# Table parsing을 이용한 정보검색시스템의 효율향상

김영순, 권혁철
부산대학교 정보시스템공학과

# Implementation of Information Retrieval System by Table-parsing

Young-Sun Kim, Hyuk-Chul Kwon
Dept. of Computer Engineering, Pusan Nat'l University
E-mail : admin, hckwon@pusan.ac.kr

요 약

인터넷 문서에서 구조정보의 대표적인 예라 할 수 있는 표(table)는 의미있는 정보를 가지고 있는 경우가 많다. 하지만 인터넷상의 표는 여러 가지 형태이며, 이것에 맞게 표를 효과적으로 parsing하는 방법이 필요하다. 이렇게 parsing한 표의 정보를 이용하여, 인터넷 문서, 특히 전자상거래 문서에 있는 표를 표준화한 틀에 따라 개념화하여, 의미있는 정보를 추출해 낼 수 있다.

## 1. 서론

인터넷 문서는 문서 내에 포함되어 있는 문장에 대한 색인 후 나오는 단어에 대한 정보 뿐만 아니라 HTML 문서의 구조에 따른 특성도 유용한 정보가 될 수 있다. 예를 들어 시각화의 효과를 위해 웹 상에 정리된 방식으로 보여주는 표(table)라든지, 하이퍼링크의 특성을 이용한 링크정보, 표(table)에 대한 정보, 제목, 파일이름과 디렉토리 이름 따위의 주변 정보를 분석하여 색인어를 추출하여 사용할 수 있다.

일반적으로 표는 도식화된 정보를 표현하고 제공하는 데 유용한 수단이 되며 표의 왼쪽 첫 열과 위쪽 첫 행은 의미를 가지고 가로, 세로가 만나는 지점에 정보를 저장하는 경우가 많다. 이러한 표의 특성 때문에 과거처럼 단순히 문장을 색인하는 것으로는 모자라는 경우가 많다.

본 연구는 정보통신부에서 시행한 우수신기술 지정 지원사업의 결과임.

또한 표 안의 내용은 단순 문서 색인에서는 별로 의미가 없는 숫자나 특정 기호일 때도 있다.

이와 동시에, 유용한 정보를 표(table)를 이용하여 나타내는 홈페이지나 사이트들이 많이 존재하며, 나날이 늘어가고 있는 추세이다.

특히 전자상거래 사이트와 같은 경우는 표준화된 표를 이용하는 경우가 많다. 이런 사이트에 대한 개념화된 정보를 이끌어 내기 위해서는 테이블에 포함되어 있는 정보의 의미를 알아내는 것이 필요하며, 이것을 위해서 table parsing은 필수적이다.

## 2. Table parsing

HTML 문서에서 표는 그 일부분이며 표가 있는 Web Page 주변의 정보를 이용하면 표에 대한 정확한 정보를 얻을 수 있다. 이는 단순히 표의 주변 정보를 색인기를 이용하여 분석하는 방법에서부터 표의 의미적 특성을 활용하기 위하여 전문 분야에 대한 전문 용어 사전 구축, 표의 특성을 반영한 불용어 사전 구축 등을 적용한다.

인터넷상에서 존재하는 표는 일반적인 표의 형식과

똑같다. 하지만 내부 구조는 HTML tag로 존재하므로 HTML tag에 대한 특성 파악이 필요하다. 더불어 표의 주변 정보를 수집하는 것도 표의 내용을 파악하는 데 도움이 된다. 표의 주변 정보는 표가 HTML 문서에서 차지하는 비중, HTML 문서의 서두에 나오는 제목으로 생각되는 문장, table의 caption tag, table 바로 다음에 나와서 table의 설명 문구로 생각되는 문장, 파일이름, 디렉토리 이름 등이 있다. 하지만 이러한 정보들은 각각의 문서 양식에 따라 다양한 형식으로 유지되어 있다. 그러므로 표를 사용하는 다양한 문서에서 특징을 파악하여 표의 형식을 분석하는 것이 가장 중요한 부분이다. 전자성거래의 문서를 예로 들었을 때, 전체적으로 공통적인 문서의 특징이 있으며, 가격, 제품명 등 공통적인 항목들이 존재한다. 이런 공통적인 표의 구조와 항목에 대해서 공통적인 틀을 구성할 수가 있다. 표의 이러한 공통정보에 얼마나 일치하는가 하는 것이 표의 또하나의 특성이 되고, 분석하는 방법이 된다. 이런 공통적인 형식과 내용을 제외한 나머지는 다시 예외적인 형식과 내용으로 다른 처리 기법을 사용하게 된다. 표에 대한 특성 분석은 개개의 문서에서 전체적인 특성으로 통합해서 분석하고, 다시 개개의 특성을 다시 분석할 수 있게 분석해 나간다.

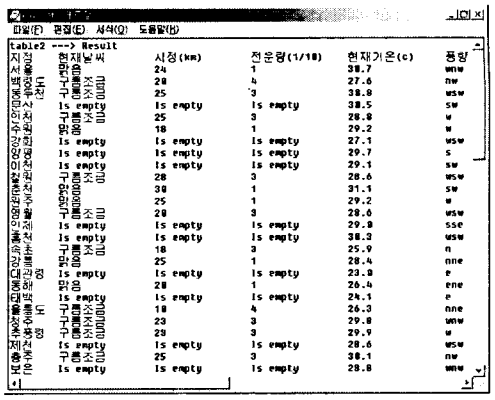
◆ 정형화된 표 parsing

정형화되었다는 말은 우리가 분석한 공통된 특성에 일치하는 표를 말한다. 정형화된 표는 기본적인 방법으로 HTML tag의 분석을 통해서 표의 구조에 맞는 항목을 얻어오고, 그 항목별로 정보들을 추출해서 2차원적인 배열의 형식으로 저장한다. 다음 [그림 1][그림 2]는 정형화된 표를 parsing한 예이다.

전국현재날씨 2001년 8월 24일 17시

| 지점 | 현재날씨 | 시정 (km) | 전운량 (1/10) | 현재기온 (C) | 풍향 | 풍속 (m/s) | 습도 (%) | 불쾌지수 | 일강수량 (mm) | 해면기압 (hPa) |
|---|---|---|---|---|---|---|---|---|---|---|
| 서울 | 맑음 | 24 | 1 | 30.7 | WNW | 2.0 | | | | |
| 백령도 | 구름조금 | 20 | 4 | 27.6 | NW | 2.0 | | | | |
| 울주간 | 구름조금 | 25 | 3 | 30.8 | WSW | 2.0 | | | | |
| 부산 | | | | 30.5 | SW | 1.0 | | | 0.1 | |
| 인천 | 구름조금 | 25 | 3 | 28.8 | W | 4.5 | | | | |
| 수원 | 맑음 | 18 | 1 | 29.2 | W | 3.0 | | | | |
| 강화 | | | | 27.1 | WSW | 3.5 | | | | |
| 양평 | | | | 29.7 | S | 1.0 | | | | |
| 이천 | | | | 29.1 | SW | 2.0 | | | | |
| 철원 | 구름조금 | 28 | 3 | 28.6 | WSW | 2.0 | | | | |
| 춘천 | 맑음 | 30 | 1 | 31.1 | SW | 1.5 | | | | |
| 원주 | 맑음 | 25 | 1 | 29.2 | W | 1.5 | | | | |
| 영월 | 구름조금 | 20 | 3 | 28.6 | WSW | 3.0 | | | | |
| 인제 | | | | 29.0 | SSE | 2.0 | | | | |
| 홍천 | | | | 30.3 | WSW | 0.5 | | | | |
| 속초 | 구름조금 | 18 | 3 | 25.9 | N | 1.5 | | | | |
| 강릉 | 맑음 | 25 | 1 | 28.4 | NNE | 2.0 | | | | |
| 대관령 | | | | 23.0 | E | 1.0 | | | | |

[그림 1] 정형화 된 표의 예 - 기상청 홈페이지

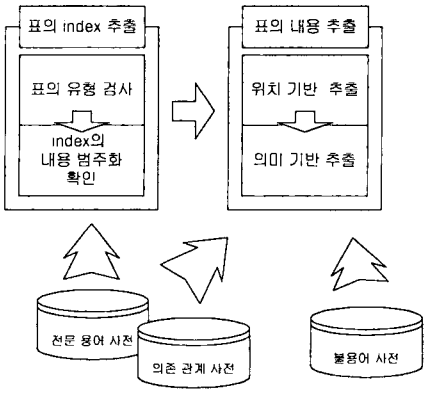



[그림 2] 기상청 홈페이지 Parsing 결과

◆ 비정형화된 표 parsing

비정형화된 표를 Parsing하는 방법은 다음의 몇 단계를 거쳐서 분석하게 된다. 비정형화된 표는 우리가 정한 공통된 특징에 맞지 않는 표를 공통된 형식(Frame)에 항목을 입력할 수 있도록 재가공하여 Parsing 할 수 있다. 기본적인 2차원 배열의 형식으로 표를 저장하기 위해서는 항목에 해당하는 부분을 찾는 것이다. 이를 위해서 표의 내용이 가로로 나열되어 있는지 세로로 나열되어 있는지를 찾는다. 다음으로 항목에서 그 표의 인덱스에 해당하는 항목을 선택해야 한다. 인덱스에 해당하는 항목(상품명, 도시명, 제목)은 90% 정도는 첫 번째 항목에 나타난다. 그 외의 경우는 문서의 제목과 단어를 색인해서 의미에 기반해서 선택하게 된다.

비정형화된 표에서 이러한 항목을 선택한 후에 표의 내용 중 해당하는 부분을 추출해 낸다. 이때 정형화되어 있는 표처럼 가로나 세로의 위치를 이용해서 파악하며, 이것이 효과적이지 않을 경우는 표의 내용과 인덱스 항목의 의미적 연관성을 고려한다. 예를 들어서 "가격"이라는 인덱스에 나타나는 항목은 "숫자"일 가능성이 높다. 이를 이용해서 내용에서 필요한 정보를 걸러낸다. 이를 위하여 표의 주변 정보와 표의 특성을 개념화한 각종 사전 혹은 전문용어 사전, 불용어 사전을 구축 사용한다. 다음 [그림 3]은 비정형화된 표 Parsing에 대한 전체 흐름도이다.




[그림 3] 비 정형화된 표의 Parsing

◆ 표 주변 정보를 이용한 Parsing의 정확도 개선

HTML 문서는 태그 중심으로 구조가 이루어져 있다. 이 태그로 문서의 구조와 모양 내용까지 표현한다. 그리고 표에 해당되는 태그에서도 글자 모양, 색깔, 크기, 이미지 같은 정보가 내부에 포함되어 있으므로 파싱단계에서 필요한 내용에 해당하는 정보만을 분리한다.

이러한 방법을 이용하여 문서 전체에서 필요한 정보만을 추출한다. 그리고 표에 대한 정보 중 가장 중요한 정보는 표를 나타낼 수 있는 키워드이다. 즉, 표가 무엇을 나타내는지를 결정해야 한다. 이 내용은 표의 내부(caption tag)에서 찾을 수도 있고 외부에서 찾을 수도 있다. 그 외에, 표의 앞부분 내용이나 문서의 제목, 디렉토리 이름 등에 정보가 있을 경우, 이러한 정보들은 우선 표의 앞, 뒤와 디렉토리 정보, 문서의 제목 등이 표의 제목이 될 수 있으므로 이를 문서에서 분리한다.

HTML 문서에 대한 분석을 간단한 예를 통해서 살펴보면, [그림 4]는 표를 포함한 HTML 문서에 대한 일반적인 개략도이다. <TITLE> tag는 전체적인 문서의 정보를 알려주며, 이것이 전자상거래 문서라는 것을 나타낸다. 테이블 위에 문자열로 "17인치"라는 것이 있다. 이렇게 테이블의 앞과 뒤에 테이블을 설명하는 문구가 자주 들어가는 것을 볼 수 있다. 다음으로 table caption tag를 볼 수 있는데 표에 대한 전체적인 설명을 기술하는 곳이다. 이렇듯 표의 주변 정보는 표에 대한 보충 설명을 해주며 표가 속한 범주를 알려주는 역할도 한다.

```
<TITLE>상품 리스트</TITLE>
...
<H2>17인치</H2>
<B>현재위치 : Computers &gt; 컴퓨터 주변기기 &gt;
모니터 &gt; 17인치</B>
<TABLE>
   <CAPTION>상품정보</CAPTION>
      <TR><TD><A HREF="http://...">일반모니터
</A></TD></TR>
...
</TABLE>
<A HREF="http://...">이용약관</A>
<A HREF="http://...">개인보호정책</A>
<A HREF="http://...">고객센터</A>
...
```

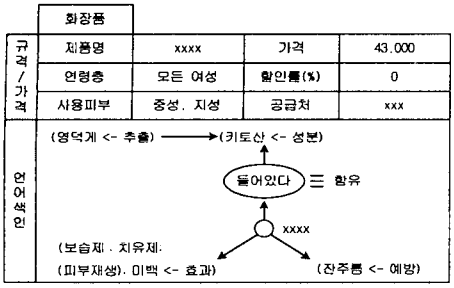
[그림 4] TABLE의 주변정보

## 3. 슬롯 끼워넣기

슬롯 끼워넣기는 HTML 문서의 Parsing 결과를 다른 format으로 바꾸는 작업을 말한다. 즉, 표의 내용을 개념에 기반해서 재구성함으로써 정보검색시스템에서 효율적으로 접근할 수 있게 재구성하는 것이다.

◆ 한국어 문서분석 결과통합

표의 내용을 이용해서 시스템에서 효율적으로 사용할 수 있게 표의 내용과 한국어의 개념을 이용하여 하나의 새로운 프레임(Frame)으로 구성한다. 즉, 표의 정보와 각 응용분야의 표준규격(예, 한국전자산업진흥회의 전자부품 표준화 규격) 등을 이용하여 응용분야의 특성에 맞게 구현한다. 제품의 이름으로 응용분야를 결정하고 응용분야의 특성화된 Frame으로 표의 정보를 슬롯 끼워 넣기 방식으로 저장한다. 다음의 [그림 5]는 표의 정보를 한국어 정보와 통합한 예이다. 여기서 설명에 대한 색인은 한국어 문장의 개념화 방법에 따른다. 위에서 '추출', '성분' 등은 이끈말(governor)이며, '영덕게', '키토산' 등은 매인말(dependent)이다. 또 '들어있다'는 '성분', '효과'와 '예방'을 이끈다. 그리고 '성분이 들어있을 때'는 의미가 '함유'의 뜻임을 보여준다.




[그림 5] 표와 한국어 정보의 통합

각 응용 영역 특성에 따라 사용자의 검색 요구는 달라진다. 전자상거래에서 전자제품은 특정 상표를 중심으로 검색할 것이고, 생활용품에서는 가격으로 검색할 것이다. 그리고 의류와 신발 분야에서는 스타일과 색깔 같은 것으로 검색을 할 것이다. 각 응용분야별로 프레임을 구성할 때 이런 각 응용분야의 특성을 분석하여 사용자의 검색요구에서 항목별 가중치를 설정하여 효율적인 검색이 가능하다.

## 5. 결론

사건, 개념 또는 의미를 효과적으로 표현하는 방법은 인공지능 연구에서 다양하게 연구되고 있다. 하지만 상식의 표현이 쉽지 않다는 것은 인공지능연구자들에게는 잘 알려져 있는 사실이다. 효율적인 검색을 위해서는 시소러스 방식을 넘어서 개념을 표현하는 인공지능적인 방법이 필요하다. 이제 웹의 확장과 더불어 정보는 빠르게 증가하고 있으며 다양한 방법으로 저장되고, 표현되고 있다. 그리고 무엇보다도, 의미있는 정보는 이제 단순히 문장으로 텍스트화 되어 있는 것이 아니라, 표와 같은 형식으로 저장되어 있는 경우가 많다. 이러한 추세는 앞으로도 더욱 가속화되어 표를 이용한 정보검색을 하지 않을 경우, 표로 작성되어 있거나 표에 저장되어 있는 가치있는 정보들을 얻어내지 못하게 된다. 표에 내포되어 있는 문장들에 대한 검색, 또는 표 내부에 위치한 표에 대한 검색, 그리고 표가 있는 table의 주변 정보와 표 내부의 단어를 이용한 검색 등에 대한 논의와 연구가 아직 미흡한 실정이다. 표 정보를 이용한 웹 문서간의 관계 개념화의 응용 방법에 대한 연구는 앞으로도 긴밀하게 요구되는 부분이다.

## [참고문헌]

[Salton] G. Salton and M.J.McGill, Introduction to Modern Information Retrival, McGraw-Hill, 1983

[Korfhage] Robot R. Korfhage, "Information Storage and Retrieval" Wiley Computer Publishing, 1997

Daniel Jurafsky and James H. Martin (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 12장

[이준영] 이준영, "다중색인과 압축저장에 의한 정보검색 시스템 개발에 관한 연구", 이학석사 학위논문, 부산대학교 전자계산학과, 1997

[박민경] 박민경, "한국어 어구 색인을 이용한 색인 효율의 향상", 이학석사 학위논문, 부산대학교 인지과학협동과정, 1998

[김영관] 김영관, 권혁철, "미리내 검색시스템의 명사 추출 시스템", 제 11회 한글 및 한국어정보처리 학술대회 한국어 형태소 분석기 및 품사태거 평가 워크숍 논문집, p89-91, 1999년

[김민정] 김민정, "규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거", 이학박사 학위논문, 1997년 2월, 부산대학교

[박민식] 박민식, "링크정보를 이용한 웹 기반 정보검색 시스템 구현", 이학석사 학위논문, 2000년 8월, 부산대학교