한국시뮬레이션학회 논문지
Vol. 32, No. 1, pp. 35-44 (2023. 3)
http://doi.org/10.9709/JKSS.2023.32.1.035
ISSN 1225-5904

토픽모델링을 이용한 약어 중의성 해소

이운교 · 김자희 † · 양준기

Abbreviation Disambiguation using Topic Modeling

Woon-Kyo Lee · Ja-Hee Kim† · Junki Yang

ABSTRACT

In recent, there are many research cases that analyze trends or research trends with text analysis. When collecting documents by searching for keywords in abbreviations for data analysis, it is necessary to disambiguate abbreviations. In many studies, documents are classified by hand-work reading the data one by one to find the data necessary for the study. Most of the studies to disambiguate abbreviations are studies that clarify the meaning of words and use supervised learning. The previous method to disambiguate abbreviation is not suitable for classification studies of documents looking for research data from abbreviation search documents, and related studies are also insufficient. This paper proposes a method of semi-automatically classifying documents collected by abbreviations by going topic modeling with Non-Negative Matrix Factorization, an unsupervised learning method, in the data pre-processing step. To verify the proposed method, papers were collected from academic DB with the abbreviation 'MSA'. The proposed method found 316 papers related to Micro Services Architecture in 1,401 papers. The document classification accuracy of the proposed method was measured at 92.36%. It is expected that the proposed method can reduce the researcher's time and cost due to hand work.

Key words : Word Sense Disambiguation, NMF, Text Analysis, Micro Services Architecture

요 약

최근 텍스트 분석으로 트렌드 분석이나 연구 동향 분석을 하는 연구 사례가 많다. 텍스트 분석을 위한 자료 수집에 사용되는 검색어가 약어일 때 약어의 특성상 의미 중의성 해소가 필요하다. 다수의 연구에서는 연구에 필요한 자료를 찾기 위해 수작업으로 자료를 하나씩 읽어 문서를 분류하고 있다. 약어의 의미 중의성 해소를 위한 연구는 단어의 의미를 명확화하는 연구가 대부분이고 지도학습을 이용하고 있다. 약어 중의성 해소를 위한 선행 방법은 약어로 검색된 자료에서 연구 대상 자료를 찾는 문서 분류에는 적합하지 않으며 관련 연구도 부족하다. 본 연구에서는 데이터 전처리 단계에서 비지도 학습 방법인 비음수 행렬 분해 방법으로 토픽 모델링을 진행하여 약어로 수집된 문서를 반자동으로 분류하는 방법을 제시한다. 이를 검증하기 위해 'MSA'라는 약어 검색어로 학술 데이터베이스에서 논문 자료를 수집했다. 수집된 논문 1,401편에서 제안된 방법으로 316편의 Micro Services Architecture와 관련된 논문을 찾았다. 제안된 방법의 문서 분류 정확도는 92.36%로 측정되었다. 제안된 방법이 수작업에 따른 연구자의 시간과 비용을 줄일 수 있기를 기대한다.

주요어 : 단어 중의성 해소, 비음수 행렬 분해, 텍스트 분석, 마이크로 서비스 아키텍처

1. 서론

트랜드 분석이나 주제 분석을 위하여 문서를 수집하는 경우 키워드를 이용해서 문헌을 검색하고 수집된 문서를 연구자가 일일이 확인하여 대상 문서를 선정하는 과정을 수행한다. 연구 동향을 분석하는 연구(나상태 등(2016), 정명석 등(2017), 김효선(2021))를 보면 연구자가 연구와

Received: 24 November 2022, **Revised**: 27 February 2023, **Accepted**: 27 February 2023

† **Corresponding Author**: Ja-Hee Kim
E-mail: jahee@seoultech.ac.kr
Graduate school of Public Policy and Information Technology, Seoul National University of Science & Technology, Seoul, Korea

관련성이 없는 논문을 찾기 위해 수집된 논문의 제목과 영문 요약문을 일일이 읽어 제거한다는 것을 알 수 있다. 검색된 논문이 수천 건이 넘는 경우 연구자가 일일이 요약문을 읽어 연구 대상 논문을 찾는 작업은 시간과 노력이 필요하다. 특히 약어로 논문을 검색하는 경우 약어의 의미 중의성 문제로 여러 주제의 논문이 함께 검색되어 논문의 수도 많고 대상 논문을 찾기도 어렵다.

약어로 검색된 논문의 분류가 어려운 것은 약어의 특성상 여러 의미를 함께 가지고 있어 하나의 약어 검색으로 다양한 주제의 문서가 검색되기 때문이다. 약어의 의미 중의성 해소 연구는 대부분 단어의 의미를 명확화하기 위한 연구이며 1990년대 이후 말뭉치 통계 정보를 이용하고 있고 말뭉치 학습 여부에 따라 지도적 방법과 비지도적 방법으로 나누어지며 지도학습 방법의 어려움으로 최근에는 비지도적 방법으로 연구가 많이 진행되고 있다(Bevilacqua, M. et al.(2021), Navigli, R. (2009)). 하지만 약어로 검색된 문서를 분류하는 연구는 미흡하고 수집단계에서 수작업 양이 많아 최소화할 방법을 찾는 연구는 의미가 있다.

본 연구에서는 약어로 검색된 논문을 분류하는 방법을 제안한다. 분류 방법은 3단계로 진행한다. 첫 번째 단계는 요약문을 TF-IDF(Term Frequency-Inverse Document Frequency)로 벡터화한다. 두 번째 단계는 비음수 행렬 분해 방법으로 토픽 모델링하여 토픽별 주요 키워드로 토픽을 분류한다. 세 번째 단계는 토픽에 해당하는 논문을 찾아 분류한다. 관련 없는 논문을 제거하고 3단계를 반복하여 진행한다. 제안된 방법은 사례를 통하여 수작업으로 분류한 경우와 비교하여 검증한다. 비교 분석은 혼돈 매트릭스를 활용하여 분류 결과를 평가하고 토픽 모델링을 진행하여 도출된 주제 영역을 비교한다.

본 연구는 텍스트 마이닝의 전처리 과정에서 수작업을 줄여주는 방법을 제안한다. 특히 약어로 검색되어 다양한 주제가 혼합된 경우 효과적인 방법이다.

2장에서 관련 선행 연구를 분석하고 3장에서 제안하는 분류 방법을 설명한다. 4장에서 사례를 통하여 제안된 분류 방법으로 문서를 분류하고 수작업과 비교한다. 마지막 5장에서 연구 내용을 요약하고 향후 발전 방향을 논의한다.

2. 선행 연구

약어 검색으로 문서를 수집하는 경우 단어 중의성 문제로 다양한 문서가 혼합되어 수집되는 문제가 발생한다. 단어 중의성 문제를 해결하기 위한 연구를 2.1에서 살펴보고 다양한 주제의 혼합된 문서를 주제별로 분류하는 토픽 모델링 방법인 비음수 행렬 분해를 2.2에서 살펴본다.

2.1 단어 중의성 해소(WSD; Word Sense Disambiguation) 연구

단어 중의성 해소 연구는 자연어 처리(NLP)나 인공지능(AI)연구에서 오래된 과제로, 1950년대부터 기계 번역 관점에서 주요 관심사였다. 현재에도 단어 중의성 해소는 자연어 처리나 인공지능 분야의 도전적 과제 중 하나로 남아 있다(Bevilacqua, M. et al.(2021)).

단어 중의성 해소 연구는 단어의 의미를 명확히 하기 위해 두 가지 방법으로 연구되어왔다. 하나는 사전과 같은 지식 기반의 방법이고 다른 하나는 대량의 말뭉치 데이터에서 통계 정보를 기반한 방법이다. 1990년대 이후부터 말뭉치 통계 정보를 기반한 연구가 활발히 진행되었다. 해당 방법은 단어 의미 모호성 문제를 통계적 분류 문제로 단순화하여 머신러닝 기법을 적용하여 문제를 해결하려 했다. 단어의 의미를 명확히 하기 위해 단어 의미 태그 말뭉치(sense-tagged corpora)를 학습에 사용하는가에 따라 지도적 방법과 비지도적 방법으로 나눈다(Navigli, R. (2009)). 지도학습 방법을 이용한 단어 의미 모호성 해소 방법은 비지도 학습 방법에 비하여 좋은 성능을 나타내지만 많은 시간과 비용을 요구하고 대량의 의미 태그 말뭉치를 만들기 쉽지 않다(Kim, M.& Kwon, H.-C.(2021)).

특히, 과학 논문에는 일반적으로 특정 영역에 대해 통상적으로 사용되거나 임시로 구성되는 약어를 사용하기 때문에 논문 검색이나 정보 추출과 같은 작업으로 약어의 의미를 찾는 것이 어렵다. 일반적으로 논문에서 약어에 대한 정의를 찾을 수 있지만, 정의를 자동으로 찾는 것은 간단하지 않으며 관련 연구도 많이 진행되었다. 그러나 추출 방법이 실패하거나 논문에서 약어의 정의를 찾을 수 없는 경우가 많아 어려운 과제로 남아 있다(Charbonnier, J.&Wartena, C.(2018)). 약어의 의미를 명확히 하기 위한 최근 연구 방법으로 단어 임베딩(word embedding)의 하나인 TF-IDF(Li, C., Ji, L.&Yan, J.(2015)), 사전 훈련 모델인 BERT를 이용하거나(Zhong, Q., Zeng et al.(2021)), 비지도 학습 방법(Ciosici, M. et al.(2019))을 적용하는 등 다양한 연구가 진행되어왔다.

단어 중의성 해소 연구는 주로 단어의 의미를 명확히 하기 위한 다양한 연구가 진행되었다. 하지만, 단어 중의성 문제로 약어로 검색한 문서를 분류하는 연구는 다소

미흡하다. 약어 사전이 준비되지 않거나 시간과 비용이 충분하지 못한 경우 간단하게 문서를 분류하는 연구는 필요하다. 문서를 주제별로 분류할 수 있는 토픽 모델링 방법인 비음수 행렬 분해를 2.2에서 살펴본다.

2.2 비음수 행렬 분해(NMF; Non-Negative Matrix Factorization)를 이용한 문서 분류 연구

NMF는 행렬식을 이용하여 차원 축소의 문제점을 보완한 방법이다(Lee, D.&Seung, H. S.(2000)). 0과 양수 값으로 구성된 원 데이터를 의미특징 행렬과 의미변수 행렬로 분해한다. 분해된 의미특징 행렬과 의미변수 행렬이 0과 양수 값으로 표현되기 때문에 원 데이터를 파악하기가 쉽다(박선(2007)).

NMF 알고리즘은 원 데이터인 비음수 행렬 A를 비음수 행렬 W와 비음수 행렬 H로 분해한다. 분해된 행렬 W와 행렬 H의 행렬 곱의 값이 원 데이터인 행렬 A에 근사값을 갖도록 목적함수를 구성하여 수렴할 때까지 W와 H를 반복적으로 업데이트하여 W와 H를 계산한다(S. Wild et al.(2003), 박선(2007), Mifrah, S. & Benlahmar, E. H. (2020)).

$$A \approx WH \qquad \text{식(1)}$$

목적함수에서 유클리드 거리를 기반으로 A와 W와 H의 곱의 차이를 측정한다. 목적함수를 사용하여 W와 H의 값을 업데이트한다. 변경된 값은 병렬 연산으로 계산하여 새로운 W와 H를 사용하여 차이를 계산하고 이 과정을 원 데이터 행렬 A와 수렴할 때까지 반복한다. 최종 선택된 W와 H 행렬을 이용하여 A 행렬의 의미특징과 의미변수를 알 수 있다(S. Wild et al.(2003), 박선(2007)).

비음수 행렬 분해 방법으로 의미특징 행렬과 의미변수 행렬로 분해된 행렬을 분석하여 문서를 분류하는 문서 클러스터링 연구는 다수 진행되었다. 문서 클러스터링은 문서를 효율적으로 탐색할 수 있도록 돕는 텍스트 마이닝에서 중요한 작업이다. 문서 클러스터링의 한 방법으로 토픽 모델링은 문서가 토픽에 근접한 정도를 토픽의 가중치 조합으로 표현하여 이용되었다. NMF는 문서 클러스터링이나 토픽 모델링 방법으로 잘 사용되고 있다(Kuang, D. et al.(2015)).

문서 클러스터링에서 클러스터가 명확하게 구분되는 경우 다른 클러스터링인 K-means 방법보다 좋은 성능을 보였다(Kuang, D. et al.(2015)). 약어로 검색한 문서는 서로 명확하게 구분되는 문서가 혼합되어 다른 클러스터링 방법과 비교해 NMF 방법이 더 적합하다고 할 수 있다. 토픽 모델링에서는 NMF는 LDA(Latent Dirichlet Allocation)모델과 비교되었다(Kuang, D. et al.(2015), Mifrah, S. & Benlahmar, E. H. (2020), Latif, S. et al. (2021)). 토픽 모델은 주로 학습된 주제(topic)가 사람이 판단한 결과와 일치하는 정도에 초점을 맞추어 다양한 방법이 제안되었고, 모델에 대한 평가는 사용자나 응용프로그램의 설정에 따라 다른 결과가 나왔다. LDA는 주제의 일관성이나 주제의 핵심 단어는 잘 찾으나 주제의 분류에서는 NMF가 좋은 결과가 나왔다.

약어 검색의 혼합된 문서에서는 주제를 분류하는 것이 중요하므로 NMF가 더 적합하다고 할 수 있다. NMF 방법을 이용한 약어 검색으로 수집된 문서를 분류하는 방법을 3장에서 자세히 설명한다.

3. 약어 검색 문서 클러스터링 접근 방법

약어로 수집된 문서는 약어의 특성상 다양한 주제의 문서가 혼합되어 수집된다. 연구자는 수집된 문서에서 연구 주제에 맞는 문서를 찾아야 한다. 연구자는 연구 주제에 적합한 문서를 찾기 위해 많은 시간을 소비하고 있다(정명석 등(2017), 김효선(2021)). 다양한 주제가 혼합되어 수집된 문서 집합에서 원하는 주제의 문서를 찾는 방법으로 TF-IDF(Term Frequency-Inverse Document Frequency)와 비음수 행렬 분해(NMF; Non-Nagative Matrix Factorization)을 이용한 문서 분류 방법을 Fig. 1과 같이 제안한다.

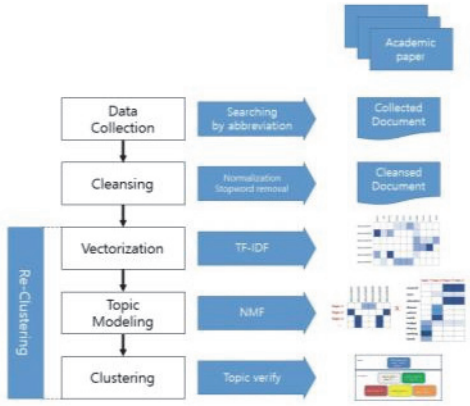

Fig. 1. Document classification method procedure using topic modeling

약어 검색어로 수집된 문서를 정규화(Normalization),

불용어 처리(Stopword removal), 토큰화(Tokenization) 처리한 정제된 문서를 만든다. 정제 문서의 단어를 TF-IDF로 벡터화하여 문서-단어로 구성된 행렬을 구성한다. NMF로 문서-단어 행렬을 문서-토픽, 토픽-단어의 행렬로 분해한다. 문서-토픽 행렬에서 문서를 토픽별로 분류하고, 토픽-단어 행렬에서 주제의 주요 단어를 파악하여 토픽이 연구 주제와 맞는 주제인지 확인한다. 연구 주제에 맞지 않은 논문을 제외하고 TF-IDF와 NMF를 반복하여 연구 주제에 맞는 논문을 찾는다. 제안 방법으로 약어 검색으로 다양한 주제가 혼합된 문서에서 원하는 주제의 문서를 분류하는 방법으로 TF-IDF와 토픽 모델링의 NMF 방법을 상세히 설명한다.

3.1 TF-IDF를 이용한 문서-단어 행렬 구성

문자로 작성된 문서를 분류하기 위해서는 문자를 숫자로 표현하는 벡터화 작업이 필요하다. 벡터화는 문서의 복잡성을 줄이고 문서의 특성을 표현할 수 있다. 문서 벡터 방법은 다양한 방법이 연구에 사용되고 있다(김지은(2017)). 그중 단어 빈도만으로 문장의 중요 단어를 분석하는 단점을 보완한 문장 내 단어의 가중치를 부여하여 단어의 중요도를 파악하는 방법으로 많이 사용되는 방법이 TF-IDF이다(Salton, G.&Buckley, C. (1988), 송민(2017), Lee, J. H. et al.(2019)).

TF-IDF는 문서 내 단어 빈도로 단어의 중요도를 측정하는 TF의 단점을 보완하여 문서 전체에 단어가 나타나는 빈도를 로그값으로 계산한 IDF의 곱으로 측정한다. 문서 내에 단어가 많을수록, 전체 문서에서 해당 단어가 포함된 문서가 적을수록 가중치 값이 커진다. TF-IDF 기법은 주제와 관련된 의미 있는 단어를 잘 파악 할 수 있다(Salton, G.&Buckley, C. (1988), Lee, J. H. et al. (2019)).

TF-IDF 과정을 진행하면 약어가 포함된 각 문서 내 단어의 가중치 값을 계산하여, 문서를 찾기 위해 사용된 검색어는 TF-IDF의 특성상 가중치 값이 작아져 분석에서 자연스럽게 사라지는 효과가 있다((Salton, G.&Buckley, C. (1988), Lee, J. H. et al. (2019)). 약어로 검색된 문서의 특성상 동일한 단어가 여러 문서에 나타나게 되는데 TF-IDF 계산과정에서 빠지게 된다. 단어의 가중치를 주제와 관련된 단어에 가중치를 높게 부여하여 약어로 검색된 문서의 의미 있는 단어를 찾아 주제의 특징을 더 잘 찾을 수 있다. TF-IDF 계산과정의 결과는 문서-단어 행렬이 되고 행렬을 이용하여 문서를 주제별로 분류하는 방법으로 토픽 모델링의 비음수 행렬 분해 방법을 적용한다.

3.2 토픽 모델링의 비음수 행렬 분해를 이용한 문서 주제 분류

약어로 검색된 문서는 약어의 의미 모호성 문제로 다양한 주제의 문서가 혼합되어 있다. 이러한 다양한 주제가 혼합된 문서 집합에서 주제별로 문서를 분류하는 방법으로 토픽 모델링의 NMF 방법이 적합하다. 문서 클러스터링 측면에서도 여러 주제로 혼합된 문서의 클러스터링에서도 NMF 방법이 좋은 결과를 보였다(Xu, W. et al. (2003), Kuang, D. et al.(2015)), Mifrah, S. & Benlahmar, E. H.(2020)).

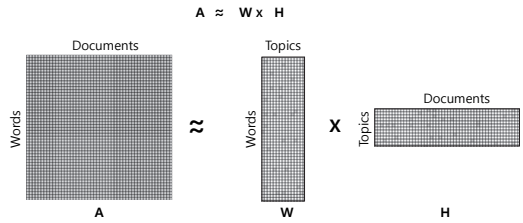

Fig. 2. Matrix decomposition of documents using NMF

Fig. 2 그림처럼 단어와 문서 행렬 A는 단어와 주제가 있는 행렬 W와 주제와 문서가 있는 행렬 H로 분해된다. 원 데이터인 비음수 행렬 A는 TF-IDF로 벡터화된 값을 넣어 비음수 행렬 W와 비음수 행렬 H로 분해한다. 비음수 행렬 A는 어떤 단어가 어떤 문서에 포함되었는지 나타내는 행렬이 된다. 비음수 의미특징 행렬 W는 주제를 구성하는 단어의 가중치를 표현한다. 비음수 의미변수 행렬 H는 문서에서 발견된 주제로 클러스터링 된다(Lee, D.&Seung, H. S.(2000)).

NMF 방법을 통하여 문서를 주제별로 클러스터링하고 각 주제에 의미 있는 단어를 찾을 수 있다. 각 주제의 가중치가 높은 단어를 확인하여 원하는 주제인지, 원하지 않은 주제인지 파악하여 주제별로 문서를 분류한다. 원하는 주제에 원하는 문서가 모두 포함되지 않을 수 있다. 원하는 문서를 최대한 많이 찾기 위해서는 문서 분류를 향상할 필요가 있다. TF-IDF와 NMF 과정을 반복하여 원하는 주제의 문서를 좀 더 많이 찾는 방법을 설명한다.

3.3 반복과정으로 문서 분류 향상

NMF를 통한 문서 분류 작업을 통하여 원하는 주제의 문서를 선별한다. 명확하게 제외할 주제와 원하는 주제를 알 수도 있지만 혼합되어 모호한 주제도 있을 수 있다. 확실하게 제외할 주제의 문서는 제외하고 클러스터링 과정을 반복한다. 즉, 명확하게 원하는 주제와 모호한 주제

의 문서를 새로운 수집된 문서 집합으로 선택하고 문서-단어 행렬을 구성하고, 토픽 모델링을 반복하여 진행한다.

반복 작업에서는 불용어 처리 등 정제작업은 이전 작업 결과를 그대로 사용한다. TF-IDF와 NMF 과정을 반복 진행하여 클러스터링 주제에 모호한 주제가 없도록 반복하여 진행한다.

약어로 검색된 문서 전체를 수작업으로 읽고 분류하는 것보다 토픽 모델링 작업으로 확인된 주제의 중요 단어를 확인하여 원하는 주제의 문서를 분류하는 것이 소요 시간과 노력을 줄 일 수 있다. 제안 방법으로 원하는 주제의 문서를 분류할 수 있는지 확인하기 위하여 4장에서 사례를 통하여 제안된 방법으로 찾은 문서 결과와 수작업 찾은 문서 결과를 비교하여 확인해본다.

4. 약어 검색 문서 분류 방법 사례: MSA 중심

약어로 검색한 문서를 수집하여 제안된 방법으로 문서를 분류하여 토픽 모델링을 진행한 결과와 수작업으로 분류한 문서에서 토픽 모델링을 진행한 결과를 비교해 보자고 한다. 검색할 약어는 'Microservices Architecture'의 'MSA'로 검색하여 수집한다. 위키피디아에 약어 'MSA'는 기술 및 과학 분야에서 측정 시스템 분석(Measurement systems analysis), 마이크로스트립 안테나(Microstrip Antenna), 다계통위축(Multiple system atrophy), 현대 표준 아랍어(Modern Standard Arabic)등 다양한 부분에 참조되고 있다. 마이크로 서비스 아키텍처는 애플리케이션을 확장이 용이하고 민첩하게 개발하는 방법으로 IT 기업에서 클라우드 환경에서 많이 적용하고 있어 'MSA'를 적용사례로 선정했다. 약어 검색 및 진행 과정의 개념도는 그림 Fig. 3과 같다.

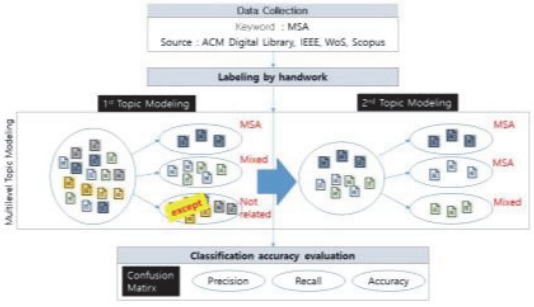

Fig. 3. How to classify documents for abbreviation search using topic modeling

4.1 약어 검색으로 문서 수집 및 전처리 작업

약어 논문 검색을 위하여 해외 학술 데이터베이스 중 ACM Digital Library, IEEE Xplorer, Web of Science, Scopus를 선정하였다. 논문 검색을 위한 검색어는 "MSA" OR "micro service*" OR "micro-service*"로 입력했다. 2009년부터 Microservices Architecture 개념이 나온 것을 고려하여 검색 기간은 2009년부터 2021년으로 지정했다. 수집된 논문은 ACM Digital Library 15편, IEEE Explorer 414편, Web Of Science 567편, Scopus 405편으로 총 1,401건이 수집되었다.

수작업 분류와 토픽 모델링 분류의 비교를 위하여 수집된 논문 1,401편에 대한 라벨링 작업을 진행했다. 수집된 논문의 제목과 요약문을 모두 읽어 Microservices Architecture 관련 여부를 확인했다. "Microservices Architecture"이라는 단어가 있어도 논문이 직접적으로 관련이 없다면 동료 연구자와 의논하여 Microservices Architecture 관련 여부에서 제외하였다. 예를 들어 사례로 언급되거나, 다른 아키텍처와 비교하기 위하여 사용되는 경우는 제외하였다. 최종 수작업으로 분류한 Microservices Architecture 관련 논문은 1,401편 중 376편(27%)을 선정하였다.

수집된 논문의 요약문에 불용어 처리, 대소문자 통합 등의 정제작업을 진행했다. 정제된 요약문을 이용하여 토픽 모델링을 통한 분류 작업을 2회 반복하여 진행하였다. 분류 작업 과정을 4.2에서 자세히 살펴본다.

4.2 TF-IDF와 NMF을 이용한 문서 분류

정제된 요약문을 TF-IDF를 수행하고, 토픽 모델링을 위한 토픽 개수 선정 작업 후 비음수 행렬 분해를 이용한 토픽 모델링을 진행하여 각 토픽의 주요 키워드를 확인하여 분류 작업을 진행했다. 분류 작업은 1차, 2차로 2회 반복하여 진행했다. 토픽 개수 선정은 파이썬의 Gensim 패키지를 활용해 perplexity와 coherence를 계산했다. 토픽 수를 2에서부터 50까지 토픽 모델링을 반복 시행하면서 일관성 점수를 계산하여 최소의 값이 나온 것을 선정했다. 일관성 점수 측정값이 0에 가까울수록 잘 모델링된 것으로 판단할 수 있다(Mimno et al.(2011)). 1차, 2차의 계산된 일관성 점수는 Fig. 4와 같고, 최적 토픽 수는 1차, 2차 각각 23개와 24개로 선정되었다.

1차 분류 작업은 정제된 요약문을 대상으로 토픽수 23개로 설정하고 사이킷런 패키지를 활용한 비음수 행렬 분해을 진행했다. 토픽별로 Microservice Architecure 관련 토픽, 제외할 토픽, 혼합된 토픽, Microservice Architecture

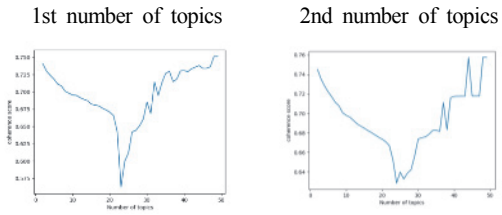

Fig. 4. Calculation of the appropriate number of topics

관련 없는 토픽으로 분류하기 위하여 TF-IDF 값이 큰 상위 10개 단어로 각 토픽의 의미를 파악했다.

토픽별 논문 수와 Microservice Architecture와 관련 여부는 Table 1과 같다. 관련된 토픽은 Topic2, Topic14, Topic22로 분류되어 해당 논문 185편을 찾았다.

Table 1. 1st Clustering Result

Topic	Number of papers	Result
Topic1	243	not related
Topic2	51	related
Topic3	90	not related
Topic4	66	not related
Topic5	54	not related
Topic6	65	not related
Topic7	37	mixed
Topic8	25	not related
Topic9	72	mixed
Topic10	15	mixed
Topic11	40	not related
Topic12	41	not related
Topic13	96	mixed
Topic14	55	related
Topic15	23	not related
Topic16	16	mixed
Topic17	8	not related
Topic18	64	not related
Topic19	23	not related
Topic20	62	mixed
Topic21	33	not related
Topic22	79	related
Topic23	143	mixed
Sum	1,401	

1차 분류 작업의 결과 중 Microservice Architecture와 관련이 없는 토픽에 해당하는 논문을 대상 목록에서 제외하였다. 1,401편의 논문 중 1차 분류 작업에서 not related로 분류된 775편의 논문을 2차 분류 작업에서 제외하고 진행하였다. 혼합된 토픽과 Microservice Architecture 관련 토픽에 해당한 논문 626편을 대상으로 2차 분류 작업을 진행하였다.

2차 분류 작업도 벡터화, 토픽 모델링, 클러스터링 작업을 동일한 방법으로 진행했다. 최적의 토픽 수를 24개로 설정하고 비음수 행렬 분해를 진행해 토픽별로 상위 10단어로 토픽을 분류하였다. “Microservices Architecture”와 관련된 토픽은 Topic1, 3, 4, 5, 6, 7, 8, 11, 12, 14, 15, 17, 22, 23으로 해당하는 논문 363편을 찾았다. 혼합된 토픽은 제외하였다. 2차 분류 작업의 결과는 Table 2와 같다.

Table 2. 2nd Clustering Result

Topic	Number of papers	Result
Topic1	18	related
Topic2	66	not related
Topic3	29	related
Topic4	13	related
Topic5	5	related
Topic6	45	related
Topic7	55	related
Topic8	61	related
Topic9	9	not related
Topic10	23	not related
Topic11	9	related
Topic12	4	related
Topic13	13	not related
Topic14	18	related
Topic15	5	related
Topic16	39	mixed
Topic17	28	related
Topic18	5	mixed
Topic19	30	not related
Topic20	25	mixed
Topic21	30	not related
Topic22	33	related
Topic23	40	related
Topic24	23	not related
Sum	626	

제안된 방법으로 Microservices Architecture 관련 논문을 1차 분류에서 185편, 2차 분류에서 363편을 찾았다. 수작업으로 분류한 376편과 제안된 방법으로 분류한 결과를 혼동 매트릭스(Confusion Matrix)와 토픽 모델링을

진행하여 비교 분석한다. 4.3에서 비교 결과를 설명한다.

4.3 약어 검색 문서의 분류 결과 비교

MSA 약어 검색을 통해 수집된 논문의 분류 결과가 적합한지 확인하기 위하여 혼동 매트릭스(Confusion Matrix)를 활용한 평가와 실제 토픽 모델링을 진행하여 선정된 토픽을 분석하였다. 분류 결과를 평가하기 위해 수작업으로 분류한 결과와 제안된 방법으로 분류한 결과를 혼동 매트릭스(Confusion Matrix)에 대입하여 평가 지표를 확인하였다. 평가 지표로 정밀도(precision), 재현율((recall), 정확도(accuracy), F1-Score를 계산했다(Kang, S. et al. (2021)). 분류 결과 분석을 위한 다른 방법으로 수작업과 제안된 방법으로 분류한 결과를 토픽 모델링 진행하여 분류된 논문에서 주제 영역을 도출하여 비교했다. 4.3.1에서 혼동 매트릭스 평가 지표를 살펴보고 4.3.2에서 토픽 모델링 주제 영역을 비교한다.

4.3.1 혼동 매트릭스를 이용한 결과 평가

1차 2차 분류 작업의 결과는 혼동 매트릭스(Confusion Matrix)를 이용하여 Table 3과 Table 4와 같이 정리되었다.

Table 3. 1st Topic Modeling Confusion Matrix

		Predicted		Sum
		True	False	
Actual	Positive	171	205	376
	Negative	14	1,011	1,025
Sum		185	1,216	1,401

정밀도(Precision) = 171 / (171 + 14) = 92.43%
재현율(Recall) = 171 / (171 + 205) = 45.48%
정확도(Accuracy) = (171+1,011) / (171+205+14+1,011)
= 84.37%
F1-Score = 0.6096

Table 4. 2nd Topic Modeling Confusion Matrix

		Predicted		Sum
		True	False	
Actual	Positive	316	60	376
	Negative	47	978	1,025
Sum		363	1,038	1,401

정밀도(Precision) = 316 / (316 + 47) = 87.05%
재현율(Recall) = 316 / (316 + 60) = 84.04%
정확도(Accuracy) = (316 + 978) / (316 + 47 + 60 + 978)
= 92.36%
F1-Score = 0.8552

토픽 모델링으로 분류한 결과와 수작업 분류한 결과가 얼마나 같은지 측정한 정밀도는 92.4%로 계산되었다. 수작업 분류가 관련된 논문으로 분류한 것 중 토픽 모델링으로 분류한 결과가 관련 논문으로 분류한 재현율은 45.48%로 계산되었다. 전체 논문 중 수작업 분류 대비 토픽 모델링이 분류한 결과가 일치하는 정도인 정확도는 84.3%로 계산되었다. 2차 토픽 모델링의 분류 결과 정밀도는 87.05%, 재현율은 84.04%, 정확도는 92.36%로 계산되었다.

Table 5. Result of Confusion Matrix

Value	1st.	2nd	Change
Recall(%)	45.48	84.04	38.56
Accuacy(%)	84.37	92.36	7.99
Precision(%)	92.43	87.05	-5.38
F1-Score	0.6096	0.8552	0.2456
Success(Paper)	171	316	145

1, 2차 토픽 모델링 결과 비교는 Table 5와 같다. 재현율은 1차와 비교하여 38.56% 증가했으며, 정확도는 1차와 비교하여 7.99% 증가하였다. 정밀도는 -5.38% 감소하였으나 F1-Score는 0.2456 증가하였다. 수작업 논문과 일치하는 Microservices Architecture와 관련된 논문이 171편에서 316편으로 145편 증가했다.

재현율이 증가한 것은 1차에서 찾지 못한 논문을 2차에서 추가로 찾았기 때문이다. 1차 분류 작업에서 혼합된 토픽을 제외하고 2차 분류 작업에서 혼합된 토픽에서 관련 논문을 추가로 찾아 재현율이 증가했다. 정밀도가 감소한 것도 1차에 제외되었던 혼합된 토픽이 2차 분류 작업에서 관련 없는 논문이 포함되어 정밀도가 감소했다. 비교를 위한 수작업 분류과정에서도 “Microservices Architecure”라는 단어가 있어도 관련 없는 논문을 제외했기 때문에 토픽 모델링으로 분류하는 작업의 정밀도가 감소했다.

반복하여 진행된 토픽 모델링 분류 결과 관련된 논문 편수도 증가하였고 정확도 평가 지표도 양호한 결과를 보였다. 약어 검색 논문의 분류에 반복 진행하는 제안 방법이 효과가 있음을 알 수 있었다.

4.3.2 토픽 모델링의 주제 영역 비교

수작업으로 분류한 Microservices Architecutre와 관련된 논문 376편과 제안된 방법으로 분류한 논문 316편

에 대하여 토픽 모델링을 진행하여 결과를 비교하였다. 적정 토픽 수를 일관성 점수를 통하여 계산하였다. 결과는 Fig. 5와 같다.

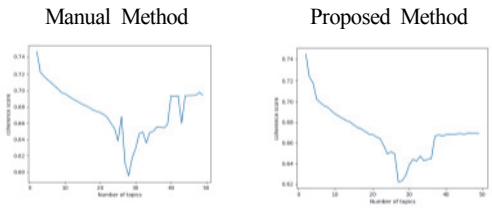

Fig. 5. Calculation of the appropriate number of topics by method

적정 토픽 수는 수작업 분류는 28개, 제안된 방법의 분류는 27개로 선정되어 1개의 차이가 났다. 분류 방법별로 선정된 토픽 수를 설정하여 토픽 모델링을 진행하였다.

Table 6. Topic by method

Manual Method	Proposed Method
iot	iot
business	business
software	software
big data	big data
cloud	cloud
digital	digital
platform	platform
auto scaling	auto scaling
container	container
docker	-
-	agile
workflow	workflow
grid power	-
SOA	SOA

두 방법으로 분류한 논문을 토픽 모델링을 진행하여 각 토픽에 속하는 논문이 20개 이상인 토픽만을 선택했다. 선택된 토픽의 주요 키워드를 확인하여 각 토픽의 주제를 동료 연구자와 협력하여 Table 6과 같이 정했다. 수작업으로 분류한 논문에서 선정된 키워드 중 'docker', 'grid power'가 제안된 방법에서는 선정되지 않았고, 반대로 제안된 방법으로 분류한 논문에서 선정된 키워드 중 'agile'이 수작업 방법에서는 선정되지 않았다. 선정된 토픽이 크게 차이가 없음을 알 수 있다.

5. 결론

약어를 이용하여 문서를 검색하는 경우 약어의 의미 중의성 문제로 다양한 주제가 혼합된 문서를 수집하게 된다. 수집된 문서를 수작업으로 분류하는 시간과 노력을 줄이는 방법을 제안했다. TF-IDF, 비음수 행렬 분해, 클러스터링하는 3단계 분류과정을 반복 수행하는 방법이다. 분류 방법을 검증하기 위하여 'MSA' 단어로 검색해 문서를 수집하고 Microservices Architecture와 관련된 주제의 문서를 찾아가는 과정을 사례로 진행했다. 제안된 방법으로 분류한 결과와 수작업으로 분류한 결과를 비교하여 검증했다. 사례의 분류 결과 반복과정을 통하여 재현율과 정확도 지표가 개선되었음을 알 수 있다. 또한, 토픽 모델링을 진행한 주제 영역이 제안된 방법과 수작업 분류 간에 유사한 결과를 나타냈다.

본 연구에서 제안된 방법이 약어로 검색된 문서를 완벽하게 분류하는 것에는 다소 부족하나 다수의 연구에서 대상 문서를 분류하기 위해 일일이 수작업으로 분류하는 노력을 줄일 수 있다는 점에서 연구의 의미가 있다. 분류 과정을 반복 수행하여 분류 결과가 향상됨을 알 수 있다. 향후 다양한 연구에서 약어로 검색된 문서를 분류하는 수작업의 수고를 줄일 수 있기를 기대한다.

본 연구에서는 TF-IDF와 비음수 행렬 분해로 진행하였으나 관련 분야에 다양한 방법이 연구되고 있어 향후 여러 방법을 조합하여 개선할 필요가 있다. 반복 회수에 따른 성능 변화 확인도 필요하다. 사례 검증 면에서 제안된 방법이 본 연구의 사례 검증만으로 일반화하기 부족하며 다양한 사례 연구를 통한 추가 검증이 필요하고 영문 문서 외에 한글 문서 사례 연구도 필요하다.

References

Bevilacqua, M., Pasini, T., Raganato, A., and Navigli, R. (2021), Recent trends in word sense disambiguation: A survey, *In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conference on Artificial Intelligence, Inc.*

Charbonnier, J., and Wartena, C. (2018), Using word embeddings for unsupervised acronym disambiguation, *In Proceedings of the 27th International Conference on Computational Linguistics,* pp. 2610-2619.

Chung, M. S., Park, S. H., Chae, B. H., and Lee, J. y.

(2017), Analysis of major research trends in artificial intelligence through analysis of thesis data, *Journal of Digital Convergence*, 15(5), 225-233.
(정명석, 박성현, 채병훈, 이주연 (2017), 논문데이터 분석을 통한 인공지능 분야 주요 연구 동향 분석, *디지털융복합연구*, 15(5), pp. 225-233.)

Ciosici, M., Sommer, T., and Assent, I. (2019), Unsupervised Abbreviation Disambiguation Contextual disambiguation using word embeddings, *arXiv preprint arXiv:1904.00929*.

Kang, S., Chang, S. R. and Suh, Y. (2021), Machine Learning Approach to Classifying Fatal and Non-Fatal Accidents in Industries, *Journal of the Korean Society of Safety,* 36(5), pp. 52-60.

Kim, H. (2021), Analysis of the Research Trends on Business Archives: Focusing on the Topic Modeling Analysis, *Journal of Korean Society of Archives and Records Management*, 21(3), 163-186.
(김효선 (2021), 기업 아카이브에 관한 연구 동향 분석: 토픽모델링 분석을 중심으로, *한국기록관리학회지,* 21 (3), pp. 163-186.)

Kim, J. (2017), Keyword and topic analysis on the college and university structural reform evaluation using big data, *Ph.D.Thesis, Seoul National University.*
(김지은 (2017), 빅데이터를 활용한 대학구조개혁 평가의 키워드 및 토픽분석, *서울대학교대학원 박사학위논문*.)

Kim, M. and Kwon, H.-C. (2021), Word Sense Disambiguation Using Prior Probability Estimation Based on the Korean WordNet, *Electronics,* 10, 2938. https://doi.org/10.3390/electronics10232938

Kuang, D., Choo, J. and Park, H. (2015), Nonnegative matrix factorization for interactive topic modeling and document clustering, *In Partitional clustering algorithms, Springer, Cham,* pp. 215-243.

Lee, D. and Seung, H. S. (2000), Algorithms for non-negative matrix factorization, *Advances in neural information processing systems,* 13.

Lee, J. H., Lee, M., and Kim, J. W. (2019). A study on Korean language processing using TF-IDF, *The Journal of Information Systems,* 28(3), pp. 105-121.

Latif, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access,* 9, pp. 127531-127547.

Li, C., Ji, L., and Yan, J. (2015), Acronym disambiguation using word embedding, *In Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 29, No. 1.

Mifrah, S. and Benlahmar, E. H. (2020), Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus, *International Journal of Advanced Trends in Computer Science and Engineering,* pp. 5756-5761.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011), Optimizing semantic coherence in topic models, *In Proceedings of the 2011 conference on empirical methods in natural language processing,* pp. 262-272.

Na, S. T., Kim, J. H., Jung, M. H., and Ahn, J. E. (2016), Trend Analysis using Topic Modeling for Simulation Studies, *Journal of The Korea Society for Simulation*, Vol. 25, No. 3, pp. 107-116.
(나상태, 김자희, 안주언, 정민호 (2016), 토픽 모델링을 이용한 시뮬레이션 연구 동향 분석, *한국시뮬레이션학회 논문지,* 제25권, 제3호, pp. 107-116.)

Navigli, R. (2009), Word sense disambiguation: A survey, *ACM computing surveys (CSUR),* 41(2), pp. 1-69.

Park, S. (2007), Automatic Document Summarization Using Non-negative Matrix Factorization, *Ph.D.Thesis, Inha University.*
(박선 (2007). 비음수 행렬 분해를 이용한 자동 문서 요약, *인하대학교 박사학위논문*.)

Salton, G. and Buckley, C. (1988), Term-weighting approaches in automatic text retrieval, *Information processing & management,* 24(5), pp. 513-523.

Song, M. (2017), Text Mining. *Seoul: Chungram.*
(송민 (2017), 텍스트 마이닝, *서울: 청람.)*

S. Wild, J. Curry and A. Dougherty (2003), Motivating Non Negative Matrix Factorizations, *In proceedings of Society for Industrial and Applied Mathematics Conference on American Library Association (ALA'03).*

Xu, W., Liu, X. and Gong, Y. (2003), Document clustering based on non-negative matrix factorization, *In Proceedings of the 26th annual international ACM*

SIGIR conference on Research and development in informaion retrieval, pp. 267-273.

Zhong, Q., Zeng, G., Zhu, D., Zhang, Y., Lin, W., Chen, B., and Tang, J. (2021), Leveraging Domain Agnostic and Specific Knowledge for Acronym Disambiguation, *arXiv preprint arXiv:2107.00316.*

이 운 교 (ORCID : https://orcid.org/0009-0002-4594-9247 / johntato@seoultech.ac.kr)

1997 가톨릭관동대학교 전자계산공학과 학사
2018 서울과학기술대학교 IT정책전문대학원 석사
2021~ 현재 서울과학기술대학교 IT정책전문대학원 박사과정
2007~ 현재 ㈜한국금융아이티 본부장

관심분야 : 요구공학, 텍스트마이닝, 문서분류

김 자 희 (ORCID : https://orcid.org/0000-0002-6863-0821 / jahee@seoultech.ac.kr)

1995 KAIST 전산학과 학사
1997 KAIST 전산학과 석사
2003 KAIST 산업공학과 박사
2005~ 현재 서울과학기술대학교 IT정책전문대학원 교수

관심분야: 요구공학, 반도체 제조 스케줄, 스마트 그리드

양 준 기 (ORCID : https://orcid.org/0009-0009-1050-4513 / javamania@gmail.com)

2021 서울과학기술대학교 IT정책전문대학원 석사
2014~ 현재 SK㈜C&C Cloud Transformation 그룹 팀장

관심분야 : 요구공학, 마이크로 서비스 아키텍처, 텍스트마이닝