http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.23, No.2, May 2018, pp.83-96
https://doi.org/10.7838/jsebs.2018.23.2.083

# Word2Vec 기반의 의미적 유사도를 고려한 웹사이트 키워드 선택 기법

## Web Site Keyword Selection Method by Considering Semantic Similarity Based on Word2Vec

이동훈(Donghun Lee)*, 김관호(Kwanho Kim)**

### 초 록

문서를 대표하는 키워드를 추출하는 것은 문서의 정보를 빠르게 전달할 수 있을 뿐만 아니라 문서의 검색, 분류, 추천시스템 등의 자동화서비스에 유용하게 사용 될 수 있어 매우 중요하다. 그러나 웹사이트 문서에서 출현하는 단어의 빈도수, 단어의 동시출현관계를 통한 그래프 알고리즘 등의 기반으로 키워드를 추출할 경우 웹페이지 구조상 잠재적으로 주제와 관련이 없는 다양한 단어를 포함하고 있는 문제점과 한국어 형태소 분석의 정확성이 떨어지는 형태소 분석기 성능의 한계점 때문에 의미적인 키워드를 추출하는데 어려움이 존재한다. 따라서 본 논문에서는 의미적 단어 위주로 구축된 후보키워드들의 집합과 의미적 유사도 기반의 후보 키워드를 선택하는 방법으로써 의미적 키워드를 추출하지 못하는 문제점과 형태소 분석의 정확성이 떨어지는 문제점을 해결하고 일관성 없는 키워드를 제거하는 필터링 과정을 통해 최종 의미적 키워드를 추출하는 기법을 제안한다. 실 중소기업 웹페이지를 통한 실험 결과, 본 연구에서 제안한 기법의 성능이 통계적 유사도 기반의 키워드 선택기법보다 34.52% 향상된 것을 확인하였다. 따라서 단어 간의 의미적 유사성을 고려하고 일관성 없는 키워드를 제거함으로써 문서에서 키워드를 추출하는 성능을 향상시켰음을 확인하였다.

### ABSTRACT

Extracting keywords representing documents is very important because it can be used for automated services such as document search, classification, recommendation system as well as quickly transmitting document information. However, when extracting keywords based on the frequency of words appearing in a web site documents and graph algorithms based on the co-occurrence of words, the problem of containing various words that are not related to the topic potentially in the web page structure, There is a difficulty in extracting the semantic keyword due to the limit of the performance of the Korean tokenizer. In this paper, we propose a method to select candidate keywords based on semantic similarity, and solve the problem that semantic keyword can not be extracted and the accuracy

이 논문은 2017년도 정부(미래창조과학부) 한국연구재단이 후원하는 개인기초연구사업(No. NRF-2017R1D1A1B03035639)과 한국연구재단이 후원하는 X-mind Corps 프로그램의 지원을 받아 수행된 연구임.

* Lead Author, Dept. of Industrial and Management Engineering, Incheon National University(dhlee@inu.ac.kr)
** Corresponding Author, Dept. of Industrial and Management Engineering Incheon National University (khokim@inu.ac.kr)

Received: 2018-04-11, Review completed: 2018-05-25, Accepted: 2018-05-29

of Korean tokenizer analysis is poor. Finally, we use the technique of extracting final semantic keywords through filtering process to remove inconsistent keywords. Experimental results through real web pages of small business show that the performance of the proposed method is improved by 34.52% over the statistical similarity based keyword selection technique. Therefore, it is confirmed that the performance of extracting keywords from documents is improved by considering semantic similarity between words and removing inconsistent keywords.

**키워드** : 키워드 추출, 키워드 선택, Word2Vec, 텍스트 마이닝, 의미적 유사도
Keyword Extraction, Keyword Selection, Word2Vec, Text Mining, Semantic Similarity

## 1. 서 론

문서를 대표하는 키워드를 추출하는 것은 문서의 정보를 빠르게 전달할 수 있을 뿐만 아니라 문서의 검색, 분류, 추천시스템 등의 자동화 서비스에 유용하게 사용 될 수 있어 매우 중요하다. 현재 문서의 키워드는 문서 상호간 비교를 위해 중요하게 활용되고 있으며[9], 텍스트 마이닝 분야의 전처리 과정에서 효율적인 속성으로 사용되고 있다[3].

키워드를 통한 자동화 서비스는 사용자가 의사결정을 하는데 도움을 줄 수 있다. 정보기술 발달로 인해 수많은 컨텐츠 속에서 사용자는 상품을 선택하여 구매하는 데에 어려움을 겪는다. 그러므로 사용자들은 쇼핑몰, 소셜 커머스 등의 사용자들이 반응하는 리뷰, 평점 정보를 통해 상품의 정보를 얻는다. 따라서 리뷰정보는 사용자의 선호도 정보를 포함하고 있어 리뷰정보의 키워드는 사용자가 좀 더 빠른 의사결정을 내리는데 바탕이 될 수 있는 중요한 도구의 역할을 할 수 있다.

그리고 사용자는 웹사이트 문서의 키워드 정보를 통해 짧은 시간에 문서의 정보를 파악 할 수 있다. 웹사이트에 존재하는 일반적인 기사, 뉴스 등의 문서는 발생한 사건이나 사실을 있는 그대로 알려주는 설명문으로써 정확한 정보만을 포함하고 있다. 또한 알기 쉬운 단어나 표현을 쓰므로 키워드를 통해 모든 사용자들이 이해할 수 있다. 그렇기 때문에 키워드 정보는 문서의 정보를 모든 사람이 이해할 수 있게 요약 제공할 수 있어 문서의 내용을 쉽게 파악할 수 있다.

그러나 문서의 키워드를 추출하기 위해서 많이 사용되고 있는 방법인 문서에서 출현하는 단어의 빈도수, 단어의 동시출현관계를 통한 그래프 알고리즘 등의 기반으로 키워드를 추출할 경우 문서의 키워드가 추출되지 않는 문제점이 있다. 웹사이트 문서에서는 구조상 <Title>, <Menu>, <Header> 태그 등에 포함 되어있는 의미 없는 구문이 반복적으로 등장한다. 따라서 잠재적으로 주제와 관련이 없는 다양한 단어를 포함하고 있어 기존의 방법으로 문서의 키워드를 추출할 경우 주제성을 내포하고 있지 않는 키워드가 추출되는 문제점이 발생한다.

게다가 형태소 분석기 성능의 한계점 때문에 한국어 형태소 분석의 정확성이 떨어진다[8, 14]. 형태소 분석은 기분석기 사전에 의해 분석되는데 현실적으로 모든 어절을 등록할 수는 없는 한계가 있다. 또한 기분석 말뭉치의 어절에서 부분어절을 자르고 자동으로 분석하여 알

아내는 것은 쉬운 일이 아니다[17]. 그러므로 형태소 분석과정을 통해 웹페이지 문서에서 의미적인 키워드를 추출하는데 어려움이 존재한다.

이와 같은 문제를 해결하기 위해 지도학습과 비지도 학습을 통한 키워드 추출이 대두되고 있다. 지도 학습이란 훈련데이터의 결과 값(label)을 미리 정하여 학습하는 방법을 말하며, 비지도 학습은 결과 값(label)을 먼저 구축하지 않고 학습하는 방법을 말한다[20]. 그러므로 많은 연구들이 시간과 비용에 면에서 효율적인 비지도 학습 기반의 키워드 추출 연구에 집중하고 있다.

Wen et al.[19]의 연구에서는 비지도 학습기반의 Word2Vec 모델을 사용하여 기존의 문서에 출현한 단어들로 구성된 그래프를 활용하여 키워드를 추출하였으며, Hu et al.[4]의 연구에서는 비지도 학습기반의 모델과 문서에서 단어의 출현빈도수 기반의 TF-IDF 기법과 분류기법에서 사용되는 로지스틱 리그레션(Logistic Regression)을 활용하여 키워드를 추출하였다. 하지만 이와 같은 기법들은 형태소 분석기의 한계점을 극복하지 못하였고 일관성 없는 키워드를 제거 하지 못하는 문제점이 발생하였다.

따라서 본 논문에서는 먼저 의미적 단어 위주로 구축된 후보키워드들의 집합과 Word Embedding 모델을 통해 의미적 유사도 기반의 후보 키워드를 선택함으로써 기존 연구에서 발생되었던 형태소 분석기의 정확성이 떨어지는 문제점과 주제성을 내포하지 않는 키워드가 추출되는 문제점을 해결한다. 또한 최종 필터링 과정을 통해 문서와 일관성이 없는 키워드를 제거하고 키워드 추출 성능을 향상 시킬 수 있는 의미적 유사도를 고려한 키워드 추출 기법을 제안한다.

제안된 기법은 크게 4가지 단계로 구성된다. 첫째, 웹사이트 문서에서 단어의 출현빈도수 기반으로 추출하는 프로파일 과정, 둘째, 단어를 특정 차원의 벡터 공간에 Embedding하기 위한 Word Embedding 과정, 셋째, 웹사이트 문서에서 단어의 출현빈도수에 의해 추출된 웹사이트 프로파일 중 관련 없는 단어를 제거하고 의미적 키워드로 변환하기 위한 키워드 선택 과정, 마지막으로 선택된 후보키워드 중 일관성 없는 후보키워드를 제거하기 위한 필터링 과정을 통해 최종 키워드를 추출한다.

본 논문의 구성은 다음과 같다. 제 2장에서 기존 연구에 대해서 언급하고, 제 3장에서 제안 기법에 대한 단계별 내용을 설명한다. 제 4장에서는 제안한 모델의 키워드 추출 성능 향상에 따른 타당성을 검증하고 마지막으로 제 5장에서 결론을 기술한다.

## 2. 관련 연구

키워드 추출과 관련된 기존연구는 제목, 방법론에 따라 <Table 1>과 같이 요약할 수 있다. 기존의 연구에서는 문서의 주요 키워드를 추출하기 위해 다양한 기법들이 연구 되었다. 예를 들어[10, 18]의 연구에서는 단어의 동시출현관계를 찾은 후 그 동시 출현 빈도를 저장하는 행렬(Co-occurrence Matrix)과 계층적 그래프 모델을 통해 키워드를 추출한다.

Cho and Lee[2]는 LDA 기법을 통한 문서의 잠재 키워드를 찾는 방법을 제안하였다. 주어진 문서와 유사한 문서의 키워드를 후보 키워드로 선택하고, 후보 키워드를 구성하는 개별 단어들의 등장 확률을 이용하여 중요도를 평가하였다.

〈Table 1〉 Previous Research Related on Keyword Extraction

| Analysis Methods | Feature Considered | References |
|---|---|---|
| Word Co-occurrence, Hierarchical Graph Model | Frequency | [18] |
| Word Co-occurrence | | [10] |
| Modified TF-IDF, Word Co-occurrence | | [6, 7] |
| Modified TF-IDF, Social Network | | [13] |
| LDA | | [2] |
| Modified Text Rank | | [16] |
| Text Rank Model, Word2Vec | Semanticity | [19] |
| TF-IDF, Graph Model, Word2Vec | | [1] |
| TF-IDF, Logistic Regression, Word2Vec | | [4] |

Lee and Kim[6, 7], Noh et al.[13] 연구에서는 기존의 TF-IDF 기법에서 벗어나 TF 값은 문서 집합 내에서의 단어의 출현빈도로 정의하고 정규화 하는 방식으로 TF-IDF를 변형한 키워드 추출방법 제시하여 더 나은 성능을 보였다.

Rose et al.[16] 연구에서는 기존에 사용되었던 그래프 기반의 Text Rank 알고리즘에서 변형을 주어 키워드 추출 시 후보 키워드의 길이가 길수록 가중치를 주어 추출함으로써 성능을 향상시켰다.

Cao et al.[1], Hu et al.[4], Song and Kim[18] 연구에서는 TF-IDF, Text Rank, Word Co-occurrence 등 기존의 많이 사용되었던 기법과 함께 비지도 학습 기반의 Word2Vec 기법을 사용하였다. 이 기법을 통해 통계적 유사도를 고려했던 방법과는 달리 단어의 의미적 유사도를 고려하여 키워드를 추출함으로써 성능을 향상시켰다.

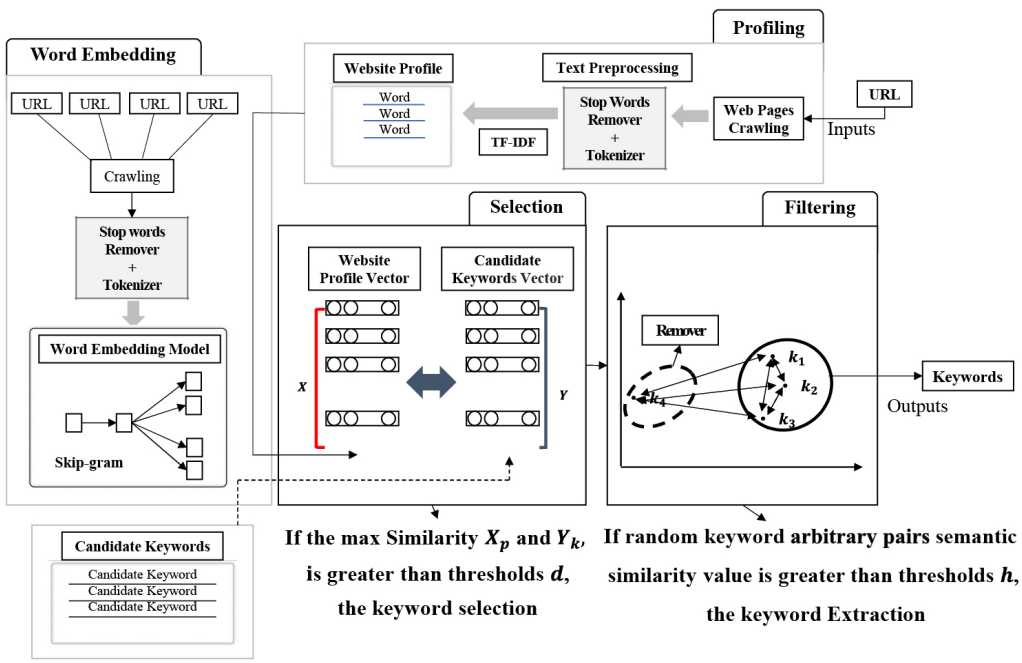
하지만 위의 기존 연구들에서는 키워드 추출 시 일관성 없는 키워드를 제거하기 위한 연구는 존재하지 않았으며 문서에서 함께 등장하는 단어와 출현 빈도수를 통해 키워드를 추출하는데 초점을 두고 있다.

따라서 본 연구에서는 키워드 추출의 성능을 향상시키기 위해 Word2Vec 기반의 Word embedding 모델을 구축하여 웹사이트 문서의 키워드를 출현빈도수에 의해 추출하고 일관성 없는 키워드를 제거하기 위한 의미적 유사도 기반의 키워드 선택 기법 프레임워크를 제안한다.

## 3. 제안 기법

제안된 기법은 <Figure 1>과 같이 크게 4단계로 나뉜다. 첫 번째, 프로파일 과정에서는 임의의 웹페이지에서 텍스트 데이터를 수집한다. 수집된 데이터들은 불용어 제거 및 형태소 분석기를 통해 전처리 과정을 마친 후 문서에 있는 단어의 출현 빈도수 기반으로 문서의 대표 키워드를 추출한다.

두 번째, 특정 차원의 벡터공간에서 단어 간 의미적 유사도를 계산할 수 있는 비지도 학습 기반의 Word Embedding 모델을 구축한다. 이를 위해 먼저 비정형 텍스트 데이터를 수집한다. 수집한 데이터는 불용어 제거 및 형태소




〈Figure 1〉 Proposed Framework

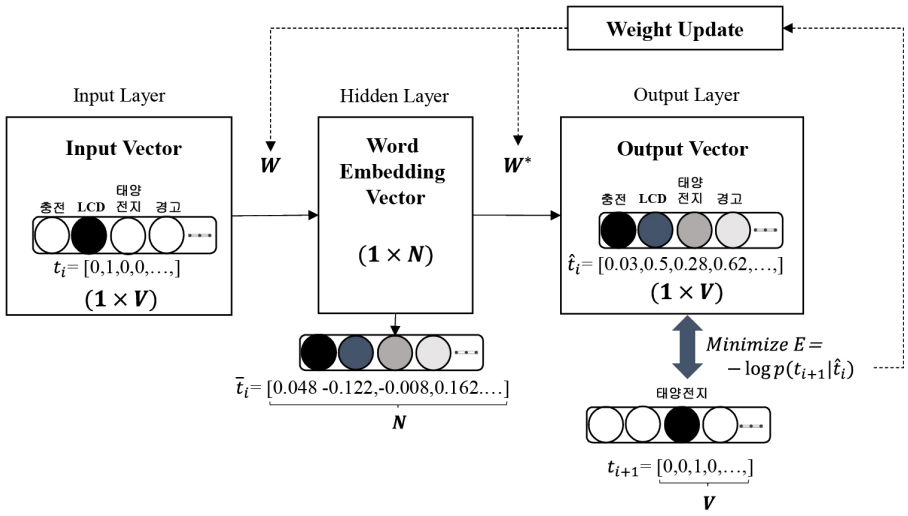
분석기를 통해 전처리 과정을 마친 후 명사 텍스트 데이터를 이용하여 파라미터 Iteration, Windows size, Dimension 각각 $A, B, C$로 설정하여 학습한다. Iteration은 학습 횟수, Windows size는 입력데이터와 함께 등장하는 단어의 개수, Dimension은 사용자가 Embedding 하려는 차원의 수를 나타낸다.

세 번째, 키워드 선택 과정에서는 Word Embedding 모델을 통해 웹사이트 프로파일과 후보키워드와의 의미적 유사도 계산을 통해 후보키워드를 선택한다. 프로파일 과정에서 추출 된 키워드는 주제성을 내포하고 있지 않는 키워드가 포함되어 있을 수 있기 때문에 후보 키워드를 선택함으로써 주제성을 내포하고 있지 않는 키워드를 제거한다. 후보 키워드는 사전에 의미적 키워드 위주로 구축하였다.

마지막으로 필터링 과정에서는 키워드 선택 과정에서 선택 된 키워드 중 후보 키워드와의 유사도는 높지만 문서와는 일관성이 없는 키워드가 선택될 수 있다. 따라서 선택 된 후보 키워드 중 일관성 없는 키워드를 제거하기 위해서 모든 후보키워드 쌍의 대해 의미적 유사도를 계산하고 일관성 없는 키워드를 제거하고 최종 키워드를 추출한다.

## 3.1 프로파일링

프로파일 과정은 크롤링 기법을 통해 웹사이트 문서 텍스트 데이터를 수집한다. 텍스트 데이터들은 분석에 필요한 데이터로 정제하기 위해 불용어 제거 및 형태소 분석기를 통해 전처리 추출 된 단어는 TF-IDF 기법[5, 15]을 통해 추출된 웹페이지의 대표키워드 집합인 $X$를 구성한다.




〈Figure 2〉 Word Embedding Model for "LCD"

### 3.2 Word Embedding 모델

Word Embedding 모델은 Word2Vec 기반의 Skip-gram 모델을 사용하였으며, 이 모델은 단어로부터 등장하는 문맥을 유추하는 방법으로 학습해 나가는 비지도 학습 기반의 모델이다.

Skip-gram 모델은 입력층, 은닉층, 출력 층으로 구성되어있다. <Figure 2>는 Word embedding 과정을 나타낸 그림이다. 입력 값으로는 $i$번째, one-hot-vector $t_i, i=1,\cdots, V$로 입력된다. one-hot-vector는 $i$번째가 1이고 나머지는 0인 행렬로 구성되며 $V$는 총 고유 단어의 수이다.

입력층에서 은닉층을 잇는 은닉층 가중치 $W$는 $V\times N$행렬로 구성되며 $N$은 사용자가 Embeeding하려는 차원의 수이다. 은닉층에서 출력층을 잇는 출력층 가중치 $W^*$는 $N\times V$행렬로 구성되어 있다[11].

입력층에서 주어진 $t_i$는 $W$와의 계산을 통해 $N$인 행벡터 $\bar{t}_i$가되고 $\bar{t}_i$는 $W^*$와의 계산을 통해 $V$인 행벡터가 된다. 마지막으로 출력층에서 식 (1)과 같이 Softmax 계산을 통해 $\hat{t}_i$가 나올 확률 값을 계산한다. 식 (1)은 입력 값이 $t_i$일 때, $V$개의 단어 중 $\hat{t}_i$가 나올 확률 값을 나타낸다. 식 (2)는 $\hat{t}_i$와 $t_{i+1}$와의 조건부 확률을 나타낸다. 식 (2)를 통해 $\hat{t}_i$와 $t_{i+1}$와의 차이를 최소화 하는 $W, W^*$의 값을 찾기 위해 반복적으로 학습한다[11, 12].

$$P(\hat{t}_i)=\frac{e^{t_i}}{\sum_{q=1}^{V} e^{t_q}}, q=1,\cdots,V \qquad (1)$$

$$E=-\log p(t_{i+1}|\hat{t}_i) \qquad (2)$$

### 3.3 키워드 선택 과정

<Figure 3>과 같이 키워드 선택 과정은 먼저 사전에 구축된 Word Embedding 모델을 통해 의미적 유사도 기반의 키워드를 선택하는 과정이다. 먼저 프로파일 과정에서 추출된

| | $Y_1$ | $Y_2$ | $Y_3$ | ··· | $Y_k$ | ··· | $Y_m$ |
|---|---|---|---|---|---|---|---|
| $X_1$ | 0.211 | 0.381 | **0.687** | ··· | 0.434 | ··· | 0.549 |
| $X_2$ | 0.512 | 0.222 | 0.349 | ··· | **0.721** | ··· | 0.432 |
| $X_3$ | **0.811** | 0.633 | 0.189 | ··· | 0.218 | ··· | 0.351 |
| ⋮ | ⋮ | ⋮ | ⋮ | ··· | ⋮ | ··· | ⋮ |
| $X_p$ | $Sim(X_p,Y_1)$ | $Sim(X_p,Y_2)$ | $Sim(X_p,Y_3)$ | ··· | $Sim(X_p,Y_k)$ | ··· | $Sim(X_p,Y_m)$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ··· | ⋮ | ··· | ⋮ |
| $X_n$ | $Sim(X_n,Y_1)$ | $Sim(X_n,Y_2)$ | $Sim(X_n,Y_3)$ | ··· | $Sim(X_n,Y_k)$ | ··· | $Sim(X_n,Y_m)$ |

$$\begin{cases} \boldsymbol{argmax\ Sim(X_p,Y_k)} & \boldsymbol{if\ the\ Similarity\ between\ X_p\ and\ Y_k > Thresholds\ d} \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases}$$

〈Figure 3〉 Examples of Keyword Selection based on Semantic Similarity

집합 $X$ 중 $p$번째 프로파일 $X_p$, $p=1, \cdots, n$와 후보키워드 집합 $Y$ 중 $k$번째 후보키워드 $Y_k$, $k=1, \cdots m$는 $N$차원의 벡터 공간에서 식 (3)과 같이 유사도 점수를 계산한다. $X_p \cdot Y_k$는 두 단어 벡터간의 내적 값을 뜻한다. $|X_p|, |Y_k|$ 각각은 벡터의 크기를 뜻하며, $|X_p \| Y_k|$는 두 벡터크기의 곱셈 계산을 뜻한다.

$$Sim(X_p, Y_k) = \frac{X_p \cdot Y_k}{|X_p \| Y_k|} \qquad (3)$$

만약 임계 값 $d$보다 큰 유사도 점수 값을 갖는 $Y_k$ 중 가장 큰 유사도 값을 갖는 $Y_{k^*}$를 후보키워드로 판단하고 선택한다. 모든 $X_k$와 $Y_p$쌍과의 키워드 선택 과정을 실시해 $Y_{k^*}$의 집합인 $Z$를 구성한다.

## 3.4 필터링

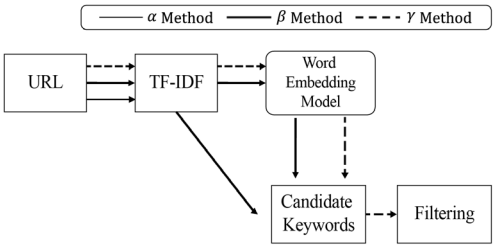
<Figure 4>에서 설명하는 바와 같이 필터링 과정을 통해 최종 키워드를 추출 한다. 입력 데이터는 $Z$이며, 출력 데이터는 최종 키워드 집합 $Z'$이다. 서로 다른 모든 임의의 $Y_{k'a}$, $Y_{k'b}$, $a, b = 1, \cdots, l$쌍에 대하여 식 (3)과 같은 방법으로 $Sim(Y_{k'a}, Y_{k'b})$ 계산을 통해 임계 값 $h$ 이상일 경우 점수를 계산하여 집합 $Z'$에 점수와 $Y_{k'a}$를 포함시키고 $Y_{k'a}, Y_{k'b}$가 같을 경우와 $h$ 이하일 경우에는 점수를 0으로 계산한다. 최종적으로 $Z'$에 포함된 $Y_{k'a}$를 최종 키워드로 정의하고 추출한다.

**Input** : Set of Z
**Output** : Final extracted keywords Z′
**Begin** keyword index $a, b = 1$
**Repeat**
  **If** $Y_{k'a} \neq Y_{k'b}$, $sim\,(Y_{k'a}Y_{k'b}) > Thresholds\ h$
    score = $sim\,(Y_{k'a}Y_{k'b})$
    **Contain** the corresponding keyword $Y_{k'a}$, and score into List $Z'$
  **End if**
  **If** $Y_{k'a} = Y_{k'b}$
    score = 0
  **End if**
  **If** $Y_{k'a} \neq Y_{k'b}$, $sim\,(Y_{k'a}Y_{k'b}) < Thresholds\ h$
    score = 0
**Until** length of $l$
**End**

〈Figure 4〉 Proposed Filtering Algorithm

### 3.5 키워드 선택 기법들

본 연구에서 제안한 기법을 응용하여 <Figure 5>와 같이 $\alpha$-Method, $\beta$-Method, $\gamma$-Method를 구현하였다. $\alpha$-Method는 후보키워드와의 통계적 유사도를 통한 키워드를 선택하는 방법이다. $\beta$-Method는 Word Embedding 모델을 통해 특정 벡터공간에서 후보키워드와의 의미적 유사도를 통한 키워드를 선택하는 방법이다. 마지막으로 $\gamma$-Method는 의미적 유사도를 통한 키워드를 선택 한 후 그 중 일관성 없는 키워드를 제거하고 키워드를 추출하는 방법이다.




〈Figure 5〉 Proposed Methods

## 4. 실험 및 평가

### 4.1 실험 환경

Word Embedding 모델 구축을 위한 데이터는 <Table 2>와 같이 한국 위키피디아, 나무위키, 중소기업뉴스, 중소기업 웹사이트 문서를 이용하였으며, 1,092,523개의 문서의 텍스트 데이터를 수집하였다. 텍스트 데이터들은 사용자 사전을 포함한 형태소 분석기를 통해 명사 데이터만을 추출하였다.

추출 된 명사 데이터는 $A, B, C$ 각각 5, 10, 200으로 설정하여 학습함으로써 Word Embedding 모델을 구축하였다. 사전에 구축 된 후보키워드 집합 $Y$는 통계분류포털 웹사이트에 있는 산업분류표와 색인어를 참고하여 의미적 단어 1,500개 위주로 구성하였다. 웹사이트 프로파일 집합 $X$는 12개를 추출하였다.

〈Table 2〉 Dataset for Words Embedding Model

| Corpus Documents | Number of Document |
|---|---|
| Wikipidia | 395,937 |
| Namu Wiki | 501,686 |
| Small Business News | 44,900 |
| Small Business Web Document | 150,000 |
| Total | 1,092,523 |

키워드 선택 과정에서 임계값 $d$는 경험의 의해 최적의 값으로 0.48로 설정하였다. $d$가 커질수록 관련 있는 후보키워드를 선택하지 못하는 경우가 발생하고, 작아질수록 관련 없는 후보키워드가 많이 선택되는 문제점이 발생하여 경험의 의해 최적의 값으로 설정하였다. 필터링 과정에서의 임계 값 $h$는 예비 실험에 의해 결정하였으며[10], 0.3에서 0.5로 설정하여 0.01간격으로 변화를 주어 실험하였다.

본 연구에서 제안한 기법의 성능을 평가하기 위해 50개의 중소기업 웹사이트를 무작위로 선택하여 실험데이터로 설정하고 2가지 방법을 통해 성능 평가를 실시한다. 첫 번째, $\alpha$, $\beta$, $\gamma$-Methods에 따른 성능을 비교한다. 두 번째로는 $h$값 변화에 따른 $\gamma$-Method 정밀도를 통해 성능을 평가한다.

웹사이트 문서에서 키워드는 정량적으로 평가하기 어려우므로 웹사이트 문서의 키워드라고 판단할 수 있는 키워드를 정성적으로 평가하여 선택하였다. 정밀도(Precision) 방법은 식(5)에서 $TP$는 실제 값의 클래스 값과 예측 값

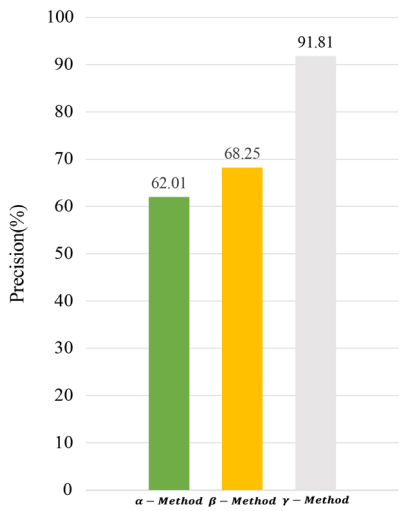
의 클래스 값이 일치 했을 때의 값을 나타내며 $FP$는 실제 값의 클래스 값과 예측 값의 클래 값이 불일치했을 때의 값을 나타낸다. 따라서 식 (5)와 같이 키워드 추출의 성능은 정밀도 관점에서 평가된다.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

## 4.2 키워드 추출 결과

<Figure 6>에서 볼 수 있듯이 $\alpha$-Method, $\beta$-Method, $\gamma$-Method 각각 62.01%, 68.25%, 98.01%의 성능을 보였다. <Table 3>과 같이 실험을 통해 $\beta$-Method는 $\alpha$-Method보다 10.06%의 성능을 향상시켰음을 알 수 있었고 본 연구에서 제안한 $\gamma$-Method는 $\beta$-Method보다 34.52%의 성능을 향상시켰음을 확인하였다. 이를 통해 단어의 의미를 고려하여 키워드를 선택하여 추출하는 기법이 더 높은 성능을 보인다는 것을 알 수 있었으며 그 중 일관성 없는 키워드를 제거하고 키워드를 추출 할 경우 더 높은 성능을 보인다는 것을 확인하였다.

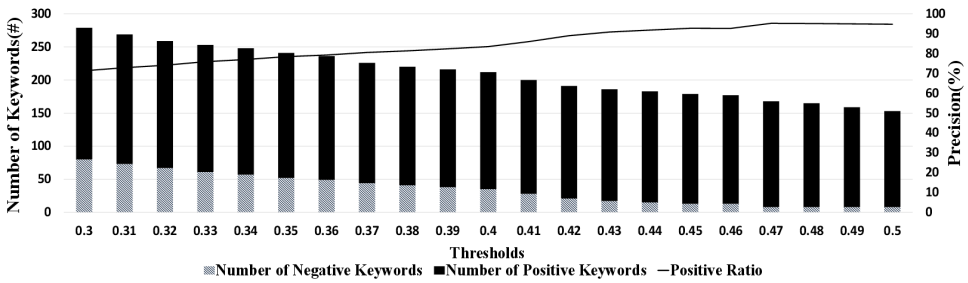
다음으로 <Figure 7>의 결과는 $h$값 변화에 따른 정밀도 결과와 일관성 키워드 개수의 변화를 나타낸다. Positive, Negative, Ratio 각각은 일관성 있는 키워드, 일관성 없는 키워드, 정답률을 나타낸다. 먼저 $h$값이 0.3일 경우 Ratio는 71.33%를 보이며 Positive의 개수는 199개, Negative의 개수는 80개를 나타낸다. $h$값이 증가할수록 Ratio는 증가하고 Positive, Negative의 개수는 감소하였다. 그러나 $h$값이 0.47부터는 Ratio, Positive는 감소하지만 Negative는 일정하게 유지되었다.




〈Figure 6〉 $\alpha$, $\beta$, $\gamma$ of Keywords Extraction Performance by Methods Type

〈Table 3〉 Improved Ratio of Performance Compared to $\alpha$, $\beta$, $\gamma$

| $\alpha$-Method | $\beta$-Method | $\gamma$-Method |
|---|---|---|
| 62.01% | 10.06% 향상 | 34.52% 향상 |




〈Figure 7〉 Precision Result and Numbers of Extracted Keywords Under Thresholds

또한 $h$값이 0.46일 때, 특정 웹사이트 문서에서의 Positive가 모두 제거되는 문제점이 발생하였다. Ratio 측면에서는 증가하고 있어 키워드 추출 성능이 향상된 것으로 보이나 $h$값이 0.46보다 작을 때 웹사이트 문서에서 의미적인 키워드를 추출하였음에도 불구하고 $h$값을 증가시키는 것으로 인해 문서의 키워드가 추출되지 않는 결과가 발생하는 경우를 실험을 통해 확인하였다.

이를 통해 본 연구에서 제안한 모델은 사용자의 환경에 따라 적절한 $h$값을 설정하는 것이 중요하며 $h$값을 자유롭게 설정 할 수 있는 유동성을 가지고 있는 장점이 있는 것을 확인하였다. 그러나 $h$값이 0.47 이상 일 경우 Negative가 제거 되지 않고 Positive가 제거되는 되는 문제가 발생하였고, $h$값이 0.46일 때 특정 웹사이트 문서에서 키워드가 모두 제거되는 문제점이 발생하였으므로 본 연구에서 제안한 모델은 $h$값을 0.46 이하로 설정하는 것이 중요하다는 것을 실험을 통해 확인하였다.

## 4.3 웹사이트 문서 적용 사례

<Table 4>는 중소기업 웹사이트 실 데이터에 적용한 결과이다. $\alpha$-Method는 프로파일 과정과 같이 단어의 출현빈도수 기반으로 키워드

〈Table 4〉 Examples of the Selected Keywords by Using the Proposed Methods for Company Web Sites(Bold and Underline Means Correct Keywords)

| Company Names | Website URLs | Methods Applied | | |
|---|---|---|---|---|
| | | $\alpha$-Method | $\beta$-Method | $\gamma$-Method |
| 대영합판㈜ | http://www.daeyoung.com | **MDF**, **합판**, 영상기기, 모니터 | 솔레노이드, **합판**, **MDF**, 기술개발, 모니터, 번역 | **MDF, 합판** |
| 비노텍㈜ | http://www.vinotec.co.kr | 정보검색, **폐기물**, 보호장비, **슬러지**, 청소, 정보관리시스템, 소각로, 교통시설물, 기록계 | **폐기물**, **슬러지**, 미술품, **분쇄기**, 건설, 설계 | 설계, **슬러지**, 건설, **폐기물** |
| 서울아스콘㈜ | http://saholdings.co.kr | **아스콘, 레미콘**, 정보검색, 기술연구 | **아스콘**, **레미콘**, 전선, 라디오, 대학교, 수산 | **아스콘**, **레미콘** |
| 명성금속 | http://www.safetypin.co.kr | 정보검색, **제조설비, 머리핀,** 보호장비, **옷핀** | **제조설비**, 서스펜스, **머리핀**, 법무사, **옷핀** | None |
| 성광기계 | http://www.skct.net | **냉각탑**, 매트리스, 포장기, **원형냉각탑**, 파이프, **냉각수**, 정보시스템, 설계 | **냉각탑**, **원형냉각탑**, 금융, 물리학, 미술품, **냉각수**, 타이어, 설계 | **냉각탑, 원형냉각탑, 냉각수** |
| 세무법인정석 | https://jstax.modoo.at | **세무회계, 법무법인**, 검색서비스, 노무사, 전기장판, 음성인식 | **세무회계**, 수산, 클라우드, **세무사** | **세무회계, 세무사** |
| (주)에어텍 | http://www.ateng.co.kr | **냉난방기**, 현장, 탄소, 원문, 상쇄, 중립, 마크, **항온항습기** | **냉난방기**, 리튬, 번역, **항온항습기**, 신물, 재생에너지 | **냉난방기, 항온항습기** |
| ㈜지씨아이글로벌 | http://gciglobal.co.kr | **슬라이드**, 슬라이딩, **거울**, 설명서, **조명, LED**, 스토리, 갤러리, 패턴, **글라스**, 군내, 마무리 | **슬라이드**, 범퍼, 렌즈, 시트, **조명**, 정보검색, 서스펜스, 미술품, **글라스**, **LED** | 시트, **LED**, 범퍼, **조명**, 렌즈, **글라스** |
| 피플라이프 금융센터 K지점 | www.peoplelife.co.kr | 나이프, **보험**, 법무법인, 정보검색, **재무관리**, 종합병원, 경영지원, **컨설팅** | **보험**, 수산, **재무관리**, 금융, **부동산, 컨설팅** | **부동산, 컨설팅, 재무관리, 보험**, 금융 |
| 승화정밀 | http://www.ishpre.co.kr | **도금**, 크롬강, **기계부품**, **연마**, **금속**, 표면처리, **금형제작** | 현미경, 철판, 신문, **도금, 알루미늄, 기계부품, 연마, 금속** | 현미경, 철판, 알루미늄, **연마, 도금, 금속** |

를 추출하였을 때의 결과를 나타낸다. 웹사이트 문서와 관련 없는 단어와 일관성 없는 단어가 많이 추출 되는 문제점이 발생하였다. $\beta$-Method는 키워드 선택 과정과 같이 후보 키워드와의 의미적 유사도를 기반으로 후보 키워드를 선택하여 추출 된 결과를 나타낸다. 웹사이트 문서와 관련 없는 단어들은 많이 제거되고 추출되었지만 일관성 없는 단어들이 추출 되는 문제가 발생하였다. 마지막으로 $\gamma$-Method 필터링 과정을 통해 일관성 없는 키워드를 제거하고 추출함으로써 가장 높은 성능을 보였다. 따라서 본 연구에서 제안한 기법이 웹사이트 문서에서 키워드를 추출하는데 가장 높은 성능을 보이는 것을 확인하였다. 그러나 명성금속 기업에서는 필터링 과정에서 키워드간의 의미적 유사도가 설정한 $h$값 보다 낮아 키워드가 추출되지 않는 문제점이 발생하였다. 또한 지씨아이글로벌과 승화정밀 기업에서는 일관성 없는 키워드가 $h$값 보다 높아 추출되는 문제점이 발생하였다.

## 5. 결 론

본 연구에서는 비지도 학습기반의 Word Embedding Model을 통한 의미적 유사도를 고려한 키워드 선택 기법을 제안한다. 웹사이트에서 문서에서 단어의 출현빈도수에 의해 추출된 키워드 중 관련 없는 단어와 일관성 없는 단어를 제거하고 의미적 키워드를 추출함으로써 키워드 추출 성능을 향상 시킬 수 있다. 본 연구에서 제안한 기법의 성능을 측정하기 위해 $\alpha$, $\beta$, $\gamma$-Methods에 따른 성능 비교 실험과 $\gamma$-Method의 $h$값 변화에 따른 성능의 변화 실험을 진행하였다. 실험을 통해 본 연구에서 제안한 기법이 보다 높은 성능을 보임을 증명하였고, $h$값 변화에 따른 성능의 차이가 있다는 것을 확인하였다.

이를 바탕으로 본 연구에서 제시한 키워드 추출 기법은 문서의 검색, 분류, 추천시스템 등의 자동화서비스에 유용하게 활용 될 수 있을 것이다. 게다가 뉴스, 기사 등의 키워드를 제공하는 서비스에 활용함으로써 구독자에게 편리함을 제공하거나 키워드 정보를 기반으로 트렌드 분석에 활용할 수 있을 것이다.

향후 연구에서는 제안한 기법을 보완하기 위해 단어 간 의미적 유사도를 더 정확하게 계산하는 방법, 그리고 특정문서에서 일관성 키워드가 제거되지 않는 문제점을 보완하기 위해 특정문서가 아닌 모든 문서에서일관성 없는 키워드를 제거하기 위한 기법 등에 대해 연구할 예정이다.

## References

[1] Cao, J., Jiang, Z., Huang, M., and Wang, K., “A Way to Improve Graph-Based Keyword Extraction,” Proceedings of IEEE International Conference on Computer and Communications, pp. 166-170, 2015.

[2] Cho, T. and Lee, J.-H., “Latent Keyphrase Extraction Using LDA Model,” Journal of Korean Institute of Intelligent Systems, Vol. 25, No. 2, pp. 180-185, 2015.

[3] Choi, D. J., Lee, S. W., Kim. J. K., and Lee, J. H., “A Study on Graph-Based

Topic Extraction from Microblogs," Journal of Korean Institute of Intelligent Systems, Vol. 21, No. 5, pp. 564-568, 2011.

[4] Hu, J., Jin, F., Zhang, G., Wang, J., and Yang, Y., "A User Profile Modeling Method Based on Word2Vec," Proceedings of IEEE International Conference on Software Quality, Reliability and Security Companion, pp. 410-414, 2017.

[5] Lee, K-H., Lee, K-C., and Kim, K-Ok., "Ranked Web Service Retrieval by Keyword Search," The Journal of Society for e-Business Studies, Vol. 13, No. 2, pp. 213-223, 2008.

[6] Lee, S. and Kim, H. J., "News Keyword Extraction for Topic Tracking," Proceedings of IEEE Networked Computing and Advanced Information Management, Vol. 2, pp. 554-559, 2008.

[7] Lee, S.-J. and Kim, H-J., "Keyword Extraction from News Corpus using Modified TF-IDF," The Journal of Society for e-Business Studies, Vol. 14, No. 4, pp. 59-73, 2009.

[8] Lee, Y. J., "Korean Morphological Analysis Algorithmas for Automatic Idexing," Proceedings of the Annual Conference on Human and Cognitive Language Technology, pp. 240-246, 1989.

[9] Lott, B., "Survey of Keyword Extraction Techniques," UNM Education, 2012.

[10] Matsuo, Y. and Ishizuka, M., "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information," International Journal of Artificial Intelligence Tools, Vol. 13, No. 1, pp. 157-169, 2004.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., "Distributed Representations of Words and Phrases and Their Compositionality," Advances in Neural Information Processing Systems, pp. 3111-3119, 2013.

[12] Mikolov, T., Chen, K., Corrado, G., and Dean, J., "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv, pp. 1301-3781, 2013.

[13] Noh, Y., Lim, J., Bok, K., and Yoo, J., "Hot Topic Prediction Scheme using Modified TF-IDF in Social Network Environments," Journal of Korean Institute of Information Scientists end Engineers, Vol. 23, No. 4, pp. 217-225, 2017.

[14] Oh, J. Y. and Cha, J. W., "High Speed Korean Dependency Analysis using Cascaded Chunking," Journal of the Korea Society for Simulation, Vol. 19, No. 1, pp. 103-111, 2010.

[15] Robertson, S. E., "Term Specificity," Journal of Documentation, Vol. 28, No. 1, pp. 164-165, 1972.

[16] Rose, S., Engel, D., Cramer, N., and Cowley, W., "Automatic Keyword Extraction from Individual Documents, Text Mining: Applications and Theory," pp. 1-20, WILEY, 2010.

[17] Shin, J.-C. and Ock, C.-Y., "A Korean Morphological Analyzer using a Pre-ana-

lyzed Partial Word-phrase Dictionary," Journal of Software and Applications, Vol. 39, No. 5, pp. 415-424, 2012.

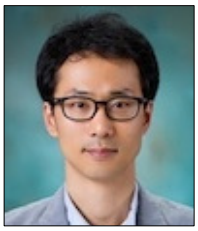
[18] Song, G. H. and Kim, Y.-S., "Automatic Keyword Extraction using Hierarchical Graph Model Based on Word Co-occurrences," Journal of Korean Institute of Information Scientists end Engineers, Vol. 44, No. 5, pp. 522-536, 2017.

[19] Wen, Y., Yuan, H., and Zhang, P., "Research on Keyword Extraction Based on Word2-Vec Weighted TextRank," Proceedings of IEEE International Conference on Computer and Communications, No. 2, pp. 2109-2113, 2016.

[20] Yarowsky, D., "Unsupervised word sense disambiguation rivaling supervised methods," Proceedings of the Association for Computational Linguistics, pp. 189-196, 1995.

## 저 자 소 개

이동훈 (E-mail: dhlee@inu.ac.kr)
2016년 단국대학교 산업공학과 (학사)
2017년~현재 인천대학교 산업경영공학과 (석사과정)
관심분야 텍스트 마이닝, 통계적 기계학습



김관호 (E-mail: khokim@inu.ac.kr)
2006년 동국대학교 정보시스템전공 (학사)
2012년 서울대학교 산업공학과 (박사)
2013년 경희대학교 (연구박사)
2014년~현재 인천대학교 산업경영공학과 교수
관심분야 통계적 기계학습, 빅데이터