JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 8, pp. 47-54, August 2021
https://doi.org/10.9708/jksci.2021.26.08.047

Adjusting Weights of Single-word and Multi-word Terms for Keyphrase Extraction from Article Text

In-Su Kang*

*Associate Professor, Dept. of Computer Science, Kyungsung University, Busan, Korea

[Abstract]

Given a document, keyphrase extraction is to automatically extract words or phrases which topically represent the content of the document. In unsupervised keyphrase extraction approaches, candidate words or phrases are first extracted from the input document, and scores are calculated for keyphrase candidates, and final keyphrases are selected based on the scores. Regarding the computation of the scores of candidates in unsupervised keyphrase extraction, this study proposes a method of adjusting the scores of keyphrase candidates according to the types of keyphrase candidates: word-type or phrase-type. For this, type-token ratios of word-type and phrase-type candidates as well as information content of high-frequency word-type and phrase-type candidates are collected from the input document, and those values are employed in adjusting the scores of keyphrase candidates. In experiments using four keyphrase extraction evaluation datasets which were constructed for full-text articles in English, the proposed method performed better than a baseline method and comparison methods in three datasets.

▸ **Key words**: Keyphrase, Keyphrase extraction, Score adjustment, Type-token ratio, Unsupervised keyphrase extraction

[요 약]

핵심구 추출은 문서의 내용을 대표하는 주제 용어를 자동 추출하는 작업이다. 비지도 방식 핵심구 추출에서는 문서 텍스트로부터 핵심구 후보 용어가 되는 단어나 구를 추출하고 후보 용어에 부여된 중요도에 기반하여 최종 핵심구들이 선택된다. 본 논문에서는 비지도 방식 핵심구 후보 용어 중요도 계산에서 단어 유형 후보 용어와 구 유형 후보 용어의 중요도를 조정하는 방법을 제안한다. 이를 위해 핵심구 추출 대상 문서 텍스트로부터 후보 용어 집합의 타입-토큰 비율과 고빈도 대표 용어의 정보량을 단어 유형과 구 유형으로 구분하여 수집한 후 중요도 조정에 활용한다. 실험에서는 영어로 작성된 full-text 논문을 대상으로 구축된 4개 서로 다른 핵심구 추출 평가집합들을 사용하여 성능 평가를 수행하였고, 제안된 중요도 조정 방법은 3개 평가집합들에서 베이스라인 및 비교 방법들보다 높은 성능을 보였다.

▸ **주제어**: 핵심구, 핵심구 추출, 중요도 조정, 타입-토큰 비율, 비지도 방식 핵심구 추출

• First Author: In-Su Kang, Corresponding Author: In-Su Kang
*In-Su Kang (dbaisk@ks.ac.kr), Dept. of Computer Science, Kyungsung University
• Received: 2021. 06. 10, Revised: 2021. 07. 29, Accepted: 2021. 08. 04.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

핵심구 추출(keyphrase extraction)은 문서로부터 문서에서 다루어지는 중요한 주제 용어들을 추출하는 작업이다[1,2]. 핵심구 추출을 위해 지도(supervised) 및 비지도(unsupervised) 방식의 접근법들이 기존 연구에서 시도되었으며, 본 논문에서는 비지도 방식의 핵심구 추출 방법을 다룬다.

비지도 방식 핵심구 추출은 문서로부터 추출된 핵심구 후보 용어들에 핵심구 중요도를 부여한 후 중요도가 높은 상위 핵심구들을 추출하는 과정으로 진행된다. 핵심구 후보 용어 추출에서는, 문서 텍스트를 품사 태깅한 후 제한된 명사구의 품사 패턴을 갖는 단어 나열을 후보 용어로 추출하거나, 문서 텍스트로부터 불용어 및 문장부호를 포함하지 않는 단어 n-gram들을 후보 용어로 추출하는 방법이 많이 사용된다. 핵심구 중요도 계산의 경우 많은 기존 연구들에서 통계적 접근법과 그래프 기반 접근법이 시도되었다. 통계 기반 접근법[3,4]에서는 후보 용어와 관련하여 문서 내외부에서 수집된 통계치들을 활용하여 핵심구 후보 용어에 중요도를 부여한다. 그래프 기반 접근법[5,6,7]에서는 문서 내 출현 용어를 노드로 갖는 그래프에 대해 노드 랭킹 알고리즘을 적용한 후 얻어지는 노드 랭크를 활용하여 핵심구 후보 용어에 중요도를 부여한다.

일반적으로 핵심구 후보 용어 집합에는 1개 단어로 이루어진 단어 유형의 용어들과 2개 이상 단어로 이루어진 구 유형의 용어들이 포함될 수 있다. 핵심구 후보 용어의 중요도 계산에서 후보 용어의 문서 내 출현 빈도수가 중요한 요소이며, 대부분의 경우 단어가 구보다 문서 내 출현 빈도가 높다는 점[3]을 고려할 때, 핵심구 후보 용어들의 중요도 계산에서 단어와 구 중요도 간 적절한 조정이 없는 경우 고빈도 단어들로의 핵심구 선택 편향 문제가 발생할 수 있다. 이와 관련하여 KP-Miner에서는 단어 유형 용어 중요도와의 균형을 위해 구 유형 용어의 중요도를 높이는 문서 의존적 boosting factor를 도입하였다[3].

본 연구에서는 고빈도 단어 유형 후보 용어들로의 핵심구 편향 문제를 다루기 위해 문서 내 단어 및 구 유형 용어들의 출현 정보에 기반하여 핵심구 후보 용어의 중요도를 조정하는 방법을 제안한다. 이를 위해 개별 문서 단위로 단어 및 구 유형 용어 집합들의 타입-토큰 비율과 고빈도 대표 단어 및 대표 구의 출현 확률에 기반한 정보량을 활용한다. 특히 본 연구에서는 역문헌빈도 등 문서 외부 정보의 사용으로 인한 성능 변화 효과를 배제하기 위해, 핵심구 추출 대상 문서 내부의 정보만을 사용하여 핵심구 추출을 수행한다. 실험에서는 기존 핵심구 추출 연구들에서 사용된 4개 서로 다른 데이터셋들에 대해, 제안된 방법의 성능 평가 결과를 제시한다.

논문의 구성은 다음과 같다. 2장에서는 핵심구 추출의 기존 연구들을 기술하고, 3장에서는 본 논문에서 제안하는 핵심구 후보 용어 중요도 조정 방법에 대해 기술한다. 4장에서는 실험 방법 및 제안된 방법의 성능 평가 결과를 기술하고, 5장에서 결론을 맺는다.

II. Related Works

비지도 방식 핵심구 추출의 많은 연구들은 통계 기반 방법과 그래프 기반 방법으로 진행되었다. 통계 기반 방법에서는 용어의 문서 내 출현 횟수, 출현 위치, 출현 문맥 혹은 외부 문서 집합에서의 출현 정보 등을 사용하여 핵심구 후보 용어의 중요도를 계산한다. 통계 기반 방법의 베이스라인으로 고려되는 TF-IDF(term frequency-inverse document frequency) 방법에서는 후보 용어의 문서 내 출현 빈도수와 외부 문서 집합에서의 역문헌빈도수의 곱으로 중요도를 계산한다[2]. 대표적 통계 기반 방법인 KP-Miner[3]에서는 문서 텍스트로부터 불용어와 문장 부호로 분리되지 않는 단어 나열들을 핵심구 후보 용어로 추출한 후, 최소 출현 빈도수 제약(예: 문서 내 최소 3회 이상 출현할 것)과 최초 출현 위치 제약(예: 문서의 첫 400 단어 이내 출현할 것)을 만족하지 못하는 용어들은 후보 용어에서 배제하였다. 이후 TF, IDF, boosting factor 및 위치 가중치를 고려하여 용어 중요도를 계산하였다. Boosting factor는 문서 내 전체 후보 용어들의 개수와 구 유형 후보 용어들의 개수의 비에 비례하는 값으로 계산된다. KP-Miner는 SemEval-2010 핵심구 평가 태스크에서 비지도 방법 중 가장 높은 성능을 보였으며[8], 최근 핵심구 추출 방법들의 비교 평가 실험에서도 우수한 성능을 보이고 있다[9]. 최근 시도된 통계 기반 방법 중 하나인 Yake에서는 후보 용어의 대소문자 표기 정보, 빈도수, 출현 위치, 출현 문맥 정보를 고려하여 핵심구 중요도를 계산하였다[4].

그래프 기반 방법에서는 문서 텍스트로부터 추출된 용어 집합과 용어 간 공기 정보 등을 바탕으로 문서 텍스트에 대한 그래프 표현을 생성한 후 노드 랭킹 알고리즘을 적용하여 용어 순위화를 수행한다. 그래프 기반 방법은 Mihalcea와 Tarau에 의해 시도된 TextRank 방법[5] 이후로 최근까지 다양한 방법들이 제안되고 있다. PositionRank[6]에서는 단어를 노드로 갖는 그래프에 대

해 문서 내에서 보다 앞쪽에 출현한 단어들이 선호되도록 random walk을 수행하였다. 이를 위해 각 단어의 문서 내 모든 출현 위치의 역수들의 합을 정규화하여 preference vector로 구성하는 방식을 사용함으로써, 핵심구 추출을 위한 그래프 기반 방법에서 위치 정보를 효과적으로 결합하였다. Multipartite 그래프 기반 방법[7]에서는 문서에서 추출된 핵심구 후보 용어들을 k개 군집으로 클러스터링한 후, 핵심구 후보 용어를 노드로 고려하여 구성된 k-partite 그래프에서 위치 정보가 결합된 노드 간 가중치를 사용하여 노드 랭킹을 수행하였다.

많은 기존 연구들은, 통계 기반 및 그래프 기반 방법을 통해 문서에 출현한 단어 단위로 중요도를 부여한 후, 구 후보 용어의 중요도 계산을 위해, 구 용어를 구성하는 내부 단어의 중요도를 합산하는 방식을 사용하였다[2,6,10]. Florescu와 Caragea는 이러한 합산 방식의 경우 핵심구 후보 용어의 길이가 최종 중요도에 큰 영향을 미칠 수 있다는 점을 지적하면서 핵심구 용어를 구성하는 내부 단어 중요도들의 조화평균과 핵심구 용어의 빈도수를 곱한 값을 최종 핵심구 중요도로 사용하는 방식을 제안하였다[11]. 이 방법은 El-Beltagy와 Rafea의 연구[3]나 본 논문의 방법과 같은 단어-구 유형 간 직접적 중요도 조정 방법은 아니나 구를 구성하는 구 유형 용어의 중요도를 구의 길이에 과의존되지 않게 계산하는 방법이라는 관점에서 단어 및 서로 다른 길이의 구 용어 간 중요도 조정 방법으로 볼 수 있다. 본 논문에서는 전술한 합산 및 평균을 이용한 중요도 계산 방식을 실험에서 비교 방법들로 사용한다.

III. The Proposed Method

이 장에서는 비지도 방식 핵심구 추출에서 고빈도 단어로의 핵심구 편향 문제를 다루기 위해 후보 용어의 중요도를 조정하는 방법을 제안한다. 그림 1의 수식들은 핵심구 중요도 조정에 활용되는 통계치들을 제시한 것으로 이 통계치들은 핵심구 추출 대상 문서로부터 수집된다. 그림 1의 수식들에서 tf(t)는 용어 t의 문서 내 빈도수이다. T_s는 문서 내 단어 유형 후보 용어들의 집합이며, T_m은 문서 내 구 유형 후보 용어들의 집합이다. 또한 $top_s(k)$는 문서 내 상위 k개 고빈도 단어 유형 후보 용어들의 집합이며, $top_m(k)$는 문서 내 상위 k개 고빈도 구 유형 후보 용어들의 집합이다.

$$TTR_s = \frac{|T_s|}{\sum_{t \in T_s} tf(t)}$$

$$TTR_m = \frac{|T_m|}{\sum_{t \in T_m} tf(t)}$$

$$P(top_s(k)) = \frac{\frac{\sum_{t \in top_s(k)} tf(t)}{k}}{\sum_{t \in T_s} tf(t)}$$

$$P(top_m(k)) = \frac{\frac{\sum_{t \in top_m(k)} tf(t)}{k}}{\sum_{t \in T_m} tf(t)}$$

$$I(top_s(k)) = \log \frac{1}{P(top_s(k))}$$

$$I(top_m(k)) = \log \frac{1}{P(top_m(k))}$$

Fig. 1. A list of equations which are used in the proposed method.

타입-토큰 비율(type-token ratio)은 일반적으로 텍스트에 출현한 단어 타입의 총 개수와 단어 토큰의 총 개수의 비율로 정의된다[12]. 이 정의를 따라 그림 1의 TTR_s는 문서 내 단어 유형 후보 용어들의 타입-토큰 비율로 정의되며, 이는 입력 문서에 출현한 모든 단어 유형 후보 용어들의 타입의 총 개수를 모든 단어 유형 후보 용어들의 토큰의 총 개수로 나눈 값으로 계산된다. 유사한 방식으로 TTR_m은 입력 문서 내에 출현한 모든 구 유형 후보 용어들의 타입의 총 개수를 모든 구 유형 후보 용어들의 토큰의 총 개수로 나눈 값으로 정의한다. 타입-토큰 비율을 어휘다양성 척도로 고려[13]할 때, TTR_s 및 TTR_m은 각각 핵심구 추출 대상 문서 내에서의 단어 유형 후보 용어의 어휘다양성 및 구 유형 후보 용어의 어휘다양성 정도를 계량화한 것에 해당한다.

$P(top_s(k))$는 문서 내 상위 k개 고빈도 단어 유형 후보 용어들의 평균 빈도수를 문서 내 모든 단어 유형 후보 용어 토큰들의 총 개수로 나눈 값으로 정의한다. 유사하게 $P(top_m(k))$는 문서 내 상위 k개 고빈도 구 유형 후보 용어들의 평균 빈도수를 문서 내 모든 구 유형 후보 용어 토큰들의 총 개수로 나눈 값으로 정의한다. $P(top_s(k))$는 평균적 고빈도 단어 후보 용어의 출현확률이며 $P(top_m(k))$는 평균적 고빈도 구 후보 용어의 출현확률에 해당한다. $I(top_s(k))$ 및 $I(top_m(k))$는 각각 평균적 고빈도 단어 후보 용어 및 구 후보 용어에 해당하는 정보량[14]을 계산하기 위해 정의되었다.

식 (1)은 본 논문에서 제안하는 핵심구 중요도 수식으로 s(t)는 후보 용어 t의 핵심구 중요도 값에 해당한다. 식 (1)

에서 s0(t)는 후보 용어 t에 대해 중요도 조정 이전에 부여된 중요도 값이며, |t|는 t를 구성하는 단어의 개수를 의미한다. 즉 |t|=1은 t가 단어 유형 후보 용어인 경우를 의미하며, |t|>1은 구 유형 후보 용어인 경우를 의미한다.

$$s(t) = \begin{cases} s0(t) \times TTR_s \times I(top_s(k)) & ,|t| = 1 \\ s0(t) \times TTR_m \times I(top_m(k)) & ,|t| > 1 \end{cases} \quad (1)$$

식 (1)에서는 핵심구 추출을 위한 후보 용어의 중요도 조정 관점에서 단어 및 구 유형 용어 집합들의 어휘다양성을 반영할 필요가 있다고 가정하고, 단어 및 구 유형 용어 중요도 수식에 단어 및 구 유형 용어 집합의 어휘다양성 정도를 표현한 값들인 TTR_s 및 TTR_m을 각각 적용하고 있다. 또한 식 (1)에서는 고빈도 단어 및 구 유형 대표 용어의 정보량에 비례하여 후보 용어의 중요도를 조정하기 위해, 단어 및 구 유형 용어 중요도에 단어 및 구 유형 대표 용어의 정보량을 각각 추가 적용하였다.

$$s0(t) = \frac{tf(t)}{FirstSentenceNo(t)} \quad (2)$$

본 연구에서는 식 (1)에서의 s0(t) 수식으로 식 (2)를 사용한다. 식 (2)에서 FirstSentenceNo(t)는 문서 내에서 용어 t가 최초 출현한 문장의 순서 번호이다. 핵심구 후보 용어 t에 대해, 식 (2)는 t의 문서 내 빈도수가 크고, t가 문서 내에서 첫 출현한 문장이 문서의 시작 위치에 가까울수록 큰 중요도 값이 부여되도록 한다. 식 (2)는 본 논문에서 새롭게 제안하는 방법이 아니며 핵심구 추출의 기존 많은 방법들에서 활용된 두 자질을 결합 사용한 것이다. 식 (2)에서는 용어의 위치 값으로 문서 내 단어 순번 대신 용어가 출현한 문장 순번을 사용하고 있는데 이는 기존 Yake 시스템[4]의 방법을 참조한 것이다.

다음은 4개 문장(s1, s2, s3, s4)으로 구성된 가상의 예제 문서의 후보 용어 목록에 대해 k=2로 가정하고 후보 용어 index의 핵심구 중요도 계산 과정을 보인 것이다.

s1: query, deep learning, document retrieval
s2: query, word embedding
s3: query, index, deep learning
s4: index, recall, word embedding
TTR_s=3/6
$top_s(2)$={query, index}
$P(top_s(2))$=((3+2)/2)/(3+2+1)
s0(index)=2/3
s(index)=(2/3)*(3/6)*log(6/2.5)

IV. Experiments

1. Experimental Setup

제안된 방법의 성능 평가를 위해 영어 full-text 논문 집합을 대상으로 구축된 Citeulike[15], NUS[16], SemEval[8], Krapivin[17]의 4개 핵심구 추출 평가집합들을 사용한다. 핵심구 추출 평가집합에는 문서 텍스트 및 문서에 대한 정답 핵심구들이 포함되어 있다. Citeulike 데이터셋 내 논문들은 대부분 생물정보학 분야에 속하며, NUS 데이터셋은 과학 분야 논문들로 구성되어 있다. SemEval 데이터셋 내 논문들은 분산시스템, 정보탐색/검색, 분산인공지능(다중에이전트시스템), 사회/행동과학(경제학) 분야에 속하며, Krapivin 데이터셋의 경우 computer science 분야 논문들로 이루어져 있다. 실험에 사용된 평가집합들의 정보는 표 1과 같으며 이 평가집합들은 [18]의 사이트로부터 다운로드되었다.

Table 1. Keyphrase evaluation datasets.

Dataset	Number of documents	Average number of keyphrases
Citeulike	183	17.4
NUS	209	12.0
SemEval	243	15.6
Krapivin	2304	5.3

핵심구 추출 성능에 대한 평가지표로는, 시스템이 출력하는 상위 순위 핵심구 후보 용어들에 대해 정답 핵심구들과의 비교를 통해 계산되는 F1 값을 사용한다. 핵심구 후보 용어와 정답 핵심구 용어 간 비교 시 포터스테머[19]를 용어에 적용하여 얻어지는 stemming 결과에 대해 일치 여부를 검사하였다. 본 실험에서는 핵심구 추출 방법들에 의해 출력되는 상위 5개, 10개, 15개, 20개 핵심구 후보 용어들에 대한 F1 값을 핵심구 추출 방법들의 성능 결과로 제시한다.

제안된 방법과의 성능 비교를 위해, 다음의 네 가지 방법들을 사용하였다.

(1) s0(t) 기반 방법
(2) Boosting factor 기반 중요도 조정 방법
(3) Sumup 방법
(4) HmeanTF 방법

s0(t) 기반 방법은 식 (2)의 s0(t)에 기반하여 핵심구 후보 용어를 추출하는 방법으로, 후보 용어의 단어 및 구 유형 여부에 따른 중요도 조정이 적용되지 않은 베이스라인 방법이다.

$$BF(t) = \begin{cases} s0(t) & ,|t| = 1 \\ s0(t) \times BoostingFactor & ,|t| > 1 \end{cases}$$
$$BoostingFactor = \min\left(\sigma, \frac{\sum_{t \in T} tf(t)}{\alpha \sum_{t \in T_m} tf(t)}\right) \quad (3)$$

Boosting factor 기반 방법은 식 (3)의 BF(t)에 기반하여 핵심구 후보를 추출하는 방법으로, KP-Miner 시스템에서 사용된 boosting factor 수식을 식 (2)와 결합하여 만든 중요도 조정 방법에 해당한다. 식 (3)에서 T는 문서에 출현한 전체 후보 용어의 집합이며, σ 및 α의 값은 El-Beltagy와 Rafea의 연구[3]를 따라 각각 3과 2.3을 사용하였다.

$$Sumup(t) = \sum_{w \in t} s0(w) \quad (4)$$

Sumup 방법[2,6,10]은 식 (4)의 Sumup(t)에 기반하여 핵심구 후보를 추출하는 방법으로, 문서에 출현한 후보 용어 t에 대해, t를 구성하는 단어들의 단어 중요도들의 총합을 t의 중요도로 부여한다. 식 (4)에서 w는 용어 t를 구성하는 개별 단어이며 단어 중요도 수식으로 식 (2)를 사용한 것이다.

$$HmeanTF(t) = tf(t) \times \frac{|t|}{\sum_{w \in t} \frac{1}{s0(w)}} \quad (5)$$

HmeanTF 방법[11]은 식 (5)의 HmeanTF(t)에 기반하여 핵심구 후보를 추출하는 방법으로, 문서에 출현한 후보 용어 t에 대해, t를 구성하는 단어들의 단어 중요도들의 조화평균과 t의 빈도수를 곱한 값을 t의 중요도로 부여한다. Sumup(t)에서처럼 식 (5)에서 w는 용어 t를 구성하는 개별 단어를 의미하며 단어 중요도 s0(w)에 대해 식 (2)를 사용하였다.

실험에 사용된 모든 방법들에 공통적으로, 문서 텍스트에 대해 spaCy 패키지[20]를 이용하여 품사태깅을 수행하고 제한된 명사구 품사패턴에 해당하면서 불용어를 포함하지 않는 단어나열을 추출한 후 그 내부 단어를 스테밍하여 얻어지는 stemmed 단어나열을 후보 용어 표현으로 사용하였다. 또한 문서 내 출현 빈도수 3회 이상의 후보 용어들에 대해서만 중요도를 계산하였고, 시스템이 출력하는 최종 핵심구 표현으로는 중요도 기준으로 상위 순위의 stemmed 용어에 해당하는 최초 단어 나열(들) 중에서 빈도수가 가장 높은 단어나열을 사용하였다.

제안된 방법의 경우, $I(top_s(k))$ 및 $I(top_m(k))$ 계산을 위한 $top_s(k)$ 및 $top_m(k)$ 집합은 k=5로 설정하여 각각 문서 내 상위 5개 고빈도 단어 용어들과 상위 5개 고빈도 구 용어들로 구성하였다.

2. Experimental Results

제안된 방법의 중요도 조정 동작을 분석하기 위해 식 (1)을 핵심구 순위화 관점에서 동일한 식 (6)으로 변경한 후, 실험 문서집합들에서 식 (6)의 SingleTermWeight의 값들을 수집하였고, 그 결과를 box plot 형식으로 그림 2에 제시하였다.

$$s(t) = \begin{cases} s0(t) \times SingleTermWeight & ,|t| = 1 \\ s0(t) & ,|t| > 1 \end{cases}$$
$$SingleTermWeight = \frac{TTR_s}{TTR_m} \times \frac{I(top_s(k))}{I(top_m(k))} \quad (6)$$

모든 실험 문서들에서 SingleTermWeight 값은 0.6 미만이었으며, 가장 많은 문서로 이루어진 Krapivin 데이터셋의 SingleTermWeight 값들이 가장 넓은 분포를 보였다. SingleTermWeight의 평균 값의 경우, Citeulike가 가장 높았고 다음으로 NUS, SemEval, Krapivin 순이었다. 그림 2로부터 제안된 방법은 실험 문서 내 단어 유형 용어들의 중요도를 감소시키는 방식으로 동작한다는 것을 알 수 있다.

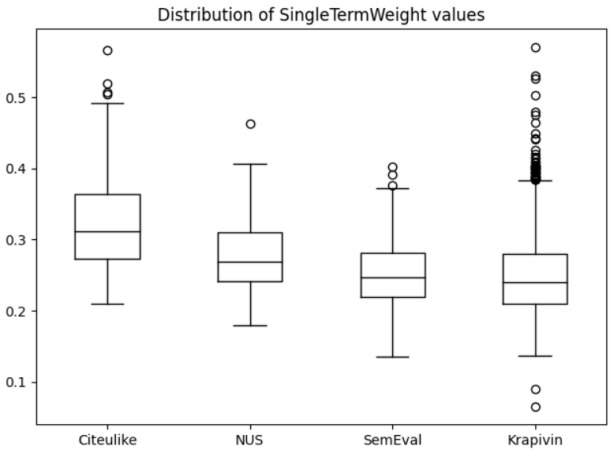

Fig. 2. Distribution of SingleTermWeight values over different datasets.

제안된 방법의 단어 유형 용어 중요도 감소에 따른 효과를 알아보기 위해 중요도 조정을 적용하기 전과 적용한 후의 상위 핵심구 목록 내 단어 유형 용어의 포함 비율 변화를 살펴보았고 그 결과를 그림 3에 제시하였다. 그림 3에서 가는 점선들은 s0(t)를 사용하여 추출된 상위 순위 핵심구 목록에서 단어 유형 용어들이 차지하는 비율에 해당하며, 굵은 점선들은 s(t)를 사용한 경우의 비율에 해당한다.

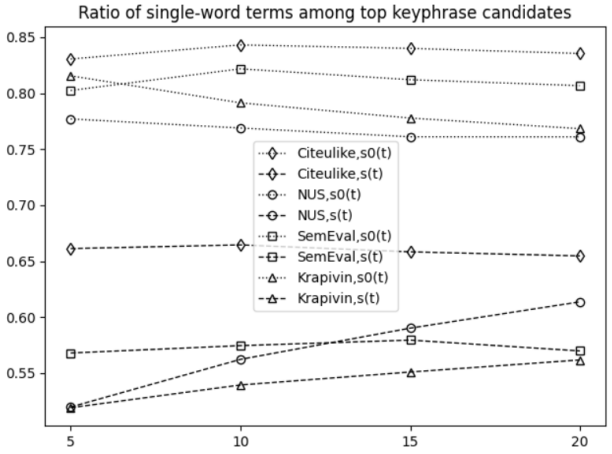

Fig. 3. Ratio of single-word terms among top keyphrase candidates.

그림을 통해 모든 실험 문서 집합들에서 제안된 중요도 조정 방식은 상위 핵심구 목록 내 단어 용어의 포함 비율을 감소시키고 있음을 알 수 있다. 다음은 SemEval 데이터셋 내 한 논문에 대해 베이스라인 방법과 제안된 방법의 상위 5개 핵심구 목록들을 비교 제시한 것으로, 제안된 방법의 경우 구 유형 용어 포함 비율이 베이스라인 방법에 비해 상대적으로 높음을 확인할 수 있다. 아래 목록에서 정답 핵심구들은 볼드체로 표시되었다.

- s0(t) 상위 핵심구: ['network', '**market barriers**', 'communication', 'agents', 'product']
- s(t) 상위 핵심구: ['**market barriers**', 'network', 'communication', '**marketing systems**', 'agents']

그림 4~7은 제안된 방법과 비교 방법들을 4개 각 핵심구 평가집합에 적용하여 얻은 상위 5, 10, 15, 20 순위의 핵심구 후보 용어 목록들에 대해 F1 성능을 제시한 것이다. 그림 4~7에 공통적으로 마커 +, ×, ○, △, □는 각각 제안된 방법, 베이스라인 방법, boosting factor 기반 방법, Sumup 방법, HmeanTF 방법에 해당한다.

중요도 조정이 적용되지 않은 베이스라인 방법과 비교할 때, 제안된 방법 및 boosting factor 기반 방법은 공통적으로 NUS, SemEval, Krapivin의 데이터셋들에서 베이스라인 방법보다 높은 성능을 보였다. 제안된 방법과 boosting factor 기반 방법을 비교한 경우, 제안된 방법은 NUS, SemEval, Krapivin 데이터셋들의 모든 상위 순위 용어 목록들에서 boosting factor 기반 방법보다 높은 성능을 보였다. Sumup 및 HmeanTF 방법과 비교한 경우에도 제안된 방법은 NUS, SemEval, Krapivin의 평가집합들에서 더 높은 성능을 보였다. 이러한 결과는 제안된 방법에서 시도된 단어-구 유형 용어 집합의 어휘다양성 및 고빈도 대표 용어의 정보량에 기반한 후보 용어의 중요도 조정 방법이 핵심구 추출 성능 향상에 긍정적 효과가 있음을 보여주는 결과이다.

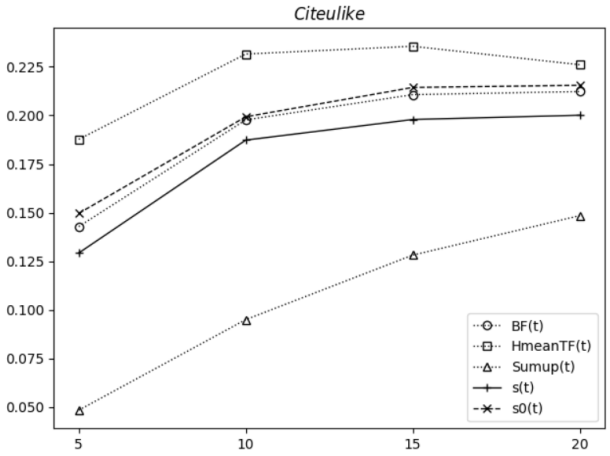

Fig. 4. F1 performance of top-k candidates on Citeulike dataset.

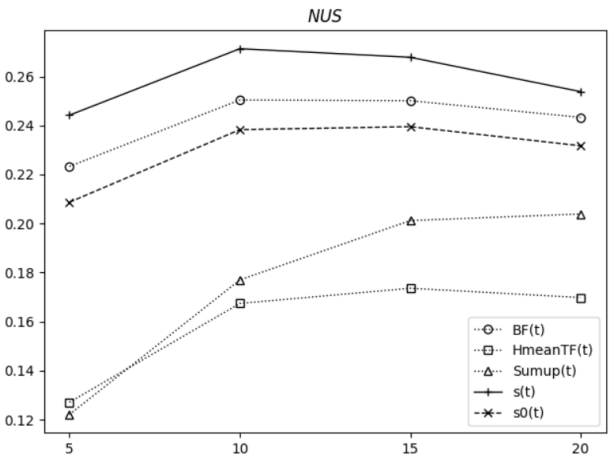

Fig. 5. F1 performance of top-k candidates on NUS dataset.

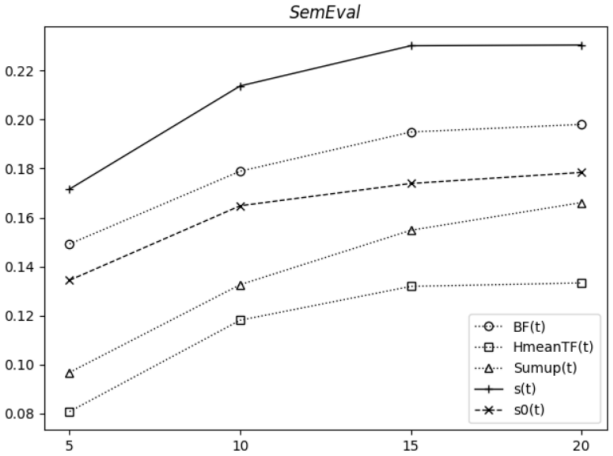

Fig. 6. F1 performance of top-k candidates on SemEval dataset.

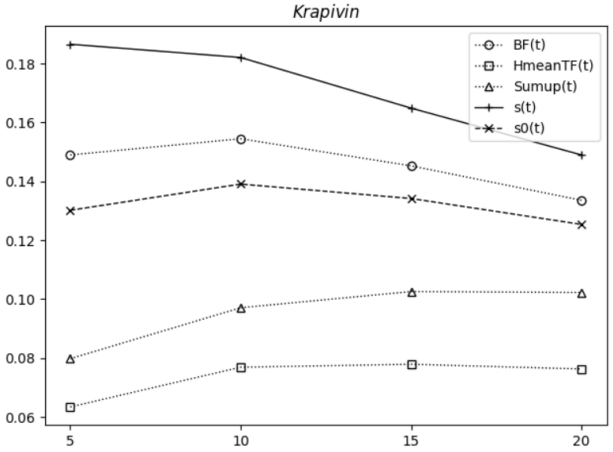

Fig. 7. F1 performance of top-k candidates on Krapivin dataset.

그러나, 제안된 방법은 Citeulike 데이터셋에서는 나머지 데이터셋들과 달리, 베이스라인, boosting factor 기반 방법 및 HmeanTF 방법보다 낮은 성능을 보였다. 그 이유 중 하나는 Citeulike 데이터셋 내 정답 핵심구 중 단어 유형 키워드의 높은 비율이 제안된 방법에서의 단어 유형 용어 중요도 감소 효과와 상충되기 때문인 것으로 보인다. 표 2에 보인 바와 같이 Citeulike 데이터셋의 경우 다른 데이터셋들과 비교하여 정답 핵심구 목록 내에 포함된 단어 유형 용어의 비율이 상대적으로 높다.

Table 2. Ratio of single-word and multi-word terms among the gold-standard keyphrases.

Dataset	Gold standard keyphrases	
	Ratio of single-word terms	Ratio of multi-word terms
Citeulike	0.77	0.23
NUS	0.28	0.72
SemEval	0.20	0.80
Krapivin	0.19	0.81

V. Conclusions

본 논문에서는 핵심구 추출을 위해 단어 및 구 유형 후보 용어의 중요도를 조정하는 방법을 제안하였다. 이를 위해 단어 및 구 유형 후보 용어 집합에 대한 어휘다양성과 단어 및 구 유형 고빈도 대표 용어의 출현확률에 기반한 정보량을 결합 사용하였다. 핵심구 추출 평가 집합들을 사용한 분석을 통해 제안된 방법은 단어 유형 후보 용어의 중요도를 감소시키는 효과가 있음을 보였다. 성능 평가를 통해 제안된 방법은 중요도를 조정하지 않은 베이스라인 방법 및 다른 비교 방법들보다 NUS, SemEval, Krapivin 데이터셋들에서 높은 성능을 보였다(그림 5,6,7 참조). 그러나 Citeulike 데이터셋의 경우 제안된 방법은 Sumup 방법을 제외한 나머지 방법들보다 낮은 성능을 보였는데(그림 4 참조), 이는 제안된 방법의 단어 유형 용어 중요도 감소 효과(그림 2,3 참조)가 단어 유형의 정답 핵심구 비율이 상대적으로 높은 Citeulike 데이터셋에서 부정적으로 동작한 이유에 기인한 것으로 분석되었다. 향후에는 새로운 중요도 조정 방법들을 고안하여 현재 방법을 보다 개선할 계획이다.

REFERENCES

[1] P. Turney, "Learning Algorithms for Keyphrase Extraction," Information Retrieval, Vol. 2, No. 4, pp. 303-336, 2000.

[2] K. Hasan, and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1262-1273, 2014.

[3] S. El-Beltagy, and A. Rafea, "KP-Miner: Participation in SemEval-2," Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 190-193, 2010.

[4] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," Information Sciences, Vol. 509, pp.

257-289, 2020.

[5] R. Mihalcea, and P. Tarau, "TextRank: Bringing Order into Texts," Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404-411, 2004.

[6] C. Florescu, and C. Caragea, "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1105-1115, 2017.

[7] F. Boudin, "Unsupervised Keyphrase Extraction with Multipartite Graphs," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 667-672, 2018.

[8] S. Kim, O. Medelyan, M. Kan, and T. Baldwin, "SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles," Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21-26, 2010.

[9] E. Papagiannopoulou, and G. Tsoumakas, "A review of keyphrase extraction," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 10, No. 2, 2020.

[10] X. Wan, and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge," Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, pp. 855–860, 2008.

[11] C. Florescu, and C. Caragea, "A New Scheme for Scoring Phrases in Unsupervised Keyphrase Extraction," Advances in Information Retrieval - 39th European Conference on IR Research, pp. 477-483, 2017.

[12] W. Johnson, "Studies in language behavior: A program of research," Psychological Monographs, Vol. 56, No. 2, pp. 1-15, 1944.

[13] B. Richards, "Type/Token Ratios: what do they really tell us?," Journal of Child Language, Vol. 14(2), pp. 201-209, 1987.

[14] C. Shannon, "A mathematical theory of communication," Bell System Technical Journal, Vol. 27, No. 3, pp. 379-423, 1948.

[15] O. Medelyan, E. Frank, and I. Witten, "Human-competitive tagging using automatic keyphrase extraction," Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1318-1327, 2009.

[16] T. Nguyen, and M. Kan, "Keyphrase Extraction in Scientific Publications," Proceedings of the 10th International Conference on Asian Digital Libraries, pp. 317-326, 2007.

[17] M. Krapivin, A. Autaeu, and M. Marchese, "*Large Dataset for Keyphrases Extraction*," University of Trento, Tech Report # DISI-09-055, 2009.

[18] Datasets of Automatic Keyphrase Extraction, *https://github.com /LIAAD/KeywordExtractor-Datasets*

[19] M. Porter, "An Algorithm for Suffix Stripping," Program, Vol. 14, No. 3, pp. 130-137, 1980.

[20] SpaCy. *https://spacy.io/*

Authors

In-Su Kang received his bachelor's degree from Kyungpook National University in 1995, and master's and doctoral degrees from POSTECH, in 1999, and 2006, respectively.

He is an associate professor at the Department of Computer Science, Kyungsung University. He is interested in natural language processing and information retrieval.