테이블 객체 모델링을 이용한 웹문서의 제목추출

박세종° 윤주형 이승욱 한영석
수원대학교 컴퓨터과학과

An Object Model of Korean Web Pages for Title Identification

Se-Jong Park °, Ju-Hyung Yun, Seung-Yuk Lee, Young-Suk Han
Suwon University Computer Science

요 약

한국어 웹 문서에는 일반적으로 제목이 명시가 되어 있음으로 텍스트를 요약하는 방식의 제목추출과는 달리 여러 테이블 형태로 이루어진 웹 문서의 특성을 고려하여 제목에 해당하는 테이블 객체를 찾아내야 한다. 웹 문서를 테이블 객체의 리스트로 보고, 이들을 휴리스틱 규칙에 의해서 본문 후보와 이를 기준으로 하는 제목 후보 객체들로 구분하는 단계와 제목 후보들 간의 확률적 분포 값과 본문과의 언어적 유사도를 이용하여 제목 개체를 결정하는 단계를 통하여 제목을 인식한다. 인식의 정확성에 기여하는 것은 제목과 본문 객체를 구분하는 규칙 그리고 제목의 확률분포 및 언어적 유사정도 등이며 이들 각 정보가 정확성에 기여하는 정도를 실험하였다. 무작위로 추출된 500개의 다양한 양식의 웹문서를 대상으로 실험한 결과 제목인식 정확성은 95.1%였다.

1. 서 론

웹 문서에서 제목을 추출하는 것은 웹 문서를 관리하는 측면과 다양한 고 부가가치 정보 및 지식을 구축하는 노력의 전 단계로서 의미가 있다. 그러나 웹 문서에서 제목을 추출하는 것은 기존에 일반 텍스트 정보에서 제목을 추출하는 것과는 다르다. 우선 웹 문서는 불필요한 정보들이 많이 포함 되어 있으며, 이들 정보들이 대개 테이블 객체로 나뉘어져 있다. 무엇보다도 제목이 이미 명시가 되어 있음으로 이 제목정보를 활용하는 것이 중요한데, 본문 테이블과 기타 테이블 객체들 사이에 섞여 있는 제목테이블을 찾아 내는 것은 다른 문제가 된다.

본 연구에서는 규칙정보를 이용한 1단계와 확률 및 언어정보를 이용하는 2단계를 통해 실제적인 제목 인식이 가능한 모델을 제안한다. 1단계에서는 웹 페이지 상의 불필요한 태그(Tag) 정보를 걸러낸 후, 테이블(Table) 단위로 객체를 생성한다. 생성된 객체를 중심으로 휴리스틱 규칙에 의해 본문후보와 제목후보를 추출한다. 2단계에서는 추출된 후보들 간의 위치에 대한 확률 분포 값을 적용하고, 제목 후보와 본문 후보간의 언어적 유사도를 코사인 계수를 통해 구하여 최종적으로 두가지 데이터에 대한 가중치를 구하여 제목을 인식한다. 마지막으로 후처리 단계를 거쳐 추출한 제목을 정규화 한다.

제안하는 방법에 의해서 다양한 객체 패턴을 가진 웹 문서들을 대상으로 실험한 결과 본문 인식 규칙, 확률값 적용, 언어적 유사도 규칙등이 제목객체와의 유효한 관계에 있음을 알 수 있었다.

정보추출은 어떤문서의 중심적 의미를 나타내는 특정 구성 요소를 인식하여 추출하는 작업이다[1]. 정보추출의 기술은 자연어 처리기반의 방식[2]과 Wrapper 기반의 방식[3]으로 나눌 수 있다. 이중 웹 문서의 구조를 이용한 정보추출은 주로 Wrapper 기반의 방식을 적용 한다. 비록 Wrapper 방식이 자연어 처리방식에 비해 구조적(structured)이거나 반구조적(semi-structured)인 문서에서 유연함과 확장성을 제공하지만, 단지 특정 웹 문서의 구조에 기인한 규칙을 보여줌으로 수많은 웹페이지의 형식에 대응하기엔 복잡도가 높다. 또한 새로운 형태의 웹 문서가 등장할 경우 Wrapper 시스템 자체가 무효화 될수도 있다는 문제가 있다.

본 논문의 구성은 다음과 같다. 2장에서는 테이블 객체에 의한 웹문서 모델에 대해 기술한다. 3장에서는 실제 본문과 제목 후보 인식을 위한 규칙 모델을 제시한다. 4장에서는 제목후보와 본문 후보간의 확률 모델 및 언어적 유사도에 대한 관계에 대해서 설명한다. 5장에서는 실제 제목 추출을 위한 1단계 2단계 모델들의 각각의 실험과 최종 제목 인식의 정확성에 대한 실험 및 후처리를 통한 제목의 유효성에 대한 실험결과를 제시한다. 마지막으로 6장에서는논문의 방법론 및 실험의 한계점에 대해 기술하고, 향후 방향에 대해 살펴본다.

2. 테이블 객체에 의한 웹 문서 추출 프로세스

본 연구에서 제시하는 제목추출 방법은 휴리스틱 규칙을 적용하여 제목과 본문 후보객체들을 생성하는 1단계와 각 본문 후보에 대한 여러 개의 제목 후보의 확률분포 및 언어적 유사도를 계산하여 최종적으로 제목을 결정하는 2단계를 거쳐 후처리를 통한 정규화 과정으로 이루어 진다.

1단계의 결과물인 제목과 본문 후보 객체들은 <table>태그에 의해서 인식되며, 기본적으로 각 본문후보 개체 위에 1개 이상의 제목후보 객체가 있도록 구성되어 진다. 2단계에서는 각 본문후보 객체와 그에 속하는 제목후보 객체간의 확률 값(표 1) 및 언어적 유사도를 적용하여 제목의 가능성을 평가하게 되며, 최종적으로 가장 높은 값을 가진 제목 후보가 제목으로 선정된다. 마지막 후처리 단계에서는 제목의 유효성을 평가하여 불필요한 기호 및 단어를 제거하는 정규화 과정을 거친다.

<표 1> 270개의 실험자료로부터 추출된 확률분포

No. of main candinate object s= 1

	P(main1) 1.00
P(title1)	0.80
P(title2)	0.06
P(title3)	0.04
P(title4)	0.04
P(title5)	0.02

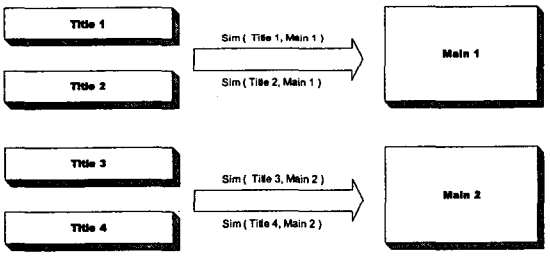
No. of main candinate object s= 1

	P(main1) 0.98	P(main2) 0.02
P(title1)	0.75	0.61
P(title2)	0.11	0.23
P(title3)	0.08	0.16
P(title4)	0.0	0.00
P(title5)	0.0	0.00

웹 태깅 언어는 웹문서에서 보여지는 모든 형태를 정의하여 문서의 중요한 특성을 표현한다. 그러나 하나의 웹 문서상에서도 비슷한형태의 태그와 수많은 텍스트가 존재하여 태그만으로 제목을 추측하는 것은 무리가 있다. 그렇지만 <table>은 문단의 형성, 문단의 위치 등 웹 문서를 구조화 시켜주는데 많이 쓰이는 태그로 제목을 추출하는데 있어 하나의 단위가 될 수 있다. <table>안에 존재하는 <td>는 웹 에 보여지는 텍스트가 존재하고 그 만의 의미 혹은 특성이 있으므로 각각의 객체로 볼 수 있다. (그림 1) 그러나 이 구조는 웹 문서 전체에 걸쳐서 제목 과 본문 및 기타 다른 텍스트들의 표현 방식이 될 수 있으므로, 유효한 정보만을 찾아 내는것이 중요하다. 언어의 유사도 모델은 본문 후보별로 위쪽에 존재하며, 바로 위 본문 후보보다는 아래에 있는 제목 후보들과의 유사도를 이용한다. (그림 2)

```
<table>
<tr>
  <td width="20%"  bgcolor="f6f6f6" align="center">
    <font class="t1" color="353535"><b>금주다운</b></font>
  </td>
  <td  bgcolor="fefefe">
    <font class="t1">4</font>
  </td>
  <td  width="20%"  bgcolor="f6f6f6" align="center">
    <font class="t1" color="353535"><b>Uninstall</b></font>
  </td>
  <td  bgcolor="fefefe">
    <font class="t1">지원안함</font>
  </td>
</tr>
</table>
.
.
```

<그림 1> 실제 웹 문서의 테이블 구조



<그림 2> 제목 후보 객체 별 본문과의 유사 관계

3. 본문과 제목 후보 인식을 위한 규칙 및 후처리 규칙

본문과 제목의 후보객체를 인식하는 단계는 객체의 문서 내 위치 등을 반영하는 html 태그를 활용하는 규칙에 의존한다.

(1) <td>와 </td> 사이의 문자열에서 하이퍼링크 (Hyper Link), <option>, 라디오 버튼, 체크 박스를 한 개의 객체로 보고 후보의 자격을 준다

(2) 제목(후보) 객체는 반드시 본문 객체와 쌍으로 존재해야하며 없으면 더미 객체를 만들어 준다. 제목 후보는 4바이트 이상 73바이트 이하의 문자열이고 74바이트 이상의 객체와 4바이트 이하의 연속된 후보들을 본문 후보로 한다. 기준이 되는 74바이트는 473개의 웹 문서에서 제목과 본문의 구분을 가장 적은 오류로 나눌 수 있는 값으로 결정 되었다 (그림 3).

(3) 웹 문서에서 흔히 발견되는 불필요한 텍스트를 걸러내 주기 위해 불용어(stop word)를 제거 한다.

본문후보 객체를 인식하는 작업은 기본적으로 텍스트의 길이로 찾게 된다. 규칙은 다음과 같다.

(1) 텍스트의 길이가 74바이트 이상인 경우에 본문후보 객체가 된다.

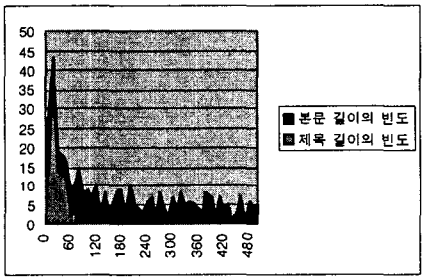
(2) 위에서 불용어를 찾을 수 있었다면 그 위치로부터 가까운 거리에서 80바이트 이상의 텍스트를 찾게 되고 해당 텍스트가 없으면 가상의 본문 후보를 생성하여 작업을 수행한다. 테이블 거리는 테이블에 매겨진 번호의 간격을 말하며 웹 문서의 하단일수록 테이블 값은 커진다.

후처리 단계에서의 정규화 과정은 기본적으로 형태소 분석[4]을 통해 이루어진다.

(1) 형태소 분석 이전에 전처리 과정으로 불필요한 문자를 제거한다. 괄호 묶음의 경우, 제목의 보충 설명이거나 대분류 이므로 묶음 자체를 제거한다.

(2) 형태소 분석을 통해 삭제하여도 정보의 손실이 없는 관형사와 부사, 감탄사등의 독립언[5]을 삭제한다.

(3) 명사로 판정이 되었더라도, 의미가 명확하지 않는 경우 불용어 사전을 검색하여 단어를 삭제한다.



최적 길이 73 BYTE

<그림 3> 제목과 본문의 길이 분포

4. 제목 확률 모델 및 언어적 유사도

본문 후보 객체의 수에 따른 실제 본문의 분포와, 이 본문 객체 위에 존재하는 제목 후보 객체 중에서 실제 제목의 분포를 조사하여 확률을 계산하였다. (표 1)

또한 제목 후보 객체들과 본문후보 객체 사이의 언어 유사성을 적용함으로써 정확도를 높일 수 있다. 모든 객체들을 바이그램(bigram)방식[6]으로 나눈 후, 각각의 유사도를 구한다.

두 과정을 거치면 각 후보 객체는 확률 값과 유사도를 가진다. 이 두 수치의 비율을 달리하면서 합한 후 정확률을 비교하여 최적의 비율을 찾아낸다.

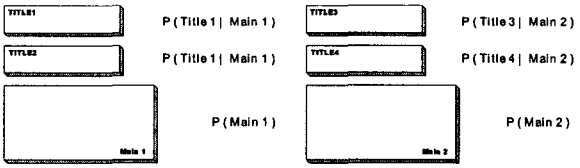
$$e * p(t) + (1-e) * sim(t, m)$$

5. 실 험

확률 값 추출을 위해서 테이블 구조를 가진 웹 문서 약 270개를 대상으로 본문과 제목 후보를 인식하는 규칙을 적용하여 후보로 선정된 텍스트를 두고 수작업으로 실제 제목, 본문을 선정하고 이로부터 본문객체대비 제목객체의 빈도수를 계산하였다. (그림 4)

수작업으로 실제 제목과 본문을 구하고, 이 위치가 후보들 중에의 상대적으로 차지하는 위치를 통계를 낸다. 이런 구조의 문서가 100개중, 첫 번째 본문 후보가 실제 본문인 문서가 75개 였다면 P(main1) = 0.75가 된다. 그리고 이 때, 제목이 될 수 있는 후보는 다시 title1,title2로 좁혀지게 되고, 같은 방식으로 확률을 구하게 된다. 실험은 본문 후보의 개수 만큼 진행될 수 있지만, 3개까지 제한하였다.

실험을 위한 자료는 같은 방법으로 테이블 구조를 가진 위와 다른 웹 문서 약 500개를 추출하였다. 실험의 정확성에 적용한 방법은 본문 인식 규칙, 확률 값 적용, 언어적 유사도의 3가지 이고 각기 어느 정도의 정확성을 가지고 있나 계산하였다. (표 2)



<그림 4> 확률 객체 모델

확률 값과 언어적 유사성	본문 인식과 확률 값	본문 인식과 언어적 유사성
65.4%	85.2%	93.8%

<표 2> 실험을 통한 결과

축약 결과를 살펴보면, 본문 객체의 인식 규칙과 언어적 유사도를 사용하였을 경우가 가장 높은 정확성을 보이고 있는데, 이것은 제목 후보의 객체 수가 줄어든 만큼 언어의 유사성에서 본문이 제목이 아닌 것과 높은 유사성을 가질수 있는 오류를 줄였기 때문이다.

따라서 본문 인식의 결정이 확률 값과 언어적 유사성의 작용에 큰 영향이 있으며, 확률 값 보다는 언어적 유사성의 결과에 좀 더 가중치를 두어야 한다는 것을 알 수 있다.

확률적 모델의 가중치와 언어 유사성에 의한 가중치 모두 적용할 경우에, 두 정보가 정확도에 기여하는 정도가 다를 것임으로 두 정도를 각기 다른 가중치로 배합하였다.(표 3)

<표 3> e 의 크기에 따른 정확도

$e * p(t) + (1-e) * sim(t, m)$

e	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
정확도	95.1%	92.7%	88.3%	86.5%	86.4%	86.6%	85.8%	85.4%	85.1%

표3에 의하면 확률 값을 10% 정도 반영하고 90%는 언어적 유사성을 반영할 때 가장 높은 정확성을 나타낸다.

후처리 단계에서는 추출된 제목 중에 정규화가 필요한 제목의 비율과 정규화 한 제목의 유효성에 대한 수치를 나타내었다. 추출된 500개의 문서에서 변환이 수행된 제목의 숫자는 370개로 전체의 74%가 정규화 대상이 되었다. 이중 정규화로 인한 제목의 손실은 20개로 18.7%정도의 손실을 보였다. 이처럼 다소 높은 손실률을 보인 이유는 문법에 크게 어긋난 문장 때문에 형태소 분석이 제대로 이루어지지 않았기 때문으로 본다.

6. 결 론

본 논문에서는 테이블 객체에 의한 위치 정보로 선정한 제목 후보와 본문후보에 확률 모델과 언어적 유사계수를 적용하여 웹 페이지에서 중심 제목을 추출하는 방법에 대해 기술하였다. 이 시스템을 실제 웹 페이지 500개에 적용한 결과 95.1%의 정확성을 얻을 수 있었다.

웹 문서는 일반 텍스트와는 또 다른 언어적 특성을 담고 있다. 본 연구의 결과로 다음과 같은 관찰 결과를 얻을 수 있었다.
(1) 본문 객체의 인식규칙에 의해서 정확성이 크게 올릴 수 있음에 비추어, 제목은 본문을 기준으로 일정 거리에서 존재한다.
(2) 본문을 기준으로 일정거리에 존재하되, 특정위치에서 주로 발생하는 분포를 가지고 있다.
(3) 제목과 본문간의 언어적 유사도 역시 다른 기존의 연구에서와 마찬가지로 유용한 정보를 제공한다.

참고문헌

[1] N. Kushmerick. Gleaning the Web, IEEE Intelligent Systems, vol.14, no.2, pp. 20-22, 1999
[2] Lancaster, F. Wilfrid; Warner, Amy J. Information Retrieval Today. Arlington, Virginia, Information Resource Press. 314p. 1993
[3] L. Eikvil. Information Extraction from World Wide Web: A Survey, Report No. 945, ISBN 82-539- 0429-0, July, 1999
[4] 강승식. 한국어 분석 모듈 HAM version 5.0.0 국민대학교
[5] 백혜승. 한국어 문서 축약 시스템. 한국 과학 기술원, 석사 학위 논문 1992
[6] 안정수, 한글문서의 효과적인 검색을 위한 N-Gram 기반의 색인 방법. 한국 과학 기술원, 석사 학위 논문 1995