자동문서분류를 위한 워드넷 기반 특징 가공 기법

A WordNet-based Feature Engineering Method for Text Classification

노준호(Jun-ho Roh)*, 김한준(Han-joon Kim)**, 장재영(Jae-young Chang)***
loece@naver.com, khj@uos.ac.kr, jychang@hansung.ac.kr

초 록

자동문서분류시스템의 성능을 높이기 위해서는 최적의 특징 집합을 구성하여 분류모델을 구축하는 것이 중요하다. 본 논문에서는 분류모델의 개선을 위해 워드넷(WordNet)을 이용하여 의미정보가 풍부한 특징을 생성하는 기법을 제안한다. 기본 아이디어는 학습문서집합 내에서 중요도가 높은 특징을 선정하여, 선정된 특징의 동의어 및 상위어를 추가함으로써 특징의 의미 확장을 수행하는 것이다. 또한 특징집합의 확장 가공을 위해 단어와 클래스간의 유사도를 계산함으로써 문서분류에 도움이 될 수 있는 단어들을 선별하였다. 결과적으로 제안한 특징 가중치 기법을 이용하여 나이브 베이즈 분류기의 성능을 개선하였다. 제안 기법을 평가하기 위해 표준 테스트 집합인 Reuters-21578 문서집합을 이용하여 실험을 수행하였다.

1. 서론

최근 들어 웹 문서의 급격한 증가에 따라 자동문서분류의 중요성이 점점 높아져 가고 있다. 자동문서분류(text classification)란 학습문서를 이용하여 분류모델을 구축하고, 새로 유입된 문서를 분류모델을 통해 자동으로 클래스를 예측하는 것을 의미한다. 텍스트 마이닝(text mining) 연구분야에서 자동문서분류 시스템의 정확도를 높이는 것이 중요한 문제 중 하나이다.

최근 자동문서분류 기법은 주로 기계학습(machine learning) 기술을 사용한다. 일반적인 문서분류를 위한 기계학습 방법으로는 나이브 베이즈(Naïve Bayes), 지지벡터기계(Support Vector Machine), 인공 신경망(Neural Network) 등이 있다. 기본적으로 이러한 기법들은 학습문서 집합 내에서의 단어의 빈도수를 이용하여 분류모델을 구축한다. 만약 워드넷(WordNet)과 같은 어휘사전을 이용하여 특징(feature)의 동의어(synonym) 및 상위어(hypernym) 등의 의

* 서울시립대학교 전자전기컴퓨터공학부 석사과정
** 서울시립대학교 전자전기컴퓨터공학부 교수
*** 한성대학교 컴퓨터공학과 교수

미관계를 고려한 분류모델을 구축한다면, 문서분류의 향상을 기대할 수 있다. 그런데 워드넷을 사용하여 분류모델을 구축할 때 특별한 가공 없이 학습문서 내의 모든 단어의 의미관계 단어를 추가하였을 경우에는 오히려 문서분류의 정확도를 떨어뜨리는 효과를 낳는다. 이는 학습문서 내의 모든 특징의 중요도가 동일하지 않고, 워드넷을 통해 특정 클래스에서 확장된 의미관계 단어들 중 상당수가 다른 클래스에도 확장되어 오히려 클래스를 분별할 수 없게 만들기 때문이다. 또한 워드넷에서의 의미관계 단어들 중에서 클래스 내에서의 중요도가 떨어지는 단어들도 많이 존재한다. 이러한 문제를 해결하기 위한 문서분류에 도움이 되는 의미관계 단어 선별 과정이 주요 연구 이슈가 된다.

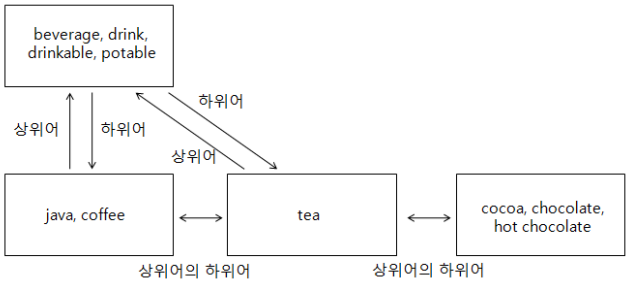
본 논문에서는 이 문제에 초점을 맞추어 워드넷에 기반한 새로운 특징 가공 기법을 제안한다. 제안기법의 기본 아이디어는 특징 선택을 통하여 중요 특징을 선정하고, 선정된 특징들을 대상으로 워드넷을 사용하여 의미관계 단어를 확장하는 것이다. 워드넷을 통해 확장되는 의미관계 단어가 실제 문서분류에 도움이 되는 지를 판별하기 위해 클래스를 대표하는 특징벡터를 정의하고 이 클래스 특징벡터와 확장되는 의미관계 단어와의 유사도를 계산하여 특정 의미관계 단어를 선별하게 된다. 이 과정을 통해 선별된 의미관계 단어 집합이 분류모델에 반영되어 결과적으로 자동문서분류시스템의 성능을 높일 수 있다. 본 연구에서는 제안기법을 평가하기 위해서 Reuters-21578과 20Newsgroups 문서 집합을 이용한 실험을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 워드넷 관련연구와 자동문서분류에 가장 많이 이용되는 나이브 베이즈 분류모델에 대해 설명한다. 3장에서는 선별된 특징들의 의미관계 단어를 선별하고, 이를 이용하여 나이브 문서분류기의 분류모델을 개선시키는 특징 가공 기법을 기술한다. 4장에서는 제안 기법의 실험 결과를 기술하고 5장에서는 결론과 향후 연구 과제를 제시한다.

2. 관련연구

2.1 워드넷

워드넷은 일종의 영어 어휘간 관계정보를 담은 사전으로서 단어의 의미를 synset 단위로 표현한다. 또한 워드넷에서는 각 의미에 대한 동의어, 상위어 같은 의미관계(relation)를 알 수 있다. 예를 들어 java는 3 가지의 의미를 가지는데, 컴퓨터 언어인 java, 커피를 의미하는 java, 섬을 뜻하는 java가 그것이다. 그 중 커피의 의미로 사용된 java의 유사어는 coffee가 되고 상위어는 beverage가 된다. 그림 1 은 이러한 워드넷 계층구조를 나타낸다.



<그림 1> 워드넷 의미관계

워드넷을 사용하여 문서분류시스템의 성능을 향상시키기 위한 연구가 활발하다. [2]에서는 감독자가 개입하여 주어진 특징에 해당하는 워드넷 의미관계 단어들 중에서 가장 적합한 동의어를 골라내어 이를 분류

모델에 반영하여 문서분류 정확도를 높였다. [3]에서는 모든 특징을 대상으로 단어의미 중의성 해소기법을 통해 분류모델을 개선하는 연구를 하였다. [4]에서는 높은 단어 빈도수를 가지는 특징의 동의어와 상위어를 추가하여 좋은 성과를 보여주었다. [5]에서는 모든 특징의 동의어 및 상위어를 추가하여 분류모델을 구축하였다. 그 결과는 문서집합에 따라 성능 변동이 나타났다.

대부분의 기존 연구에서는 모든 특징을 사용하여 의미관계 단어를 추가하였으며, 또한 모든 동의어 및 상위어를 추가하였다. 하지만 이는 분류에 도움이 되지 않는 단어가 다수 추가되는 문제점이 있다. 본 논문에서는 특징 선택을 통해 선정된 특징에 해당하는 의미관계 단어만을 추가하는 방안을 제안한다. 또한 2 차적으로 선별된 의미관계 단어 중 특징 가공 기법을 사용하여 분류모델의 개선에 도움이 되는 특징을 선별한다.

2.2 나이브 베이즈 문서분류 기법

나이브 베이즈 문서분류기에서는 학습문서로부터 클래스에 소속될 분류모델에 대한 인자들을 추정하게 된다. 추정된 분류 모델 $\hat{\theta}_{NB}$는 다음과 같은 두 개의 인자로 구성되어 있다.

$$\hat{\theta}_{NB} = (\hat{\theta}_{w|c}, \hat{\theta}_{c}) \qquad (1)$$

여기서 $\hat{\theta}_{w|c}$는 클래스 c에 속하는 문서집합에서 임의의 단어가 w일 확률값을 나타내고, $\hat{\theta}_{c}$는 전체문서집합에서 임의 추출한 문서가 클래스 c일 사전확률값을 나타낸다.

나이브 베이즈 알고리즘에 의한 문서분류는 새로운 문서 d_i의 클래스를 예측하기 위해 베이즈 정리에 의해 다음과 같이 클래스의 사후확률값을 추정함으로써 이루어진다.

$$\Pr(c_j, d_i) = \frac{\Pr(c_j)\Pr(d_i|c_j)}{\Pr(d_i)} \qquad (2)$$

여기서 $\Pr(c_j)$는 전체문서집합에서 임의 추출한 문서가 클래스 c_j에 속할 사전확률값을 의미하며, $\Pr(d_i|c_j)$는 클래스 c_j에 속하는 문서집합에서 임의로 추출한 문서가 d_i일 확률값을 의미한다. $\Pr(d_i)$는 전체문서집합에서 임의 추출한 문서가 d_i일 확률값을 의미한다. 이 식을 이용하여 문서 d_i는 전체 클래스 집합 C중에서 사후확률값으로 $\mathrm{argmax}_{c_j \in C} \Pr(c_j, d_i)$ 인 클래스 c_j로 분류된다. 여기서 문서 d_i는 단어들의 다중집합인 $(w_{i1}, w_{i2}, \ldots, w_{i|di|})$로 표현된다. 나이브 베이지안 분류기는 문서에 출현하는 단어들은 서로 독립적이라는 가정을 한다. 이로 인해 분류함수는 다음과 같이 간단히 표현할 수 있다.

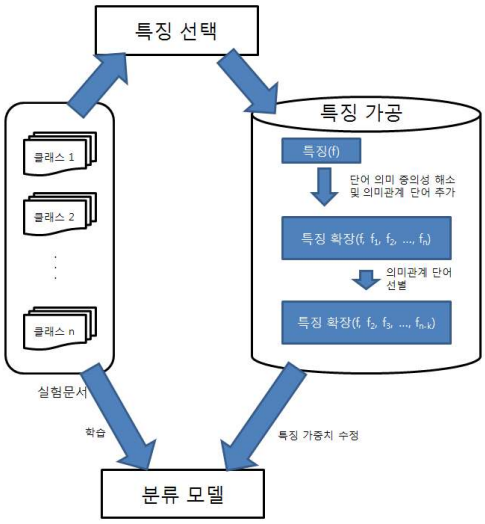
$$\Phi_{\hat{\theta}_{NB}} = \mathrm{argmax}_{c_j \in C} \Pr(c_j, d_i) \qquad (3)$$
$$= \mathrm{argmax}_{c_j \in C} \prod_{k=1}^{|d_i|} \Pr(w_{ik}|c_j) \Pr(c_j)$$

나이브 베이즈 학습기법은 문서분류를 위해 주어진 학습문서 집합을 한 번의 작업으로 인자들을 결정할 수 있다. 또한 기존 모델의 인자를 변경함으로써 쉽게 분류모델을 개선할 수 있다. 본 논문에서는 워드넷을 기반으로 학습문서로부터 워드넷의 의미관계 단어를 선별하고, 선별된 단어들을 이용하여 분류모델 인자를 변경함으로써 문서분류기의 성능을 향상시키는 방법을 고안하였다.

3. 워드넷 기반 특징 가공 기법

앞에서 언급한 바와 같이 분류모델의 기본요소는 학습문서의 특징이다. 특징은 문서 내의 단어(word)나 용어(term)등을 뜻한다.

본 논문에서 문서분류 성능 향상을 위해 세가지 방안을 제안한다. 첫째, 특징 선택을 통해 클래스 내에서 중요한 단어들을 선정하고, 선정된 특징에 대해 워드넷을 이용하여 특징 가공을 수행한다. 워드넷을 이용할 때, 이전연구에서 사용한 단어 의미 중의성 해소(Word Sense Disambiguation) 기법을 적용하여 특징의 의미를 정하고, 워드넷 의미관계 중 동의어와 상위어를 사용한다. 둘째, 워드넷의 의미관계 단어들 중 클래스 내에서 중요한 단어들을 선별한다. 이는 단어와 클래스간의 유사도 함수를 통해 이루어진다. 셋째, 나이브 베이지안 분류법을 개선시키기 위해 모델 인자인 특징 가중치를 수정하는 것이다. 나이브 베이지안 문서분류기는 $\mathrm{argmax}_{c_j \in C} \Pr(c_j, d_i)$를 추정함으로써 이루어진다. 따라서 문서 내의 단어 w가 클래스 c_x에서 중요한 단어라면 다른 클래스 c_y에서의 확률값보다 항상 높은 것이 바람직하다. 그래서 단어와 클래스간의 유사도를 계산하여, 만약 단어 w와 c_x의 유사도가 c_y의 유사도 보다 높지만 확률값 $\Pr(w, c_x)$가 $\Pr(w, c_y)$ 보다 낮을 경우에는, c_x에서의 단어 w를 추가하여$\Pr(w, c_x)$의 확률값을 높인다. 이는 해당 클래스에서의 단어 w의 특징 가중치를 수정함으로써 분류모델을 개선하는 효과를 준다. 그림 2는 본 논문에서는 제안하는 기법을 도시한 것이다.



<그림 2> 특징 가공 기법

3.1 의미관계 확장을 위한 특징의 선정

특징 가공을 수행할 때 문서 내에 모든 용어들의 의미를 확장하는 것은 오히려 문서 분류의 성능을 저하시킨다. 본 논문에서는 특징 선택을 통해 중요도가 높은 단어들에 대해서만 의미관계를 확장한다. 우선 특징선택 방법 중 많이 사용되는 χ^2-통계량을 이용하여 각 클래스 내에서의 단어의 중요도를 산정하고, 이 중 상위 10%의 단어들을 선정하여 특징의 의미관계 확장에 사용하였다.

χ^2-통계량을 이용한 특징 선택

본 논문에서 사용한 특징 선택 방법은 χ^2-통계량 방식으로, χ^2-통계량은 모든 특징에 대해 문서집합의 각 클래스의 주제와의 연관성을 평가하여 문서와 가장 연관성이 큰 특징을 선택할 수 있게 된다. 식 (4)는 χ^2-통계량 $\chi^2(c, w)$을 표현한 것이다. 여기서 주어진 단어 w와 클래스 c의 관련성 정도를 산정하며, 값이 작을수록 서로 독립적인 것을 의미하고 값이 클수록 상호 연관

성이 크다는 것을 나타낸다.

$$\chi^2(c,w) = \frac{N\times(DF(w,c)\times DF(\bar{w},\bar{c}) - DF(w,\bar{c})\times DF(\bar{w},c))^2}{(DF(w,c)+DF(\bar{w},c))\times(DF(w,c)+DF(w,\bar{c}))\times(DF(w,\bar{c})+DF(\bar{w},\bar{c}))\times(DF(\bar{w},c)+DF(\bar{w},\bar{c}))} \quad (4)$$

여기서 $DF(w,c)$는 w를 포함하는 문서 중 클래스 c에 속하는 문서의 빈도수를 나타내며, $DF(\bar{w},c)$는 클래스 c에 속하는 문서 중 w를 포함하지 않는 문서의 빈도수를 나타낸다. $DF(w,\bar{c})$는 w를 포함하는 문서 중 클래스 c에 속하지 않는 문서의 빈도수를 나타내며, $DF(\bar{w},\bar{c})$는 클래스 c에 속하지 않는 문서 중 w를 포함하지 않는 문서의 빈도수를 나타낸다. N은 총 문서의 수를 나타낸다.

$\chi^2(c,w)$ 값을 통해 클래스 내에서의 모든 특징의 중요도를 산정한 후, 이 중에서 중요도가 높은 단어들을 워드넷을 이용하여 동의어, 상위어 등의 의미관계 단어를 추가함으로써 특징집합이 확장된다.

3.2 워드넷 의미관계 단어 선별

문서 분류에 도움이 되는 단어는 클래스의 주제를 나타낼 수 있는 단어들이다. 본 논문에서는 이러한 단어를 판별하기 위해 단어와 클래스 간의 유사도 함수를 제안한다. 유사도 값이 임계값 α 이상일 때, 해당 단어는 문서분류에 도움이 된다고 가정한다.

단어와 클래스 간의 유사도는 단어간의 유사도 계산을 응용하여 산정할 수 있다. 이 유사도 함수는 주어진 단어 w와 클래스 c의 모든 단어집합 $\{f_1, f_2, \ldots, f_n\}$와의 유사도를 각각 계산하고 계산된 유사도의 평균값으로 정의한다. 여기서 클래스에 존재하는 모든 단어와의 유사도를 계산하는 것은 질적으로 그리고 시간적으로 효율을 떨어뜨린다. 그래서 클래스를 대표할 수 있는 단어들과의 유사도를 계산하여 이 유사도의 평균을 구하는 것이 합리적이다. 이를 위해 본 논문에서는 χ^2 -통계량을 통해 수치가 높은 M개의 단어로 클래스를 대표하는 단어를 선정한다.

단어간 유사도 계산

문서집합 내에서 특정 단어 w_x와 이와 다른 단어 w_j사이의 유사도를 계산 할 수 있다. 예를 들어, 표 1에서 단어 w_x에 대한 문서벡터 $< w_x >$ 는 {3, 4, ..., 2}이 되고, w_j에 대한 문서벡터 $< w_j >$는 {10, 5, ..., 1}이 된다. 이 두 단어 간의 유사도는 식 (5)로 계산한다.

<표 1> 문서집합 내에서의 단어 빈도수

	w_1	w_2	...	w_x	...	w_j
문서 1	2	1	...	3	...	10
문서 2	6	0	...	4	...	5
...	...	...	...	...	...	
문서 N	9	10	...	2	...	1

$$sim(< w_x >, < w_j >) = \frac{<w_x>\cdot<w_j>}{|<w_x>|\times|<w_j>|} \quad (5)$$

이 방식을 응용하여 단어와 클래스간의 유사도를 계산할 수 있다.

단어와 클래스간의 유사도 계산

본 논문에서는 단어 w_x와 클래스 c_i간의 유사도 함수 $Score(w_x, c_i)$는 식 (6)과 같이 정의한다.

$$Score(w_x, c_i) = \sum_{j=1}^{M} \frac{sim(<w_x>, <w^*_j>)}{M} \quad (6)$$

여기서 w^*_j는 클래스 c_i를 대표하는 M개의 단어 중 j번 째 단어를 의미한다. 단어 w_x와 M개의 클래스를 대표하는 단어와의 유사도를 계산한 후, 그 값들의 평균을 유사도 값으로 정의하는 것이다. 결국 이는 해당 단어 w_x가 어느 정도 클래스 c_i와 연관성이 있는지 평가하는 방법이다. 점수가 높을수록 클래스에서의 해당 단어의 중요도는 높아진다. 그리고 유사도가 주어진 임계값 α 이상의 의미관계 단어들을 선별하여 이를 분류 모델의 구성에 참여시킨다.

3.3 나이브 베이즈 분류모델의 특징 가중치 개선

본 논문에서 자동문서분류를 위해 채택한 나이브 베이즈 분류기는 단어의 빈도수가 분류모델의 주요 인자가 된다. 그러므로 특징확장 과정을 통해 선별된 의미관계 단어들을 추가하여 분류모델을 개선하는데, 여기서 클래스 별로 특징을 추가하는 비율을 결정해야 한다. 비율을 결정 할 때에는 의미관계 단어의 클래스를 고려하여야 한다. 예를 들어 의미관계 단어가 하나의 클래스에서만 존재할 경우는 가장 이상적인 경우로 해당 클래스의 주제와 밀접한 연관이 있는 단어이다. 이 경우에는 해당 단어를 클래스에서 가장 높은 빈도수를 가지는 단어의 빈도수만큼 추가 한다. 그렇지 않은 경우에는 유사도를 기준으로 유사도 순으로 각 클래스에서의 확률값이 높아지도록 보정해 준다. 예를 들어 단어 w와 클래스 c_x에서의 유사도가 c_y 보다 높다면 $\Pr(w, c_x) > \Pr(w, c_y) * \delta$ 가 성립되도록 클래스 c_x에 단어 w를 추가한다. 여기서 δ는 가중치 정도를 결정하는 인자로 본 논문에서는 1.02로 고정하였다.

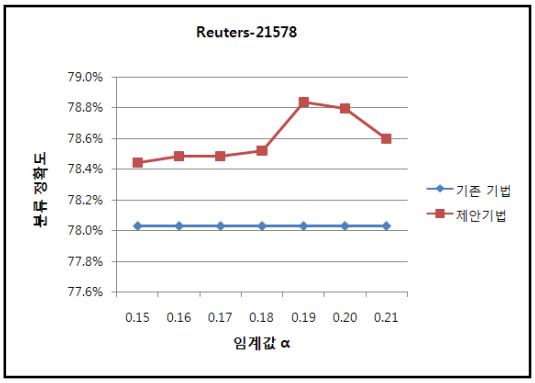
4. 성능 분석

4.1 실험 환경

본 논문에서 제안한 기법의 성능을 검증하기 위해서 Reuter-21578 문서집합을 이용한 실험을 수행하였다. Reuters-21578는 135개의 클래스로 묶여진 21,578개의 뉴스 기사들로 구성되어 있다. 이 문서집합은 문서들의 분포가 균형 잡히지 못하고 특정 클래스들에 집중되는 경향을 보인다. 따라서 본 논문에서는 이러한 불균형을 해결하기 위해서 클래스를 20개로 줄여서 실험을 수행하였다. 각 문서집합에 대해서 70%는 학습을 위해 사용하였고 나머지 30%는 분류 성능을 평가하는데 사용하였다. 분류시스템은 MALLET[6]을 이용하였고, 제안 기법의 분류기를 평가하기 위해 기존 분류기와의 정확도를 비교하였다.

4.1 실험 결과

그림 3은 Reuters-21578 문서집합에서 임계값 α의 변화에 따른 분류 정확도를 보여주고 있다. α가 0.19일 때 분류 정확도가 가장 높은 것을 확인할 수 있다. α가 0에 가까우면 모든 의미관계 단어집합을 추가하므로 오히려 문서분류 정확도가 낮아질 가능성이 높고, 반대로 너무 크면 의미관계 단어집합의 개수가 너무 적어져서 성능 향상을 기대하기 어렵다.



<그림 3> 임계값 변화에 대한 분류 정확도 (Reuters-21578)

5. 결론

본 논문에서는 워드넷에 기반한 의미관계 단어집합에서 특징을 선별하고, 나이브 베이즈 분류기의 성능을 향상하기 위한 특징 가공 기법을 제안하였다. χ^2-통계량을 이용하여 중요 특징의 의미관계 단어집합을 구성한 후, 단어와 클래스간의 유사도 함수를 이용하여 문서분류에 도움이 되는 특징을 선별하였다. 향후, 제안 기법의 성능을 더욱 높이기 위하여 워드넷의 동의어, 상위어 외의 다른 의미관계를 연구할 것이며, 더욱 정밀한 특징 가중치 보정함수를 고안할 계획이다.

6. 감사의 글

이 논문은 2010년 정보(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것이며(과제번호: 2010-0025212), 또한 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임.(과제번호: NRF-2011-0022445).

참고문헌

[1] 김한준, 장재영, “점진적 특징 가중치 기법를 이용한 나이브 베이즈 문서분류기의 성능 개선”, 한국정보처리학회 논문지, 15권 5호, 2008.

[2] de Buenaga Rodriguez, M. Gomez-Hidalgo, J. and DiazAgudo, B., "Using WordNet to complement training information in text categorization", In Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP'97), 150-157, 1997.

[3] Hotho, A., and Bloehdorn, S., "Boosting for text classification with semantic features", In Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 70-87, 2004.

[4] Jensen, L., and Martinez, T., "Improving text classification by using conceptual and contextual features", In Proceedings of the Workshop on Text Mining at the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00), 101-102, 2000.

[5] Scott, S., and Matwin, S., "Text classification using WordNet hypernyms", In Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (Coling-ACL'98), 45-52, 1998.

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