Word2Vec을 이용한 웹 문서 클러스터링 시스템 구현*

이현석, 안성훈, 이용환, 천명재, 박혁주, 박미화, 이용규
동국대학교 컴퓨터공학과-서울
e-mail : lazuki@dongguk.edu

Implementation of a Web Document Clustering System Using Word2Vec

Hyun Seok Yi, Sung Hun Ahn, Yong Hwan Lee, Myung Jae Cheon,
Hyeok Ju Park, Mee Hwa Park, Yong Kyu Lee
Department of Computer Science and Engineering, Dongguk University-Seoul

요 약

웹 문서 추천 시스템에서는 유사한 내용의 문서임에도 불구하고 URL이 달라서 다른 문서로 인식하여 사용자에게 추천하는 데이터 희소성 문제가 있다. 여기서 기존 연구들은 이 문제에 대한 해결 방법으로 TF-IDF를 이용하였으나 비용 및 시간의 한계가 있으며 유의어 분류 문제가 있다. 본 논문에서는 Word2Vec을 이용한 웹문서 학습 시스템을 통해 문제를 해결한다. 제안 시스템은 언론사의 뉴스를 수집하고 이를 정형화된 형식으로 분석하여 가공하는 전처리 과정을 거친 후 Word2Vec 학습을 통해 문서 벡터를 생성하고 이를 K-Means 클러스터링으로 유사 문서군으로 분류한다. 이 시스템을 이용하면 데이터 희소성 문제를 해결할 뿐만 아니라 연산량이 TF-IDF에 비해 줄어들고 유의어 분류 시 유사도가 높아지는 강점이 있다.

1. 서론

추천 시스템에서 주로 사용하는 협업 필터링[1]은 현재 사용자와 유사한 패턴을 보이는 다른 사용자를 찾아 그들이 평가했던 아이템 중 현재 사용자가 좋아할만한 아이템을 찾아내는 것이다. 이러한 협업 필터링의 문제점은 사용자가 평가한 기존 아이템의 수가 적거나, 사용자에 비해 아이템의 수가 많아 사용자들 사이의 교집합을 찾을 수 없는 것이다. 이러한 문제를 데이터 희소성[2]이라고 한다.

하지만 웹 문서 추천에서의 데이터 희소성 문제는 조금 다르다. 인터넷 뉴스 서비스의 경우, 하나의 사건에 대해 여러 언론사에서 글을 작성하기 때문에, URL(이하 ID)은 다르지만 같은 내용의 뉴스(이하 아이템)들이 다수 존재하게 된다. 이 경우 동일하거나 유사한 내용의 아이템이지만 다른 아이템으로 인식하여 사용자에게 추천해주는 문제점이 있다.

기존 시스템에서는 이러한 문제점을 해결하기 위해 문서에 속한 단어들의 TF-IDF[3] 값을 이용하여 문서 벡터를 만든다. 그 후에 각 문서들의 유사도를 계산하여 문서들을 군집화한다. 이를 통해 유사한 내용의 문서들이 같은 군집에 속하게 되고, 군집을 하나의 아이템으로 판단하면 사용자가 본 아이템들의 교집합이 커지기 때문에 데이터 희소성 문제가 완화된다. 하지만 계산하는 데에 시간이 오래 걸리고, PCA나 SVD와 같은 차원 축소 방법이 수행되어야 한다. 또한 정확한 Stemming 작업이 이루어지지 않거나, 동일한 단어가 아닌 유의어가 사용된다면 정확한 분류가 이루어지지 않을 수 있다.

Word2Vec은 문서를 이루는 말뭉치를 얕은 신경망을 통해 학습시키는 Word Embedding 방법이다. Hierarchical Softmax기법[4]을 사용하여 연산량이 줄어들고, 유의어의 영향이 적은 장점이 있다.

본 논문에서는 Word2Vec을 이용하여 웹 문서를 학습하는 시스템을 제안한다. 이를 통해 기존의 방식보다 빠르고 유의어의 영향이 적은 방식으로 클러스터링하여 데이터 희소성을 감소시킬 수 있다.

2. 관련 연구

2.1 협업 필터링

협업 필터링은 많은 사용자들로부터 얻은 기호정보를 이용하여 사용자들에게 아이템을 추천해주는 방법이다.

협업 필터링은 사용자 기반 방식과 아이템 기반 방식으로 구분할 수 있다. 사용자 기반 방식은 비슷한 패턴을 보이는 사용자들끼리 군집화하여 현재 사용자가 속한 군집의 사용자들이 높은 점수를 준 아이템을 추천해주는 방식이다. 이와 달리, 아이템 기반 방식은 비슷한 아이템으로 군집화하여 군집을 생성하고 그 군집 안에서 현재 사용자의

* 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 서울어코드 활성화지원사업의 연구결과로 수행되었음(IITP-2016-R0613-16-1147)
* 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학지원 사업의 연구결과로 수행되었음(R7116-16-1014)

패턴을 이용하여 다른 아이템을 추천해주는 방식이다[5].

본 논문에서는 주제에 적합한 사용자 기반의 협업 필터링을 사용한다.

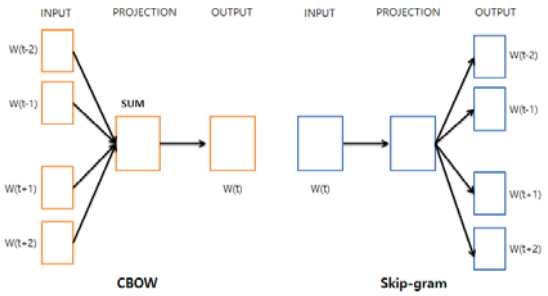
2.2 Word2Vec

Word2Vec은 Neural Network 기반의 Continuous Word Embedding 학습 모형이다. 단어가 가지는 의미 자체를 다차원 공간에서 벡터로 표현하여 벡터 연산을 통해 단어 사이의 관계를 파악하거나 추론을 내릴 수 있다.

Word2Vec의 학습 모델은 2가지가 있다. CBOW 모델[4]과 Skip-gram 모델[6]이다.

CBOW 모델은 여러 개의 단어들로부터 그 단어들과 가까운 특정 단어를 예측하는 모델이다. Skip-gram 모델은 특정 단어로부터 그 단어와 가까운 여러 단어들을 예측하는 모델이다.

(그림 1)은 CBOW 모델과 Skip-gram 모델의 구조를 나타낸다.



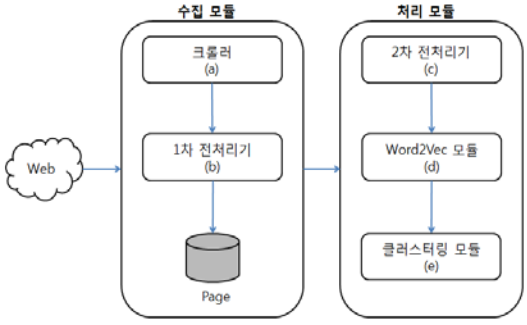
(그림 1) CBOW 모델과 Skip-gram 모델

특정문서에 대해 유사한 문서를 찾기 위해 본 논문에서는 Skip-gram 모델을 사용하였다.

3. 웹 문서 수집 및 Word2Vec 클러스터링 시스템

3.1 시스템 구성도

본 논문에서 제안하는 시스템은 (그림 2)와 같이 수집 모듈과 처리 모듈로 구성된다.



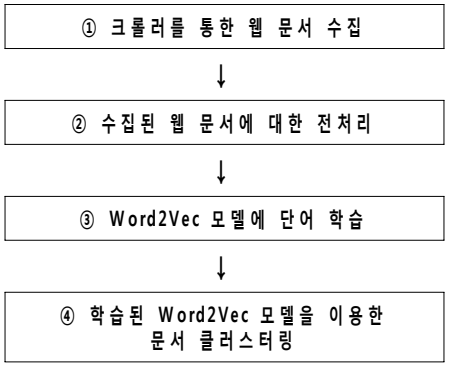
(그림 2) 시스템 구성도

수집 모듈은 언론사 홈페이지의 뉴스 데이터를 수집하기 위한 크롤러(Crawler)와 수집된 뉴스 데이터 내의 학습이 불가능한 단어 제거, 광고 문구 제거 작업을 수행하는 1차 전처리기(Preprocessor)로 구성되어 있다. 1차 전처리를 수행한 데이터는 학습을 위해 데이터베이스에 저장된다.

처리 모듈은 Word2Vec 학습을 위해 문장 단위로 문서를 재구성하고 스테밍(Stemming)을 수행하는 2차 전처리기, Word2Vec의 Skip-gram 모델로 단어를 학습시키고 학습된 단어 벡터를 이용하여 문서 벡터를 생성하는 Word2Vec 모듈, 문서 벡터를 이용하여 클러스터링을 수행하는 클러스터링 모듈로 구성되어 있다.

3.2 동작흐름

(그림 3)은 본 논문에서 제안하는 시스템에 대한 순서도이다.



(그림 3) 전체 시스템 순서도

① (그림 2)의 (a)에 해당하는 크롤러는 지정된 언론사의 뉴스를 수집한다. 수집하는 정보는 문서 작성일, 문서 내용, 문서 제목, 문서 URL이다.

② 수집된 정보 중 데이터베이스에 저장할 수 없는 단어(ex>이모티콘)나 광고 문장 등의 불필요한 정보를 제거하는 작업을 (그림 2)의 (b)에 해당하는 1차 전처리기에서 수행한다. 1차 전처리 과정이 완료되면 문서 단위로 데이터베이스에 저장한다. (그림 2)의 (c)에 해당하는 2차 전처리기는 데이터베이스에 저장된 문서들을 로드하여 각 문서를 문장별로 분리하고, Stemming 작업을 거쳐 Word2Vec 학습의 입력 데이터로 사용할 수 있도록 한다.

③ 전처리 과정을 거친 데이터를 (그림 2)의 (d)에 해당하는 Word2Vec 학습 모듈의 입력으로 사용하여 Skip-gram 모델을 학습시킨다. 학습된 단어 벡터들을 다시 문서 기준으로 그룹핑 한 후 그룹 내 단어 벡터의 합을 평균하여 문서 벡터를 생성한다.

④ (그림 2)의 (e)에 해당하는 클러스터링 모듈에서는 3번에서 생성한 문서 벡터에서 임의로 K개의 문서벡터를

비복원추출하고 K-Means 클러스터링을 이용하여 유사 문서군으로 분류한다.

3.3 구축 환경

시스템 구축 환경은 아래 <표 1>과 같다.

<표 1> 시스템 구축 환경

H/W	프로세서	Intel Xeon Family 8코어
	메모리	16GB RAM
	네트워크	200MB/s
S/W	운영체제	우분투 14.04-LTS
	라이브러리	Apache Spark, Jsoup, Okhttp3, Twitter Korean-txt

3.4 데이터베이스 구조

데이터베이스는 수집하여 분석된 정보를 저장하는 Page 테이블로 이루어져 있다.

이 테이블은 (그림 4)와 같이 문서ID, 문서에 대한 URL, 문서 작성일, 문서 제목, 문서 내용으로 구성된다.

문서 ID	URL	문서 작성일	문서 제목	문서 내용

(그림 4) Page 테이블 구조

4. 구현결과 및 성능 분석

4.1 구현결과

본 논문에서는 5000개의 데이터를 크롤링하여 사용하였으며, Skip-gram 모델의 벡터 크기 200, Window 크기 5, 반복 횟수 12회를 기준으로 학습하였다. 200차원의 단어 벡터를 생성하여 웹문서 추천을 위한 클러스터를 생성하였다.

① 문서벡터 생성

각 문서에 속한 단어와 4.2에서 얻은 데이터를 단어 기준으로 JOIN하여 각 문서에 속한 단어 벡터를 얻을 수 있으며 이에 대한 평균을 구함으로 문서 벡터를 계산한다. 이 때 한 문장 내에 동일한 단어가 여러 번 등장할 경우 중복된 횟수만큼 단어의 벡터를 더한다. 이러한 방식으로 계산하는 경우 문서 벡터의 값이 해당 단어의 공간으로 치우치게 된다. 이는 TF-IDF의 TF와 동일한 역할을 하게 된다. 즉, 여러 번 등장하는 단어일수록 문장 벡터 위치에 큰 영향을 주게 된다.

② 문서 클러스터링

전체 문서 중에서 k개를 비복원추출을 수행한다. 전체 문서에 대하여 추출된 k개의 문서와의 유클리디안 유사도를 계산한다. 각 문서별로 k개의 문서 중 가장 가까운 문서를 선택한다. 이는 모든 문서가 k개의 그룹으로 군집화 되었음을 의미한다.

생성된 k개의 그룹에 대해 다시 중심점을 계산하여 중심점과 가장 가까운 문서를 추출하고, 전체 문서에 대해 동일한 연산을 반복 수행한다.

(그림 5)는 클러스터링 한 군집들 중에서 하나의 군집에 속한 문서들을 보여준다.

ID	URL	Title
1438	http://news.joins.com/article/17069137	"'동물농장' 김상중, 특별 내레이션…백호로 빙의해 "그런데 말입니다""
1443	http://news.joins.com/article/17069735	"'동물농장' 김상중, 특별 내레이션…백호로 빙의해 "그런데 말입니다""
1461	http://news.joins.com/article/17070704	"'동물농장' 김상중, 특별 내레이션…백화 탐정 "완벽 빙의" 폭소"
1479	http://news.donga.com/List/3/all/20150201/69414687/13	동물농장 그것이알고싶다 버전, 김상중 종추한 목소리로 "그런데 말입니다"
1480	http://news.donga.com/List/3/all/20150201/69414687/15	김상중 내레이션으로 동물농장 경력 등장 "그런데 말입니다~"
1491	http://news.joins.com/article/17070566	"'동물농장' 김상중, 백호 탐정 빙의해…"그런데 말입니다" 폭소"
1498	http://news.joins.com/article/17070112	"'동물농장' 김상중, 두 번째 "특별 내레이션"…"그런데 말입니다" 폭소"
1504	http://news.chosun.com/site/data/html_dir/2015/02/01/2015	'동물농장 아저씨' 신동엽, 송해가 보인다 [동물농장 700회]
1535	http://news.joins.com/article/17069231	"'동물농장' 김상중, 백호 사파리 폭행사건 파헤치다…"그런데 말입니다""
1537	http://news.joins.com/article/17069554	"'동물농장' 김상중, 특별 내레이션…탐정 호랑이 변다'
1541	http://news.joins.com/article/17069829	"'동물농장' 김상중, 목소리 출연했다…백호 탐정 빙의 "그런데 말입니다""
1549	http://news.joins.com/article/17069399	"'동물농장' 김상중, 특별 내레이션…"그런데 말입니다" 아침부터 "소름""
1603	http://news.donga.com/List/3/all/20150201/69414687/10	김상중, 동물농장 700회 특집 경력 등장…여지없이 "그런데 말입니다"
1604	http://news.donga.com/List/3/all/20150201/69414687/12	그것이알고싶다 버전 동물농장…김상중 내레이션 "많이 놀라셨죠?"
1632	http://news.donga.com/List/3/all/20150201/69414687/4	김상중, 동물농장 700회 특집 경력 등장…여지없이 "그런데 말입니다"
1649	http://news.donga.com/List/3/all/20150201/69414687/3	김상중, 동물농장 특별 출연 "내 나이가 어때서~"
1650	http://news.donga.com/List/3/all/20150201/69414687/7	김상중, 동물농장 특별 출연 "내 나이가 어때서~"
1653	http://news.donga.com/List/3/all/20150201/69414687/6	그것이알고싶다 김상중, 동물농장 특별 출연 "그런데 말입니다"
1655	http://news.donga.com/List/3/all/20150201/69414687/5	동물농장 김상중, 백호 사연 소개하며 "그런데 말입니다"
1658	http://news.donga.com/List/3/all/20150201/69414687/11	동물농장 김상중 "일요일 아침부터 놀라셨죠?…그런데 말입니다"

(그림 5) 군집화 된 문서들

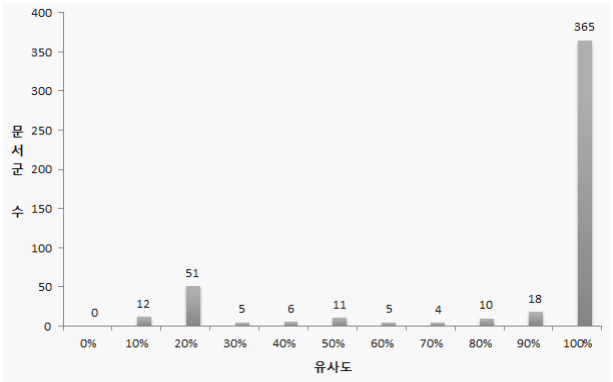
4.2 성능 분석

클러스터링 결과가 정확한지 검사하기 위해서 5000개의 데이터에 대해 직접 라벨링 과정을 수행하였다.

웹문서 추천 시스템의 정확도를 높이기 위해서는 클러스터링 결과로 생성된 각 클러스터 내부의 문서들은 서로 유사해야 한다. 이를 위한 유사도 공식은 아래와 같다.

$$\text{유사도} = \frac{(a)\text{실제 유사한 문서의 수}}{(b)\text{해당 문서군에 속한 문서 수}}$$

(a): 클러스터 내부 문서중 유사한 문서의 수
(b): 클러스터에 속한 전체 문서의 수



(그림 6) 군집별 유사도 현황

(그림 6)은 분류된 군집에 대해 유사도별로 문서군 수를 표현한 그래프이다. x축은 유사도를 나타내고, y축은 문서군 수를 나타낸다.

문서 5000개에 대해서 487개 군집으로 분류하였다. 분류한 군집의 평균 유사도는 약 85%의 유사도를 보였다.

5. 결론

기존 방법은 실시간으로 추가되는 뉴스 추천에서의 데이터 희소성 문제를 완화하는데 계산 및 시간 비용이 높아지는 한계를 가지고 있었다. 본 논문에서 제안하는 시스템에서는 수집된 문서를 Word2Vec 알고리즘에 의한 학습을 통해 각각의 문서벡터를 생성한다.

이를 이용하여 유사 문서를 분류한 결과, 85% 이상의 정확도를 갖는 클러스터를 생성할 수 있었다. 이 분류 기법을 웹 문서 추천에서 활용하면 유사한 내용으로 다른 URL을 가지고 있는 페이지를 다른 내용으로 인식하여 추천하는 문제를 해결할 수 있다.

뉴스 데이터에서의 스테밍 과정에서 단어가 올바르게 추출되지 않아 문서의 군집화가 제대로 이루어지지 않는 경우가 발생하고 있어 향후 정확한 단어 추출을 위한 스테밍 연구를 진행할 것이다.

참고문헌

[1] Gediminas Adomavicius, Alexander Tuzhilin. “Toward the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions”, Foundations and Trends in Human-Computer Interaction, Vol 4, No. 2, pp. 91, 2010.

[2] K. Goldberg, T. Roeder, D. Gupta and C. Perkins, “Eigentaste: A Constant Time Collaborative Filtering Algorithm”, Information Retrieval, Vol 4, No. 2, pp.133-151, 2001.

[3] Su. Xiaoyuan, and Taghi M. Khoshgoftaar, "A survey of collaborative filtering techniques", Advances in artificial intelligence 2009, Vol 2009, pp.5-8, 2009.

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient estimation of word representations in vector space”, arXiv preprint arXiv, Vol 1301, No. 3781, 2013.

[5] Singhal, Amit, "Modern information retrieval: A brief overview", IEEE Data Eng. Bull. Vol 24, No. 4, pp.35-43, 2004.

[6] T. Mikolov, J. Dean, “Distributed representations of words and phrases and their compositionality”, Advances in neural information processing systems, 2013.