청킹 기반 특징 추출을 통한 문서 분류 시스템의 성능 향상

Improving text classification systems through chunking-based feature extraction

이아람(Aram Lee)*, 김한준(Han-joon Kim)**
chigaya06@gmail.com, khj@uos.ac.kr

초 록

기계학습을 이용하는 자동문서분류 시스템은 분류모델의 구성을 위해서 개별적인 단어(word)를 특징(feature)으로 사용한다. 그래서 중의성 또는 모호성을 가지는 단어는 분류모델의 성능을 저하시키는 요인으로 작용한다. 문서분류의 성능을 높이는 방법 중의 하나는 분류모델의 특징으로 사용되는 단어가 가질 수 있는 중의성을 제거하는 것이다. 본 논문에서는 청킹(chunking) 기술을 이용하여 분류모델을 개선하는 방안을 제시한다. 청킹이란 자연어 처리 기법 중의 하나로서 정보를 의미 있는 단위로 묶어주는 기술이다. 청킹을 통하여 단어가 가지는 중의성을 최소화하여 자동 분류에 도움이 되는 특징을 만들어 낼 수 있다. 청킹을 통해 생성된 새로운 특징 집합에서 문헌 빈도수를 기준으로 최적의 특징을 선정하고, 선정된 최적의 특징 집합을 이용하여 다음의 세가지 방법으로 분류모델을 개선하였다. (1) 학습할 문서 집합에서의 선정된 특징들의 출현빈도수를 높여주거나, (2) 테스트 집합에서 선정된 특징을 가진 문서를 찾아 특징의 출현빈도수를 높여주는 방법, (3) 테스트 집합의 일부로 사전 분류를 실행하여 분류의 오류 정보를 얻어 점진적으로 학습 모델을 개선하는 방법을 사용하였다..

1. 서론

초고속 인터넷의 빠른 보급과 스마트폰의 대중화로 인터넷은 문서 정보가 넘쳐 나고 있다. 블로그(blog), 뉴스 등과 같은 온라인 문서들이 꾸준한 증가 추세를 보임에 따라 자동문서분류 시스템은 스팸 메일 분류, 사용자의 선호도를 고려한 뉴스 기사의 분류 등 다양한 분야에서 응용되고 있다.

본 연구는 2010년 정보(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행되었음. (과제번호 : 2010-0025212)
* 서울시립대학교 전자전기컴퓨터공학과 석사과정
** 서울시립대학교 전자전기컴퓨터공학부 부교수

최근의 자동문서분류 기법은 주로 기계학습(machine learning)기술을 사용한다. 기계학습 방식은 분류를 위해 학습문서 집합으로부터 각 카테고리에 출현하는 특징 집합을 주요 인자로 하여 문서를 분류 할 수 있는 모델을 만든다. 대표적인 관련 알고리즘은 나이브 베이즈(Naïve Bayes)[1], 지지벡터머신(Support Vector Machine)[2] 등을 포함한다. 특히 나이브 베이즈 알고리즘은 분류모델의 단순성에 비하여 성능이 우수한 편으로 평가되어 자동 문서 분류 시스템의 개발을 위해 자주 활용되고 있다.

일반적인 기계학습에 기반을 둔 문서 분류 알고리즘에서 분류성능에 영향을 미치는 문제들 중 우리가 주목할 것은 특징 선택(feature selection)의 문제이다. 특징 선택은 '차원의 저주(curse of dimensionality)'라고 불리는 문제를 해결하기 위한 방법으로 제시 되었다. 차원의 저주란 문서의 수가 증가할수록 특징이라고 불리는 단어(word) - 또는 용어(term) - 의 수가 지수적으로 증가하는 문제를 말하며, 특징 선택이란 이러한 특징들 중에서 카테고리를 가장 잘 표현할 수 있는 일부 특징을 추출하는 과정을 말한다. 일반적으로 문서들은 단어들의 다중집합(multi-set 또는 bag)으로부터 추출된 특징들로 표현되며, 모든 문서에 대한 특징들로 구성된 특징 공간(feature space)은 결국 수많은 단어들이 모인 어휘(vocabulary)를 구성하게 된다. 따라서 문서분류의 성능을 향상시키기 위해서는 특징 선택을 통해서 특징 공간을 줄이고 왜곡된 특징들을 삭제하는 과정이 필수적이다.[3]

자동문서분류에 있어서 본 연구가 주목하는 이슈는 문서내의 특정 단어가 여러 의미를 갖는 경우에 올바른 분류를 하지 못하는 중의성 문제이다. 중의성을 갖는 단어는 문맥의 앞 뒤 내용을 파악해야만 그 정확한 의미를 가려 낼 수 있다. 만약 한 단어가 어떤 단어와 일정 수준 이상 빈번하게 공기(cooccurrence)하면, 그 단어들은 서로 연관이 있는 단어라 할 수 있다.[4] 이러한 공기 관계를 활용하기 위해 본 논문에서는 자연어 처리 기법의 하나인 청킹(chunking) 기술을 이용하여 문서내의 단어가 가지는 중의성을 최소화하는 특징 확장 방법을 제안하고자 한다. 청킹은 주변의 단어와 공기 관계를 고려할 수 있다.

본 논문의 구성은 다음과 같다. 우선 2장에서 자연어 처리 기법의 하나인 청킹에 대하여 살펴보고, 3장에서는 청킹을 이용한 분류 모델을 개선하는 방안을 제안한다. 4장에서는 실험 결과를 제시하고 5장에서는 결론과 이후 연구에 대한 향방을 제시한다.

2. 배경연구

2.1 청크(Chunk)

청크(Chunk)란, 언어학적으로 본다면 말모듬이라는 뜻으로 언어 학습자가 한번에 하나의 단위처럼 배울 수 있는 어구를 뜻한다. 심리학에서는 기억 대상이 되는 자극이나 정보를 서로 의미 있게 연결 시키거나 묶는 것을 지칭한다. 또한 청킹(chunking)은 정보를 의미 있게 묶어 청크를 만드는 과정을 뜻한다. 본 연구에서는 의미 있는 단어를 결합하여 새로운 특징을 만드는 과정을 청킹이라 정의한다. 주어진 학습 문서 집합에 중의성을 나타내는 단어들에 대하여 청킹 과정을 수행함으로써 분류 모델의 구성에 도움이 되는 특징을 추출하고자 한다.

2.2 청킹(Chunking)

예를 들어, ‘data mining’ 이라는 용어를 통해 청킹을 설명하겠다. data 라는 단어는 정보처리나 컴퓨터 분야 외의 넓은 범위에서 사용되고 있는 단어이다. 그리고 mining 이란 단어는 사전적 의미로 ‘채광, 발굴’등의 의미를 갖는다.

data mining (일반 단어) datamining (청크 특징)

하지만 data와 mining이 함께 쓰인 ‘datamining’ 이라는 용어는 컴퓨터 과학의 한 분야를 가리키는 용어가 된다. 본 논문에서는 이와 같이 단어가 가지는 의미의 범위를 좁혀 중의성과 모호성을 최소화 하는 과정을 청킹이라 정의하고, 청킹을 통하여 얻어진 ‘datamining’ 과 같은 단어를 청크 특징이라고 하겠다.

3. 청킹 기반 특징 추출과 분류 모델의 개선

중의적인 의미로 인하여 광범위하게 사용되는 일반단어에 비하여 청크 특징은 상대적으로 그 사용범위가 좁다. 그만큼 의미를 강화한다는 의미로, 분류에 있어서 좋은 기준점이 될 수 있을 것이다. 이 장점을 이용하여 분류 정확도를 향상시키고자 청크 특징의 출현 빈도수에 변환을 주어 분류 모델을 개선하였다.

3.1 중의성 해소를 위한 청크 특징 추출

영어 문장의 경우에 그 문장의 의미(주제)를 가장 잘 표현하는 품사는 명사이다. 명사는 다른 품사들과 달리 해당 단어가 연속적으로 연결되어도 문법상 문제가 없으며, 명사만의 나열만으로도 의미를 나타낼 수 있다. 이러한 이유로 명사를 서로 연결 시켜 의미를 가질 수 있는 정보가 될 수 있다. 본 연구에서는 청킹 과정을 통해 연속적 명사군을 찾아냄으로써 중의성을 가지는 명사 특징을 제거하고자 한다.

본 연구에서는 명사군의 청킹 과정은 세 단계를 거친다. 첫 번째는 문장을 구(phrase) 단위로 나눈다. 이 중 명사를 포함하고 있는 명사구만 추려낸다. 두 번째 단계에서는 명사구 내에서 품사를 구분한다. 마지막으로 단어 사이에 명사 외의 다른 품사가 존재하지 않으며 2회 이상 명사가 연속되는 단어를 결합하여 청크 특징을 만든다.

사용할 수 있는 청크 특징은 두 종류가 있다. 하나의 카테고리에서만 나타나는 청크 특징과 여러 카테고리에서 나타나는 청크 특징이다. 전자의 경우 하나의 카테고리에서만 나타나기 때문에 카테고리를 대표하는 특징이 될 수 있다. 후자의 경우 1~2개 정도로 적은 수의 카테고리에서 확연히 높은 문헌 빈도수를 보이는 청크 특징이 카테고리의 특징을 보일 수 있을 것이다.

3.2 주요 청크 특징의 추출

사실 모든 청크 특징이 문서 분류에서 유용하게 사용되는 것은 아니다. 희귀한 단어, 즉 출현 빈도수가 매우 적은 단어의 경우 좋은 분류모델의 구성에 방해가 된다. 이러한 방해 요소를 배제하기 위하여 적정 수준 이상의 출현 빈도수를 보이는 특징을 골라내야 한다..

단어의 출현 빈도수를 측정하는 방법은 두 가지가 있다. 하나는 단어가 출현하는 문서의 수를 의미하는 문헌 빈도수(document frequency)이고, 다른 하나는

문서의 수와 관계없이 한 문서 내의 단어의 출현 횟수를 의미하는 단어 빈도수(term frequency)이다. 카테고리의 구분 없이 문헌 빈도수가 높은 단어는 모든 카테고리에서 자주 출현함을 의미함으로 해당 단어가 특징으로 작용하지 못한다. 그러나 하나의 카테고리 안에서만 특정 단어의 문헌 빈도수가 높다면, 해당 단어는 그 카테고리에서는 폭넓게 쓰이고 있지만 다른 카테고리에서는 사용되지 않는다 것을 의미한다. 그러므로 문헌 빈도수를 계산할 때 카테고리 별로 나누어서 계산한다.

학습할 문서 집합의 적어도 하나 이상의 카테고리에서 일정 비율 이상의 문헌 빈도수를 보임으로써 문서 분류 시 영향력 있는 중요한 청크 특징을 SCF(Significant Chunk Feature)라 명명하겠다. 본 연구에서는 SCF를 이용하여 학습 모델을 개선할 것이다.

3.3 출현 빈도수의 확대

청크 특징은 일반적으로 쓰이는 단어들 중에서도 몇 가지 조건을 만족하는 단어들을 모아 만들어지기 때문에 청크 특징이 아닌 일반 단어에 비하여 그 출현 빈도수가 적다. 이 부분을 보완하기 위해 출현 빈도수를 의도적으로 늘려주는 방식을 취할 수 있다.

출현 빈도수를 높이기 위해 학습할 문서 집합에 해당되는 SCF을 추가로 넣어 학습한다. 실험에서 사용되는 SCF는 아래와 같이 두 가지로 구분하여 정의한다.

single_class_feature : 하나의 카테고리에만 출현하는 청크 특징

multi_class_feature : 여러 카테고리에 출현하는 청크 특징

Only_feature는 하나의 카테고리에서만 등장하는 SCF를 이용한다. Multi_feature는 여러 카테고리에서 등장하지만 1-2개의 카테고리에서 확연히 높은 문헌 빈도수를 가지는 SCF를 이용한다. 위에서 설명한 두 가지 방법은 편의상 아래와 같이 구분된 실험을 적용한다.

SCF_one : 해당 SCF를 1회씩만 추가
SCF_ratio : 문헌 빈도수 비율만큼 추가

구체적으로 SCF_one 방법은 해당 SCF를 1회씩만 추가하는 것으로 추가되는 특징의 수가 가장 적은 보수적인 접근법 이다. SCF_ratio 방법은 문헌 빈도수 비율에 따라 추가하는 것이다. 해당 SCF의 문헌 빈도수를 주어진 문서 집합에서의 문헌 빈도수 최소값으로 나누어 그 횟수만큼 추가한다.

3.4 주요 청크 특징 기준의 문서 검색을 통한 분류 모델 개선

청크 특징은 그 수가 일반적으로 사용되는 단어들 보다 출현빈도수가 적다. 이것은 분류대상이 되는 문서들에서도 마찬가지이다. 이 문제를 극복하기 위해 테스트 집합의 청크 특징의 출현 빈도수를 높여주고자 한다. 테스트 집합에서 SCF를 가진 문서를 검색한다. 검색된 문서에 해당 문서가 가지고 있는 SCF를 1회 추가하여 준다. 학습할 문서집합이 아닌 분류할 문서에 직접 추가하여 준다.

3.5 점진적 학습 기법을 이용한 분류 모델 개선

이번 절에서는 분류 모델의 개선을 위해 추출된 SCF을 활용한 점진적 학습 (incremental learning)기법을 제시한다.

알고리즘 : 점진적 학습 기법을 이용한 분류 모델 개선	
입력 : 분류모델 m P_i, i번째 선분류 집합 eP_i, i번째 선분류 집합 오분류 정보 T, 테스트 집합 출력 : 개선된 모델 m'	
	BEGIN
1	while(| P_i| < |T|*0.5) {
2	while(accuracy(m'(P_i) − accuracy(m(P_{i-1})> 0) {
3	// m(P_i) : m으로 P_i를 분류함을 의미
4	eP_i = m(P_i)
5	// eP_i : P_i 의 오분류 정보
6	m' = learning(m+eP_i)
7	// m+eP_i을 학습하여 m'을 생성
8	}
9	P_{i+1} = P_i ∪ S
10	// S = T의 부분집합
11	// |S| ≤ |T|*0.1
12	}
13	개선된 분류모델 m' 획득
	END

<그림 1> 점진적 학습 알고리즘

우선은 테스트 집합의 일부로 선분류 집합을 구성한다. 점진적으로 모델을 개선해 나갈 것이기 때문에 최초의 선분류 집합의 크기는 작을수록 좋다. <그림 1>의 4 처럼, 선분류 집합 P를 분류모델 m으로 분류하여 오분류 된 문서에 대한 정보를 얻는다. 그 다음으로 <그림1> 6을 진행한다. 이 과정은 기존의 학습 모델에 오분류 된 문서가 가진 SCF를 추가하고, 다시 학습하여 개선된 분류모델 m'을 만드는 과정을 보여준다. <그림 1>의 2에서처럼 선분류 집합 P를 분류했을 때 그 결과에 대한 정확도의 변화가 0 또는 0에 가깝다면, <그림 1>의 9를 수행하여 선분류 집합 P의 크기를 늘리고, 앞의 과정인 <그림 1> 2-8을 반복한다. 증가된 선분류 집합 P의 크기가 테스트 집합 T의 50% 보다 커지면 모든 반복과정을 종료한다. 가장 마지막으로 생성된 분류모델 m'은 점진적 학습을 통해 개선된 분류 모델이다.

4. 실험

4.1 실험 환경

제안한 기법의 성능을 평가하기 위하여 20Newsgroup 을 이용한 실험을 실시하였다. 20Newsgroup는 문서분류의 성능을 평가하기 위해 일반적으로 많이 사용되는 문서집합이다. 20개의 카테고리를 갖는 19,997개의 기사로 구성되어 있으며, 카테고리 간에 문서들이 비교적 균형 있게 분포되어 있다. 단어의 품사 구분을 위하여 Illinois pos Tagger와 Illinois Chunker[5]를 사용하였다. 전체 문서의 70%를 학습을 위해 사용하였고, 나머지 30%는 분류성능을 평가하는데 사용하였다. 분류는 나이브 베이지안(Naïve Bayseian) 문서분류기의 하나인 Mallet 시스템[6]을 이용하였다. 분류의 성능은 각 문서가 속할 카테고리를 얼마나 정확하게 분류하는 가를 기준으로 평가하였다.

4.2 실험 결과

<표 1>, <표 2>, <표 3>에서 기준 기법이라 함은 실험에 사용되는 문서에 청킹을 적용하여 청크 특징을 가지고는 있는 상태로, 제안한 실험들은 적용되지 않은 것을 뜻한다. 기준 기법을 기준 정확도로 설정하고 제안한 실험들의 성능 비교에 사용하였

다.

표1은 SCF 중 single_class_feature를 이용하여 출현 빈도수를 확대해준 실험 결과이다. 보는 바와 같이 SCF_one의 경우 약간의 성능 향상을 보였으나, SCF_ratio의 결과 처럼 특징을 추가하는 횟수가 늘어날수록 성능은 떨어졌다. 표2는 테스트 집합에서 SCF를 가진 문서를 검색하여 SCF를 추가하여 주는 방법에 대한 실험 결과이다. SCF를 1회만 추가하였을 경우 약 1% 의 분류 정확도 상승을 보였으나 SCF의 추가 횟수가 늘어날 수록 성능이 떨어졌다. 표3은 선분류를 통하여 점진적 학습을 진행하여 개선된 분류 모델을 얻어내는 실험의 결과이다. SCF와 일반 단어를 오분류 정보로 이용하였을 경우가 SCF만을 이용한 경우보다 미미하지만 좋은 성능을 보였다.

대부분의 실험에서 괄목할만한 분류 성능의 향상은 보이지 못하였다. 하지만 SCF를 기준으로 테스트 문서집합에서 문서를 검색을 통한 분류 모델의 개선 방법은 자동분류 시스템의 성능 향상 가능성을 보여주었다.

<표 1> SCF의 출현 빈도수 확대 를 통한 분류 모델 개선 (단, single_class_feature)

실험	분류 정확도
기준 기법	78.08
SCF_one	78.13
SCF_ratio	75.51

<표 2> SCF 기준의 문서 검색을 통한 분류 모델 개선

실험	분류 정확도
기준 기법	78.08
SCF 1회 추가	79.24

<표 3> 선분류를 통한 점진적 모델 개선

실험	분류 정확도
기준 기법	78.08
SCF만 추가	78.45
SCF와 일반단어 추가	78.47
일반 단어만 추가	77.98

5. 결론

기존의 나이브 베이지안 분류 방법은 통계적 수치만을 고려하기 때문에 단어가 가지는 중의성으로 인하여 분류 성능에 영향을 받았다. 본 연구에서는 청킹을 통하여 단어가 가지는 중의성을 최소화 하는 청크 특징을 만들고 이를 이용하여 분류 모델을 개선하고자 하였다. 청크 특징 중 분류 시스템에 영향을 줄 수 있을 것이라 예상되는 특징들을 추출하고 분류 모델 개선에 이용하였다. 본 실험에서의 성능 향상은 미미하였으나 20Newsgroup 문서집합에 대한 실험을 통하여, 제안 기법의 가능성을 볼 수 있었다. 향후, 특징 추출과 활용 방법을 확장하는 연구를 진행하여 자동문서분류 시스템의 성능을 향상시킬 수 있도록 연구를 진행할 것이다.

참고문헌

[1] T.M. Mitchell, “Bayesian Learning,” Machine Learning, McGraw-Hill, pp.154-200, 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", 정보처리학회논문지 B 제15-B권 제5호, 2008.

[4] Wilks, Y., Fass, D. Guo, C., McDonald, J. Plate T., and Slator, B. "Providing machine tractable dictionary tools.", Machine Translation, 5, pp99-154, 1990.

[5] Illinois chunker, University of Illinois at urbana champaign Congnitve Computation Group, http://cogcomp.cs.illinois.edu/

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