Change>+0.84</Change>
    <Change>0.01%</Change>
  </DELL>
</CompanyName>
</table>
```

(b)

그림 1. HTML 테이블의 XML 변환 예: (a) HTML 테이블의 예, (b) 추출된 XML 문서

DATE	OPPONENT	TIME	SCORE
1	Washington	7:30	4-2
2	Ottawa	7:30	7-3
6	Edmonton	7:30	5-2
8	TampaBay	7:30	5-5
9	Florida	7:30	4-2
12	Buffalo	7:30	3-0
14	Boston	7:30	2-1(OT)
16	N.Y.Rangers	7:30	8-3
19	St.Louis	7:30	4-2
21	Philadelphia	7:30	8-3
22	Hartford	7:00	7-1
27	Montreal	7:30	2-2
30	Boston	7:30	6-1

(a)

Temperature	66°
Humidity	75%
Precip. today	n/a
Wind	SE at 13 mph
Barometer	30.02 in
Dewpoint	58°
Visibility	10 mi
Sunrise	5:31 a.m
Sunset	8:15 p.m

(b)

Breeding Pair Combinations	CLEAR MALE	CARRIER MALE	AFFECTED MALE
CLEAR MALE	100% Clear	50/50 Carrier/Clear	100% Carrier
CARRIER FEMALE	50/50 Carrier/Clear	25/50/25 Clr./Carr./Affctd.	50/50 Carrier/Affected
AFFECTED FEMALE	100% Carrier	50/50 Carrier/Affected	100% Affected

(c)

Category	State	Years of Experience	Base Compensation	Bonus	Total	Level
Corporate Salaried	Colorado	3-5	$80K	$5K	$85K	Average
			$60K	$0K	$60K	Low
			$95K	$10K	$100K	High
Category	**State**	**Years of Experience**	**Base Compensation**	**Bonus**	**Total**	**Level**
1099 Contractor	Arkansas	1-2	$30/hr	n/a	n/a	Average
			$20/hr	n/a	n/a	Low
			$40/ht	n/a	n/a	High
Category	**State**	**Years of Experience**	**Base Compensation**	**Bonus**	**Total**	**Level**
W-2 Contractor	Texas	6-10	$50/hr	n/a	n/a	Average
			$40/hr	n/a	n/a	Low
			$65/hr	n/a	n/a	High
Category	**State**	**Years of Experience**	**Base Compensation**	**Bonus**	**Total**	**Level**
Agency Recruiter	Illinois	10+	$45K	$25K	$60K	Average
			$30K	$10K	$40K	Low
			$60K	$40K	$100K	High

(d)

Last: 60.31	Change: +0.87	Open: 59.95	High: 60.65	Low: 59.66	Volume: 21,941,000
	Percent Change: +1.46%	Yield: n/a	P/E Ratio: 56.37	52 Week Range: 47.50 to 76.15	

(e)

그림 2. 테이블의 분류: (a) 열 방향 테이블의 예, (b) 행 방향 테이블의 예, (c) 타임 테이블의 예, (d) 복합 테이블의 예, (e) 혼합-셀 테이블의 예

며 이로부터 XML 문서를 생성한다. 제안된 영역구분 방법의 성능평가를 위하여 1,180개의 HTML 테이블을 대상으로 실험한 결과, 86.7%의 정확도를 보여 기존 연구보다 우수하였다.

본 논문의 구성은 다음과 같다. 2절에서는 HTML 테이블의 구조분석 관련 기존 연구의 특징을 간략히

기술하고, 3절에서는 제안된 방법을 영역구분과 구조분석의 두 단계로 구분한 후, 이를 자세히 기술한다. 4절에서는 실험결과를 통하여 제안된 방법의 성능을 기존 연구와 비교 및 분석하며, 마지막으로 5절에서는 결론 및 향후 연구방향을 기술한다.

2. 관련 연구

HTML 테이블의 구조분석은 일반적으로 포맷팅, 레이아웃, 그리고 온톨로지의 세 가지 정보를 사용한다. 포맷팅 및 레이아웃 정보는 각각 데이터와 테이블의 모양을 나타내는 정보로서 시각적 차이에 의해 속성과 값 영역을 구분하기 위하여 사용된다. 또한 속성과 값을 정의한 온톨로지는 테이블의 의미적 특징을 사용하여 영역을 구분하는데 사용된다. <표 1>은 HTML 테이블의 구조분석에 관한 기존 연구의 특징을 요약한 것이다.

Jung 등 [1] 은 테이블의 각 셀에 7개의 휴리스틱한 규칙을 적용하여 얻어진 이진 행렬의 값 차이에 따라 속성과 값 영역을 구분한다. 그러나 본 방법은 제한된 수준의 휴리스틱한 규칙을 적용하기 때문에 다양한 종류의 속성 영역을 포함하는 복잡한 테이블을 처리하는데 한계를 갖는다. 또한 잡음 영역의 제거 및 정규화 과정을 거치지 않기 때문에 정확한 영역구분이 어려울 수 있다.

Pivk 등 [2] 은 테이블의 레이아웃과 컨텐츠 타입 등을 사용하여 FTM(Functional Table Model)을 생성한 후 WordNet 등 기존에 구축된 온톨로지를 사용하여 의미적인 연관관계를 추출한다. Li 등 [3-6] 은 HTML 테이블을 제안된 개념모델(conceptual model)로 변환한 후 이로부터 XML 문서를 생성한다. 이때 속성과 값 영역은 폰트 크기 및 스타일 등의 포맷팅 정보를 사용하여 구분한다. 한편 한개 이상의 캡션(caption)과 분리된 헤딩(separate heading)을 가지는 테이블을 복합 테이블(composite table)로 정의하며, 이를 다수의 서브 테이블로 분리한 후 제안된 변환규칙을 적용한다. 그러나 본 방법은 제한된 포맷팅 정보만을 사용하여 속성과 값 영역을 구분한다.

Masuda 등 [7] 은 행과 열간의 유사도 차이에 기반하여 속성과 값 영역을 구분한다. 특히, 107개의 특징을 사용하여 행(혹은 열) 방향에 존재하는 셀 간의 유사도와 행(혹은 열) 간의 유사도를 계산한다. 한편 제안된 방법이 사용하는 특징의 대부분은 일본어에 의존적이다. Itai 등 [8] 은 동일한 도메인에 존재하는 여러 가지 형태의 테이블을 단일 DTD (Document Type Definition)를 가지는 XML 문서로 통합한다. 이를 위하여 테이블의 셀에 해당하는 블록(block)을 SVM(Support Vector Machine)과 HMM (Hidden Markov Model) 기법을 사용하여 속성과 값 영역으로 구분한다.

Yang 등 [9] 은 3개의 휴리스틱한 규칙을 적용하여 속성과 값 영역을 구분한다. 또한 속성과 값이 단일 셀에 존재하는 혼합-셀 테이블(mix-cell table)을 인식하기 위하여 속성과 값의 시각적인 차이와 구두점을 이용한다. Yoshida [10] 는 온톨로지와 HMM 기법을 사용하여 테이블의 셀을 나타내는 블록의 역

표 1. HTML 테이블의 구조분석 방법

연도	저자	특징	XML 생성여부
2006	Jung 등 [1]	7개의 휴리스틱한 규칙을 사용하여 영역구분	×
2005	Pivk 등 [2]	레이아웃 및 컨텐츠 타입을 사용하여 계층구조 생성	×
2004	Li 등 [3-6]	포맷팅 정보에 기반하여 속성 및 값 영역 추출	○
2004	Masuda 등 [7]	107개의 규칙을 사용하여 행과 열간의 유사도 계산	×
2003	Itai 등 [8]	SVM과 HMM 기법 적용	○
2002	Yang 등 [9]	3개의 휴리스틱한 규칙을 사용하여 영역구분	×
2002	Yoshida [10]	온톨로지와 HMM 기법 적용	×
2002	Hu 등 [11-14]	레이아웃 및 구문적 일관성의 차이를 이용하여 영역구분	×
2001	Masuda 등 [15]	컨텐츠 인식과 구조 인식의 두 단계를 통하여 테이블의 구조인식	×
2000	Lim 등 [16][17]	태그 TH와 TD를 사용하여 영역구분	○
2000	Wang 등 [18]	온톨로지를 사용한 영역구분	×
2000	Chen 등 [19]	태그속성 span에 기반에 기반하여 속성과 값의 쌍 추출	×

할을 구분한다. 즉, 입력된 블록 시퀀스(block sequence)에 해당하는 가장 적절한 스태이트 시퀀스(state sequence)를 HMM 기법을 사용하여 추정한다. 이때 스태이트란 블록의 역할, 즉, 속성 또는 값을 의미한다.

Hu 등 [11-14] 은 레이아웃 정보를 사용하여 추출된 값 후보 영역에 대하여 구문적 일관성을 검사한다. 알파벳과 비 알파벳의 두 가지 타입을 사용한 일관성 검사는 값 후보 영역에 포함된 이질적인 타입을 가진 속성 영역을 제거한다. Masuda 등 [15] 은 컨텐츠 인식(contents recognition)과 구조 인식(structure recognition)의 두 단계를 통하여 테이블의 구조를 분석한다. 컨텐츠 인식은 테이블의 속성이 가지는 3가지 일반적인 특징, 즉, 순차적인 수(혹은 일자, 등급 등), 제한된 문자길이, 그리고 구두점 등에 기반하여 행 속성의 개수와 열 속성의 개수를 계산하고, 이로부터 테이블을 타임 테이블(time table), 세로방향 테이블(lengthways table), 가로방향 테이블(sideways table)로 분류한다. 한편, 행 속성의 개수와 열 속성의 개수를 사용하여 테이블의 분류가 불가능한 경우에는 좌상단 셀의 데이터 유무와 열의 개수를 사용하여 테이블을 분류한다.

Lim 등 [16,17] 은 HTML 문서를 XML 문서로 변환하는 휴리스틱한 방법을 제안한다. 제안된 방법은 먼저 테이블을 태그 TH와 TD를 식별하여 속성과 값 영역으로 구분한 후 테이블을 정규화 한다. 정규화된 테이블은 제안된 규칙을 적용하여 계층구조를 가지는 DAH(DAta Hierachy)로 변환되며 이로부터 XML 문서가 생성된다. Wang 등 [18] 은 온톨로지를 사용하여 테이블의 셀을 개념(concept)과 인스턴스(instance)로 구분하고, 이로부터 테이블을 1차원, 2차원, 그리고 컴플렉스 테이블(complex table)로 분류한다. Chen 등 [19] 은 테이블의 우하단 셀과 태그 속성 스팬(span)이 사용된 행(혹은 열)을 경계로 속성과 값의 쌍을 추출한다. 이후 나머지 부분에 대하여 동일한 과정을 반복함으로써 테이블의 모든 데이터를 속성과 값의 쌍으로 구분한다. 그러나 본 방법은 다양한 형태를 가지는 테이블에 적용하기 어렵다.

테이블 구조분석에 관한 기존 연구들의 대부분은 제한된 포맷팅 및 레이아웃 정보를 사용하거나 혹은 특정 언어와 도메인에 의존적이다. 또한 잡음 영역을 갖는 테이블, 즉, 주석을 포함하는 테이블, 레이아웃을 위하여 공백 행(혹은 열)이 삽입된 테이블 등에 대하여 그 영역을 제거하지 않기 때문에 정확한 구조분석이 어렵다. 뿐만 아니라 비정상적으로 편집된 테이블에 대해서는 기존 연구들의 방법으로 분석이 어렵다. 이를 위하여 본 논문에서는 잡음영역 제거와 정규화 과정을 통하여 테이블의 구조분석을 용이하게 하며, 다양한 시각적 정보와 의미적 일관성을 사용함으로써 효과적인 영역구분 및 연관관계 추출이 가능하게 한다. 뿐만 아니라 합성 테이블의 구조분석을 위하여 휴리스틱한 규칙을 제안한다.

3. 제안된 방법

본 절에서는 테이블의 속성 및 값 영역을 구분하여 구조를 분석하고 이로부터 XML 문서를 생성하는 효율적인 방법에 대하여 기술한다. 제안된 방법은 <그림 3>과 같이 영역구분과 구조분석의 두 단계로 구성된다. 특히 영역구분 단계는 전처리, 시각적 일관성 검사, 의미적 일관성 검사, 그리고 후처리의 네 부분으로, 구조분석 단계는 테이블 모델에 기반한 구

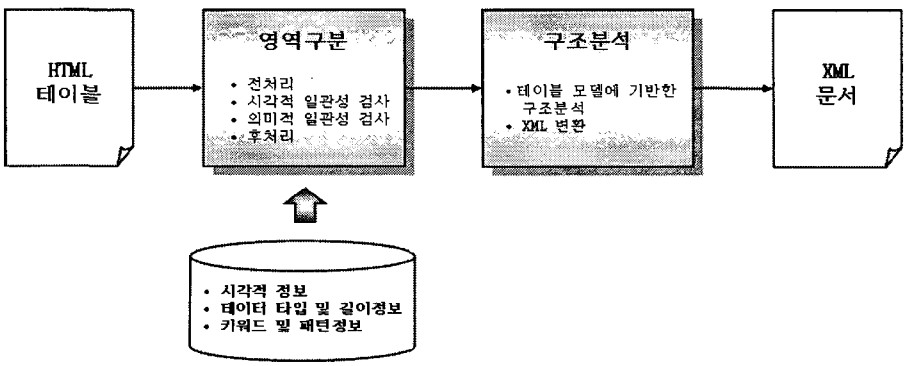

그림 3. 테이블의 구조분석 및 XML 변환 과정

조분석과 XML 변환의 두 부분으로 이루어진다.

제안된 방법은 테이블의 영역구분을 위하여 잡음 영역 제거 및 정규화 과정 등의 전처리 단계를 거친 후 시각적 및 의미적 일관성 검사를 통하여 속성과 값 영역을 구분한다. 구조분석 단계에서는 구분된 영역에 제안된 테이블 모델을 적용하여 계층구조를 추출하며 이로부터 XML 문서를 생성한다. 각 단계에 대한 자세한 설명은 다음과 같다.

3.1 영역구분

영역구분 단계에서는 테이블의 정확한 구조분석을 위하여 전처리 단계를 거친다. 전처리 단계에서는 4개의 정규화 규칙을 적용하여 테이블의 잡음 영역을 제거하고 태그속성 스팬(span)을 정규화 한다. 정규화된 테이블은 시각적 및 의미적 일관성 검사를 통하여 속성과 값 영역으로 구분된다. 한편 합성 테이블은 후처리 단계를 통하여 영역이 구분된다.

3.1.1 전처리

전처리 과정은 주석을 포함하거나 다수의 셀이 통합되어 표현된 테이블을 정규화 한다. 이를 위하여 제안된 방법은 <그림 4>와 같이 정규화 규칙을 적용한다. 예를 들어, <그림 5(a)>는 'A-76 Studies Initiated in FY 1997'의 캡션을 가지며, 'Functions' 및 '#FTEs'의 값을 갖는 속성 셀이 값 영역을 구성하는 다수의 셀과 통합되어 표현된 테이블이다. 이러한 테이블은 전처리 규칙 2와 4에 의해 <그림 5(b)>와 같이 캡션이 분리되고, 하나의 셀에 하나의 값을 갖는 테이블로 정규화 된다.

(1) IF 행(또는 열)의 데이터가 존재하지 않음
THEN 해당 행(또는 열) 제거
(2) IF 첫 번째 행(또는 열)의 태그속성 스팬 값이 열(혹은 행)의 수와 동일
THEN 해당 행(또는 열)을 캡션으로 간주
(3) IF 태그속성 스팬의 값이 1이 아님
THEN 해당 셀 데이터를 복사하여 셀 추가
(4) IF 셀들의 데이터가 일정한 패턴을 가지고 반복됨
THEN 패턴에 따라 셀 분리

그림 4. 제안된 전처리 규칙

3.1.2 시각적 일관성 검사

정규화된 테이블은 시각적 일관성 검사를 통하여 속성과 값 영역이 구분된다. 시각적 일관성 검사란 HTML 테이블이 속성과 값 영역을 구분하기 위하여 서로 다른 포맷팅 정보를 사용한다는 사실에 기반한 검사로서, HTML에 정의된 다양한 태그를 사용한다.

일반적으로 테이블의 포맷팅을 위하여 사용된 태그들 중에는 그 사용 여부에 따라 테이블의 영역을 구분할 수 있는 태그가 존재한다. 제안된 방법은 <표 2>의 HTML 태그를 식별함으로써 영역을 구분한다. 태그에 의한 영역구분이 불가능할 경우에는 포맷팅 일관성 검사를 통하여 영역을 구분한다. 포맷팅 일관성이란 셀의 포맷을 결정하는 태그가 행(또는 열)에서 사용된 빈도수를 나타낸 값으로(식 (1) 참조) 행(또는 열)간 포맷팅 일관성의 값 차이를 이용

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	350

(a)

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	350

(b)

그림 5. 전처리 규칙 2와 4를 적용하여 정규화된 테이블의 예: (a) 전처리 전의 테이블, (b) 전처리 후의 테이블

표 2. 영역구분을 위한 태그

순번	HTML 태그	의 미
1	<Thead>	테이블의 머리말(속성)에 해당함
2	<Tbody>	테이블의 본문(값)에 해당함
3	<Tfoot>	테이블의 꼬리말에 해당함

하여 영역을 구분한다. 포맷팅 일관성을 위하여 사용된 태그는 <표 3>과 같다.

$$\text{포맷팅 일관성} = \frac{\text{시각적 특징을 나타내는 태그를 가진 셀의 수}}{\text{행(또는 열)의 전체 셀 수}} \quad (1)$$

한편 <그림 6>은 행과 열 모두 동일한 포맷팅 정보를 사용하기 때문에 포맷팅 일관성을 사용한 영역 구분이 어렵다. 이러한 테이블은 구문적 일관성 검사를 통하여 영역을 구분한다[20]. 구문적 일관성이란 타입 및 길이의 일관성을 나타낸 값(식 (2) 참조)으로 제안된 방법은 우하단 셀을 기준으로 아래에서 위(bottom-up)로 그리고 오른쪽에서 왼쪽(right-left) 방향으로 이웃 셀과 일관성을 계산한다.

$$\text{구문적 일관성} = \frac{\text{타입 및 길이 일관성을 갖는 셀의 수}}{\text{행(혹은 열)의 셀 수}} \quad (2)$$

표 3. 포맷팅 일관성 검사를 위한 주요 태그

종류	의미
<i>, , <var>, <cite>	글자를 이탤릭체로 표현함
, <b>	글자를 굵게 나타냄
<u>	밑줄 그은 글자체로 표현
<h1>~<h6>	글자 크기를 변경
<big>	기본 글꼴보다 한 단계 큼
<font color="">	글자 색을 설정
<font size="">	글자 크기를 설정
<font face="">	글꼴을 변경
<tr bgcolor="">, <td bgcolor="">	행 혹은 셀의 배경색을 설정

Class	Price
Residental/small non-residential	$0.05689/kWh
Meduim/non-residential	$0.05732/kWh
Large/non-residential	$0.0613/kWh

그림 6. 구문적 일관성 검사를 적용하여 영역구분이 가능한 테이블의 예

·구문적 일관성은 기준이 되는 셀에 이웃한 셀을 하나씩 추가하면서 계산된다. 예를 들어, <그림 6> 테이블의 열 방향 구문적 일관성은 '$0.0613/kWh'의 값을 갖는 셀과 세로방향으로 이웃한 셀, 즉, '$0.05732/kWh'의 값을 갖는 셀의 일관성을 계산하고, 이후 동일한 방법으로 이웃한 셀을 하나씩 추가하면서 계산된다. 이와 같이 계산된 각 단계의 일관성 값 차이는 속성과 값 영역을 구분하는 기준이 된다. 즉, 각 단계의 일관성 차이가 가장 큰 부분을 속성과 값 영역의 경계로 간주한다.

3.1.3 의미적 일관성 검사

시각적 일관성 검사에 의해 영역이 구분되지 않은 테이블과 2×n 열 방향 및 n×2 행 방향, n≥1, 테이블은 셀 데이터의 의미적 일관성을 사용하여 영역을 구분한다. 의미적 일관성 검사란 임의의 속성 값으로 올 수 있는 키워드 및 패턴 정보를 이용하여 대응하는 속성과 값이 의미적으로 부합하는지를 판단하는 검사로서 부합하는 속성과 값의 위치에 따라 속성과 값 영역을 구분한다. <표 4>는 제안된 방법에서 사용된 속성의 키워드와 가능한 값의 패턴 및 키워드이며, <그림 7>은 의미적 일관성 검사를 통하여 영역이 구분된 테이블의 예이다. <그림 7>에서 보면 'E-mail', 'Telephone', 그리고 'Web Site'의 값을 가진 셀은 <표 4>의 속성에 포함된 키워드에 포함되어 속성 영역으로 구분된다. 한편, 'citizenspark@ aiken.net', '(803)642-7760', 그리고 'www.@aiken.net'의 값을 가진 셀은 해당 속성에 대응되는 값의 패턴을 가지므로 제안된 방법에 의하여 값 영역으로 구분된다.

3.1.4 후처리

후처리 단계에서는 속성과 값이 단일 셀에 존재하는 혼합-셀 테이블의 영역구분과 함께 합성 테이블의 여부 판단 및 단순 테이블로의 분리 역할을 담당한다. 제안된 방법은 우선 혼합-셀 테이블의 여부를 판단하기 위하여 ':' 과 '=' 등의 구분자를 식별한다. 만약 테이블 내의 모든 셀에 구분자가 존재하며, 모든 셀의 구분자 좌우 데이터 타입이 각각 동일한 경우 제안된 방법은 해당 테이블을 혼합-셀 테이블로 간주하며, 구분자를 중심으로 속성 및 값 쌍을 추출한다.

한편 제안된 방법은 혼합-셀 테이블을 제외한 모든 테이블에 대하여 합성 테이블 여부를 판단한다.

표 4. 의미적 일관성 검사를 위한 키워드와 패턴 정보

분류	속성에 포함된 키워드	가능한 값의 패턴	가능한 값의 키워드
버전정보	Version	\d+(\.\d+)+	-
이메일	E-mail, email	\b[A-Z0-9._%-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b	-
전화번호	Telephone, Phone, TEL, FAX, Contact	(\+\d+\s)?(\(\d+\))?(\s)?((\d+)(\s|-|\.))?(\d+)(\s|-|\.)(\d{4})	-
날짜	period, Date, Create, Revise, Deadline, Start, End	(19|20)\d\d[-/.](0[1-9]|1[012])[-/.](0[1-9]|[12][0-9]|3[01]), (19|20)\d\d[-/.](Jan|January)[-/.](0[1-9]|[12][0-9]|3[01]),…, (19|20)\d\d[-/.](Dec|December)[-/.](0[1-9]|[12][0-9]|3[01])	-
시간	time	(0[1-9]|1[0-9]|2[0-4])[-/.:]([0-5][0-9])[-/.:]([0-5][0-9]), ((N|n)oon), ((M|m)idnight)	am, pm, hour(hr), minute(min), second(sec)
월	Month	(0[1-9]|1[012])	January(Jan), … , December(Dec)
일	Date, Day	(0[1-9]|[12][0-9]|3[01])	Monday(Mon),…, Sunday(Sun)
URI	Web Site, Online, URI	("http://"|"mailto:"|"ftp://")[^ \n\r\"\<\ \]+	http, www
운영체제	OS	-	Windows, NT, XP, UNIX, LINUX, Tru64
길이	length, height, width	-	cm, km, mile, ft, inch, yd
가격	Cost, Asset, Purchases, profit, value, Price	-	$, £, ¥, \, USD, CAD
확률	Humidity, Hum, Probability	-	%
직업	Company	-	Corp. Inc, Ltd.

이를 위하여 테이블의 추출된 속성 영역의 개수를 사용하는데, 속성 영역이 2개 이상인 경우에는 합성 테이블로 판단하여 이를 기준으로 다수의 단순 테이블로 분리한다. 만약 추출된 속성 영역이 1개인 테이블은 태그속성 스팬을 사용하여 복합 테이블의 여부를 판단하게 되고, 복합 테이블인 경우 스팬을 기준으로 다수의 단순 테이블로 분리한다. 이때, 단순 테이블이 속성 영역을 가지지 않는다면 합성 테이블의 속성 영역을 복사한다. 입력된 테이블이 1개의 속성 영역을 가지고, 복합 테이블이 아니라면 제안된 방법은 이를 단순 테이블로 간주한다. 제안된 합성 테이블의 영역구분 절차는 <그림 8>과 같다.

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

그림 7. 의미적 일관성 검사를 통한 영역구분의 예

3.2 구조분석

속성과 값 영역이 구분된 테이블은 구조분석 단계에서 XML 문서로 변환된다. 구조분석 단계에서는 테이블을 구조 및 속성의 위치에 따라 분류하고, 각각 테이블 모델을 적용하여 계층구조를 추출한 후 깊이우선 탐색과 병합규칙을 적용하여 XML 문서를 생성한다.

3.2.1 테이블 모델에 기반한 구조분석

테이블 모델에 기반한 구조분석 단계에서는 테이블의 구조 및 속성의 위치에 따라 열 방향, 2×n 열 방향, 행 방향, n×2 행 방향, 그리고 타임 테이블의 5가지로 분류한 후 각각 테이블 모델을 적용하여 계

```
입력 : 혼합-셀 테이블을 제외한 모든 테이블 (Table)
출력 : 속성과 값 영역이 구분된 다수의 단순 테이블
       (Simple table)

함수 및 변수 정의 :
Table[] SimpleTables
        ::= 다수의 단순 테이블 저장
Table[] SegmentedSimpleTables
        ::= 속성과 값 영역으로 구분된 다수의 단순 테
            이블 저장
Table[] DivideComplexTable()
        ::= 합성 테이블을 단순 테이블로 분리

방법:
1:  If (테이블의 추출된 속성 영역이 2개 이상) {
2:    SegmentedSimpleTables[] = DivideComplexTable();
3:    // 속성 영역을 기준으로 테이블 분리
4:  }
5:  else If (태그속성 스팬의 값이 행(혹은 열)의 수와
    같은 셀이 두 개 이상 존재) {
6:    // 입력 테이블을 복합 테이블로 간주
7:    SimpleTables[] = DivideComplexTable(); // 스팬
      을 기준으로 테이블 분리
8:    If (분리된 단순 테이블이 합성 테이블의 속성과
      같은 값을 가짐) {
9:      SegmentedSimpleTables[] = SimpleTables[];
10:   }
11:   else {
12:     SegmentedSimpleTables[] = 합성 테이블의 속성
        을 단순 테이블에 복사;
13:   }
14: }
15: else {
16:     SegmentedSimpleTables[] = Table;
17: }
```

그림 8. 합성 테이블의 영역구분 절차

층구조를 추출한다. 테이블 모델(table model)이란 셀들이 갖는 계층구조를 도식화한 것으로서, 속성 $H_1,\cdots,H_n$ 과 값 $(D_{1,1},\cdots,D_{1,n}),\cdots,(D_{m,1},\cdots,D_{m,n})$을 가진다. <그림 9>는 테이블 모델과 계층구조를 나타낸다.

제안된 방법은 입력된 테이블을 5가지로 분류한 후 적절한 테이블 모델을 적용하여 속성-값 연관관계 및 계층구조를 추출한다. 특히, 속성 영역을 나타내는 $H_1,\cdots,H_n$은 주 속성 및 부 속성으로 표현이 가능하다. <그림 9(a)>와 <그림 9(b)>는 열 방향 테이블의 테이블 모델과 계층구조를 나타낸다. <그림 9(a)>는 일반적인 열 방향 테이블의 테이블 모델로서, 값 영역이 2개 이상 존재하는 테이블의 테이블 모델이다. 이러한 테이블의 값 영역은 가로방향으로 의미적인 연관관계를 갖는데, 이를 표현하기 위한 적절한 계층구조를 갖는다. 반면 열 방향이면서 2×n의 형태를 가진 테이블의 값 영역은 서로 연관관계를 갖지 않으며 <그림 9(a)>와는 다른 계층구조를 갖는다(<그림 9(b)> 참조). <그림 9(c)>와 <그림 9(d)>는 행 방향 테이블의 테이블 모델로서 각각 열 방향 테이블과 동일한 형태를 가진다. 한편 <그림 9(e)>와 같이 타임 테이블의 테이블 모델은 열 방향과 행 방향의 테이블 모델을 합한 것과 같다. 즉, 하나의 셀 데이터는 2개의 계층구조를 가지게 되는데, 이는 정보 추출을 위한 질의 시 기준 속성을 어디에 두는지에 따라 계층구조가 달라지기 때문이다. <그림 11>은 <그림 10>의 열 방향 입력 테이블에 대한 테이블 모델 적용의 예이다. 'TODAYS'와 'YESTERDAYS'의 값을 갖는 셀은 주 속성으로, 'OPEN'과 'CHANGE'의 값을 갖는 셀은 부 속성으로 간주되어 적절한 계층구조가 생성된다.

3.2.2 XML 변환

구조가 분석된 테이블은 XML 변환단계에서 XML 문서로 변환된다. XML 문서는 적격성(well-formed) 요건, 즉, XML 선언 및 모든 태그가 중첩되지 않아야 하는 등의 조건,을 만족해야 하므로, 이를 위하여 제안된 방법은 깊이우선 탐색(depth first search) 기법을 사용한다. 우선, HTML 테이블을 XML 문서로 변환하기 위해서 XML 선언부를 정의한다. XML 문서의 선언부는 "<?xml ?>" 형태를 가지는 문장으로 해당 문서의 버전 및 인코딩 방식 등을 설명하며, 항상 문두에 위치한다. XML 문서를 선언한 다음에는 생성된 계층구조를 깊이우선 탐색하여 적절한 XML 요소를 생성한다. 이때 스택을 사용하여 적절한 시작 태그와 끝 태그를 생성함으로서 적격성 요건을 만족하도록 한다. 또한 생성된 XML 문서의 불필요한 요소의 중복 제거를 위하여 Li 등 [3-6] 이 제안한 병합규칙을 사용한다.

<그림 12>는 <그림 10>의 테이블을 XML 문서로 변환한 예이다. <그림 12(a)>는 구조분석 단계에서 생성된 계층구조에 스택을 적용한 깊이우선 탐색 결과로 중복태그가 존재한다. 이를 제거하기 위하여 제안된 방법은 병합규칙을 적용하며, 그 결과 적격성 요건의 만족과 동시에 문서를 간결화 한다. <그림 12(b)>는 생성된 최종 XML 문서의 예이다.

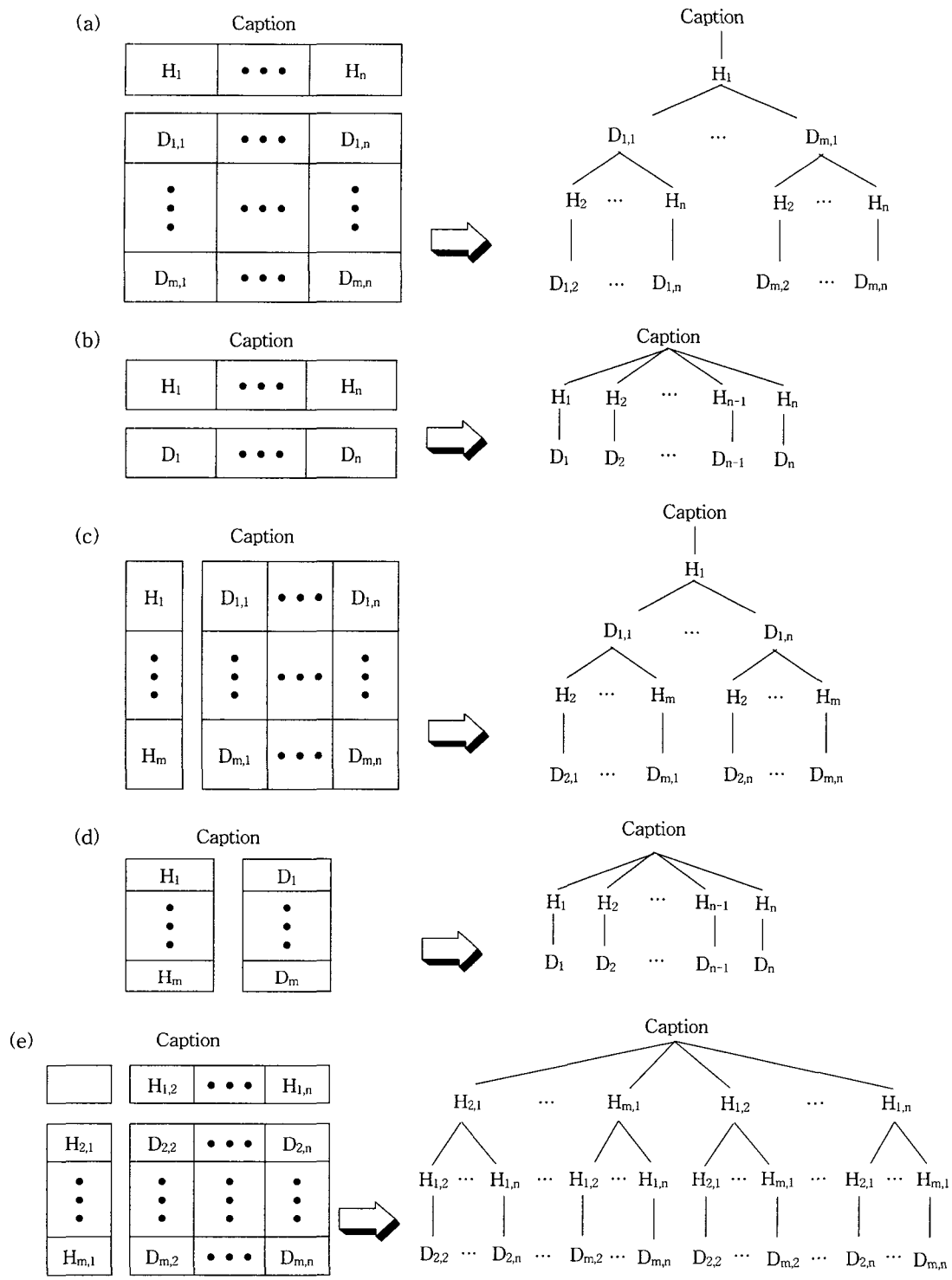

그림 9. 테이블의 분류에 따른 테이블 모델 및 계층구조: (a) 열 방향 테이블의 테이블 모델 및 계층구조, (b) 2×n 열 방향 테이블의 테이블 모델 및 계층구조, (c) 행 방향 테이블의 테이블 모델 및 계층구조, (d) n×2 행 방향 테이블의 테이블 모델 및 계층구조, (e) 타임 테이블의 테이블 모델 및 계층구조

COMPANY	TODAYS		YESTERDAYS	
	OPEN	CHANGE	OPEN	CHANGE
BLUE INC	75 1/2	+1 1/8	74 9/16	-4 1/4
GREEN COM	89 1/4	+2	88 5/8	-2 13/16
RED INC	22 1/4	+5/16	21 13/16	-3/8
YELLOW LTD	103 3/8	-1 13/16	101	-4
PURPLE INC	27 11/16	-2 5/8	27 5/8	-1 1/8
BROWN COM	68	+11/16	66 11/16	-1 5/8
PINK LTD	130 7/16	+1 1/16	130	-2 3/8

그림 10. 입력 테이블의 예

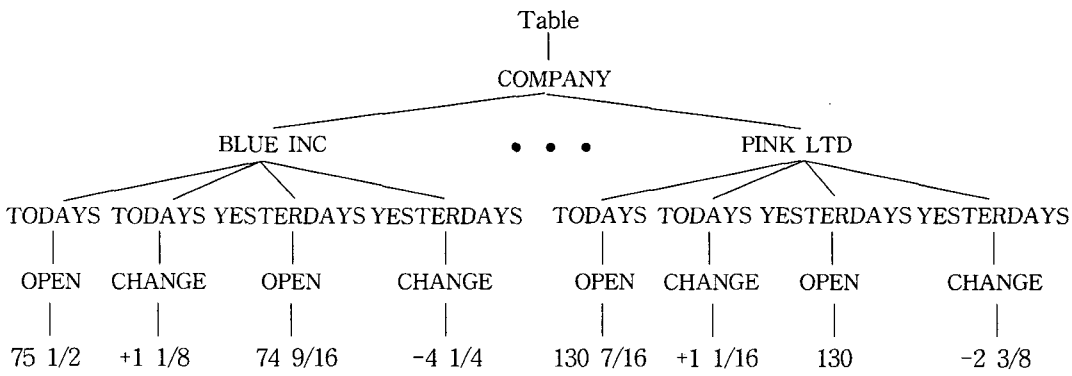

그림 11. 〈그림 10〉 입력 테이블의 계층구조

4. 실험 결과

제안된 방법은 테이블을 열 방향, 행 방향, 타임, 그리고 합성 테이블로 분류하여 각각 실험하였다. 제안된 방법의 성능 평가기준은 <표 5>와 같다.

4.1 성능 평가

제안된 방법의 성능을 평가하기 위하여 Wang가 Hu [21]의 연구에서 사용한 11,477개의 테이블 중 1,180개의 진짜 테이블을 대상으로 실험하였다. 제안된 방법의 성능을 정량적으로 평가한 결과는 <표 6>과 같다.

제안된 방법은 <표 6>과 같이 86.7%의 정확도를 보여 우수한 결과를 보였다. 이는 본 논문이 테이블의 구조를 분석하기 위하여 보다 체계적이며 정교한 방법에 기반하기 때문이다. 제안된 방법은 먼저 전처리 단계로서 잡음 영역을 제거하고 테이블을 정규화한다. 또한 시각적 일관성은 물론이고 일관성 검사가 어려운 테이블에 대하여 의미적 일관성을 검사함으로서 보다 정교한 영역구분이 가능하다. 예를 들어, 크기가 2×2인 테이블의 경우 규칙에 기반한 기존 연구들의 방법을 통해서는 해결이 불가능한데 반하여 제안된 방법은 속성과 값 영역 사이의 의미적 일관성의 유무를 검사함으로서 영역구분이 가능하다.

표 5. 성능 평가기준

기준	정 의
정확도	정확하게 영역이 구분된 테이블의 수 / 진짜 테이블의 총 수

표 6. 성능 평가

구분	테이블 개수	정답 개수	오류 개수	정확도
열 방향	857	788	69	91.59%
행 방향	143	117	26	81.82%
타임	18	18	0	100%
합성	162	100	62	61.73%
합계	1180	1023	157	86.69%

```
<BLUEINC>
  <TODAYS><OPEN>75 1/2</OPEN></TODAYS>
  <TODAYS><CHANGE>+1 1/8</CHANGE></TODAYS>
  <YESTERDAYS><OPEN>74 9/16</OPEN></YESTERDAYS>
  <YESTERDAYS><CHANGE>-4 1/4</CHANGE></YESTERDAYS>
</BLUEINC>
```

⇩

```
<BLUEINC>
  <TODAYS><OPEN>75 1/2</OPEN><CHANGE>+1 1/8</CHANGE></TODAYS>
  <YESTERDAYS><OPEN>74 9/16</OPEN><CHANGE>-4 1/4</CHANGE></YESTERDAYS>
</BLUEINC>
```

(a)

```
<?xml version="1.0" encoding="UTF-8" ?>
<table>
  <COMPANY>
    <BLUEINC>
      <TODAYS>
        <OPEN>75 1/2</OPEN>
        <CHANGE>+1 1/8</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>74 9/16</OPEN>
        <CHANGE>-4 1/4</CHANGE>
      </YESTERDAYS>
    </BLUEINC>
    <GREENCOM>
      <TODAYS>
        <OPEN>89 1/4</OPEN>
        <CHANGE>+2</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>88 5/8</OPEN>
        <CHANGE>-2 13/16</CHANGE>
      </YESTERDAYS>
    </GREENCOM>
    <REDINC>
      <TODAYS>
        <OPEN>22 1/4</OPEN>
        <CHANGE>+5/16</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>21 13/16</OPEN>
        <CHANGE>-3/8</CHANGE>
      </YESTERDAYS>
    </REDINC>
    <YELLOWLTD>
      <TODAYS>
        <OPEN>103 3/8</OPEN>
        <CHANGE>-1 13/16</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>101</OPEN>
        <CHANGE>-4</CHANGE>
      </YESTERDAYS>
    </YELLOWLTD>
    <PURPLEINC>
      <TODAYS>
        <OPEN>27 11/16</OPEN>
        <CHANGE>-2 5/8</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>27 5/8</OPEN>
        <CHANGE>-1 1/8</CHANGE>
      </YESTERDAYS>
    </PURPLEINC>
    <BROWNCOM>
      <TODAYS>
        <OPEN>68</OPEN>
        <CHANGE>+11/16</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>66 11/16</OPEN>
        <CHANGE>-1 5/8</CHANGE>
      </YESTERDAYS>
    </BROWNCOM>
    <PINKLTD>
      <TODAYS>
        <OPEN>130 7/16</OPEN>
        <CHANGE>+1 1/16</CHANGE>
      </TODAYS>
      <YESTERDAYS>
        <OPEN>130</OPEN>
        <CHANGE>-2 3/8</CHANGE>
      </YESTERDAYS>
    </PINKLTD>
  </COMPANY>
</table>
```

(b)

그림 12. 병합규칙 적용사례 및 생성된 XML 문서: (a) 병합규칙 적용사례, (b) 병합규칙을 적용하여 최종 생성된 XML 문서의 예

표 7. 오류분석 결과

구분	오류내용	개수(%)
속성 → 값	포맷팅 일관성이 없음	3 (2%)
값 → 속성	포맷팅 일관성이 없음	125 (80%)
	구문적 일관성이 없음	3(2%)
구분불가	합성 테이블에서 각 테이블의 속성이 모두 다름	8 (5%)
	2×n, n×2, 2×2 테이블의 시각적, 의미적 일관성이 없음	8 (5%)
	합성 테이블로서 다양한 형태를 가진 테이블이 섞여있음	8 (5%)
	비정상적인 테이블 편집으로 인한 오류	2 (1%)
합계		157 (100%)

	U. S. President						
School	Nader Laduke	Browne Olivier	Buchanan Foster	Bush Cheney	Gore Lieberman	Hagelin Goldhaber	Phillips Frazier
Abbott Loop Elementary				52	38		
Airport Heights Elementary	22	8	17	99	54	8	7
Alpenglow Elementary	2		5	155	37		4
Anch 7th Day Adv Jr Academy				12	4	1	
Anchor Lutheran School	8	2	6	77	12	3	4

그림 13. 속성 영역을 값 영역으로 잘못 식별 한 예

실험 결과, 제안된 방법의 영역구분 오류는 <표 7>과 같으며 각각에 대한 자세한 설명은 다음과 같다. <그림 13>, <그림 14>, <그림 15>는 제안된 방법이 테이블의 구조를 정확하게 인식하지 못한 테이블의 예이다. <그림 13>은 속성 영역의 포맷팅 일관성이 존재하지 않아서 잘못 구분된 경우이다. 이러한 경우 첫 번째 행과 두 번째 행의 포맷팅 정보가 같다면 시각적 일관성으로 영역구분이 가능하다. 또한 <그림 14>는 값 영역을 속성 영역으로 잘못 구분한 예로서, <그림 14(a)>는 <그림 13>과 같이 값 영역이 속성 영역과 동일한 포맷팅 정보를 가져 잘못 구분된 경우이고, <그림 14(b)>는 값 영역의 구문적 일관성이 존재하지 않아서 속성 영역으로 잘못 구분된 예이다.

Seq. #	Contract Month	Product Code	First Trade Date	Last Trade Date	Settlement Date	Delete Date
1	Jun 2001	EBM1	5/22/00	5/18/01	5/18/01	5/21/01
2	Jul 2001	EBN1	3/26/01	6/22/01	6/22/01	6/25/01
3	Aug 2001	EBQ1	4/23/01	7/20/01	7/20/01	7/23/01
4	Sep 2001	EBU1	8/21/00	8/24/01	8/24/01	8/27/01
5	Dec 2001	EBZ1	11/20/00	11/23/01	11/23/01	11/26/01
6	Mar 2002	EBH2	2/20/01	2/15/02	2/15/02	2/19/02

(a)

Date of Forecast	Medium in millions	
	2025	2050
1980	NoCount	NoCount
1982	8177	
1990	8504	
1994	8294	
Feb.1998	7900	9400
Oct.1998	7800	8900

(b)

그림 14. 값 영역을 속성 영역으로 잘못 식별 한 예: (a) 속성과 값 영역의 포맷팅 일관성 존재, (b) 값 영역의 구문적 일관성 부재

Pass Rate (%) Comparison			
Oct-2000			
Examination	1st Time Takers	Repeat Takers	Overall
Chemical	64	28	48
Civil	56	33	44
Electrical	41	18	28
Environmental	77	59	72
Mechanical	55	27	42
Structural I	65	35	55
Agricultural	47	42	45
Control Systems	86	59	81
Fire Protection	42	23	35
Industrial	56	32	46
Manufacturing	79	67	73
Metallurgical	89	80	86
Mining/Mineral	56	44	50
Nuclear	100	100	100
Petroleum	71	41	59
Naval Architecture/ Marine Engineering	41	NA	41
(last given October 1999)			
Examination	AM Only	PM Only	Total Exam
Structural II	38	33	20

(a)

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

(b)

8:00	REGISTRATION ? CONTINENTAL BREAKFAST ? EXHIBIT HALLS OPEN		
9:00	MEDICARE TOPIC TO BE ANNOUNCED Presented by Empire Medicare Services ? Medicare Part B Speaker: Jim Barnes, Manager ? EMC Marketing		
	TRACK A	TRACK B	TRACK C	TRACK D
10:00	Exhibitor workshop to be announced	Exhibitor workshop to be announced	Exhibitor workshop to be announced	Exhibitor workshop to be announced
10:30	COFFEE BREAK	COFFEE BREAK	COFFEE BREAK	COFFEE BREAK
11:00	Exhibitor workshop to be announced	Exhibitor workshop to be announced	Exhibitor workshop to be announced	Exhibitor workshop to be announced
11:30	Exhibitor workshop to be announced	Exhibitor workshop to be announced	Exhibitor workshop to be announced	Exhibitor workshop to be announced
12:00	LUNCH	LUNCH	LUNCH	LUNCH
2:00	CMC Program Topics to be announced (Members of Medicare of Genesee County)	Exhibitor workshop to be announced	CEO Program Topics to be announced (PHYSICIAN)	Exhibitor workshop to be announced
2:30		Exhibitor workshop to be announced		Exhibitor workshop to be announced
3:00	COFFEE BREAK	COFFEE BREAK	COFFEE BREAK	COFFEE BREAK
3:30	CMC (Physicians)	Exhibitor workshop to be announced	CEO (Hospitals)	Exhibitor workshop to be announced
4:00		Exhibitor workshop to be announced		Exhibitor workshop to be announced
4:30		Exhibitor workshop to be announced		Exhibitor workshop to be announced

(d)

Table 1

Summary of the *Energy−10* Weather File for Portland, Maine (portland.et1)

Latitude: 43.7 | Winter Design Day Dry Bulb (2.5%): -18.4 &degC

Longitude: 70.3 | Summer Design Day Dry Bulb (97.5%): 28.9 &degC

Elevation: 13 m | Summer Design Day Wet Bulb (97.5%): 21.7 &degC

Month	TAA	TMXA	TMNA	TMX	TMN	TWBA	RH	WSA	HS	HDD	CDD
January	-5.4	-0.7	-11.0	11.1	-24.4	-6.9	63.4	14.5	1896	740	0
February	-4.5	-0.0	-10.2	11.1	-22.8	-6.0	65.5	13.0	3008	647	0
March	0.4	4.7	-4.5	10.6	-19.4	-1.7	62.6	13.9	4001	556	0
April	7.4	12.3	2.1	27.8	-2.8	4.7	69.4	16.5	4855	327	2
May	11.3	16.3	5.9	28.7	1.7	8.9	75.5	16.5	5636	214	0
June	17.3	23.4	10.9	32.2	2.2	14.0	73.4	14.3	5051	46	20
July	20.7	26.4	15.4	33.3	10.0	17.3	74.4	11.1	6177	9	98
August	19.8	24.8	14.9	30.0	8.3	16.7	75.4	13.9	5357	13	71
September	15.1	20.2	9.3	26.7	2.8	13.0	80.9	14.2	4286	102	6
October	8.7	14.4	3.0	25.6	-4.4	6.3	71.5	13.9	2931	288	0
November	4.0	8.2	-0.4	22.8	-8.9	2.1	72.0	14.7	1836	422	0
December	-3.5	-0.0	-7.7	8.9	-22.8	-5.1	64.6	13.5	1515	678	0
Year	7.6	12.5	2.3	33.3	-24.4	5.3	70.7	14.2	3963	4039	197

TAA: Avg Dry Bulb Temp, &degC

TMXA: Avg Daily Max Dry Bulb Temp, &degC

TMNA: Avg Daily Min Dry Bulb Temp, &degC

TMX: Maximum Dry Bulb Temperature, &degC

TMN: Minimum Dry Bulb Temperature, &degC

TWBA: Average Wet Bulb Temperature, &degC

WSA: Avg Wind Speed, kmph

HS: Avg Daily Horizontal Solar Radiation, Wh/m2

RH: Relative Humidity, %

HDD: Heating Degree Days, Base 18.0 &degC

CDD: Cooling Degree Days, Base 18.0 &degC

(c)

그림 15. 속성과 값 영역을 구분하지 못한 테이블의 예: (a) 각 테이블의 속성이 모두 다름, (b) 2×n 테이블로서 의미적 일관성 부재, (c) 합성 테이블로서 영역구분 불가, (d) 잘못된 편집으로 인한 오류

한편 <그림 15>는 속성과 값 영역을 구분하지 못한 테이블의 예이다. <그림 15(a)>는 다수의 테이블이 각각 다른 속성을 가져 영역을 구분하지 못한 경우이다. 또한 <그림 15(b)>는 테이블의 크기가 2×2로 의미적 일관성 부재로 인하여 영역을 구분하지 못한 경우이다. 그러나 만약 속성과 값의 키워드로 각각 "document type"과 "interview"를 추가 한다면 영역 구분이 가능하다. <그림 15(c)>와 <그림 15(d)>는 각각 형태가 다른 다수의 테이블이 섞여있는 합성 테이블과 잘못된 편집으로 인한 테이블로 제안된 방법으로는 영역구분이 불가능하다.

4.2 기존 연구와의 비교

Jung 등은 <표 8>과 같이 영역구분을 위하여 서로 다른 두 개의 성능평가 기준을 적용하였다. 본 연구에서는 Jung 등의 실험 데이터를 입수할 수 없었기 때문에 <표 8>과 같이 동일한 성능평가 기준에 의하여 간접적으로 비교하였다.

<표 9>는 테이블을 기준으로, <표 10>은 속성 셀을 기준으로 성능을 비교한 값이다. <표 9>와 <표 10>과 같이 제안된 방법은 동일한 기준을 적용한 Jung 등의 방법보다 우수한 결과를 보였다. 이는 속성 및 값 영역의 구분을 위하여 Jung 등은 포맷팅 정보에 기반한 휴리스틱한 규칙을 사용한데 반하여 제안된 방법은 잡음 영역을 제거하고 테이블을 정규화한 후 보다 정교한 시각적 및 의미적 일관성 검사를 실시하였기 때문이다. 또한 Jung 등의 방법으로 영역구분이 어려운 합성 테이블에 대해서도 제안된 방법은 후처리를 적용함으로서 영역구분이 가능하였기 때문이다.

표 8. Jung 등의 성능 평가 기준

대상	기준	정의
테이블	정확도	본 논문과 동일
셀	정확도	시스템이 속성으로 판단한 셀 중 진짜 속성 셀의 수 / 시스템이 속성으로 판단한 셀의 수
	재현률	시스템이 속성으로 판단한 셀 중 진짜 속성 셀의 수 / 진짜 속성 셀의 수

5. 결론 및 향후 연구방향

최근 들어 웹을 통하여 새롭게 생성되는 정보의 양이 급속도로 증가하면서 웹으로부터 유용한 정보를 추출하는데 많은 관심이 모아지고 있다. 특히 테이블은 연관된 정보를 효과적으로 표현할 수 있는 대표적인 방법으로서 폭넓게 사용된다. 한편, HTML은 문서를 시각적으로 랜더링 하기위한 용도로 제안된

표 9. 제안된 방법과 Jung 등의 실험결과 비교 (테이블 기준)

제안된 방법					Jung 등의 방법			
종류	테이블 개수	정답 개수	오류 개수	정확도	테이블 개수	정답 개수	오류 개수	정확도
열	857	788	69	91.59%	2,565	2,105	551	82.1%
행	143	117	26	81.82%				
타임	18	18	0	100%				
합성	162	100	62	61.73%				
합계	1180	1023	157	86.69%				

표 10. 제안된 방법과 Jung 등의 실험결과 비교 (셀 기준)

제안된 방법						Jung 등의 방법	
종류	속성 셀의 총수	시스템이 찾은 수	찾은 것 중 맞는 수	P	R	P	R
열	5,047	5,685	4,900	86.19%	97.09%	86.2%	88.4%
행	1,542	1,475	1,367	92.67%	88.65%		
타임	364	364	364	100%	100%		
합성	2,619	2,194	1,953	89.02%	74.57%		
합계	0	0	0	88.33%	89.68%		

포맷이기 때문에 컴퓨터로 하여금 유용한 정보를 추출 및 재가공 하기에는 부적합하다. 이를 위하여 논리적인 구조정보를 표현할 수 있는 XML 문서로의 변환이 필수적이다. 본 논문에서는 HTML 문서에 포함된 진짜 테이블의 영역을 구분하고 구조를 분석하여 XML 문서로 변환하는 효과적인 방법을 제안하였다.

제안된 방법은 HTML 테이블의 XML 변환을 위하여 영역구분과 구조분석의 두 단계로 구성된다. 먼저 속성과 값 영역을 구분하는 영역구분 단계는 잡음 영역을 제거하고 태그속성 스팬을 정규화하는 등의 전처리 과정을 거친다. 정규화된 테이블은 이후 시각적인 일관성 검사를 통하여 영역이 구분된다. 제안된 방법은 이를 위하여 포맷팅 일관성과 구문적 일관성을 제안하였다. 한편, 시각적 일관성이 없거나 혹은 검사가 불가능한 테이블은 다시 의미적인 일관성 검사를 통하여 영역을 구분한다. 이는 속성과 값이 의미적으로 부합한다는 사실에 기반한다. 또한 비정상적으로 편집된 테이블과 합성 테이블의 영역구분을 위하여 후처리 단계를 거친다. 실험 결과, 제안된 방법은 기존 연구와 비교하여 우수한 성능을 보였다. 특히 기존 연구와는 달리 휴리스틱한 규칙에 기반하지 않고, 다양한 테이블에 적용이 가능하기 때문에 효과적이다.

향후 본 연구에서는 기 구축된 온톨로지 정보나 WordNet 등을 활용하여 보다 정교한 영역을 구분함과 동시에 정보의 재사용 및 자동 온톨로지 구축, 정보 검색 시스템 등의 다양한 응용분야에의 활용에 대하여 연구를 진행 할 예정이다.

참 고 문 헌

[1] S.W. Jung and H.C. 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.

[2] A. Pivk, P. Cimiano, and Y. Sure, "From Tables to Frames," *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 3, Issue 2-3, pp. 132-146, 2005.

[3] S. Li, M. Liu, G. Wang, and Z. Peng, "Capturing Semantic Hierarchies to Perform Meaningful Integration in HTML Tables," *Proc. The Asia Pacific Web Conference*, pp. 899-902, 2004.

[4] S. Li, Z. Peng, and M. Liu, "Extraction and Integration Information in HTML Tables," *Proc. Fourth Int'l Conf. Computer and Information Technology*, pp. 315-320, 2004.

[5] S. Li, M. Liu, T.W. Ling, and Z. Peng, "Automatic HTML to XML Conversion," *Proc. Fifth Int'l Conf. Web-Age Information Management*, pp. 714-719, 2004.

[6] S. Li, M. Liu, and Z. Peng, "Wrapping HTML Tables into XML," *Proc. Fifth Int'l Conf. Web Information Systems Engineering*, pp. 147-152, 2004.

[7] H. Masuda, S. Tsukamoto, and H. Nakagawa, "Recognition of HTML Table Structure," *Proc. First Int'l Joint Conf. Natural Language Processing*, pp. 183-188, 2004.

[8] K. Itai, A. Takasu, and J. Adachi, "Information Extraction from HTML Pages and Its Integration," *Proc. Int'l Symposium on Applications and the Internet*, pp. 276-281, 2003.

[9] Y. Yang and W.S. Luk, "A Framework for Web Table Mining," *Proc. Fourth Int'l Workshop on Web Information and Data Management*, pp. 36-42, 2002.

[10] M. Yoshida, "Extracting Attributes and Their Values from Web Pages," *Proc. The ACL-02 Student Research Workshop*, pp. 72-77, 2002.

[11] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Evaluating the Performance of Table Processing Algorithms," *Int'l Journal on Document Analysis and Recognition*, Vol. 4, No. 3, pp. 140-153, 2002.

[12] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Experiments in Table Recognition," *Second Int'l Workshop on Document Layout Interpretation and its Applications*, 2001.

[13] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Table Structure Recognition and Its Evaluation," *Proc. SPIE Document recognition and*

Retrieval Ⅷ, Vol. 4307, pp. 44-55, 2001.

[14] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "A System for Understanding and Reformulating Tables," *Proc. Fourth IAPR Int'l Workshop on Document Analysis Systems*, pp. 361-372, 2000.

[15] H. Masuda, D. Yasutomi, and H. Nakagawa, "How to Transform Tables in HTML for Displaying on Mobile Terminals," *Proc. 6th NLPRS Workshop of Automatic Paraphrasing: Theories and Applications*, pp. 29-36, 2001.

[16] S.J. Lim and Y.K. Ng, "A Heuristic Approach for Converting HTML Documents to XML Documents," *Proc. First Int'l Conf. Computational Logic*, pp. 1182-1196, 2000.

[17] S.J. Lim and Y.K. Ng, "An Automated Approach for Retrieving Hierarchical Data from HTML Tables," *Proc. 8th Int'l Conf. Information and Knowledge Management*, pp. 466-474, 1999.

[18] H.L. Wang, S.H. Wu, I.C. Wang, C.L. Sung, W.L. Hsu, and W.K. Shih, "Semantic Search on Internet Tabular Information Extraction for Answering Queries," *Proc. Ninth Int'l Conf. Information and Knowledge Management*, pp. 243-249, 2000.

[19] H.H. Chen, S.C. Tsai, and J.H. Tsai, "Mining Tables from Large Scale HTML Texts," *Proc. 18th Int'l Conf. Computational Linguistics*, pp. 166-172, 2000.

[20] Y.S. Kim and K.H. Lee, "Detecting Tables in Web Documents," *Engineering Applications of Artificial Intelligence*, Vol. 18, No. 6, pp. 745-757, 2005.

[21] 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, 2002.

김 연 석

2003년 명지대학교 전자정보통신공학부 졸업(학사)
2005년 연세대학교 컴퓨터과학과 졸업(석사)
2006년~현재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 웹문서 분석, 정보 추출 및 통합, XML, 모바일 웹 서비스

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 부교수
관심분야 : 멀티미디어 문서처리, XML, 웹 서비스