웹문서 자동 분류를 위한 하이퍼링크 기반 특징 가중치 부여 기법

이아람, 김한준
서울시립대학교 전자전기컴퓨터공학부
e-mail:chigaya06@gmail.com, khj@uos.ac.kr

A Hyperlink-based Feature Weighting Technique for Web Document Classification

A-Ram Lee, Han-Joon Kim,
School of Electrical and Computer Engineering, University of Seoul

요 약

기계학습을 이용하는 문서 자동분류 시스템은 분류모델의 구성을 위해서 단어를 특징으로 사용한다. 자동분류 시스템의 성능을 높이기 위해 보다 의미있는 특징을 선택하여 분류모델을 구성하기 위한 여러 연구가 진행되고 있다. 특히 인터넷상에서 사용되는 웹문서는 단어 외에도 태그정보, 링크정보를 가지고 있다. 본 논문에서는 이 두 가지 정보를 이용하여 웹문서 자동분류 시스템의 성능을 향상 시키는 방법 제안 한다. 태그 정보와 링크 정보를 이용하여 적절한 특징을 선택하고, 각 특징의 중요도를 계산하여 가중치를 구한다. 계산된 가중치를 각 특징에 부여하여 분류 모델을 구성하고 나이브 베이지안 분류기를 통하여 성능을 평가하였다

1. 서론

최근 스마트폰의 대중화와 SNS의 급격한 성장으로 인터넷은 문서 정보가 넘쳐 나고 있다. 블로그(blog), 페이스북(facebook)등을 통해 정보를 연결시킴으로써 종전의 TV, 신문과 같은 매체보다 빠르게 정보를 전달할 수 있다. 이러한 정보 연결 관계에서 연결 고리가 되는 단어를 하이퍼링크(Hyperlink)라하며, 하이퍼링크를 포함하고 있는 문서를 하이퍼텍스트문서(Hypertext document)라 한다. 흔히 인터넷상에 생성되는 웹문서(Web document)를 하이퍼텍스트 분서라 한다. 웹문서가 인터넷상에서 정보 전달에 큰 부분을 차지하면서 자연스럽게 웹문서의 자동분류시스템에 대한 관심도 증가하였다.

일반적으로 자동문서분류는 기계학습(Machine learning) 기법을 사용한다. 기계학습 방식은 분류를 위해 학습문서 집합으로부터 각 카테고리에 출현하는 특징 집합을 주요 인자로 하여 문서를 자동 분류 할 수 있는 패턴 또는 모델을 만드는 것이다. 대표적인 알고리즘은 나이브 베이즈(Naïve Bayes)[1], 지지벡터머신(Support Vector Machine)[2] 등을 들 수 있다. 특히 나이브 베이즈 알고리즘은 분류모델의 단순성에 비하여 성능이 우수한 편으로 평가되어 문서의 자동분류 시스템에 활용되고 있다.

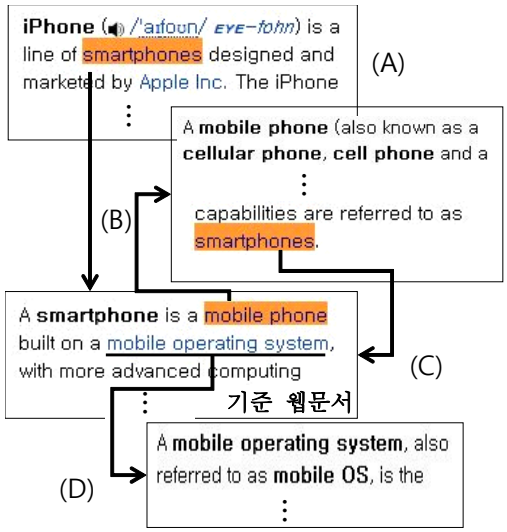
기계학습에 기반을 둔 자동분류 알고리즘에서 분류성능에 영향을 미치는 요소는 알고리즘 자체보다는 특징 선택(Feature selection)의 문제라 할 수 있다. 특징 선택이란 전체 특징 집합으로부터 카테고리를 가장 잘 표현할 수 있는 특징 부분 집합을 추출하는 과정을 말한다. 일반적으로 문서들은 단어들의 다중집합(Multi-set 또는 bag)으로부터 추출된 특징들로 표현되며, 모든 문서에 대한 특징들로 구성된 특징 공간(Feature space)으로부터 분류를 위한 모델을 만들게 된다. 따라서 문서분류의 성능을 향상시키기 위해서는 특징 선택을 통해서 특징 공간의 복잡도를 줄이고 왜곡된 특징들을 삭제하는 과정이 필수적이다.[3]

웹문서는 단어 이외에도 하이퍼링크(이하 링크)정보, 태그(Tag)정보를 갖고 있다. 링크 정보를 이용하여 연결된 웹문서의 단어를 특징으로 이용하거나, 웹문서 간의 관계를 고려하여 특징에 가중치를 부여하여 좋은 특징을 선택함으로써 자동문서 분류의 성능을 향상 시킬 수 있다.

2. 배경 지식 및 관련 연구

우선 링크로 연결된 웹문서간의 링크 관계 및 관련 용어를 살펴보자. 연결 관계의 기준이 되는 웹문서를 기준 웹문서(Target web document)라 하고, 기준 웹문서와 직접으로 링크하고 있는 웹문서를 인접 웹문서(Neighborhood web document)라 한다. 기준 웹문서를 링크하고 있는 웹문서를 선임문서(Predecessor)라 하며, 이는 기준 웹문서와 Inlink 관계를 맺고 있다고 부른다. 반대로 기준 웹문서가 링크하고 있는 문서를 후임문서(Successor)라 하며, 이는 기준 웹문서와 Outlink 관계를 맺고 있다고 부른다. 상호링크를 하는 경우 상호링크(co-linked) 웹문서라 하며, 해당 웹문서는 기 준 웹문서와 연관성이 높은 것으로 간주하여 선임문서이면서 후임문서로 중복하여 이용한다. (그림 1)에서 기준 웹문서 (C)는 “Smartphone”에 관한 문서이다. (A) 웹문서는 “iPhone”에 대한 웹문서로 “smartphones”라는 링크를 통하여 기준 웹문서로 연결된다. (A) 웹문서는

교신저자 : 김한준 (khj@uos.ac.kr)



(그림 1) 링크로 연결된 웹문서의 관계

(C) 웹문서의 선임문서가 되며, (C) 웹문서와 Inlink 관계를 맺고 있다. 같은 원리로 (D) 웹문서는 (C) 웹문서의 후임문서로, “mobile operating system”라는 링크를 통하여 연결된다. (B) 웹문서는 “mobile phone”와 “smartphones”라는 링크를 통하여 (C) 웹문서와 상호링크 관계를 맺고 있는 상호링크 웹문서이다.

웹문서의 자동분류에서는 인접 웹문서의 단어를 특징으로 이용할 수 있다. 하지만 인접 웹문서의 모든 단어를 특징으로 이용하는 것 불필요한 정보를 과도하게 포함하게 되어 분류 성능이 저하 된다. 반대로 인접 웹문서의 카테고리 정보만을 이용하는 경우도 필요한 정보가 누락되어 문서 분류에 좋은 성능을 보이지 못하였다.[4]

웹문서간의 관계를 계산할 수 있는 알고리즘으로 구글의 페이지랭크(PageRank)알고리즘[5]이 있다. 페이지랭크 알고리즘은 래리 페이지와 세르게이 브린이 검색 엔진에 대한 연구 기획의 일부로 개발한 것으로, 링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 알고리즘이다. 페이지랭크는 상대적으로 중요한 웹문서는 더 많은 다른 사이트로부터 링크를 받는다는 관찰에 기초하고 있다. 중요도가 높은 웹문서 일수록 다른 웹문서에서 링크를 건 Inlink의 수가 많다. 실제 페이지랭크 알고리즘으로 그 값을 계산하는 경우 사전에 계산 되어있는 웹문서의 중요도와 거미줄처럼 얽힌 링크 관계를 고려하여 계산되기 때문에 시간 복잡도가 높다. 효율적으로 이용하기 위해서는 이 시간 복잡도를 줄여야 할 필요가 있기 때문에 페이지랭크 알고리즘의 권위도(Prestige)를 단순한 계산법으로 바꾸어 사용한다.

3. 인접 웹문서를 이용한 특징 선택과 가중치 부여

웹문서의 자동분류의 성능을 높이기 위해 활용할 수 있는 웹문서의 대표적인 특징은 태그 정보와 링크 정보가 있다. 링크정보를 특징으로 활용한다는 것은 웹문서 간의 관계성을 이용하여 자동분류에 도움이 되는 특징을 추출함을 의미한다. 또한 두 정보를 통하여 특징의 중요도를 계산하여 그 중요도에 따라 각 특징에 가중치를 부여(Feature weighting)하였다.

태그 정보는 웹문서 내에서 단어의 중요도를 가늠할 수 있는 척도가 될 수 있다. 웹문서의 제목은 <Title>, 내용은 <Body>, 링크정보를 가진 단어는 <a>등으로 단어가 나타내는 정보에 따라 태그가 붙는다. 이 중에서도 중요한 정보를 담고 있는 제목 태그 <Title>, 앵커텍스트 태그 <a>를 활용하여 상대적으로 중요하다 간주되는 특징에 높은 가중치를 부여하였다.

링크 관계에 따라 웹문서의 중요도인 권위도를 계산할 수 있다. 계산된 권위도 값을 통하여 분류정확도를 높이기 위한 가중치를 계산하였다.

3.1 태그 정보를 이용한 특징 추출

웹문서에 포함된 태그는 단어가 웹브라우저에 표현되는 형태 및 의미를 결정하므로, 태그에 포함된 단어는 주요 특징으로 활용할 수 있다.

제목은 글의 내용을 대표한다. 그렇기 때문에 제목 태그 <Title>의 단어는 다른 태그 단어들보다 더 중요하다고 할 수 있기 때문에 특징으로 사용할 수 있다. 하지만 통계적 수치를 기반으로 하는 기계학습 자동분류에서는 특징의 등장빈도수가 곧 특징의 중요도를 나타내기 때문에, 제목 태그의 특징이라 하더라도 등장 빈도수가 적은 경우 중요 특징이 되지 못한다. 이러한 상황을 방지 하기 위해 제목 태그의 단어는 다른 태그의 단어보다 높은 가중치를 부여하여 중요도를 높여준다.

앵커텍스트는 웹문서를 연결시켜주는 링크를 가진 단어로 연결 관계를 대표하는 단어다. 다시 말하자면 앵커텍스트는 기준 웹문서와 인접 웹문서의 접점이 되는 단어다. 이 단어는 기준 웹문서의 내용을 담고 있으면서, 인접 웹문서의 내용도 포함하는 단어이다. 그렇기 때문에 앵커텍스트는 문서 분류를 위한 특징으로 유용하게 이용될 수 있다. 인접한 웹문서의 단 어를 특징으로 이용한 여러 실험 중에서는 앵커텍스트를 포함하여 앞뒤로 10개의 단어를 특징으로 사용한 경우가 가장 좋은 분류 성능을 보였다.[4] 본 연구에서도 앵커텍스트를 뒷받침해줄 수 있는 주변의 단어를 특징으로 이용하였다.

본 실험에서는 중요하다고 여겨지는 제목태그 <Title>, 앵커텍스트 태그 <a>, <a> 주변 단어 10개의 단어를 특징으로 활용하였다.

3.2 권위도

본 연구에서 웹문서의 자동분류를 위한 효과적인 특징을 추출하기 위해 웹문서의 중요도를 활용하고자 한다. 가장 저명한 방법으로 페이지랭크 알고리즘이 있으나, 기술적으로 페이지랭크값을 빠른 시간 내에 계산하기 어렵다. 그래

서 본 연구에서는 페이지랭크 알고리즘의 권위도를 단순한 계산법으로 변경하였다. 권위도의 계산 방법은 다음과 같다. 웹문서에서의 하나의 후임문서는 -1로, 하나의 선임문서는 +1로 간주하여 계산한다. 예를 들어 기준 웹문서는 6개의 선임문서와 3개의 후임문서를 가지고 있다고 하자. 이를 계산을 하면 -3+6 = 3으로 권위도는 3이 된다. 권위도가 높을수록 중요한 문서임을 뜻한다.

웹문서의 권위도가 음수가 될 수도 있다. 이 경우 그 웹문서가 다른 웹문서로부터 참조되는 경우보다 다른 웹문서를 참조하는 경우가 더 많다는 뜻이다. 결국 0보다 작은 권위도 값을 가지는 웹문서는 그것의 중요도가 낮다는 것이다. 특징 가중치 부여를 하기 위한 관점에서 중요도가 낮다는 것은 그 웹문서는 쓸모가 없다는 의미로 볼 수 있다. 결론적으로 권위도가 음수인 문서의 특징은 문서분류를 위한 특징으로 이용하기에 부적합하다. 그렇기 때문에 권위도가 양수인 경우의 웹문서의 특징만을 차용한다.

3.3 특징 가중치 부여

특징 가중치 부여 과정은 크게 두 단계로 나뉜다. 첫번째 단계에서는 특징의 웹문서 정보에 따른 권위도를 기반으로 한 가중치와 특징의 태그 정보에 따른 가중치로부터 최종 가중치를 계산한다. 두번째 단계에서는 계산된 최종 가중치를 통해 해당 특징의 단어 빈도수(Term Frequency)를 높여준다. 단어 빈도수의 확 장에 있어서는 보수적인 접근이 필요하다. 과도하게 높은 단어 빈도수는 중요한 특징을 불필요한 특징으로 만들어 자동문서분류의 성능을 저하시킬 수 있기 때문이다. 그렇기 때문에 가중치를 적절한 범위 내의 값으로 변환하여 사용하는 것이 필요하다.

분류에 사용되는 특징은 해당 웹문서의 권위도에 따라 특징 가중치 부여를 할 수 있다. 권위도가 음수인 경우는 고려하지 않는다.

$$P=\{P|0\le P\le P_{Max}\},\quad W_P=\{W_P|1\le W_P\le M\},\quad f: P\rightarrow W_P$$

<수식 1> 권위도 기반 특징 가중치의 변환

권위도는 0부터 최대 권위도값($P=\{P|0\le P\le P_{Max}\}$)을 가진다. 이 값을 1부터 M으로 변환하여 권위도 기반 특징 가중치 W_P를 구한다.($W_P=\{W_P|1\le W_P\le M\}$) M은 최대 권위도값 보다 작은 수로 변환된 가중치 W_P의 최대값이다.

태그정보에 따른 가중치를 W_T라 한다. 본 연구에서는 <Title> 태그 단어는 3, 앵커텍스트는 2, 앵커텍스트 주변 단어는 1.5로 각 태그에 따라 차등적으로 가중치를 부여하였다.

1차함수는 권위도와 동일한 비율로 가중치를 높여 주는 방법이다. 2차함수의 경우 권위도가 높을수록 가중치가 증가 비율이 낮아진다. 이는 권위도가 높은 웹문서들의 가중치는 상대적인 차이가 적고, 권위도가 낮은 웹문서들의 가중치 차이가 크다는 것을 의미한다. 반대로 1/2차함수의 경우 권위도가 높을수록 가중치 증가 비율이 높아진다. 권위도가 높은 웹문서들의 가중치 차이가 권위도가 낮은 웹문서들의 가중치 차이보다 크다는 것을 의미한다.

$P=4$일 때, $W_P=M$

1차 함수: $W_P=\frac{M-1}{4}P+1$

2차 함수: $W_P=\frac{M-1}{16}P^2+1$

1/2차 함수: $W_P=\frac{M-1}{2}\sqrt{P}+1$

<수식 2> 권위도 가중치 변환 공식 3가지

$$\sum(W_P+W_T)=W_F,\quad W_F=\{W_F|1.5\le W_F\le W_{FMax}\},\quad W=\{W|1\le W\le N\},\quad f: W_F\rightarrow W$$

<수식 3> 최종 가중치 변환

계산된 최종가중치 W를 통해 특징의 단어 빈도수를 높여준다. 예를 들어 현재 단어 빈도수가 4인 특징 A 가 있다. 특징 A의 최종 가중치 가 1.43 이라면, 단어 빈도수를 1.43배 확장한다. 최종 가중치 를 적용한 특징 A의 단어 빈도수는 5.72가 된다. 하지만 단어 빈도수는 자연수 정수이므로 반올림하여 계산한다. 최종적으로 특징 A의 단어 빈도수는 6이 된다.

4.성능평가

4.1 실험 환경

본 연구에서는 제안 기법의 성능을 평가하기 위하여 Web-Kb을 이용한 실험을 실시하였다. Web-Kb는 웹문서 분류의 성능을 평가하기 위해 보편적으로 사용되는 문서 집합이다. 1997년 CMU text learning group이 만든 Web-kb는 여러 대학들의 컴퓨터 과학과 웹문서를 모아 놓은 데이터 집합이다. 본 실험에서는 Cornell(7개의 카테고리)을 실험 평가를 위한 데이터로 사용하였다. 전체 문서의 70%를 학습을 위해 사용하였고, 나머지 30%는 분류 성능을 평가하는데 사용하였다. 분류는 나이브 베이지안 문서분류기의 하나인 Mallet 시스템[6]을 이용하였다. 분류의 성능은 각 문서가 속할 카테고리를 얼마나 정확하게 분류하는가를 평가하였다. 이를 분류 정확도라 한다.

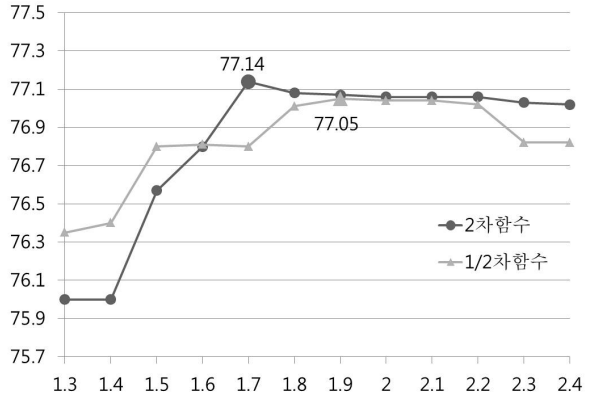
4.2 실험 결과

통계적인 방법을 이용하는 자동문서분류기법에서는 단어 빈도수의 확장에 대해 보수적인 접근이 필요하다. 단어 빈도수의 확장에 적절한 값을 구하기 위해 가중치의 범위를 변환하여 사용하였다. 권위도 가중치의 변환 공식에 있어서 1차, 2차,1/2차 함수를 이용하였다. 가중치의 변환 공식은 본 연구에서 사용한 1차, 2차, 1/2차 함수 외에도 여러 함수를 사용할 수 있다. 하지만 변환 공식을 사용하는 목적이 가중치의 상대적인 값을 확장한다는 면에서 볼 때 모든 함수에 대하여 실험을 하는 것은 큰 의미가 없다.

<표 1>은 권위도 가중치 W_P의 최대값 M을 2로, 최종

<표 1> 권위도 가중치의 변환 공식에 따른 자동문서분류 정확도

실험 구분	변환 범위	분류 정확도(%)
Baseline 1		72.81
Baseline 2	M = 2	75.90
1차함수	M = 2, N = 2	76.12
2차함수	M = 2, N = 2	77.06
1/2차함수	M = 2, N = 2	77.04



(그림 2) 최종 가중치의 최대값 N에 따른 자동문서 분류 정확도

가중치 W의 최대값 N을 2로 설정하고 권위도 가중치 W_P의 변환 공식에 따른 자동문서분류 정확도를 비교한 결과이다. Baseline 1은 링크정보와 태그정보를 고려하지 않은 문서에 대한 자동분류기법이다. Baseline 2는 웹문서에서 태그 기반 가중치를 적용한 기법이다. 다른 실험과 동일하게 최종 가중치의 최대값은 M이다. 2차함수로 변환하였을 때 77.06%로 가장 높은 분류 정확도를 보였다. 1/2차함수로 변환한 경우의 분류 정확도는 77.04%로 2차함수를 이용한 경우 보다는 낮지만 매우 유사한 분류 정확도를 보였다.

(그림 2)는 권위도 가중치 W_P의 범위 변환의 두 공식을 비교한 결과이다. X축은 최종가중치 W의 최대값 N을 의미하고, Y축은 분류 정확도를 의미한다. 권위도 가중치 W_P의 최대값 M이 2일 때 최종 가중치 W의 최대값 N에 따른 분류 정확도를 볼 수 있다. 변환 공식으로 2차함수를 이용하는 경우, 최종 가중치 W의 최대값 N이 1.7일 때 가장 높은 분류정확도인 77.14%를 보여 Baseline 2보다 1.24% 향상되었다. 변환 공식으로 1/2차함수를 이용하는 경우 최종 가중치 W의 최대값 N이 1.9일 때 분류정확도 77.05%를 보였다. 최종가중치 W의 최대값 N이 커질수록 분류 정확도는 하락했다.

단어빈도수의 변화가 클수록 분류정확도가 떨어지는 결과를 보였다. 이는 단어빈도수의 큰 변화가 왜곡된 특징을 만드는 것이라 볼 수 있다. 권위도 가중치의 변환 공식에 따른 문서분류의 정확도는 1차함수를 이용하여 변환한 결과보다 2차, 1/2차함수를 이용하여 변환한 결과가 좋은 성능을 보였다. 이는 권위도에 따른 가중치의 값의 절대치가 아닌 상대적인 차이를 확대해 주는 것이 문서 분류의 성능을 향상 시킬 수 있는 방법이 된다는 것을 보여준다.

5. 결론

본 논문은 웹문서의 특징인 태그 정보와 웹문서 간의 연결 관계 정보를 이용하여 자동문서분류의 성능을 향상시키는 방안을 제시하였다. 링크 정보를 이용하여 인접 웹문서의 단어를 특징으로 사용하고, 웹문서의 권위도를 계산하여 특징의 가중치를 계산하였다. 태그 정보를 이용하여 적절한 특징을 선택하고 특징의 중요도에 따라 가중치를 부여하였다. 웹문서의 중요도는 페이지랭크 알고리즘보다 단순한 방법인 권위도로 계산하였다. 가중치를 통하여 단어 빈도수를 확장하는 방법으로 모델을 개선하고자 하였다. 단어 빈도수의 과도한 확장을 막기 위하여 가중치를 일정 범위로 변환하여 사용하였다.

성능이 어느 정도 향상되는 지를 평가하기 위하여 Web-Kb의 집합 중 7개의 카테고리를 갖는 Cornell을 실험 데이터로 사용하고, 나이브 베이지안 문서 분류기로 분류정확도를 계산하였다. 그 결과 태그 정보만을 이용했던 기존 기법인 75.90%보다 최대 1.16% 향상된 77.06%의 분류 정확도를 보였다. 이는 태그 정보만을 이용하는 것보다 링크 정보를 함께 사용하는 것이 분류 성능 향상에 좋은 영향을 준다는 것을 시사한다. 향후 링크 정보를 활용하는 방안을 연구하여 분류 성능을 더 높일 수 있을 것이다.

6. 감사의 글

이 논문은 2010년 정보(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (과제번호: 2010-0025212)

참고문헌

[1] T.M. Mitchell, “Bayesian Learning,” Machine Learning, McGraw-Hill, pp154-200, 1997.

[2] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proceedings of the 10th European Conference on Machine Learning (ECML’98), pp137-142, 1998.

[3] 김한준, 장재영, “점진적 특징 가중치 기법을 이용한 나이브 베이즈 문서분류기의 성능 개선,” 정보처리학회 논문지B 제15-B권 제5호, 2008.

[4] H. Utard and J. Furnkranz, “Link-Local Features for Hypertext Classification,” Semantics, Web and Mining: Joint International Workshops, EWMF|KDO. Lecture Notes in Computer Science, val.4289, pp 58-69 Springer, 2005.

[5] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998.

[6] Mallet (MAchine Learning for LanguagE Toolkit), http://mallet.cs.umass.edu/