Journal of the Korean Institute of Industrial Engineers
Vol. 41, No. 4, pp. 381-388, August 2015.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2015.41.4.381

<Original Research Paper>

대칭 조건부 확률과 TF-IDF 기반 텍스트 분류를 위한 N-gram 특질 선택

최우식 · 김성범†

고려대학교 산업경영공학과

N-gram Feature Selection for Text Classification Based on Symmetrical Conditional Probability and TF-IDF

Woo-Sik Choi · Seoung Bum Kim

Department of Industrial Management Engineering, Korea University

The rapid growth of the World Wide Web and online information services has generated and made accessible a huge number of text documents. To analyze texts, selecting important keywords is an essential step. In this paper, we propose a feature selection method that combines a term frequency-inverse document frequency technique and symmetrical conditional probability. The proposed method can identify features with N-gram, the sequential multiword. The effectiveness of the proposed method is demonstrated through a real text data from the machine learning repository, University of California, Irvine.

Keywords: Text classification, Feature selection, N-gram, Term frequency-Inverse document frequency, Symmetrical conditional probability

1. 서 론

텍스트마이닝 분야에서 텍스트 데이터를 분류하는 문제는 오래 전부터 논의가 되었으며, 대표적인 예로는 기사나 웹 페이지 등의 카테고리 자동 분류, 스팸 메일 분류, 사용자 성별 분류 등이 존재한다(Nigam *et al*., 2000; Mukherjee and Liu, 2010). 또한 근래에 들어 데이터가 빠른 속도로 급증하고 있기 때문에(Cho and Kim, 2012) 이러한 분류 문제를 효율적으로 해결할 수 있어야 하는 상황이다(Burger *et al*., 2011). 이를 위해서는 텍스트 데이터를 상황에 맞게 구조화(structuring) 즉, 숫자로 표현 되는 행렬로 변환해야 하며, 여기에 사용되는 특질(feature)들을 적절하게 선택해야 한다.

텍스트 데이터를 구조화하는 방법 중 대표적으로 쓰이는 것은 텍스트 전체, 혹은 텍스트 내에 있는 문장에 포함된 단어의 빈도를 이용하는 것이다(Feldman and Sanger, 2007). <Figure 1>을 확인해보자. (a)는 예시 문장이고, (b)는 문장 내에 있는 단어 빈도의 일부를 나타내고 있다. (a)의 문장에서 'text', 'mining', 'chapter'는 3번씩, 'introduction'은 2번, 'book' 1번 사용되었는데, 이러한 방식으로 각 단어들의 빈도를 계산하여 (a)가 (b)와 같은 형태로 요약이 되는 것이다. 이를 일부 문장이 아닌 하나의 텍스트 전체로 적용하게 되면 그 텍스트에서 자주 쓰인 단어를 확인해볼 수 있기 때문에, 이는 해당 텍스트가 어떤 내용을 담고 있는지 쉽게 확인해볼 수 있는 수단이다.

이와 같은 방식으로 각각의 단어를 이용하여 텍스트 분류를 하고자 할 경우에, 특질 선택을 위하여 사용하는 방법이 여러 가지 존재한다. 이 중 TF-IDF(term frequency-inverse document frequency) 방법은 잠재 디리클레 할당(latent dirichlet allocation) 방법과 함께 가장 널리 쓰이고(Chemudugunta and

제10회 석사논문경진대회 수상논문.
† 연락저자 : 김성범 교수, 136-701 서울시 성북구 안암로 145 고려대학교 산업경영공학과, Tel : 02-3290-3769, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2015년 2월 9일 접수; 2015년 4월 20일 수정본 접수; 2015년 5월 11일 게재 확정.

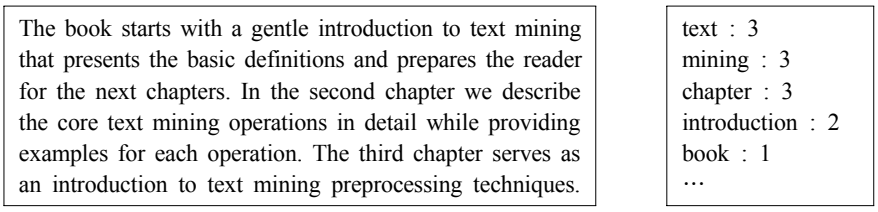

Figure 1. Example of structuring text data based on term frequency

Steyvers, 2007), 동시에 이해하기 쉬우며 효율적이다(Ramos, 2003).

여기서 말하는 특질은 꼭 단어 한 개에 국한 될 필요는 없고, 여러 개의 단어가 연합되어 있는 즉, 어절 형태로도 표현할 수 있다. 이를 N-gram이라고 하며 N은 연합되는 단어의 수를 나타낸다(Brown *et al*., 1992). 경우에 따라서는 명사를 수식하는 단어를 제외하여 묶는 방법도 있으나(Sidorov *et al*., 2014), 보통은 연속된 단어를 묶어서 사용한다. <Figure 1>의 (a)를 N을 3으로 하는 N-gram(3-gram)을 추출하면 'the book starts', 'book starts with', …, 'mining preprocessing techniques'으로 형성된다. N-gram을 사용하는 이유는 텍스트 내에 특정 인물의 이름이나 'World Cup'과 같이 자주 붙어서 등장하는 단어를 찾아냄으로써 단어의 의미를 보다 잘 표현할 수 있기 때문이다(da Silva and Lopes, 1999).

N-gram을 활용한 텍스트 분류 연구는 다수 진행되었는데, 그 중 하나는 개별 단어를 이용하여 텍스트들을 분류할 때 사용하는 TF-IDF를 N-gram에 단순 적용하여 사용하는 것이고(Zaki *et al*., 2014), 또 하나는 대칭 조건부 확률(symmetrical conditional probability)을 이용한 지역 최대화 방법(local maxima method)이다(da Silva and Lopes, 1999). 그러나 전자의 경우 N-gram을 빈도에 따라서만 선택하였기 때문에 N-gram에 속하는 단어들의 조합(collocation)을 반영하지 못한다는 단점이 존재하며(Smadja *et al*., 1996), 후자의 경우 구조화할 때에 사용하는 특질의 수가 매우 많아 차원의 저주가 발생하여 텍스트 분류에 많은 시간을 투자하는 단점이 있다(Li and Jain, 1998).

이러한 단점을 해결하기 위하여 지역 최대화 방법과 TF-IDF를 혼합한 연구가 등장하였다(Silva and Lopes, 2010). 해당 연구에서는 대칭 조건부 확률을 이용한 지역 최대화 방법으로 텍스트를 구조화한 이후, TF-IDF를 적용하여 N-gram에 속하는 단어의 조합과 N-gram의 중요도를 동시에 고려하였다. 그러나 이 방법은 TF-IDF를 특질 선택이 아닌 N-gram의 중요도를 평가하는 요소로만 사용하여, 기존의 지역 최대화 방법과 마찬가지로 차원의 저주가 발생하는 단점이 있다. 또한 대칭 조건부 확률과 TF-IDF를 결합하여 N-gram 특질을 선택한 연구는 존재하지 않았다.

본 논문에서는 기존의 TF-IDF와 지역 최대화 방법을 사용한 논문들의 단점을 보완하고, 장점을 융합하기 위한 새로운 방법으로 SCP TF-IDF(symmetrical conditional probability with term frequency-inverse document frequency)를 제안하고자 한다. 이에 따라 텍스트 데이터를 분류하는 데에 필요한 N-gram 특질을 효율적이고 알맞게 선택하여, 이를 바탕으로 기존의 방법보다 더 정확하게 텍스트들을 분류하는 것에 목표를 두고 있다.

제 2장에서는 기존의 방법으로 중요 단어를 선택할 때 자주 쓰이는 TF-IDF와 N-gram에서 특질 선택을 할 때에 응용되는 대칭 조건부 확률에 대해서 소개하고, 제 3장에서는 본 논문에서 제안하는 방법론인 SCP TF-IDF에 대해 설명한다. 제 4장에서는 이 두 방법론을 비교하기 위하여 실험에 사용할 데이터와 분석 알고리즘을 설정하고, 이에 대한 결과를 제 5장에서 서술한다. 마지막으로 제 6장에서 본 논문의 결론을 서술한다.

2. 텍스트 분류를 위한 특질 선택 방법

2.1 TF-IDF 방법

앞서 설명한 바와 같이, TF-IDF는 텍스트 마이닝에서 중요한 단어를 추출할 때 쓰이는 가장 대표적인 방법 중 하나이다. TF-IDF는 한 텍스트 내 특정 단어의 빈도를, 그 단어의 전체 텍스트 출현 빈도로 나누어준다는 의미이다(Salton and McGill, 1983). 각각의 단어에 대하여, TF-IDF 값은 식 (1)을 통해 계산된다.

$$(TF-IDF)_{w,d} = f_{w,d} \cdot \log\left(\frac{|D|}{f_{w,D}}\right) \tag{1}$$

여기서 w는 특정 단어, d는 특정 텍스트, D는 전체 텍스트 데이터를 의미하며, $f_{w,d}$는 d에 나타나는 w의 빈도, $f_{w,D}$는 w가 등장한 텍스트의 수, $|D|$는 전체 텍스트 데이터의 수를 의미한다. 따라서 특정 단어가 다수의 텍스트에 등장을 하면 $(TF-IDF)_{w,d}$값이 감소하고, 소수의 텍스트에 등장하면 증가한다. 예를 들어 관사인 'the', 'a', 'an'의 경우 모든 텍스트에 등장할 가능성이 매우 높으므로 $(TF-IDF)_{w,d}$값은 0에 근사한 값을 가진다.

일반적인 TF-IDF 방법은 식 (1)을 적용하여 형성된 행렬을 분류 알고리즘에 바로 도입한다. 그러나 특질 선택을 위하여 단어 별 중요도를 판단하고 텍스트 분류에 적절한 단어를 선별하는 과정이 필요하다. 이를 위하여 전체 텍스트에 대한 단

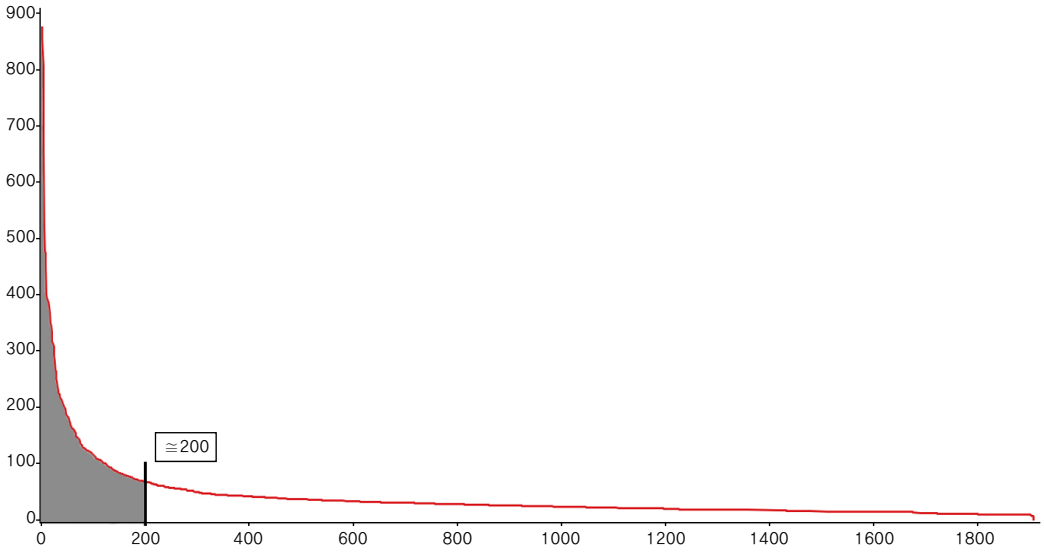

Figure 2. Selecting features at the elbow point of TF-IDF scores

어 별 점수를 계산하며, 수식은 다음과 같다.

$$(TF-IDF)_{w,D} = \sum_{d \in D} (TF-IDF)_{w,d} \quad (2)$$

$(TF-IDF)_{w,d}$는 전체 텍스트에 대한 단어 별 TF-IDF 점수를 의미하며, 특정 단어 w의 $(TF-IDF)_{w,d}$ 값을 모든 텍스트에 대하여 합하여 계산한다. 다른 수식 없이 단순히 합하여 사용하므로 평균으로 계산하여도 결과는 동일하다(Tang *et al.*, 2005). 이 값들을 모든 단어에 대해 내림차순으로 정렬하고, 팔꿈치 지점(elbow point)까지의 단어 특질만 선택하여 사용한다(Jing *et al.*, 2002). <Figure 2>는 TF-IDF 점수를 이용하여 특질을 선택하는 것을 보여주고 있으며, x축은 $(TF-IDF)_{W,D}$ 점수의 내림차순으로 정렬된 특질, y축은 $(TF-IDF)_{W,D}$ 점수이다. <Figure 2>의 경우 팔꿈치 지점이 약 200에 있으므로 분석에 사용할 특질의 수를 200개로 설정한다. 때때로 팔꿈치 지점이 명확하게 나타나지 않을 수 있는데 이럴 경우에는 사용자가 관련 지식을 이용하거나 전문가의 도움을 받아 임의로 지점을 선정할 수 있다.

2.2 대칭 조건부 확률

대칭 조건부 확률은 지역 최대화 방법 사용 시 N-gram을 선택할 때 기준이 되는 값 중 하나이다(da Silva and Lopes, 1999). 여기서 지역 최대화 방법은, 특정 N-gram의 주변 N-gram과 비교하였을 때, 특정 N-gram의 기준 값이 가장 클 경우에만 선택하는 방법을 지칭한다. 대칭 조건부 확률은 식(3)과 같다.

$$SCP(x,y) = p(y \mid x) \cdot p(x \mid y) \quad (3)$$
$$= \frac{p(x,y)}{p(x)} \cdot \frac{p(x,y)}{p(y)} = \frac{p(x,y)^{2)}}{p(x) \cdot p(y)}$$

즉, $SCP(x,y)$는 x, y의 대칭 조건부 확률 값으로, x가 발생했을 때 y가 발생할 조건부 확률과 y가 발생했을 때 x가 발생할 조건부 확률을 곱해준 것이다. 따라서 x, y가 같이 발생할 확률이 각각 독립적으로 발생할 확률에 비하여 높다면 대칭 조건부 확률의 값이 높아진다. 이를 2-gram에 도입하게 되면 식(3)은 다음과 같이 표현된다.

$$SCP(w_1, w_2) = \frac{p(w_1, w_2)^2}{p(w_1) \cdot p(w_2)} \quad (4)$$

여기서 w_1, w_2는 연속되어 있는 각각의 단어를 의미하며, $SCP(w_1, w_2)$는 w_1, w_2가 순차적으로 포함되어 있는 2-gram의 대칭 조건부 확률 값으로 해석할 수 있다. 그러나 N이 3 이상일 경우 즉, 3-gram, 4-gram 등에서는 단어를 나눌 수 있는 지점 즉, 분할 지점이 여럿 존재하기 때문에, 이를 전부 고려할 수 있도록 일반화된 수식으로 변형해야 하며, 이를 식(5)에서 보여주고 있다.

$$SCP(w_1 \cdots w_N) = \frac{p(w_1 \cdots w_N)^2}{\frac{1}{(N-1)} \sum_{i=1}^{N-1} p(w_1 \cdots w_i) \cdot p(w_{i+1} \cdots w_N)} \quad (5)$$

식(5)의 분모는 해당 N-gram이 가지고 있는 모든 분할 지점에 대해서 도출되는 전체 확률 값 조합의 평균이다(Houvardas and Stamatatos, 2006). 예를 들어 $N=3$일 경우 $\frac{1}{2}\{p(w_1) \cdot p(w_2, w_3) + p(w_1, w_2) \cdot p(w_3)\}$이 되므로, 첫 번째 분할 지점으로 나누어 $p(w_1)$과 $p(w_2, w_3)$을 곱한 값과 두 번째 분할 지점으로 나누어 $p(w_1, w_2)$와 $p(w_3)$을 곱한 값을 더하고 평균을 계산한 것이다. 이를 활용하여 특정 N-gram에 포함 되는 단어 혹은 N-gram이 텍스트 내에 각각 존재할 확률과, 특정

N-gram이 텍스트 내에 존재할 확률을 비교하게 된다. 따라서 해당 N-gram 내에 있는 단어 혹은 N-gram이 같이 존재할 확률이 높을수록 $SCP(w_1 \cdots w_N)$ 값이 증가한다. $SCP(w_1 \cdots w_N)$ 값이 크다는 것은 특정 N-gram을 분할하였을 때에 비해 통합하였을 때의 발생 확률이 비교적 높다는 것으로, 특정 N-gram의 발생 빈도에 대해서는 고려하지 않는다.

참고로, N이 2인 경우에 식(5)는 식 (4)와 동일하며, N이 1인 경우는 고려하지 않는다. 또한 N-gram의 특성상 특정 N-gram이 분리되어 등장하는 경우가 없을 때는 $SCP(w_1 \cdots w_N)$ 값이 1보다 큰 경우가 존재하며, 이 때의 값은 1로 설정한다.

3. 제안 방법

이 장에서는 본 논문에서 제안하는 N-gram 추출 방법에 대하여 설명한다. 제안 방법인 SCP TF-IDF는 제 2.1절에서 설명한 TF-IDF와 제 2.2절에서 설명한 대칭 조건부 확률을 동시에 적용한 것이다. 이에 따라 제안 방법은 N-gram에 속하는 단어의 조합을 고려함과 동시에 중요한 특질만을 선택할 수 있는 특성을 갖도록 하였다. 수식은 식 (1)과 식 (5)를 곱하는 형태가 되는데, 다음과 같이 표현된다.

$$(SCP\ TF-IDF)_{w,d} = \frac{p(w_1 \cdots w_n)_d^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \cdots w_i)d \cdot p(w_{i+1} \cdots w_n)_d} \cdot f_{w,d} \cdot \log\left(\frac{|D|}{f_{w,D}}\right) \quad (6)$$

w는 단어 $w_1 \cdots w_n$을 연속적으로 포함하는 N-gram을 의미하며, 이 외의 기본적인 표기법은 제2장과 동일하다. $(SCP\ TF-IDF)_{w,d}$는 특정 텍스트 d에서 N-gram w의 SCP TF-IDF 값을 의미한다. 텍스트마다 각각의 N-gram이 등장하게 되는 배경이 다를 수 있기 때문에 텍스트 별로 해당 값을 다르게 산정해야 한다.

제안 방법은 TF-IDF와 마찬가지로 각각의 N-gram에 대해 전체 텍스트에 대한 점수를 계산하며, 수식은 다음과 같다.

$$(SCP\ TF-IDF)_{w,D} = \sum_{d \in D}(SCP\ TF-IDF)_{w,d} \quad (7)$$

$(SCP\ TF-IDF)_{w,D}$는 전체 텍스트에 대한 N-gram별 SCP TF-IDF 점수를 의미하며, 이 또한 TF-IDF와 동일하게 $(SCP\ TF-IDF)_{w,d}$ 값을 모든 텍스트에 대하여 합한다. 이 값을 토대로 <Figure 2>에서와 같이 팔꿈치 지점까지의 N-gram을 특질로 사용하여 텍스트 데이터를 분류한다.

제안한 방법을 통해 얻을 수 있는 이점은 두 가지가 있다. 하나는 대칭 조건부 확률과 TF-IDF를 동시에 적용하면서도 적절한 양의 특질만을 사용하여 차원의 저주를 저지할 수 있다는 것이고, 또 하나는 N-gram에 속하는 단어의 조합을 고려하여 특질 선택에 반영한다는 것이다. 이를 통해 TF-IDF처럼 사용할 특질의 수를 크게 감소시킴과 동시에 각각의 텍스트가 가지고 있는 특성을 보다 잘 반영할 수 있는 N-gram을 선택하도록 하였다.

4. 실 험

제 2장, 제 3장에서 각각 서술한 방법이 실제로 어느 정도의 차이가 있는지 확인하기 위하여 현실 데이터를 이용하여 실험 결과를 비교하였다. 사용한 언어는 파이썬이며(Python Software Foundation, 2010), 분류 알고리즘은 파이썬 패키지인 Scikit-learn을 이용하여 구현하였다(Pedregosa *et al.*, 2011). Scikit-learn 패키지는 나이브 베이지안 분류기, 의사 결정 나무 등 다양한 종류의 분류 알고리즘을 포함하며, 사용자가 보다 쉽게 알고리즘을 구현할 수 있도록 편의를 제공한다. 실험 과정은 <Figure 3>과 같다. 화살표는 진행 방향이며, 점선의 경우 반복 가능한 위치이다. 먼저 사용할 현실 텍스트 데이터를 선택한 후 특수기호 제거, 형태소 분석 등의 전처리 과정을 수행하고, 이를 바탕으로 대칭 조건부 확률 계산 및 학습 데이터 구조화를 한다. 구조화 데이터 및 대칭 조건부 확률 값을 사용하여, 앞서 설명한

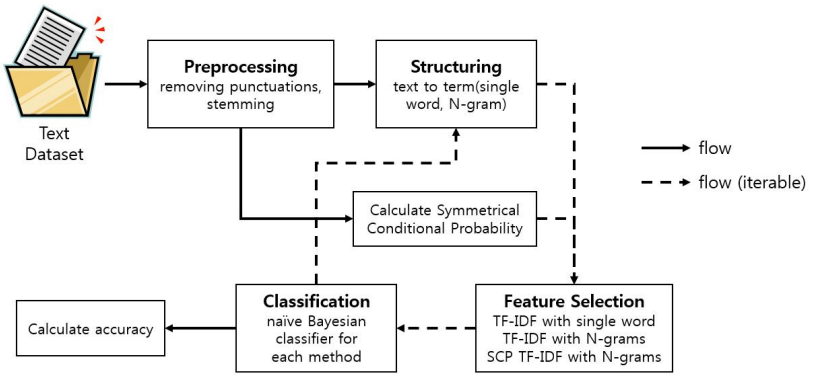

Figure 3. Overview of experiment for comparison

Table 1. Example of the Opinosis Opinion/Review Dataset

Domain	Product	Function	Review Example
Hotel	Holiday Inn	Service	It is not 4 star hotel service
Car	Honda	Performance	No problem with quality, performance, etc.
Car	Honda	Comfort	Overall performance is good but comfort level is poor
Electronics	Kindle	Battery-Life	Battery life is very good, even with the wireless on constantly
Electronics	IOPod	Battery-Life	The battery dies faster than my previous 2nd Gen Nano
Electronics	1005ha	Battery-Life	This netbook has plenty of power for my needs and the battery life is great

특질 선택 방법을 기반으로 특질 선택을 하고, 각각의 방식에 따라 분류 모델을 구축한다. 각 분류 모델에 실험 데이터를 적용하여 정확도를 계산하는데, 필요에 따라 학습 데이터와 실험 데이터를 다시 나누어 구조화부터 반복하여 진행한다. 반복을 완료하면 반복 과정에 의해 도출되는 정확도의 평균을 계산하며 특질 선택 방법에 따라 이를 비교한다.

4.1 데이터

실험을 위한 실제 데이터는 UCI(University of California at Irvine) Machine Learning Repository에 있는 (Bache and Lichman, 2013) Opinosis 데이터셋(Opinosis Opinion /Review Dataset)을 이용하였다(Ganesan *et al.*, 2010). Opinosis 데이터셋은 51개의 리뷰 텍스트로 구성되어있으며, 이는 각각 차량, 전자제품, 호텔 총 3가지의 도메인으로 구분된다. 또한 10개의 제품명, 34개의 기능으로 분류가 가능하다. 각각의 텍스트들은 평균 약 100개의 문장으로 구성되어 있으며, 그 일부는 <Table 1>과 같다.

제품명과 기능에 따라서도 차이가 있으나, 어떤 도메인에 속해 있는지에 따라서도 텍스트마다 등장하는 단어, 문장 등에 큰 차이가 있음을 확인할 수 있었다. 따라서 도메인을 각 텍스트의 클래스로 지정하여 텍스트 분류를 시도하였다.

4.2 전처리

제 4.1절에서 설명한 Opinosis 데이터셋을 구조화하기 전에 전처리 과정인 특수기호 제거, 어간 추출을 적용하였다. N-gram의 경우 전처리 없이 구조화해도 큰 문제는 없으나 (Houvardas and Stamatatos, 2006), 개별 단어를 사용했을 때와 원활하게 비교하기 위하여 전처리를 한 이후에 구조화하였다. 관사, 전치사, 접속사 등 불용어(stopword)의 경우 개별 단어 특질 선택 시 제거가 되므로 전처리 과정에서 고려하지 않았다.

4.3 데이터 구조화

텍스트 분류에 대한 예측 정확도를 확인하기 위하여 데이터를 학습 데이터와 실험 데이터로 임의대로 나누었고, 각각의 비율은 9 : 1, 8 : 2 두 가지 상황을 설정하였다. 비율을 나누는 기준은 딱히 없으며 보통 9 : 1과 8 : 2를 많이 사용한다. 사용하는 특질의 경우 학습 데이터에서만 선정하였다. 또한 개별 단어를 사용할 경우와 N-gram을 사용할 경우 각각 독립적으로 구조화를 진행하였다. 이 때 사용한 N은 2, 3, 4이고, N이 5 이상일 경우는 희소행렬(sparse matrix) 형태를 보였기 때문에 텍스트를 분류하기에는 부적절하므로 제외하였다(Phan *et al.*, 2008).

4.4 특질 선택

개별 단어를 이용한 구조화 데이터는 제 2.1절에서 서술한 TF-IDF를 적용하였고, N-gram을 이용한 구조화 데이터는 TF-IDF, 그리고 본 연구에서 제안한 SCP TF-IDF 두 가지를 적용하여, 총 3가지 방식으로 특질 선택을 하였다. 3가지 방식으로 도출된 특질에 대한 점수를 각각 내림차순으로 정렬한 결과를 <Figure 4>에서 보여주고 있다. <Figure 4>는 각 방식에 대해 학습 데이터를 3회 선출하고, 내림차순 정렬 결과를 도식화한 것으로, <Figure 2>와 마찬가지로 x축은 점수의 내림차순으로 정렬된 특질이며, y축은 각 특질의 점수이다. 모든 경우 팔꿈치 지점이 약 50에 해당하는 위치에 존재하므로 50개의 특질을 사용하였다.

4.5 분류

텍스트 데이터를 도메인에 따라 나누기 위해 분류 알고리즘을 적용하여 모델을 구축하였다. 본 논문에서는 나이브 베이지안 분류기(naïve Bayesian classifier)를 사용하였는데, 이는 일반적으로 텍스트 분류 시 가장 보편적으로 쓰이는 알고리즘으로 알려져 있기 때문이다(Ting *et al.*, 2011).

나이브 베이지안 분류기는 강한 독립을 가정한 베이즈 정리를 기반으로 하고 있는 분류기이다. 관측치가 들어왔을 때, 기존에 가지고 있는 사전 확률을 이용하여 다음 수식을 활용한다.

$$p(C \mid F_1, \cdots, F_n) = \frac{p(C) \cdot p(F_1, \cdots, F_n \mid C)}{p(F_1, \cdots, F_n)} \quad (8)$$

여기서 $p(C|F_1, \cdots F_n)$은 $F_1 \cdots F_n$의 N-gram의 값이 주어질 때 특정 클래스 C에 속할 확률이라 볼 수 있다.

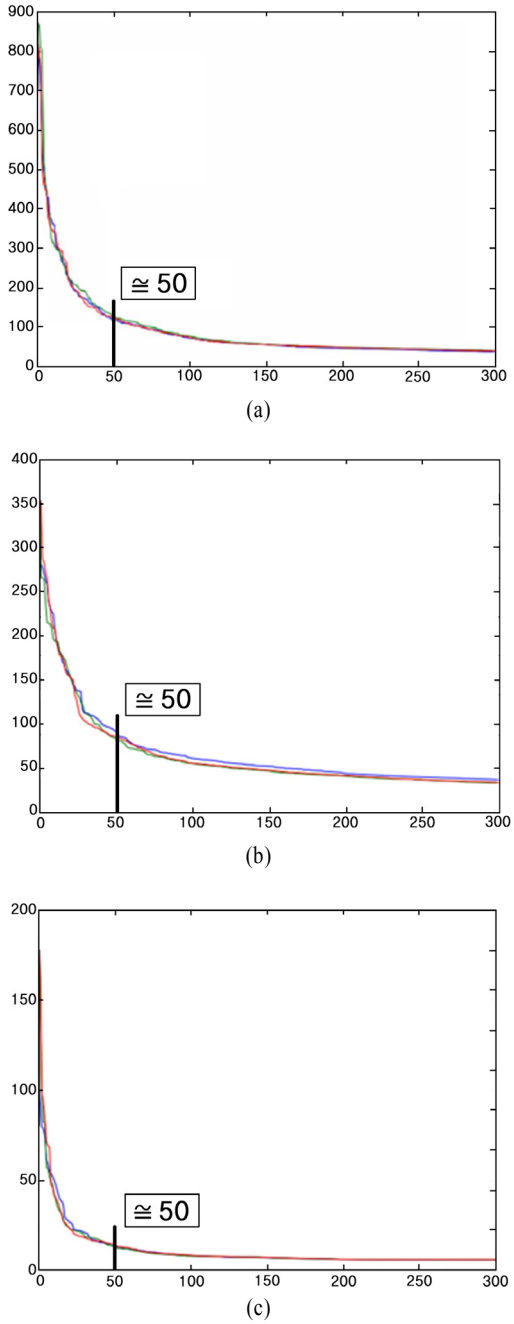

Figure 4. Selecting features for (a) TF-IDF with single words, (b) TF-IDF with 2~4-grams, (c) SCP TF-IDF with 2~4-grams

4.6 정확도

데이터 구조화 및 특질 선택을 한 이후나이브 베이지안 분류기를 이용하여 분류 모델을 구축하면 각각의 방식에 따른 예측 정확도를 계산할 수 있으며,그 식은 다음과 같다.

$$Accuracy = \frac{\#\ of\ testing\ data\ classified\ correctly}{\#\ of\ testing\ data} \quad (9)$$

정확도 계산 후에는 필요에 따라 구조화 과정으로 돌아가서 실험을 반복 진행한다.본 논문에서는 50회 반복하여 실험하였다. 반복 종료 후 각각의 특질 선택 방식 및 학습 데이터와 실험 데이터의 비율에 따라 평균 정확도를 계산하고,이를 비교한다.

5. 결 과

제 4장에서 서술한 실험 설계를 이용하여 각각의 특질 선택 방식들에 따른 텍스트 분류 결과를 확인하였다.이를 위하여 각 방법 별 예측 정확도를 비교하였고,N-gram 특질 선택 방법에 따라 선택된N-gram 특질의 차이를 확인하였다

5.1 예측 정확도

<Table 2>와 <Table 3>은 학습 데이터와 실험 데이터를 각각 9 : 1과 8 : 2로 나누었을 때의 예측 정확도 비교 결과이다.두 경우 모두 제안 방법인SCP TF-IDF를 이용하여 특질 선택을 했을 때의 예측 정확도가 가장 높음을 확인할 수 있었다

Table 2. Testing accuracy with 9 : 1 training and testing data ratio

Feature Selection Method	Testing Accuracy
TF-IDF w/single word	93.50%
TF-IDF w/2~4-gram	96.00%
SCP TF-IDF w/2~4-gram (Proposed Method)	**97.67%**

Table 3. Testing accuracy with 8 : 2 training and testing data ratio

Feature Selection Method	Testing Accuracy
TF-IDF w/single word	95.09%
TF-IDF w/2~4-gram	96.55%
SCP TF-IDF w/2~4-gram (Proposed Method)	**97.45%**

<Table 2>와 <Table 3>의 결과에 따르면, 본 논문에서 제안하는 SCP TF-IDF 방법에 따른N-gram 특질 선택 방식은TF-IDF 방법에 따른 특질 선택 방식에 비해 보다 좋은 예측 정확도를 보여준다고 할 수 있다. 이는 제안 방법이 텍스트 분류 관점에서 기존 방법보다 올바른 특질을 선택하였음을 보여준다

5.2 선택된 N-gram 특질

<Table 2>는 N-gram 특질 선택 방법에 따라 선택된 상위10개의 N-gram을 나타낸다. ‘ga mileag’와 같이 어색한 형태가 존재하는데, 이는 어간 추출을 사용하여 전처리를 하였기 때문이다. 50회 반복 실험에서50개의 특질 중 동일한 특질이 선택

된 비율은 평균51.2%로 대략 절반은 서로 다른 특질을 선택하였고, 또한 TF-IDF는 3~4-gram 선택 비율이 평균0.1%, SCP TF-IDF는 평균 4.8%로 N이 2보다 클 경우 선택된N-gram의 차이가 큼을 확인하였다

<Table 4>의 결과에 따르면, TF-IDF를 이용한N-gram 특질 선택 방법은 단어의 조합을 고려하지 않기 때문에,관사, 접속사 등의 불용어가 포함된N-gram을 다수 선택하였다. 이는 불용어를 삭제하지 않을 경우TF-IDF는 특정 인물의 이름이나 'World Cup'과 같은 의미 있는N-gram을 선택할 가능성이 감소함을 의미한다.반면 SCP TF-IDF를 이용한N-gram 특질 선택 방법은 단어의 조합을 잘 고려하여'tuscan inn', 'speed limit'과 같이 각 텍스트의 특성을 잘 반영하는 적합한N-gram을 선택함을 확인할 수 있다

Table 4. Selected ten features based on each N-gram feature selection method

TF-IDF w/2~4-gram	SCP TF-IDF w/2~4-gram
the room	batteri life
the staff	ga mileag
batteri life	front desk
the hotel	tuscan inn
ga mileag	cabl car
great locat	the staff
staff wa	the room
the keyboard	the batteri
room were	speed limit
clean and	san francisco

6. 결 론

본 논문에서는 연속된 단어인N-gram을 이용하여 텍스트를 분류할 수 있는 새로운 특질 선택 방법을 제안하였다.제안 방법을 사용하여 특질을 선택하였을 때 텍스트를 얼마나 잘 분류하는 지 확인하기 위하여 기존 방법인TF-IDF를 이용하여 비교하였으며, 분류 알고리즘인 나이브 베이지안 분류기를 이용하여 예측 정확도를 산출한 결과 제안 방법이 기존 방법보다 높은 정확도로 텍스트를 분류할 수 있는 특질을 선택함을 확인하였다. 또한 선택된N-gram 특질을 확인한 결과 제안 방법이 기존 방법보다 적합한N-gram을 선택함을 확인하였다

한편, 본 논문에서는 텍스트의 길이에 따른 가중치를 고려하지 않았고, 적은 양의 텍스트를 가지고 있는 데이터셋을 활용하였다는 한계점이 있다. 이에 따라, 관측치가 더 많은 데이터셋을 이용하고 동시에 텍스트의 길이에 따른 가중치를 고려하였을 때, 제안한 특질 선택 방법에 따른 특질들이 기존의 방법에 비해 어느 정도로 텍스트를 정확하게 분류하는지에 대해 확인하는 것을 향후 과제로 계획하고 있다

참고문헌

Bache, K. and Lichman, M. (2013), *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Science, http://archive.ics.uci.edu/ml.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992), Class-Based N-gram Models of Natural Language, *Computational linguistics*, **18**(4), 467-479.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011), Discriminating Gender on Twitter, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1301-1309, Association for Computational Linguistics.

Chemudugunta, C. and Steyvers, P. S. M. (2007), Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model, *Advances in Neural Information Processing Systems 19 : Proceedings of the 2006 Conference*, 19, MIT Press.

Cho, S. G. and Kim, S. B. (2012), Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining, *Journal of the Korean Institute of Industrial Engineers*, **38**(1), 67-73.

da Silva, J. F. and Lopes, G. P. (1999), A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units, *Sixth meeting on the Mathematics of Language*, 369-381.

Feldman, R. and Sanger, J. (2007), *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press.

Ganesan, K., Zhai, C., and Han, J. (2010), Opinosis : A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions, *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China.

Houvardas, J. and Stamatatos, E. (2006), N-Gram Feature Selection for Authorship Identification, *Artificial Intelligence : Methodology, Systems, and Applications*, 77-86, Springer Berlin Heidelburg.

Jing, L., Huang, H., and Shi, H. (2002), Improved Feature Selection Approach TFIDF in Text Mining, *Proceedings of the First International Conference on Machine Learning and Cybernetics*, 944-946, Beijing, China.

Li, Y. H. and Jain, A. K. (1998), Classification of Text Documents, *The Computer Journal*, **41**(8), 537-546.

Mukherjee, A. and Liu, B. (2010), Improving Gender Classification of Blog Authors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 207-217, Association for Computational Linguistics.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000), Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, **39**(2-3), 103-134.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay. E. (2011), Scikit-learn : Machine Learning in Python, *Journal of Machine Learning Research*, **12**, 2825-2830.

Phan, X., Nquyen, L., and Horiguchi. S. (2008), Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections, *Proceedings of the 17th International Conference on World Wide Web*, 91-100, ACM.

Python Software Foundation (2010), *Python Language Reference*, Ver-

sion 2.7, http://www.python.org/.

Ramos, J. (2003), Using TF-IDF to Determine Word Relevance in Document Queries, *Proceedings of the First Instructional Conference on Machine Learning*.

Salton, G. and McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernandez, L. (2014), Syntactic N-grams as Machine Learning Features for Natural Language Processing, *Expert Systems with Applications*, **41**, 853-860.

Silva, J. and Lopes, G. (2010), Towards Automatic Building of Document Keywords, Proceedings of the 23rd *International Conference on Computational Linguistics : Posters*, 1149-1157, Association for Computational Linguistics.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996), Translating Collocations for Bilingual Lexicons : A Statistical Approach, *Computation Linguistics*, **22**(1), 1-38.

Tang, B., Shepherd, M., Milios, E., and Heywood, M. I. (2005), Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering, *Proceeding of SIAM International Workshop on Feature Selection for Data Mining*, 17-26.

Ting, S. L., Ip, W. H., and Tsang, A. H. C. (2011), Is Naïve Bayes a Good Classifier for Document Classification?, *International Journal of Software Engineering and Its Applications*, **5**(3), 37-46.

Zaki, T., Es-saady, Y., Mammass, D., Ennaji, A., and Nicolas, S. (2014), A Hybrid Method N-Grams-TFIDF with Radial Basis for Indexing and Classification of Arabic Documents, *International Journal of Software Engineering and Its Applications*, **8**(2), 127-144.