Topic Signature와 동시 출현 단어 쌍을 이용한 문서 범주화

배원식 ○, 한요섭*, 차정원

국립창원대학교 컴퓨터공학과

wonsigi529@changwon.ac.kr, jcha@changwon.ac.kr

* 한국과학기술연구원(KIST)

emmous@kist.re.kr

Text Categorization using Topic Signature and Co-occurrence Features

Won-Sik Bae○, Yo-Sub Han*, Jeong-Won Cha

Changwon National University

* Korea Institute of Science and Technology(KIST)

요 약

본 논문에서는 문서 내에서 동시에 출현하는 단어 쌍을 자질 추출 단위로 하는 문서 범주화 시스템에 대하여 기술한다. 자질 추출 단위를 단어 쌍으로 정의한 것은 문서에서 빈번하게 동시에 출현하는 단어들은 서로 연관관계가 높으며, 단어 하나보다는 연관관계가 높은 단어들의 쌍이 특정 범주의 문서에서만 나타날 확률이 높아지므로 문서 분류 능력을 높이는데 좋은 요인으로 작용할 수 있을 것이라는 가정 때문이다. 그리고 문서 요약 분야에서 제안된 Log-likelihood Ratio를 기반으로 하는 Topic Signature Term Extraction 방법을 사용하여 자질 추출을 하고, Naïve Bayes 분류기를 이용하여 문서를 분류한다. 본 연구는 Reuters-21578 문서 집합을 이용한 성능평가에서 좋은 결과를 보였으며, 이는 앞으로의 연구에도 기여할 수 있을 것이라 기대한다.

1. 서론

문서 범주화는 텍스트 형태의 문서를 미리 정의되어 있는 하나 이상의 범주에 할당하는 문제를 다루는 분야로 주로 정보 검색 분야와 기계 학습 분야에서 활발하게 연구되고 있다. 문서 범주화의 핵심은 자질 추출 과정과 문서 분류 과정인데, 이들 과정에서 어떤 방법을 사용하느냐에 따라 시스템의 성능이 달라진다. 자질 추출 과정에서는 학습 문서에서 나타나는 많은 단어들 중에서 문서의 범주를 판단하는데 유용하게 쓸 수 있는 단어를 추출하는 과정으로 주로 TF-IDF, 상호 정보(Mutual Information), 카이 제곱 통계량(χ^2 Statistics), 정보 획득량(Information Gain) 등의 방법이 사용된다[1]. 문서 분류 과정에서는 자질 추출 과정에서 추출된 단어들을 이용하여 문서의 범주를 판단할 수 있는 학습 모델을 만들고, 이 학습 모델을 이용하여 문서의 범주를 할당한다. 학습 모델은 주로 기계 학습 방법에 의해 만들어지는데, 주로 대표적으로 Naïve bayes 모델[2][3], 지지 벡터 기계(Support Vector Machine)[4][5], K-NN(K- Nearest Neighbor Classification)[6], 신경망(Neural Network)[7] 등이 사용된다.

기존 연구들에서 자질 추출 단위는 일반적으로 문서에서 나타나는 단어 중에서 불용어(Stop-words)를 제거하고, 스테밍(Stemming) 후 남는 단어의 어간(Stem)을 사용한다. 그러나 본 연구에서는 자질 추출 단위로 문서 내에서 빈번하게 동시에 출현하는 단어의 쌍으로 정의한다. 이는 문서에서 빈번하게 동시에 출현하는 단어들은 서로 연관관계가 높으며, 단어 하나보다는 연관관계가 높은 단어들의 쌍이 특정 범주의 문서에서만 나타날 확률이 높아지므로 문서 분류 능력을 높이는데 좋은 요인으로 작용할 수 있을 것이라는 생각으로부터 정의되었다.

본 논문의 구성은 다음과 같다. 2장에서는 Reuters-21578 문서 집합과 Reuters 문서 집합을 사용했던 이전 연구들에 관련된 내용을 소개한다. 3장에서는 본 연구에서 제안하는 시스템에 대하여 자세히 설명하고, 4장에서는 실험 준비를 위한 설정을 소개하고, 실험의 결과와 결과에 대해 분석하고, 5장에서는 결론과 향후 과제를 다룬다.

2. 관련연구

문서 범주화에서 전통적으로 벡터 공간 모델(Vector

Space Model)은 문서를 표현하는 방법으로 이용되고 있다[8]. 벡터 공간 모델에서 문서는 자질의 벡터로 표현되며, 자질은 TF(Term Frequency)와 IDF(Inverse Document Frequency)에 의해 표현된다. 이 때, TF는 자질이 나타난 위치에 관계 없이 단순히 나타난 횟수만으로 결정된다. 고영중[9]은 문서의 각 문장은 문서의 내용을 식별하는데 서로 다른 중요도를 가지며, 중요한 문장일수록 중요하지 않은 문장보다 더 높은 가중치를 부여하여 문서를 문장의 중요도에 따라 다른 가중치를 갖는 자질 벡터로 표현하는 새로운 방법을 제시하였다. [9]에서 문서의 중요도는 두 가지 방법에 의해서 결정된다. 첫 번째 방법은 일반적으로 문서의 중요한 내용을 요약하고 있을 것이라고 여겨지는 문서의 제목과 각 문장과의 유사도(Similarity)를 계산하여 유사도가 높은 문장이 중요도가 높다고 판단하여 가중치를 부여한다. 단, 이 방법은 제목의 질에 의존적이므로 제목이 별다른 의미를 갖지 않거나 제목이 없는 경우에는 사용할 수 없다. 두 번째 방법은 카이 제곱 통계량과 TF, IDF를 사용하여 용어의 중요도를 구하고, 중요도가 높은 용어를 포함하는 문장에 가중치를 부여한다. 이 방법은 제목과 유사하지 않더라도 중요도가 높은 용어를 많이 포함하는 문장은 중요도가 높은 문장일 가능성이 높기 때문에 사용된다. 이 두 가지 방법을 조합하여 최종 문장 중요도를 계산하여 문장의 중요도에 따라 다른 가중치로 TF를 구한다. 가중치가 있는 TF와 IDF를 이용하여 문서를 색인하고, 이를 여러 문서 분류기를 이용하여 문서 범주화를 한 결과, 문장의 중요도를 고려하지 않았을 때보다 실험에 사용된 모든 문서 분류기에서 어느 정도 성능이 향상됨을 보였다. Sebastiani는 [10]에서 Reuters-21578 문서 집합의 여러 학습/실험 분리 방법을 이용하여 실험하였다. [표 1]은 Sebastiani에 의해 보고된 결과 중에서 "ModApte" 분리 방법을 이용한 결과 중에서 좋은 결과만 요약해서 정리한 표이다. Micro-averaged breakeven point(BEP) 성능 평가 방법을 사용하여 성능을 평가한 것이라 F_1-measure를 이용한 성능과 직접적으로는 비교할 수 없겠지만 주목할만한 결과를 보이고 있다.

표 1. 기존 시스템들의 성능(BEP)

Results reported by	R(90)	Top-10
(Joachims, 1998)	86.4	-
(Dumais et al., 1998)	87.0	92.0
(Weiss et.al., 1999)	87.8	-

[표 2]는 좀 더 최근에 보고된 결과로 Micro-averaged F_1-measure를 사용하여 성능을 평가하였다. [11]에서 보고된 성능 표에 [11]에서 제안된 시스템의 성능 평가 결과를 추가하였다.

표 2. 기존 시스템들의 성능(F_1-Measure)

Results reported by	R(90)	Top-10
(Gao et al., 2003)	88.42	93.07
(Kim et al., 2005)	87.11	92.21
(Gliozzo and Strapparava, 2005)	-	92.80
(Zelaia et al., 2006)	87.27	**93.57**

본 장의 남은 절에서는 일반적으로 영어 문서 분류 시스템의 평가에 이용되는 Reuters-21578 문서 집합의 특성에 대해 다룬다.

2.1. Reuters-21578 문서 집합

영어 문서 범주화 시스템의 성능 평가에는 Reuters-21578 문서 집합이 많이 이용된다[1][10][11][12]. Reuters-21578 문서 집합은 다중 범주(Multi-class)와 중복 범주(Multi-label)의 특성을 가지고 있는데, 다중 범주는 범주화를 할 수 있는 범주의 수가 두 개 이상 존재한다는 것을 의미하며, 중복 범주는 한 문서에 두 개 이상의 범주가 할당될 수 있다는 것을 의미한다. Reuters-21578 문서 집합은 총 135개의 범주와 각 문서당 평균 1.2개의 범주가 할당되어 있다[11].

기존 시스템들과의 성능 비교를 위해서는 동일한 문서 집합을 이용하는 것은 물론이거니와 문서 집합을 동일한 학습문서와 실험문서로 분리하여 성능을 평가해야 하는데, Reuters-21578 문서 집합은 문서 집합에서 학습문서와 실험문서를 분리하는 몇 가지 방법을 제시하고 있다. 그 중에서 일반적으로 사용되며 권장되는 방법은 "ModApte" 분리 방법[12]이다.

3. 시스템

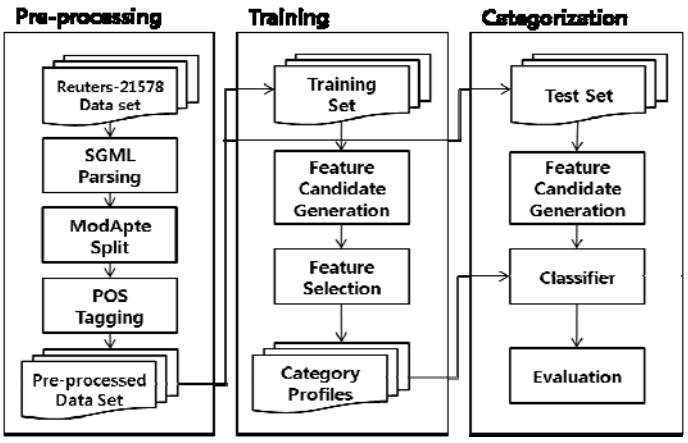

그림 1. 본 연구에서 제안된 시스템의 구성도

본 장에서는 본 연구에서 사용된 방법에 대하여 자세하게 설명한다. 본 연구에서 제안된 시스템의 전체

구성도는 [그림 1]과 같다.

3.1. 전처리 과정

SGML 분석기(Parser)를 사용하여 Reuters-21578 문서 집합으로부터 문서의 제목과 본문을 추출하여 별개의 파일로 분리하고, "ModApte" 분리 방법에 의해 학습 문서와 실험 문서로 분리한다. 그리고 품사 부착기(POS-tagger)를 사용하여 품사가 부착된 문서로 만든다. 이 품사가 부착된 문서를 학습과 실험에 사용한다. 품사 부착기는 일반적으로 단어의 어간을 찾기 위해 사용하는 스테머를 대신하여 단어의 원형을 복원하는데 사용하였고, 품사 부착기로부터 얻어진 단어의 품사 정보는 불용어 사전을 대신하여 불용어를 제거하고, 자질 후보로 추출할 단어와 배제할 단어를 구분하기 위하여 사용하였다.

3.2. 동시 출현 단어 쌍 단위로 자질 후보 생성

본 연구에서의 자질 추출 단위는 [그림 2]와 같이 기준 단어(Basis Word)와 그 단어를 기준으로 앞뒤로 일정 크기의 윈도우 내의 단어와의 쌍이 된다. 예를 들어 윈도우의 크기가 4이고, 기준 단어가 w_3라고 한다면, w_3와 쌍이 될 수 있는 단어는 w_1, w_2, w_4, w_5가 되며, 이로부터 4개의 단어 쌍 (w_3, w_1, N), (w_3, w_2, N), (w_3, w_4, P), (w_3, w_5, P)이 생성된다. 이와 같은 방식으로 문서의 처음에 나타난 단어 w_1부터 마지막 단어인 w_n까지 슬라이딩 윈도우 기법을 이용하여 자질을 생성한다. 여기서 자질의 후보가 되는 단어는 문서에서 나타난 모든 단어가 아니라, 그 단어의 품사가 동사나 명사, 형용사에 해당되는 단어이며, 그 외의 단어는 사용하지 않는다. 이렇게 생성된 단어 쌍들 중에서 일정 빈도수 이상으로 출현한 단어 쌍을 자질 후보로 생성한다.

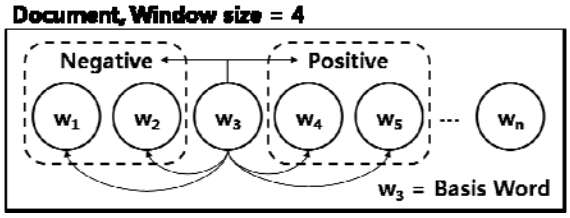

그림 2. 동시 출현 단어 쌍의 개념

3.3. Topic Signature Term Extraction

본 연구에서는 자질 추출 방법으로 일반적으로 문서 범주화 시스템에서 자질 추출에 사용되는 카이 제곱 통계량이나 정보 획득량을 대신하여, 문서 요약 분야에서 Chin-yew Lin[13]이 제안한 Log-likelihood Ratio를 기반으로 하는 Topic Signature Term Extraction 방법을 사용한다.

만약 길이가 n인 문서를 D = {w_1, w_2, w_3, ..., w_i, w_{i+1}, ..., w_n}라고 정의하면, Topic Signature 방법으로 학습하기 위한 자질 후보의 추출 단위는 다음 식(1)과 같다

$$(w_i, w_j, A),\ A = \{P, N\},\ i-\alpha \le j \le i+\alpha \quad (1)$$

여기서 w_i는 기준 단어이고, w_j는 w_i를 기준으로 일정 크기의 윈도우 내에 존재하는 단어이며, A는 기준 단어의 앞쪽의 단어와의 쌍(P)인지, 뒤쪽의 단어와의 쌍(N)인지 여부를 나타낸다. 그리고 α는 윈도우의 크기 절반 값이다.

표 3. Contingency 표

	R	~R
t_i	O_{11}	O_{12}
~t_i	O_{21}	O_{22}

특정 자질이 한 범주에만 나타나는 경우, 즉 위 Contingency 표의 O_{12}가 0이 되는 경우, Topic Signature 값이 매우 커지는데, 이 때, O_{11}은 Topic Signature 값의 크기에 영향을 미치지 못한다. 그리고 O_{11}이 매우 작으면서 O_{12}가 0인 경우가 O_{11}이 매우 크고, O_{12}가 매우 작은 경우보다 훨씬 큰 값을 갖는데, 이는 문서 분류를 방해하는 요소로 작용할 수 있다. 이 문제를 보완하기 위하여 Topic Signature 값을 그대로 사용하지 않고 TF를 곱해줌으로써 O_{11}이 Topic Signature 값의 크기에 영향을 줄 수 있도록 하였다.

3.4. 제목 나타난 자질 후보에 가중치 부여

일반적으로 문서의 제목은 문서의 내용을 요약하고 있으므로 제목에 나타난 단어들을 본문에 나타난 단어보다 중요하게 처리하는 것이 문서 분류에 도움이 되는 요소로 작용할 수 있을 것이다. 그래서 위 Contingency 표에서 O_{11}과 O_{12}를 각각 제목과 본문 중에 나타난 위치에 따라 O_{111}, O_{112}, O_{121}, O_{122}로 다시 나눈다. O_{111}은 특정 범주의 제목에서 나타난 t_i의 빈도수, O_{121}은 특정 범주 외의 제목에서 나타난 t_i의 빈도수, O_{112}는 특정 범주의 본문에서 나타난 t_i외의 용어의 빈도수, O_{122}는 특정 범주 외의 제목에서 나타난 t_i외의 용어의 빈도수이다. 우리는 제목에서 나타난 용어의 정보량은 본문에서 나타난 용어들의 정보량과 같다고 가정하여 제목에 나타난 용어에 대한 가중치 w를 다음과 같이 구하고, O_{11}과 O_{12}를 다음과 식(2)와 같이 수정한다.

$$O_{11} = O_{111} \times w + O_{112},\ O_{12} = O_{121} \times w + O_{122}$$
$$w = \frac{\text{본문에서 나타난 고유한 용어의 수}}{\text{제목에서 나타난 고유한 용어의 수}} \quad (2)$$

3.5. 평탄화(Smoothing)

학습되지 않은 동시 출현 단어 쌍이 실험 문서에서 자질 후보로 나타나는 경우, 이에 대처하기 위해 식(3)을 사용한다.

$$TS_c = TS_c(w_i,*,*) \quad \text{if } TS_c(w_i,w_j,A) = 0,\ A = \{P, N\} \quad (3)$$

여기서 TS_c는 범주 c일 때, 실험 문서로부터 생성된 동시 출현 단어 쌍 자질 후보가 갖게 될 Topic Signature 값이고, $TS_c(w_i, w_j, A)$는 범주 c일 때, 학습 문서로부터 학습된 동시 출현 단어 쌍 자질의 Topic Signature 값이다. 그리고 $TS_c(w_i, *, *)$는 실험 문서로부터 생성된 동시 출현 단어 쌍 자질 후보가 학습 문서에서 학습되지 않은 자질일 경우, 평탄화를 위해 사용되는 값으로 동시 출현 단어 쌍에서 기준 단어의 Topic Signature 값이다. Topic Signature 값의 절대적인 크기는 O_{21}, O_{22}의 값에 영향을 받으며, 단일 단어에 비해 자질의 수가 많은 동시 출현 단어 쌍에 기반한 Topic Signature의 값이 상대적으로 너무 큰 값을 가져 평탄화 효과가 제대로 발휘되기 힘든 문제가 있다. 따라서 동시 출현 단어 쌍의 O_{21}, O_{22}를 단일 단어의 O_{21}, O_{22} 값으로 대신하여 두 값의 상대적인 크기 차이를 너무 나지 않도록 하였다. 이는 동시 출현 단어 쌍이 단일 단어에 비해서 한 단어를 기준으로 윈도우 크기만큼 더 생성된 여분 자질(Dummy Feature)에 의해 O_{21}, O_{22} 값이 커져 Topic Signature의 값도 커진 것이므로, 여분 자질을 제거하고 Topic Signature 값을 계산하는 것이다.

3.6. 문서 범주화

문서 범주화 과정에서는 학습 문서로부터 생성된 범주별 프로파일과 실험 문서로부터 학습 과정과 같은 방식으로 생성된 자질 후보를 이용하여 문서를 분류한다. 실험 문서에서 생성된 자질 후보와 모든 범주별 프로파일을 비교하여 각각의 자질 후보와 문서의 Topic Signature 값을 구하고, Naïve Bayes 분류기를 사용하여 문서의 범주를 할당한다.

4. 실험

4.1. 실험 데이터

Reuters-21578 문서 집합에 존재하는 총 135개의 범주 중에서 문서수가 많은 순서대로 상위 10개의 범주에 대하여 실험하였다.

4.2. 실험 파라미터(Parameter)

4.2.1. 단일 단어 파라미터

단일 단어는 절대 빈도와 상대 빈도 파라미터를 갖는데, 절대 빈도는 범주별 학습 문서량에 관계 없이 절대적인 빈도수를 기준으로, 상대 빈도는 범주별 학습 문서량에 비례하여 상대적인 빈도수를 기준으로 기준보다 높이 출현하는 단어만 학습에 이용한다. 빈도수 F = {2, 5, 7, 10, 15, 20}의 6개로 하여 실험을 진행하였으며, 상대 빈도의 경우 (범주별 총 고유 단어 수 * F * 0.0001)로 하여 상대적인 빈도수를 결정하였다.

4.2.2. 동시 출현 단어 쌍 파라미터

동시 출현 단어 쌍은 파라미터로 윈도우의 크기 W = {10, 20, 30, 40, 50, 60}의 6개와 절대 빈도 F = {2, 5, 7, 10}의 4개로 하여 실험을 진행하였다. 상대 빈도로는 실험을 하지 않았다.

4.2.3. 문서 범주 할당 임계값

Reuters-21578 문서 집합은 한 문서에 여러 범주가 할당될 수 있는 중복 범주의 특성을 가지고 있으므로, 본 연구에서는 [11]과 유사한 방식으로 순위화된 범주 목록 1, 2등의 Topic Signature 값을 이용하여 다음 식(4)를 만족하면 2등의 범주도 문서에 할당한다.

$$c_{rank1} >= c_{rank2} \times t \quad t = \{0.01, 0.02, ..., 0.99, 1.00\} \quad (4)$$

4.3. 실험 결과

표 4 성능 결과 표(Micro-averaged F_1-measure)

실험 그룹	Origin	TF	log(TF)
(1) 단일, TS, 절대	85.14	82.87	**86.39**
(2) 단일, TS, 상대	86.93	85.74	**87.41**
(3) 단일, χ^2, 절대	**82.29**	70.09	80.01
(4) 단일, χ^2, 상대	**85.42**	71.43	83.34
(5) 단어 쌍, TS	92.32	90.03	**92.50**
(6) 단어 쌍, χ^2	**87.79**	75.29	86.12
(7) (5)+가중치(제목)	92.69	91.93	**93.13**
(8) (7)+평탄화	92.67	91.59	**93.24**

실험은 총 8개의 그룹으로 구성되어 있다. (1)번부터 (4)번까지의 실험은 단어 하나를 자질 추출 단위로 한 것이며, 나머지는 동시 출현 단어 쌍을 자질 추출 단위로 한 것이다. 또한 절대와 상대는 4.2.1절에서 언급한 절대 빈도와 상대 빈도를 나타낸다. TS는 Topic Signature, χ^2는 카이 제곱 통계량을 각각 자질 추출 방법으로 사용하여 실험하였음을 나타낸다. 이처럼 각각의 실험 그룹은 4.2절에서 언급한 실험 파라미터를 조절하며 실험을 진행하였다. [표 4]는 8가지 그룹의 실험 집합에서 가장 성능이 높은 실험의 결과만을 표로 정리한 것이다. 성능은 Micro-averaging 기법을 이용한 F1-measure(Micro-averaged F_1-measure)를 사용하여 측정하였다.

4.3.1. 자질 추출 단위

[표 4]의 (1)~(4)번 실험과 (5)~(7) 실험을 통해 자질 추출 단위에 대한 비교를 할 수 있다. (1)~(4)번 실험은 기존에 사용되는 단일 단어에 대한 실험이고, (5)~(7)번 실험은 본 논문에서 제안한 동시 출현 단어 쌍에 대한 실험이다. [표 4]를 보면 동시 출현 단어 쌍을 이용하는 실험이 단일 단어를 이용하는 것보다 모두 높은 성능을 보이고 있다. 그리고 단일 단어를 자질 추출 단위로 하였을 때는 절대 빈도보다는 상대 빈도로 한 실험이 성능이 조금 더 높은 결과를 보였다.

4.3.2. 자질 추출 방법(TS vs. χ^2)

[표 4]의 (1)번과 (3)번, (2)번과 (4)번, (5)번과 (6)번 실험을 통해 자질 추출 방법에 대한 성능 비교를 할 수 있다. 본 연구에서는 자질 추출 방법으로 Topic Signature(TS)를 사용하는 실험이 χ^2 통계량을 사용하는 실험보다 훨씬 높은 성능을 보였다.

4.3.3. TF-TS

[표 4]의 각 실험 그룹에 대한 3개의 열(Origin, TF, log(TF))을 보면 이 실험의 결과를 알 수 있다. 모든 실험 그룹에 대하여 TF를 Topic Signature 값이나 χ^2 값에 곱하면 학습할 때 TF가 너무 높은 자질이 도움이 되기보다 오히려 문서 분류에 방해 요인으로 작용하여, 성능이 떨어지는 결과를 보였다. 그래서 TF에 log를 취한 실험에서 TF를 곱하지 않았을 때보다 높은 성능을 보였다. 그러나 χ^2 통계량을 사용한 실험에서는 TF를 곱하지 않고 원래 값을 그대로 사용하는 경우가 더 성능이 높았다. log(TF)를 곱하는 것은 Topic Signature를 사용하는 경우에 유효한 방법인 것으로 보인다

4.3.4. 제목에 가중치 부여

[표 4]의 (5)번과 (7)번 실험을 통해 제목을 비중 있게 다루는 것이 그렇지 않은 것보다 문서 분류에 도움이 된다는 것을 알 수 있다. 단, 이는 제목의 질에 영향을 받으므로 모든 경우에 일반화하기는 힘들 것으로 보이나 어느 정도는 성능 향상에 도움이 되는 요인이 될 수 있을 것으로 보인다.

4.3.5. 평탄화

[표 4]의 실험 (7)번과 (8)번을 비교해보면 Topic Signature에 평탄화를 한 것과 하지 않은 것의 성능을 비교할 수 있다. log(tf)를 곱하는 경우 성능 향상에 도움이 되지만 TF를 곱하지 않은 순수 Topic Signature(Origin)와 TF를 곱하는 두 실험에서는 성능이 떨어지는 결과를 볼 수 있다. 그리고 [표 5]의 평탄화 실험을 살펴보면 평탄화를 하는 것이 하지 않는 것보다 잠재적 정확도가 향상되는 것을 볼 수 있다. 따라서 평탄화는 성능 향상에 도움이 되는 요인이며, 평탄화의 방법에 따라 성능 변화에 다소 차이가 있을 수 있다는 것을 알 수 있다.

4.3.6. Potential Top3 F_1-measure

Potential Top3 F_1-measure는 [14]에서 시스템의 잠재적 정확도(Potential Accuracy)를 측정하기 위한 성능 평가 방법으로 사용되었다. 문서에 할당된 범주가 순위화된 범주 목록에서 3등 내에만 존재한다면 제대로 문서의 범주를 할당하였다고 평가하는 이 방법은 향후 시스템이 얼마나 성능 향상의 여지가 있는가를 나타내는 척도이다. [표 5]에서 살펴본 본 연구의 최고 잠재적 정확도는 (8)번 실험의 99.45%로 3등 이내에는 거의 모든 문서의 범주가 할당된다고 볼 수 있다. 같은 방식으로 Potential Top2 F_1-measure를 이용하여 본 연구를 평가해 본 결과, (8)번 실험이 98.01%로 최고 잠재적 정확도를 보였다.

표 5 잠재 정확도 표(Potential Top3 F_1-measure)

실험 그룹	Origin	TF	log(TF)
(1) 단일, TS, 절대	97.25	**97.71**	97.62
(2) 단일, TS, 상대	98.15	97.81	**98.31**
(3) 단일, χ^2, 절대	**97.55**	95.07	97.11
(4) 단일, χ^2, 상대	**97.82**	93.59	97.36
(5) 단어 쌍, TS	98.89	98.86	**99.12**
(6) 단어 쌍, χ^2	**98.68**	94.51	97.92
(7) (5)+가중치(제목)	99.04	99.04	**99.37**
(8) (7)+평탄화	99.20	99.11	**99.45**

5. 결론 및 향후 과제

5.1. 결론

실험을 통해 자질 추출 단위를 단어 하나로 할 때보다 문서에서 동시에 출현하는 단어의 쌍으로 할 때 문서 범주화 시스템의 성능이 향상되는 결과를 확인할 수 있었다. 또한 본 연구에서 사용한 Topic Signature 방법이 기존에 자질 추출 방법으로 좋은 성능을 보이는 χ^2 통계량 방법에 못지 않은 성능을 발휘한다는 것을 확인할 수 있었다. 그리고 제목에 나타나는 자질을 본문에 나타나는 자질보다 중요하게 다룬 것과 Topic Signature의 단점을 보완하기 위해 TF를 곱한 것, 평탄화를 하는 것도 성능 향상에 좋은 영향을 미치는 것을 확인할 수 있었다. 현재까지 보고된 Reuters-21578 문서 집합을 이용한 문서 범주화 시스템 중에서 최고의 성능을 내지는 못하였으나 높은 성능을 보였으며, 본 연구의 잠재적 정확도를 볼 때, 다른 자질을 추가하거나 평탄화 방법이나 다른 방법들을 개선한다면 충분히 더 좋은 성능을 낼 수 있을 것이다.

5.2. 향후 과제

본 연구가 Reuters-21578 문서 집합에서 높은 성능을 보이는 것은 문서 집합의 특성과 서로 잘 맞았기 때문이라고 볼 수도 있으므로, 향후 본 연구에서 사용한 방법을 다른 문서 집합에 적용하여 성능을 평가해보고 본 연구의 일반성을 확인할 필요성이 있다. 그리고 동시 출현 단어 쌍을 단위로 자질 후보를 생성하면 단일 단어에 비해 훨씬 많은 자질 후보가 생성되며, 윈도우의 크기가 크면 클수록 더욱 많은 자질 후보가 생성되게 되는데, 그 자질 후보 중에서는 불필요한 자질 후보 또한 단일 단어에 비해 훨씬 많이 생성될 것이다. 그러므로 문서 집합의 특성에 맞는 최적 윈도우 크기를 결정하는 방법과 불필요한 자질 후보를 줄여 계산 비용을 줄이는 방법에 관한 연구도 필요할 것으로 보인다. 끝으로 영어 문서가 아닌 한글 문서에도 본 연구 방법을 적용하여 한글에 맞도록 개선하여 한글 문서 분류에도 이용할 수 있도록 할 필요성이 있다.

참고문헌

[1] Y.Yang and J.O.Pederson, “A comparative study on feature selection in text categorization”, In Proceedings of the 14th International Conference on Machine Learning, pp. 412-420, 1997.

[2] David D.Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval”, In European Conference on Machine Learning, pp. 4-15, 1998.

[3] Andrew McCallum and Kamal Nigram, “A Comparison of Event Models for Naive Bayes Text Classification”, AAAI ‘98 Workshop on Learning for Text Categorization, pp. 41-48, 1998.

[4] C.Cortes and V.Vapnik, “Support vector networks,” Machine Learning, 20, pp. 273-297, 1995.

[5] Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, In European Conference on Machine Learning (ECML), pp. 137-142, 1998.

[6] Y.Yang, “Expert netword: Effective and efficient learning from human decisions in text categorization and retrieval”, In 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’94), pp. 13-22, 1994.

[7] E.Wiener, J.O.Pedersen and A.S. Weigend, “A neural network approach to topic spotting”, In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR’95), pp. 317-332, 1995.

[8] Salton G., Yang C. and Wang A., “A vector space model for automatic indexing”, Communications of the ACM, Vol. 18, No. 11, pp. 613-620, 1975.

[9] Youngjoong Ko, Jinwoo Park and Jungyun Seo, “Automatic Text Categorization using the Importance of Sentences”, In Proceedings of the 19th COLING, Taipei, pp. 474-480, 2000.

[10] Sebastiani, F.: “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, 34(1), pp. 1-47, 2002.

[11] Ana Ana Zelaia, Inaki Alegria, Olatz Arregi and Basilio Sierra, “A Multiclassifier based Document Categorization System: profiting from the Singular Value Decomposition Dimensionality Reduction Technique”, In Proceedings of the workshop on Learning Structured Information in Natural Language Applications 11th EACL, pp. 25-32, 2006.

[12] Lewis, D.D., Yang, Y., Rose, T.G. and Li, F, “RCV1: A New Benchmark Collection for Text Categorization Research”, Journal of Machine Learning Research, 5, pp. 361-397, 2004.

[13] Chin-Yew Lin and Eduard Hovy, “The Automated Acquisition of Topic Signatures for Text Summarization”, In Proceedings of the 18th COLING, Strasbourg, France, pp. 495-500, 2000.

[14] Y. Yoon, G.G. Lee, “Efficient implementation of associative classifiers for document classification”, Information Processing and Management 43, pp. 393 - 405, 2007.