The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 18, No. 5, pp.25-30, Oct. 31, 2018. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2018.18.5.25
JIIBC 2018-5-4

Word2Vec를 이용한 한국어 단어 군집화 기법

Korean Language Clustering using Word2Vec

허지욱*

Jee-Uk Heu*

요 약 최근 인터넷의 발전과 함께 사용자들이 원하는 정보를 빠르게 획득하기 위해서는 효율적인 검색 결과를 제공해주는 정보검색이나 데이터 추출등과 같은 연구 분야에 대한 중요성이 점점 커지고 있다. 하지만 새롭게 생겨나는 한국어 단어나 유행어들은 의미파악하기가 어렵기 때문에 주어진 단어와 의미적으로 유사한 단어들을 찾아 분석하는 기법들에 대한 연구가 필요하다. 이를 해결하기 위한 방법 중 하나인 단어 군집화 기법은 문서에서 주어진 단어와 의미상 유사한 단어들을 찾아서 묶어주는 기법이다. 본 논문에서는 Word2Vec기법을 이용하여 주어진 한글 문서의 단어들을 임베딩하여 자동적으로 유사한 한국어 단어들을 군집화 하는 기법을 제안한다.

키워드 : Word2Vec, 단어임베딩, 한국어, 군집화

Abstract Recently with the development of Internet technology, a lot of research area such as retrieval and extracting data have getting important for providing the information efficiently and quickly. Especially, the technique of analyzing and finding the semantic similar words for given korean word such as compound words or generated newly is necessary because it is not easy to catch the meaning or semantic about them. To handle of this problem, word clustering is one of the technique which is grouping the similar words of given word. In this paper, we proposed the korean language clustering technique that clusters the similar words by embedding the words using Word2Vec from the given documents.

Key Words : Word2Vec, WordEmbedding, Korean, Clustering

Ⅰ. 서 론

최근 인터넷의 빠른 보급과 관련 기술의 발달로 인하여 사용자들이 웹상에서 접근할 수 있는 정보는 제한이 없으며 생성되고 있는 문서의 양은 무한하다. 이와 더불어 페이스북, 트위터 등과 같은 대표적인 사회 관계망 서비스 (Social Network Service: SNS) 뿐만 아니라 각종 인터넷 포럼, 리뷰, 댓글, 신문기사, 블로그, 이메일까지 웹상에서 생성되는 문서의 형태 및 구조 그리고 내용 또한 다양하다. 하지만 사용자들은 자신이 원하는 정보들을 쉽게 찾기 위하여 끊임없이 생성되는 문서들을 검색해야 하며 때때로 검색하고자 하는 단어 이외의 의미상 유사한 단어들의 무지로 인하여 효과적인 검색을 할 수 없는 결과를 가져오게 된다. 또한 사용자들의 세대가 변형 되면서 자연스럽게 새롭게 생겨나는 유행어 등과 같은 신조어들의 사용이 증가하기 때문에 이러한 단어들이 포함한 문장들을 분석하기란 쉬운 작업이 아니다.

이러한 문제들을 해결해 주고 효율적인 검색 결과를 제공해 주기 위하여 주어진 문서들의 핵심 단어들을 추출하는 기법들이 많이 연구되고 있다[1]. 특히 이러한 핵

*정회원, 한양대학교 컴퓨터공학과
접수일자 : 2018년 8월 21일, 수정완료 : 2018년 9월 21일
게재확정일자 : 2018년 10월 5일

Received: 21 August, 2018 / Revised: 21 September, 2018 / Accepted: 5 October, 2018
*Corresponding Author: jeeukheu@gmail.com
Dept. of Computer Engineering, Hanyang University, Korea

심단어들을 효과적으로 사용자들에게 제공해주기 위하여 추출된 단어들의 의미적인 유사성을 분석하여 연관된 관련 단어들의 군집화를 생성하는 연구들에 대한 관심이 많아지고 있다. 생성된 단어 군집화는 특정 단어를 접한 비전문가들이 해당 단어에 대한 개념을 쉽게 이해시킬 수 있을 있는 효과를 줄 수 있으며 특정 단어에 대한 의미적인 유사성 분석을 용이하게 해줄 수 있게 한다. 또한 자연어 처리 또는 인공지능, 정보검색 분야에서는 추천 시스템[2], 한글 번역기 등에서 한국어 군집화는 많은 분야에서 활용이 가능하다. 따라서 한국어로 구성된 문장이나 문단에 존재하는 단어들을 분석하고 이를 바탕으로 추후 활발한 연구를 하기 위해서는 연관 단어들로 이루어진 한국어 말뭉치 구축에 대한 연구가 절대적으로 필요하다고 볼 수 있다.

해외의 경우에는 언어학 적인 연구를 위하여 다양한 집단에서 언어나 문서에 대한 분석을 위한 자료들을 생성해서 제공을 해주고 있으며, 영어 뿐 만 아니라 중국어, 스페인어 등과 같은 다양한 언어에 대한 자료들을 제공하지만 국내에는 한글을 대상으로 생성한 한국어 단어 구축 연구들은 환경적인 한계 때문에 활발하지 않아 현재로서는 한국어 단어들 간의 의미적인 분석을 위한 말뭉치자료들이 부족하다[3].

본 논문에서는 Word2Vec를 통하여 단어 임베딩을 하고 k-평균 군집화 알고리즘을 이용하여 주어진 한국어 문서들에 대한 한국어 단어 군집화를 구축하는 기법을 제안한다.

II. 관련연구

1. 텍스트마이닝과 언어 군집화

기존부터 단어의 의미적 또는 중요도등을 추출 다양한 텍스트 마이닝 기법 들이 SNS, 뉴스, 블로그, 전자우편 등과 같은 다양한 문서들을 대상으로 진행되고 있다[4]. 텍스트 마이닝 기법으로는 TF-IDF(Term Frequency - Inverse Document Frequency)와 같이 단어와 문서간에 존재하는 빈도수를 분석하여 추출하는 통계적인 방법과, SVM(Support Vector Machine), Naive Bayesian, LDA(Located Dichirict Allocation) 등과 같은 선행, 비선행 학습들을 통한 기계학습 기법들을 활용한 다양한 텍스트 마이닝 기법들에 대한 연구들이 진행되고 있다. 군집화 알고리즘으로는 같은 확률 기반의 기법과 , 벡터공간의 거리 계산을 통한 K-최근접 알고리즘과 K-평균 알고리즘과 같은 다양한 텍스트 마이닝 기법들을 기반으로 연구가 진행되고 있다.

해외에서는 각국에 해당 하는 단어에 대한 의미 분석을 위해 다양한 연구가 진행되고 있으며, 말뭉치에 대한 배포활동이 활발하다. DUC(Document Understand Conference)와 TAC(Text Analysis Conference)에서는 매년 영문으로 구성된 문서나 단어의 의미적인 실험을 위한 다양한 데이터 들을 제공해주고 있으면, 사진공유 사이트인 플리커(Flickr)애서 사용자가 사진에 직접 입력한 태그들을 바탕으로 유사한 단어들의 집합인 태그 클러스터를 API로 서비스 해주고 있다. 그 외에

일본어[5], 중국어[6,7], 아랍어[8,9], 스페인어[10] 등 다양한 나라에서 각국의 언어를 군집화를 하기 위하여 각국의 특징들을 고려하여 군집화를 생성하기위한 연구들이 진행되고 있다.

한국에서도 한국어 단어의 의미적인 분석을 위한 말뭉치를 제작하고 있다. 국립국어원[11]에서는 한국어 말뭉치 자료를 제작하여 배포를 해왔지만 2014년 그 이후로 배포된 말뭉치에 대한 자료가 미미하다. 또한 언어정보연구원[12]에서는 1900년 초 부터 지속적으로 한국어에 관련된 말뭉치 사전을 제작 배포를 하였지만 최근 배포한 말뭉치는 2011년 10월 한 달 간 작성된 트위터를 바탕으로 제작된 자료로써 최근에 생성되는 단어에 대한 의미분석에 한계가 있다고 할 수 있다. 세종21[13]에서는 21세기세종계획에 의해 '세종시맨틱검색시스템"이 제작 되었다. 세종시맨틱검색시스템은 1997년부터 2007년까지 약 10년간의 데이터를 수집하여 온톨로지형태의 전자사전을 기반으로 구축하여 배포를 하고 있다. 하지만 서론에서도 서술 했듯이 국내에서는 한국어 단어 분석을 위한 관련된 자료생산에 많은 노력과 시간을 투자 하였지만 현재로서는 그 연구에 한계가 있다는 점이 현실이다.

2. 단어 임베딩(Word Embedding)

단어 임베딩은 단어들을 표현을 하기 위해서 수치화 하여 벡터의 형태로 변경하는 방법이다. 대표적으로 사용되는 방법은 주어진 단어들을 리스트로 만들어서 해당 단어의 유무를 '1' 또는 '0'으로 표현하는 BOW(Bag of Words)가 있다. 하지만 BOW는 단순한 단어의 유무를 수치적으로 표현한 방법으로 실제적으로 그 단어의 특징

이나 의미적인 부분 까지는 표현을 하지 못한다. 뿐만 아니라 빈도수가 낮은 단어를 표현할 경우 이를 위하여 생성되는 BOW의 크기가 확장이 되어야 하기 때문에 이를 이용하여 분석을 할 때 불필요한 공간을 차지 한다는 문제점이 발생하게 된다.

T. Mikolov(2013)의 연구에서는 주어진 문서에 있는 각 단어들의 빈도와 위치 등을 파악하고 단어의 확률적인 출연 까지 고려하여 학습하는 Word2Vec 알고리즘을 제안하였다[14]. Word2Vec은 신경망[15] 을 기반으로 하여 구축을 하며, 효율적인 계산을 위하여 subsampling, negative sampling 기법을 사용하였다. 이 연구에서는 Word2Vec를 구성하기 위하여 CBOW(Continue Bag of Words)와 Skip-gram의 두 가지 모델을 제안하고 있다. CBOW 모델은 주어진 문장의 연속된 단어들을 순서대로 입력받아 문장에서 한 단어에 앞뒤로 붙어 있는 단어들을 통해서 해당 단어의 유사도로를 유추하는 방법이다. 반면에 Skip-gram 모델은 주어진 단어를 가지고 주위에 존재하는 나머지 단어들에 대한 존재여부를 유추하여 유사도를 구하는 방법이다. 두 모델은 입력받은 문서의 양에 따라 속도와 정확도의 차이가 생기게 되는데, 일반적으로 Skip-gram 모델이 단어 임베딩 시 속도나 느리지만 그 정확도가 높아 많이 사용되고 있다.

Ⅲ. 3장 Word2Vec를 활용한 한국어 단어 군집화 기법

본 연구에는 주어진 문서들을 분석하고 유사한 단어들로 이루어진 군집화를 생성하기 위하여 Word2Vec 기법을 사용한다. 그림[1] 은 본 논문에서 제안하는 기법 4단계의 순서를 도식화 한 것이다.

우선 주어진 문서들을 입력받아 문서내에 존재하는 단어들을 전처리 과정을 통하여 분석을 하기 용이한 형태로 변경을 해준다.

그 후 Word2Vec를 사용하여 전체 단어 대한 임베딩을 하여 벡터로 생성을 하고, K-means 알고리즘을 통하여 각 유사한 단어들 끼리 군집화를 한다. 마지막으로 각 군집화에 존재하는 단어들 중 빈도 및 유사도 값을 고려한 분석을 통하여 가장 의미 있는 단어를 해당 군집화의 대표 단어로 선정한다.

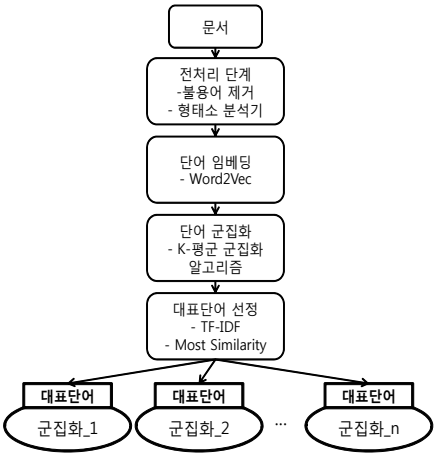

그림 1. 한국어 군집화 순서도
Fig. 1. The process of korean word clustering

1. 단계 전처리 과정

전처리 과정에서는 주어진 문장들의 의미 있는 정보만을 남기기 위하여 숫자, 조사, 기호등과 같은 분석에 불필요한 불용어들을 제거하고 Park, Eunjeong L(2014) 연구에서 배포한 형태소 분석기를 사용하여 2단계에서 군집화를 위한 작업을 한다[16]. 형태소 분석기는 한국어 문장을 읽어 각 단어에 해당하는 명사, 동사, 조사 등과 같은 품사들로 분류 해주는 기능을 해주며, 본 논문에서는 형태소 분석기를 이용하여 명사, 형용사 위주로 추출을 하여 군집화를 진행하기 위한 과정에 사용한다.

2. Word2vec를 이용한 단어 임베딩

전처리 과정을 통하여 나온 단어들은 Word2Vec 을 통하여 단어들을 벡터화 한다. Word2Vec는 앞서 기술하였듯이 Cbow와 Skip-gram 두 가지 모델 기법을 적용하여 단어들에 대한 벡터가 생성한다. 일반적으로 Skip-gram 모델이 속도나 정확도 면에서 어 우수한 성능을 보여준다[14]. 따라서 본 논문에서는 skip-gram 기법을 사용하여 단어 임베딩을 진행한다.

3. 군집화 작업

2단계에서 Word2Vec에 의해서 임베딩 된 단어들의 벡터를 사용하여 유사한 단어들로 묶어주는 군집화 작업을 진행한다. 본 논문에서는 가장 대표적인 군집화 알고리즘인 K-평균 군집화 알고리즘을 사용하여 단어 군집화를 진행하였다.

K-평균 군집화 알고리즘은 2차원 이상의 벡터공간으로 표현된 데이터들 입력 받아서 K개로 군집화로 묶어주는 알고리즘이다. 이를 위해서 주어진 데이터에 대한 군집화를 진행하기 위하여 우선 임의의 K개의 중심점(centroid)을 설정한다. 그 후 이를 기반으로 각 중심점과 각 데이터 들 간의 거리를 구하여 그 값이 최소가 되는 중심점에 해당 데이터를 할당하고 이를 반복한다. 이때 중심점과 각 데이터들 간의 거리를 구하기 위하여 코(Cosine similarity)사인 유사도나 유클리디언 유사도(Euclidean similarity)를 사용한다. 본 논문에서는 거리 측정 시 코사인 유사도를 적용하여 군집화를 진행 하였다.

4. 대표 단어 선정

3단계에서 생성되어진 각 군집화 된 결과물은 Word2Vec와 K-means 알고리즘에 의하여 의미적으로 유사한 단어들의 집합으로 이루어져 있다. 일반적으로 K-means 알고리즘은 비지도 학습으로 군집화 수행 시 각 군집화에 대한 라벨링 작업 까지 하지 않는다. 그렇기 때문에 보다 의미적인 군집화 결과를 제공해주기 위하여 각 군집화에 대한 대표 단어를 선정해주어야 한다. 본 논문에서 이를 위하여 각 군집화에 있는 단어들의 TF-IDF 값과 Word2Vec를 통하여 획득할 수 있는 각 단어들 간의 유사도 분석을 하여 획득한 단어들로 해당 군집화의 대표 단어들을 선정해준다. 그 계산은 아래의 식과 같이 진행 된다.

$$ClusterLabel_i = \underset{i,j \in Cluster}{ArgMax\ Value}(MostSimilarity_{i,j} + TF-IDF_{i,d}) \quad (1)$$

ClusterLabel은 생성된 클러스터를 대표하는 단어를 나타내고 있으며, ArgMaxValue 각 단어의 TF-IDF의 값과 Word2Vec에서 제공해주는 다른 단어와의 유사도 값을 합산하여 구한다. 그 후 최고점을 획득한 군집화내의 단어 중 3, 4개를 대표단어로 선정을 하게 된다.

Ⅳ. 실험 및 결과

1. 실험 데이터 및 방법

본 논문에서는 실험을 위하여 2018년 1월부터 2018년 7월까지 약 6개월간 생성된 6662개 네이버 블로그를 수집하였다. 블로그의 주제는 네이버 데이터 랩(https://datalap.naver.com)을 참조 하였으며, 6개월간 '월드컵', '정상회담', '미세먼지', '김정은', '트럼프'와 같은 단어들을 비롯하여 그동안 화제가 되었거나 주목을 받았던 단어들을 선정하였다. 매월 선정된 블로그의 주제 단어와 수집된 블로그의 수는 표[1]과 같다.

표 1. 수집된 키워드와 블로그 데이터
Table 1. Collected Blog Key Words and Data

월	주제어	블로그의 수
1월	신년,, 최두호 정유라 현송월 윤식당2.	763
2월	평창올림픽, 개막식, 영미, ...	1131
3월	정봉주, 무한도전, 강원랜드, 미세먼지,	831
4위	정상회담, 김정은, 평양공연, ...	1690
5월	트럼프, 싱가포르, 정현, 테니스 어벤져스, ...	1017
6월	러시아, 월드컵, 스웨덴, 프랑스, 호날두, ...	1230

본 논문에서 제안하고 있는 Word2Vec를 활용한 군집화 기법의 평가를 위하여 수집된 데이터에 대하여 미리 선정해 놓은 정답 데이터와 비교 분석을 하였다. 또한 기존의 군집화 알고리즘과의 우수성을 증명하기 위하여 K-평균 군집화 알고리즘 결과 값을 비교하였다. 군집화 된 결과를 정량적인 수치로 평가하기 위하여 결과 값의 정확율(Precision), 재현율(Recall), F-Measure 값을 측정하였으며 그 식은 아래와 같다.

$$Precision_{i,j} = \frac{Nmber\ of\ Words\ of\ topic_j\ in\ Cluster_i}{Nmber\ of\ Words\ in\ Cluster_i} \quad (2)$$

$$Recall_{i,j} = \frac{Nmber\ of\ Words\ of\ topic_j\ in\ Cluster_i}{Nmber\ of\ Words\ of\ topic_j\ on\ Dataset} \quad (3)$$

$$F-Measure_{i,j} = \frac{2 \times (Precision_{i,j} \times Recall_{i,j})}{Precision_{i,j} + Recall_{i,j}} \quad (4)$$

2. 실험 결과

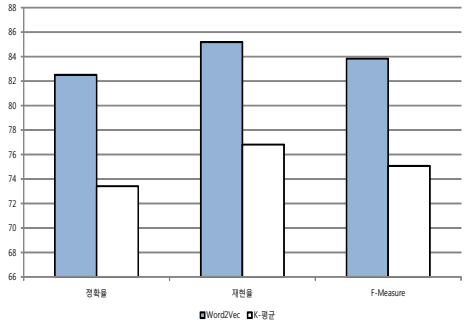

그림 2. 정확율, 재현율, F-Measure 비교
Fig. 2. Comparison of Precision, Recall, and F-Measure

그림 [2]는 Word2Vec을 활용한 군집화와 K-평균 군집화 알고리즘의 정확율, 재현율, F-Measure를 비교한 결과를 보여주고 있다. 그림[2]에서 볼 수 있듯이 Word2Vec를 이용한 군집화의 결과를 보이고 있으며, 각각 12.4%, 10.91%, 11.65% 의 향상된 성능을 보여주고 있다. 이것은 주어진 문서를 단순 BOW 기반으로 구성된 벡터공간을 이용하여 단어들 간의 거리를 구하는 K-평균 군집화 알고리즘 보다 Skip-gram 모델을 이용하여 벡터공간을 표현한 Word2Vec기법 단어들간의 유사도를 측정하는 결과가 더 우수하다는 것을 보여주고 있다. Skip-gram은 문서내의 있는 단어들 간의 위치를 고려하여 벡터공간을 구성하기 때문에 문맥상 유사한 단어들이 벡터공간 내에서 거리가 가까워지게 된다. 따라서 군집화 알고리즘을 수행할 때 벡터공간상에서 거리에 의한 유사도 측정 시 더 정확한 결과 값이 나오게 되었다고 볼 수 있다.

표 2. Word2Vec를 이용한 한국어 단어 군집화 결과
Table 2. The resault of Korean word Clustering by using Word2Vec

군집 1	미국, 북한, **트럼프**, 대통령, **김정은**, 미, 핵, 회담, **비핵화**, **정상회담**, 정부, 북, 선언, 협상, 주장, 관계, 국제, 제재, 체제, 미사일, 대북, 대화, 핵무기, 요구, 미군, 강조, 합의, 언급, 전문가, 약속, 군사, 발언, 조치, 외교, 유해, 의지, 종전, 폐기, 조건, 보장, 장관, 추진, 행정부, 동맹, 중단, 시도, 폼페이, 전달, 제안, 북핵, 송환, 논의, 주한미군, 철수, ...
군집 2	경기, **월드컵**, 선수, 팀, 프랑스, 크로아티아, 축구, **프랑스 우승**, 이후, 대회, 승리, 감독, 안정환, 잉글랜드, **손흥민**, 결승, 기록, 결승전, 이탈리아, 출전, 무대, 승부, 패배, 경기장, 준우승, 승부차기, 아이슬란드, 연장, 만주키치, 호날, 폴 포그바, **조현우**, 맞대결, 투지, 2002 월드컵, ...
군집 3	영화, **아이언맨**, **어벤져스**, 토르, 마블, 등장, **인피니티 워**, **타노스**, 캡틴 아메리카, 캐릭터, 시리즈, 개봉, 히어로, 로키, 스토리, 헐크, 아스가르드, 토니, 배경, 앤트맨, 윈터, 액션, 퓰니르, 라그나로크, 솔져, 변역, 자막, 죽음, 비전, 쿠키, 오브, 빌런, 토니 스타크, 루소, 닥터 스트레인지, 만화, 예고편, 시빌 워, 관객, 가모라, 스타크, 헬라, 버키, 악당, 코믹스, 결말, 애초, 복수, 팔콘, 소울, 블랙 위도우, 영화관, '블랙 팬서', ...

표 [2]는 Word2Vec를 이용하여 군집화 된 3개에 대한 결과 예시들을 보여주고 있다. 굵은 글씨로 표현된 단어들은 각 군집화시 4단계를 통하여 선정된 대표하는 단어들이다. 각 군집화에서 구성된 대부분의 단어들은 객관적으로 유사하다고 볼 수 있으며, 몇 개 의 단어들만이 의미적으로 유사하지 않는 모습들이 보이고 있다. 이것은 K-평균 알고리즘을 실행 시 그 결과 값이 수렴이 될 때 그 과정에서 포함된 노이즈 데이터로 보이며 이로 인하여 군집화 결과의 정확도가 떨어지는 결과가 나온 것으로 해석된다.

V. 결 론

본 논문에서는 주어진 한글 문서들을 기반으로 생성된 한국어 말뭉치 구축을 위하여 Word2Vec기법을 이용한 군집화 기법을 제안하였다. 주어진 문서의 단어를 임베딩하기 위하여 단순히 통계적인 방법을 기반으로 한 BOW 기법은 단어들 간의 의미적인 부분을 고려하지 않기 때문에 단어들 간의 유사도 측정이 부정확하게 된다. 반면 Word2Vec는 단어임베딩 시 Skip-gram 모델을 이용하여 문맥상 유사한 의미의 단어들을 가깝게 하여 벡터공간을 구축하기 때문에 유사도 측정시 정확한 결과를 보여주게 된다. 뿐만 아니라 생성된 각 군집화내에 존재하는 각 단어의 TF-IDF와 단어들 간의 유사도를 통하여 각 군집화를 대표하는 단어들을 선정하였다. 제안하는 기법의 우수성을 확인하기 위하여 수집된 약 6천여 개의 블로그들을 대상으로 군집화를 하였으며 기존 K-평균 군집화 알고리즘과의 비교 실험을 하였다. 비교 실험 결과 Word2Vec를 이용한 단어 군집화가 우수한 정확율과 재현율 결과를 보여주었다.

본 기법은 단어임베딩 시 명사만을 추출하기 때문에, 추후 형용사, 부사, 동사등과 같은 여러 개의 품사 들을 고려하여 한국어 군집화를 할 수 있는 기법들을 연구할 계획이며 특히 군집화시 무관한 단어들로 인한 노이즈들을 배제하기 위하여 Word2Vec기법을 개선하여 다양한 실험을 진행할 예정이다. 또한 한국어 말뭉치를 구축하기 위하여 블로그 뿐 만 아니라 신문기사, 트위터, 페이스북, 인스타그램과 같은 다양한 SNS에서의 데이터 수집을 통하여 방대한 양의 자료들을 기반으로 한 연구를 계속적으로 진행할 예정이다.

References

[1] M. Sun, H, Um, "The Study on Recent Research Trend in Korean Tourism Using Keyword Network

Analysis," Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 17, No. 9, pp. 68-73, 2016.

[2] E. Bae, S. Yu, "Keyword-based Recommender System Dataset Construction and Analysis, "Journal of KIIT. Vol. 16, No. 6, pp. 91-99, 2018. DOI : 10.14801/jkiit.2018.16.6.91.

[3] http://www.bloter.net/archives/260569

[4] Jae-Young Chang, "A Study on Research Trends of Graph-Based Text Representations for Text Mining", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 13, No. 5, pp. 37-47, Oct 2013.
DOI: http://dx.doi.org/10.7236/JIIBC.2013.13.5.37

[5] Shirai, Kiyoaki, and Makoto Nakamura. "JAIST: Clustering and classification based approaches for Japanese WSD." Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, pp. 379-382, 2010.

[6] Chen, Qian, Zengru Jiang, and Jinqiang Bian. "Chinese keyword extraction using semantically weighted network." In Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on, Vol. 2, pp. 83-86. IEEE, 2014.

[7] Xu, G. X., W. Sun, and X. P. Peng. "Clustering Research across Tibetan and Chinese Texts." Journal of Digital Information Management Vol. 13, No. 3, pp. 163-168, 2015

[8] Abuaiadah, Diab, Dileep Rajendran, and Mustafa Jarrar. "Clustering Arabic tweets for sentiment analysis." In Computer Systems and Applications (AICCSA), 2017 IEEE/ACS 14th International Conference on, pp. 449-456. IEEE, 2017.

[9] Sahmoudi, Issam, and Abdelmonaime Lachkar. "Formal Concept Analysis for Arabic Web Search Results Clustering." Journal of King Saud University-Computer and Information Sciences 29, No. 2, pp 196-203. 2017

[10] Copara, Jenny, Jose Ochoa, Camilo Thorne, and Goran Glavaš. "Exploring unsupervised features in Conditional Random Fields for Spanish Named Entity Recognition." In Intelligent Systems (BRACIS), 2016 5th Brazilian Conference, pp. 283-288. IEEE, 2016.

[11] https://ithub.korean.go.kr

[12] https://ilis.yonsei.ac.kr

[13] http://www.sejong21.org

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," In Proceedings of workshop at ICLR, pp. 1 - 12, 2013.

[15] M. Kim, T. Kang,"Proposal and Analysis of Various Link Architectures in Multilayer Neural Network,"Journal of KIIT. Vol. 16, No. 4, pp. 11-19, 2018. DOI : 10.14801/jkiit.2018.16.4.11

[16] Park, Eunjeong L., and Sungzoon Cho. "KoNLPy: Korean natural language processing in Python." Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology. pp. 133-136, 2014.

저자 소개

허 지 욱(정회원)

• 2007년 한림대학교 컴퓨터공학과 학사 졸업 . 2009년 한양대학교 컴퓨터공학과 석사 졸업, 2016년 한양대학교 컴퓨터공학과 박사 졸업

<주관심분야 : 멀티미디어 정보검색, SNS 분석, 단어 군집화, 다중문서요약 등>