http://dx.doi.org/10.3745/KIPSTB.2012.19B.4.237

Modified ECCD 및 문서별 범주 가중치를 이용한 문서 분류 시스템

한 정 석[†] · 박 상 용[†] · 이 수 원[††]

요 약

웹 문서 정보 서비스는 관리자의 효율적 문서관리와 사용자의 문서검색 편의성을 위해 문서 분류 시스템을 필요로 한다. 기존의 문서 분류 시스템은 분류하고자 하는 문서 내 선택된 자질어의 개수가 적거나, 특정 범주의 문서 비율이 높아 그 범주에서 대부분의 자질어가 선택되어 모델이 생성된 경우 분류 정확도가 저하되는 문제점을 가진다. 이러한 문제점을 해결하기 위해 본 논문에서는 'Modified ECCD' 기법 및 '문서별 범주 가중치' 특징 변수를 사용한 문서 분류 시스템을 제안한다. 실험 결과, 제안 방법인 'Modified ECCD' 기법이 χ^2 및 ECCD 기법에 비해 높은 분류 성능을 보였으며, '문서별 범주 가중치' 특징 변수를 'Modified ECCD' 기법으로 선택된 자질어 변수에 추가하여 학습하였을 경우에 더 높은 분류 성능을 보였다.

키워드 : 문서 분류, 자질어, 범주, ECCD

A Document Classification System Using Modified ECCD and Category Weight for each Document

Chungseok Han[†] · Sangyong Park[†] · Soowon Lee[††]

ABSTRACT

Web information service needs a document classification system for efficient management and conveniently searches. Existing document classification systems have a problem of low accuracy in classification, if a few number of feature words is selected in documents or if the number of documents that belong to a specific category is excessively large. To solve this problem, we propose a document classification system using 'Modified ECCD' feature selection method and 'Category Weight for each Document'. Experimental results show that the 'Modified ECCD' feature selection method has higher accuracy in classification than χ^2 and the ECCD method. Moreover, combining the 'Category Weight for each Document' feature value and 'Modified ECCD' feature selection method results better accuracy in classification.

Keywords : Document Classification, Feature Selection, ECCD

1. 서 론

웹 문서 정보 서비스는 서비스 사업자에 따라 서로 다른 문서 분류 체계(범주)를 가지고 있다. 웹 문서 정보 서비스를 제공하기 위해서는 각각의 문서를 수동적으로 해당 분류 체계에 맞게 전문가가 직접 분류해야 하는 번거로움이 있다. 이러한 방식은 문서의 양이 점점 많아지고 복잡해질수록 많은 시간과 노력을 필요로 한다. 따라서 관리자의 효율적 문서관리와 사용자의 문서검색 편의성을 위해 범주별 문서 자동 분류 시스템이 요구된다.

자동 문서 분류 기법은 Salton[1]에 의해 체계화되었다. 자동 문서 분류 기법은 사전에 범주가 정의된 문서에서 선택된 자질어 변수를 학습하여 분류 모델을 생성하며, 생성

※ This work (Grants No. 00045616) was supported by Business for Academic-industrial Cooperative establishments funded Korea Small and Medium Business Administration in 2011, the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2011-0027668), and the Industrial Strategic Technology Development Program (10035348, Development of a Cognitive Planning and Learning Model for Mobile Platforms) funded by the Ministry of Knowledge Economy(MKE, Korea).
† 준 회 원 : 숭실대학교 컴퓨터학과 박사과정
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수
논문접수 : 2012년 4월 10일
수 정 일 : 1차 2012년 5월 22일
심사완료 : 2012년 5월 22일
* Corresponding Author : Soowon Lee(swlee@ssu.ac.kr)

된 분류 모델로 새로운 문서를 예측된 범주에 할당하는 방식이다. 하지만 분류 모델 생성에 사용될 문서의 수가 적거나, Fig. 1과 같이 특정 범주의 문서 수가 다른 범주에 비해 상대적으로 많은 차이를 갖는 경우 분류 정확도가 저하되는 문제가 있다[1][2].

그림 1. 연합뉴스의 범주별 문서 수
Fig. 1. Number of documents for each category in Yonhapnews

이러한 문제점을 해결하기 위하여 본 연구에서는 'Modified ECCD' 기법과 '문서별 범주 가중치'를 이용한 문서 분류 시스템을 제안한다.

본 논문에서는 제안하는 'Modified ECCD' 기법은 범주별로 비슷한 수의 문서 비율을 가정하는 'ECCD' 기법에 범주별 문서 비율을 고려하여 자질어를 선택한다. 또한 선택된 자질어가 예측될 문서 내에 출현하지 않거나 문서 내 출현하는 자질어가 문서 작성 시점에 따라 범주가 변경될 경우를 위하여 '문서별 범주 가중치'를 고려한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련된 기존 연구를 소개하고, 3장에서는 전체 시스템 구조와 본 논문에서 제안된 'Modified ECCD' 기법 및 '범주별 문서 가중치'에 대하여 기술한다. 4장에서는 3장에서 제안된 아이디어의 검증을 위한 실험 및 성능평가 결과를 분석한다. 마지막 5장에서는 본 연구의 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

문서 분류는 문서 및 문서의 범주를 대표할 수 있는 자질어 집합을 이용해 새로운 문서를 해당 범주로 자동 분류하는 기법이다. 일반적으로 문서 분류에서 분류 대상인 문서는 자질어의 집합으로 표현된다. 따라서 문서 분류의 정확도는 '자질어 선택 기법'과 '분류 모델 학습 알고리즘'에 많은 영향을 받는다.

2.1 자질어 선택 기법

자질어 선택기법은 문서 내 출현단어들 중 문서 분류에 효과적인 자질어들을 선정하며, 이는 효과적인 분류 모델을 생성하기 위함이다. 일반적으로 많이 사용되는 자질어 선택 기법은 Mutual Information(MI), Information Gain(IG), χ^2 (Chi-square), 그리고 ECCD(Entropy based Category Coverage Difference criterion) 등이 있다[3][4][5].

Mutual Information을 이용한 자질어 선택 기법은 해당 범주 c와 범주 내 존재하는 단어 t간 연관성을 이용하지만 IG, χ^2, ECCD 등을 이용한 자질어 선택 기법은 해당 범주 c와 범주 내 존재하지 않는 단어 t간 연관성도 고려한다. 특히 ECCD 자질어 선택 기법은 단어의 해당 범주 내 존재 여부뿐만 아니라 각 문서의 단어 출현 빈도(frequency)를 고려한 Shannon entropy를 사용함으로써 보다 정확한 자질어의 평가가 가능하다[4][5].

하지만 위에서 언급한 자질어 선택 기법들은 자질어 선택에 사용된 문서의 범주별 비율이 유사해야 하며, 예측될 문서 내에 선택된 자질어들 중 최소 한 개 이상의 자질어가 존재해야하는 문제점이 있다.

2.2 문서 분류 모델 학습 알고리즘

문서 분류 모델 학습에 사용되는 알고리즘은 SVM (Support Vector Machine)[7], C4.5[8][9], Naive Bayesian 등이 존재하며, 각 알고리즘은 서로 다른 분류 특징(선형, 규칙, 확률기반 등)이 있기 때문에 도메인에 맞는 알고리즘을 사용해야 최적의 분류기를 생성할 수 있다. 그러나 위에서 언급된 알고리즘은 문서의 길이 및 문서 내 자질어의 개수를 반영하기 어렵다는 문제점을 갖고 있다.

이러한 문제점을 해결하기 위해 Navie Baysian MultiNominal 알고리즘이 제안되었다[10][11]. Naive Bayesian MultiNominal 모델 학습은 문서 내 자질어의 개수와 각 문서의 길이를 반영할 수 있다. 따라서 타 분류 모델에 비해 각 문서의 길이 차이에 강인한 분류 모델을 생성한다.

3. 문서 분류 시스템

3.1 문서 분류 시스템 구조

본 연구에서 제안하는 문서 분류 시스템의 구조는 Fig. 2와 같다.

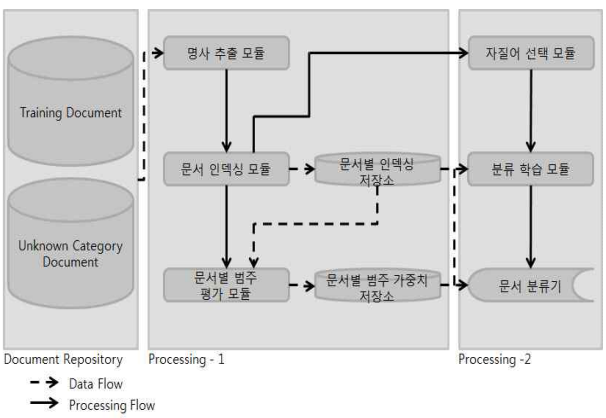

그림 2. 제안 문서 분류 시스템 구조
Fig. 2. Structure of the proposed document classification system

Document Repository 영역은 Crawling된 문서를 저장하는 영역이다. Crawling된 문서는 범주가 Labeling된 Training Document와 Labeling되지 않은 Unknown Category Document로 구분된다.

Processing-1 영역의 경우 Training Document가 수집될 때마다 Incremental하게 처리되며, '명사 추출 모듈', '문서 인덱싱 모듈' 및 '문서별 범주 평가 모듈'로 구성된다.

Processing-1의 수행 절차는 수집된 문서로부터 '명사 추출 모듈'에서 명사 원형만을 추출하고 추출된 원형을 '문서 인덱싱 모듈'로 전달한다. '문서 인덱싱 모듈'에서는 문서별 인덱싱 정보를 '문서별 인덱싱 저장소'에 저장한다. '문서별 범주 평가 모듈'에서는 '문서별 인덱싱 저장소'의 정보를 이용해 '문서별 범주 가중치' 특징 변수를 생성하고, 이를 '문서별 범주 가중치 저장소'에 저장한다. Processing-1 영역에서의 Incremental의 의미는 각각의 Training 문서가 추가될 때마다 '문서별 인덱싱 저장소' 및 '문서별 범주 가중치 저장소'의 정보가 업데이트되는 것을 의미한다.

Processing-2 영역의 경우 관리자가 정한 분류 모델의 업데이트 기간마다 주기적으로 수행되는 '자질어 선택 모듈' 및 '분류 학습 모듈'로 구성된다.

Processing-2의 수행 절차는 두 가지 과정을 수행한다. 첫째, '문서별 인덱싱 저장소'의 정보를 입력으로 한 '자질어 선택 모듈'에서 3.2절에 제안된 방법인 'Modified ECCD' 기법을 이용해 의미 있는 단어 집합 즉, 자질어 집합을 생성한다. 둘째, '분류 학습 모듈'은 생성된 자질어 집합 정보와 '문서별 인덱싱 저장소' 정보로부터 추출한 문서별 자질어 출현 빈도 값과 '문서별 범주 가중치 저장소'의 문서별 범주 가중치를 학습하여 문서 분류기를 생성한다.

3.2 Modified ECCD 기법

'Modified ECCD' 자질어 선택 기법은 기존 ECCD 자질어 선택 기법에 범주별 문서 비율을 고려하여 적정 개수의 자질어를 선택한다.

'Modified ECCD' 자질어 선택 기법은 식(1)의 ECCD 자질어 선택 기법으로 Shannon entropy $E(t_j)$가 고려된 범주별 자질어의 평가 값($ECCD(t_j,c_k)$)을 구한다.

- n_j^k : 전체 문서 중 j번째 단어의 k번째 범주에서의 출현 횟수
- $E(t_j)$: 단어 t_j의 Shannon entropy 값
- A : c_k에 속하며, t_j를 갖는 문서의 개수
- B : c_k에 속하지 않으며, t_j를 갖는 문서의 개수
- C : c_k에 속하며, t_j를 갖지 않는 문서의 개수
- D : c_k에 속하지 않으며, t_j를 갖지 않는 문서의 개수

$$
\begin{aligned}
tf_j^k &= \frac{n_j^k}{\sum^k n_j^k} \\
E(t_j) &= -\sum_{k=1}^{r}(tf_j^k)\times(\log_2(tf_j^k)) \\
ECCD(t_j,c_k) &= (P(t_j|c_k)-P(t_j|\overline{c_k}))\times\frac{E_{\max}-E(t_j)}{E_{\max}} \\
ECCD(t_j,c_k) &\approx \frac{AD-BC}{(A+C)(B+D)}\times\frac{E_{\max}-E(t_j)}{E_{\max}}
\end{aligned}
\tag{1}
$$

최종 자질어 집합의 생성은 범주별로 $ECCD(t_j,c_k)$ 값을 내림차순하여 정렬하고 범주별로 적정 개수의 자질어를 선택한 후, 각 범주별로 선택된 자질어를 union해 최종 자질어 집합을 생성한다. 범주 c_k에 대한 적정 자질어 개수 $SelectedFeature_{count}(c_k)$는 식(2)와 같이 구한다.

- $Document_{ratio}(c_k)$: 전체 문서 중 범주 c_k의 문서 비율
- $Init\,TotalFeature_{count}$: 초기 자질어 개수 설정 값

$$
\begin{aligned}
&SelectedFeature_{count}(c_k) \\
&= Document_{ratio}(c_k)\times Init\,TotalFeature_{count}
\end{aligned}
\tag{2}
$$

3.3 문서별 범주 가중치 특징 변수

문서 분류에서 자질어의 사용이유는 선택된 자질어가 특정 범주를 대표할 수 있기 때문이다. 문서 분류 모델 학습에서는 많은 자질어를 사용할수록 분류 성능은 오르지만 지나치게 많은 자질어 선택의 경우(모든 문서에 나오는 대부분의 단어가 자질어가 될 경우) 선택된 자질어 중 일부가 모든 범주에서 출현할 수 있기 때문에 분류 성능이 낮아지는 문제가 발생된다. 또한 잘 선택된 자질어도 자질어가 사용된 문서 작성 시간에 따라 다른 범주를 대표하는 자질어일수도 있다. 예를 들어 "안철수"라는 자질어의 경우 2011년 초에는 IT/과학 범주에 주로 출현하지만 2011년 후반에는 대선 후보에 언급되면서 정치 쪽 범주에 주로 출현하는 현상을 보인다.

본 논문에서는 이러한 문제점을 해결하기 위해 '문서별 범주 가중치'라는 특징 변수를 제안한다. '문서별 범주 가중치'는 문서마다 연산되며, 이는 분류 모델 학습에 이용한다.

문서별 범주 가중치는 다음 두 Step으로 연산된다.

- Step 1 : 본 논문에서 정의된 RDC(Retrieved Document Count)함수는 문서 단위로 인덱싱 된 각각의 단어들을 query로 포털(Naver)에서 제공되는 범주별, 문서 작성 날짜별 웹 뉴스 문서 검색 결과 개수를 추출한다.
- Step 2 : 본 논문에서 정의된 DCW(Document Category Weight) 함수는 Step1의 RDC함수 결과인 범주별 문서 검색 개수를 전체 범주 문서 검색 개수로 나누어 정규화하며, 문서 단위로 연산된 DCW를 제안된 '문서별 범주 가중치' 특징 변수로 사용한다.
  - $RDC(T(d_i),t_{i,j}c_k)$: 문서 d_i의 j번째 단어$t_{i,j}$를 query로 하였을 때, 문서 d_i의 작성 날짜 $T(d_i)$에 k번째 범주c_k에서 검색된 문서의 개수
  - $DCW(d_i,c_k)$: 범주c_k의 문서 d_i에서의 가중치

'문서별 범주 가중치' $DCW(d_i,c_k)$는 식(3)에 따라 연산되며, 문서 분류 모델 학습에서 자질어 이외의 추가 학습 변수로 사용된다. (t_i는 i번째 문서에서의 단어벡터)

$$
\begin{aligned}
&DCW(d_i,c_k)= \\
&\sum_{j=1}^{|t_i|}\left\{RDC(T(d_i),t_{i,j},c_k)/\left\{\sum_{k=1}^{|c_k|}RDC(T(d_i),t_{i,j},c_k)\right\}\right\} \\
&\qquad\qquad |t_i|
\end{aligned}
\tag{3}
$$

3.4 Naive Bayesian Multinominal 분류 모델

제안한 문서 분류 시스템에 사용된 Naive Bayesian Multinominal 알고리즘은 많은 수의 자질어가 사용되거나, 문서길이에 독립적인 문서 분류 시스템에서 평균적으로 높은 성능을 보이는 알고리즘이다[12][13]. 해당 알고리즘은 [11]에서 제안된 알고리즘이며, 알고리즘(Naive Bayesian) 특성상 별도의 파라미터 설정이 필요 없다.

4. 실험 결과 및 분석

4.1 실험 환경

본 연구에서는 제안하는 문서 분류 시스템의 성능 평가를 위해 인터넷 연합뉴스 데이터(2011년7월4일 - 2011년7월31일)를 수집하였다. 총 문서의 수는 3,806개이다(범주별 분포는 Fig. 1 참조). H/W 및 개발 환경은 Table 1과 같다.

표 1. H/W 및 개발 환경
Table 1. H/W and development environment

H/W 환경	CPU	Intel(R) i7 3.07GHz
	RAM	12G
개발 환경	OS	Windows 7 64bit
	개발 툴	eclipse
	개발 언어	Java
	형태소 분석기	KLT-2010 [12]
	학습 알고리즘	Weka 3.7

실험은 크게 2가지로 진행되었다. 첫 번째로 본 논문에서 제안한 'Modified ECCD' 자질어 선택 기법을 검증하였으며, 두 번째로 본 논문에서 제안한 '문서별 범주 가중치' 특징 변수의 유의성을 검증하였다.

분류 알고리즘으로 Naive Bayesian Multinominal을 사용하였으며, Training Document 내 자질어 변수의 값은 문서당 자질어의 출현빈도 값을 사용하였다[12][13].

4.2 실험 결과

'자질어 선택 기법' 및 '문서별 범주 가중치'를 사용한 문서 분류 정확도 검증은 10 fold cross validation을 이용해 Precision, Recall 및 F-Measure(F1-Score) 척도로 평가하였다.

4.2.1 자질어 선택 기법 검증 실험

본 논문에서 제안하는 'Modified ECCD' 자질어 선택 기법 검증을 위해 χ^2 및 ECCD 자질어 선택 기법과 비교하였다.

Table 2는 χ^2와 ECCD 자질어 선택 기법을 이용해 각 범주별 자질어 평가값을 구한 후 내림차순으로 정렬한 자질어 순위 결과이다.

Table 3은 본 논문에서 제안된 'Modified ECCD' 자질어 선택 기법으로 추출된 자질어 선택결과의 범주별 분포표이다. 'Modified ECCD'의 $Init\,TotalFeature_{count}$ 값이 100이라면, Table 3에서 표시된 셀과 같이 각 범주별 $SelectedFeature_{count}(c_k)$가 사용된다.

표 2. χ^2와 ECCD feature Selection 별 자질어 순위
Table 2. Ranking by χ^2 and ECCD feature selection

$\chi^2 Max$	$\chi^2 Mean$	$ECCD\ Max$	$ECCD\ Mean$
한나라당	대회	한나라당	경찰
대회	선수	대회	협의
선수	한나라당	선수	재판부
출전	출전	출전	서울
우승	우승	우승	부장검사
...	...	...	...

표 3. Modified ECCD를 이용한 범주별 자질어 순위
Table 3. Feature ranking by each category using Modified ECCD

Modified ECCD								
IT/과학 (2개)	인물 (2개)	사회 (24개)	전국 (9개)	정치 (13개)	문화 (8개)	국제 (10개)	스포츠 (14개)	경제 (15개)
아이폰	부친상	경찰	여수	한나라당	환자	보도	대회	가격
애플	발인	협의	박람회	홍준표	제품	테러	선수	증가
사용자	모친상	재판부	여수세계박람회	최고위원	병원장	베링	출전	분기
앱스토어	전보	서울	주민	민주당	병원	노르웨이	우승	실적
방송통신위원회	장인	부장검사	전남	의원	건강	브레이비크	경기	매출
스마트폰	삼성서울병원	부장판사	서귀포시	청와대	섭취	뉴욕타임스	시즌	직장인
운영체제	서울아산병원	민주노총	시민	최고위원회의	출시	총리	결승	기록
이통사	별세	노조	여수지역	대표	원장	이슬람	금메달	하반기
업계	승진	검찰	해군기	북한	치료	현지시각	감독	한국은행
위치정보	개인사업	산사태	도내	대통령	예방	영국	라운드	상승
...	...	...	...	...	...	...	...	...

'Modified ECCD' 자질어 선택 기법 검증 실험 결과 Fig. 3에서와 같이 제안 방법인 'Modified ECCD'가 평균적으로 좋은 성능을 보였다. 특히 자질어 선택의 수가 100개~5000개 사이일 경우 뚜렷한 성능 차이를 보였다. 그 이유는 웹 뉴스 데이터의 특성상 범주별 문서의 빈도 차이가 많이 나기 때문에 최대/평균 평가함수를 통해 선정된 자질어보다 범주별 문서 빈도를 고려하여 자질어 집합을 생성하는 것이 더 좋은 성능을 보이기 때문이다.

기존 'ECCD' 기법을 이용한 자질어 집합의 생성은 범주별로 $ECCD(t_j, c_k)$ 값을 내림차순하여 정렬하고 범주별로 동일한 개수의 자질어를 선택하므로 범주별 문서의 빈도 차이가 많이 나는 환경에 적합하지 않다. 특히 'ECCD Mean'의 성능이 상대적으로 낮은 이유는 단어 t_j에 대한 $ECCD(t_j, c_k)$의 평균 평가 값이 대부분 유사하기 때문에 올바른 자질어 선택이 어렵기 때문이다.

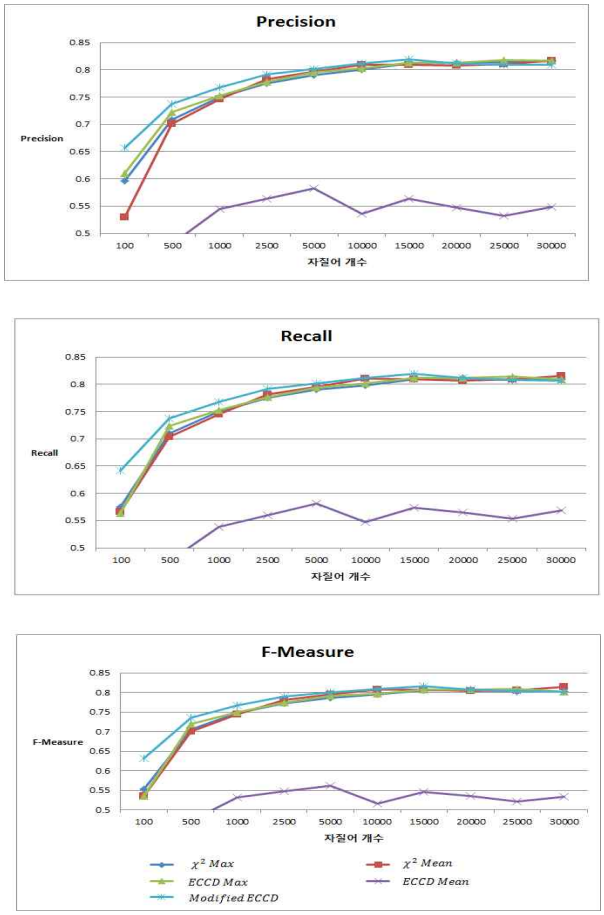

그림 3. Modified ECCD 실험 결과
Fig. 3. Experimental results of Modified ECCD

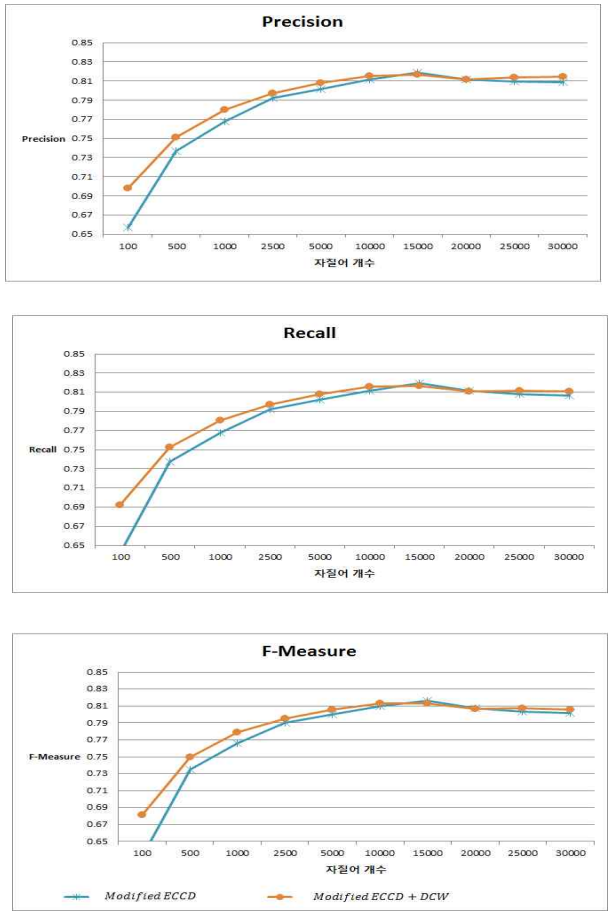

그림 4. 문서별 범주 가중치 검증 실험
Fig. 4. Experiments of category weight for each document

분류 교차표(Confusion Matrix) 확인 결과 문서 비율이 높은 범주인 사회, 전국, 정치, 문화, 국제, 스포츠 및 경제에서 'Modified ECCD' 자질어 선택 기법이 χ^2와 ECCD 자질어 선택 기법에 비해 정확한 분류 성능을 보였으며, 이는 전체 평균 정확도 향상 효과를 가져 올 것으로 해석된다.

4.2.2 문서별 범주 가중치 특징 변수 검증 실험

본 논문에서 제안하는 '문서별 범주 가중치'의 유의성 검증을 위해 첫 번째 실험에서 가장 좋은 결과를 보인 'Modified ECCD' 자질어 선택 기법에 제안한 '문서별 범주 가중치' 특징 변수를 추가하여 비교 평가하였다.

Fig. 4에서와 같이 제안 방법인 Modified ECCD + DCW('범주별 문서 가중치')가 평가 결과 평균적으로 좋은 성능을 보였다. 특히 자질어 선택의 수가 100개~2500개 및 20000~30000개 사이일 경우 뚜렷한 성능 차이를 보였다.

첫 번째로 100개~2500개 사이의 자질어 선택의 경우, 자질어 만으론 문서 분류정확도 확보가 어렵다. 하지만 '문서별 범주 가중치' 특징 변수를 추가해 줌으로써 분류 성능의 향상 효과를 볼 수 있었다. 두 번째로 20000~30000개 사이의 자질어를 선택할 경우 지나치게 많은 자질어 선택으로 인해 분류성능이 저하된다. 하지만 '문서별 범주 가중치' 특징 변수를 추가해 줌으로써 분류 성능 향상 효과를 볼 수 있었다. 세 번째로 5000~15000개 사이의 자질어를 사용할 경우 두 방식의 성능 차이가 거의 없는 이유는 자질어 만으로 획득할 수 있는 최대 분류 성능을 보이기 때문에 '문서별 범주 가중치' 특징 변수가 추가된 효과가 분류 성능에 영향을 미치지 못한다고 볼 수 있다.

5. 결론 및 향후연구

본 논문은 웹 뉴스 도메인에 맞게 개선된 자질어 선택 기법인 'Modified ECCD' 기법을 제안하였다.

첫째, 웹 뉴스 도메인의 범주별 문서 비율 차이를 고려한 'Modified ECCD' 기법은 기존 자질어 선택 기법보다 범주별 학습 문서 비율 차이에 강건하고 문서 분류 정확도가 높은 자질어 집합을 생성하는 것으로 나타났다. 둘째, '문서별 범주 가중치' 특징 변수를 추가함으로써 자질어 변수만 사용하였을 경우 보다 자질어 집합의 개수가 적은 상황에도 우수한 문서 분류 성능을 나타내었고, 지나치게 많은 수의 자질어가 선택될 경우 발생되는 분류기 성능의 저하를 개선하였다. 또한 자질어 후보의 대표 범주가 시간에 따라 변화되는 상황에도 좋은 문서 분류 성능을 나타내었다. 마지막

으로 학습한 문서의 범주별 분포와 실제 예측할 문서의 범주별 분포가 시간에 따라 다를 경우에도 시간이 고려된 '문서별 범주 가중치' 특징 변수를 사용하기 때문에 문서 분류 성능을 확보할 수 있었다.

향후 연구로서, 3.3절에서 언급한 시간에 따른 특정 자질어의 대표 범주가 변하는 문제 해결을 검증하기 위해 좀 더 많은 기간의 데이터가 확보된 실험이 필요하다. 또한, 제안 방법으로 생성된 웹 뉴스 분류 모델을 이용해 SNS 메시지 분류 연구를 진행할 계획이다.

참 고 문 헌

[1] Salton, G. "Automatic processing of foreign language documents." *Journal of the American Society for Information Science*, 21(3), pp.187-194, 1970.

[2] Kil-Hong Joo, Eun-young Shin, Joo_Il Lee, Won-Suk Lee, "Hierarchical Automatic Classification of News Articles based on Association Rules" *Journal of Korea Multimedia Society*, Vol.14, No.6, pp.730-741, June, 2011.

[3] Sanasam Ranbir Singh, Hema A. Murthy, Timothy A. Gonsalves, "Feature Selection for Text Classification Based on Gini Coefficient of Inequality" *JMLR: Workshop and Conference Proceedings*, pp.76-85, 2010.

[4] Christine Largeron, Christophe Moulin, Mathias Gery, "Entropy based feature selection for text categorization" *Proceedings of the 2011 ACM Symposium on Applied Computing*, pp.924-928, 2011.

[5] C. E. Shannon, "A mathematical theory of communication" *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol.5 Issue 1, January, 2001.

[6] Haichao Dong, Siu Cheung Hui, Yulan He*, "Structural Analysis of Chat Messages for Topic Detection" *Online Information Review*, pp.496-516, 2006.

[7] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design" *Journal Neural Computation*, Vol.13 Issue 3, March, 2001.

[8] Steven L. "C4.5: Programs for Machine Learning" *Book Review, Machine Learning*, 16, pp.235-240, 1994.

[9] P.Winstron, http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html, 1992.

[10] Hang Li, Kenji Yamanishi, "Document classification using a finite mixture model", *EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997.

[11] McCallum, A., Nigam, K., "A comparison of event models for Naive Bayes text classification.", *AAAI-98 Workshop on Learning for Text Categorization*, pp.41-48, 1998.

[12] KLT2010, http://nlp.kookmin.ac.kr/

[13] Dan-Ho Park, Won-Sik Choi, Hong-Jo Kim, Seok-Lyong Lee, "Web Document Classification System Using the Text Analysis and Decision Tree Model", *Proceedings of The Korean Institute of Information Scientists and Engineers 2011 fall*, Vol.38, No.2(A), pp.248-251, 2011.

한 정 석

e-mail : jshan97@mining.ssu.ac.kr
2004년 관동대학교 컴퓨터공학과(학사)
2007년 숭실대학교 일반 컴퓨터학과(석사)
2007년~현 재 숭실대학교 컴퓨터학과 박사과정
관심분야 : 데이터마이닝, 인공지능, 기계학습 등

박 상 용

e-mail : silverwing86@naver.com
2010년 남서울대학교 컴퓨터학과(학사)
2011년~현 재 숭실대학교 컴퓨터학과 박사과정
관심분야 : 데이터마이닝, 인공지능, 기계학습 등

이 수 원

e-mail : swlee@ssu.ac.kr
1982년 서울대학교 자연과학대학 계산통계학과(학사)
1984년 한국과학기술원 전산학과(석사)
1994년 University of Southern California 전산학과(박사)
1995년~현 재 숭실대학교 컴퓨터학부 교수
2008년~현 재 한국BI데이터마이닝학회 부회장
2008년~2009년 한국정보과학회논문지(SA) 편집위원장
관심분야 : Data Science, 기계학습, 인공지능 등