그래프 합성 신경망 기반 연구 논문 분류 시스템의 설계 및 구현

딥또 비스와스[1], 변태영[2], 길준민[3,*]

[1]대구가톨릭대학교 컴퓨터소프트웨어학과, [2]컴퓨터소프트웨어학부, [3]제주대학교 컴퓨터공학과

dipto.biswas94@gmail.com, tybyun@cu.ac.kr, *jmgil@jejunu.ac.kr

Design and Implementation of Research Paper Classification based on Graph Convolutional Network

Dipto Biswas[1], Byun Tae-Young[2], Gil Joon-Min[3,*]

[1]Dept. of Computer Software Eng., [2]School of Computer Software Eng., Daegu Catholic Univ., [3]Dept. of Computer Engineering, Jeju National Univ.

요 약

텍스트 분류는 자연어 처리 분야에서 중요한 문제로, 고전적인 연구 주제 중에 하나이다. 현재, 다양한 딥러닝 모델들이 텍스트 분류에 사용되고 있다. 그 중에서도 그래프 합성곱 신경망(GCN)은 텍스트를 좀 더 유연하게 분류하며, 데이터를 그래프로 처리하여 비순차적인 관계들로 이루어진 데이터에서 효과적으로 특징을 추출할 수 있다. 이 논문에서는 텍스트 그래프 합성곱 신경망(TextGCN)이 문서 내 단어 관계를 파악할 수 있다는 장점을 활용하여 연구 논문 데이터에 TextGCN을 적용한 연구 논문 분류 시스템을 설계하고 구현한다. 이를 위해 BERT와 Word2Vec의 임베딩 특징 벡터를 TextGCN에 적용하여 모델의 테스트 정확도를 평가하고 다양한 유형의 F1-Score 방법(F1-Macro, F1-Micro, F1-Weighted)을 사용하여 상위 N개 단어를 기반으로 비교·분석한다. 실험 결과, BERT 기반 TextGCN 모델이 기존의 Word2Vec 기반 TextGCN 모델보다 연구 논문의 분류 및 추천에 있어서 우수한 성능을 가짐을 보인다.

Ⅰ. 서 론

텍스트 분류는 기사 주제 분류, 스팸 필터링, 감정 분석과 같은 다양한 응용에 사용되며 자연어 처리(NLP: Natural Language Processing)에서 중요한 작업이다[1]. 전통적으로 텍스트 분류 작업에는 합성곱 신경망(CNN)[2]이 사용되었지만, 최근에는 그래프 합성 신경망(GCN)[3]이 주목을 받고 있다. GCN은 데이터를 그래프 형태로 처리하는 방법을 제공하며, 단어를 노드로, 단어 간의 관계는 에지로 표현한다. 이 구조에 의해서 GCN은 복잡한 데이터 연결을 가지는 기법보다 데이터를 더 효과적으로 처리할 수 있다. 따라서, GCN은 텍스트 분류뿐만 아니라 화학적 특성 예측, 네트워크 분석 등 다양한 분야에도 유용하게 사용되고 있다. GCN의 여러 가지 모델 중에 TextGCN[3] 모델은 그래프에서 문단과 단어를 모두 노드로 처리하며 텍스트 분류에 GCN을 적용한다.

본 연구의 목적은 연구 논문을 효과적으로 분류하기 위해 사용자 선호도에 맞게 맞춤형 연구 논문을 추천하는 시스템을 설계하는 것이다. 이를 위해 BERT 모델의 특징 벡터를 사용하여 TextGCN 모델을 학습시켜 연구 논문을 분류하고 추천한다.

Ⅱ. 연구 배경 및 시스템 모델

제안 시스템은 다음과 같이 데이터 수집 단계으로 시작하여 모델 학습 및 평가 단계까지 단계별로 수행한다.

• **데이터 수집 및 전처리**: 먼저 Selenium과 BeautifulSoup를 사용하여 Science Direct 웹사이트에서 2024년에 출판된 논문지의 데이터를 웹 스크래핑을 통해 수집한다. "Computer Science(CS-24)"과 "Social Science(SS-24)"와 관련된 다양한 저널의 연구 논문 초록을 수집하여 데이터 세트를 구성한다. 전처리를 통해 텍스트를 정리하고 NLTK(Natural Language Toolkit)를 사용하여 불필요한 단어, 구두점, 숫자, URL 및 웹사이트 링크 등을 제거한다. 또한 동사와 부사는 해당 단어를 명사로 변환한다. 표 1은 전처리를 수행한 후 데이터 세트의 전반적인 통계를 보여준다.

• **그래프 구성 및 GCN 모델**: GCN 모델의 경우, 사전 처리된 데이터가 문서와 단어가 노드로 처리될 수 있도록 설계된다. BERT 임베딩을 사용해서 초록의 각 단어에 대한 특징 벡터를 생성함으로써 단어 간의 의미적 및 맥락적 관계를 포착한다. 이러한 임베딩은 그래프 노드의 초기 특징 벡터로 사용된다. 단어 동시 발생 및 TF-IDF(Term Frequency-Inverse Document Frequency) 값을 기반으로 에지를 생성하며 텍스트 내 단어 간의 관계와 중요성을 반영하여 에지 값을 설정한다. 이런 방식으로 텍스트 그래프를 구성한 후 TextGCN 모델은 GCN 모델을 사용하여 텍스트 그래프 구조를 처리할 때 그래프 내의 종속성과 상호작용을 효과적으로 포착할 수 있도록 설계한다.

• **임베딩 및 TextGCN 모델 설계**: TextGCN 모델의 경우, 첫 번째 합성곱 계층의 임베딩 크기를 200으로 하고, 슬라이딩 윈도우의 크기를 20으로 설정한다. 학습률은 0.01, 드롭아웃률은 0.2, 손실 가중치는 0으로 설정한다. 학습 세트 중 10%를 무작위로 선택하여 검증 세트로 사용한다. Adam 옵티마이저를 사용하여 최대 150개 에포크 동안 TextGCN 모델을 학습하였고, 15개 연속 에포크 동안 검증 손실이 변경하지 않으면 학습을 중단한다.

TextGCN 모델의 최종 계층은 2개로 구성한다. 또한 TextGCN 모델의 특징을 추출하기 위해 BERT 임베딩 모델을 사용한다. BERT 임베딩 모델에는 사전학습된 'bert-base-uncased'를 적용한다. 토큰 수는 512로 설정하고, 임베딩 차원 수는 782로 설정하였다. 반면, Word2Vec 임베딩의

표 1. 데이터 세트의 전반적인 데이터 통계.

데이터 세트	#문서	#학습 데이터	#테스트 데이터	#단어	#노드	문서의 길이 (평균)
CS-24	5,857	4,685	1,172	10,206	16,063	1334.66
SS-24	5,111	4,088	1,023	12,047	17,158	1565.75

경우에는 300차원을 사용하고, 윈도우 크기는 5로 설정하였다. 최소 단어 수는 50으로 설정하고, worker 수는 4로 설정하였다. Word2Vec 차원을 명확하게 활용하기 위해 CBOW는 0으로, Sg를 1로 설정하였다.

• **모델 학습 및 평가**: TextGCN 모델은 전처리된 그래프 데이터를 사용하여 학습한다. 학습을 위한 타겟 값은 데이터 세트에서 가장 자주 발생하는 상위 15(Top N=15)개의 단어를 사용하였으며, 이렇게 선택된 타겟 값이 GCN 모델을 학습하는데 사용된다. 모델의 성능은 분류 정확도와 여러 유형의 F1-Score 방법(F1-Macro, F1-Micro, F1-Weighted)을 사용하여 평가되었다. F1-Macro 방법은 모든 클래스를 동일한 중요도로 평가하므로, 샘플 수의 불균형이 평가에 큰 영향을 줄 수 있다. F1-Micro 방법은 전체 모델의 성능을 평가하는데 초점이 맞추어져 있으며, 클래스 간 샘플 수의 불균형을 고려한다. F1-Weighted 방법은 각 클래스의 중요도를샘플 수에 따라 가중치를 두어 평가하여, 불균형한 데이터 세트에서도 각 클래스의 영향력을 조정하여 반영할 수 있다.

최종 예측을 위해 상위 N개 단어를 평가하여 분류 모델의 분류 정확도 측면에서 평가한다. 예측 단계에서는 상위 N개 레이블에서 계산된 가장 높은 예측 점수를 기반으로 사용자의 검색과 관련된 연구 논문을 식별한다. 또한, 각 논문과 관련된 주요 단어를 표시하여 연구 논문을 추천하는 데 도움이 되도록 한다.

Ⅲ. 실험 결과

표 2는 두 가지 데이터 세트(CS-24와 SS-24) 관점에서 세 가지 모델의 정확도를 보여준다. 표 2의 결과를 살펴보면, BERT 기반 TextGCN 모델은 다른 두 모델(TextGCN-CBOW과 TextGCN-Sg) 보다 우수한 결과를 나타낸다. 이는 TextGCN 모델이 그래프 구조를 활용하여 단어 간의 복잡한 관계와 종속성을 포착하는 능력 덕분에 텍스트 내에 존재하는 맥락 정보를 효과적으로 모델링할 수 있기 때문이다. 반면, TextGCN-CBOW와 TextGCN-Sg 모델은 TextGCN-BERT 다소 낮은 성능을 보여준다.

표 2. 모델의 테스트 정확도.

모델	CS-24(%)	SS-24(%)
TextGCN-CBOW	72	61
TextGCN-Sg	77	71
TextGCN-BERT	93	91

그림 1은 각 모델의 F1-Score 성능 측정 결과를 나타낸다. 이 그림에서 볼 수 있듯이, TextGCN-BERT 모델은 다양한 F1-Score(F1-Macro, F1-Micro, F1-Weighted) 측면에서 우수한 성능을 보여준다. 또한 전체 결과를 분석해 보면, SS-24 보다 CS-24 데이터 세트에서 TextGCN-BERT 모델이 다른 두 모델에 비해 좋은 성능을 가짐을 확인할 수 있다.

Ⅳ. 결론

본 논문에서는 연구 논문 분류를 위해 BERT 기반 TextGCN 모델을 설계 및 구현하고 기존의 Word2Vec 기반 TextGCN 모델과 성능 비교를 제

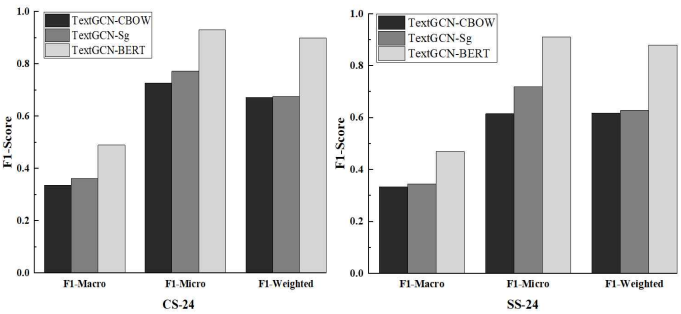

그림 1. 두 가지 데이터 세트에 대한 각 모델의 F1-Score 결과

시하였다. 우선, 전체 연구 논문의 초록 데이터에 대한 이기종 단어 문서 그래프를 구축하고 연구 논문 분류를 노드 분류 문제로 나타내었다. TextGCN은 단어 동시 발생 정보를 포착하고 상위 N개 레이블이 지정된 문서를 잘 활용할 수 있다. 2계층 TextGCN 모델은 두 개의 벤치마크 데이터 세트에 대해서 다른 두 모델에 비해 우수한 성능을 도출하였다. 따라서 BERT 기반 TextGCN 모델이 연구 논문 분류에 더 적합함을 알 수 있었다.

한편, 본 연구의 향후 연구는 새로운 유형의 텍스트 그래프 프레임워크을 개발하고 그래프 어텐션 네트워크(GAT), 그래프 합성 순환 신경망(GCRN), 그리고 일반 회귀 신경망(GRNN)을 사용하여 연구 논문 분류 시스템의 성능을 개선하는 것이다.

ACKNOWLEDGMENT

본 결과물은 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신사업의 결과입니다(2023RIS-009). 그리고 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022R1A2C1092934).

참 고 문 헌

[1] Ghiassi, M., Sean L., and Swati, R. G. "Sentiment analysis and spam filtering using the YAC2 clustering algorithm with transferability." Computers & Industrial Engineering, vol. 165 p.107959, 2022.

[2] Kim, D. "Text Classification Based on Neural Network Fusion." Tehnički glasnik vol. 17, no. 3, pp. 359-366, 2023.

[3] Yao, L., Chengsheng M., and Yuan L. "Graph convolutional networks for text classification." In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 7370-7377. 2019.