http://dx.doi.org/10.3745/KIPSTD.2012.19D.4.263

한글 형태소 및 키워드 분석에 기반한 웹 문서 분류

박 단 호[†] · 최 원 식[††] · 김 홍 조[††] · 이 석 룡[†††]

요 약

최근 초고속 인터넷과 대용량 데이터베이스 기술의 발전으로 웹 문서의 양이 크게 증가하였으며, 이를 효과적으로 관리하기 위하여 문서의 주제별 자동 분류가 중요한 문제로 대두되고 있다. 본 연구에서는 한글 형태소 및 키워드 분석에 기초한 문서 특성 추출 방법을 제안하고, 이를 이용하여 웹 문서와 같은 비구조적 문서의 주제를 예측하여 문서를 자동으로 분류하는 방법을 제시한다. 먼저, 문서 특성 추출을 위하여 한글 형태소 분석기를 사용하여 용어를 선별하고, 각 용어의 빈도와 주제 분별력을 기초로 주제 분별 용어인 키워드 집합을 생성한 후, 각 키워드에 대하여 주제 분별력에 따라 점수화한다. 다음으로, 추출된 문서 특성을 기초로 상용 소프트웨어를 사용하여 의사 결정 트리, 신경망 및 SVM의 세 가지 분류 모델을 생성하였다. 실험 결과, 제안한 특성 추출 방법을 이용한 문서 분류는 의사 결정 트리 모델의 경우 평균 Precision 0.90 및 Recall 0.84 로 상당한 정도의 분류 성능을 보여 주었다.

키워드 : 문서 특성 추출, 문서 분류, 형태소 분석, 키워드 빈도 분석

Web Document Classification Based on Hangeul Morpheme and Keyword Analyses

Dan-Ho Park[†] · Won-Sik Choi[††] · Hong-Jo Kim[††] · Seok-Lyong Lee[†††]

ABSTRACT

With the current development of high speed Internet and massive database technology, the amount of web documents increases rapidly, and thus, classifying those documents automatically is getting important. In this study, we propose an effective method to extract document features based on Hangeul morpheme and keyword analyses, and to classify non-structured documents automatically by predicting subjects of those documents. To extract document features, first, we select terms using a morpheme analyzer, form the keyword set based on term frequency and subject-discriminating power, and perform the scoring for each keyword using the discriminating power. Then, we generate the classification model by utilizing the commercial software that implements the decision tree, neural network, and SVM(support vector machine). Experimental results show that the proposed feature extraction method has achieved considerable performance, i.e., average precision 0.90 and recall 0.84 in case of the decision tree, in classifying the web documents by subjects.

Keywords : Document Feature Extraction, Document Classification, Morpheme Analysis, Keyword Frequency Analysis

1. 서 론

초고속 인터넷과 고성능 컴퓨팅 시스템 및 대용량 데이터베이스와 같은 정보기술의 발전으로 웹 문서가 양산되고 있으며, 이에 따라 증가하는 웹 문서를 효율적으로 관리하고 검색할 수 있는 방법이 중요한 문제가 되고 있다. 수많은 웹 문서를 효율적으로 관리하고 검색하기 위해서 일반적으로 문서의 주제에 따른 분류 방식을 사용하며 문서의 양이 크게 증가함에 따라 문서 분류를 위한 노력 역시 증가하게 되었다. 특히 인터넷 신문, 잡지 등의 웹 문서와 같은 비구조적(non-structured) 문서의 경우, 문서의 주제를 판단하기 위해 상세한 태그가 있는 구조적 문서에 비해서 훨씬 많은 노력이 요구되기 때문에 효율적인 문서 관리를 위해서는 주제 판단을 자동 처리할 수 있는 방법이 필요하다.

문서를 분류하는 방법에는 문서 클러스터링과 문서 범주화가 있다[1]. 본 논문에서는 주제 분류에 대한 사전 정의가 이루어진 경우, 새로운 웹 문서를 어떠한 주제로 분류할 것인가에 대한 판단에 기초하여 문서를 자동으로 분류하는 문서 범주화 방식을 다룬다. 본 논문에서 사용하는 용어 중 '

※ 이 연구는 2012학년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임.
† 준 회 원 : 한국외국어대학교 산업경영공학과 석사과정
†† 준 회 원 : 한국외국어대학교 산업경영공학과 학사
††† 정 회 원 : 한국외국어대학교 산업경영공학과 교수
논문접수 : 2012년 1월 9일
수 정 일 : 1차 2012년 5월 30일
심사완료 : 2012년 6월 20일
* Corresponding Author : Seok-Lyong Lee(sllee@hufs.ac.kr)

주제'는 문서를 분류하기 위하여 사전에 정의된 문서의 클래스 (예를 들면, 경제, 사회, 스포츠, 정치 등)를 말하며, '문서 범주화'는 주어진 문서에 사전에 정의된 주제를 할당하는 과정이다. 또한, '키워드'는 주제를 분별하기 위하여 문서로부터 추출된 주제 분별 용어라 정의한다.

문서 범주화는 특성 추출(feature extraction)과 분류 모델 생성의 두 가지 핵심 요소로 이루어지며, 이들 과정에서 사용되는 방법에 따라 시스템의 성능이 크게 달라진다[2]. 일반적으로 특성은 문헌 내의 용어(term) 또는 색인어(index term)로 표현되며[3], 문서의 주제 판단에 유용하게 사용될 수 있는 용어를 나타낸다. 이러한 특성 추출 과정에는 TF-IDF (term frequency-inverse document frequency), 상호 정보 (mutual information), 카이 제곱 통계량($\chi 2$), 정보 획득량 (information gain) 등의 방법이 사용된다[4]. 특성 추출 과정을 통해 파악된 특성들을 기반으로 문서 주제 판단을 위한 분류 모델을 구성하게 되는데, 주로 나이브 베이즈(naïve Bayes), SVM, K-NN (K-nearest neighbor), 신경망 등의 방법이 사용된다.

본 논문은 문서 특성 추출 방법의 개발에 초점을 두었으며, 분류 모델 생성은 기존의 상용 소프트웨어를 사용하였다. 기존의 문서 특성 추출 방법들은 주로 문서에서 사용된 색인어 또는 표제어 등의 중요 용어에 대하여 각각의 빈도를 파악하고, 중요성에 따라 가중치를 부여하는 방식을 많이 사용한다. 이때 용어의 빈도는 각각의 문서에서 몇 번 출현하였는가를 기준으로 계산된다. 그러나 보다 정확한 분류를 위해서는 각 문서에 출현하는 용어의 빈도 뿐 아니라, 각 용어가 해당 주제의 문서들에서 나타나는 누적 빈도도 같이 고려해야 한다.

본 논문에서는 먼저, 문서 특성 추출을 위하여 한글 형태소 분석기를 사용하여 문서로부터 용어를 선별하고, 각 용어의 빈도와 주제 분별력을 기초로 주제 분별 용어인 키워드 집합을 생성한다. 다음으로, 각 키워드에 대하여 해당 주제의 문서들에서 나타나는 누적 빈도를 고려한 주제 분별력을 기초로 이를 점수화하여 문서 특성을 추출한다. 마지막으로, 추출된 문서 특성을 기초로 상용 소프트웨어를 사용하여 정확한 분류 방법으로 널리 알려진 의사 결정 트리, 신경망 및 SVM의 세 가지 분류 모델을 생성하여 문서를 분류하였다.

본 논문의 구성은 다음과 같다. 2장에서는 문서 특성 추출 및 분류 모델 생성에 대한 관련 연구들을 소개한다. 3장에서는 문서 특성 추출 과정 및 웹 문서에서 주제 분별 용어 집합을 구성하고 점수화하는 과정을 제시하고, 주제 분별 용어의 분별력과 빈도를 기반으로 학습 모델 생성을 위한 과정을 기술한다. 4장에서는 시스템 구현에 관하여 설명하며, 5장에서는 실험 환경과 자료 구성을 소개하고, 실험 결과에 대해 분석한다. 6장에서는 결론 및 향후 연구 방향에 대하여 언급한다.

2. 관련 연구

폭발적으로 증가하는 문서의 양으로 인해 효율적이며 정확한 문서 범주화 결과에 대한 수요가 급증하는 추세이다[5]. 정보의 양적 증가와 이에 따른 분류 작업의 양적 증가는 기존의 수작업으로 관리하기 어려운 수준에 이르렀으며, 분류 대상이 되는 정보의 유형도 기존의 책자 형태의 정보와 함께 전자 형태의 정보까지 확대되었다[6]. 이에 따라 문서를 자동으로 분류하는 과정은 문서의 효율적 관리 측면에서 매우 중요하다. 본 논문은 문서 범주화의 핵심 요소 중 특성 추출 방법에 초점을 두고 있으므로 이에 관한 국내외 관련 연구를 주로 살펴보고, 분류 모델 생성에 관해서는 간단히 언급하기로 한다.

특성 추출 관련 연구는 용어에 기초하는 방법(통계적 방법)과 문장에 의존하는 방법(의미적 해석), 문헌 구조에 의존하는 방법(구조적 방법)으로 구별된다[7]. 이 가운데 문헌에 나타난 용어의 특성 분석을 바탕으로 하는 자동 분류 시스템의 연구가 활발하게 이루어지고 있으며, 한글과 영어는 형태론적 구조가 다르므로 형태소 추출 과정은 다르지만 추출된 용어의 빈도 등의 특성 분석은 본질적으로 유사하다. 용어의 특성 분석을 바탕으로 하고 있는 국내 연구로서, 성기윤, 윤보현의 연구[1]는 정교한 언어 분석을 통한 개체명 인식 결과를 바탕으로 인명, 지명, 회사명, 물품명 등 개념어 기반의 기법을 제안하였다. 그러나 개체명에 키워드를 추가하여 수행하였기 때문에 개체명만의 클러스터링의 우수함을 보이지는 못하였다. 남영준, 김규환의 연구[6]는 추출한 특성을 벡터로 표시하여 이를 SVM에 적용함으로서 문서 범주화를 수행하였고, 최종적으로 유사어 사전을 부여하기 전과 부여한 후의 자동 범주화 실험 수행 결과를 비교 분석하였다. 그러나 일차원적인 범주 모델을 확대하여 계층적인 자동 범주 및 자동 분류 모델의 개발이 필요하다는 과제를 남기고 있다. 배원식 외 2인의 연구[2]는 문서 내에서 동시에 출현하는 단어 쌍을 특성 추출 단위로 하는 문서 범주화 시스템에 대하여 연구하였다. 실험 결과, 문서 범주화 시스템의 성능이 향상되는 것을 보여주었으나 동시 출현 단어 쌍을 단위로 특성 후보를 생성하면 단일 단어에 비해 훨씬 많은 특성 후보가 생성되어 계산 비용이 증가하는 문제점이 있다.

용어의 특성 분석을 바탕으로 하고 있는 국외 연구로서 Ludovic et al.[8]은 문서에 나오는 용어들을 적합 용어 집합과 부적합 용어 집합으로 나누고, 적합 단어 집합의 용어들에 대하여 가중치 접근 방법을 사용하여 문서 범주화를 시도하였다. Chen et al. [9]은 관련 있는 용어들의 군집화 집단으로 문서를 구분하고, 문서의 특성을 이용하여 각 군집을 분류하는 방법을 사용하여 문서를 범주화하였다. 또한, Thanaruk [10]은 웹에서 추출한 웹 문서를 대상으로 텍스트 마이닝(text mining)을 수행하기 위하여 문서를 윈도우 방식의 문단으로 구분하였고, 문단 정보와 각 문단에서 용어의 상호 출현 빈도의 두 가지 정보를 이용하여 문서를 범주화하였다.

한편, 분류 모델 생성 관련 연구는 다양한 기법에 기초를 두고 있다. 널리 알려진 국외 연구로서, Carnegie Mellon 대학에서는 Naive-Bayes 이론을 이용하여 분류 모델을 생성하였고 [11], 사례 기반 추론이나 인공 신경망, 의사 결정 트리, 그리고 SVM 등을 이용한 분류 모델 생성 방법들이 널리 이용되고 있다[12].

본 논문에서는 주제 분별력이라는 개념을 도입하여 키워드 집합을 생성하고, 각 키워드에 대하여 누적 빈도를 고려하여 점수화함으로서 문서 특성을 표현하였다. 또한, 용어 선정에 있어서는 개체명과 키워드를 구분 짓지 않음으로서 기존 연구의 단점을 극복하였고, 분류 성능을 높였다.

3. 문서 특성 추출 및 분류 모델

본 연구에서 제안하는 문서 특성 추출 및 분류 모델 생성은 그림 1에 도시되어 있는 과정을 거쳐서 수행된다. 문서를 효과적으로 분류하기 위해서 문서 집합을 크게 학습용 집합(training set)과 검증용 집합(testing set)의 두 종류로 나눈다. 학습용 집합은 다시 분류 모델의 생성을 위한 모델 생성용과 모델 평가를 위한 모델 평가용 집합으로 나누어지고, 검증용 집합은 분류 모델의 성능을 검증하기 위하여 사용한다.

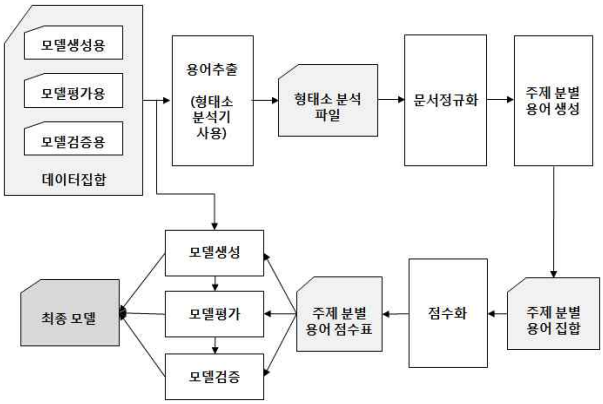

그림 1. 문서 특성 추출 및 분류 모델 생성 과정
Fig. 1. Document feature extraction and classification model generation

본 연구에서 제안한 문서 특성 추출은 먼저, 한글 형태소 분석기를 사용하여 각각의 문서 집합에서 사용되는 용어를 추출하고 각 용어의 빈도를 파악한 후, 각각의 문서의 크기를 고려하여 용어의 빈도를 정규화한다. 다음으로 추출한 용어 중 주제를 정확하게 분류할 수 없는 용어를 제외함으로써 주제 분별 용어 집합을 구성한다. 그리고 모델 생성용 집합에서 나타나는 주제 분별 용어의 빈도와 분별력을 기초로 각 주제 분별 용어에 대하여 점수화하여 주제 분별 용어 점수표(score table)를 생성한다.

분류 모델 생성은 주제 분별 용어 점수표에 기초하여 높은 성능을 보이는 것으로 알려진 기존의 의사 결정 트리 및 신경망 기법, 그리고 SVM 을 구현한 상용 소프트웨어 SPSS Clementine 14.1 을 사용하였다. 각각의 방법에 의해 생성된 모델에 대하여 평가 및 검증 수행하고, 이를 통해 우수한 성능을 보이는 모델을 선택하여 최종 모델을 확정한다.

3.1 문서 특성 추출

3.1.1 용어 추출 및 빈도 파악

문서 범주화를 위해 사용되는 특성은 일반적으로 문서의 표제어 또는 색인어, 본문, 참고문헌 제목 등에서 사용되는 용어를 기반으로 한다. 한글은 영어와 달리 교착어이므로 어근에 여러 첨가가 발생하기 때문에 이러한 용어 기반의 특성을 정확히 추출하기 위해서는 조사나 어미의 구분이 필요하다. 이러한 한글의 용어 분석은 상당한 노력을 요구하게 되므로 이를 기계적으로 처리할 수 있는 방법에 대한 연구가 많이 진행되고 있다. 본 연구에서는 문서 내의 용어를 추출하기 위해서 한국어 형태소 분석기[13]를 이용하였다. 이를 통해서 각 문서 내에 포함되어 있는 용어와 각각의 빈도를 파악할 수 있다. 그림 2는 한국어 형태소 분석기를 이용하여 얻어지는 용어와 빈도가 기록된 텍스트 파일의 예이다.

그림 2. 한국어 형태소 분석기를 이용한 용어와 빈도 분석 결과
Fig. 2. Keyword and its frequency analysis using Hangeul morpheme analyzer

3.1.2 용어 빈도 정규화

하나의 문서 내에 포함된 용어의 빈도를 바탕으로 서로 다른 주제에 속하는 문서들 사이에 용어 빈도의 차이를 비교하려면 먼저 이들 문서의 용어 빈도가 상호 비교할 수 있는 수준이어야 한다. 즉, 문서의 크기가 크면 그 문서에서 나타나는 용어들의 절대 빈도도 증가하므로 용어의 빈도를 문서 크기에 따라 정규화 해야 한다. 이를 위해 각 문서 별로 출현 용어의 총 누적 빈도를 분석하고, 다음으로 각 해당 용어의 빈도를 구해 문서 크기에 따라 정규화된 문서별 용어 빈도 TN_{ijk}를 각각의 용어에 대하여 도출한다. 즉, TN_{ijk}는 주제 i에 속하는 문서 k에서 나타나는 용어 j의 문서 크기에 따른 용어 빈도 정규화 값을 의미하며, 이를 수식으로 표현하면 다음 식(1)과 같다.

$$TN_{ijk} = \frac{\text{문서}k\text{에서 해당 용어}j\text{의 빈도}}{\text{주제}i\text{속 해당문서}k\text{의 총 용어 빈도 합}} \quad (1)$$

3.1.3 주제 분별 용어 선정

문헌 내에는 수많은 용어가 사용되기 때문에 모든 용어를

그대로 이용할 경우 계산 및 메모리 공간의 효율적인 측면에서 문제점이 발생하며 문서 범주화 성능에도 역시 좋지 않은 결과를 초래한다. 따라서 적절한 수준으로 특성을 축소(reduction)할 필요가 있다[3]. 여기서는 문서의 주제 분류에 도움이 될 수 있는 특성의 기준을 수립하여 축소된 특성을 주제 분별 용어, 혹은 키워드라 부르며, 각 용어는 주제 분별력에 대한 수치 값을 갖는다.

주제 분별 용어는 특정 용어가 특정 주제에서 자주 나타날 경우 해당 용어의 빈도 정보를 이용하여 문서의 주제를 판단할 수 있을 것이라는 아이디어에서 출발한다. 따라서 각각의 용어가 개별적인 문서에서 나타나는 빈도가 아닌 각 주제에서 나타나는 모든 용어의 누적 빈도를 분석해야 한다. 주제 i에서 나타나는 용어 j의 정규화된 빈도의 총 누적치 CTN_{ij}는 다음 수식(2)로 표현된다.

$$CTN_{ij} = \sum_{j=1}^{n} TN_{ij} \tag{2}$$

위 식에서 TN_{ij}는 주제 i에서 나타나는 용어 j의 정규화된 빈도를 나타낸다. 이때 용어의 총 개수를 n이라 하면, 주제 i에서 나타나는 용어 j의 총 누적 정규화 빈도 CTN_{ij}는 TN_{i1}부터 TN_{in}의 총 합이 된다. 문서 집합은 사전에 정의된 주제 분류 체계에 대한 정보를 갖고 있다. 각 주제에 속하는 문서의 합이 같더라도 주제마다 용어의 종류와 크기가 다르기 때문에 각각의 주제 내 문서들에 대하여 해당 용어의 총 누적 정규화 빈도(CTN_{ij})를 고려하여 정규화 값을 얻는다. 즉, 같은 주제 상의 모든 문서에 출현하는 각 단어들의 정규화된 빈도 값을 가지고 평균과 표준편차를 이용하여 정규화된 값 $ZCTN_{ij}$을 구할 수 있다.

$$ZCTN_{ij} = \frac{CTN_{ij} - Mean(CTN_i)}{Stdv(CTN_i)} \tag{3}$$

위 수식(3)에서 정규화한 수치 $ZCTN_{ij}$는 주제 i에서 나타나는 용어 j의 총 누적 정규화 빈도 CTN_{ij}에서 주제 i에서 나타나는 용어들의 총 누적 용어 정규화 빈도의 평균 $Mean(CTN_i)$을 뺀 후, 주제 i에서 나타나는 용어들의 총 누적 용어 정규화 빈도의 표준편차 $Stdv(CTN_i)$)로 나눈 값으로 표현된다. 각 주제에 속한 용어들 가운데 주제별로 유사한 빈도를 가지는 경우에는 특정 주제에 대한 분별력이 떨어진다고 할 수 있다. 따라서 이러한 용어들은 제거의 대상이 되고, 차원을 축소(dimension reduction)할 수 있는 기회를 제공한다. 수식(3)을 통해서 각 용어들은 모든 주제에 대하여 정규화한 수치($ZCTN_{ij}$)를 가지게 된다. 따라서 각 용어의 $ZCTN_{ij}$를 이용하여 표준편차를 구할 수 있는데 이는 주제 분별력을 나타낸다. 주제 분별력 V_j는 다음 수식(4)로 표현될 수 있다.

$$V_j = \sqrt{\frac{\sum_{i=1}^{p} (ZCTN_{ij} - Mean(ZCTN_{ij}))^2}{p}} \tag{4}$$

식(4)를 통해 각 용어들의 주제 분별력을 산출하면, 각 용어들에 대해서 주제별로 점수화할 수 있다. 이렇게 주제별로 나타낸 주제 분별력 점수에 대한 평균 $mSTD$을 식(5)를 통해 구하고, $mSTD$에 임의의 배수를 곱하여 산출한 값을 기준 (축소 기준 C)으로 하여 주제 분별에 효과적인 용어만을 추출하게 된다. 식(5)에서 $stdv(ZCTN_{ij})$는 각 용어의 주제 분별력 점수이고, $ZCTN_{ij}$의 표준편차를 의미하며, $Mean$은 $stdv(ZCTN_{ij})$의 평균을 의미한다. 축소 기준 C를 수식으로 표현하면 식(6)과 같다. 즉, $mSTD$에 임의의 배수 x 를 곱한 값이 각 용어의 주제 분별력 V_j보다 크면 그 용어는 제거됨을 의미한다. 그렇게 함으로써 주제를 정확하게 분류할 수 없는 용어를 제거하여 주제 분별 용어만을 추출하고, 동시에 차원을 축소하여 효율적인 처리가 가능하게 할 수 있다.

$$mSTD = mean(stdev(ZCTN_{ij})) \tag{5}$$

$$\text{축소기준 } C = mSTD \times x \tag{6}$$

3.1.4 점수화

앞서 생성한 주제 분별 용어 집합을 이용하여 문서별로 주제 분별 용어 점수표를 생성한다. 각 문서 a의 주제 분별 용어 점수를 TS_a라 할 때, 이 점수는 다음 식으로 표현된다.

$$TS_a = DF_{ak} \times V_k \tag{7}$$

위의 식(7)에서 TS_a는 문서 a에서 추출된 주제 분별 용어 k가 나타나는 빈도인 DF_{ak}와 주제 분별 용어 k가 나타내는 주제 분별력 (V_j에서 축소 기준에 따라 제거 후 남은 용어의 주제 분별력의 의미)인 V_k의 곱으로 표현된다. 점수표의 각 행은 하나의 문서가 되고 각 열은 주제 분별 용어의 점수로 표현된다.

3.2 분류 모델의 생성과 평가 및 검증

주제 분별 용어 점수표가 만들어지면 이를 이용하여 주제 분류 모델을 생성한다. 먼저, 학습용 집합을 이용하여 분류 모델을 생성하고 평가한다. 학습용 집합을 이용한 분류 모델의 생성 및 평가 절차는 다음과 같다. 먼저, 3.4 절에서 구한 각 문서 별 주제 분별 용어 점수를 입력 변수로 하고 주제 필드를 출력 변수로 설정한 후, 상용 소프트웨어 SPSS Clementine 에 포함된 의사 결정 트리와 신경망 모델 및 SVM 방법을 이용하여 각각의 분류 모델을 생성한다. 의사 결정 트리의 경우 C5.0 알고리즘을 사용하였고, 신경망 모델은 다층 퍼셉트론 (multi-layer perceptron, MLP) 기법을 적

용하였다. SVM 방법에서는 커널 함수를 선택하는 것이 성능에 결정적인 영향을 미치며, 본 연구에서는 문서 분류에서 높은 성능을 보이는 RBF (radial basis function) 커널을 사용하여 분류 모델을 생성하였다. 다음으로, 생성된 분류 모델에 대하여 학습용 집합 중 모델 생성에 이용되지 않은 나머지 데이터를 사용하여 모델을 평가하며, 평가 결과가 최적이 되도록 반복하여 입력 파라미터들을 튜닝 (tuning)하는 과정을 거친다.

분류 모델의 검증의 경우에도 모델 생성의 경우와 마찬가지로 검증용 집합의 데이터에 대하여 용어 추출 및 용어 빈도 정규화 등의 과정을 거쳐 주제 분별 용어를 추출하고 점수표를 작성한 후, 의사 결정 트리와 신경망 모델 및 SVM 방법을 이용하여 생성된 각 분류 모델의 정확도를 검증한다.

4. 시스템 구현

4.1 주제 분별 용어 분석기의 구현

한국어 형태소 분석기[13]를 이용하여 그림 2와 같은 텍스트 결과물을 얻은 후, 이러한 결과 파일을 이용하여 수식(2)와 수식(3)의 값을 구하고, 그 값을 통해 주제 분별 용어 집합과 주제 분별 용어 점수표를 생성하는 과정을 수행하는 주제 분별 용어 분석기를 개발하였다. 이 분석기는 자바 언어로 개발되었으며, 표 1과 같이 4개의 클래스로 이루어져 있다.

표 1. 주제 분별 용어 분석기의 구현
Table 1. Implementation of keyword frequency analyzer

순번	클래스명	역할
1	ConvertTerm Freq	형태소 분석기를 이용하여 얻는 결과 파일에서 필요한 정보만 요약
2	TextNorm	문서 크기에 따른 정규화를 실행하여 정규화된 용어 빈도 계산
3	SeekKeyword	주제 분별 용어 집합 생성
4	ScoringKeywords	문서별 주제 분별 용어 점수표 생성

각각의 클래스는 1에서 4의 순서로 순차적으로 호출된다. 첫 번째 ConverTermFreq 클래스를 이용하여 형태소 분석기 결과 파일에서 용어와 빈도, 문서명, 주제 정보로 이루어진 데이터를 생성한다. 두 번째 TextNorm 클래스는 ConvertTermFreq 클래스의 결과 파일을 입력 데이터로 사용하여 각각의 문서의 크기에 따른 용어 빈도를 정규화하여 기존 빈도를 정규화된 값으로 교체한다. 이를 위해 수식(1)을 이용한다. 이 두 개의 클래스는 학습용 데이터와 검증용 데이터 모두에 적용된다. 세 번째 SeekKeyword 클래스는 학습용 데이터에만 적용되며 TextNorm 클래스의 결과 파일을 입력 데이터로 사용하여 주제 분별 용어를 추출하고, 각각의 주제 분별 용어별 주제 분별력을 계산한다. 해당 클래스에서는 수식(2)를 이용하여 주제별로 정규화된 용어의 누적 빈도를 계산하고, 각각의 용어와 주제별로 존재하는 수식(2)의 값을 통해 수식(3)의 값을 산출한다. 수식(3)의 값은 다시 용어 별로 수식(4)에 적용되어 주제 분별력을 구하는데 이용된다. 이렇게 구해진 주제 분별력은 수식(5)와 (6)을 통해 임의의 기준을 만족하는지 확인하기 위해 이용되며, 여기에 적합한 용어만을 선택하여 주제 분별 용어 집합 파일을 생성한다.

마지막으로 ScoringKeywords 클래스에서는 학습용 데이터와 검증용 데이터를 이용하여 생성된 TextNorm 클래스의 결과 파일과 주제 분별 용어 집합 파일을 통해 문서의 주제 분별 용어별 점수표를 생성한다. 점수표의 행은 하나의 문서를 의미하며 각각의 열은 해당 주제 분별 용어의 주제별 점수를 나타낸다. 수식(7)을 이용하여 문서 별 주제 분별 용어의 점수를 계산하여 점수표를 생성하며, 이 점수표는 모델 생성을 위한 입력 데이터 및 검증용 데이터로 이용된다.

4.2 분류 모델의 생성

4.2.1 자료 집합의 구성

모델의 생성, 평가 및 검증을 위한 자료 집합을 구성하기 위해, 2009년 10월 1일부터 2011년 7월 11일 사이의 동아, 조선, 경향 신문의 기사 중 720개의 웹 문서를 임의로 선택하였다. 720개의 웹 문서는 경제, 사회, 스포츠, 정치의 4가지 주제별 기사를 포함하며, 각각의 주제는 180개의 기사를 포함한다. 이 중 400 개는 분류 모델을 생성하고 평가하기 위한 학습용 집합으로 사용하고, 나머지 320개는 분류 모델의 성능 검증을 위한 검증용 집합으로 사용하였다. 학습용 집합 중 짝수 번째 웹 문서 200 개를 이용하여 모델을 생성하고, 모델 생성에 사용되지 않은 홀수 번째 문서 200 개를 이용하여 모델을 평가하였다. 일반적으로, 720개의 문서는 분류 모델의 생성 및 검증에 충분한 양은 아니지만, 기존에 이미 정제되어 제공되고 있는 대량의 웹 문서를 사용하는 대신, 직접 구성한 최근의 자료 집합을 사용하였다. 이는 웹 문서에 사용되는 용어에 최근 많은 변화가 발생하고 있어 기존에 제공되고 있는 웹 문서를 사용할 경우, 최근 정보에 대한 분류 정확도가 저하될 우려가 있고, 또한 제안한 시스템이 경제, 사회, 스포츠, 정치 등의 소수의 범주에 대하여 분류를 수행하도록 개발되었으므로 성능 검증에 대한 신뢰성은 상당 부분 확보할 수 있기 때문이다. 향후 보다 다량의 문서를 확보하여, 보다 다양한 범주로 문서를 분류하는 방법을 연구하는 것이 필요하다.

4.2.2 분류 모델의 생성

SPSS Clementine의 의사 결정 트리 (c5.0)와 신경망 기법 및 SVM 을 통하여 분류 모델을 생성한다. 생성된 분류 모델을 평가용 집합으로 평가하여 결과가 수준 이하일 경우 데이터의 전처리 과정의 파라미터 값들을 튜닝(tuning)한다. 해당 파라미터는 문서 빈도 정규화 알고리즘에 사용되는 입력 값들과 주제 분별용 용어 선정의 기준 값 등이다. 파라미터 튜닝은 일정한 값에 수렴할 때까지 반복적으로 수행되며, 최종 평가 결과, 분류 정확도가 임계 값 이하로 나타날 경우에는 해당 분류 방법은 최종 모델에서 배제한다. 본 연

구에서는 임계 정확도를 70%로 설정하였다. 최종 평가 결과, 의사 결정 트리 및 신경망 모델이 각각 200개의 웹 문서 중 168개를 정확하게 분류하여 84%의 정확도를 보였으며. SVM의 경우 79%의 정확도를 보여 주었다. 세 가지 방법 모두 분류 정확도가 임계 값 이상이므로 모두 채택된다. 생성된 분류 모델의 분석을 위하여 널리 사용되는 이득 도표(gain chart)를 사용하여 의사 결정 트리와 신경망 모델 및 SVM 방법을 비교하였다. 이득 도표는 해당 등급에 따라 계산 된 이득(gain) 값을 연속적으로 연결한 도표로서, 그림 3에서 왼쪽 아래에서 오른쪽 위로 이어지는 대각선은 모델 비교의 기준선으로서, 모델의 성능이 나쁠수록 이 기준선에 가까워지는 특성이 있다. 도표에서 관찰할 수 있는 바와 같이 의사 결정 트리, 신경망 모델, 그리고 SVM 순으로 성능이 좋은 것으로 나타났다.

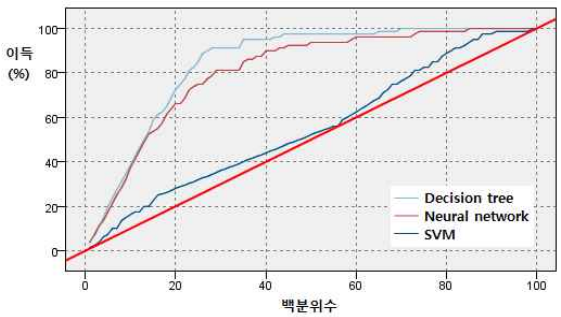

그림 3. 이득 도표를 통한 세 가지 분류 모델의 비교
Fig. 3. Comparison of three classification models through Gain chart

5. 실험 및 결과 분석

5.1 제안한 방법의 성능 검증

분류 모델의 성능은 320개의 웹 문서를 포함하고 있는 검증용 집합을 사용하여 검증하였으며, 다음 표 2~4는 각각 경제, 사회, 스포츠, 정치의 4가지 주제에 대하여 의사 결정 트리와 신경망 분석 및 SVM을 통한 분류 모델의 성능 검증 결과를 나타낸 것이다. 검증 결과 표 2~4에서 볼 수 있는 바와 같이 의사 결정 트리의 분류 결과가 84.06%의 정확도를 보여 신경망 분류의 정확도(77.19%)와 SVM 분류의 정확도(69.06%)보다 높은 것을 알 수 있다. 의사 결정 트리의 분류 결과는 모델 평가 시 나타났던 84%와 근소한 차이를 보이고 있는 반면, 신경망 분류와 SVM 분류의 경우 정확도가 모델의 평가시보다 저하됨을 알 수 있다.

표 2. 의사결정트리를 사용한 분류 결과 교차표
Table 2. Crosstab of classification results using decision tree

실제 \ 분류	경제	사회	스포츠	정치	합계
경제	72	5	0	3	80
사회	11	53	0	16	80
스포츠	1	2	76	1	80
정치	1	9	2	68	80
합계	85	69	78	88	320

표 3. 신경망 모델을 사용한 분류 결과 교차표
Table 3. Crosstab of classification results using neural network model

실제 \ 분류	경제	사회	스포츠	정치	합계
경제	63	7	1	9	80
사회	16	42	4	18	80
스포츠	4	1	73	2	80
정치	5	5	1	69	80
합계	88	55	79	98	320

표 4. SVM을 사용한 분류 결과 교차표
Table 4. Crosstab of classification results using SVM

실제 \ 분류	경제	사회	스포츠	정치	합계
경제	67	11	0	2	80
사회	15	46	1	18	80
스포츠	11	2	65	2	80
정치	34	3	0	43	80
합계	127	62	66	65	320

웹 문서 분류 결과를 좀 더 구체적으로 분석하기 위하여 문서 검색 분야에서 널리 사용되고 있는 측정 지표인 Precision과 Recall의 지표를 이용하여 분류 결과를 제시한다. Precision은 분류 모델이 주제 범주에 따라 분류한 문서 중 정확하게 분류해 낸 문서의 비율을 나타내며, Recall은 전문가가 주제 범주에 맞게 분류해 놓은 문서 중 분류 모델이 정확하게 분류해 낸 문서의 비율을 의미한다. RET가 모델이 주제 범주에 따라 분류한(retrieved) 문서 집합이고, REL이 전문가가 주제 범주에 맞게 분류한(relevant) 문서 집합이라 하면, Precision과 Recall은 각각 다음 식으로 정의된다.

$$\mathrm{Precision} = \frac{\#(RET \cap REL)}{\#RET} \tag{8}$$

$$Recall = \frac{\#(RET \cap REL)}{\#REL} \tag{9}$$

표 2의 의사 결정 트리 모델의 교차표에서 경제의 경우 $\#REL$=80, $\#RET$=85 이다. 이는 분류 모델이 문서를 '경제'라는 주제 범주로 분류한 85개의 웹 문서 중에서 실제 전문가가 경제라는 주제 범주에 맞게 분류한 문서가 72개라는 의미이다. 또한 전문가가 '경제'라는 주제 범주로 분류해 놓은 80개의 웹 문서 중에서 모델이 맞게 분류한 웹 문서가 72개이다. 따라서 의사 결정 트리 모델의 '경제' 주제에 대한 경우, Precision=72/85(0.85), Recall=72/80(0.90)이다.

표 5-7은 각각 표 2~4의 의사 결정 트리와 신경망 모델 및 SVM 모델의 교차표에 근거하여 주제별로 Precision 과 Recall 을 계산한 결과이다. 표 에서 관찰할 수 있는 바와 같이 신경망 모델과 SVM 모델에 비하여 의사 결정 트리 모델이 경제, 사회, 스포츠, 정치 등 모든 주제에 대하여 Precision과 Recall 비율이 우수한 결과로 나타났다.

Precision과 Recall 의 평균을 보면 의사 결정 트리 모델의 경우, 각각 0.90 및 0.84 이며, 신경망 모델의 경우에는 각각 0.76, 0.77, SVM 모델의 경우에는 각각 0.73, 0.69 로서 평균적으로 의사 결정 트리 모델, 신경망 모델, SVM 모델의 순으로 나타났다. 따라서 본 연구에서는 최종적인 모델로서 의사 결정 트리 모델을 선택하기로 한다. 그러나 이러한 결론은 경제, 사회, 스포츠, 정치 등에 관한 웹 문서 분류에 있어서 의사 결정 트리 모델이 우수하다는 것을 의미하며, 일반적인 분야에 모두 적용되는 것은 아니다. 본 연구에서 도출한 의사 결정 트리 모델을 이용한 웹 문서 분류 방법의 평균 Precision 0.90 과 Recall 0.84 는 문서 분류 분야에서 상당한 수준의 정확도이며, 실제 응용에서도 활용이 가능할 것이라 기대된다.

표 5. 의사 결정 트리 모델의 Precision 및 Recall
Table 5. Precision/Recall for decision tree model

주제	#REL	#RET	Precision	Recall
경제	80	85	0.85	0.90
사회	80	69	0.77	0.66
스포츠	80	78	0.97	0.95
정치	80	88	0.77	0.85
평균	-	-	0.90	0.84

표 6. 신경망 모델의 Precision 및 Recall
Table 6. Precision/Recall for neural network model

주제	#REL	#RET	Precision	Recall
경제	80	88	0.72	0.79
사회	80	55	0.76	0.53
스포츠	80	79	0.92	0.91
정치	80	90	0.70	0.86
평균	-	-	0.76	0.77

표 7. SVM 모델의 Precision 및 Recall
Table 7. Precision/Recall for SVM model

주제	#REL	#RET	Precision	Recall
경제	80	127	0.53	0.84
사회	80	62	0.74	0.56
스포츠	80	66	0.98	0.81
정치	80	65	0.66	0.54
평균	-	-	0.73	0.69

5.2 웹 문서 분류 추가 실험

앞 절에서 보인 경제, 사회, 스포츠, 정치의 4개의 주제에 대한 웹 문서 분류 실험에 부가하여, 3개의 주제 (경제+정치, 사회, 스포츠)에 대한 분류 실험을 추가적으로 시행하였다. 이는 '경제와 정치', 또는 '경제, 정치 및 사회' 주제의 경우, 많은 기사 내용이 명확한 주제로 분류하기에 모호하여 오분류되거나 복수 개의 주제로 중복 분류될 가능성이 높기 때문에 이를 적절히 통합한다면 보다 높은 정확도를 기대할 수 있기 때문이다. 실제로 대부분의 문서 분류 시스템에서 각 주제별 특성이 뚜렷할수록 문서를 명확하게 구분할 수 있는 주제 분별 용어가 많이 발견될 수 있으며, 이는 분류의 정확도 상승으로 이어지기 때문이다. 3 개의 주제에 대한 분류 모델은 이전 절에서 성능이 우수한 것으로 평가된 의사 결정 트리를 이용하여 생성하였고, 실험에 사용된 웹 문서의 배정은 4 개의 주제에 대한 실험과 동일하게 하였다.

3 개의 주제 (경제+정치, 사회, 스포츠)에 대한 분류 실험에서 도출한 교차표와 Precision 및 Recall 의 계산 결과가 각각 표 8 및 표 9에 나타나 있다. 표 8에서 볼 수 있는 바와 같이 320개의 웹 문서 중 294개를 정확히 분류하여 정확도가 91.88% 이고, 26개를 오분류하여 8.12%의 오류율를 나타내었다. 3개의 주제로 분류 시 사회를 정치로 오분류하는 경우가 많았다. 이는 사회 주제 중에는 정치 기사에 포함되는 부분이 일부 존재하기 때문이라고 판단된다. Precision과 Recall 의 경우, 표 9에서 볼 수 있는 바와 같이 각각 평균 0.92, 0.92 로서 4 개의 주제의 경우보다 우수한 성능을 보임을 관찰할 수 있다.

표 8. 3개의 주제 분류 시 분류 결과 교차표
Table 8. Crosstab for 3-subject classification

실제 \ 분류	경제	스포츠	정치/사회	합계
경제	72	0	8	80
스포츠	1	76	3	80
정치/사회	12	2	146	160
합계	85	78	157	320

표 9. 3개 주제별 Precision 및 Recall
Table 9. Precision/Recall for 3-subject classification

주제	#REL	#RET	Precision	Recall
경제	80	85	0.85	0.90
스포츠	80	78	0.97	0.95
정치/사회	160	157	0.93	0.91
평균	-	-	0.92	0.92

6. 결론 및 향후 과제

본 연구에서는 한글 웹 문서에 사용된 한글 형태소 및 키워드의 빈도에 기초하여 문서의 특성을 추출하는 방법을 제시하였고, 이를 기초로 비구조적인 문서의 주제를 예측하여 자동적으로 웹 문서를 분류하는 효과적인 방법을 제시하였다. 문서 분류 모델로서 성능이 검증된 의사 결정 트리, 신경망 모델 및 SVM 방법을 사용하였으며, 의사 결정 트리의 경우 평균 Precision과 Recall 비율은 4 가지 주제의 분류 체계에서 각각 0.90 및 0.84로서 상당한 수준의 정확도를 달성하였다. 또한, 3 가지 주제에 대한 실험에서는 평균

Precision과 Recall 비율이 각각 0.92, 0.92, 로서 주제 간의 구분이 명확할수록 검색의 정확도는 증가함을 보였다. 이는 주제 간의 구분을 가능한 한 명확하게 유지해야 하며, 주제 분별 용어의 선택이 중요함을 의미한다. 따라서 보다 정교한 주제 분별 용어의 선택 방법에 대한 추가적인 연구가 필요하다.

또한 현재 계속 증가하고 있는 웹 문서들의 주제 분별 용어를 계속적으로 갱신하는 과정이 필요하며, 이를 위해 학습용 및 검증용 자료 집합을 주기적으로 최근의 웹 문서로 재구성하는 과정이 필요하다. 즉, 새로운 문서의 주제를 분류할 때, 기존 문서에서 추출한 용어가 아닌 새로운 용어가 출현할 경우 분류 성능이 저하된다는 문제점을 극복해야 하기 때문이다. 따라서 새로운 용어의 출현 시 적응적(adaptive)으로 이를 분류 시스템에 반영하는 방법에 관한 추가적인 연구가 필요하다.

참 고 문 헌

[1] K.Y. Sung and B.H. Yun, "Topic based Web Document Clustering using Named Entities," Journal of Korean Contents, Vol.10, No.5, 2010, pp.29-36.

[2] W.S. Bae, Y. S. Han, and J. W. Cha, "Text Categorization using Topic Signature and Co-occurrence Features," Proc. of KCC2008, Vol.35, No.1, 2008, pp.1-8.

[3] E.K. Chung, "A Semantic-Based Feature Expansion Approach for Improving the Effectiveness of Text Categorization by Using WordNet," Journal of the Korean Society for information Management , Vol.26, No.3, 2009, pp.261-278.

[4] Y. Yang and J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. of 14th Int. Conf. on Machine Learning, 1997, pp.412-420.

[5] Forman, G., "An extensive empirical study of feature selection metrics for text classification," J. Mach. Learn. Res., Vol.3 (2003), pp.1289-1305.

[6] Y.J. Nam and K.H. Kim, "A Study on Automatic Text Categorization of Web-Based Query Using Synonymy List," Journal of information management, Vol.35, No.4, 2004, pp.81-105.

[7] Y.J. Nam, "A study of Korean automatic indexing by morphological analysis," Ph.D Thesis, Chungang University, Seoul, 1995.

[8] D. Ludovic G. Patrick, and Z. Hugo, "HMM-based Passage Models for Document Classification and Ranking," 23rd BCS European Annual Colloquium on Info. Retrieval, 2001.

[9] W. Chen, X. Chang, H. Wang, J. Zhu, and T. Yao, "Automatic Word Clustering for Text Categorization Using Global Information," LNCS Vol.3411, 2005, pp.1-11.

[10] T. Theeramunkong, "Applying passage in Web text mining," Int. J of Intelligent Systems - Intelligent Technologies, Vol.19, Issue 1-2, 2004, pp.149-158.

[11] http://www.cs.cmu.edu/~mccallum/bow /rainbow/

[12] J.S. Lee and J.W. Lee, "A Hangeul Document Classification System using Case-based Reasoning," Asia Pacific Journal of Information Systems, Vol.12, No.2, 2002, pp.179-195.

[13] S.S. Kang, KLT version 2.2.0., http://nlp.kookmin.ac.kr, Korean Language Processing and Information Retrieval Laboratory, 2010.

박 단 호

e-mail : shiningm67@gmail.com

2012년 한국외국어대학교 산업경영공학과 (학사)

2012년~현 재 한국외국어대학교 산업경영공학과 석사과정

관심분야 : 최적화, 메타휴리스틱, 데이터마이닝, 정보검색

최 원 식

e-mail : yuriwonsik@hanmail.net

2012년 한국외국어대학교 산업경영공학과 (학사)

관심분야 : 생산관리, 데이터마이닝, 경영전략

김 홍 조

e-mail : nanunnana@hanmail.net

2012년 한국외국어대학교 산업경영공학과 (학사)

관심분야 : 데이터마이닝, 응용통계, 품질경영

이 석 룡

e-mail : sllee@hufs.ac.kr

1984년 연세대학교 기계공학과(학사)

1993년 연세대학교 산업공학과 전자계산 전공(석사)

2001년 한국과학기술원 정보및통신공학과 (박사)

1984년~1995년 한국IBM 소프트웨어연구소 선임연구원

2002년~현 재 한국외국어대학교 산업경영공학과 교수

관심분야 : 멀티미디어 데이터베이스, 데이터마이닝, 정보검색, 영상분석