Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 2, April 2015, pp. 180-185
http://dx.doi.org/10.5391/JKIIS.2015.25.2.180

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

LDA 모델을 이용한 잠재 키워드 추출

Latent Keyphrase Extraction Using LDA Model

조태민 · 이지형†
Taemin Cho, and Jee-Hyong Lee†

성균관대학교 정보통신대학 전자전기컴퓨터공학과
Department of Electrical and Computer Engineering, Sungkyunkwan University

요 약

인터넷 미디어의 발달과 함께 온라인 문서의 양이 급격하게 증가함에 따라, 문서 요약과 정보 검색 등 다양한 분야에 활용가능한 키워드를 자동으로 찾고자하는 연구가 활발히 진행되고 있다. 하지만 기존의 키워드 추출 연구들은 문서에서 나타나는 키워드만을 대상으로 하고 있어, 문서에서 등장하지 않는 잠재 키워드를 추출하지 못하는 한계를 갖고 있다. 잠재 키워드는 실데이터 키워드의 1/4 이상을 차지하고 있으며, 문서에서 나타나지는 않지만 문서의 중요한 개념이나 내용을 함축하고 있어 문서 요약 및 정보 검색에 중요한 역할을 차지할 수 있다. 특히 SNS와 같이 내용이 적어 키워드가 명시적으로 나타나기 어려운 문서에서 유용하게 활용될 수 있다. 본 논문에서는 잠재 키워드를 추출하기 위해 주어진 문서와 유사한 문서의 키워드를 후보 키워드로 선택하고 후보 키워드를 구성하는 개별 단어들을 이용해 후보 키워드의 중요도를 평가하는 방법을 제안한다. 실험을 통해, 제안 기법이 잠재 키워드를 합리적인 수준으로 추출할 수 있음을 보였다.

키워드 : 잠재 키워드, 잠재 디리클레 할당(LDA), 키워드 추출, 이웃 문서

Abstract

As the number of document resources is continuously increasing, automatically extracting keyphrases from a document becomes one of the main issues in recent days. However, most previous works have tried to extract keyphrases from words in documents, so they overlooked latent keyphrases which did not appear in documents. Although latent keyphrases do not appear in documents, they can undertake an important role in text summarization and information retrieval because they implicate meaningful concepts or contents of documents. Also, they cover more than one fourth of the entire keyphrases in the real-world datasets and they can be utilized in short articles such as SNS which rarely have explicit keyphrases. In this paper, we propose a new approach that selects candidate keyphrases from the keyphrases of neighbor documents which are similar to the given document and evaluates the importance of the candidates with the individual words in the candidates. Experiment result shows that latent keyphrases can be extracted at a reasonable level.

Key Words : Latent Keyphrase, Latent Dirichlet Allocation(LDA), Keyphrase Extraction, Neighbor Document

Received: Sep. 14, 2014
Revised : Sep. 28, 2014
Accepted: Mar. 24, 2015
†Corresponding author(john@skku.edu)

이 논문은 2014년도 미래창조과학부의 재원으로 한국연구재단-차세대정보 컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. NRF-2014M3C4A7030503). 또한 미래창조과학부 및 정보통신기술진흥센터(IITP)의 SW컴퓨팅 산업융합원천기술개발사업의 일환으로 수행되었음 [B0101-15-0559, 디지털 소상공인 지원을 위한 지역 비즈니스 전략 분석 및 맞춤형 영상 홍보 창작 SW 플랫폼 개발]

1. 서 론

인터넷 미디어의 발달과 함께 온라인 문서의 양이 급격하게 증가함에 따라, 이로부터 유용한 정보를 획득하는 기술의 필요성이 증대되고 있다. 키워드는 문서의 내용을 나타내는 가장 간결한 표현 방법으로, 적합한 키워드는 문서의 내용을 정확하게 함축하여 나타낼 수 있다. 또한 이는 문서 요약(Text summarization)과 정보 검색(Information retrieval)과 같은 텍스트 마이닝(Text mining) 분야에서의 주요 속성으로 활용될 수 있다. 하지만 극소수의 문서들만이 저자가 부여한 키워드를 갖고 있으며, 각 문서의 키워드를 제 3자가 지정하는 일은 신뢰성이 낮고 어렵다. 따라서 주어진 문서에서 자동으로 키워드를 추출하는 연구가 활발히 진행되고 있다[1-16].

대부분의 키워드 추출 연구들은 크게 두 단계 과정을 거쳐 키워드를 선별하게 된다. 이들은 주어진 문서에서 후보 키워드를 선택한 후, 후보 키워드가 등장한 위치와 횟수를 기반으로 후보 키워드의 중요성을 평가하여 키워드를 추출한다. 하지만 기존 방법들은 주어진 문서에서 나타나는 구문만을 후보 키워드로 선택하기 때문에 문서에서 나타나지 않는 구문은 선택하지 못하는 한계가 있다. 또한 후보 키워드를 평가하는 데 있어, 후보 키워드가 주어진 문서에서 나타나는 것을 가정하고 있기 때문에 문서에서 나타나지 않는 후보 키워드를 평가하기에도 적합하지 않다. 따라서 기존 방법들은 문서에서 나타나지 않는 잠재 키워드를 추출하

기 어렵다. 잠재 키워드는 실데이터 [2, 17-18] 키워드의 1/4 이상을 차지하고 있으며, 문서에서 나타나지는 않지만 문서의 중요한 개념이나 내용을 함축하고 있어 문서 요약 및 정보 검색에 중요한 역할을 차지할 수 있다. 특히 SNS와 같이 내용이 적어 키워드가 명시적으로 나타나기 어려운 문서에서 유용하게 활용될 수 있다.

따라서 본 논문에서는 주어진 문서의 잠재 키워드를 찾고자, 문서에서 나타나지 않는 구문을 후보 키워드로 선택하고 이를 평가하는 방법을 제안한다. 이를 위해, 주어진 문서와 유사한 문서들을 구성하여 신뢰성이 높은 정보들을 얻고, 이들의 키워드를 후보 키워드로 선택한다. 이와 같이 선택된 후보 키워드는 주어진 문서에서 나타나지 않으므로 후보 키워드 자체를 평가하지 않고, 후보 키워드를 구성하는 개별 단어들을 이용해 후보 키워드의 중요성을 평가한다. 후보 키워드의 선택 및 평가에는 LDA를 기반으로 생성된 주제 모델을 활용한다.

본 논문의 구성은 다음과 같다. 2장에서는 키워드 추출 기법 연구들을 소개하고, 3장에서는 제안한 방법의 배경 지식에 대해 기술하고, 4장에서는 본 연구에서 제안하는 방법을 서술한다. 5장에서는 실험을 통하여 제안한 방법의 효용성을 분석하고, 6장에서는 결론을 도출하고 향후 연구에 대해 논의한다.

2. 관련 연구

이 장에서는 키워드 추출과 관련된 기존 연구들과 그 한계점에 대해 논의한다. 앞 장에서 언급했듯이 대부분의 키워드 추출 연구는 크게 후보 키워드 선택과 후보 키워드 평가를 통한 최종 키워드 추출, 두 가지 과정을 거치게 된다.

첫 번째 단계는 후보 키워드를 선택하는 과정으로, 기존 연구에서 가장 많이 사용되는 방법에는 n-gram으로 선택된 모든 구문을 후보 키워드로 활용하는 방법이 있다 [1]. 이 방법은 일반적으로 문장부호와 불용어를 포함하는 후보 키워드를 제거한다. Hulth [2]는 대다수의 키워드가 명사구로 이루어져 있다는 점에 착안하여, n-gram 기법에서 선택된 후보 키워드 중에서 명사로 끝나지 않는 구문을 제거해 후보 키워드 목록을 구성했다. 또한 You [3]는 문서에서 가장 자주 등장하는 단어들을 핵심 단어로 지정하고, 주어진 문서에서 핵심 단어를 포함하는 명사구들을 후보 키워드로 사용했다.

하지만 이와 같은 기법들은 문서에서 등장하는 구문만을 후보 키워드로 선택하므로 문서에서 나타나지 않는 구문이 최종적인 잠재 키워드에서 배제되는 문제가 있다. 따라서 잠재 키워드를 추출하기 위해서는 문서에서 나타나지 않는 양질의 구문을 후보 키워드로 선택하는 방법이 필요하다.

두 번째 단계는 후보 키워드를 평가하여 최종 키워드를 선별하는 단계로, 대부분의 기존 연구들은 모델을 만든다. 모델을 구축하는 방법에는 크게 지도 학습과 비지도 학습이 있다. 지도 학습은 후보 키워드를 속성 형태로 가공한 후, 나이브 베이즈 (Naive bayes) [1-2], 지지 벡터 기계 (Support vector machine) [4]와 조건부 임의 필드 (Conditional random field) [5]와 같은 기계 학습 (Machine learning) 기법을 이용하여 후보 키워드가 키워드인지 아닌지 이진 분류하는 방법이다. 지도 학습에서 일반적으로 많이 사용되는 후보 키워드의 속성은 TF-IDF [6]와 문서에서 처음으로 등장한 상대적인 위치 [1]와 문서 제목에서의 등장여부 [4]와 같은 속성 등이 있다. 또한 Haddoud [7]는 어떤 구문이 다른 구문에 반복적으로 포함되어 나타나면 중요하지 않다는 것에 착안하여 후보 키워드 간의 중첩성을 고려하는 DPM (Document phrase maximality) 속성을 고안했다. 비지도 학습은 정답 키워드를 학습하지 않고, 후보 키워드들이 갖는 관계를 바탕으로 후보 키워드의 중요도를 평가한 후 가장 중요한 키워드 c개를 추출하는 방법이다. Mihalcea [8]는 문서에서 등장하는 후보 키워드를 정점(vertex)으로 두고, 각 구문이 한 문장에서 동시에 발생하는 횟수를 간선(link)으로 나타낸 그래프 모델을 생성한 후, 정점간의 관계를 나타내는 간선으로 각 정점이 얼마나 중요한 구문인지 평가하여 키워드를 추출했다. 향후, 이 방법은 다양한 방법으로 확장되게 되었다 [9-10]. Matsuo [11]는 의미 없는 구문은 문서의 내용과 관계없이 일정하게 나타나는 것에 착안하여, 문서에서 자주 등장하는 단어가 나타나는 횟수의 확률 분포와 후보 키워드가 문서에서 자주 등장하는 단어와 함께 나타나는 횟수의 확률 분포를 비교하여 후보 키워드를 평가하여 키워드를 추출했다.

하지만 이와 같은 기법들은 후보 키워드가 주어진 문서에서 나타나는 것을 가정하여 후보 키워드를 평가하는 문제가 있다. 지도 학습은 후보 키워드를 속성 형태로 가공하는데 있어 후보 키워드의 등장 위치와 횟수를 주로 활용하며, 비지도 학습은 후보 키워드가 나타나는 것을 바탕으로 후보 키워드간의 관계를 추론하였다. 따라서 이 방법들은 문서에서 나타나지 않는 구문을 평가하기에는 적합하지 않다. 잠재 키워드를 추출하기 위해서는 문서에서 나타나지 않는 구문을 평가하는 새로운 기법이 필요하다.

3. 배경 지식

본 장에서는 후보 키워드 선택과 후보 키워드 평가에 사용되는 기계학습 기법과 문서를 벡터 형태로 표현하는 방법에 대해 기술한다. 또한 논문에서 정의하는 잠재 키워드에 대해 살펴본다.

3.1 Latent Dirichlet Allocation (LDA)

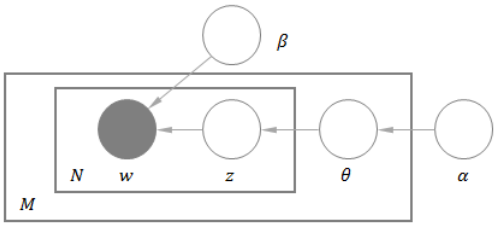

그림 1. 잠재 디리클레 할당 모델
Fig. 1. Latent dirichlet allocation model

LDA는 주어진 문서가 잠재적으로 갖는 주제들을 추론해내는 확률 모델로, generative topic model로도 불린다. Generative topic model이란 주어진 문서에 대한 주제들의 확률 분포θ와 각 주제에 대한 단어들의 확률 분포z가 주어졌을 때, 문서를 구성하는 주제를 확률적으로 선택하고, 선택된 주제에 존재하는 단어를 확률적으로 선택하는 과정을 반복함으로써 임의의 문서를 생성해낼 수 있는 모델을 뜻한다. LDA는 주어진 문서 셋에서 사전에 정의되는 α와 β 등의 파라미터 값들을 활용하여 θ와 z를 확률적으로 추론하는 모델이다.

3.2 Bag-of-words model

Bag-of-words model이란 문서를 단어 벡터 형태로 간략화시켜 표현해내는 방법으로, 텍스트 마이닝 분야에서 널리 사용된다.

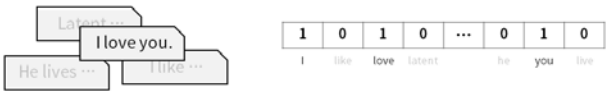

그림 2. 문서의 단어 벡터
Fig. 2. Word vector of document

예를 들어, 그림 2의 "I love you." 문서는 "I", "love"와 "you" 단어가 각각 1번씩 출현하였으므로, (1, 0, 1, 0, ···, 0, 1, 0)와 같은 벡터 형태로 나타내어진다.

3.3 잠재 키워드

본 논문에서 정의하는 잠재 키워드는 문서에서 나타나지 않는 키워드를 지칭한다. 기존 연구에서는 잠재 키워드를 추출하지 않았으므로, 이를 부르는 용어가 없었다. 그림 3은 통상적으로 정의되는 키워드와 잠재 키워드를 비교하는 그림을 나타낸 것이다.

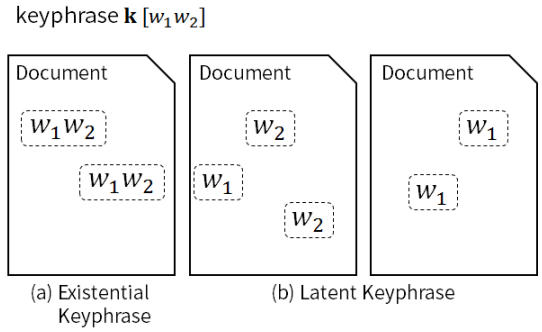

그림 3. 일반 키워드와 잠재 키워드의 비교
Fig. 3. Comparison between existential keyphrase and latent keyphrase

그림 3에서 키워드 k는 두 단어로 구성되어 있다. (a)에서 k는 w_1과 w_2가 연속해서 나타나고 있다. 이는 일반적으로 생각할 수 있는 키워드의 형태이다. 반면에, (b)에서 k는 w_1과 w_2가 서로 떨어져 있으며, 경우에 따라서는 w_2는 나타나지 않고 w_1만 나타나는 경우도 있다. 우리는 (b)와 같은 키워드를 잠재 키워드로 정의한다. 잠재 키워드는 실데이터[2, 17-18] 키워드의 1/4 이상을 차지하고 있으며, 문서에서 나타나지는 않지만 문서의 중요한 개념이나 내용을 함축하고 있어 문서 요약 및 정보 검색에 중요한 역할을 차지할 수 있다. 특히 SNS와 같이 내용이 적어 키워드가 명시적으로 나타나기 어려운 문서에서 유용하게 활용될 수 있다.따라서 본 논문에서는 이와 같은 잠재 키워드들을 추출하고자 한다.

4. 제안 방법

제안하는 기법은 문서 셋의 특징을 고려하여 의미 없는 공통어를 제거한다. 그리고 주어진 문서와 유사한 이웃 문서를 활용하여 후보 키워드를 선택하고, 후보 키워드를 구성하는 개별 단어들을 이용해 후보 키워드의 중요도를 평가하여 최종 키워드를 선별한다. 후보 키워드 선택 및 평가 과정에는 LDA를 기반으로 생성된 주제 모델이 활용된다. 일련의 과정에 대한 구조는 그림 4과 같다.

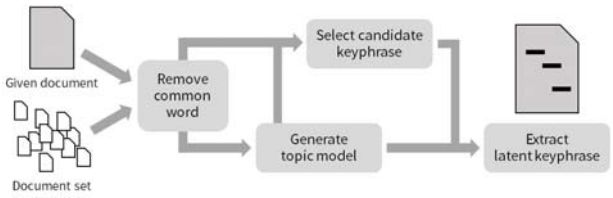

그림 4. 잠재 키워드 추출 과정
Fig. 4. Procedure of extracting latent keyphrase

4.1 공통어 제거

일반적으로 문서 셋은 같은 종류의 문서들로 구성되어 있으며, 같은 종류의 문서들은 불용어가 아님에도 자주 사용되는 공통어를 갖는다. 이는 문서를 분석하는 데 있어 불용어와 유사하게 노이즈로 작용하게 된다. 따라서 본 논문에서는 불용어 제거와 같은 전처리 과정과 별개로 문서 셋에서 자주 등장하는 r개의 공통어들을 제거한다.

4.2 주제 모델 생성

주제란 다수의 문서에서 공통적으로 기술하고 있는 추상적인 내용을 뜻한다. 이러한 관점에서 문서는 한 가지의 주제에 대해서만 기술하고 있다기보다는 여러 가지의 주제를 부분적으로 기술하고 있다고 볼 수 있다. LDA는 문서 셋에서 공통적으로 기술하는 주제들을 찾고, 각 문서들이 갖는 주제θ를 확률적으로 나타낸다. 또한 각 주제에 속하는 단어들의 확률 분포z도 나타낼 수 있다. θ는 4.3절에서 이웃 문서의 구성에 활용되며, z는 4.4절에서 후보 키워드를 구성하는 개별 단어의 등장 확률의 평가에 활용된다. 따라서 4.3절과 4.4절에 들어가기에 앞서 주어진 문서 셋에 대해 LDA를 기반으로 주제 모델을 생성한다.

4.3 후보 키워드 선택

후보 키워드란 최종적인 잠재 키워드로 선별될 가능성이 구문으로, 최종 키워드는 후보 키워드들 중에서 선별된다. 따라서 문서에서 나타나지 않는 잠재 키워드가 추출되기 위해서는, 문서에서 나타나지 않는 구문이 후보 키워드로 선택되어질 수 있어야 한다. 하지만 주어진 문서는 한정된 정보만을 제공하여, 문서에서 나타나지 않는 다양한 구문을 선택할 수 없다. 따라서 후보 키워드를 선택하는 데 있어 주어진 문서 이외의 다른 정보를 활용해야할 필요성이 있다. 본 논문에서는 신뢰성 있는 정보를 활용하기 위해 주어진 문서와 유사한 이웃 문서의 정보를 활용한다.

이웃 문서는 다양한 방법으로 정의될 수 있다. 첫 번째 방법은 주어진 문서와 내용이 유사한 문서를 찾는 것이다. 문서의 내용은 bag-of-words 모델에 의해 단어 벡터로 나타내어질 수 있으므로, 주어진 문서와 단어 벡터가 유사한 문서들을 내용이 서로 비슷한 문서로 구성할 수 있다. 두 번째 방법은 주어진 문서와 유사한 주제를 갖는 문서를 찾는 것이다. 우리는 주제 모델에서 추론된 각 문서의 주제 벡터를 이용하여, 주어진 문서와 주제 벡터가 유사한 문서들을 이웃 문서로 구성한다.

두 벡터 간의 유사도는 코사인 유사도(식 1)를 통해 산출한다. 주어진 문서d_g는 문서 셋에 포함되는 모든 문서D와 유사도가 계산되며, 이 중에서 유사도가 가장 높은 n개의 문서가 주어진 문서의 이웃 문서로 구성된다. 그리고 이웃 문서의 키워드는 주어진 문서의 후보 키워드로 선택된다.

$$sim(\overrightarrow{d_g}, \overrightarrow{d_n}) = \frac{\overrightarrow{d_g} \cdot \overrightarrow{d_n}}{\|\overrightarrow{d_g}\| \times \|\overrightarrow{d_n}\|},\ d_n \in D \quad (1)$$

4.4 잠재 키워드 추출

4.3절에서 선택된 후보 키워드는 주어진 문서에서 나타날 수도 있으며 나타나지 않을 수도 있으므로, 우리는 두 가지 유형의 후보 키워드를 모두 평가할 수 있어야 한다. 이를 위한 한 가지 방법은 후보 키워드를 구성하는 단어w_i 별로 중요도를 평가하여 이를 종합하는 것이다. 본 논문에서는 주제 모델에서 추론된 각 문서에 대한 주제들의 확률 분포와 각 주제에 대한 단어들의 확률 분포를 이용하여 w_i가 주어진 문서d_g에서 등장할 확률을 계산한다. w_i는 여러 가지 주제t에서 중복하여 등장할 수 있는데, 각 주제는 서로 독립이므로 식 2와 같이 각 주제별로 w_i가 생성될 확률을 더해주게 된다. 또한 LDA는 d_g와 w_i가 독립임을 가정하므로 식 2의 $p(w_i|t,d_g)$의 d_g는 생략될 수 있다.

$$p(w_i|d_g) = \sum_t p(w_i|t,d_g) \cdot p(t|d_g) = \sum_t p(w_i|t) \cdot p(t|d_g) \quad (2)$$

후보 키워드w의 중요도는 식 3과 같이 w를 구성하는 단어 w_i가 주어진 문서에서 등장할 확률들의 기하 평균으로 산출된다. 이는 후보 키워드w를 구성하는 단어 w_i가 주어진 문서에서 등장할 확률이 높으면 중요하다는 직관적인 생각에 근거한다.

$$p(w|d_g) = \sqrt[k]{\prod_i^k p(w_i|d_g)} \quad (3)$$

후보 키워드는 이웃 문서의 키워드로부터 선택되었으므로, 하나의 이웃 문서에 소속[1]된다. 그리고 유사도가 높은 이웃 문서에서 선택된 후보 키워드가 더 중요할 수 있으므로 주어진 문서와 후보 키워드가 소속된 문서의 유사도(식 1)를 식 3의 가중치로 활용하는 것이 가능하다. 식 1을 식 3의 가중치로 표현하면 식 4와 같이 나타내어진다.

$$p(w|d_g) = \sqrt[k]{\prod_i^k p(w_i|d_g)} \cdot sim(\overrightarrow{d_g}, \overrightarrow{d_n}) \quad (4)$$

최종적으로, 후보 키워드는 식 3과 식 4에 의해 평가되며, 가장 중요도가 높은 c개의 후보 키워드가 문서의 잠재 키워드로 추출된다.

1) 후보 키워드는 여러 문서에서 나타날 수 있으나, 여기에서는 가장 유사도가 높은 이웃 문서에 소속됨

5. 실 험

이 장에서는 제안한 잠재 키워드 추출 성능 평가의 실험 환경에 대해 알아보고, 그 결과를 분석한다. 실험은 후보 키워드 선택 방법을 비교하는 실험과 이웃 문서의 유사도 활용여부를 비교하는 실험, 두 종류로 진행했다.

5.1 실험 환경

실험에 사용된 데이터는 Hulth [4]의 데이터로, Inspec DB에서 수집된 자연과학 분야 논문의 제목, 초록과 키워드를 포함하는 영어 문서 2,000개로 구성되어 있다. 각 문서의 키워드는 미리 정의된 사전 안에서 부여된 키워드와 자유롭게 부여된 키워드 두 가지가 있으며, 본 논문에서는 전자의 키워드를 사용했다. 모든 문서는 아래와 같은 전처리 과정을 통해 거친 후, 실험에 사용되었다.

- 초록의 단어가 100개 미만인 문서는 충분한 정보를 제공하지 못하므로 제외되었다.
- 대소문자 정보는 활용되지 않으므로, 문서 내의 모든 문자는 소문자로 변경되었다.
- 의미가 없는 온점, 쉼표와 따옴표 등의 문장 부호는 삭제되었다.
- 전치사, 대명사와 관사 등 영어에서 일반적으로 통용되는 불용어는 제거되었다 [19].
- 마지막으로, 영어 전처리 과정의 핵심인 어간 추출 방법이 적용되었다 [20].

주제 모델의 생성에는 JGibbLDA [21]를 사용하였으며, 문서 셋에 나타나는 주제의 개수 k와 하나의 주제가 포함하는 단어의 개수 $twords$는 각각 실험을 통해 가장 좋은 값을 산출하는 16과 80으로 설정했다. 그리고 주어진 문서와 유사한 이웃 문서의 개수 n은 12로 설정되었다. LDA 모델은 확률 모델이기에 무작위성을 동반하므로, 모든 결과는 10번의 실험을 통해 계산된 평균값으로 채택했다.

5.2 실험 결과

본 실험에서 각 문서의 후보 키워드는 주제 모델을 이용해 평가되었으며, 이 중에서 가장 중요도가 높은 3개의 구문을 최종 잠재 키워드로 선별했다. 추출된 키워드는 각 문서의 정답 키워드와 비교해 정답유무를 확인하였으며, 실험은 두 종류로 진행했다.

첫 번째 실험에서는 word vector로 구성된 이웃 문서에서 선택된 후보 키워드와 topic vector로 구성된 이웃 문서에서 선택된 후보 키워드에서 추출되는 잠재 키워드 기법의 성능을 비교했다. 또한, 문서 셋에서 자주 등장하는 공통어 가 제거되는 개수 r이 변화함에 따라 보이는 잠재 키워드 추출 성능도 비교하였다. 그림 5는 그 결과를 보여준다.

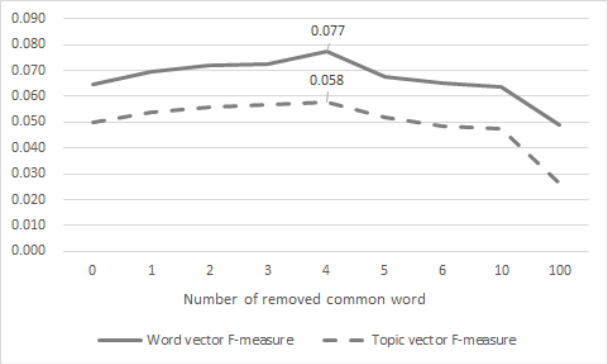

그림 5. 후보 키워드 선택 방법의 결과 비교
Fig. 5. Result comparison between methods of selecting candidate keyphrase

실험 결과, word vector를 이용하여 추출된 잠재 키워드의 정확률(F-measure)이 topic vector를 이용하여 추출된 잠재 키워드의 정확률보다 높음을 보였다. 이는 문서의 내용이 내포하는 주제를 추론하여 얻은 이웃 문서보다는 문서의 내용을 가공하지 않고 이웃 문서를 추론해내는 것이 이웃 문서를 통한 후보 키워드 선택에 더 좋음을 알 수 있다. 이 결과는 모든 경우에 통용되는 결과는 아니며, 잠재 키워드를 추출하는 경우에 한정된다. 또한, 그림 5는 문서 셋에서 자주 등장하는 공통어를 제외하는 경우에 잠재 키워드 추출 성능이 증가할 수 있음을 보였다. 공통어는 4개가 제거되었을 경우에 성능이 가장 좋았으며, 제거되는 공통어가 지나치게 많아지면 오히려 잠재 키워드 추출 성능이 하락함을 볼 수 있었다. 이는 문서에서 중요한 단어들이 제외되면 이웃 문서를 정확하게 선택하지 못하기 때문임을 알 수 있다. 제거되는 공통어는 "system", "model", "control"과 "method"이다. 이 단어들은 자연과학 분야 논문에서 자주 나타나기는 하지만 습관적으로 사용되는 단어로, 문서의 핵심적인 내용을 내포하고 있지 않는 단어임을 확인할 수 있었다.

두 번째 실험은 식 3과 식 4를 비교하는 실험으로, 두 식은 후보 키워드의 중요도를 평가하는데 있어 이웃 문서 유사도의 활용여부에 차이가 있다. 두 번째 실험에서는 첫 번째 실험 결과에 따라 공통어 4개를 제외한 상태로 진행하였으며, 결과는 표 1과 같다.

실험 결과, 후보 키워드의 중요도를 평가하는데 있어 이웃 문서의 유사도를 활용하는 것이 더 좋은 결과를 보였다. 이는 같은 이웃 문서일지라도 유사성이 높은 이웃 문서에서 선택된 후보 키워드가 더 중요할 가능성이 높다는 것을 시사한다. 그리고 이 경우 잠재 키워드 추출 결과는 정밀도(Precision) 0.097, 재현율(Recall) 0.075과 정확률(F-measure) 0.087이다.

표 1. 이웃 문서 유사도의 활용여부에 따른 결과 비교
Table 1. Result comparison between whether or not utilizing neighbor document similarity

	Unutilizing neighbor document similarity	Utilizing neighbor document similarity
Precision	0.088	0.097
Recall	0.069	0.075
F-measure	0.078	0.087

본 실험에서 추출하는 잠재 키워드는 관련 연구가 없어 베이스라인이 없다. 하지만 일반적인 환경에서 키워드 추출의 정확률이 0.2에서 0.4정도 나오는 것을 고려하면, 문서에서 나타나지 않는 잠재 키워드를 0.087의 정확률로 추출해내는 것은 낮지 않은 수치임을 알 수 있다.

6. 결 론

본 논문에서는 LDA 모델을 이용하여 문서에서 중요한 잠재 키워드를 추출하는 방법을 제안하였다. 이를 위한 핵심 아이디어는 주어진 문서와 유사한 문서의 키워드를 후보 키워드로 선택하고, 후보 키워드를 구성하는 개별 단어들의 등장 확률을 이용하여 후보 키워드의 중요도를 평가하는 것이다. 실험을 통해, 제안한 기법이 잠재 키워드를 합리적인 수준으로 추출하는 것을 볼 수 있었다. 향후에 잠재 키워드를 추출하기 위한 맞춤형 LDA 모델을 개발한다면 더 좋은 결과를 기대할 수 있을 것이다.

References

[1] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-specific keyphrase extraction," *Proceedings of the 16th international joint conference on artificial intelligence*, pp 668-673, 1999.

[2] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," *Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics*, 2003.

[3] W. You, D. Fontaine, and J. P. Barthès, "An automatic keyphrase extraction system for scientific documents," *Knowledge and information systems*, vol. 34, no .3, pp. 691-724, 2013.

[4] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword exraction using support vector machine," *Proceedings of the 7th international conference on web-age information management*, pp 86-96, 2006.

[5] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information System*, vol. 4, no. 3, pp. 1169-1180, 2008.

[6] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11-21, 1972.

[7] M. Haddoud, and S. Abdeddaïm, "Accurate keyphrase extraction by discriminating overlapping phrases," *Journal of Information Science*, 2014.

[8] R. Mihalcea, and P. Tarau, "Textrank: bringing order into texts," *Association for Computational Linguistics*, 2004.

[9] X. Wan, and J. Xiao, "Single Document Keyphrase Extraction Using Neighborhood Knowledge," *Association for the Advancement of Artificial Intelligence*, vol. 8, 2008.

[10] Z. Liu, W. Huang, Y. Zheng, and M. Sun, "Automatic

keyphrase extraction via topic decomposition," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2010.

[11] Y. Matsuo, and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 1, pp. 157-169, 2004.

[12] J. Park, J. Kim, J. Lee, and J. H. Lee, "Keyword extraction for blogs based on content richness," *Journal of Information Science*, vol. 40, no.1, pp. 38-49.

[13] T. Cho, H. Cho, and H. J. Lee, "Latent Keyphrase Generation by Combining Contextually Similar Primitive Words," *Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems*, pp. 600-604, 2014.

[14] M. G. Kim, N. G. Kim, and I. H. Jung, "A Methodology for Extracting Shopping-Related Keywords by Analyzing Internet Navigation Patterns," *Journal of Intelligence and Information Systems*, vol. 20, no. 2, pp. 123-136, 2014.

[15] J. Go, J. W. Son, H. J. Song, and S. Y. Park, "Personalized Keyword Extraction using Dialogue History," *Journal of the Korean Institute of Information Scientists and Engineers: Computing Practices and Letters*, vol. 18, no. 12, pp. 896-900, 2012.

[16] D. J. Choi, S. W. Lee, J. K. Kim, and J. H. Lee, "A Study on Graph-based Topic Extraction from Microblogs," *Journal of The Korean Institute of Intelligent System*, vol. 21, no. 5, pp. 564-568, 2011.

[17] M. Krapivin, A. Autaeu, and M. Marchese, "Large dataset for keyphrases extraction," *Technical Report DISI-09-055*, 2009.

[18] S. N. Kim, O. Medelyan, M. K. Kan, and T. Baldwin, "Semeval-2010 task 5: automatic keyphrase extraction from scientific articles," *Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics*, 2010.

[19] lextek, "Stop Word List 1," Available: http://www.lextek.com/manuals/onix/stopwords1.html, [Accessed: March 10, 2015].

[20] M. F. Porter, "An algorithm for suffix stripping," *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130-137, 1980.

[21] X. H. Phan and C. T. Nguyen, "Jgibblda," Available: http://jgibblda.sourceforge.net, [Accessed: January 16, 2015].

저 자 소 개

조태민 (Taemin Cho)
2014년: 성균관대학교 전자전기컴퓨터공학과 학사
2014년~현재: 성균관대학교 대학원 전자전기컴퓨터공학과 석사과정

관심분야 : Text Mining, Machine Learning
Phone : +82-31-290-7987
E-mail : tmchojo@skku.edu

이지형 (Jee-Hyong Lee)
1993년: 한국과학기술원 전산학과 학사
1995년: 한국과학기술원 전산학과 석사
1999년: 한국과학기술원 전산학과 박사
2002년~현재 : 성균관대학교 전자전기컴퓨터공학과 교수

관심분야 : Fuzzy Theory and Application, Intelligent System, Machine Learning
Phone : +82-31-290-7154
E-mail : john@skku.edu