단어의 의미와 순서를 고려하는 문서색인방법을 이용한 CNN 기반 한글문서분류

김남훈† · 양형정† 1)
† 전남대학교 전자컴퓨터공학대학원

Classification of Korean Documents Based on CNN Using Document Indexing Method based on Word Meaning and Order

Nam-Hun Kim† · Hyung-Jeong Yang†
† Dept. of Computer Science, Chonnam National University

요 약

본 논문에서는 컨볼루션 신경망 네트워크(CNN:Convolution Neural Network)을 기반으로 단어의 의미와 순서를 고려하는 문서 색인 방법을 이용하여 한글 문서 분류 방법을 제안한다. 먼저 문서를 형태소 분석하여 어절 단위로 분리 한 후, 불용어를 처리 하고, 문서의 단어 의미를 고려하는 문서 표현하고, 문서의 단어 순서까지 고려하여 CNN의 입력으로 사용하였다. 실험결과 CNN 분류기를 기반으로 본 논문에서 제안하는 문서 색인 방법은 TF-IDF를 이용하는 방법보다 4.2%, Word2vec만 단독으로 사용하는 것보다 1.4%의 성능 상승을 이루었다. 이러한 결과를 통해 본 논문에서 제안하는 방법이 문서 범주화 데이터 셋에서 문서 분류 성능향상에 영향을 미친다는 것을 확인하였다.

1. 서 론

많은 정형 혹은 비정형 데이터들이 인터넷, 뉴스, 웹 페이지 등을 통해 텍스트 형태로 생산이 되고 있다. 그러나 이러한 데이터들은 대부분 정형화 되어 있지 않고, 주어진 형식도 없이 무작위로 매우 빠르게 데이터가 증가하고 있다. 그러나 이 데이터들이 어떤 종류의 데이터이고 어떤 분야에 속하는지에 대한 정보를 가져야 사용자에게 제대로 의미 있는 데이터가 될 수 있으므로, 이것을 파악하는 것은 중요해지고 있다[1].

먼저 문서를 컴퓨터가 이해하는 언어로 바꾸려면 색인이라는 작업을 해야 한다. 색인은 해당 단어가 문서에서 어느 정도의 가중치를 가지고 있는지를 기준으로 우선순위를 부여한다. 대표적으로 벡터공간모델인 TF-IDF(Term Frequency-Inverse Document Frequency)는 문서에 등장하는 단어들의 중요도를 나타내는 특성 값을 사용하여 문서를 벡터형태로 바꿔준다[6]. 이런 단어의 출현률을 기반으로 문서의 특징벡터를 추출하고 특징 벡터들을 문서 분류에 사용하는 방법이 가진 문제점은 문서에 포함된 단어들만이 문서의 특징으로 영향을 미친다는 것이다. 이러한 TF-IDF 벡터공관모델은 문맥과 의미를 고려하지 않기 때문에 성능에 한계가 있음을 알 수 있다.

[2]의 연구에서는 이러한 단점을 보완하기 위해 문서 내의 단어 간 유사도를 사용하여 출현하지 않은 단어의 특징 값을 간접적으로 평가하고 같은 문서 내에서 출현한 유사한 단어들은 가중치를 주는 방법을 이용하였다. 이 방법은 단어가 가지는 의미를 고려할 수 있는 장점이 있다.

그러나 [2]의 연구는 문서에서 나타나는 한 단어와 그 주변단어들을 함께 학습하여 단어와 단어의 관계만을 모델링 하는 지역적 한계점이 존재하고 단어의 순서나 문서 전체의 내용 흐름을 반영하지 못하는 단점이 있다. [3]에서는 [2]가 가지지 못하는 단어의 순서를 고려하는 장점을 가질 뿐 만 아니라 지역적 한계성을 벗어나 문서 전체 내용의 특징을 반영하는 장점을 보인다[3].

자동으로 문서를 분류하기 위해서는 어느 범주로 문서를 분류할 것 인지를 결정하기 위해서 문서 분류 규칙을 정한다. 문서 분류를 위해 학습에 사용되는 알고리즘은 규칙 기반 방법, 확률 기반 방법, 결정트리를 이용하는 방법, 서포트벡터머신(SVM: Support Vector machine)을 이용하는 방법 등 다양한 방법이 있다[4].

1) 교신저자

규칙기반 방법은 사용자가 직접 규칙을 미리 정의하고 규칙에 따라 문서를 분류 하는 것인데 이 방법은 수작업으로 규칙을 구축하므로 다른 분야에 이용하거나 시스템을 크게 확장 시킬 때 많은 시간과 비용이 요구된다. 확률기반 방법은 단어들이 특정 범주에 출현할 확률 값을 계산 하고 그것을 통해서 새로운 문서가 들어왔을 때 그 범주를 맞추는 것인데, 이 방법은 좋은 성능을 가질 수 있으나 많은 학습시간을 가지는 문제가 있다[13]. 서포트벡터머신은 학습 문서를 통해서 추출되는 긍정과 부정을 벡터 공간으로 표현한다. 이 둘 간의 차이가 극대화 되는 벡터를 찾는 방법이다. 컨볼루션 신경망 분류기(CNN)는 기존 텍스트 분류 기술에서 널리 사용되는 분류기에 비해 주제 예측에서 뛰어난 성능을 보였다[1]. 그러나 이 방법은 문서에 나타난 단어의 순서 정보를 반영하지 못하고, L2 정규화를 적용하지 않을 경우 과적합의 가능성을 가지는 단점이 있다. 따라서 본 논문은 이를 해결하기 위한 방법을 제안한다.

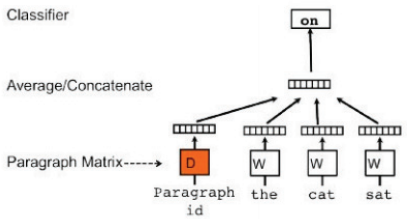
본 논문에서는 문서의 색인 과정에서 서로 다른 범주를 가지는 문서의 차이점을 효율적으로 표현할 수 있는 방법으로 문서의 지역특징과 전역특징을 반영하도록 단어의 의미를 표현하고 순서를 고려하는 방법을 문서의 색인에 사용하여 문서를 벡터로 표현하였다. 문서 분류는 주제 예측에 높은 성능을 보이는 컨볼루션 신경망 분류기를 이용하였다. 성능 비교를 위해 한국일보-20000 문서범주화 실험문서집합의 데이터 셋[5]를 이용한 실험결과 CNN 분류기를 기반으로 본 논문에서 제안하는 문서 색인 방법이 TF-IDF를 이용하는 방법보다 4.2%, Word2vec만 단독으로 사용하는 것보다 1.4%의 성능 향상을 이루었다.

2. 본 론

본 논문에서는 문서 내의 단어의 의미와 순서를 문서 벡터 표현에 넣기 위해 단어의 의미를 고려하는 Word2vec과 단어의 순서까지 고려하는 Doc2vec을 이용하여 문서를 표현하고, 컨볼루션 신경만 분류기를 이용하여 분류하는 방법을 제안한다.

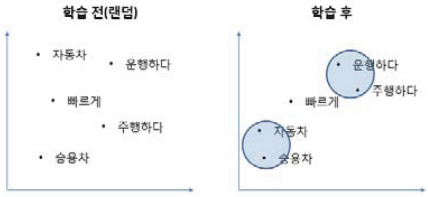
문서를 벡터로 표현하면 각 차원의 값들은 특징과 연결되며, 그 정보는 단어나 문서의 의미 또는 문법에 관한 정보도 가질 수 있다[6]. 자연어 처리에서 분류기의 성능을 향상 시키는 일반적이고 가장 중요한 방법은 문서를 벡터로 적절하게 표현하는 것이다.

최근에는 Word2vec와 Doc2vec 방법이 떠오르고 있는데, 둘 다 비지도 학습 방법으로 피드포워드 신경망을 이용하여 학습을 하는 표현방식으로, 학습 데이터에서 어휘를 형성하고 각 단어와 각 문서의 벡터 표현을 학습한다. 표현한 벡터의 로그우도(log-likelihood)를 높여가는 과정에서 단어에 대한 의미론적 정보와 문서에 대한 의미론적 정보를 표현 가능하다.



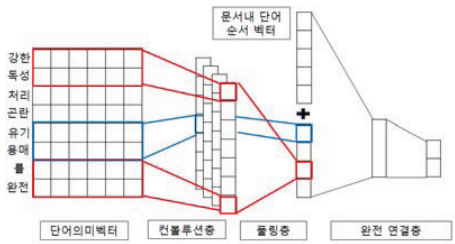
[그림 1] Doc2vec의 DM모델[3]

Doc2vec는 Word2vec의 CBOW(Continuous Bag of Words)모델과 비슷한 DM(Distributed Memory) [7] 모델이 있다. 이 모델이 Word2vec의 CBOW모델과 다른 점은 문서 매트릭스(matrix) D 가 추가되었다는 것이다. 예를 들어 문서에 “the cat sat on”이라는 문장이 존재하면, DM 모델은 “the cat sat ”3개의 단어 벡터와 문서 벡터를 평균한 벡터를 4번째 단어를 “on”을 예측하는데 사용한다. 예시는 위에 나타난 [그림 1]와 같다. Word2vec 모델로 단어들을 학습 시켰을 경우 학습 후 단어 벡터의 변화는 아래 [그림 2]과 같다.



[그림 2]학습전,후 Word2vec의 단어벡터 위치[15]

제안하는 방법은 [그림 3]과 같이 문서 내 단어 순서벡터를 이용하기 위해 Doc2vec를 사용해 생성한 벡터가 문서 분류 수행 할 때 한 번 사용되고, 문서 내 단어들의 의미를 벡터로 표현하기 위해 Word2vec를 이용하여 생성한 벡터 배열이 컨볼루션 층과 풀링층을 지나 문서 내 단어 순서 벡터와 결합하여 마지막 벡터 배열을 만들고 이것이 완전 연결층의 입력으로 전달이 된다.



[그림 3] 시스템 구조도

3. 실험 및 결과

3.1 데이터 집합

본 논문에서 분류시스템의 학습과 분류실험을 위하여 사용된 데이터 셋은 '한국일보-20000'(HKIB-20000)이다. 한국일보-20000 실험문서집합은 한국일보-40075 집합의 기사 중 20,000건을 추출하여 분류체계를 보다 수립하였으며, 3단계 분류체계의 모든 노드에 기사를 할당하여 구축한 계층적 분류체계의 문서범주화용 실험문서집합이다. 본 논문에서는 문서 분류를 위하여 각 범주 당 140개 이상의 문서를 가진 10개의 범주를 선택하여 실험에 사용하였다. 실험에 사용된 범주는 아래 표 1과 같다. 학습문서와 실험문서의 개수를 달리하며 5겹 교차검증(5-fold cross validation)을 적용하였다.

[표 1] 실험에 사용한 범주

범주 이름	문서개수
건강의학/의약학/질병(암외의질병)	196
경제금융/은행(금유업계동향)	282
과학/자연과학/화학	188
문화종교/스포츠/야구	325
문화종교/공연/방송연예	518
문화종교/종교/불교	141
사회질서/군대	774
산업건설업/토목	272
산업제조업/컴퓨터	481
여가실외/여행관광	182

3.2 실험방법

실험을 위해 전체 실험 데이터에 형태소 분석을 수행하기 위해 Konlpy(Korean NLP in Python)[8]를 사용해 문장을 어절 단위로 분리하여 명사만 추출하였다. 사용된 형태소 분석기는 Konlpy에 내장된 형태소 분석기 Kkma, Komoran, Hannanum, Twitter, Mecab 중 대용량 문서에서 분석기 실행 시간이 짧고, 한글 신조어 등을 가장 많이 보유하고 있는 형태소 분서기 Twitter를 이용하여 어절을 생성하였다[10]. 그 후 Term-Document-Matrix를 생성하여 DF (Document Frequence)값이 높은 단어를 추출해 전체 문서에서 50%이상 출현한 단어들은 각 문서들을 구분하는데 방해가 되는 불용어로 처리하였다 불용어 처리 기준은 [11]에 나온 기준에 따라 설정하였다. 단어와 문서의 벡터 표현을 만들기 위해 위의 방법으로 전 처리한 문서를 가지고 각 어절들을 Word2vec와 Doc2vec 라이브러리를 활용하여 문서와 단어의 벡터 표현을 생성한다. 생성된 Word2vec의 벡터들에 대한 컨볼루션 층과 풀링(Pooling) 층의 출력 벡터와 Doc2vec의 벡터를 더하여 평균 값을 구한 벡터를 완전 연결층의 입력 값으로 사용 하였다.

훈련 데이터는 전체 데이터의 90%, 실험 데이터는 10% 이다. 컨볼루션 신경만 분류기의 설정은 [1]에서 문서 분류에 가장 높은 성능을 보인 파라미터 값을 사용했다. 사용한 설정은 다음 표 2와 같다.

[표 2] 컨볼루션 신경만 분류기 설정 값

설정명	설정값
filter region size	(3,4,5)
feature maps	100
activation function	ReLU
pooling	1-max pooling
dropout rate	0.5

위의 설정 이외에도 문서 당 단어 수는 300개, 상위 50만개 단어를 사전에 저장하도록 하고, 학습률은 0.02, 감소는 선형방식으로, 채널 크기, 커널 크기, 히든 크기, 임베딩 크기는 각각 2500, 4, 1000, 300으로 설정했다. 그리고 L2 정규화의 람다 값을 훈련 데이터의 과적합을 방지하기 위하여 0.0001로 설정하였다[12].

비교 실험의 성능 평가를 위해 정확도를 이용하였다. 정확도는 다음 식 1과 같이 표현된다.

$$\text{정확도} = (\text{True-positive}+\text{False-negative})/\text{총개수} \quad (1)$$

3.3 실험결과

여러 분류기와 다른 색인 방법을 가지고 한국일보-20000 문서 범주화 실험 데이터 셋의 정확도를 구하였을 때 그 값은 다음 표 와 같다.

[표 3] 색인과 분류기를 다르게 적용한 범주화 실험 정확도

색인 / 분류기	TF-IDF	Word2vec[2]	**Word2vec+ Doc2vec**
K-NN[9]	0.635	0.624	0.630
BPNN[4]	0.720	0.728	0.736
서포트벡터 머신[14]	0.747	0.751	0.761
CNN[1]	0.762	0.821	**0.832**

각 분류기에 문서 색인 방법을 TF-IDF, Word2vec, Word2vec+Doc2vec 세 가지 방법으로 적용하여 정확도를 비교하였다. [표 3]에 나타난 K-NN 방법은 문서 간의 거리를 비교하여 가까운 거리에 있는 문서들이 많은 범주에 문서를 분류한다. 그러나 K-NN은 비정형 문서에서 좋은 성능을 보여주지 못하는 단점이 있다. 또한 역전파 신경망 알고리즘인 BPNN(Back Propagation Neural Network)은 특이값분해(SVD:Singular Value Decomposition)를 결합하여 차원축소를 수행해 이용하는 방법으로 입력과 출력 사이의 관계가 사용자에 대하여 불투명하고, 학습시간이 길다는 단점이 있다.

위의 결과와 같이, 본 논문에서 제시하는 방법인 문서 내 단어의 의미와 순서를 고려하는 방법으로 Word2vec과 Doc2vec을 이용하여 색인하고 문서 분류에서 CNN 분류기를 이용하는 방법이 [1]의 결과 82.1% .보다 1.4% 가량 정확도가 높은 것을 볼 수 있다.

4. 결 론

본 논문에서는 문서 내 단어의 의미와 순서를 고려하기 위하여Word2vec와 Doc2vec를 이용하여 문서를 색인하는 방법을 한국일보-20000 문서 범주화 실험문서 집합에 적용하는 방법을 제안하였다. 실험 결과를 통해 CNN을 기반으로 하는 Word2vec와 Doc2vec를 결합한 방법이 타 방법에 비해 높은 정확도를 가졌다.

향후 연구로는 본 논문에서 제안한 방법을 문서 분류 뿐 만 아니라 정보검색에도 적용해보고자 한다. 또한 문서의 길이가 논문처럼 긴 문서에 대해서 적용 할 수 있는지 연구하고자 한다. 문서의 길이와 양이 늘어날 경우 훈련에 걸리는 시간도 기하급수적으로 늘어나는데 분류기의 실험 시간을 줄이는 방법에 대해서도 체계적인 분석이 필요하다.

Acknowledgement

본 성과물은(논문, 산업재산권, 품종보호권 등)은 농촌진흥청 연구사업(세부과제번호: PJ011823022017)의 지원에 의해 이루어진 것임.

참 고 문 헌

[1] 조휘열, 김진화, 윤상웅, 김경민, & 장병탁. (2015). 컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술. 한국정보과학회 학술발표논문집, 792-794.

[2] Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

[3] Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368.

[4] 리청화, 변동률, & 박순철. (2010). 한글문서 분류에 SVD 를 이용한 BPNN 알고리즘. 한국산업정보학회논문지, 15(2), 49-57.

[5] http://www.kristalinfo.com/TestCollections/readme_hkib.html

[6] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[7] Kimothi, D., Soni, A., Biyani, P., & Hogan, J. M. (2016). Distributed representations for biological sequence analysis. arXiv preprint arXiv:1608.05949.

[8] Park, E. L., & Cho, S. (2014, October). KoNLPy: Korean natural language processing in Python. In Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology (pp. 133-36).

[9] Tan, S. (2006). An effective refinement strategy for KNN text classifier. Expert Systems with Applications, 30(2), 290-298.

[10] Lee, Y. G. (2015). An Experimental Study on Open Source Korean Morphological Analyzers for Evaluating Noun Extraction. 한국도서관정보학회 동계 학술발표회, 365-382.

[11] 강승식. (2004). 한글 문서의 색인어와 색인기법. 정보과학회지, 22(4), 72-77.

[12] Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning (p. 78). ACM.

[13] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2), 103-134.

[14] Tsoumakas, G., & Katakis, I. (2006). Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3(3).

[15] 김우주,김동희,&장희원. (2016). Word2vec을 활용한 문서의 의미 확장 검색방법.한국콘텐츠학회논문지, 16(10), 687-692.