ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

문서 클러스터를 위한 워드넷기반의 대표 레이블 선정 방법☆

Representative Labels Selection Technique for Document Cluster using WordNet

김 태 훈[1] 손 미 애[1*]
Tae-Hoon Kim Mye Sohn

요 약

본 연구에서는 문서 클러스터링 결과 도출된 개별 클러스터가 함축하고 있는 의미를 파악하는 데 필요한 어휘들의 정보량을 활용한 문서 클러스터 레이블링(Documents Cluster Labeling) 방법을 제안하였다. 이를 위해, 클러스터에 포함된 어휘들이 해당 클러스터에서 얼마나 중요한 비중을 차지하고 있는지 파악하기 위하여 각 어휘의 출현 빈도와 정보량을 이용한 어휘의 가중치를 계산한 후, 워드넷을 이용하여 클러스터에 포함된 어휘들의 최근접 공통 상위어를 후보 레이블로 식별하였다. 이상의 과정을 거쳐 식별된 후보 레이블의 정보량과 클러스터내에서의 중요도 가중치를 활용해, 해당 클러스터의 의미와 특징을 포괄적으로 표현할 수 있는 대표 레이블을 결정하였다. 본 연구의 우수성을 입증하기 위해 다음과 같은 실험을 수행하였다. 실험은 본 연구에서 제안한 방법에 따라 선정된 레이블과 후보 레이블을 워드넷에 프로젝션한 후, 워드넷상에서 이들 레이블의 위치(깊이)를 확인하였다. 또한 선정된 후보 레이블을 상위어로 갖고 있는 클러스터 내 어휘의 수를 도출하여, 휴리스틱 방법에 따라 선정된 레이블을 전문가가 찾은 대표 레이블과의 비교를 수행하였다. 평가지표로 후보 레이블의 적합성($Suitability_{cl}$)과 대표 레이블의 적절성($Appropriacy_{rl}$)을 활용하였다. 실험 결과, 본 연구에서 제안한 방법을 적용해 문서 클러스터 레이블링을 수행할 경우, 후보 레이블의 적합성의 경우 기존의 방법보다 약간 감소하지만 계산량이 기존 방법의 약 20% 정도로 감소하였으며, 대표 레이블의 적절성의 경우 기존의 방법보다 우수한 결과를 도출하는 것을 확인하였다.

☞ 주제어 : 문서 클러스터 레이블링, 정보량, 워드넷, 유사도 계산

ABSTRACT

In this paper, we propose a Documents Cluster Labeling method using information content of words in clusters to understand what the clusters imply. To do so, we calculate the weight and frequency of the words. These two measures are used to determine the weight among the words in the cluster. As a nest step, we identify the candidate labels using the WordNet. At this time, the candidate labels are matched to least common hypernym of the words in the cluster. Finally, the representative labels are determined with respect to information content of the words and the weight of the words. To prove the superiority of our method, we perform the heuristic experiment using two kinds of measures, named the suitability of the candidate label ($Suitability_{cl}$) and the appropriacy of representative label ($Appropriacy_{rl}$). In applying the method proposed in this research, in case of suitability of the candidate label, it decreases slightly compared with existing methods, but the computational cost is about 20% of the conventional methods. And we confirmed that appropriacy of the representative label is better results than the existing methods. As a result, it is expected to help data analysts to interpret the document cluster easier.

☞ keyword : Documents Cluster Labeling, Information content, WordNet, Similarity Calculation

1. 서 론

문서 클러스터 레이블링(document cluster labeling)이란 클러스터링된 문서들을 대상으로, 문서 클러스터의 의미를 나타내 주는 대표 레이블을 선정하는 절차를 의미한다 [4, 8, 9]. 방대한 양의 텍스트 문서를 대상으로 수행한 클러스터링 결과로부터 생성된 클러스터의 의미를 해석하고 각각의 클러스터에 맞는 레이블을 선정하기 위해서는 개별 클러스터가 포함하고 있는 문서들의 의미 해석

[1] Department of Industrial Engineering, Sungkyunkwan University, Korea.
* Corresponding author (myesohn@skku.edu)
[Received 10 October 2016, Reviewed 28 October 2016, Accepted 28 February 2017]
☆ 본 논문은 성균관대학교 석사학위논문 '어휘 정보량을 이용한 문서 클러스터 레이블 선정 방법 연구 [25]'을 수정 및 보완한 논문임.
☆ 이 논문은 산업통상자원부의 국민안전증진기술개발 사업의 지원을 받아 수행하였습니다 (과제번호:1005_0810)

이 필요하며, 그로 인한 전문가들의 분석이 필수적으로 수반되어야 한다. 그러나 클러스터에 포함된 문서 수의 방대함으로 인해 전문가들이 모든 문서를 직접 확인한 후, 클러스터의 의미를 도출하는 것은 현실적으로 불가능하다. 이러한 어려움을 극복하기 위해, 문서 클러스터가 포함하고 있는 어휘들을 기반으로 개별 문서 클러스터의 의미를 도출할 수 있는 대표 레이블을 선정한 후, 이를 이용해 개별 문서 클러스터에 대한 이해를 제고시키는 문서 클러스터 레이블링 연구가 등장하였다[4, 8].

일반적인 문서 클러스터 레이블링 연구의 경우, 전체 문서들에 대하여 연관성이 떨어지는 수 많은 단어들을 모두 고려할 수 없기 때문에 문서와 연관성이 높은 단어들로 먼저 후보 레이블을 선정한 후 전체 클러스터를 대표할 수 있는 후보 레이블을 대표 레이블로 선정한다. 이때 후보 레이블이란 문서 클러스터와 관련된 어휘를 의미하며, 후보 레이블중 문서 클러스터와 연관성이 높은 어휘들을 선정 해 대표 레이블로 지정한다.

그 중 문서 클러스터 레이블링과 관련된 초기 연구에서는 클러스터의 내부 정보만을 활용해 클러스터에 대한 레이블링을 수행하였다[1-5]. 이들은 한 문장에 특정 어휘가 동시에 포함되면 두 어휘 사이에는 시맨틱 관계가 존재한다는 가정하에, 클러스터에 존재하는 어휘를 후보 레이블로 지정한 후, 후보 레이블과 클러스터에 포함된 어휘들의 동시 출현 빈도에 기반한 시맨틱 관계를 통해 대표 레이블을 결정한다. 그러나 어휘들의 동시 출현 빈도를 활용해 어휘들간의 시맨틱 관계를 식별하게 되면, 어휘들간 계층관계나 동의어 관계 등과 같은 언어학적 시맨틱 관계를 고려하지 못하는 한계가 노정된다[11].

이러한 문제를 해결하기 위해, 클러스터의 내부 정보뿐만 아니라 언어학적 시맨틱 관계를 모델링하고 있는 외부 정보를 추가적으로 활용하는 문서 클러스터 레이블링 연구가 활발히 수행되고 있다[6-10]. 이들은 문서 클러스터에 대한 레이블링에 필요한 대표 레이블을 결정하기 위해, 우선 외부 정보를 이용해 문서 클러스터가 포함하고 있는 어휘들을 포괄할 수 있는 어휘를 식별하고, 이후 선정된 후보 레이블과 클러스터에 포함된 어휘들간의 관계를 통해 대표 레이블을 결정한다. 외부 정보를 추가적으로 활용할 경우, 문서 클러스터의 내부 정보로만을 확인할 수 없는 어휘들간의 시맨틱 관계를 충분히 고려할 수 있으며, 외부 정보로부터 추가되는 어휘들을 이용해 문서 클러스터의 의미를 정확히 전달해줄 수 있는 어휘들을 후보 레이블로 선정할 수 있다는 장점이 있다. 결과적으로 내부 정보만을 이용하는 방법보다는 문서 클러스터 레이블링에 효과적이다. 그러나 외부 정보를 추가적으로 활용한 연구도 다음과 같은 한계가 존재한다. 첫째, 클러스터에 포함된 개별 어휘의 중요도에 대한 고려가 미흡하다. 예를 들어, 문서 클러스터에 포함된 어휘들은 특정 클러스터와의 높은 관련성으로 인해 출현 빈도가 높은 경우와 일반적인 어휘이기 때문에 출현 빈도가 높은 경우가 발생할 수 있음에도 불구하고, 어휘의 출현 빈도만 고려한다면 발생 빈도가 높은 이유를 정확히 파악할 수 없게 된다. 둘째, 어떤 어휘를 후보 레이블로 식별할 것인가에 대한 고려가 미흡하다. 예를 들어, 외부 정보를 이용해 식별한 모든 어휘를 후보 레이블로 추가하게 되면, 대표 레이블을 결정에 소요되는 계산량의 부담이 현저히 증가하게 된다. 이러한 부담을 완화하기 위해 클러스터에 존재하는 어휘들과 유사성이 높은 관련 어휘만 후보 레이블로 추가한다면, 후보 레이블의 부재로 인한 대 레이블 선정에 왜곡이 발생할 수 있다. 마지막으로, 대표 레이블 결정시, 기존 연구들은 후보 레이블이 클러스터의 전체 의미를 포괄하는 정도를 고려하지 않는다. 만약 일반적인 어휘가 대표 레이블에 선정된다며, 클러스터의 전체 의미를 포괄하는 정도는 높아질 수 있으나, 클러스터의 특징을 구체적으로 표현하는 능력은 낮아질 수 있다. 반대로, 구체적인 어휘를 대표 레이블로 선정한다면 클러스터에 포괄적인 의미에 대한 설명력은 낮아지는 반면 클러스터의 특징에 설명력은 높아지는 상쇄관계가 나타난다.

이에 본 연구에서는 클러스터의 전체적인 의미를 포괄하면서 클러스터의 특징을 대표할 수 있는 대표 레이블을 결정하기 위한 방법으로 어휘의 정보량을 이용한 문서 클러스터 레이블 선정 방법을 제안하고자 한다. 이때 어휘의 정보량은 한 어휘가 전달 가능한 정보의 양을 정량화한 값으로 정의하며, 어휘의 정보량이 낮을 경우 일반적인 어휘로, 어휘의 정보량이 높을 경우 구체적인 어휘로 간주할 수 있다. 이러한 정보량을 사용하기 위해서는 어휘 간 계층 관계의 정보가 필요하다. 이를 위해, 외부 정보로 워드넷을 사용하였다.

본 논문은 구성은 다음과 같다. 2장에서는 관련 연구를 요약하고, 3장에서는 본 논문에서 제시하는 정보량 기반의 문서 클러스터 레이블 선정 방법에 대한 전체 아키텍처 구조와 방법을 기술한다. 4장에서는 실험을 통해 본 논문에서 제안하는 레이블 선정 방법의 유효성에 대해 입증하였고, 마지막 5장에서는 본 논문의 결론과 향후 연구 방향에 대해 약술하였다.

2. 관련연구

2.1 어휘의 정보량과 워드넷

어휘의 정보량이란 한 어휘가 전달 가능한 정보의 양을 정량화한 값이다[11, 12] 이는 주어진 어휘 a의 집합 내에서 전체 어휘의 하위어 중 해당 어휘와 그 하위어의 비율 ($P(a)$)에 대한 음수 로그 값이며 계산식은 (1)과 같다.

$$IC_a = -\log P(a) \quad (1)$$

어휘의 정보량은 어휘들간의 시맨틱 유사성을 계산하기 위해 사용되며, 두 어휘 a와 b의 하위어 중 공통으로 출현한 하위어의 비율 $LCS(a,b)$를 통해 계산한다. 계산식은 식 (2)와 같다. 즉 두 어휘 a와 b의 모든 하위어 중에서 공통인 하위어의 비율을 도출하고, 전체 하위어 중 공통인 하위어의 비율을 이용해 두 어휘의 유사성을 계산한다.

$$Similarity(a,b) = IC(LCS(a,b)) \quad (2)$$

또한 어휘의 계층구조에서 해당 어휘의 하위어 수가 적을수록 구체적인 어휘 즉 정보량이 많은 어휘이며, 하위어의 수가 많을수록 일반적인 어휘 즉 정보량이 적은 어휘이다. 본 연구에서는 어휘의 정보량을 도출하기 위해 어휘들간의 계층관계를 정의하고 있는 워드넷을 외부 정보로 활용하였다. 워드넷은 영어의 의미 어휘 목록으로 약 15만 개의 어휘, 11만 5천 개의 동의어 집합, 20만 개의 어휘-의미 쌍을 제공한다[13-15]. 워드넷에는 어휘 간 상하위 관계로 표현되는 어휘 간 계층관계가 존재한다. 즉 상하위 관계를 통해 어휘 간의 포함관계 및 어휘의 일반적 구체적 정도를 파악하는데 활용 가능하다. 또한 워드넷에는 어휘의 정보량을 계산하기 위한 대부분의 어휘가 포함되어 있으며 어휘의 의미 및 어휘간 관계 계산에 효과적이다[12].

2.2 클러스터 레이블링

클러스터에 적합한 대표 레이블을 선정하기 위한 기존 연구는 크게 내부 정보를 활용한 연구[1-5]와 내부 정보와 외부 정보를 동시에 활용하는 연구[6-10]로 분류된다. 기존 연구에서 고려한 요소들에 대한 분류와 대표 레이블을 선정하기 위해 사용한 방법은 (Table 1)에 요약하였다.

전자는 두 어휘가 한 문장에서 동시에 출현하는 빈도가 많을 경우 서로 시맨틱 관계가 존재한다는 가정을 기반으로 연구를 수행하였다. 그러나 동시 출현 빈도를 통한 어휘들의 시맨틱 관계로는 실제 어휘의 의미와 관련된 언어학적 시맨틱 관계에 대한 고려가 불가능하다는 한계가 존재한다[11]. 이와 같은 한계를 극복하기 위해 내부 정보와 외부 정보를 동시에 활용하는 연구가 수행되었다. 이들은 어휘의 의미를 고려하기 위해 어휘 간 관계가 미리 정의되어 있는 정보(위키피디아, 워드넷 등)을

(표 1) 문서 클러스터 레이블링 관련 연구
(Table 1) Related research for document cluster labeling

	Semantic of words	Semantic relation between words	Candidate label	Method of scoring the representative label	Ref.
Internal Information-based	-	Co-occurrence	Words in cluster	Conditional probability	[1]
	-	Co-occurrence	Words in cluster	PageRank	[2], [3]
	-	Co-occurrence	Words in cluster	Centroid of graph	[4]
	-	Co-occurrence	Phrases in cluster	Cosine similarity	[5]
External Information-based	Wikipedia	Wikipedia	Word in cluster & Word in Wikipedia	Centroid of graph	[6]
	Wikipedia	Wikipedia	Word in cluster & Word in Wikipedia	Conditional probability	[7]
	WordNet	Co-occurrence	Phrases in cluster using lexical chain	Conditional probability	[8]
	WordNet	WordNet	Word in cluster & Hypernym in WordNet	Heuristic	[9], [10]

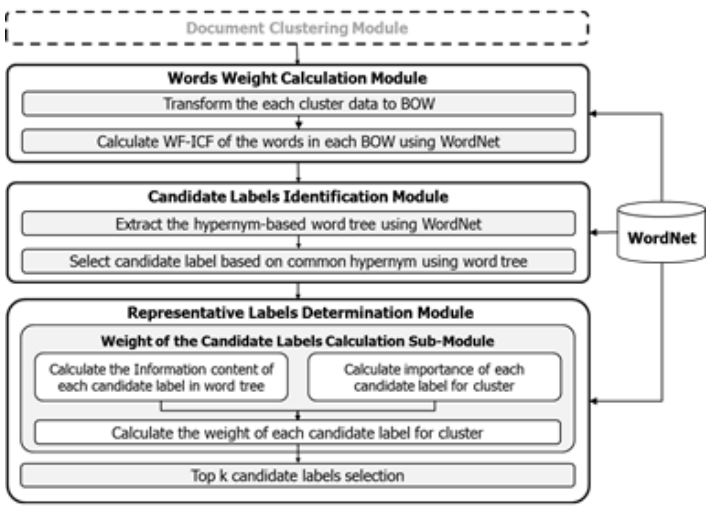
활용한다. [6,7]은 클러스터에 포함된 어휘들과 관련이 있으며 동시에 위키피디아에 존재하는 어휘를 활용하여 후보 레이블을 선정한 후, 위키피디아에 포함된 어휘들의 관계를 표현한 그래프를 활용하여 클러스터에 포함된 어휘와의 관계가 가장 많은 후보 레이블을 대표 레이블로 선정하였다.

그러나 위키피디아의 경우 어휘들 간의 포함관계 및 동의어 관계 등과 같은 어휘들 사이의 시맨틱을 정확히 파악할 수 없다는 한계가 있으며, 이를 극복하기 위해 워드넷을 활용한 연구가 수행되었다. [9,10]은 클러스터에 존재하는 어휘들의 상위어를 후보 레이블로 선정하기 위해 워드넷을 활용한 대표적인 연구이다. 선정된 후보 레이블들은 워드넷 상에서의 어휘의 깊이와 클러스터에 존재하는 어휘들 중에서 후보 레이블의 하위어로 존재하는 개수를 활용한 휴리스틱 방법을 통해 대표 레이블로 결정하였다. 이 연구들에서는 워드넷을 이용해 클러스터에 포함된 어휘의 의미를 포함한 상위어를 후보 레이블로 확장하였으며, 후보 레이블 중에서 클러스터의 의미를 정확히 표현할 수 있는 구체적인 어휘를 대표 레이블로 선정하기 위해 워드넷에서의 어휘의 깊이를 고려하였다. 그러나 워드넷에서 같은 크기의 깊이일지라도 어휘가 포함하고 있는 구체적 정도는 다를 수 있다. 예를 들면, 과일이란 뜻의 'fruit'의 경우 워드넷에서의 깊이가 9이지만 호박 파이라는 뜻의 'pumpkin pie'도 워드넷에서의 깊이가 9이다. 즉 워드넷에서의 깊이를 가지고 어휘의 일반적 구체적 정도를 고려할 경우 일반적 구체적 정도의 구분이 불가능한 경우가 발생할 수 있다.

본 연구에서는 이와 같은 한계를 극복하고자 어휘의 정보량을 통한 어휘의 일반적 구체적 정도를 활용하였다. 어휘의 정보량을 활용할 경우 어휘 그 자체가 포함하고 있는 하위어를 통해 어휘의 일반적 구체적 정도를 활용하기 때문에 워드넷에서의 깊이보다 어휘의 일반적 구체적 정도를 파악하는데 효과적이라고 할 수 있다[12].

3. 전체 아키텍처 및 세부 모듈

어휘 정보량을 이용한 문서 클러스터 레이블링 시스템의 아키텍처를 (Figure 1)에 도식화하였다. 본 시스템은 문서 클러스터링 모듈(Document Clustering Module, DCM), 어휘 가중치 계산 모듈(Words Weight Calculation Module, W2CM), 후보 레이블 식별 모듈(Candidate Labels Identification Module, CLIM) 및 대표 레이블 결정 모듈(Representative Labels Determination Module, RLDM) 로 구성되어 있다. DCM은 문서를 입력 받아 문서-어휘 행렬로 변환한 뒤 비슷한 어휘의 패턴을 가진 문서들을 각각의 클러스터로 분류하는 기능을 수행한다. 그러나 문서의 클러스터링은 본 연구의 범위가 아니므로 더 이상의 논의를 하지 않는다.



(그림 1) 어휘 정보량을 이용한 문서 클러스터 레이블링의 전체 아키텍처

(Figure 1) Overall architecture of information content-based document cluster labeling

W2CM은 문서-어휘 행렬로 이루어진 개별 클러스터를 어휘와 어휘의 빈도로 이루어진 BOW(bag of words)로 변환한 뒤, 출현 빈도가 사전에 정의한 역치(δ)이하인 어휘와 워드넷에서 명사(noun)로 정의되지 않은 어휘를 필터링하는 기능을 수행한다. 필터링 결과를 이용해, 해당 클러스터에 포함된 개별 어휘의 중요도를 알기 위해 각각의 어휘에 대하여 해당 클러스터 내에서의 전체 출현 빈도와 어휘의 정보량을 이용하여 가중치(TF-ICF)를 계산한다.

CLIM은 W2CM을 통해 선택된 클러스터 내 어휘들을 이용해 후보 레이블을 식별하는 기능을 수행한다. CLIM은 클러스터 내 어휘들의 공통 상위어를 통해 클러스터의 내용을 잘 전달해 줄 수 있는 상위어를 식별하여 후보 레이블로 선정한다.

RLDM은 CLIM을 통해 생성된 후보 레이블 중 클러스터의 의미를 가장 잘 포괄하는 동시에 특징을 잘 표현할 수 있는 대표 레이블을 결정하는 기능을 수행한다. 이를 위해, 클러스터에 대한 후보 레이블의 중요도와 후보 레이블이 가지고 있는 정보량 기반의 가중치를 통해 상위 k개의 후보 레이블을 대표 레이블로 결정한다.

3.1 어휘 가중치 계산 모듈(W2CM)

본 연구의 목적은 클러스터의 특징을 대표하는 동시에 전체적인 내용을 포괄할 수 있는 대표 레이블을 결정하는 것이다. 이를 위해, 클러스터링의 결과로 생성된 문서-어휘 형태의 클러스터로부터 어휘를 추출하여 추출된 어휘의 가중치를 계산해야 한다. 문서 클러스터에 포함된 어휘는 특정 클러스터에서 차지하는 비중과 중요도에 따라 해당 클러스터와 관련성의 평가할 수 있으며. 이를 위해 클러스터에 포함된 어휘들의 가중치 계산이 선행되어야 한다.

어휘들의 가중치 계산은 다음과 같이 수행된다. 첫째, 특정 클러스터에 포함된 문서로부터 모든 유형의 명사형 어휘들을 추출한다. 둘째, 이들 어휘들이 특정 클러스터에 출현한 빈도를 도출하여 어휘 주머니에 저장한다. 이때 클러스터로부터 추출된 모든 어휘들중에서 클러스터와 의미적 연관관계가 있는 어휘들만을 선별하기 위해 워드넷에 존재하는 명사 어휘만 선별하여 어휘의 빈도와 함께 어휘 주머니에 저장한다. 문서 클러스터와 어휘 주머니에 대한 정의는 다음과 같다.

정의 1 i^{th} 문서 클러스터(document cluster) 클러스터링의 결과로 생성된 문서-어휘 행렬 형태의 i^{th}문서 클러스터 DC_i는 다음과 같이 정의된다.

$$DC_i = \begin{matrix} \\ d_i^1 \\ \vdots \\ d_i^k \end{matrix} \begin{matrix} t_{i1} \cdots \ t_{in} \\ \begin{bmatrix} f_{11} \cdots f_{1n} \\ \vdots \ \ddots \ \vdots \\ f_{k1} \cdots f_{kn} \end{bmatrix} \end{matrix} \quad (3)$$

이때, d_i^h는 DC_i에 포함된 h^{th}문서이며, t_{ij}는 d_i^h에 포함된 j^{th}어휘로 정의된다. 또한 f_{hj}는 h^{th}문서에 포함된 j^{th}어휘의 빈도를 의미한다 $(i=1,2,...,h=1,2,...,k,j=1,2,,n)$.

정의 2 i^{th} 어휘 주머니(bag of words) i^{th} 클러스터 DC_i에 포함된 어휘 t_{ij}와 그 어휘의 클러스터 내 전체 빈도 tf_{ij}로 이루어진 어휘-빈도 집합인 i^{th} 어휘 주머니 BOW_i 는 다음과 같이 정의된다.

$$BOW_i = \{(t_{ij}, tf_{ij}) | t_{ij} \in T_i, tf_{ij} \in TF_i\} \quad (4)$$

이때, T_i는 i^{th} 어휘 주머니 BOW_i에 속해있는 어휘 집합이며, TF_i는 i^{th} 어휘 주머니 BOW_i에 속해 있는 어휘들의 빈도 집합으로 정의된다$(i=1,2,...,j=1,2,...n)$.

또한 워드넷에 존재하지 않으며 해당 클러스터에 포함된 문서의 수 보다 출현 빈도가 적은 어휘는 BOW_i에서 필터링 되며 이때 남은 어휘 ct_{ij}에 대하여 BOW_i'는 다음과 같이 재정의된다.

$$BOW_i' = \{(ct_{ij}, tf_{ij}) | ct_{ij} \in CT_i, tf_{ij} \in TF_i\} \quad (5)$$

이때, CT_i는 T_i중 필터링 되고 남은 어휘의 집합으로 정의된다$(i=1,2,...,j=1,2,...m, m \le n)$.

BOW_i'에 저장된 어휘들은 해당 클러스터에서 일정 이상의 출현 빈도를 갖고 있다. 일반적으로 특정 어휘는 그 출현 빈도가 높을수록 중요도가 높다고 할 수 있다. 그러나 개별 어휘들의 출현 빈도로 중요도를 계산할 경우, 상용 어휘들에 대한 중요도가 높게 나올 수 있다. 따라서 클러스터의 특징을 포함하는 구체적인 어휘와 일반

적인 어휘를 구분하여 클러스터에 포함된 어휘의 중요도를 고려하기 위하여 어휘의 빈도에 추가적으로 어휘의 정보량을 활용하였다. 개별 어휘들이 일반적인 의미를 가지고 있을 경우 그 어휘의 정보량은 낮아지고, 각각의 어휘들이 구체적인 의미를 가지고 있을 경우 그 어휘의 정보량은 높아진다. 이를 클러스터 내 각 어휘의 출현 빈도와 함께 고려하여 해당 클러스터의 각각의 어휘 별 가중치를 계산한다.

BOW_i'에 포함된 어휘 ct_{ij}에 포함된 하위어의 수는 $n(Hypo_{ct_{ij}})$라고 정의한다. 그리고 BOW_i'에 포함된 어휘 ct_{ij}에 대한 정보량 $icf_{ct_{ij}}$은 다음과 같이 계산한다.

$$icf_{ct_{ij}} = \log\left(\frac{\max_{\forall j}(Hypo_{ct_{ij}})}{1+n(Hypo_{ct_{ij}})}\right) \quad (6)$$

이때, $n(Hypo_{ct_{ij}})$는 BOW_i'에 포함된 어휘 ct_{ij}의 하위어 수를 의미한다.

또한 BOW_i'에 포함된 어휘 ct_{ij}의 가중치는 tw_{ij}는 다음과 같이 계산한다.

$$tw_{ij} = tf_{ij} \times icf_{ct_{ij}} \quad (7)$$

이때, tf_{ij}는 BOW_i'에 포함된 어휘 ct_{ij}의 빈도를 의미한다.

BOW_i'에 포함된 어휘 ct_{ij}와 어휘의 가중치 tw_{ij}는 개선된 어휘 주머니 안에 저장된다. 개선된 어휘 주머니는 다음과 같이 정의된다.

정의 3 개선된 i^{th} 어휘 주머니 BOW_i'에 포함된 어휘 t_{ij}와 어휘의 빈도 tf_{ij}는 필터링 후 남은 어휘 ct_{ij}와 어휘의 가중치 tw_{ij}로 갱신되며, 개선된 어휘 주머니 $fBOW_i$에 저장되며 다음과 같이 정의된다.

$$fBOW_i = \{(ct_{ij}, tw_{ij}) | ct_{ij} \in CT_i, tw_{ij} \in TW_i\}, \\ (i = 1,2,...,j = 1,2,...m) \quad (8)$$

$fBOW_i$에 저장된 어휘 ct_{ij}는 워드넷을 통해 후보 레이블을 식별하기 위하여 사용되며 어휘의 가중치 tw_{ij}는 클러스터에 대한 후보 레이블의 중요도를 계산하는데 활용된다.

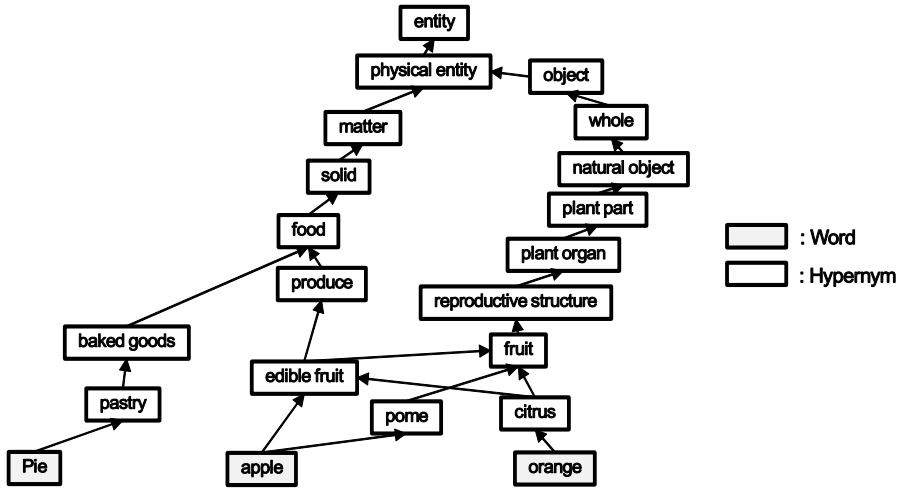
3.2 후보 레이블 식별 모듈(CLIM)

후보 레이블 식별 모듈은 어휘 가중치 계산 모듈에서 추출된 어휘들의 상위어를 통해 후보 레이블을 식별하는 모듈이다. 만약 후보 레이블을 식별하지 않을 경우 모든 단어 중에서 대표 레이블을 선정해야 하므로 그 범위가 너무 넓어 계산량이 많아지므로 후보 레이블 선정을 통해 문서 클러스터와 관련된 단어로 대표 레이블의 후보를 식별해야 한다.

이를 위해 먼저 $fBOW_i$에 포함된 어휘 ct_{ij}와 그 상위어를 워드넷에 맵핑시켜 워드 트리를 추출한다. 워드 트리란 어휘와 워드넷 상의 상위어들 간의 계층 관계를 트리 형태로 표현한 그래프이다. 워드넷 기반 워드 트리의 예를 (Figure 2)에 도식화하였다.

워드 트리를 이용하여 $fBOW_i$에 포함된 임의의 어휘들의 조합이 갖는 최근접 공통 상위어(Least Common Hypernym, LCH) 집합을 1차 후보 레이블로 식별한다. 최근접 공통 상위어란 두 어휘의 공통 상위어 중 두 어휘와의 거리가 가장 가까운 상위어이다. 예를 들면, 아래 워드 트리에서 'pie'와 'apple'간의 최근접 공통 상위어는 'food'고, 'apple'과 'orange'사이의 최근접 공통 상위어는 'edible fruit'이다. 이러한 최근접 공통 상위어를 통해 $fBOW_i$ 내 여러 어휘들을 포함하면서 그 중 가장 구체적인 의미를 가지고 있는 상위어를 1차 후보 레이블로 식별할 수 있다. 하지만 $fBOW_i$에 포함된 어휘 ct_{ij}들 간의 최근접 공통 상위어만 사용할 경우 어휘 ct_{ij}의 상위어 경로에 따라 식별되지 않는 유의한 상위어가 생길 수 있다. 따라서 식별된 1차 후보 레이블과 $fBOW_i$에 포함된 어휘 ct_{ij}의 최근접 공통 상위어를 통해 후보 레이블을 확장한다. 이는 기존의 방법으로 식별되지 않는 상위어로 추가 확장 가능하여 클러스터의 내용을 잘 전달해 줄 수 있는 후보 레이블 식별이 가능하다. 후보 레이블의 식별은 아래의 3단계로 수행된다.

Step 1. $fBOW_i$에 포함된 어휘 ct_{ij}와 그 상위어를 워드넷에 맵핑시켜 워드 트리를 생성한다. 이때, ct_{ij}를 기



(그림 2) 워드넷 기반 워드 트리(예)
(Figure 2) WordNet-based words tree (illustrative example)

준으로 워드넷으로부터 ct_{ij}의 모든 상위어를 추출한다.

Step 2. 워드 트리 WT_i에서 $fBOW_i$ 내 임의의 어휘 ct_{ij}의 조합이 갖는 최근접 공통 상위어 집합을 1차 후보 레이블로 선정한다. 후보 레이블을 선정하는 방법은 각 상위어가 포함하고 있는 ct_{ij}의 개수를 카운팅하며, 카운팅 과정에서 ct_{ij}의 개수가 바뀌는 지점의 상위어를 후보 레이블로 선정한다.

Step 3. 후보 레이블 Cl_i을 확장하는 단계이다. 1차 후보 레이블과 $fBOW_i$내 임의의 어휘 ct_{ij}의 조합이 갖는 말단 LCH 집합을 추가한다. 이때, 어휘의 경로에 따라 선정되지 않는 상위어만이 추가 대상이 된다.

마지막으로 각 상위어가 포함하고 있는 1차 후보 레이블 cl_{il} 개수를 카운팅하며, 카운팅 된 ct_{ij}의 개수와 합하여 카운팅 수가 바뀌는 지점의 상위어를 후보 레이블에 추가함으로써 후보 레이블 선정 과정이 종료된다. 이상의 과정을 알고리즘으로 요약하면 (Figure 3)과 같다.

3.3 대표 레이블 결정 모듈(RLDM)

대표 레이블 결정 모듈은 어휘 가중치 계산 모듈에서 계산된 어휘의 가중치와 후보 레이블 식별 모듈에서 추출된 워드 트리를 이용하여 클러스터에 대한 후보 레이블 cl_{il}의 가중치를 계산한다. 다음으로 계산된 후보 레이블의 가중치를 이용하여 top-k 개의 후보 레이블을 대표 레이블로 결정한다.

먼저 후보 레이블의 가중치를 계산하기 위해 워드 트리 내 모든 후보 레이블 cl_{il}의 정보량 $icf_{d_{il}}$을 계산한다. 후보 레이블의 정보량은 다음과 같이 도출된다.

$$icf_{d_{il}} = \log\left(\frac{\max_{\forall l}\left(Hypo_{d_{il}}\right)}{1+n\left(Hypo_{d_{il}}\right)}\right) \quad (9)$$

이때 $n\left(Hypo_{d_{il}}\right)$는 후보 레이블 cl_{il}의 하위어 수를 의미한다,

해당 클러스터에 대한 후보 레이블 cl_{il}의 정보량을 계산하기 때문에 후보 레이블 cl_{il}의 하위어 $Hypo_{d_{il}}$는 클러스터에 존재하는 어휘와 그 상위어로 생성된 워드 트리 내 존재하는 어휘로 제한한다.

다음 단계는 해당 클러스터에서 각각의 후보 레이블의 비중을 파악하기 위해 클러스터에 대한 후보 레이블 cl_{il}의 중요도 $imp_{d_{il}}$를 계산한다. 클러스터에 대한 후보 레이블의 중요도 $imp_{d_{il}}$의 계산식은 다음과 같다.

$$imp_{d_{il}} = \frac{\sum_{\forall p} tw_{lp}}{\sum_{\forall j} tw_{ij}} \quad (10)$$

$\boldsymbol{Hyper}_{ct_{ij}}$: $\boldsymbol{fBOW}_i$ 에 포함된 어휘 ct_{ij}의 상위어 집합
$\boldsymbol{Hyper}_{CT_i}$: $\boldsymbol{fBOW}_i$ 에 포함된 모든 어휘 ct_{ij}에 대한 상위어 집합
$hyper^{h}_{CT_i}$: $\boldsymbol{Hyper}_{CT_i}$에 속해 있는 h 번째 상위어
$\boldsymbol{WS}_i$: $\boldsymbol{fBOW}_i$ 에 포함된 전체 어휘와 $\boldsymbol{Hyper}_{CT_i}$에 포함된 상위어의 합집합
$\boldsymbol{WT}_i$: $\boldsymbol{WS}_i$의 어휘와 워드넷을 이용하여 만들어진 워드 트리
$\boldsymbol{Cl}_i$: i^{th} 클러스터에 대한 후보레이블 집합 (cl_{il}: $\boldsymbol{Cl}_i$ 속해 있는 l^{th} 후보레이블)

```
For all j
    Hyper_{ct_ij} ←Extract hypernym of ct_ij in fBOW_i using WordNet
EndFor
Hyper_{ct_ij} =∪_{∀j} Hyper_{ct_ij}
WS_i = T_i ∪ Hyper_{T_i}
Build the word tree WT_i using WS_i and WordNet
For all h
    n(hyper^h_{CT_i}) = 0.0
    For all j
        If t_ij is hypernym of hyper^h_{CT_i}
            Then n(hyper^h_{CT_i}) = n(hyper^h_{CT_i}) + 1
        EndIf
    Map n(hyper^h_{CT_i})to hyper^h_{CT_i}in WT_i
    EndFor
EndFor

For all h
    IF n(hyper^h_{CT_i}) > max(n(hypernym of hyper^h_{CT_i})) in WT_i
        Then Cl_i ← hyper^h_{CT_i}
    EndIf
EndFor
For all h
    For all cl_il
        If tcl_{il_ij} is hyponym of hyper^h_{CT_i}
            Then n(hyper^h_{CT_i}) = n(hyper^h_{CT_i}) + 1
        EndIf
    EndFor
EndFor
For all h
    IF n(hyper^h_{CT_i}) > max(n(hypernym of hypo^h_{CT_i})) in WT_i
        Then Cl_i ← hyper^h_{CT_i}
    EndIf
EndFor
```

(그림 3) 후보 레이블 식별 알고리즘
(Figure 3) Identification algorithm of candidate labels

이때, tw_{lp}는 후보 레이블 cl_{il}을 상위어로 갖는 p^{th} 어휘의 가중치이다. ($p \leq j$)

마지막으로 해당 클러스터의 의미를 가장 잘 포괄하면서 클러스터의 특징을 잘 표현할 수 있는 후보 레이블을 대표 레이블로 결정하기 위하여 앞서 계산한 후보 레이블 cl_{il}의 정보량 $icf_{cl_{il}}$과 후보 레이블 cl_{il}의 중요도 $imp_{cl_{il}}$를 통해 클러스터에 대한 후보 레이블 cl_{il}의 가중치 $w(cl_{il})$를 계산한다. 후보 레이블의 가중치 $w(cl_{il})$은 다음과 같이 도출한다.

$$w(cl_{il}) = icf_{cl_{il}} \times imp_{cl_{il}} \tag{11}$$

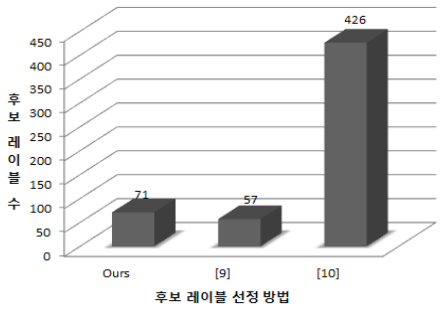
식 (11)을 이용해 해당 클러스터에 존재하는 모든 후보 레이블 cl_{il}에 대한 가중치 $w(cl_{il})$가 계산되면, 해당 클러스터에 존재하는 모든 후보 레이블 cl_{il}을 가중치 $w(cl_{il})$에 따라 내림차순으로 정렬한다. 그 후 정렬된 후보 레이블 cl_{il}중 가중치가 높은 상위 k개의 후보 레이블을 대표 레이블로 결정한다.

4. 실험 및 평가

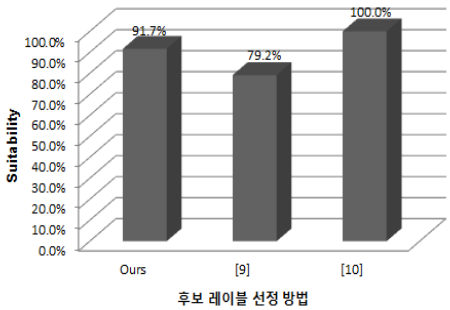
본 연구의 절차에 따라 결정된 대표 레이블의 적절성을 검증하기 위한 데이터로 Reuters-21578의 부분 집합을 활용하였다. Reuters-21578는 경제 관련 도메인과 관련 있는 문서의 집합으로 한 문서당 약 133개의 어휘를 포함하고 있다. 각각의 문서는 최소 하나의 카테고리로 분류

(표 2) 후보 레이블 수 비교
(Table 2) Comparison of the number of candidate labels by methods

Category	Words in $fBOW$	Ours	[9]	[10]
Earn	Ct, loss, net, quarter, shr	7	5	34
Acquisition	Company, corporation, group, pct, share, stake ,stock	8	7	53
Money-fx	Bank, currency, dollar, economy, exchange, finance, firm, market, money	9	7	59
Grain	Corn, crop, grain, pct, product, stock, ton	10	9	47
Crude	Barrel, company, crude, market, oil , OPEC, petroleum	11	8	45
Trade	Billion, export, import, market, minister, office, surplus, tariff, trade	10	8	89
Interest	Bank, bill, interest, market, pct, rate	9	7	57
Ship	Gulf, port, ship, tanker, union, vessel	7	6	42



(그림 4) 식별된 후보 레이블 수 비교
(Figure 4) Comparison of identified labels



(그림 5) 후보 레이블의 suitability
(Figure 5) Suitability of candidate labels

되어 있으며 약 90개의 카테고리가 존재한다. 본 연구에서는 그 중 8개의 카테고리로 분류되어 있는 클러스터에 속해있는 문서들을 활용하여 실험을 진행하였다.

본 연구의 우수성을 입증하기 위하여, 본 연구의 절차에 따라 선정된 레이블과 후보 레이블의 워드넷에서의 깊이와 후보 레이블을 상위어로 갖고 있는 클러스터 내 어휘의 수를 기반으로 한 휴리스틱 방법[9, 10]에 따라 선정된 레이블을 전문가가 찾은 대표 레이블과 비교하였다. 경제 관련 도메인과 관련된 문서가 포함된 Reuter-21578의 각 카테고리에 속해있는 어휘를 파악하고 적절한 레이블을 선정하기 위하여 경제, 경영 전공의 전문가 17명이 카테고리 별 어휘를 보고 각 카테고리 별 3개의 대표 레이블을 선정하였다.

본 연구에서는 두 가지 평가 지표를 통해 연구의 우수성을 평가하였다. 먼저 본 연구의 방법을 통해 선정된 후보 레이블을 전문가가 선정한 대표 레이블과 비교하여 후보 레이블의 적합성을 평가하였다. 후보 레이블의 적합성($Suitability_d$)이란 전문가가 선정한 대표 레이블 R_{User}중 각 방법으로 식별한 후보 레이블 CL이 있을 확률이며 다음과 같은 식으로 계산된다.

$$Suitability_d = p(CL|R_{User}) \quad (12)$$

후보 레이블의 적합성을 평가하기 위하여 먼저 기존 방법을 이용해 카테고리 별 후보 레이블을 식별한다. (Table 2)는 카테고리 별 어휘와 본 연구에서 제안한 방법과 기존 방법을 이용해 식별한 후보 레이블의 빈도를 나타낸다.

본 연구에서 제안한 방법과 기존 방법을 이용해 식별한 후보 레이블의 수는 (Figure 4)와 같으며, 이들의 $Suitability_d$는 (Figure 5)와 같다. 이를 통해 본 논문에서 제안한 방법을 적용해 후보 레이블을 선정할 경우 약 92%의 $Suitability_d$가 있음을 확인했으며, 이는 [10]의 방법보다 $Suitability_d$가 약간 낮지만 대표 레이블을 찾는 계산량을 기존의 약 20% 정도로 감소시키는 효과가 있으므로 $Suitability_d$값이 일부 하향되는 것에 대한 보상효과가 충분하다. 즉 기존의 방법으로 추출한 후보 레이블과의 적절성의 차이는 크지 않지만 그에 비해 수 많은 데이터를 분석하고 해석할 경우 발생하는 계산량이 감소하므로 효과적이라고 할 수 있다.

다음으로 본 연구의 절차에 따라 결정된 대표 레이블과 전문가가 선정한 대표 레이블을 비교하여 대표 레이블의 적절성($Appropriacy_{rl}$)을 평가하였다. 대표 레이블의 적절성이란 전문가가 선정한 대표 레이블 R_{User}과 해당 방법으로 결정된 대표 레이블 R_d의 일치 확률을 통해 알 수 있으며 계산식은 다음과 같다.

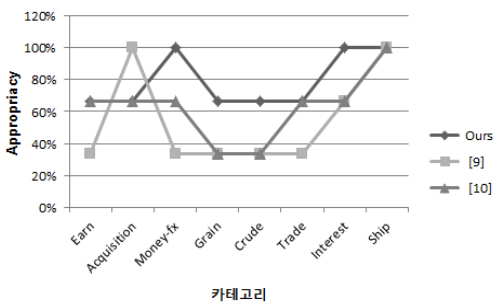
$$Appropriacy_{rl} = p(R_d|R_{User}) \quad (13)$$

본 연구에서 제안한 방법을 적용해 결정된 대표 레이블과 [9]와 [10]을 통해 결정된 각각의 대표 레이블을 (Table 3)에 요약했다.

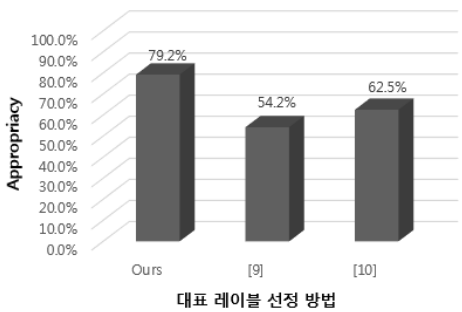
또한 본 연구에서 제안한 방법과 기존 방법을 이용해 결정한 카테고리 별 대표 레이블의 $Appropriacy_{rl}$과 전체 카테고리에 대한 대표 레이블의 $Appropriacy_{rl}$을 (Figure 6)과 (Figure 7)에 도식화하였다. 이를 통해 본 연구의 방법은 전문가가 선정한 대표 레이블과 약 79% 정도 대표 레이블이 일치하는 것을 확인할 수 있으며, 전체적으로 [9]보다는 약 25% 정도, [10]보다는 약 17% 정도 대표 레이블의

(Table 3) 대표 레이블 비교
(Table 3) Comparison of representative labels by the methods

Category	Ours	[9]	[10]
Earn	1. Asset	1. Possession	1. Possession
	2. Possession	2. Relation	2. Relation
	3. Relation	3. Measure	3. Asset
Acquisition	1. Share	1. Stock	1. Stock
	2. Stock	2. Asset	2. Share
	3. Asset	3. Organization	3. Asset
Money-fx	1. Commercial enterprise	1. Act	1. Finance
	2. Finance	2. Enterprise	2. Commercial enterprise
	3. Monetary system	3. Commercial enterprise	3. Act
Grain	1. Food Stuff	1. Seed	1. Seed
	2. Seed	2. Product	2. Grain
	3. Grain	3. Food stuff	3. Product
Crude	1. Oil	1. Lipid	1. Organic compound
	2. Fossil fuel	2. Organic compound	2. Lipid
	3. Organic compound	3. Oil	3. Oil
Trade	1. Place of business	1. Unit	1. Place of business
	2. Trade good	2. Artifact	2. Establish
	3. Act	3. Place of business	3. Trade good
Interest	1. Possession	1. Possession	1. Relation
	2. Transferred property	2. Relation	2. Possession
	3. Currency	3. Transferred property	3. Transferred property
Ship	1. Vessel	1. Craft	1. Ship
	2. Craft	2. Vessel	2. Vessel
	3. Ship	3. Ship	3. Craft



(그림 6) 카테고리 별 대표 레이블의 appropriacy
(Figure 6) Appropriacy of representative labels by categories



(그림 7) 대표 레이블의 appropriacy
(Figure 7) Appropriacy of representative labels

$Appropriacy_{rl}$이 높다는 것을 확인할 수 있다. 이를 통해 기존의 방법보다 대표 레이블의 $Appropriacy_{rl}$이 높다는 것을 알 수 있다.

5. 결론 및 추후 연구

본 연구에서는 먼저 클러스터에 포함된 어휘들이 해당 클러스터에서 얼마나 중요한 비중을 차지하고 있는지 어휘의 빈도와 정보량을 이용하여 해당 어휘의 가중치를 계산한다. 다음으로, 클러스터 내 포함된 어휘들의 상위 어를 통해 후보 레이블로 식별하고 그 중 해당 클러스터의 의미를 가장 잘 포괄하면서 클러스터의 특징을 잘 표현할 수 있는 대표 레이블을 결정한다. 실험은 전문가가 선정한 대표 레이블과 비교하는 방식으로 진행하였고 본 연구의 절차에 따라 식별된 후보 레이블의 경우 기존 방

법과 적절성의 차이는 크지 않지만 그에 비해 수 많은 데이터를 분석하고 해석할 경우 발생하는 계산량이 기존의 20% 정도로 크게 감소하여 보상효과가 충분하다 판단할 수 있으며, 기존의 방법과 본 연구의 방법을 통해 선정한 대표 레이블과 전문가가 선정한 대표 레이블의 비교를 통해 본 연구의 대표 레이블이 적절하다는 것을 입증하였다.

본 연구가 문서 클러스터 레이블 선정 방법과 관련하여 기여하는 바는 다음과 같다. 첫째, 정보량을 이용하여 클러스터에 포함된 어휘의 일반적 구체적 정도를 클러스터 내 어휘의 출현 빈도와 함께 적용하여 클러스터에 포함된 각각의 어휘의 중요도를 계산하였다. 둘째, 다중 경로를 가지고 있는 어휘들의 상위어로 후보 레이블 식별 시 가능한 적절한 상위어 선택에 대한 방법론을 고안하였다. 마지막으로, 식별된 후보 레이블 중 해당 클러스터에 적절한 대표 레이블을 찾기 위해 후보 레이블의 정보량과 중요도에 따른 가중치를 계산하여 top-k의 후보 레이블을 결정하였다.

본 연구는 추후 여러 방향으로 확장하여 추후 연구를 수행할 수 있다. 첫 번째, 각각의 클러스터에서 어휘를 추출할 때, 어휘의 도메인을 활용해 도메인 별 어휘를 추출한다면 클러스터에 알맞은 도메인과 그 도메인에 속한 어휘를 찾을 수 있을 것이다. 또한, 본 연구에서 제안한 방법은 워드넷의 계층 관계를 통해 어휘 간 상위어, 하위어 관계를 고려하였는데, 워드넷은 일반적인 단어들과 특정한 도메인에 속한 단어를 구분하기 어렵다. 따라서 특정한 도메인에 대한 어휘 간 계층 정보를 활용한다면, 추후 워드넷이 가지고 있는 한계를 개선할 수 있을 것이다. 마지막으로, 본 연구에서 제안한 방법을 활용하여 실제 클러스터링 후 생성된 클러스터를 입력 받아 대표 레이블을 결정할 수 있는 레이블링 자동화 시스템을 구현할 필요가 있다. 문서 클러스터링 후 생성된 개별 클러스터에 대해 대표 레이블을 자동적으로 결정해준다면 클러스터에 대한 사용자의 이해를 제고시킬 수 있을 것이다.

참 고 문 헌(Reference)

[1] Q. Mei, X. Shen, and C. Zhai, “Automatic labeling of multinomial topic models,” *In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 490-499, 2007. https://doi.org/10.1145/1281192.1281246

[2] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts,” *Association for Computational Linguistics*, 2004. http://digital.library.unt.edu/ark:/67531/metadc30962/

[3] W. Lu, Q. Cheng and C. Lioma, “Fixed versus dynamic co-occurrence windows in TextRank term weights for information retrieval,” *In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,* pp. 1079-1080, 2012. https://doi.org/10.1145/2348283.2348478

[4] F. Role and M. Nadif, “Beyond cluster labeling: Semantic interpretation of clusters’ contents using a graph representation,” *Knowledge-Based Systems*, vol. 56, pp. 141-155, 2014. http://dx.doi.org/10.1016/j.knosys.2013.11.005

[5] C. T. Nguyen, X. H. Phan, S. Horiguchi, T. T. Nguyen and Q. T. Ha, “Web search clustering and labeling with hidden topics,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, issue. 3, pp. 12, 2009. https://doi.org/10.1145/1568292.1568295

[6] Z. S. Syed, T. Finin and A. Joshi, “Wikipedia as an Ontology for Describing Documents,” *In ICWSM*, 2008. http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-024.pdf

[7] D. Carmel, H. Roitman and N. Zwerdling, “Enhancing cluster labeling using Wikipedia,” *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 139-146, 2009. https://doi.org/10.1145/1571941.1571967

[8] Z. Li, J. Li, Y. Liao, S. Wen and J. Tang, “Labeling clusters from both linguistic and statistical perspectives: A hybrid approach,” *Knowledge-Based Systems*, vol. 76, pp. 219-227, 2015. http://dx.doi.org/10.1016/j.knosys.2014.12.019

[9] Y. H. Tseng, “Generic title labeling for clustered documents,” *Expert Systems with Applications*, vol. 37, issue. 3, pp. 2247-2254, 2010. http://dx.doi.org/10.1016/j.eswa.2009.07.048

[10] C. Bouras and V. Tsogkas, “A clustering technique for news articles using WordNet,” *Knowledge-Based Systems*, vol. 36, pp. 115-128, 2012.

http://dx.doi.org/10.1016/j.knosys.2012.06.015

[11] W. H. Gomaa and A. A. Fahmy, “A survey of text similarity approaches,” *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13-18, 2013. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.5446&rep=rep1&type=pdf

[12] D. Sánchez, M. Batet, D. Isern and A. Valls, “Ontology-based semantic similarity: A new feature-based approach,” *Expert Systems with Applications*, vol. 39, issue. 9, pp. 7718-7728, 2012. http://dx.doi.org/10.1016/j.eswa.2012.01.082

[13] G. A. Miller, “WordNet: a lexical database for English,” *Communications of the ACM*, vol. 38, issue. 11, pp. 39-41, 1995. https://doi.org/10.1145/219717.219748

[14] T. Pedersen, S. Patwardhan and J. Michelizzi, “WordNet: Similarity: measuring the relatedness of concepts,” *In Demonstration papers at HLT-NAACL 2004*, pp. 38-41, 2004. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1614037

[15] WordNet, “A lexical database for the English language,” Cognitive Science Laboratory, Princeton University. 2004. http://wordnet.princeton.edu

[16] P. Treeratpituk and J. Callan, “Automatically labeling hierarchical clusters,” *In Proceedings of the 2006 international conference on Digital government research*, pp. 167-176, 2006. https://doi.org/10.1145/1146598.1146650

[17] H. Anaya-Sánchez, A. Pons-Porrata and R. Berlanga-Llavori, “A new document clustering algorithm for topic discovering and labeling,” *In Iberoamerican Congress on Pattern Recognition*, pp. 161-168, 2008. https://link.springer.com/chapter/10.1007/978-3-540-85920-8_20

[18] T. Okuoka, T. Takahashi, D. Deguchi, I. Ide and H. Murase, “Labeling news topic threads with Wikipedia entries,” *11th IEEE International Symposium on Multimedia*, pp. 501-504, 2009. https://doi.org/10.1109/ISM.2009.67

[19] X. L. Mao, Z. Y. Ming, Z. J. Zha, T. S. Chua, H. Yan and X. Li, “Automatic labeling hierarchical topics,” *In Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2383-2386, 2012. https://doi.org/10.1145/2396761.2398646

[20] J. H. Lau, K. Grieser, D. Newman and T. Baldwin, “Automatic labelling of topic models,” *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1,* pp. 1536-1545, 2011. http://dl.acm.org/citation.cfm?id=2002658

[21] I. Hulpus, C. Hayes, M. Karnstedt and D. Greene, “Unsupervised graph-based topic labelling using dbpedia,” *In Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 465-474, 2013. https://doi.org/10.1145/2433396.2433454

[22] H. Roitman, S. Hummel and M. Shmueli-Scheuer, “A fusion approach to cluster labeling,” *In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 883-886, 2014. https://doi.org/10.1145/2600428.2609465

[23] A. Panchenko and O. Morozova, “A study of hybrid similarity measures for semantic relation extraction,” *In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pp. 10-18, 2012. http://dl.acm.org/citation.cfm?id=2388634

[24] S. Hingmire, S. Chougule, G. K. Palshikar and S. Chakraborti, “Document classification by topic labeling,” *In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 877-880, 2013. https://doi.org/10.1145/2484028.2484140

[25] T. H. Kim, “A study of Document Cluster Labeling using Information Content of words”, Master Dissertation of Sungkyunkwan Unversity, 2016. http://dcollection.skku.edu/jsp/common/DcLoOrgPer.jsp?sItemId=000000096202

◐ 저 자 소 개 ◑

김 태 훈 (Tae-Hoon kim)
2015년 성균관대학교 시스템경영공학과 학사
2016년 성균관대학교 산업공학과 석사
관심분야 : 온톨로지, 웹데이터 마이닝

손 미 애 (Mye Sohn)
1985년 성균관대학교 산업공학과
1988년 한국과학기술원 산업공학 석사
2002년 한국과학기술원 경영정보공학 박사
1998년~2004년 한국국방연구원
2004년~현재 성균관대학교 산업공학과 교수
관심분야 : IOT (Internet of Things), LOD (Linked Open Data), 온톨로지 및 상황인지 서비스