합성곱 신경망을 활용한 개인 맞춤형 연관 이슈 추천 시스템

Personalized recommendation system for related issues using convolutional neural network

저자
(Authors)

김성진, 김건우, 이동호
Sung-Jin Kim, Gun-Woo Kim, Dong-Ho Lee

출처
(Source)

한국정보과학회 학술발표논문집 , 2017.06, 764-766 (3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07207374

APA Style

김성진, 김건우, 이동호 (2017). 합성곱 신경망을 활용한 개인 맞춤형 연관 이슈 추천 시스템. 한국정보과학회 학술발표논문집, 764-766.

합성곱 신경망을 활용한 개인 맞춤형 연관 이슈 추천 시스템[1)]

김성진[O1] 김건우[2] 이동호[1]
한양대학교 컴퓨터공학과
tedisagood@naver.com, kgwhsy@hanyang.ac.kr, dhlee72@hanyang.ac.kr

Personalized recommendation system for related issues using convolutional neural network

Sung-Jin Kim[O1] Gun-Woo Kim[2] Dong-Ho Lee[1]
Dept of Computer Science, Hanyang University

요 약

현재 웹 서비스에서는 방대한 데이터가 생성되고 있다. 데이터가 많아질수록 사용자는 원하는 정보를 찾는데 어려움을 겪고 있다. 대부분의 웹서비스에서는 사용자에게 적절한 정보를 제공하기 위해 검색어 순위 및 연관 검색어를 제공하지만 다수의 사용자가 관심을 적게 가지는 주제는 상대적으로 적절한 정보를 제공해 주지 못하므로 사용자는 직접 모든 데이터를 살펴봐야 한다. 본 논문에서는 이러한 문제점을 극복하기 위해 특정 사용자의 관심 주제를 파악하고 관심주제와 관련된 최근 뉴스기사를 수집 및 분석하여 해당 주제와 관련된 최근 이슈들을 추출해 사용자에게 제공하는 시스템을 구축한다. 이를 위해 딥러닝의 한 종류인 합성곱 신경망을 활용하여 수집된 한글 뉴스기사들을 분석한 후 개인이 관심을 갖고 있는 주제들과 관련된 단어들을 연관이슈로 자동 추천하는 방법을 제안한다. 또한 실험을 통하여 합성곱 신경망을 활용하는 것이 기존의 다른 기계학습 기법보다 우수하다는 것을 보였다.

1. 서 론

현재 웹서비스에서는 방대한 데이터가 빠르게 생성되고 있다. 국내 웹서비스들은 많은 데이터 중 사용자에게 적절한 정보를 제공해 주기 위해 검색어 순위 및 연관 검색어 등을 제공 하고 있다. 이러한 서비스는 최근 발생한 다양한 이슈와 관련하여 많은 사용자들이 빈번히 사용한 검색어와 이와 관련된 연관 검색어들이기 때문에 특정한 개인의 관심 주제와는 거리가 멀 수 있다. 결국 사용자는 자신이 관심 있는 특정 주제 관련 최근 이슈를 얻기 위해 직접 모든 데이터를 살펴봐야 한다. 본 논문에서는 이러한 문제점을 극복하기 위해 특정 사용자의 관심주제를 파악하고 관심주제와 관련된 최근 여러 이슈들을 반영한 연관 이슈들을 추천해 주는 시스템을 구축한다.

기존의 연관단어 추천 연구는 기계학습 기법을 사용한 방법과 출현 빈도수 기반으로 연관단어를 생성하는 연구들이 진행 되었다[1],[2]. 기계학습기법은 양질의 결과를 찾기 위해 여러 특징들을 설계해야 한다. 이 특징들은 전문가가 설정해야 하며 좋은 결과가 생성 될 때까지 계속하여 설계해야 한다. 또한 그 과정에서 많은 파라미터들을 직접 설정해야 하므로 많은 시간과 노력이 든다. 기계학습 기법의 단점을 극복하기 위해 최근 딥러닝이 각광받고 있다 [3].

딥러닝은 인공신경망을 다층구조로 배치해 데이터를 분석하는 기계학습 분야의 한 종류이다. 그중 한 종류인 합성곱 신경망은 인풋 데이터를 각기 다른 필터를 통해 분석, 해당 데이터를 가장 잘 나타내는 특징들을 추출해 계속해서 학습하고 학습된 특징들을 통해 새로운 데이터를 분석한다. 주로 비전분야에 많이 쓰인 인공신경망이지만 최근 텍스트 분석에 많이 활용되고 있다. 합성곱 신경망은 재귀 신경망에 비해 인풋 데이터의 세밀한 특징을 더 잘 파악할 수 있는 강점을 가진다.

본 논문에서는 이러한 합성곱 신경망의 강점에 주목하여 사용자의 취향을 파악해 관심주제를 추출하고, 주제와 관련된 뉴스기사들의 특징들을 합성곱 신경망을 통해 찾아내어 연관이슈를 자동으로 생성하는 시스템을 구축한다. 연관이슈 추천 분야는 주로 영어문장을 분석하는 연구가 진행되었다. 본 논문에서는 한글 데이터를 분석하여 한글 연관이슈를 추천하는 시스템을 구축하였다. 이를 위해 현재 많이 사용되는 여러 형태소 분석기들 중 파이썬 한국어 분석 패키지(KoNLPy)에서 진행한 성능평가를 참조하여 가장 성능이 좋은 Mecab을 선정하여 사

1) 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A1A09918271, 다중 저장소 지속성 환경에서 빅데이터 기술을 활용한 개인 맞춤형 소셜 멀티미디어 태깅 및 태그 관리 시스템)

용하였다 [4].

2. 관련연구

연관이슈를 단어형태로 추천하는 연구는 출현 빈도수를 사용한 방법과 기계학습 기법 중 LDA를 사용한 방법 등이 연구 되었다 [1],[2]. 합성곱 신경망을 사용하여 텍스트를 분석하는 여러 연구들도 진행 되었다. 합성곱 신경망을 통해 텍스트 랭킹하는 연구, 문장을 분류하는 연구, 감성을 분석하는 연구 등이 진행되었다 [5],[6],[7]. 합성곱 신경망을 통해 텍스트 데이터를 분석하여 연관이슈를 추천하는 연구는 진행되지 않았다. 또한 위 연구들은 영어기반의 연구이며 한글 문서를 분석하여 연관이슈를 추천하는 연구는 아직까지 진행되지 않았다.

3. 제안 시스템

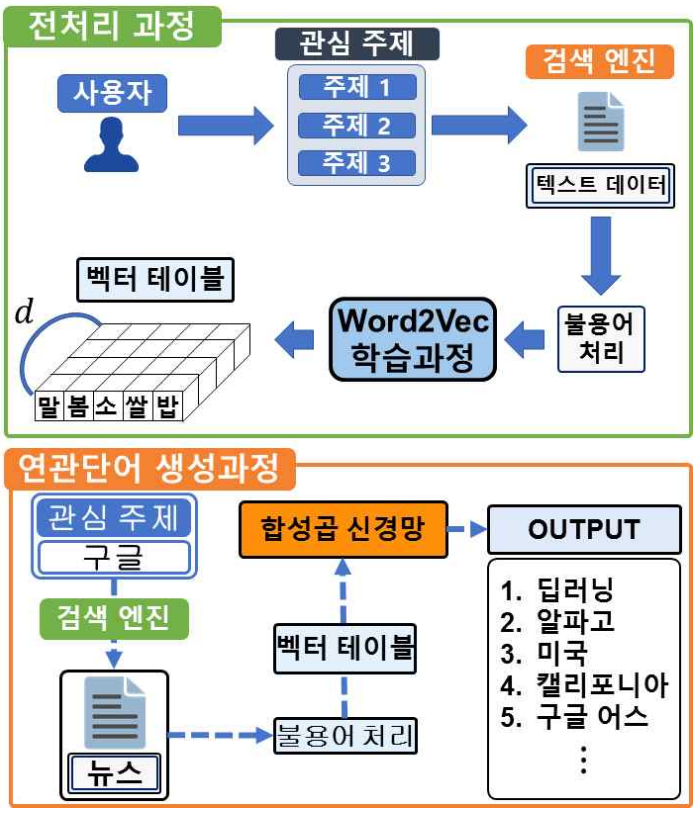

그림 1 전체 시스템 구조도

그림 1은 전체 시스템의 구조를 나타낸다. 관심 주제를 파악하지 않고 무분별한 대량의 텍스트 데이터를 학습하여 벡터테이블을 생성할 경우 다음과 같은 문제들이 발생한다. 먼저, 데이터의 크기가 커질수록 Word2Vec 로 벡터테이블을 생성하는데 걸리는 학습시간이 길어진다. 또한, 분석하는 데이터 크기에 비례하여 생성되는 벡터 테이블의 크기가 커지므로 공간효율성이 떨어지는 문제를 가지게 된다. 딥러닝은 학습하는 데이터의 크기가 많으면 많을수록 좋은 성능을 보이고 대량의 데이터의 분석은 불가피하므로 이러한 단점을 대비할 필요가 있다.

먼저, 전처리 과정에서는 Word2Vec 의 학습범위를 줄이기 위해 사용자가 평소 관심 있게 검색한 검색어의 빈도수를 파악한다. 검색어 중 빈도수가 높은 순서대로 정렬하여 관심 주제를 추출한다. 추출된 관심 주제들을 바탕으로 각종 검색엔진에서 실시간으로 생성되는 방대한 데이터들 중, 관심 주제들과 관련된 다양한 텍스트 데이터들만을 수집한다. 수집된 텍스트 데이터는 의미를 파악하는데 방해가 되는 불용어를 형태소 분석기를 사용하여 제거하는 과정을 거치고 Word2Vec를 통해 학습되어 d차원의 벡터 테이블을 생성하게 된다. 벡터 테이블은 연관단어 생성과정에서 뉴스기사의 단어들에 벡터값을 부여하는데 사용된다.

전처리 과정이 진행된 이후에 완성된 벡터테이블을 바탕으로 연관단어 생성과정을 진행한다. 사용자가 관심 있어 하는 각각의 주제들과 관련된 최근 일주일 분량의 한글 뉴스기사를 검색엔진을 통해 수집한다. 전처리과정과 마찬가지로 수집된 기사들의 불용어를 처리한다. 불용어가 제거된 각 기사의 단어들에 생성된 벡터 테이블을 통해 벡터값을 부여하고 합성곱 신경망으로 학습을 진행한다. 합성곱 신경망을 통해 기사를 분석하는 자세한 과정은 다음과 같다.

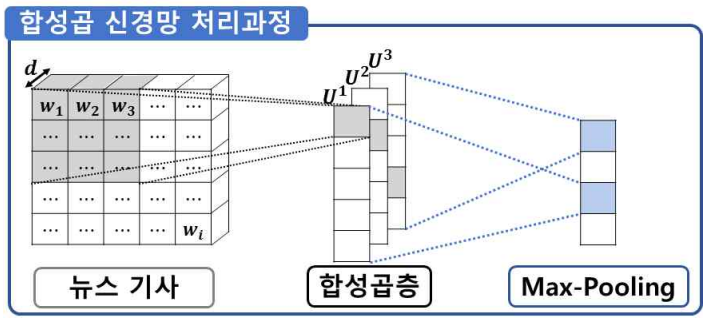

그림 2 합성곱 신경망 처리과정

d차원의 벡터값이 부여된 뉴스기사의 각 단어 $(w_1, w_2, w_3, \cdots, w_i)$들이 합성곱 신경망에 주어지면 합성곱 신경망은 단어들을 분석하여 각 기사를 잘 나타내는 특징벡터를 찾아낸다. 먼저, 합성곱층을 통해 합성곱 과정을 수행하게 된다. 합성곱 과정에서는 추출된 단어(w)와 필터의 계수(c)를 순서대로 곱하여 그 합을 사용자가 원하는 만큼 생성된 피처 맵(U_1, U_2, U_3)에 채워나간다. 합성곱 과정은 수식(1)로 나타낼 수 있다.

$$u_i = f(c \bullet w_n + b) \quad (1)$$

b는 각 필터마다 다르게 설정된 값인 바이어스 값이다. 필터의 계수(c)는 무작위 값으로 초기화 시킨 후 학습과정을 통해서 조정해 나간다. 활성화 함수(f)는 ReLU[8] 함수를 사용했다. 합성곱 과정은 수식(1)을 통해 단어로 이루어진 각 기사를 여러 부분으로 나누어 각 부분을 하나의 값(u_i)으로 표현하여 각 피처맵(U^k)에 채워넣는 과정이라고 볼 수 있다. 합성곱 과정이 진행되고 나면, Max-Pooling과정을 수행하며 그 과정은 수식(2)로

나타 낼 수 있다.

$$m_k = \max U^k(2)$$

Max-Pooling과정은 k개의 피처맵(U^k)에 저장된 합성곱 처리과정의 결과들중 최대값을 뽑아내어 k개의 결과물(m_k)을 생성한다. 합성곱 과정, ReLU, Max-Pooling과정을 원하는만큼 수행한 이후, 전결합층(Fully connected layer)을 배치하여 다시 학습을 진행 한 후, 이 과정들을 반복해서 수행한다.

특정 주제로 검색된 뉴스기사들이 들어올 때마다 계속해서 합성곱 신경망 처리과정을 통해 해당 주제를 가장 잘 나타내는 특징을 학습한다. 주어진 모든 기사를 학습한 이후에는 학습된 특징을 해당 주제의 연관 단어로 추출하여 사용자에게 제공한다. 이 과정을 사용자의 관심 주제로 생성된 다양한 주제에 대해서 반복적으로 수행한다. 수집된 일주일 분량의 뉴스기사로 생성된 연관단어는 시간이 지날수록 최신 이슈를 반영하지 못하므로 연관단어가 생성된 후 일주일 후에, 혹은 사용자가 원할 때마다 다시 연관단어 생성과정을 수행하여 최신이슈를 반영한 연관단어를 재 생성하여 사용자에게 추천한다.

4. 실험

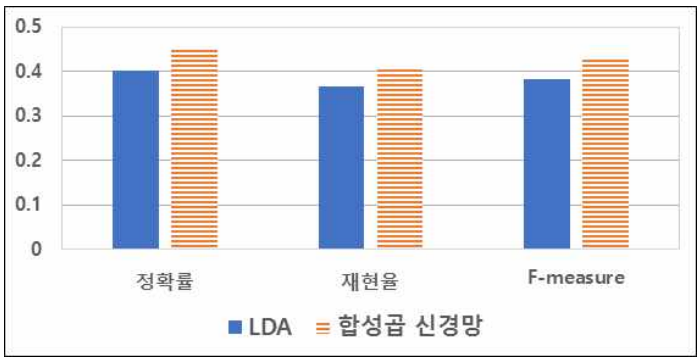

그림 3 ROUGE-1 성능평가 그래프

<표 1> ROUGE-1 성능평가 수치표

	정확률	재현율	F-Measure
합성곱 신경망	0.4484	0.4084	0.4275
LDA	0.4020	0.3661	0.3832

실험 평가를 위해 본 시스템을 통해 생성된 연관 단어와, 사람이 직접 일주일 분량의 뉴스기사를 읽고 추출한 연관 단어를 비교하여 성능 평가를 진행 하였다. 성능평가는 평가방법으로 널리 이용되고 있는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)를 이용하였다[9]. 실험은 뉴스, 블로그 등 다양한 웹서비스에서 수집한 데이터로 진행했으며 DeepLearning4J 패키지를 사용하여 시스템을 구현하였다. 연관단어 생성분야에서 좋은 성능을 보인 LDA와 그 성능을 비교 측정하였다.

그림 3 과 표 1는 본 논문에서 제안한 시스템의 성능과 LDA의 성능을 비교한 결과를 보여준다. 그림 3과 표1에서 볼 수 있는 것처럼, 본 시스템(합성곱 신경망)을 통해서 생성한 연관 단어가 LDA로 생성한 연관단어보다 더 높은 정확률, 재현율, F-Measure값을 보이는 것을 알 수 있다.

5. 결론

본 논문에서는 사용자의 검색어 기록을 통해 관심주제를 파악하고 관심주제와 관련된 뉴스기사를 합성곱 신경망을 통해 분석하여 최신 연관 이슈를 추천하는 시스템을 구축하였다. 또한, 대부분 영어 기반으로 진행 되었던 연관 단어 생성 연구와는 달리 한글 기반의 연구를 진행하였고 기존에 좋은 성능을 보인 LDA 기법과 비교실험을 진행하여 성능을 증명 하였다. 이를 통해 앞으로 진행될 연관 단어 생성 연구에 많은 기여를 할 수 있을 것이라고 기대한다.

참 고 문 헌

[1] Yang, Jinqiu, and Lin Tan. “SWordNet: Inferring semantically related words from software context.“ Empirical Software Engineering 19.6 (2014): 1856-1886

[2] Das, Pradipto, Rohini K. Srihari, and Jason J. Corso. “Translating related words to videos and back through latent topics.“ Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. “Deep learning.“ Nature 521.7553 (2015): 436-444

[4] Park, Eunjeong L., and Sungzoon Cho. “KoNLPy: Korean natural language processing in Python.“ Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology. 2014.

[5] Severyn, Aliaksei, and Alessandro Moschitti. “Learning to rank short text pairs with convolutional deep neural networks.“ Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015

[6] Kim, Yoon. “Convolutional Neural Networks for Sentence Classification“, Empirical Methods on Natural Language Processing, 2014

[7] Dos Santos, Cícero Nogueira, and Maira Gatti. “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts.“ COLING. 2014

[8] Nair, Vinod, and Geoffrey E. Hinton. “Rectified linear units improve restricted boltzmann machines.“ Proceedings of the 27th international conference on machine learning (ICML-10). 2010.

[9] Lin, Chin-Yew, and Franz Josef Och. “Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics.“ Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.