Journal of the Korea Academia-Industrial cooperation Society
Vol. 19, No. 11 pp. 758-764, 2018

https://doi.org/10.5762/KAIS.2018.19.11.758
ISSN 1975-4701 / eISSN 2288-4688

역인덱스 기반 상향식 군집화 기법을 이용한 대규모 학술 핵심어 분석

오흥선[1], 정유철[2*]
[1]한국기술교육대학교 컴퓨터공학부, [2]금오공과대학교 컴퓨터공학과

Analysis of Massive Scholarly Keywords using Inverted-Index based Bottom-up Clustering

Heung-Seon Oh[1], Yuchul Jung[2*]
[1]**School of Computer Science and Engineering, KOREATECH**
[2]**Computer Engineering, Kumoh National Institute of Technology**

요 약 특허(patent), 학술 논문(scholarly paper)과 연구 보고서(research report)와 같은 디지털 문서(digital document)에는 주제(topic)를 요약하는 저자 키워드(author keyword)가 있다. 서로 다른 문서가 동일한 키워드를 공유하고 있다면 두 문서가 동일한 주제의 내용을 기술하고 있을 가능성이 매우 높다. 문서 군집화(document clustering)는 비슷한 주제를 가지는 문서들을 비지도 학습 방법(unsupervised learning)을 이용하여 같은 군집으로 그룹(group)화 하는 것이다. 문서 군집화는 다양한 분석에 이용되지만 대용량의 문서 데이터에 적용하기 위해서는 많은 계산량이 필요함으로 쉽지 않다. 이러한 경우, 문서의 내용을 이용하는 것보다 문서의 키워드를 이용하여 군집화하면 더욱 효율적으로 대용량의 데이터를 연결할 수 있다. 기존의 상향식 군집화 방법(bottom-up hierarchical clustering)은 대용량의 키워드 군집화(keyword clustering)를 수행하는데 있어서 많은 시간이 필요하다는 문제점이 있다. 본 논문에서는 정보검색(information retrieval)에서 널리 사용되는 역인덱스(inverted-index) 구조를 상향식 군집화에 적용한 효율적인 군집화 방법을 제안하고, 제안 방법을 대용량의 키워드 데이터에 적용하였으며, 그 결과를 분석하였다.

Abstract Digital documents such as patents, scholarly papers and research reports have author keywords which summarize the topics of documents. Different documents are likely to describe the same topic if they share the same keywords. Document clustering aims at clustering documents to similar topics with an unsupervised learning method. However, it is difficult to apply to a large amount of documents event though the document clustering is utilized to in various data analysis due to computational complexity. In this case, we can cluster and connect massive documents using keywords efficiently. Existing bottom-up hierarchical clustering requires huge computation and time complexity for clustering a large number of keywords. This paper proposes an inverted index based bottom-up clustering for keywords and analyzes the results of clustering with massive keywords extracted from scholarly papers and research reports.

Keywords : Keyword clustering, Inverted-index, keyword analysis, bottom-up clustering, information retrieval

이 논문은 2018년도 한국기술교육대학교 교수 교육연구진흥과제 지원에 의하여 연구되었음. 이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2018R1C1B5031408)
*Corresponding Author : Yuchul Jung (Kumoh National Institute of Technology)
Tel: +82-42-478-7536 email: jyc@kumoh.ac.kr
Received August 16, 2018 Revised September 11, 2018
Accepted November 2, 2018 Published November 30, 2018

1. 서론

대부분의 디지털 문서(논문, 연구보고서 등) 에는 문서의 주제(topic)를 요약하는 키워드(keyword)가 있다. 문서의 저자가 직접 작성한 키워드(author keywords)는 문서의 주제를 함축적으로 나타낸다. 그러므로 두 문서가 동일한 키워드를 가지고 있을 경우에는 두 문서의 주제가 같을 가능성이 매우 높다. 키워드를 이용하면 문서의 모든 내용을 분석하지 않아도 문서 간의 주제적 연관성을 쉽게 분석할 수 있다. 그림 1은 여러 종류의 문서들이 동일한 키워드로 연결된 모습을 보여준다. 동일한 키워드를 공유하는 논문(paper)과 특허(patent) 문서는 동일한 주제를 나타낼 가능성이 매우 높다. 키워드는 정보검색(information retrieval) [1], 자연어처리(natural language processing)[2], 정보요약(text summarization)[3], 정보시각화(information visualization)[4] 등 다양한 분야에서 활용이 가능하다. 키워드를 유용하게 활용하기 위해서 키워드가 없는 문서로부터 키워드를 추출하려고 하는 다양한 연구가 수행되고 있다[5][6].

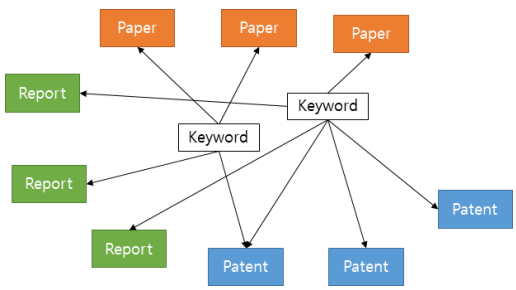

Fig. 1. Associations of digital documents keywords

대부분의 디지털 문서에서 키워드는 미리 정의된 사전(controlled vocabulary)에서 선택하여 사용되지 않고 저자에 의해서 생성된다. 저자 키워드를 사용하면 다양한 주제를 허용하는 장점도 있지만, 다음과 같은 단점들도 발생된다. 첫 번째는 다양한 키워드가 동일한 주제를 의미하는 이음동의어 문제(synonym problem)와 동일한 키워드가 문맥에 따라서 다양한 주제를 의미하는 동음이의어 문제(homonym problem)이다. 두 번째는 띄어쓰기 오류와 오탈자 문제이다. 저자가 키워드 작성 시에 잘못된 입력으로 인해서 디지털 문서에 오류가 존재할 수 있다. 세 번째는 언어의 다양성이다. 논문과 연구보고서 같은 특정한 문서 형태에서는 대부분 한글 키워드와 그에 상응하는 영어 키워드가 존재한다. 이와 같은 상황에서는 한글과 영어를 동시에 처리하는 것이 더욱 효과적이지만 각 언어에 대한 이해와 처리 방법이 다를 수 있다. 네 번째는 디지털 문서부터 추출 가능한 고유한 키워드의 수가 매우 많아서 처리하는데 큰 계산량이 필요하다.

군집화(clustering)[7]는 비지도학습 (unsupervised learning)의 대표적인 방법으로써 유사한 객체(object)들을 동일한 그룹(group)으로 군집화하는 것을 나타낸다. 계층적 군집화(hierarchical clustering)는 군집화를 수행함에 있어서 군집들(clusters)의 계층(hierarchy)을 생성하는 것이다. 계층적 군집화는 상향식(bottom-up) 방법과 하향식 방법(top-down)이 있다. 상향식 방법은 주어진 데이터의 모든 데이터 포인트에 대해서 상응하는 군집을 생성하고 유사한 군집끼리 합치는 방법이고, 하향식 방법은 데이터의 모든 데이터 포인트들을 하나의 군집에 할당한 뒤에 군집을 분리하면서 계층을 생성하는 방법이다. 계층적 군집화는 다양한 데이터 분석에 사용되고 있지만, 대용량의 데이터에 적용할 경우에 다량의 계산량 때문에, 연산시간이 많이 소요되는 문제점이 있다.

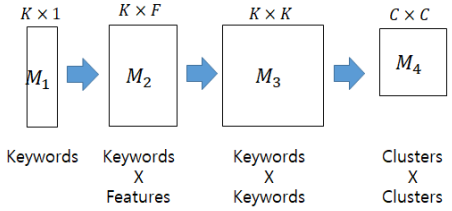

Fig. 2. Matrix representations for keyword data

그림 2는 키워드 군집화를 위한 데이터의 행렬(matrix) 표현을 나타낸다. 전체 키워드의 수가 K일 때 키워드 x 자질(feature) M_2 행렬을 생성하고 이를 이용하여 최초의 군집 유사도 행렬 M_3를 계산하고 여러 단계에 걸쳐서 유사한 군집을 병합하여 크기가 작은 M_4을 생성한다. M_4에서 유사도가 큰 두 개의 군집을 선택하여 병합하고 병합된 군집들에 대해서 다시 M_4를 생성한다. 특정 임계치를 설정하여 이 과정을 더 이상 병합할 수 없을 때 까지 반복한다. 키워드 데이터의 크기 K가 크면 초기 M_4의 행렬이 K x K의 크기를 가지므로 일반적인 상향식 방법으로 군집화를 수행할 경우에 많은 시간이 필요하다.

본 논문에서는 효과적이고 연산 시간측면에서 효율적

인 키워드 군집화 방법을 제안하고, 이를 이용한 키워드 분석을 수행하였다. 보다 구체적으로는, 정보검색(information retrieval) 분야에서 널리 사용되는 역인덱스(inverted index)[8] 구조를 이용한 상향식 군집화 방법이다. 일반적으로 키워드 대상 군집과 유사한 다른 키워드 후보 군집을 찾기 위해서는 모든 키워드 군집과의 유사도를 계산해야 한다. 이에 반해, 군집들에 대한 역인덱스를 구축하여 이용하면 공통의 자질을 가지고 있는 군집을 빠르게 찾아 군집 병합을 수행함으로 모든 군집에 대한 유사도를 계산하지 않아 효율을 높일 수 있다. 또한 문자와 단어 수준의 자질을 이용하여 띄어쓰기와 오탈자 문제에 대응할 수 있다. 이는 언어에 독립적인 자질로써 다양한 언어에 적용가능하다.

본 논문은 다음과 같이 구성된다. 2장에서는 역인덱스 기반의 상향식 군집화 방법을 소개한다. 3장에서는 제안한 방법을 대용량의 키워드 데이터에 적용한 결과를 소개하고 4장에서는 본 논문의 결론을 기술한다.

2. 제안 방법

본 논문에서 역인덱스 기반 다단계 상향식 군집화 방법을 제안하는데, 크게 5단계의 절차를 거치면서 키워드 군집화를 수행한다.

우선, 하나의 키워드는 키워드 ID, 한글 문자열, 영어 문자열로 구성된다고 가정하자.

$$k_i=(id,\ s_{kor},\ s_{eng})$$

키워드는 다음 예제처럼, 한글 또는 영어 문자열을 가지고 있다.

k_1=(729213, "공전현상", "slip phenomena")
k_2=(349038, "", "Slip phenomena")
k_3=(583717, "", "slip phenomena")

k_1은 한글과 영어 문자열을 모두 가지고 있다. 반면, k_2와 k_3 는 영어 문자열만 가지고 있으나 대문자의 유무에 따라서 다른 키워드로 사용된다.

1단계에서는 모든 군집을 초기화한다. 키워드 데이터 K에 있는 각각의 키워드에 군집을 생성한다. 각 군집 $c_i=\{k_i\}$이 된다.

2단계에서는 한글과 영어 문자열을 이용해 문자키를 생성하고 이를 이용하여 군집을 병합한다. 이때 입력 키워드 k에 대해서 문자키 생성 함수 keygen(k) = $keygen_{kor}$(k) + $keygen_{eng}$(k)를 이용한다. 한글 문자열 키 생성 함수 $keygen_{kor}$(s)는 입력 문자열의 모든 부호문자(punctuation)와 공백을 제거하여 문자키를 생성한다. 영어 문자열 키 생성 함수 $keygen_{eng}$(s)는 우선 어근추출 기술(stemming)[9]을 이용하여 단어의 어근(stem)을 추출하고 소문자로 변환한다. 어근 문자열에서 부호문자와 공백을 제거하여 문자키를 생성한다. 각 군집은 동일한 문자키를 가지고 있는 키워드를 멤버로 가지고 있다.

3단계에서는 한글 문자열을 이용해 문자키를 생성하고 이를 이용하여 군집을 병합한다. 두 군집 c1과 c2가 $keygen_{kor}$(k)을 이용하여 동일한 문자키를 가지고 있으면 병합한다.

4단계에서는 한글 문자열 검색을 이용한 군집화를 수행한다. 주어진 군집 c에 대해서 정보검색 기술을 이용하여 유사도가 높은 K개의 군집을 찾고 이들에 대해서 센트로이드 (centroid)기반의 유사도를 계산하고 임계치(threshold)를 넘는 가장 유사도가 높은 군집을 병합의 후보로 선택한다. 대상 군집 c와 유사도가 높은 후보 군집을 검색하기 위해서는 역인덱스를 이용한다. 역인덱스는 단어와 문서의 관계를 나타내는 핵심적인 자료구조이다. 단어가 출현한 문서를 리스트로 나타내고 검색 시에 모든 문서를 검색하지 않고 단어가 출현한 문서만을 대상으로 검색을 수행함으로써 효율적인 검색을 수행할 수 있게 한다. 이를 위해서 키워드로부터 q-gram을 생성하고 q-gram과 군집 사이의 역인덱스를 구축한다. q-gram은 주어진 문자열로부터 q개의 일련의 부분 문자열을 생성하는 것이다. 예를 들어, 문자열 s="공전현상"에 대한 3-gram은 {"공전현", "전현상"}이다. 그림 3은 q-gram을 이용한 역인덱스 생성을 보여준다. 각 군집 c에 있는 모든 키워드를 사용해서 gram을 생성한다.

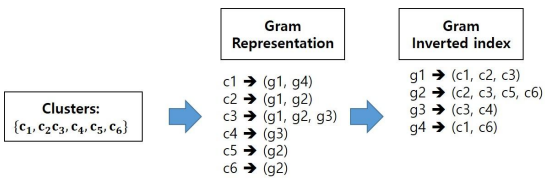

Fig. 3. Overview of constructing an inverted index using q-grams

역인덱스를 이용하여 대상 군집 c_q에서 출현한 q-gram을 가지고 있는 모든 후보 군집 C_o을 검색한다. 군집화가 진행되면 특정 군집 c는 많은 키워드를 포함하여 q-gram의 수가 증가하고 결과적으로 많은 수의 후보 군집이 검색된다. 이때 각 군집에서 중요한 소수의 q-gram을 선택하여 후보 군집을 검색하는 것이 효율적이다. q-gram의 가중치 계산 및 선택에는 정보검색에서 문서에서 단어의 가중치를 계산하는데 널리 사용되는 TF-IDF[10]를 이용하였다.

$$TFIDF(f,c) = \log(freq(f,c)) \times \log(\frac{N+1}{df(f)}) \quad (1)$$

freq(f,c)는 군집 c에서 자질 f의 출현빈도수, df(f)는 자질 f가 출현한 군집의 수, N은 전체 군집의 수이다.

TF-IDF를 이용하여 대상 군집으로부터 10개의 q-gram을 선택하고 역인덱스를 이용하여 선택된 q-gram이 출현한 군집들을 검색하여 최종 후보 군집들을 생성하였다.

마지막으로 대상 군집과 후보 군집 사이의 유사도 점수를 계산해서 점수가 가장 높은 군집 c_o을 선택해야 한다.

$$c_o^* = \arg max_{c_o \in C_o, c_o \neq c_q} score(c_q, c_o) \quad (2)$$

모든 군집을 대상 군집으로 하여 후보 군집을 선정하여 저장한다. 이때 대상 군집과 후보 군집의 유사도 점수가 임계치(τ)보다 작으면 저장하지 않는다. 본 논문에서 $\tau = 0.9$로 설정하였다.

두 군집 사이의 유사도 점수는 한글 문자열과 영어 문자열에 대한 코사인 유사도의 합을 이용하였다.

$$score(c_q, c_o) = \frac{\cos_{kor}(c_q, c_o) + \cos_{eng}(c_q, c_o)}{2} \quad (3)$$

코사인 유사도를 사용하기 위해서는 두 군집을 자질 벡터(feature vector)로 표현해야 한다. 이 자질 벡터를 센트로이드라 한다[11]. 벡터의 각 엔트리(entry)는 자질 ID와 자질의 가중치(weight) 또는 중요도를 나타낸다. 자질의 가중치는 TF-IDF을 이용하였다.

본 논문에서는 $\cos_{kor}$함수를 위해서 1-grams와 3-grams을 $\cos_{eng}$함수를 위해서 1-grams을 각각 자질로 사용했다. 아래는 4단계의 알고리즘을 정리한 것이다.

1: loop
2: build an inverted index for features x clusters
3: build centorids for clusters C
4: F={}
5: for each cq in C
6: select important q-grams using TF-IDF
7: search candidate clusters using the inverted index with the selected q-grams
8: select c_o^* by computing $score(c_q, c_o)$
9: store $(c_q, c_o^*, score)$ in F if $score(c_q, c_o^*) \geq \tau$
10: if F is empty, stop merging
11: select (c_q, c_o^*) with the highest similarity score in F and merge them

마지막, 5단계에서는 각 군집을 대표하는 레이블(label) 키워드를 선택한다. 각 군집을 대표하는 키워드를 선택하는 가장 쉬운 방법은 가장 빈도수가 많은 키워드를 선택하는 것이다[12]. 그러나 빈도수가 많으나 대표 키워드가 아닌 경우도 존재한다. 본 논문에서는 각 군집 c={k_1,k_2,...,k_n}에 대해서 군집의 센트로이드와 유사도 점수가 가장 높은 키워드를 선택한다. 적용한 유사도 점수는 4단계의 코사인 유사도 계산 방법과 유사하며, 최종적으로는 군집 c와 키워드 k에 대해서 한글 유사도와 영어 유사도를 각각 계산하여 합한다.

$$k^* = \arg max_{k \in c} score(k, c) \quad (5)$$

3. 결과 분석

3.1 데이터

본 논문에서는 학술논문(paper)과 연구보고서(research report)에 있는 키워드를 추출하여 사용하였다. 표 1은 사용한 데이터 통계를 보여준다. 대략 1M 학술논문들과 0.21M 연구보고서로부터 대략 3M 키워드와 0.5M 키워드를 추출하였다. 전체 문서 중에서 키워드를 가지고 있는 문서의 비율은 48%이다. 이 중 중복을 제외한 1,393,329개의 고유한 키워드를 확보하였다.

Table 1. Data statistics

Type	# docs	# keywords	# docs with keywords
Paper	1,323,658	2,986,414	659,902
Report	212,651	480,854	79,324

그림 4는 수집된 키워드 샘플을 보여주고 있다. 하나의 키워드는 한글과 영어 쌍으로 구성된다. 이때 한글이나 영어만 존재할 수 있다. 또한 키워드에는 오탈자가 존재한다. 그러므로 동일한 키워드에 대한 다양한 표현이 존재한다.

실단면 한계하중	Net-section limit load
실대수 G 다양체	real algebraic G variety
실대실험	Large-scale model experiment
실대크기	real scale
실대형 성토모형실험	model test of real scale embankment
실대형 시험	large scale model test
실대형 실험	
실대형 실험	large-scale model teste
실대형 터널 라이닝 섹션 하중재하 실험	
실대형 환경챔버	
실대형시험	real scale test
실대형실험	Full-scale test
실대형실험	Full-scale testing
실대형실험	Large scale model test
실데나필 유도체	Sildenafil Analogs

Fig. 4. Keyword samples

3.2 결과

K는 전체 키워드 데이터이며 $|K| = 1{,}393{,}329$이다. 최초의 군집의 수는 전체 키워드 데이터의 수와 같다. 키워드 데이터를 이용해서 역인덱스 기반 상향식 군집화를 수행하여 생성된 군집 C의 개수는 716,311이다. 그림 4는 동일한 멤버 키워드 수를 가지고 있는 군집의 수를 나타낸다. 62%의 군집은 멤버 키워드가 1개이고 19%의 군집은 멤버 키워드가 2개이다. 나머지 19%의 군집은 멤버 키워드가 3개 이상이다. 이는 사용한 데이터에 다양한 키워드가 포함되어 있다는 것을 알 수 있다.

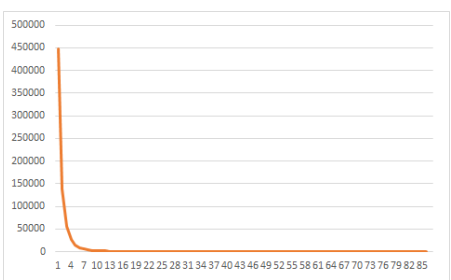

Fig. 4. Number of clusters with member counts

그림 5는 생성된 군집 중에서 멤버 키워드가 많은 상위 20개의 군집에 대한 레이블과 멤버 키워드의 수를 보여준다. 전반적으로 다양한 주제를 나타내는 키워드들이 있는 것을 볼 수 있다.

Seq	Korean	English	Members
1	유한요소법	Finite Element Method	846
2	계층분석법	Analytic Hierarchy Process	486
3	노인	Elderly	456
4	탄소섬유 강화 플라스틱	Carbon Fiber Reinforced Plastics	339
5	회귀분석	Regression Analysis	324
6	고분자 전해질막 연료전지	Polymer Electrolyte Membrane Fuel Cell	94
7	유전자 알고리즘	Genetic Algorithm	94
8	자기효능감	Self-efficacy	93
9	삶의 질	Quality of life	90
10	신호전달	signal transduction	89
11	신재생에너지	Renewable energy	87
12	만족도	Satisfaction	87
13		Taguchi Method	86
14	공동주택	Apartment	85
15	학업성취도	academic achievement	85
16	피로균열진전	Fatigue Crack Growth	82
17	다물체 동역학	Multibody Dynamics	82
18	수치해석	Numerical analysis	82
19	고로슬래그 미분말	blast furnace slag	80
20	직무만족	Job satisfaction	79

Fig. 5. Samples of top-20 cluster labels

그림 6는 생성된 군집 중 멤버 키워드를 가장 많이 포함하고 있는 군집 “유한요소법”에 대한 샘플을 보여준다. 군집의 레이블은 5단계에서 군집과 키워드 간의 유사도를 이용하여 선택하였다. 키워드는 저자가 직접 작성한 것이므로 동일한 내용에 대해서 다양하게 표현을 볼 수 있다. 표현의 다양성은 띄어쓰기(“유한요소해석”, “유한 요소 해석”, “유한 요소해석”), 오탈자(“inite Element Method”), 약어(“FEM”, “F.E.M”,“FEA”) 등의 원인이 된다. 한글과 영어가 전혀 다른 의미를 갖는 경우도 존재한다. “유한요소법”에 대한 “genetic algorithm”, “frame ratio”와 같은 영어 표현은 완전히 틀렸다고 볼 수 있다.

하나의 군집에서 비슷한 키워드일지라도 본문의 문맥에 따라서 다른 의미를 갖는 서로 다른 키워드일 수 있고 서로 다른 키워드라도 동일한 의미를 갖을 수 있다. 본 논문에서 사용한 데이터인 학술논문과 연구보고서에 있는 키워드 데이터는 다양성과 전문성이 매우 높다. 그렇기에 군집화의 성능을 판단하기 위해서는 다양한 분야의 전문지식이 필요한데 이러한 평가를 수행하기는 현실적으로 쉽지 않다.

No:	1		
ID:	621222		
Label:			
1114909	유한요소법	Finite Element Method	
Keywords:	282		
1:	1114909	("유한요소법", "Finite Element Method")	615
2:	1115058	("유한요소해석", "Finite Element Analysis")	464
3:	1114890	("유한요소법", "FEM")	349
4:	1115074	("유한요소해석", "Finite element analysis")	237
5:	1114872	("유한요소법", "")	217
6:	1115014	("유한요소해석", "")	186
7:	1114926	("유한요소법", "Finite element method")	174
8:	1115105	("유한요소해석", "finite element analysis")	140
9:	1114954	("유한요소법", "finite element method")	112
10:	1115049	("유한요소해석", "FEM")	106
273:	1115108	("유한요소해석", "finite element method stress a	1
274:	1115111	("유한요소해석", "finite element study")	1
275:	1115110	("유한요소해석", "finite element method(FEM)")	1
276:	1115113	("유한요소해석", "finite elementary analysis")	1
277:	1115112	("유한요소해석", "finite element")	1
278:	1115115	("유한요소해석", "finite-element analysis")	1
279:	1115114	("유한요소해석", "finite elements method analys	1
280:	1115117	("유한요소해석", "genetic algorithm")	1
281:	1115116	("유한요소해석", "frame ratio")	1
282:	1115118	("유한요소해석", "meshing")	1

Fig. 6. Sample cluster detail for “Finite Element Method”

4. 결론

본 논문에서는 역인덱스 기반 상향식 군집화 기법을 제안하고 이를 이용하여 대규모 학술 핵심어 분석을 실행하였다. 상향식 군집화 과정에서 역인덱스를 생성하여 공통의 자질을 포함하고 있는 후보 군집들을 검색하여 병합의 횟수를 줄여 효율을 높였으며 문자열 q-gram을 이용하여 띄어쓰기, 오탈자 등의 언어적인 문제를 처리하였다. 이 방법을 대용량의 키워드 데이터에 적용하여 대략 71만개의 다양한 주제에 대한 키워드 군집을 생성하였다. 생성된 키워드 군집을 이용하면 쉽게 문서 군집으로 확장할 수 있다.

추후 딥러닝을 이용한 키워드 임베딩(embedding)을 생성하여 적용하거나 도메인 분류를 적용한다면 더욱 효과적인 키워드 군집을 생성할 수 있을 것이다.

References

[1] O. Egozi, S. Markovitch, E. Gabrilovich, “Concept-Based Information Retrieval Using Explicit Semantic Analysis”, *ACM Transactions on Information Systems*, Vol.29, No.2, pp.1-34, 2011.
DOI: https://dx.doi.org/10.1145/1961209.1961211

[2] L. Li, R. Zhou, D. Huang, “Two-phase biomedical named entity recognition using CRFs”, *Computational Biology and Chemistry*, Vol.33, No.4, pp.334-338, 2009.
DOI: https://dx.doi.org/10.1016/j.compbiolchem.2009.07.004

[3] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, Y. Chi, “Deep Keyphrase Generation”, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.582-592, 2017.
DOI: https://dx.doi.org/10.18653/v1/P17-1054

[4] Y. G. Kim, J. H. Suh, S. C. Park, “Visualization of patent analysis for emerging technology”, *Expert Systems with Applications*, Vol.34, No.3, pp.1804-1812, 2008.
DOI: https://dx.doi.org/10.1016/j.eswa.2007.01.033

[5] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, Y. Chi, “Deep Keyphrase Generation”, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.582-592, 2017.
DOI: https://dx.doi.org/10.18653/v1/P17-1054

[6] J. Liu, J. Shang, C. Wang, X. Ren, J. Han, “Mining Quality Phrases from Massive Text Corpora”, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, pp.1729-1744, 2015.
DOI: https://dx.doi.org/10.1145/2723372.2751523

[7] C. C. Aggarwal, C. A. Zhai, Survey of Text Clustering Algorithms. In Mining Text Data, pp.77-128, Springer US, 2012.

[8] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.

[9] P. Willett, “The Porter stemming algorithm: then and now”, *Program*, Vol.40, No.3, pp.219-223, 2006.
DOI: https://dx.doi.org/10.1108/00330330610681295

[10] M. Sahami, T. D. Heilman, “A web-based kernel function for measuring the similarity of short text snippets”, *Proceedings of the 15th international conference on World Wide Web - WWW '06*, pp.377-386, 2006.
DOI: https://dx.doi.org/10.1145/1135777.1135834

[11] S. Tan, Y. Wang, G. Wu, “Adapting centroid classifier for document categorization”, *Expert Systems with Applications*, Vol.38, No.8, pp.10264-10273, 2011.
DOI: https://dx.doi.org/10.1016/j.eswa.2011.02.114

[12] T. Hasegawa, S. Sekine, R. Grishman, “Discovering relations among named entities from large corpora”, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pp.415-422, 2004.
DOI: https://dx.doi.org/10.3115/1218955.1219008

오 흥 선(Heung-Seon Oh) [정회원]

- 2006년 2월 : 한국항공대학교 컴퓨터공학과 (공학학사)
- 2009년 2월 : 한국과학기술원 전선학과 (공학석사)
- 2014년 2월 : 한국과학기술원 전산학과 (공학박사)
- 2013년 12월 ~ 2018년 2월 : 한국과학기술정보연구원 선임연구원
- 2018년 3월 ~ 현재 : 한국기술교육대학교 컴퓨터공학부 교수

<관심분야>
인공지능, 기계학습, 정보검색, 자연어처리

정 유 철(Yuchul Jung) [정회원]

- 2003년 2월 : 아주대학교 정보 및 컴퓨터공학과 (공학학사)
- 2005년 2월 : 한국과학기술원 정보통신공학과 (공학석사)
- 2011년 2월 : 한국과학기술원 전산학과 (공학박사)
- 2009년 1월 ~ 2013년 7월 : 한국전자통신연구원 선임연구원
- 2013년 8월 ~ 2017년 8월 : 한국과학기술정보연구원 선임연구원
- 2017년 8월 ~ 현재 : 금오공과대학교 컴퓨터공학과 교수

<관심분야>
인공지능, 기계학습, 정보검색, 자연어처리, 지식베이스 구축