KIPS Trans. Softw. and Data Eng.
Vol.7, No.12 pp.477~484 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2018.7.12.477

A Study on Research Paper Classification Using Keyword Clustering

Yun-Soo Lee† · They Pheaktra†† · JongHyuk Lee††† · Joon-Min Gil††††

ABSTRACT

Due to the advancement of computer and information technologies, numerous papers have been published. As new research fields continue to be created, users have a lot of trouble finding and categorizing their interesting papers. In order to alleviate users' this difficulty, this paper presents a method of grouping similar papers and clustering them. The presented method extracts primary keywords from the abstracts of each paper by using TF-IDF. Based on TF-IDF values extracted using K-means clustering algorithm, our method clusters papers to the ones that have similar contents. To demonstrate the practicality of the proposed method, we use paper data in FGCS journal as actual data. Based on these data, we derive the number of clusters using Elbow scheme and show clustering performance using Silhouette scheme.

Keywords : Classification Papers, K-Means Clustering, TF-IDF, Map-Reduce

키워드 군집화를 이용한 연구 논문 분류에 관한 연구

이 윤 수† · They Pheaktra†† · 이 종 혁††† · 길 준 민††††

요 약

컴퓨터 기술의 발전으로 힘입어 수많은 논문이 출판되고 있으며, 새로운 분야들도 계속 생기면서 사용자들은 방대한 논문들 중 자신이 필요로 하는 논문을 검색하거나 분류하기에 많은 어려움을 겪고 있다. 사용자의 이러한 어려움을 완화하기 위해 본 논문에서는 유사 내용의 논문을 분류하고 이를 군집화하는 방법을 제한한다. 본 논문의 제안 방법은 TF-IDF를 이용하여 각 논문의 초록으로부터 주요 주제어를 추출하고, K-평균 클러스터링 알고리즘을 이용하여 추출한 TF-IDF 값을 근거로 논문들을 유사 내용의 논문으로 군집화한다. 제안 방법의 실효성을 검증하기 위해 실제 데이터인 FGCS 저널의 논문 데이터를 사용하였으며, 엘보우 기법을 적용하여 클러스터 개수를 도출하고 실루엣 기법을 이용하여 클러스터링 성능을 검증하였다.

키워드 : 논문 분류, K-평균 군집화, 단어 빈도-역문서 빈도, 맵리듀스

1. 서 론

컴퓨터 기술의 발전과 더불어 관련된 수많은 논문이 출판되고 있으며, 이러한 상황에서 대부분의 사용자들은 키워드 검색이나 분야별 주요 저널을 검색하여 자신이 필요로 하는 학술논문을 찾는다. 따라서 사용자들은 방대한 학술논문들 중 자신에게 필요로 하는 학술논문을 검색하거나 분류하기에 많은 어려움을 겪고 있다.

한편, 학술논문의 여러 구성 중에 초록 부분은 논문의 요지를 축약하여 설명한 것으로 논문의 구성요소 중 가장 중요한 부분이다. 일반적으로 초록은 논문의 전체 내용 중 핵심 정보를 짧으면서도 흥미를 유발하고 이해하기 쉽게 작성된다. 초록 내용에는 연구의 특징, 의미, 결과를 간결하게 포함하고 있어야 하며 초록을 통해 논문의 전체 내용을 빠르게 파악할 수 있도록 작성된다. 따라서 사용자는 논문을 검색할 때 논문의 전체 내용을 대략적으로 빠르게 파악하기 위해서 논문제목과 함께 초록을 먼저 읽어보고 해당 논문의 주제를 파악하고 분류한다.

기술의 진보와 융합으로 새로운 연구 분야가 빠르게 탄생하고 있으며 이에 다양한 분야의 수많은 논문들이 출판되고 있다. 사용자들이 효율적으로 논문을 검색하고 이용하도록 하기 위해서는 수많은 논문에 대한 각각의 주제를 추출하고, 유사 내용의 논문을 분류할 필요가 있다. 따라서 본 논문에서는

※ 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A3B03933370).
※ 이 논문은 2018년도 한국정보처리학회 춘계학술발표대회에서 '키워드 추출과 군집화 기반의 논문 분류 시스템의 설계 및 구현'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 대구가톨릭대학교 컴퓨터정보통신공학과 석사
†† 준 회 원 : 대구가톨릭대학교 컴퓨터정보통신공학과 석사과정
††† 정 회 원 : 대구가톨릭대학교 빅데이터공학과 조교수
†††† 종신회원 : 대구가톨릭대학교 IT공학부 교수
Manuscript Received : July 6, 2018
Accepted : August 9, 2018
* Corresponding Author : Joon-Min Gil(jmgil@cu.ac.kr)

초록의 핵심을 파악하여 유사한 논문을 분류하는 방법을 제안한다. 이를 위해 논문초록에서 키워드 추출을 위해 TF-IDF (Term Frequency-Inverse Document Frequency) 기법[1-5]과 키워드 분류를 위해 K-평균 클러스터링(K-means clustering) 기법[6-9]을 사용한다. 또한 대표 키워드들의 사전을 구축하여 다른 단어 같은 의미인 동의이음어들도 논문의 TF-IDF 값을 추출할 때 함께 이용한다.

한편, 대용량의 논문에서 논문의 키워드를 효과적으로 추출·처리하기 위해 본 논문에서는 높은 확장성을 갖고 대용량 데이터를 신속하고 안정적으로 처리할 수 있는 분산 병렬 처리 프레임워크인 하둡 분산 파일시스템(HDFS; Hadoop Distributed File System)[10]에서 TF-IDF의 개별 연산을 맵리듀스(Map-Reduce) 프로그래밍 모델을 활용하여 구현하고 실행한다. 아울러, 제안 방법의 적응성과 실용성 검증을 위해, FGCS (Future Generation Computer Systems) 저널[11]의 2015~2017년도에 발행된 논문 제목과 초록을 실험 데이터로 활용한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 제안 방법에서 사용된 핵심 기술과 기존 연구에 대해서 설명한다. 3장에서는 논문 초록 내용 분석을 통한 논문 분류 시스템을 보여준다. 4장에서는 실험 환경과 결과를 나타낸다. 본 논문의 결론은 5장에서 맺는다.

2. 관련 연구

본 논문의 주제와 관련된 문서 분류(text classification)는 문서의 내용을 바탕으로 일정한 기준에 의하여 문서의 종류별로 나누어 구분하는 것을 말하며, 하나 이상의 범주에 문서를 할당하는 것으로 여러 분야에서 많이 사용되고 있다. 문서 분류에 관한 대표적 연구 분야는 다음과 같다.

- 뉴스기사 분류: 대부분의 뉴스 서비스들은 발행되는 기사의 수가 매우 대량인 경우가 많다. 이러한 뉴스기사를 수동으로 분류하고 검증하는 것은 불가능하기 때문에 대량의 뉴스기사를 자동화하여 분류하면 효과적이다[12-13].
- 오피니언 마이닝: 특정한 주제에 대한 사람들의 의견, 감성, 평가 등을 분석하는 것으로, 감정 분석이라고도 한다. 웹 페이지, 온라인 뉴스, 리뷰, 웹 블로그 및 SNS 등 여론이나 사람들의 의견을 분석하여 유용한 정보로 재가공 한 후 사용자에게 제공된다[14].
- 이메일 분류 및 스팸 필터링: 특정 주제의 내용으로 메일을 분류하거나, 스팸 메일을 분류하는 등에 활용되고 있다[15].

한편, 문서 분류 중 본 연구와 직접적으로 관련된 동일 주제의 논문 분류에 관한 연구로 Bravo-Alcobendas 등[16]은 바이오 메디컬 과학 논문 분류를 위해 Non-negative matrix factorization(NMF)로 차원 축소를 하여 특징을 추출한 뒤 K-평균 군집화 알고리즘으로 군집화 하는 알고리즘을 제안했다. 일반적으로 고차원 데이터는 클러스터링이 쉽지 않으며, 논문 집합에서 단어를 벡터로 변환하여 행렬화하면 2000~3000개의 키워드로 이루어진 고차원 데이터가 된다. 해당 연구에서는 고차원 데이터를 차원 감소 기법 중 하나인 NMF 알고리즘을 이용하여 차원 축소를 하여 클러스터링의 오차를 감소시켜 군집을 정확하게 하였다. 하지만 차원을 축소하여 대표적인 주제어들로만 군집화가 되면 여러 키워드가 아닌 대표 키워드만 사용하게 되어 세부적인 분류는 어려워진다.

Taheriyan 등[17]은 상호 인용관계를 이용한 그래프 기반 논문 분류를 제안했다. 제안한 방법은 논문들 사이의 링크 연결이 많을수록 더 좋은 결과를 얻을 수 있지만, 텍스트의 내용 분석이 아닌 인용관계를 이용한 논문 분류로 연관된 저자들의 논문들로 분류는 가능하지만, 세부적인 주제들과 관련 없는 논문들로 분류될 수 있다.

Nanba 등[18]은 연구 목적과 배경에서 키워드를 추출하고, 논문 전체 내용과 제목 초록, 서지 정보를 이용하여 논문을 분류하였다. 해당 논문의 방법 중 논문의 전체 내용을 이용하여 분류하는 것보다 초록만 이용하여 분류하는 것이 계산량과 속도 측면에서 효율적일 수 있다.

Nguyen 등[19]은 Bag-of-Word 기법과 K-최근접 이웃(K-Nearest Neighbors: KNN) 알고리즘을 이용하여 논문을 분류하는 방법을 제안하였다. 논문의 내용에서 주제를 추출하는 것에 더하여 제목에서 주제를 추출하여 사용하였다. 제목에서 추출된 주제는 논문의 내용을 비교적 잘 반영할 수 있다는 장점이 있다. 하지만 KNN 알고리즘의 경우 데이터의 양이 많을수록 분석속도가 저하될 수 있으므로, 전체 텍스트를 이용해 키워드를 추출하는 점은 상당한 계산량이 필요하다. 또한 KNN 알고리즘은 지도학습 알고리즘이며 레이블이 있는 데이터를 사용해야 하기 때문에 일정부분의 논문 레이블을 미리 지정해 주어야 한다는 단점이 있다

이상과 같이 기술한 기존 연구와 달리, 본 연구에서 제안하는 방법은 사용자가 입력한 주제어와 초록에서 주제어를 추출하여 사용한다. 사용자가 입력한 주제어는 사용자의 의도를 잘 나타내고 있기 때문에, 각 논문저자의 의도를 효과적으로 파악할 수 있다. 또한 TF-IDF 기법을 적용하여 많은 키워드의 중요도를 판단하여 각각의 논문에서 중요한 주제를 추출하고, 세부적인 주제들도 함께 판단할 수 있다. 한편, 기존의 논문 분류는 단일 시스템을 사용하였기 때문에 발표 수가 증가하는 학술 논문의 저장과 분석에 어려움을 겪는다. 따라서 본 연구에서는 빅데이터의 처리와 분석에 적합한 HDFS와 Map-Reduce를 이용하여 대량의 학술 논문을 저장, 분석하여 효율적으로 분류한다.

3. 연구 방법

3.1 연구 모델

이 절에서는 TF-IDF 모델과 하둡 맵리듀스 프레임워크에 기반하여 논문들을 주제별 유사 그룹으로 분류할 수 있는 방법에 대해서 설명한다. Fig. 1은 본 논문에서 제안하는 논문 초록 내용 분석을 통한 주제 분류 시스템의 전체적인 흐름을 나타내며, 다음과 같은 단계로 수행된다.

첫 번째, 데이터 수집 기간을 선정하여 해당 기간에 게재된 논문을 자동으로 수집한다.

두 번째, 수집된 각 논문의 키워드를 기반으로 키워드 사전을 구축한다. 전체 논문의 총 키워드 수가 많으므로, 전체 키워드 중 빈도수가 높은 상위 N개(Top-N)의 키워드를 사용한다.

세 번째, 각 논문의 초록에서 각 단어가 출현한 횟수를 계산한다.

네 번째, 키워드 사전에 정의된 각 단어들이 전체 논문의 초록에 얼마나 많이 나오는지를 계산한다.

다섯 번째, 키워드 사전에 정의된 각각의 단어가 전체 논문의 초록 중에서 몇 개 논문의 초록에 출현하는지를 계산한다.

여섯 번째, 네 번째와 다섯 번째 단계에서 구한 결과 값을 이용하여 각 논문의 키워드별 TF-IDF 값을 계산한다.

일곱 번째, TF-IDF 결과 값을 K-평균 클러스터링 알고리즘을 이용하여 유사 주제의 논문들로 군집화한다.

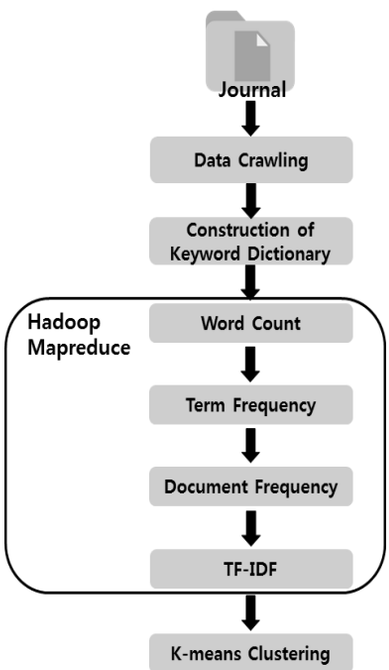


Fig. 1. System Flow

3.2 데이터 크롤링

Fig. 1의 데이터 크롤링(data crawling) 과정에서 정보를 수집하는 크롤러(crawler)는 자동으로 웹 문서를 수집하기 위한 프로그램으로 수집리스트 정보를 읽어 해당되는 조건의 논문을 수집한다. 수집한 논문 데이터는 대용량의 빅데이터의 특징을 가지고 있으므로, 확장성이 높은 HDFS에 저장되고, 맵리듀스 방식에 의해 TF-IDF 값을 계산하는데 활용된다.

3.3 키워드 사전 구축

Fig. 1의 키워드 사전 구축(construction of keyword dictionary) 단계에서는 데이터 크롤링 단계에서 추출한 키워드에 기반하여 키워드 사전을 구축한다. 본 논문에서는 유사한 의미의 키워드들을 하나의 대표 키워드로 하여 총 1394개의 대표 키워드를 선정하였다. 그러나 총 1394개의 대표 키워드 모두를 논문 분류 시 활용한다면, 많은 계산 시간을 필요로 한다. 따라서 본 논문에서는 계산 시간의 절감을 위해 대표 키워드 중 빈도수가 높은 상위 10개, 20개, 30개의 키워드만을 이용한다. Table 1은 상위 10~30개의 키워드를 보여준다.

Table 1. Top 10~30 Keywords

Top 30 (21~30)					
Top 20 (11~20)					
Top 10 (1~10)					
	Keywords		Keywords		Keywords
1	Cloud Computing	11	Map-Reduce	21	Game Theory
2	Internet of Things	12	Semantic Web	22	Data Mining
3	Big Data	13	Energy Efficiency	23	High Performance Computing
4	Security	14	Virtualization	24	Provenance
5	Scientific Workflow	15	Clustering	25	Performance Evaluation
6	Scheduling	16	Smart City	26	Machine Learning
7	Resource Management	17	Task Assignment	27	Mobile Cloud
8	Cloud Storage	18	QoS	28	Cloud Security
9	Privacy	19	Hadoop	29	Distributed System
10	Cloud	20	Distributed Computing	30	Wireless Sensor Networks

3.4 Word Count 계산

Fig. 1의 Word Count 계산 단계는 각 논문의 초록을 공백 구분자로 하여 단어를 구분한 뒤 모든 단어가 출현한 총 횟수를 계산한다. TF(Term Frequency) 값을 구하는 과정에서 논문의 크기에 따라 TF 값의 불균형을 방지하기 위해서 Word Count 값을 계산하여 문서의 길이로 이용한다.

Algorithm 1은 수집된 초록 데이터에서 모든 단어의 개수를 카운트하는 Word Count 계산의 맵리듀스 알고리즘을 나타낸 것이다.

Algorithm 1. Word Count Map-Reduce Algorithm

- Map:
  Input(Input file line offset, line content)
  output(DocName, 1)
  while(matcher.find()) {
  context.write(newText(DocName), 1)
  }
- Reduce:
  output(DocName, wc)
  wc = sum counts of for DocName

DocName: Paper Title
wc: Length of Paper Abstract

3.5 Term Frequency(TF) 계산

Fig. 1의 TF 계산 단계는 3.3절에서 정의한 키워드 사전과 각 논문의 초록을 비교하여 키워드 사전에 정의한 단어가 초록에 존재할 경우 해당 단어의 출현 횟수를 카운트한다. 다음은 TF의 계산식을 나타낸다.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

여기서, $n_{i,j}$는 단어 t_i가 문서 d_j에 출현한 횟수를 나타내며, $\sum_k n_{k,j}$는 논문 d_j에서 모든 단어가 출현한 횟수(논문의 길이)를 나타낸다($i = 1, \cdots, K,\ j = 1, \cdots, D$). K와 D는 키워드의 총 개수와 논문의 총 개수를 각각 나타낸다.

Algorithm 2는 수집된 초록 데이터의 논문별 키워드 출현 횟수를 계산하는 TF 계산의 맵리듀스 알고리즘을 나타낸다.

Algorithm 2. TF Map-Reduce Algorithm

- Map:
  Input(Input file line offset, line content)
  output((DocName+keyword), 1)
  if (word.equals(Keyword_dic))
  context.write(new Text(Doc_Name + keyword),1)
- Reduce: output((DocName+keyword),n)
  n = sum counts for keyword in documents

DocName: Paper Title
keyword: Keyword
n : Number of Keyword Occurances
d: Number of Papers with keywords among paper sets
D: Total Numbe of Papers

3.6 Document Frequency (DF) 계산

TF가 한 문서에 출현하는 키워드의 빈도를 나타낸 것이라면 DF(Document Frequency)는 전체 논문 집합 중 몇 개의 논문에서 키워드가 출현하는지 나타낸 값이다. Fig. 1의 DF 계산 단계에서는 전체 논문에서 각 키워드가 출현한 논문 개수의 비율을 계산한다. 다음은 DF의 계산식을 나타낸다.

$$DF_{i,j} = \frac{|d_j \in D : t_i \in d_j|}{|D|} \quad (2)$$

여기서, $|D|$는 논문 집합의 전체 문서의 수를 나타내며, $|d_j \in D : t_i \in d_j|$는 논문 d_j가 논문 집합 D에 속하고, 단어 t_i가 출현하는 논문의 수를 나타낸다.

Algorithm 3은 각 키워드가 나타나는 논문의 개수를 구하는 DF 계산의 맵리듀스 알고리즘을 나타낸다.

Algorithm 3. DF Map-Reduce Algorithm

- Map:
  Input(Input file line offset, line content)
  output((DocName+keyword+n))
- Reduce:
  output((DocName+keyword+wc+n),d){
  d = count papers with keyword in DocName
  }

DocName: Paper Title
keyword: Keyword
wc: Length of Paper Abstract
n : Number of Keyword Occurances
d: Number of Papers with keywords among paper sets
D: Total Numbe of Papers

3.7 TF-IDF 값 계산

DF 값이 크다는 것은 해당 단어가 많은 문서에서 등장하므로 중요한 단어가 아니라는 것을 의미한다. 그래서 DF 값의 역수인 IDF (Inverse Document Frequency)를 계산에 사용한다. 즉, IDF 값이 높으면 해당 단어가 다른 문서에 잘 등장하지 않는 특징이 있는 단어라는 것을 의미하므로 중요한 단어로 볼 수 있다. 이 절에서는 앞 단계에서 계산한 DF 값을 이용해 IDF 값을 계산한다. 다음은 IDF의 계산식을 나타낸다.

$$IDF_{t,D} = \log \frac{|D|}{|d_j \in D : t_i \in d_j|} \quad (3)$$

최종 TF-IDF 값은 Equation (1)과 (3)을 이용하여 다음과 같이 계산한다.

$$TFIDF = TF \times IDF \quad (4)$$

특정 문서에서 특정 키워드의 빈도가 높고, 전체 문서에서 해당 키워드를 포함하는 문서가 적을수록 TF-IDF 값은 높아진다. 이러한 원리를 이용하면 자주 출현하는 단어를 추출할 수 있으며, 결국 각 논문에 중요한 키워드가 어떠한 것인지 확인할 수 있다.

Algorithm 4는 전체 논문 집합의 단어 빈도-역문서 빈도를 구하는 TF-IDF 계산의 맵리듀스 알고리즘을 나타낸 것이다.

3.8 K-평균 클러스터링에 의한 논문 분류

TF-IDF 계산 단계의 결과인 TF-IDF 값은 각 논문의 특징을 나타내는 키워드의 중요도를 나타내고 있다. 그러므로 TF-IDF 값을 이용하여 논문을 분류하면, 키워드의 중요도에 따라 유사 논문들로 군집화를 할 수 있다.

본 논문에서 유사 논문의 군집화를 위한 군집화 알고리즘

Algorithm 4. TF-IDF Map-Reduce Algorithm

- Map:
  Input((DocName+keyword+wc+n),d)
  output(DocName+Keyword),TFIDF)
  D : total number of documents
  TF=n/wc
  IDF=log(D/1.0+d)
  TFIDF=TF*IDF

- Reduce:
  output(DocName+Keyword+TFIDF)

DocName: Paper Title
keyword: Keyword
wc: Length of Paper Abstract
n : Number of Keyword Occurances
d: Number of Papers with keywords among paper sets
D: Total Numbe of Papers
TFIDF: TFIDF Value

으로 K-평균 클러스터링 알고리즘을 사용한다. K-평균 클러스터링 알고리즘은 특정 영역을 대표하는 군집의 중심을 찾는 알고리즘으로, 각 논문과 클러스터 중심들 사이의 거리를 측정하여 유사도가 가장 높은 중심점에 논문을 할당한다. 같은 중심점에 할당된 논문들은 하나의 군집을 형성한다. Fig. 1의 논문 분류 단계에서 TF-IDF 값과 K-평균 클러스터링 알고리즘을 사용하여 유사 논문들의 군집화를 수행한다.

한편, K-평균 클러스터링 알고리즘은 사용자가 클러스터의 개수를 미리 지정해야 하는 단점이 있다. 따라서 본 논문에서는 엘보우 기법(Elbow scheme)[20]을 이용하여 적절한 클러스터 개수를 유도하고자 한다. 엘보우 기법은 최적의 클러스터 개수를 구하기 위한 알고리즘으로 알려져 있으며, 한 개의 클러스터를 추가했을 때, 추가하기 전보다 특정 범위 값을 넘어서는 더 좋은 결과가 나타나지 않으면 이전 클러스터의 개수를 최적의 클러스터 개수로 설정한다.

4. 실험 및 결과 분석

4.1 실험 환경

본 논문에서 제안하는 논문 초록 내용 분석을 통한 주제 분류를 위해 하둡 시스템을 데이터 처리를 위해 사용하였으며, 구체적으로 마스터 노드 서버 1대, 서브 노드 서버 1대, 데이터 노드 서버 4대로 시스템을 구성하였다. HDFS 및 맵리듀스 프레임워크(Hadoop-2.6.5 버전) 상에서 TF-IDF를 Java 언어를 이용하여 구현하였으며, Python Scikit-learn 라이브러리[21]를 활용하여 K-평균 클러스터링 알고리즘을 구현하였다. 한편, 실험에 사용한 논문 데이터는 FGCS 저널을 대상으로 3년(2015~2017) 동안 게재된 논문(543개)에서 수집한 제목, 초록, 키워드이다.

4.2 실험 결과

본 논문에서는 엘보우 기법을 이용하여 클러스터의 수를 2부터 100까지의 범위로 지정하여 결과값을 측정하였다. 측정 결과, 상위 10개의 키워드를 이용한 군집화에서는 클러스터의 개수가 24개, 상위 20개의 키워드는 33개, 상위 30개 키워드는 37개가 가장 적절한 클러스터의 개수로 계산되었다. Fig. 2, Fig. 3, Fig. 4는 각각 빈도수 상위 10개, 20개, 30개 키워드를 이용한 클러스터링에 대해서 클러스터 수의 변화에 따른 엘보우 값을 나타낸 그래프이다. 이들 그림에서 볼 수 있듯이, 엘보우 값이 급격하게 낮아지다가 완만하게 낮아지는 지점에서 최적 클러스터의 개수가 추출되었다.

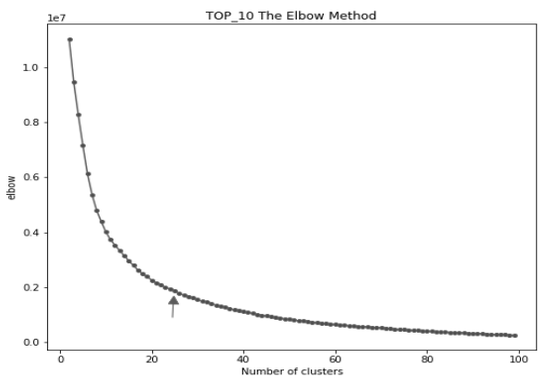


Fig. 2. Elbow Graph for Top 10 Keywords

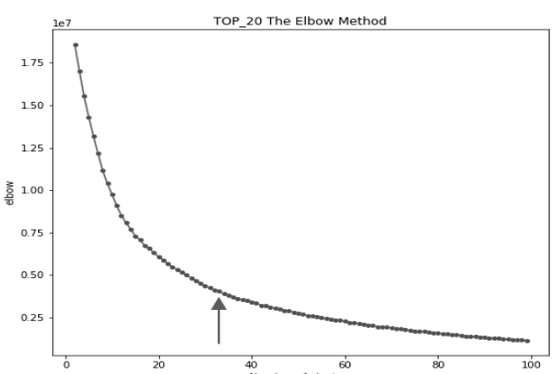


Fig. 3. Elbow Graph for Top 20 Keywords

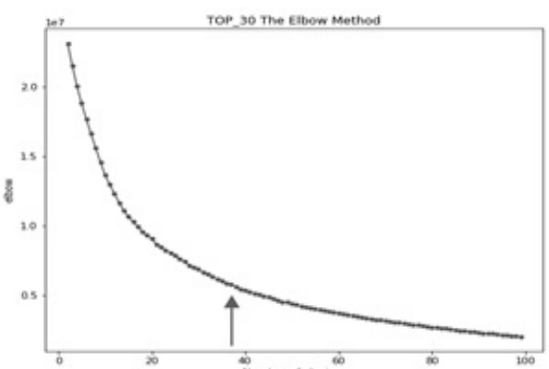


Fig. 4. Elbow Graph for Top 30 Keywords

본 논문에서는 최종으로 산출된 클러스터를 평가하기 위한 평가 척도로 내부 평가 기법 중 하나인 실루엣 기법(Silhouette scheme)[22-23]을 이용한다. 실루엣 기법에서 실루엣 값은 하나의 클러스터 내의 데이터와 다른 클러스터 내의 데이터 간의 거리를 비교하여 유사 정도를 나타내는 척도이다. 클러스터 내부 데이터 간의 거리는 가까울수록 유리하며, 다른 클러스터의 데이터와의 거리는 멀수록 유리하다. 이러한 의미에서 실루엣 값은 -1에서 1 사이의 값을 가지며 1에 가까울수록 좋은 결과를 나타낸다. 일반적으로 실루엣 값이 0.5 보다 크면 군집 결과가 타당하다고 판단한다[23]. Figs. 5~7은 빈도수 상위 10~30개 키워드를 이용하여 군집화한 결과 각 클러스터의 실루엣 값을 나타낸 그래프이다. 이들 그림에서 수직 점선은 평균 실루엣 값(0.71)을 나타낸다. 따라서 상위 10개 키워드를 이용한 군집화의 실루엣 값이 0.5 이상이므로, 유사 주제의 논문들의 군집이 타당함을 알 수 있다. 상위 20개와 30개 키워드를 이용한 군집화에서도 유사한 결과가 도출되었다. 한편, 이들 그래프에서 가장 많은 부분을 차지하는 실루엣 값은 본 논문에서 사용한 빈도수 상위 10~30개의 키워드에 해당하는 키워드가 없는 학술논문들로서 키워드 사전에 정의하지 않은 키워드로 구성된 군집이다. 본 논문에서는 이 군집이 논문 분류에 사용되지 않았으므로 이를 예외 군집으로 간주하였다.

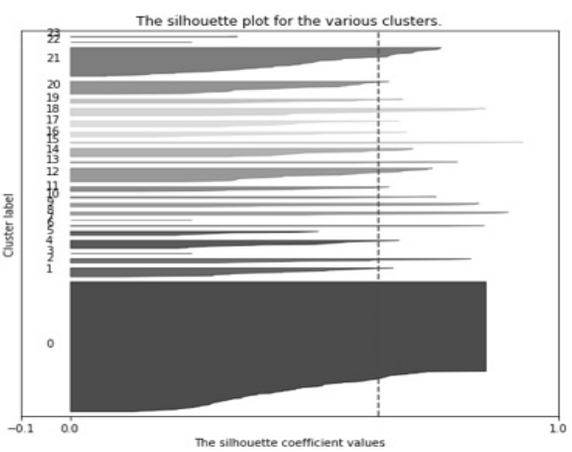


Fig. 5. Silhouette Graph for Top 10 Keywords

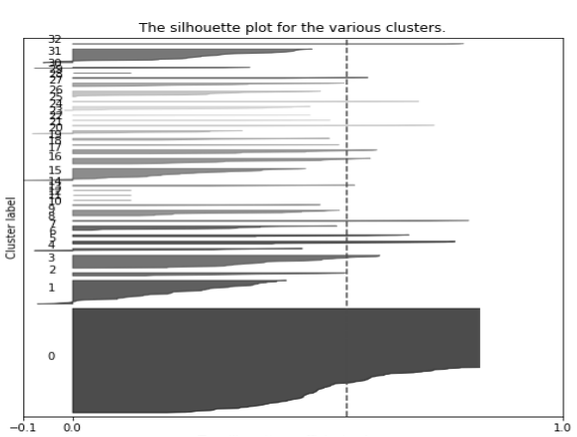


Fig. 6. Silhouette Graph for Top 20 Keywords

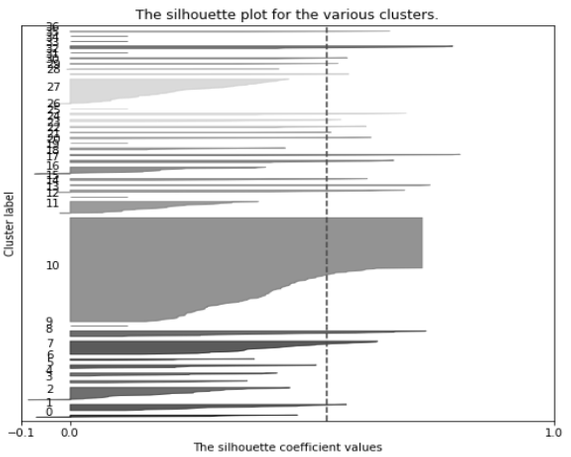


Fig. 7. Silhouette Graph for Top 30 Keywords

Table 2. Clustering Results Using the Top 10 Keywords

Cluster	Keyword	Document
1	Big Data Cloud	Uploading multiply deferrable big data to the cloud platform using cost-effective online algorithms
		Expanded cloud plumes hiding Big Data ecosystem
2	IoT Privacy	Evolving privacy from sensors to the Internet of Things
		L2P2 A location-label based approach for privacy preserving in LBS
		A comprehensive approach to privacy in the cloud-based Internet of Things
3	IoT Security Privacy	A risk analysis of a smart home automation system
		CLAPP A self constructing feature clustering approach for anomaly detection
		Midgar study of communications security among smart objects using a platform of heterogeneous devices for the Internet of Things

Table 2는 상위 10개의 키워드에 대해서 논문의 군집화 결과를 전체 클러스터 중 세 개의 클러스터만 나타낸 것이다. Table 2의 결과를 살펴보면, 1번 클러스터는 'BigData'와 'Cloud' 키워드를 2, 3 클러스터는 모두 'IoT' 키워드를 가지고 있지만, 2번 클러스터는 'Privacy' 키워드를, 3번 클러스터는 'Security'와 'Privacy' 키워드를 추가 키워드로 갖고 있다. 즉, 세 개의 클러스터는 키워드로 인해 서로 다른 클러스터로 분류되었음을 나타낸다. 한편, 상위 20개의 키워드를 이용하여 군집화한 결과, 1번 클러스터의 논문 중 "Uploading multiply deferrable big data to the cloud platform using cost-effective online algorithms" 논문에서 "Map-Reduce" 키워드가 추가되어 클러스터 결과도 기존 결과와 다르게 변경되었다. 따라서 해당되는 키워드가 많을수록 여러 주제에 연관된 클러스터에 할당되는 것을 확인할 수 있다.

5. 결 론

본 논문에서는 키워드 추출에 기반하여 유사 내용의 논문을 분류하고 이를 군집화하는 논문 분류 시스템을 설계 및 구현하였다. TF-IDF 기법을 이용하여 각 논문을 대표하는 키워드의 중요도를 계산하였고, TF-IDF 결과 값을 이용하여 K-평균 클러스터링 알고리즘으로 군집화를 수행하였다. 실제 출판된 FGCS 저널의 논문 데이터를 활용한 실험 결과는 높은 TF-IDF의 값을 갖는 키워드들이 각 논문의 주제를 잘 나타내고 있고, 군집화 결과도 연관된 주제의 논문끼리 비교적 잘 분류되었음을 확인하였다.

한편, 키워드 사전에 정의하지 않은 키워드들에 대해서는 하나의 군집을 형성하였고 이들 군집은 논문 분류에 사용되지 않는 예외가 발생하였다. 따라서 향후 연구로는 상위 빈도수 10~30개의 키워드가 아닌 전체 키워드를 이용하여 모든 키워드가 논문 분류에 활용되어 더욱 정밀하게 논문 분류가 될 수 있도록 연구를 진행할 계획이다.

References

[1] Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya, "Document clustering: TF-IDF approach," in *Proceedings of 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp.61-66, 2016.

[2] Lukáš Havrlant and Vladik Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *International Journal of General Systems*, Vol.46, No.1, pp.27-36, 2017.

[3] Bruno Trstenjak, Sasa Mikac, and Dzenana Donko, "KNN with TF-IDF based Framework for Text Categorization," *Procedia Engineering*, Vol.69, pp.1356-1364, 2014.

[4] Akiko Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, Vol.39, Iss.1, pp.45-65, Jan. 2003.

[5] Shereen Albitar, Sébastien Fournier, and Bernard Espinasse, "An effective TF/IDF-based text-to-text semantic similarity measure for text classification," in *Proceedings of International Conference on Web Information Systems Engineering (WISE 2014)*, pp.105-114, 2014.

[6] Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, Vol.100, Iss.3, pp.767-786, Sept. 2014.

[7] Rakesh Chandra Balabantaray, Chandrali Sarma, and Monica Jha. "Document clustering using K-means and K-medoids," *International Journal of Knowledge Based Computer Systems*, Vol.1, Iss.1, 2015.

[8] Rajeev Srivastava and Himanshu Gupta, "K-means based document clustering with automatic "K" selection and cluster refinement," *International Journal of Computer Science and Mobile Applications*, Vol.2, Iss.5, pp.7-13, 2014.

[9] Charu C. Aggarwal and Chandan K. Reddy, *Data clustering: algorithms and applications*, CRC press., 2013.

[10] N. K. Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework," *Journal of Big Data*, Vol.2, No.6, pp.1-18, Dec. 2015.

[11] FGCS Journal [Internet], https://www.journals.elsevier.com/future-generation-computer-systems

[12] Kil-Hong Joo, Eun-Young Shin, Joo-Il Lee, and Won-Suk Lee, "Hierarchical Automatic Classification of News Articles based on Association Rules," *Journal of Korean Multimedia Society*, Vol.14, No.6, pp.730-741, 2011.

[13] H. Cho and J.-S. Lee, "Data-driven feature word selection for clustering online news comments," in *Proceedings of 2016 International Conference on Big Data and Smart Computing (BigComp)*, pp.494-497, Jan. 2016.

[14] Anand Mahendran, Anjali Duraiswamy, Amulya Reddy, and Clayton Gonsalves, "Opinion Mining for text classification," *International Journal of Scientific Engineering and Technology*, Vol.2, Iss.6, pp.589-594, Jun. 2013.

[15] Izzat Alsmadi and Ikdam Alhami, "Clustering and classification of email contents," *Journal of King Saud University-Computer and Information Sciences*, Vol.27, Iss.1, pp.46-57, Jan. 2015.

[16] Bravo-Alcobendas and C. O. S. Sorzano, "Clustering of biomedical scientific papers," in *Proceedings of 2009 IEEE International Symposium on Intelligent Signal Processing*, pp.205-209, Aug. 2009.

[17] Mohsen Taheriyan, "Subject classification of research papers based on interrelationships analysis," in *Proceedings of the 2011 Workshop on Knowledge Discovery, Modeling and Simulation*, pp.39-44, Aug. 2011.

[18] Hidetsugu Nanba, Noriko Kando, and Manabu Okumura, "Classification of research papers using citation links and citation types: towards automatic review article generation," in *Proceedings of 11th ASIS SIG/CR Classification Research Workshop*, pp.117-134, 2011.

[19] Thien Hai Nguyen and Kiyoaki Shirai. "Text classification of technical papers based on text segmentation," *Lecture Notes in Computer Science*, Vol.7934, pp.278-284, 2013.

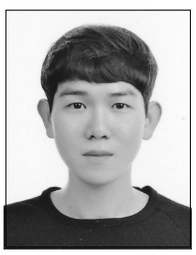
[20] Trupti M. Kodinariya and Prashant R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advanced. Researches in Computer Science and Management Studies*, Vol.1, Iss.6, pp.90-95, Nov. 2013.

[21] Scikit-Learn [Internet], http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[22] Gilberto V. Oliveira, Felipe P. Coutinho, Ricardo Campello,

and Murilo C. Naldi, “Improving k-means through distributed scalable metaheuristics,” *Neurocomputing*, Vol.246, No.12, pp.45-57, Jul. 2017.

[23] Peter J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, Vol.20, pp.53-65, Nov. 1987.



이 윤 수

https://orcid.org/0000-0003-0937-8459
e-mail : ys89@cu.ac.kr
2016년 대구대학교 통신공학전공(학사)
2018년 대구가톨릭대학교 컴퓨터정보통신공학과(석사)
관심분야 : 클라우드컴퓨팅, 빅데이터, 분산시스템, 인공지능

They Pheaktra

https://orcid.org/0000-0003-3131-6086
e-mail : pheaktra97@gmail.com
2016년 대구가톨릭대학교 컴퓨터공학전공(학사)
2017년~현 재 대구가톨릭대학교 컴퓨터정보통신공학과 석사과정
관심분야 : 클라우드컴퓨팅, 빅데이터, IoT, 분산시스템

이 종 혁

https://orcid.org/0000-0002-8163-9388
e-mail : jonghyuk@cu.ac.kr
2004년 고려대학교 컴퓨터교육과(학사)
2006년 고려대학교 컴퓨터교육학과(석사)
2011년 고려대학교 컴퓨터교육학과(박사)
2011년~2012년 University of Houston, Post-Doc.
2012년~2017년 삼성전자 클라우드플랫폼그룹 책임연구원
2017년~현 재 대구가톨릭대학교 빅데이터공학과 조교수
관심분야 : 빅데이터, 클라우드컴퓨팅, 분산시스템, 인공지능

길 준 민

https://orcid.org/0000-0001-6774-8476
e-mail : jmgil@cu.ac.kr
1994년 고려대학교 전산학과(학사)
1996년 고려대학교 전산학과(석사)
2000년 고려대학교 전산과학과(박사)
2001년~2002년 University of Illinois at Chicago, Post-Doc.
2002년~2006년 KISTI 슈퍼컴퓨팅센터 선임연구원
2006년~2010년 대구가톨릭대학교 컴퓨터교육과 교수
2010년~현 재 대구가톨릭대학교 IT공학부 교수
관심분야 : 클라우드컴퓨팅, 빅데이터, 분산시스템, 인공지능