[1]웹 문서 중 의미 있는 표의 추출

정성원[0] 이원희 김영기 권혁철
부 산 대 학 교 전 자 계 산 학 과
{swjung, whlee}@pusan.ac.kr, ykiki6292@hanafos.com, hckwon@pusan.ac.kr

Extraction of Meaningful Tables from The Web Documents

Sung-Won Jung[0] Won-Hee Lee Young-gi Kim Hyuck-Chul Kwon
Dept. of Computer Science, Busan National University

요 약

현재까지 정보 검색 시스템은 색인어 위주로써 문서의 구조적 정보를 고려하지 않았다. 글자의 크기나 글자체, 들여 쓰기, 표 등은 저자의 의도를 구체화하며, 문서를 명확하게 하는 주요한 수단이다. 이 연구에서는 특히 표에 주목한다. 표는 많은 문서에 일반적으로 쓰이며, 글을 명확하게 해 준다. 일반 문서에 비해서 웹 문서는 태그를 이용하여 정보를 추가할 수 있어 표를 쉽게 구분할 수 있다. 하지만, 웹 상의 표는 지식을 구조화하는 근본적인 목적 이외에, 단순히 화면을 정렬하려고 하는 목적으로도 많이 쓰인다. 이 연구에서는 정보 검색 시스템에 표 정보를 사용하기 위한 전처리 단계로 의미 있는 표를 추출하는 방법을 제시하며, 이를 위하여 결정 트리를 사용한다.

1. 서론

정보 검색 시스템은 이용자가 원하는 가장 적절한 정보를 빠른 시간 내에 검색해서 전달해 주기 위한 것이므로, 정보검색 시스템의 성능은 일차적으로 전체 검색 대상이 되는 문서 중에서 필요한 정보가 담겨 있는 문서를 얼마나 빨리, 그리고 많이 찾아내는가에 달려 있다고 볼 수 있다. 따라서 지금까지의 정보검색 시스템은 주로 이 재현율과 정확률을 높이기 위한 기법과 모델의 개발에 주력해 왔다. 특히 웹 문서의 경우 다수의 사용자가 폭발적으로 문서를 양산해 냄에 따라 검증되지 않은 문서들이 매우 많아지게 되었으며, 이에 따라 정보 검색 시스템의 정확률을 높이는 보다 나은 검색 기법이 요구되고 있다.

검색 시스템의 정확률을 높이기 위해서는 궁극적으로 웹 문서(Web document)에 대한 의미론적인 분석(semantics analysis)이 필요하다. 그렇지만 현재의 기술 수준으로는 이를 인터넷 정보검색 기법에 적용시키기에는 여러 가지 어려운 점이 많다. 또 다른 방법으로 저자의 의도를 파악해서 지금의 정보 검색 시스템에 추가의 정보를 덧붙이는 것인데, 그 중 하나가 웹 문서에 대한 구조적인 정보를 분석하는 것이다. 저자는 문서를 작성할 때, 저자의 의도를 독자에게 명확하게 전달하기 위하여 제목을 붙이고 단락을 나누며, 들여쓰기를 하고 제목의 앞에 번호나 기타 기호를 붙이기도 하며, 표를 사용하기도 한다.

이 논문은 웹 문서가 갖고 있는 여러 가지 구조 정보 중에서 표(table)에 초점을 맞추고 있다. 문서 상의 표는 주로 저자의 의도를 좀 더 명확하게 표현하기 위해 만들어지기 때문에 서술식으로 이루어진 문서내의 다른 문장들보다 더 중요한 정보를 담고 있다고 볼 수 있다.

[1] 이 논문은 과학 기술부(한국과학기술기획평가원)의 국가지정연구실 사업지원으로 이루어진 것임

표는 웹 문서 내에서 쉽게 찾아낼 수 있고, 의미를 추출해 내기도 쉽기 때문에, 표에 대한 분석을 통해 정보검색 시스템의 성능 향상을 기대 할 수 있을 것이다. 또한, 이 기법은 벡터 공간 모델(Vector Space Mode)이나 피놈 모델(P-norm Model) 등과 같은 현재의 정보검색 시스템의 순위화 모델(ranking model)에 추가적으로 적용시키기 용이하다.

2. 관련 연구

기존 인터넷 상의 표의 정보 추출은 특정 웹 문서 형식에 국한되어 있다. [1], [2], [3], [4]는 웹 페이지의 html 태그를 분석하여 정보를 추출하는 시스템을 구축하였다. 이 연구들에서는 먼저 웹 문서의 특별한 html 태그를 분석한 후 휴리스틱으로 미리 구축된 추출 규칙에 따라 정보를 추출한다. [5]는 웹 페이지의 표를 의미적 관점에서 이해하고, 구조가 복잡한 표를 추상 의미 모델로 기술하고, 그 결과를 트리 형태로 만든다. 이 방법들은 대체로 웹 페이지의 html 태그형식에 의존한다. 따라서, 웹 페이지의 레이아웃이나 포맷이 변할 경우 추출 시스템의 수정이 필요 하다. 따라서 실제 웹 문서에 적용하기에 적합하지 않다.

3. 웹 문서 상의 표에 대한 분류

인터넷 상에는 다양한 유형의 표가 존재한다. 일반적인 문서에서 표란 저자가 자신의 의도를 독자에게 더욱 정확하게 전달할 목적으로 서술형으로 된 문서를 격자틀 안에 구조화 한 것을 말한다. 이를 위해서 인터넷 상에 사용하는 html에서는 <table>이라는 태그를 지원한다. 하지만, 인터넷 상에 나타나는 표는 앞서 서술한 표의 본래 목적 이외에 문서를 정렬하거나 꾸미기 위한 목적으로 쓰이는 경우가 더 많다. 따라서 우리는 인터넷에서 표의 정보를 추출하는 것에 앞서서 우리가 원하는 표를 웹 문서에서 골라내는 작업이 필요하며, 이를 위해서 우리는 웹 문서 상의 표를 다음 3가지로 분류한다.

웹 문서 상의 표 정보를 추출하기 위해서는 웹 문서 상의 표 중에서 정보검색 시스템에 적용할 대상이 되는 표, 즉 의미 있는 표에 대한 이해가 필요하다. 다음 [그림 1]은 우리나라 통계청 Web page중 하나이다. 검색 시스템 사용자가 query로 '1970년의 국민 총소득' 을 입력했을 경우 검색 시스템은 [그림 1]과 같은 Web page를 찾아서 사용자에게 제공해야 한다. 이런 형태의 표를 본 연구에서는 '의미 있는 표' 라 지칭한다. 본 연구에서 고려하고 있는 인터넷 상에서 의미 있는 표는 다음과 같다.

- 가장 윗 행과 가장 좌측 열이 표의 행과 열을 대표하는 색인어를 가진다.
- 행과 열의 조합으로 추출되는 셀의 내용이 특정 정보를 가진다.
- 일반적인 색인어 추출만으로는 효과적으로 셀의 정보를 추출하기 힘들다.

이와 같은 특성을 가지는 표 이외에 우리는 흔히 [그림 2]와 같은 웹 페이지를 많이 볼 수 있다. 이 문서의 html source를 보면 <table> 태그를 문서 전체에서 골고루 사용하고 있다. 하지만, 앞서 살펴본 [그림 1]같이 의미 전달이 주 목적이 아니라 화면을 보기 좋게 꾸며 사용자에게 정돈된 시각적 효과를 주는 측면이 매우 강하다. 우리는 이런 형태의 표를 '꾸미기 위한 표' 라고 부르기로 한다.

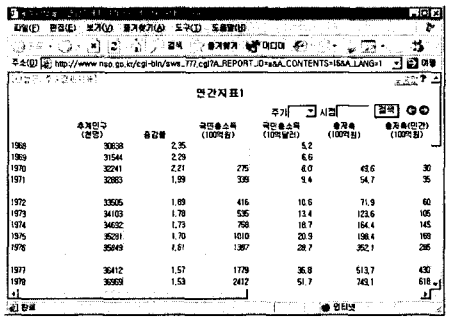

그림 1 의미 있는 표

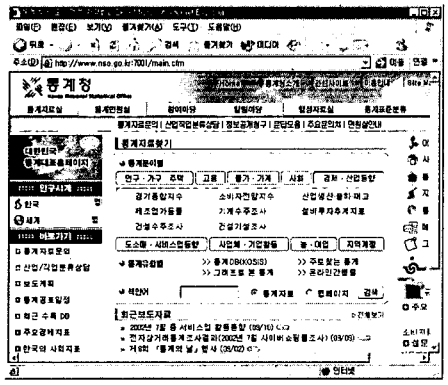

그림 2 꾸미기 위한 표

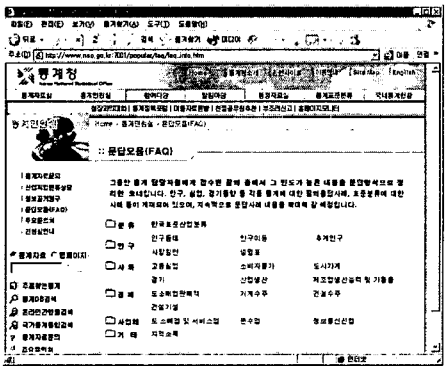

그림 3 혼합된 표

마지막으로, 앞의 두 가지 표의 특성을 모두 가지고 있는 [그림 3]과 같은 형태의 표가 있다. 우리는 이것을 '혼합된 표' 라고 부르기로 한다. 이런 표의 형태는 의미 있는 표와 같이 저자가 전달하는 바를 구조적으로 도식화 한 측면이 있지만, 표 색인어를 추출한다고 해서 의미 있는 표에서 정보를 추출한 것처럼 많은 효과를 볼 수 없다. 하지만 의미 있는 표와 형태적으로 비슷한 특성이 많아서 기계적으로 분류하기 어렵다.

이와 같은 표에 대한 분류를 전체로 인터넷 상의 실제 데이터를 이용해서 표의 통계를 수작업을 통하여 조사해 봤다. 본 연구에서는 혼합된 표는 그 특성상 구분하기 힘들기 때문에 의미 있는 표와 꾸미기 위한 표만 대상으로 한다. 수작업으로 분류한 결과는 [표 1]과 같다.

항 목	값
전체 문서 수 (A)	86475(개)
<table> 태그 포함 문서 수 (B)	67259(개)
의미 있는 표 포함 문서 수 (C)	1009(개)
문서당 표의 평균 개수	15.13(개)
전체 문서 중 <table> 태그를 포함한 표의 비율 (B/A)	77.78(%)
전체 문서 중 의미 있는 표를 포함한 문서 비율 (C/A)	1.167(%)
<table> 태그를 포함한 문서 중 의미 있는 표를 포함한 표의 비율 (C/B)	1.500(%)

표 1 웹 문서 중 표에 대한 통계

조사대상 86475건 문서 중에서 <table> 태그를 포함한 문서는 약 77.78%이다. 생각보다 매우 많은 수의 웹 문서가 <table> 태그를 사용한다. 전체 문서 중 의미 있는 표를 포함한 문서는 1.167%로 예상보다 적게 나타났다. 이것은 표가 있는 문서만을 대상으로 했을 때는 1.5% 정도이다. 이는 인터넷 상에 표를 사용하는 용도가 주 용도 보다는 웹 문서를 정렬하고 꾸미는 데 치중해 있다는 것을 알 수 있다.

4. 표에 대한 구분 특성 설정

표의 분류 작업에서 알 수 있듯이 우리가 원하는 의미 있는 표는 웹 문서에서 매우 적다. 따라서, 위의 3가지 표 중 의미 있는 표를 추출해 내는 작업이 꼭 필요하다. 하지만, 웹 문서의 양은 매우 많으므로 수작업으로 추출해 내는 것은 불가능하다. 따라서, 의미 있는 표를 구별 지을 수 있는 특징을 골라내어 그것을 기준으로 자동 처리하여 의미 있는 표인지를 판단할 수 있

어야 한다.

[그림 1]을 살펴 보면 몇 가지 특징을 얻을 수 있다. 그림의 유무라든지, 합쳐진 셀이 있는지 여부, 색인이 되는 가장 윗 행과 가장 왼쪽 열이 다른 셀들과 배경색이 다르다든지 하는 것들을 얻을 수 있다. 또한, 문서의 html 소스를 조사하여 html 태그가 어떤 형식으로 쓰여졌는지도 고려하였다. 다음은 우리가 분석해 낸 표의 특성과 그 타당성을 조사하여 얻은 구별 특성들이다.

번호	내용	값
1	<caption> 태그의 유무	0:없음 1:있음
2	<th> 태그의 유무	0:없음 1:있음
3	<thead> 태그의 유무	0:없음 1:있음
4	가장 윗 행의 cell내의 내용분류	0:없음 1:단어 2:문장 3:숫자 4:그림 5:기호 6:복합
5	가장 윗 행과 다음 행의 배경색 차이 유무	0:없음 1:있음
6	가장 윗 행과 다음 행의 폰트 색깔 차이 유무	0:없음 1:있음
7	가장 윗 행과 다음 행의 폰트 종류 차이 유무	0:없음 1:있음
8	border 속성의 유무	0:없음 1:있음
9	전체 셀에서 내용이 없는 셀의 비율 (p)	0:p>1% 1:other
10	전체 셀에서 image를 가진 셀의 비율 (p)	0:p>1% 1:other
11	전체 셀에서 href를 가진 셀의 비율 (p)	0:p>1% 1:other
12	전체 셀에서 text만으로 된 셀의 비율 (p)	0:p>1% 1:other
13	전체 셀에서 수치데이터만으로 이루어진 행이나 열의 유무	0:없음 1:있음
14	전체 셀에서 기호만으로 이루어진 셀의 비율 (p)	0:p>1% 1:other
15	table내부에 중첩된 table이 있는지 여부	0:없음 1:있음
16	table 크기에 따른 종류	0:1*n, m*1 1:m*n<10 2:10≤m*n<30 3:other
17	셀이 가진 문자열이 100자를 넘는지 여부	0:넘지 않음 1:넘음
18	표의 형태가 정확한 n*m인지 여부	0:정확함 1:정확하지 않음

표 2 표 구별을 위한 구별 특성

이 구별 특성들은 의미 있는 표를 추출하기에 어느 정도 적합해야 한다. 그에 대한 통계적인 근거는 다음 [표 3]를 통해 알 수 있다. [표 3]를 보면 각 특성 별로 꾸미는 표와 의미 있는 표의 비율이 각 특성의 특성 값 중 어느 한쪽으로 선택되는 경향이 있다. 가장 좋은 분류 특성을 보여주고 있는 특성 16, table크기에 따른 종류를 보면, 표가 1차원이거나(1*n, m*1), 표의 크기가 작을 때, 꾸미는 표일 경우가 많으며, 반대로 표가 일정크기 이상을 가지면, 의미 있는 표일 경우가 많다. 또한, 특성 1을 보면 값 0에 거의 모든 값이 몰려 있는데, 이것은 <caption> 태그를 쓰는 웹 문서가 거의 없다는 것을 의미한다. 하지만, <caption> 태그를 쓰면 거의 100% 의미 있는 표가 될 수 있는데, 이는 특성 1의 값 1에 B, 의미 있는 문서포함 비율이 0.2임을 보면 알 수 있다. 일반적인 상식으로도 <caption> 태그는 잘 쓰이지 않지만, 만약 쓰인다면, 이것은 html을 잘 아는 사람이 의미를 가지고 사용했을 가능성이 크다. 하지만, 이 비율은 매우 작은 값이므로 결정 트리에서는 이 구별 특성의 기여도가 크지 않을 것이다.. [표 3]에서 구별 특성 4는 값이 6까지 존재하지만, 지면상의 이유로 3까지만 표기하였다.

구별 특성	구별 특성 값							
	꾸미는 표				의미 있는 표			
	0	1	2	3	0	1	2	3
1	100	0			99.8	0.2		
2	99.3	0.7			91.3	8.7		
3	100	0			99.8	0.2		
4	0.1	0.1	29.4	0.1	0	1.2	48.4	1.1
5	98	2			63.1	36.9		
6	93.2	6.8			81.9	18.1		
7	98.7	1.3			94	6		
8	94.4	5.6			30	70		
9	92.5	7.5			100	0		
10	83.7	16.3			100	0		
11	88.1	11.9			97.4	2.6		
12	82.9	17.1			60.9	39.1		
13	82.7	17.3			58.2	41.8		
14	68.2	31.8			78.3	21.7		
15	66.9	33.1			99.3	0.7		
16	54.2	23.8	11.8	10.2	0	5.6	48.4	46
17	92.7	7.3			91.3	8.7		
18	87	13			52.8	47.2		

표 3 구별 특성에 대한 통계적 추출 근거

4. 결정 트리 적용

앞에서 추출된 구별 특성은 의미 있는 표를 결정짓는 기준이 될 수 있지만, 그 구별 특성들의 적용 순서도 중요하다. 이를 위하여 우리는 얻어진 구별 특성을 결정 트리를 사용하여 적용순서를 구현하였다. 결정 트리는 ID3 알고리즘[6]을 사용하였다. ID3 알고리즘은 각 단계에서 다음 단계로 나아갈 때 정보량이 최대가 되는 구별 특징을 선택하는 것이다. 다음 [그림 4]는 본 연구에서 사용한 결정 트리의 전체적인 알고리즘의 계략도이다.

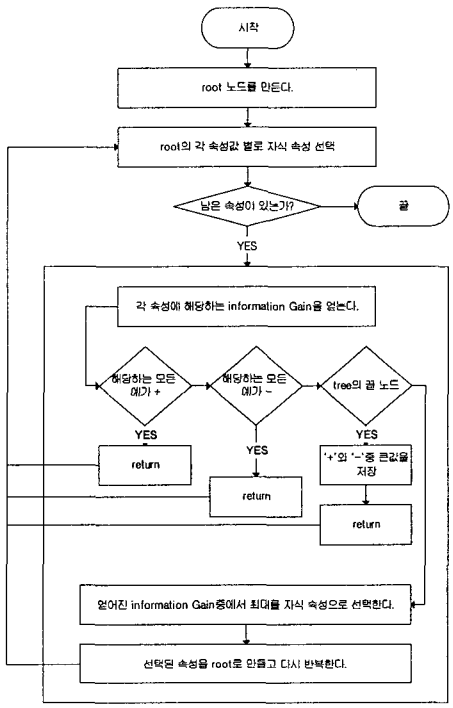

그림 4 결정 트리 알고리즘

[그림 4]의 알고리즘에서 information gain을 구하는 식은 다음과 같다.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

이 식으로 S예제 집합에서 A속성의 information gain을 구한다. v는 A속성의 속성값 중의 하나로써 각 속성값에 해당하는 entropy의 합을 S의 엔트로피에서 뺀다. 각 속성값에 해당하는 entropy가 작아 질수록, 다른 말로 하면, 분류가 잘 될수록 좋은 구분 속성이 된다. 따라서, information gain이 높을수록 좋다.

5. 시스템 구현

다음 [그림 5]는 시스템 구현 절차를 나타낸다.

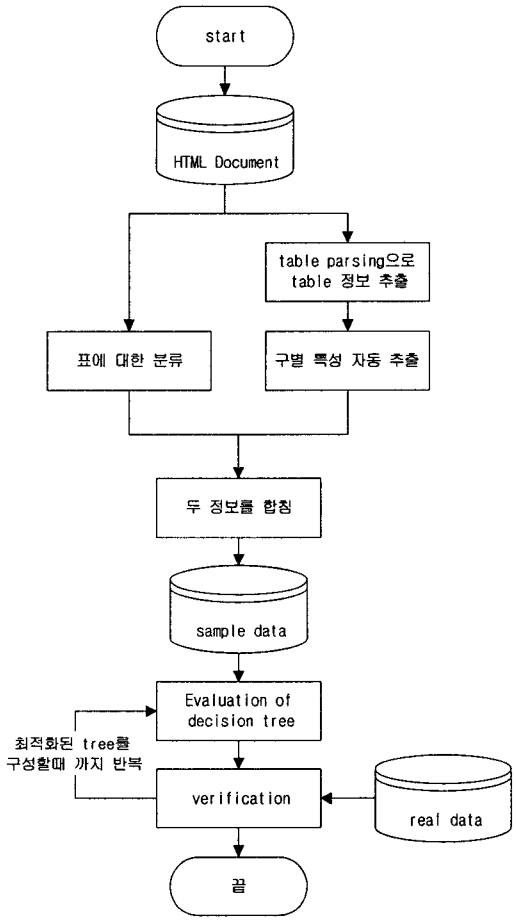

그림 5 시스템 구현 절차

전체적인 작업 과정을 요약하면 다음과 같다. 1차적으로 분석대상이 될 샘플 데이터를 수집하여 의미 있는 표를 포함한 데이터와 그렇지 않은 데이터로 분리하는 수작업이 필요하다. 우리는 이 시스템의 sample data를 구축하기 위하여 10만 건의 웹 문서를 조사하였다. 그 다음, 웹 문서를 파싱하여 그 결과에서 표 정보를 추출하고 이를 이용하여 각 표에 해당하는 구별 특성을 자동으로 추출하는 시스템을 구현하였다. 이것과 수작업으로 분류한 결과를 합쳐서 샘플데이터를 만들었다. 만들어진 샘플데이터를 사용하여 앞에서 살펴본 결정 트리 알고리즘을 사용하여 결정 트리를 구현하였으며, 만들어진 결정 트리에 실제 data에 적용하여 분류의 정확도를 살폈다.

6. 실험 결과

실험의 환경은 인텔 펜티엄 IV 2GHz CPU에 256Mbyte 주 메모리를 가진 PC를 사용하였다. 샘플 데이터는 웹 문서 10만 건에서 추출한 정보를 사용하였고, 검증을 위한 테스트 데이터는 부산대학교 인공지능 연구실에서 미리 수집해 놓은 웹 문서 약 500만 건 중에서 임의 추출하여 적용하였다.

다음 [표 4]과 [그림 6]은 결정 트리의 정확도에 대한 실험이다. 각 구별 특성에 대해서 1개만 적용해 봤을 때와 17개의 구별 특성 중 해당 되는 특성을 뺐을 때, 그리고 구별 특성을 하나씩 추가 시키면서 결정 트리의 정확도 향상 정도를 조사하였다. 5000개의 테스트 데이터를 가지고 결정 트리를 만들었으며 테스트 데이터를 만들어진 결정 트리 검증을 위해 다시 입력하여 사람이 분류한 것과 같은지를 조사하였다. 표 안의 숫자는 사람이 분류한 것과 다른 결과가 나오는 데이터의 개수이다.

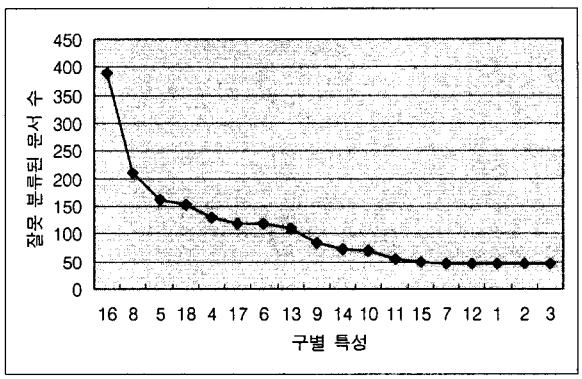

그림 6 구별 특성 누적 적용 시 정확도 향상

특성 (a)	a특성적용	18개 특성 중 a를 빼고 적용	각 특성 누적 적용
16	388	110	388
8	429	80	212
5	336	62	162
18	436	60	152
4	441	61	129
17	436	51	119
6	436	53	117
13	436	46	110
9	436	48	83
14	436	49	72
10	436	52	68
11	436	51	56
15	402	48	50
7	436	49	46
12	436	46	45
1	436	45	45
2	436	45	45
3	436	45	45
모든 데이터를 의미 없는 표로 간주			436

표 4 결정 트리의 정확도

위 [표 4]를 보면 16번 구별 특성이 구별 특성 중에서 가장 많은 기여를 했다는 것을 알 수 있다. 각 특성의 누적 적용 부분을 보면, 구분 특성을 추가해 나갈 때마다 점점 오분류한 개수가 줄어듦을 알 수 있다. 구별 특성을 좀더 추가하고, 구별 특성의 특성 값을 조절한다면 더 나은 결과를 얻을 수 있을 것이다.

다음 [표 5]는 위에서 학습한 결정 트리를 실제 data에 적용한 결과이다. 데이터는 인터넷에서 수집한 웹 문서 10000건을 선택하여 실험하였다. 결과적으로 69.1%의 분류 정확도를 얻을 수 있었으며, 추가 구분 특성 추가나 분류 정책의 보완을 통해서 결과의 정확도를 높일 수 있을 것이다.

A : 구현된 시스템에서 의미 있는 표로 추출한 개수

B : 실제 의미 있는 표의 개수

표 개수	A	B	정확도
4203	310	252	81.2%
11213	312	160	51.3%
5448	298	236	79.2%
15432	324	284	87.7%
14991	165	82	49.7%
14219	264	123	46.6%
15084	240	145	60.4%
9602	219	187	85.4%
8024	180	129	71.7%
98216	2312	1598	69.1%

표 5 실제 웹 데이터 적용 결과

7. 결론 및 향후 과제

정보 검색 시스템이 정확한 정보를 사용자에게 전달하기 위해서는 정보에 대한 평가 작업이 필요하다. 웹 문서 중 표에서 원하는 정보를 추출해 내기 위해서는 의미 있는 표와 의미 없는 표에 대한 분류 작업이 필수적이다. 일단 표가 추출되면, 그 표는 어느 정도 정형화되어 있으므로 처리하기가 수월하며, "부산의 기온" 같은 질의어가 들어 왔을 때, 표의 정보를 구조화하여 저장한 다면, 정확한 결과를 사용자에게 전달할 수 있을 것이다. 본 연구는 일반적인 정보 검색 시스템이나 표 정보가 많은 특정 영역에 적용을 위한 전처리 단계로써 의미 있는 표와 의미 없는 표를 구별하였으며, 그 결과 어느 정도 만족할 만한 성과를 얻을 수 있었다. 향후 표에 대한 더욱 세부적인 분류와 학습을 위한 정밀한 알고리즘이 필요하다. 또한, 표는 그 자체만으로 의미 있는 것이 아니라 문서의 일부분으로 문서의 내용을 보충해 주는 경우도 많으므로, 표와 문서 간의 상관관계 조사가 더욱 필요하다. 그리고 의미 있는 표의 유무가 문서의 중요도에 어떤 영향을 미치는지도 연구되어야 할 것이다.

8. 참고 문헌

[1] J. Hammer, H.Garcia-Molina, J.Cho, R.Aranha, and A.Crespo, Extracting Semistructured Information from the Web, 1997, SIGMOD Record, 26(2): 18-25

[2] Huang Yu-qing, Qi Guang-zhi, Zhang Fu-Yan, Constructing Semistructured Information extractor from the Web document, 2000, Journal of Software 11(1): 73-78

[3] Naveen Ashish and Craig Knoblock, Wrapper Generation for Semi-structured Internet Sources, 1997, SIGMOD Record, 26(4): 8-15

[4] Dan Smith and Mauricio Lopez, Information Extraction for Semi-structured Documents, 1997, In Proceedings of the Workshop on Management of Semistructured *Data, in conjunction with* PODS/SIGMOD, Tucson, AZ, USA, May, 12

[5] Ning Gu, Guowen Wu, Xiaoyuan Wu, Baile Shi, Extracting Web table information in cooperative learning activities based on abstract semantic model, 2001, Computer Supported Cooperative Work in Design, The Sixth International Conference on 2001, 492 -497

[6] Tom M. Mitchell "Machine Learning", McGraw-Hill, 1997