http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.3, August 2020, pp.77-93
https://doi.org/10.7838/jsebs.2020.25.3.077

텍스트 문서 분류에서 범주간 유사도와 계층적 분류 방법의 성과 관계 연구

A Study on the Relationship between Class Similarity and the Performance of Hierarchical Classification Method in a Text Document Classification Problem

장수정(Soojung Jang)*, 민대기(Daiki Min)**

초 록

비정형 텍스트 문서를 다중 범주로 분류하는 문제에 있어서, 계층적 분류 방법이 비계층적 분류 방법에 비하여 분류 성능이 우수한 것으로 알려져 있다. 기존 문헌과 다르게 본 연구에서는 사전에 범주들의 계층 구조가 정의된 상황에서 계층적 분류 방법과 비계층적 분류 방법의 성능을 비교하였다. 수자원 분야 기후변화 적응기술과 관련한 논문 분류 데이터와 20NewsGroup 오픈 데이터를 대상으로 계층적/비계층적 분류 방법의 성능을 비교하였다. 본 연구결과 기존 문헌과 다르게 계층적 분류 방법이 비계층적 분류 방법에 비하여 언제나 성능이 우수한 것은 아님을 확인하였다. 계층 구조의 상위/하위 수준에서의 상대적 유사도에 따라서 계층적/비계층적 분류 방법의 성능에 차이가 있음을 확인하였다. 즉, 상위 수준의 유사도가 하위 수준보다 상대적으로 낮은 경우 상위 수준에서의 오분류 감소로 계층적 분류 방법의 성능이 개선됨을 확인하였다.

ABSTRACT

The literature has reported that hierarchical classification methods generally outperform the flat classification methods for a multi-class document classification problem. Unlike the literature that has constructed a class hierarchy, this paper evaluates the performance of hierarchical and flat classification methods under a situation where the class hierarchy is predefined. We conducted numerical evaluations for two data sets; research papers on climate change adaptation technologies in water sector and 20NewsGroup open data set. The evaluation results show that the hierarchical classification method outperforms the flat classification methods under a certain condition, which differs from the literature. The performance of hierarchical classification method over flat classification method depends on class similarities at levels in the class structure. More importantly, the hierarchical classification method works better when the upper level similarity is less that the lower level similarity.

키워드 : 문서 분류, 계층적 분류 방법, 비계층적 분류 방법, 계층 구조, 유사도, SVM
Document Classification, Hierarchical Classification, Flat Classification, Hierarchy Structure, Similarity, SVM

이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A5A2A03067552).

* First Author, Graduate Student, Graduate School(Big Data Analytics), Ewha Womans University (icdi2672@gmail.com)

** Corresponding Author, Associate Professor, School of Business, Ewha Womans University(dmin@ewha.ac.kr)

Received: 2020-07-22, Review completed: 2020-08-12, Accepted: 2020-08-14

1. 서 론

1.1 연구의 배경

2015년 파리기후변화협약에 의한 신(新)기후 체제(New Climate Change Regime)는 '기후변화 적응'이라는 새로운 기술 수요를 탄생시켜 기후변화 적응 산업의 성장 기회를 제공하고 있다. 영국의 기업혁신기술부(BIS)에 의하면 2011~2012년 전 세계 기후변화 적응산업의 규모는 약 117조 원이며, 이후 연간 성장률은 7.7%에 이를 것으로 예측하는 등 기후변화 적응기술 및 해당 산업에 대한 국제적 수요가 급격히 증가하고 있다는 점에 대해서는 논란의 여지가 없다고 할 수 있다.

우리나라 또한 최근 기후변화 적응산업의 발굴과 육성을 위한 정책 연구가 추진되고 있으며, 정책 수립과정에서 국제시장의 수요를 빠르고 효과적으로 파악하는 것이 요구된다. 많은 연구에서 기후변화 적응기술에 대한 수요 조사 연구는 UNFCCC에서 정의한 기후변화 적응기술 분류 체계를 기준으로 이루어진다[2]. 특히 기후변화 적응기술의 국가별 수요를 제시하는 UNFCCC TNA(Technology Needs Assessment) 보고서에서는 UN 중심의 하향식 체계에 근거하여 수요를 체계적으로 분석·정리하고 있다. UNFCCC와 함께 UNEP[28], World Bank, Asian Development Bank 등과 같은 국제기구의 보고서와 개별 국가 수준에서 수립된 기후변화 적응 계획을 중심으로 또한 수요를 확인할 수 있다.

국가별 기후변화 적응기술 수요를 분석함에 있어 광범위한 정보 원천인 대용량 문서 자료를 많은 인적자원을 이용하여 수작업으로 정리함으로써 정합성과 효율성이 낮은 문제가 발생한다. 따라서 UNFCCC TNA와 같이 비정형 텍스트 문서로 제시되는 다수의 수요 자료를 수작업으로 분류, 분석, 정리하는 과정에서 발생하는 비용을 회피하고 효과적인 수요 자료 확보와 정합성 판단을 위하여 비정형 자료 분석의 대표적인 기법인 텍스트 분석을 이용하고자 한다. UNFCCC에서 정의하는 기후변화 적응기술의 분류 체계에 따라서 TNA 보고서와 같은 텍스트 기반 수요 문서를 효율적으로 분류하는 것은 추가적인 수요 분석을 위한 중요한 연구 주제로 판단된다. 본 논문에서는 기후변화 적응기술의 계층 구조와 같이 사전에 정의된 계층 구조를 활용하여 텍스트 문서를 분류하기 위한 문제를 고려하도록 한다.

1.2 연구의 목적

비정형 텍스트 문서 분류(unstructured text document classification)는 텍스트 문서가 어떤 종류의 범주(class)에 속하는지를 구분하는 작업을 의미한다. 최근 사회적으로 대용량 데이터의 증가로 그에 대한 분석이 다양한 방면에서 사용되고 있다[4]. 이에 따라 텍스트 분류 역시 다양한 기술을 통해 연구가 진행되어야 한다. 일반적으로 분류해야 하는 범주가 두 개인 이진분류(binary classification)가 분류해야 하는 범주가 세 가지 이상인 다중 범주 분류(multi-class classification)보다 용이한 것으로 알려져 있다. 따라서 많은 연구에서 다중 범주 분류 문제를 다수의 이진 분류 문제로 구성된 계층으로 구성하고 순차적인 분류 과정을 수행하는 방법을 적용하고 있으며[10], 이와 같은 계층적 분류 방법이 비계층적 분류 방법과

비교하여 성능이 우수한 것으로 알려져 있다[27].

계층 구조가 명확하지 않은 상황에서 계층적 분류 방법은 유사도가 높은 문서를 군집화(clustering)하여 다수 범주의 계층 구조를 생성한다[3, 8, 20, 25]. 범주의 계층 구조를 생성하기 위하여 문서 사이의 유사도에 근거한 군집화 기법인 K-means 기법을 이용하거나[8], 분류기(classifier)와 군집화 기법을 혼합 적용한 방법을 많이 사용하고 있다[20, 25]. 또한 Naik and Rangwala[22]는 전문가가 주관적으로 계층 구조를 정의함으로써 발생하는 일관성 결여 문제점을 보완하기 위하여, 문서 사이의 평균 유사도를 기준으로 군집화를 반복적으로 수행하여 전문가가 제시한 계층 구조를 수정하는 방법을 제안하였다.

본 연구에서는 범주의 계층 구조를 유사도에 근거하여 새롭게 생성하는 대신에 기후변화 적응기술의 분류 체계와 같이 범주들 사이의 계층 구조가 사전에 정의된 상황에서 계층 구조를 활용한 텍스트 문서의 분류 문제를 고려한다. 사전에 정의된 계층 구조를 이용한 문서 분류 방법과 관련한 선행 연구에서는 앞서 언급한 문서 분류 연구와 유사하게 비계층적 방법과 비교하여 계층적 방법이 상대적으로 우수한 분류 성능을 보임을 제시하고 있다. Park and Kim[23]는 계층 구조가 사전에 정의된 상황에서 계층적 분류를 진행하였으며, 분류 성능 개선을 위하여 베이즈(Bayes) 학습법 기반의 알고리즘을 제안하였다. Chen et al.[5]은 다중 범주 분류 문제에서 계층적 접근법을 사용하였으며, 각 단계별로 SVM (Support Vector Machine) 분류기를 이용한 HSVM(Hierarchical SVM) 방법을 제안하였다. Zheng et al[35]은 비정형 이미지 데이터를 대상으로 상위 범주와 여러 개의 하위 범주로 정의된 계층 구조를 이용하여 계층적 분류 성과를 개선하였다.

선행 연구에서 제시하는 바와 같이 비계층적 분류 방법과 비교하여 계층적 분류 방법의 분류 성능이 우수한 것은 분류 대상인 범주 개수가 증가하면 분류 성능이 떨어지는 현상에 근거한다. 즉, 하위 범주들의 집합인 상위 수준(Level)에서 분류 오차를 줄임으로써 하위 수준에서의 분류 오차를 줄이는 것이 가능하다. 하지만 계층적 분류 방법의 경우 상위 수준에서 오분류된 경우 하위 수준에서 오분류를 방지하거나 올바른 범주로 재분류하는 것이 어려운 Blocking 문제를 갖고 있다[27]. 즉, 계층적 분류 방법을 적용함에 있어 상위 계층에서의 분류 성능을 고려하는 것이 매우 중요하다.

본 연구에서는 계층 구조가 사전에 정의된 상황에서 비계층적 분류 대비 계층적 분류 방법의 텍스트 문서 분류 성능을 평가하고자 한다. 이때 Blocking 문제의 특성을 고려하여 계층의 상위 수준과 하위 수준에서의 범주간 유사도를 비교하여 계층 구조를 두 가지 유형으로 구분하고 유형별로 계층적 분류와 비계층적 분류 방법의 성능을 비교하도록 한다. 즉, 상위 수준의 범주간 유사도가 하위 수준의 유사도와 비교하여 상대적으로 낮은 경우에는 계층적 분류를 사용함으로써 상위 수준에서 오분류를 낮추고 결국 최종 분류 성능이 개선될 것으로 예상할 수 있다. 반대로 상위 수준에서의 유사도가 상대적으로 높은 경우에는 계층 분류에 의한 성능 개선 효과가 작아질 것으로 예상된다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 논문에서 제시하는 연구 문제를 검증하기 위한 단계별 연구 절차를 설명한다. 제3장에서는

수치실험에 사용한 데이터와 실험결과를 통하여 연구 문제를 검증한다. 마지막 제4장에서는 본 연구의 결론과 향후 연구주제를 제시한다.

2. 연구 절차

2.1 개요

본 연구의 목적은 범주의 계층 구조가 사전에 정의된 상황에서 텍스트 문서를 범주별로 분류함에 있어 비계층적 분류와 계층적 분류 방법의 성능을 비교·평가하는데 있다. 이를 위하여 <Figure 1>과 같은 연구 절차를 정의하였다.

첫째, 텍스트 문서를 분류하기 위하여 범주의 계층 구조를 정의한다. 앞서 언급한 바와 같이 계층 구조는 문서의 유사도를 근거로 새롭게 생성하는 것이 아니라 계층 구조가 사전에 정의된 문제 상황을 고려한다.

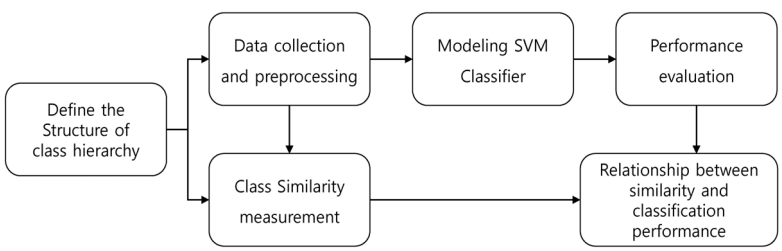
계층 구조에서 정의하는 범주로 텍스트 문서를 분류하기 위하여 수집한 데이터(즉, 텍스트 문서)를 전처리하고, 전처리한 데이터를 이용하여 분류 모형을 개발하는 과정을 수행한다. 데이터의 전처리 과정에서는 문서 집합으로부터 단어들을 추출하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency)와 같은 빈출 단어 기반의 방법을 이용하여 분류를 위한 Feature 집합을 구성한다[18]. 분류기는 텍스트 문서 분류와 관련한 선행연구에서 광범위하게 사용하고 있는 SVM을 사용하였다.

텍스트 문서의 다중 범주 분류와 함께 계층 구조의 상위/하위 수준에서의 범주 간 유사도를 기준으로 계층 구조의 두 가지 유형을 분석하였다. 계층 구조의 수준별로 범주간 유사도는 데이터의 전처리를 통하여 도출한 DTM (Document-Term Matrix)을 대상으로 Cosine similarity를 적용하여 측정하였다. 마지막으로 계층 구조의 두 가지 유형 별로 계층적/비계층적 분류 방법의 성능 차이를 분석하였다. 분석에 사용된 데이터에서 사전에 정의한 계층 구조를 제외한 단계별 주요 연구 절차는 다음 절에서 상세하게 설명하도록 한다.

2.2 계층적 분류와 비계층적 분류 모형

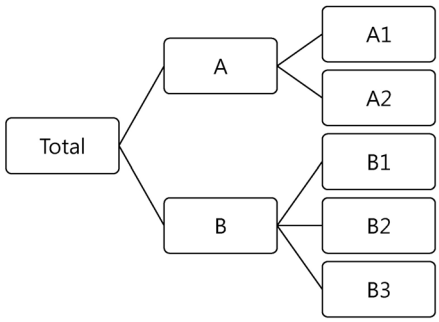
계층적 분류 방법은 텍스트 문서들을 사전에 정의된 범주의 계층 구조에 따라서 상위 수준에서 하위 수준으로 순차적으로 분류를 수행한다. 반면에 비계층적 분류 방법에서는 계층 구조를 무시하고 가장 하위 수준의 범주로 텍스트 문서를 직접 분류하게 된다.

예를 들어, <Figure 2>와 같이 2계층, 총 5개



〈Figure 1〉 Research Framework

의 세부 범주로 구성된 계층 구조에서 비계층적 분류 방법은 계층 구조를 고려하지 않고 A1, A2, B1, B2, B3 등 다섯 가지의 범주로 텍스트 문서를 직접 분류한다. 반면에 계층적 분류에서는 텍스트 문서를 먼저 상위 수준인 A와 B로 분류하고, 다음으로 A (또는 B) 범주로 분류된 문서를 대상으로 하위 수준인 A1과 A2 (또는 B1, B2, B3)로 분류를 수행한다.



〈Figure 2〉 Example of Hierarchy

텍스트 문서를 분류하기 위한 분류 모형에서 Feature 집합을 적절하게 구성하는 것이 매우 중요하다. 비계층적 분류 방법에서는 A1, A2, B1, B2, B3 등 다섯 가지의 범주를 구분하는 한 개의 Feature 집합을 구성하는 것이 필요하지만 계층적 분류 방법에서는 각 수준별로 분류 대상 범주가 달라지므로 개별적인 Feature 집합을 구성하는 것이 요구된다. 예를 들어, 상위 수준에서 A와 B로 텍스트 문서를 분류하는데 필요한 Feature 집합과 독립적으로 하위 수준에서 A1과 A2(또는 B1, B2, B3)를 분류하기 위한 Feature 집합을 구성한다.

문서 분류를 위한 Feature 집합을 구성하기 위하여 빈도기반의 방법론을 이용하였다. 우선 전처리 과정을 통하여 텍스트 문서를 워드 벡터로 구성하고 대소문자, 숫자, 불용어 등을 제거하였다. 불용어 처리는 R 소프트웨어에서 제공하는 기본 불용어 사전과 함께 본 연구에서 고려하는 수자원 분야 기후변화 적응기술의 특성을 고려하여 일부 단어를 추가로 적용하였다. 전처리한 워드 벡터를 대상으로 TF-IDF를 적용하여 DTM을 구성하였다[15]. DTM으로 정리한 단어 집합에서 SVM 모형을 구성하기 위한 Feature 집합을 도출하였다. 빈도수를 기준으로 DTM에서 포함된 전체 단어들의 90%, 95%, 99% 등 3가지 경우를 이용하여 모델을 구축하고 사전 성능평가를 진행하였다. 사전 실험 결과 99%의 단어를 Feature 집합으로 구성하는 경우 성능이 상대적으로 가장 우수하였으며, 이와 같은 사전 실험 결과를 이용하여 이후 실험은 DTM에 포함된 단어의 99%를 적용하여 진행하였다.

본 논문에서는 다중 분류 문제에 대해 상당히 좋은 성능을 보이는 것으로 알려져 있는 SVM 분류기를 사용했다[16]. SVM은 두 가지 범주로 구분되는 데이터를 분류할 때 분리경계면(Separating Hyperplane)을 학습 알고리즘을 이용해서 찾는 기법으로 최근 다양한 연구에서 다중 범주의 분류 문제에서 사용되고 있다. 최근 SVM이 다중 분류 문제에서 많이 이용되는 이유는 다음과 같이 3가지로 요약할 수 있다. 첫째, 확실한 이론적 근거에 기반을 두는 기법으로 결과를 해석하는 것이 용이하다[5]. 둘째, SVM을 사용하여 도출한 결과가 인공신경망을 통해 도출한 결과 성능과 유사하거나 그 이상으로 개선된 결과를 도출한다. 마지막으로 적은 학습 데이터로 짧은 시간 내에 분류 결과를 도출할 수 있으며, 불균형 데이터 집합에 대해서 우수한 성능을 보인다[21].

2.3 분류 성능 평가

텍스트 문서의 분류 성과는 분류 정확도, Recall, Precision, 그리고 Recall과 Precision의 균형값인 F1-Measure를 이용하여 평가한다. 정확도는 분류 대상 문서 중에서 범주를 정확하게 분류한 문서의 비율을 의미한다. 다양한 범주별로 텍스트 문서가 비슷한 비율로 분포되어 있는 균형 데이터(Balanced Data)의 경우 정확도를 이용한 성능 평가가 타당하다. 하지만 특정 범주에 속한 문서가 상대적으로 많은 불균형 데이터(Imbalanced Data)의 경우 정확도 값이 왜곡되는 결과가 발생한다. 이와 같은 문제를 보완하기 위하여 Recall과 Precision을 함께 고려하는 것이 필요하다.

Recall은 (True Positive) / (True Positive + False Negative)로 계산되며, 특정 범주 값을 갖는 문서 중에서 분류 모형이 정확하게 범주를 분류한 비율을 나타낸다. Precision은 (True Positive) / (True Positive + False Positive)로 측정되며, 특정 범주로 분류 모형이 분류한 문서 중에서 실제 해당 범주에 속하는 문서의 비율을 의미한다. Precision과 Recall은 서로 반대의 개념을 갖고 있으므로, 분류 모형의 성능을 효과적으로 표현하는 데 한계가 존재한다. 따라서 일반적으로 Recall과 Precision의 조화평균 값인 F1-Measure를 이용하여 모형의 성능을 평가한다.

2.4 범주의 유사도와 성능 비교

본 연구에서는 계층별 유사도를 기준으로 계층 구조를 두 가지 유형을 구분하였다. 첫 번째 유형은 상위계층에서의 범주 간 유사도와 비교하여 하위계층에서의 범주 간 유사도가 상대적으로 높은 경우이며, 두 번째 유형은 반대로 하위계층의 범주 간 유사도가 낮은 경우로 정의된다. 첫 번째 유형의 경우 상위계층의 범주 간 유사도가 상대적으로 낮으므로 상위계층에서 우선적으로 분류를 수행함으로써 하위계층에서의 오분류 가능성을 낮추는 것이 가능할 것으로 예상된다. 반대로 상위계층에서의 범주간 유사도가 높은 경우 오분류 가능성이 증가하며, 계층적 분류에서 Blocking 문제가 많이 발생할 가능성이 존재한다. 따라서 이 경우 계층적 분류 방법에 따른 성능 개선의 가능성이 비계층적 분류와 비교하여 상대적으로 적어질 것으로 예상된다.

범주 사이의 유사도는 분류 모형에서 사용한 Feature 집합을 대상으로 cosine similarity를 이용하여 측정하였다. 예를 들어, 앞서 제시한 <Figure 2>에서 상위계층의 2개 범주인 A와 B의 유사도는 A와 B 범주의 분류에 사용된 Feature 집합의 Cosine similarity를 이용하였다. 같은 방식으로 하위계층인 유사도는 A1과 A2(또는 B1, B2, B3)의 분류를 위하여 도출한 Feature 집합을 이용하였다.

계층별 유사도를 기준으로 구분한 두 가지 계층 구조의 유형에 따라서 계층적/비계층적 분류 방식의 성능 차이를 검증하기 위한 통계적 분석을 수행한다. 이를 위하여 계층 구조의 두 가지 유형 사이에 비계층적 분류 방법의 정확도 대비 계층적 분류 방식의 정확도 개선율에 차이가 있는지 t-검정을 수행하였다.

3. 수치 실험

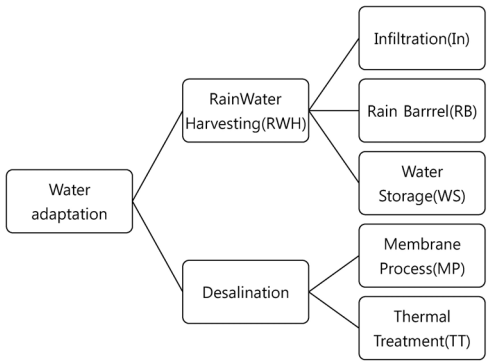
3.1 데이터

본 연구의 목적인 계층적 문서 분류 방법과

비계층적 문서 분류 방법의 성능 차이를 비교하기 위하여 두 가지 연구 데이터를 이용하였다. 먼저 앞서 연구배경에서 언급한 바와 같이 기후변화 적응기술의 수요를 효과적으로 분석하기 위하여 수자원 분야의 기후변화 적응기술을 대상으로 제안 연구방법론을 적용하였다. 수자원 분야 기후변화 적응기술의 구조는 CTCN(Climate Technology Centre and Network)에서 정의하는 기술 계층 구조를 이용하되, 문제를 단순화하기 위하여 가장 대표적인 두 가지 기술인 Rainwater Harvesting(우수 집수)과 Desalination(해수 담수화)만을 고려하였다[2]. 본 연구에서 제안한 계층적 분류 방법과 유사도에 의한 성능 변화를 검증하기 위하여 분류 대상 범주가 다양하고 분류 목적으로 많이 활용되는 데이터를 추가적으로 고려하였다. 20NewsGroup 데이터는 다양한 범주의 뉴스 기사를 포함하는 공개 데이터이며, 텍스트 문서 분류 연구에 광범위하게 사용되고 있다. 본 연구에서는 제안 연구 방법론의 검증에 문제가 없는 수준에서 실험을 단순화하기 위하여 20NewsGroup의 일부 4개 상위 범주를 활용하였다.

수자원 분야 기후변화 적응기술의 계층 구조는 <Figure 3>과 같다. 상위 수준에서는 Rainwater Harvesting과 Desalination의 2개 범주로 구성되며, 각 범주의 하위 수준은 Infiltration, Rain Barrel, Water Storage와 Membrane Process, Thermal Treatment 등으로 구성된다.

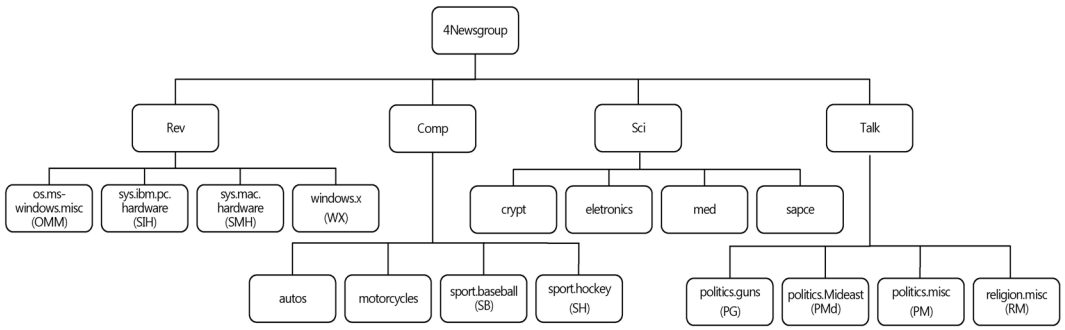
수자원 분야 기후변화 적응기술의 계층 구조에 따라서 본 연구의 목적에 적합하게 정리된 텍스트 문서 데이터를 자체적으로 구성하였다. 데이터 구성을 위하여 하위 수준의 5개 범주에 대응하는 기후변화 적응기술에 대한 CTCN의 설명으로부터 주요 키워드를 도출하고 이를 이용하여 2015년부터 2019년까지의 논문을 “Science Direct”에서 수집하였다. 학회 웹사이트의 HTML 문서로부터 학회 주제에 대한 토픽 정보를 추출하여 주제에 따라 문서를 학회에 자동 분류되는 기법을 제안한 논문[17]과 다르게 본 연구는 논문의 제목과 초록을 이용했다. 검색과정으로 수집된 논문은 전문가의 검토를 거쳐서 범주와의 관련성을 추가로 검토하였으며, 논문의 제목과 초록을 추후 분석과정에서 사용하였다.



〈Figure 3〉 Water Adaptation Technologies Structure

최종적으로 Infiltration은 376개, Rain Barrel은 307개, Water Storage는 314개, Membrane Process는 1950개, Thermal Treatment는 366개의 논문을 사용하였다.

20NewsGroup의 4개 범주로 구성된 데이터(이후 4NewsGroup)의 계층 구조는 <Figure 4>와 같다. Rev 범주의 하위 수준 범주별로 분석에 사용한 텍스트 문서의 숫자는 OMM 573개, SH 590개, SMH 577개, WX 583개와 같다. Comp의 하위 수준 범주는 autos 594개, motorcycles 598개, SB 597개, SH 600개의 데이터로 구성되어 있다. Sci의 하위에는 crypt 592개,



〈Figure 4〉 4NewsGroup Structure

electronics 590개, med 592개, space 592개의 문서가 있으며, 마지막으로 Talk 범주의 하위에는 PG 546개, PMd 564개, PM 465개, RM 377개의 데이터로 구성되어 있다.

3.2 수자원 분야 기후변화 적응기술의 논문 분류

수자원 분야 기후변화 적응기술과 관련하여 수집한 3,313개의 논문을 7:3으로 구분하여 분류 모형의 학습과 성능평가 실험에 사용하였다. 논문 분류 결과를 <Table 1>에 요약하여 제시하였으며, 상세 실험 결과를 보여주는 Confusion Matrix는 부록에 제시하였다.

단계적 문서 분류 방법을 적용한 실험 결과를 살펴보면 상위 수준에서의 분류 정확도는 95.37%로 상당히 높게 도출되었으며, 하위 수준에서의 정확도는 Rainwater Harvesting과 Desalination 범주에서 각각 68.33%와 85.27%로 나타났다. 비계층적 문서 분류를 사용하는 경우의 정확도 66.8%와 비교하여 계층적 분류 방법을 활용함으로써 분류 성능이 개선되었음을 확인하였다.

본 연구에서 사용한 수자원 분야 기후변화 적응기술 데이터의 경우 두 가지 하위 수준의 범주에 해당하는 문서가 각각 997개와 2,316개로 구성되어 불균형한 분포 특성을 갖고 있다. 이와 같은 데이터 불균형을 고려하여 성능을 평가하기 위하여 F1-Measure를 추가적으로 검토하였다. 먼저 비계층적 분류 방법의 경우

〈Table 1〉 Model Performance Measures about Water Adaptation Technologies

	Hierarchy			Flat
	Level 1	Level 2 (RWH)	Level 2 (Desalination)	
Accuracy	0.9537	0.6833	0.8527	0.6680
Precision	0.9580	0.7069	0.9113	0.5763
Recall	0.9336	0.6990	0.7027	0.3839
F1 Measure	0.9456	0.7029	0.7935	0.4608

Precision과 Recall 모두 0.6 미만의 결과를 보였으며, F1-Measure는 0.4608으로 확인되었다. 계층적 분류 방법은 Precision과 Recall 모두 비계층적 분류 방법과 비교하여 우수한 성능을 보였다. 계층적 분류 방법을 적용한 결과 F1-Measure의 평균값은 약 0.7482로 비계층적 분류 방법과 비교하여 매우 우수한 결과를 보였다.

3.3 4NewsGroup의 기사 분류

4NewsGroup에 포함된 기사 데이터를 7:3으로 구분하여 6,321개의 기사를 분류 모형의 학습에 사용하였으며, 2,709개의 기사는 성능 평가를 위한 실험에 사용하였다. 4NewsGroup 데이터의 경우 개별 범주에 포함된 데이터의 수가 약 2,000여 개로 유사하여 균형 데이터의 특성을 보이고 있다.

계층적, 비계층적 분류 방법을 적용한 실험 결과를 <Table 2>와 <Table 3>에 요약하여 제시하였다. 계층적 분류 방법의 정확도는 모든 수준과 범주에 있어서 비계층적 분류 방법의 정확도 결과인 약 81.99%와 비교하여 우수하였다. 특히 상위 수준에서의 분류 정확도는 90% 이상이었는데, 이를 통하여 상위 수준에서의 오분류 경우가 감소함으로써 하위 수준에서의 분류 정확도가 개선되었음을 알 수 있다. Precision과 Recall 그리고 F1 Measure의 경우에도 비계층적 분류와 비교하여 계층적 분류 방법에서 우수한 성능을 확인하였다.

3.4 유사도와 분류 성능 비교

계층별 유사도를 기준으로 구분한 두 가지 계층 구조의 유형에 따라서 계층적/비계층적 분류 방식의 성능 차이를 검증하였다. 계층 구

〈Table 2〉 Flat Classification Model Performance Measures: 4NewsGroup

	Hierarchy	Flat
	Level 1	
Accuracy	0.9081	0.8199
Precision	0.9078	0.8324
Recall	0.9104	0.8146
F1 Measure	0.9091	0.8234

〈Table 3〉 Hierarchical Classification Model Performance Measures: 4NewsGroup

	Hierarchy			
	Level 2(Comp)	Level 2(Rec)	Level 2(Sci)	Level 2(Talk)
Accuracy	0.8224	0.8795	0.8466	0.8622
Precision	0.8255	0.8897	0.8704	0.8664
Recall	0.8707	0.8773	0.7836	0.8831
F1 Measure	0.8475	0.8834	0.8247	0.8746

조에 있어 계층의 수준별 유사도는 범주를 분류하기 위하여 정의한 Feature 집합을 이용하여 측정하였다.

먼저 계층 구조에서 상위 수준과 하위 수준

〈Table 4〉 Similarity and Classification Performance

Label		Similarity		Improvement Rate*
		Level 1	Level 2	
4NewsGroup				
Comp	autos-motor	0.5317	0.5925	-0.0225
	autos-SB			0.0139
	autos-SH			0.0903
	motor-SB			0.0287
	motor-SH			0.1143
	SB-SH			0.1481
Rec	OMM-SIH	0.5494	0.4713	-0.0060
	OMM-SMH			-0.0042
	OMM-WX			0.0259
	SIH-SMH			-0.0208
	SIH-WX			0.0074
	SMH-WX			0.0101
Sci	crypt-elec	0.5823	0.4949	-0.0528
	crypt-med			0.0523
	crypt-space			0.0461
	elec-med			-0.0195
	elec-space			-0.0246
	space-med			0.0800
Talk	PG-PMd	0.5319	0.5291	0.1818
	PG-PM			0.0578
	PG-RM			0.0676
	PMd-PM			0.0852
	PMd-RM			0.0923
	PM-RM			-0.0194
Water Adaptation Technologies				
RWH	In-RB	0.6010	0.7200	0.2201
	In-WS			0.9360
	RB-WS			0.4821
Desalination	MP-TT	0.6010	0.7545	0.2586

* Improvement Rate = (performance of hierarchical classification - performance of flat classification) / performance of flat classification

에서 범주간 유사도를 측정한 결과를 <Table 4>에 제시하였다. 4NewsGroup 데이터의 경우 4개의 상위 수준 범주 중에서 Comp를 제외한 세 개의 범주는 상위 수준의 유사도가 하위 수준의 유사도보다 높은 유형으로 확인되었다. 반대로 수자원 분야 기후변화 적응기술 데이터의 경우 모든 범주에서 상위 수준의 유사도가 하위 수준의 유사도 보다 작은 유형으로 확인되었다. 앞서 설명한 바와 같이 문서 간 유사도가 높은 경우에는 분류 성능이 떨어지는 결과를 기대할 수 있다, 따라서 상위 수준에서 유사도가 상대적으로 높은 경우 오분류가 많아지고 이를 하위 수준에서 재분류할 수 없으므로 계층적 분류 방식의 성능이 저하되는 Blocking 문제가 발생할 것으로 예상된다.

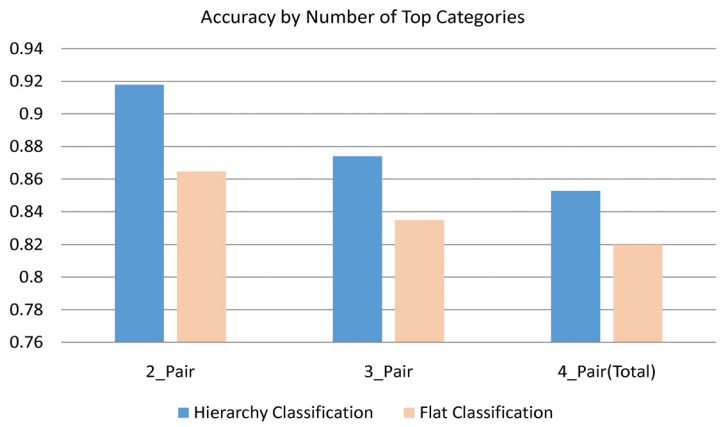
상위 수준의 유사도가 하위 수준의 유사도보다 상대적으로 높은 경우 계층적 분류 방법의 성능이 비계층적 분류 방법과 비교하여 상대적으로 저하되는 현상을 통계적으로 검증하였다. 이를 위하여 <Table 4>에 제시된 결과를 기준으로 계층 구조의 유형을 두 가지로 구분하였다. 즉, 상위 수준의 유사도가 상대적으로 더 작은 계층 구조로 Comp, RWH, Desalination을 구분하였으며, 반대로 하위 수준의 유사도가 상대적으로 더 작은 계층 구조로 Rec, Sci, Talk를 확인하였다.

계층 구조의 유형에 따라서 계층적/비계층적 분류 방법의 성능 차이를 비교하기 위하여 두 개 유형 사이에 성능 차이가 존재하는지 통계적 검증을 수행하였다. 비계층적 분류 방법 대비 계층적 분류 방법의 개선율을 의미하는 성능 개선율 결과가 두 유형 집단 사이에 차이가 있는지 t-검정을 수행하였으며, t-검정 결과 성능 개선율에 유의미한 차이가 존재함을 확인하였다(p-value = 0.04062).

5. 결 론

본 연구는 범주의 계층 구조가 사전에 정의된 상황에서 텍스트 문서의 다중 범주 분류를 위한 방법을 고려하였다. 일반적으로 계층적 분류 방법을 이용한 텍스트 문서의 분류 성과가 비계층적 분류 방법과 비교하여 우수한 분류 성능을 나타내는 것으로 알려져 있다. 하지만 계층적 분류 방법이 갖는 Blocking 문제의 특성을 고려하여 본 연구에서는 계층의 상위 수준과 하위 수준에서의 범주간 유사도를 비교하여 정의한 계층 구조의 두 가지 유형에 따라서 계층적 분류와 비계층적 분류 방법의 성능에 차이가 발생함을 검증하고자 하였다. 본 연구 목적을 위하여 텍스트 데이터의 전처리와 Feature 집합의 구성, SVM을 이용한 분류 모형을 제시하고, 두 가지 데이터를 이용한 수치실험을 수행하였다.

본 논문의 분석결과 다음과 같은 흥미로운 결과를 확인하였다. 먼저, 앞서 문헌연구에서 제시한 바와 같이 계층적 문서 분류 방법이 비계층적 분류 방법과 비교하여 우수한 분류 성능을 보였다. 특히, 균형 데이터와 비교하여 불균형 데이터에서 계층 분류 방법에 의한 성능 개선을 확인할 수 있었다. 하지만 본 연구의 주요 가설과 같이 범주간 유사도와 분류 방법의 성능 사이에서 유의미한 관계를 확인할 수 있었다. 즉, 비계층적 분류 방법 대 계층적 분류 방법의 성능 개선율은 상위 수준에서의 범주간 유사도가 하위 수준과 비교하여 상대적으로 낮은 경우에 더 높게 발생하였다. 계층적 문서 분



〈Figure 5〉 Accuracy by Number of Parent Category

류 방법의 경우 비계층적 분류와 비교하여 SVM 분류 모형의 개발과 분류 과정에서 더 많은 비용이 발생하는 것을 고려할 때, 계층 구조의 특성을 반영한 문서 분류 방법을 효과적으로 선택하는 것이 필요하다.

본 논문은 범주 간 유사도와 계층적/비계층적 분류 방법의 관계를 분석한 의의를 갖지만 다음과 같은 몇 가지 추가 연구가 필요하다. 첫째, 계층적 분류를 이용함으로써 기대되는 Blocking 문제의 개선 효과를 정량화하여 검토하는 것이 필요하다. 본 연구에서 제시하고 있는 계층적 분류 방법에 의한 분류 성능 개선 효과를 이용하여 간접적으로 Blocking 문제의 방지 효과를 확인할 수 있다. 하지만 이를 보다 정량적으로 검토함으로써 계층적 분류 방법의 개선 방안을 고려하는 것이 가능할 것이다. 둘째, 다양한 계층 구조를 갖는 경우에 대한 연구가 필요하다. 본 논문에서는 4NewsGroup 데이터의 경우 상위 수준의 범주가 4개인 경우를 고려하였다. 상위 수준의 범주가 2개, 3개, 4개인 경우 계층적/비계층적 분류 방법의 분류 정확도를 살펴보면 <Figure 5>와 같이 범주의 수가 증가할수록 분류 성능은 감소함을 확인할 수 있다. 즉, 계층 구조에 있어서 각 수준별 범주의 숫자에 따라서 분류 성능의 변화가 예상된다. 각 수준별 범주의 숫자와 함께 계층의 수준이 3단계 이상인 경우를 고려하는 것이 요구된다. <Figure 3>과 <Figure 4>와 같이 본 논문에서는 2단계 수준으로 구성된 단순한 계층 구조만을 고려하였다. 계층 구조가 3단계 이상으로 확장되는 경우 수준간 유사도 비교를 통한 계층 구조의 유형화 방안을 고려하는 것이 필요하다. 마지막으로 본 논문에서는 분류 모형으로 SVM을 고려하였다. 최근 텍스트 분석을 위한 다양한 딥러닝(Deep Learning) 기반 방법이 사용되고 있다[14]. SVM 대신에 딥러닝과 같은 최신 분류 기법을 사용함으로써 분류 성능의 개선을 기대할 수 있다.

References

[1] Agnihotri, D., Verma, K., and Tripathi,

P., “Variable global feature selection scheme for automatic classification of text documents,” Expert Systems with Applications, Vol. 81, pp. 268-281, 2017.

[2] Bertule, M., Appelquist, L. R., Spensley, J., Traerup, S. L. M., and Naswa, P., “Climate change adaptation technologies for water: A practitioner’s guide to adaptation technologies for increased water sector resilience,” CTCN publications, Copenhagen, Denmark, 2018.

[3] Beyan, C. and Fisher, R., “Classifying imbalanced data sets using similarity based hierarchical decomposition,” Pattern Recognition, Vol. 48, pp. 1653-1672, 2015.

[4] Byun, J. H., “Current Status and Perspectives of Fintech Innovation,” Journal of New Industry and Business, Vol. 26, No. 2, pp. 35-48, 2018

[5] Chen, Y., Craword, M. M., and Ghosh, J., “Integrating support vector machines in a hierarchical output space decomposition framework,” IEEE International Geoscience and Remote Sensing Symposium, Vol. 2, pp. 949-952, 2004.

[6] Cristianini, N. and Shawe-Taylor, J., “An introduction to support vector machines and other kernel-based leartning methods”, Cambridge University Press, MA, 2000.

[7] Du, Y., Liu, J., Ke, W., and Gong, X., “Hierarchy construction and text classification based on the relaxation strategy and least information model,” Expert Systems with Applications, Vol. 100, pp. 157-164, 2018.

[8] Duan, K. B. and Keerthi, S. S., “Which is the best multiclass SVM method? An empirical study,” International Workshop on Multiple Classifier Systems, Vol. 3531, pp. 278-285, 2005.

[9] Gargiulo, F., Silvestri, S., Ciampi, M., and De Pietro, G., “Deep neural network for hierarchical extreme multi-label text classification,” Applied Soft Computing, Vol. 79, pp. 125-138, 2019.

[10] Kang, S., Cho, S., and Kang, P., “Constructing a multi-class classier using one-against-one approach with different binary classifiers,” Neurocomputing, Vol. 149, pp. 677-682, 2015.

[11] Kim, P. J. and Lee, J. Y., “An experimental study on the performance improvement of automatic classification for the articles of korean journals based on controlled keywords in international database,” Journal of the Korean Society for Library and Information Science, Vol. 48, No. 3, pp. 491-510, 2014

[12] Kim, P. J., “An analytical study on automatic classification of domestic journal articles based on machine learning,” Journal of the Korean Society for information Management, Vol. 35, No. 2, pp. 37-62, 2018.

[13] Kim, Y. S. and Lee, B. Y., “Multi-class support vector machines model based clustering for hierarchical document cate-

gorization in big data environment," The Journal of the Korea Contents Association, Vol. 17, pp. 600-608, 2017.

[14] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D,, "Text classification algorithms: A survey," Information, Vol. 10, No. 4, 2019.

[15] Lee, J. H., Yi, J. S., and Son, J. W., "Unstructured construction data analytics using R programming: Focused on overseas construction adjudication cases", Journal of the Architectural Institute of Korea Structure & Construction, Vol. 32, No. 5, pp. 37-44, 2016.

[16] Lee, J. S. and Kwon, J. G., "A hybrid SVM classifier for imbalanced data sets," Journal of Intelligence and Information Systems, Vol. 19, pp. 125-140, 2013.

[17] Lee, S. K. and Kim, K., "Academic Conference Categorization According to Subjects Using Topical Information Extraction from Conference Websites," The Journal of Society for e-Business Studies, Vol. 22, No. 2, pp. 61-77, 2017.

[18] Lee, S. J. and Kim, H. J., "Keyword extraction from news corpus using modified TF-IDF," The Journal of Society for e-Business Studies, Vol. 14, No. 4, pp, 59-73, 2009.

[19] Lorena, A. C., De Carvalho, A. C., and Gama, J. M. P., "A review on the combination of binary classifiers in multiclass problems," Artificial Intelligence Review, Vol. 30, No. 19, 2008.

[20] Madzarov, G., Gjorgjevikj, D., and Chorbev, I., "A multi-class SVM classifier utilizing binary decision tree," Informatica, Vol. 33, 2009.

[21] Min, J. H. and Lee, Y. C., "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," Expert Systems with Applications, Vol. 28, pp. 603-614, 2005.

[22] Naik, A. and Rangwala, H., "Improving large-scale hierarchical classification by rewiring: A data-driven filter based approach," Journal of Intelligent Information Systems, Vol. 52, pp. 141-164, 2019

[23] Park, J. H. and Kim, J. S., "A text classification system for hierarchical categories," The Korean Institute of Information Scientists and Engineers, Vol. 27, No. 2, pp. 128-130, 2000.

[24] Silla, C. N. and Freitas, A. A., "A survey of hierarchical classification across different application domains," Data Mining and Knowledge Discovery, Vol. 22, pp. 31-72, 2011

[25] Silva-Palacios, D., Ferri, C., and Ramı́rez-Quintana, M. J., "Probabilistic class hierarchies for multiclass classification," Journal of Computational Science, Vol. 26, pp. 254-263, 2018

[26] Sun, A., Lim, E. P., Ng, W. K., and Srivastava, J., "Blocking reduction strategies in hierarchical text classification," IEEE Transactions on Knowledge and

Data Engineering, Vol. 16, pp. 1305-1308, 2004

[27] Tegegnie, A. K., Tarekegn, A. N., and Alemu, T. A., "A comparative study of flat and hierarchical classification for amharic news text using SVM," Information Engineering and Electronic Business, Vol. 3, pp. 36-42, 2017.

[28] UNEP, "Technologies for climate change mitigation," UNEP, 2011.

[29] Vapnik, V., "Estimation of Dependences Based on Empirical Data." Nauka, Moscow, 1979.

[30] Vapnik, V., "The nature of statistical learning theory", Chapter 5. Springer-Verlag, New York, 1995.

[31] Williams, T. P. and Gong, J., "Predicting construction cost overruns using text mining, numericaldata and ensemble classifiers," Automation in Construction, Vol. 43, pp. 23-29, 2014

[32] Yoon, Y. W. Lee, C. K., and Lee, G. B., "Hierarchical text categorization using support vector machine," Annual Conference on Human and Language Technology, pp. 7-13, 2013.

[33] Zhang, L., Shah, S. K., and Kakadiaris, I. A., "Hierarchical multi-label classification using fully associative ensemble learning," Pattern Recognition, Vol. 70, pp. 89-103, 2017.

[34] Zhao, Z., Wang, X., and Wang, T., "A novel measurement data classification algorithm based on SVM for tracking closely spaced targets," IEEE Transactions on Instrumentation and Measurement, Vol. 68, No. 4, pp. 1089-1100, 2019.

[35] Zheng, J., Guo, Y., Feng, C., and Chen., H., "A hierarchical neural network based document representation approach for text classification," Mathematical Problems in Engineering, Vol. 2018, 2018.

〈Appendix〉

〈Appendix 1〉 Confusion Matrix of Water Adaptation Technologies: Level 1

		Reference	
Prediction		Desalination	RainWater Harvesting
	Desalination	676	37
	RainWater Harvesting	9	272

〈Appendix 2〉 Confusion Matrix of RWH Classification

		Reference			
Prediction		RWH_In	RWH_RB	RWH_WS	Des_MP
	RWH_In	94	14	41	2
	RWH_RB	4	57	4	4
	RWH_WS	17	0	41	3
	Des_MP	0	0	0	0

〈Appendix 3〉 Confusion Matrix of Desalination Classification

		Reference				
Prediction		Des_MP	Des_TT	RWH_RB	RWH_WS	Des_MP
	Des_MP	562	67	3	30	4
	Des_TT	1	46	0	0	0
	RWH_In	0	0	0	0	0
	RWH_RB	0	0	0	0	0
	RWH_WS	0	0	0	0	0

〈Appendix 4〉 Confusion Matrix of Water Adaptation: Non-hierarchical Classification

		Reference				
Prediction		Des_MP	Des_TT	RWH_RB	RWH_WS	RWH_WS
	Des_MP	561	99	14	61	15
	Des_TT	6	14	0	0	0
	RWH_In	1	0	60	23	58
	RWH_RB	0	0	3	13	1
	RWH_WS	4	0	41	4	16

저 자 소 개

장수정 (E-mail: icdi2672@gmail.com)
2018년 숙명여자대학교 경영학부 (학사)
2018년 숙명여자대학교 빅데이터분석학(학사)
2018년~2020년 이화여자대학교 빅데이터분석학 (석사)
관심분야 텍스트 마이닝, 빅데이터 분석

민대기 (E-mail: dmin@ewha.ac.kr)
1999년 서울대학교 산업공학과 (학사)
2001년 서울대학교 산업공학과 (석사)
2010년 퍼듀대학교 산업공학과 (박사)
2001년~2006년 LG CNS
2010년~현재 이화여자대학교 경영대학 부교수
관심분야 빅데이터, 강화학습, Stochastic programming, 전력 시스템