J. Soc. Korea Ind. Syst. Eng Vol. 43, No. 2 : 25-32, June 2020
https://doi.org/10.11627/jkise.2020.43.2.025
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Technology Clustering Using Textual Information of Reference Titles in Scientific Paper

Inchae Park* · Songhee Kim** · Byungun Yoon**†

*Division of Smart Management Engineering, Hansung University
**Department of Industrial & Systems Engineering, Dongguk University

과학기술 논문의 참고문헌 텍스트 정보를 활용한 기술의 군집화

박인채* · 김송희** · 윤병운**†

*한성대학교 스마트경영공학부
**동국대학교 산업시스템공학과

Data on patent and scientific paper is considered as a useful information source for analyzing technological information and has been widely utilized. Technology big data is analyzed in various ways to identify the latest technological trends and predict future promising technologies. Clustering is one of the ways to discover new features by creating groups from technology big data. Patent includes refined bibliographic information such as patent classification code whereas scientific paper does not have appropriate bibliographic information for clustering. This research proposes a new approach for clustering data of scientific paper by utilizing reference titles in each scientific paper. In this approach, the reference titles are considered as textual information because each reference consists of the title of the paper that represents the core content of the paper. We collected the scientific paper data, extracted the title of the reference, and conducted clustering by measuring the text-based similarity. The results from the proposed approach are compared with the results using existing methodologies that one is the approach utilizing textual information from titles and abstracts and the other one is a citation-based approach. The suggested approach in this paper shows statistically significant difference compared to the existing approaches and it shows better clustering performance. The proposed approach will be considered as a useful method for clustering scientific papers.

Keywords : Technology Clustering, Technology Data Analysis, Scientific Paper, Reference Data

1. 서 론

대용량 데이터가 축적되고 분석이 가능해짐에 따라서 산업분야 전반에서 산업 데이터를 활용한 데이터 기반의 의사결정을 수행하려는 다양한 시도들이 이루어지고 있다. 각 기술분야에서의 기술동향 파악 및 향후 유망 기술 예측 등도 기술 데이터 기반의 기술기획 및 기술지능 활동이 이루어지고 있다. 특허, 과학기술논문, 연구노트, 기술 보고서 등의 다양한 형태의 문서들이 주로 기술 분석을 위한 데이터로서 활용되고 있다. 특허와 논문은 과학 및 기술의 지식이 공개된 자료로서 데이터베이스에 축적되고 관리되기 때문에 공중에 공개되지 않고 일관된 데이터베이스로 관리되지 않는 기술 보고서 및 연구보고서에 비해서 정량분석에 용이하다.

특허는 발명자, 출원인, 출원일, 등록일, 특허분류코드 등으로 구성된 서지정보와 발명의 명칭, 초록, 청구항 등

Received 25 February 2020; Finally Revised 30 March 2020; Accepted 3 April 2020
† Corresponding Author : postman3@dongguk.edu

특허의 내용이 기술된 텍스트 정보로 구성되어 있다. 논문도 마찬가지로 저자, 저자의 소속기관, 출판일, 출판 저널 등으로 구성된 서지정보와 초록 및 본문으로 구성되어 과학기술 지식을 표현하는 텍스트 정보로 구성되어 있다. 특허의 경우에는 특허를 기술분야별로 체계적으로 분류하여 심사의 효율성을 높이기 위하여 각국 특허청에서 심사과정에서 부여한 IPC(International Patent Classification)와 CPC(Cooperative Patent Classification)와 같이 전 세계적으로 공통적으로 사용하는 특허분류코드가 있다. 반면에, 논문의 경우에는 과학기술논문 데이터베이스에서 제공하는 분류가 존재하지만 제공하는 데이터베이스마다 분류가 상이하기 때문에 기술적 내용을 구분하기 위한 표준화된 분류체계가 존재하지 않는다.

일반적으로 과학기술 정보의 정량적 분석을 통한 과학기술 최신 연구동향 파악 및 유망 기술 분야의 예측을 위하여 특허와 논문 개별 데이터를 군집화(clustering)하여 기술분야를 정의하는 과정을 수행한다. 과학기술논문은 주로 과학적 지식의 기반이 되는 기초연구에 가까운 연구결과를 발표하는 매체로 비교적 상용화에 가까운 응용연구와 관련된 결과인 특허와는 과학기술분야의 동향과 시사점을 제공하는 정보로써 차이가 존재한다[9, 11, 15]. 따라서, 특허의 특징과 구별하여 기초과학기술 분야의 정량적 정보분석을 체계적으로 수행하기 위하여 과학기술 논문의 기술군집화를 위한 적절한 정보 소스가 제안될 필요가 있다.

기존 연구에서는 특허 및 과학기술 논문의 클러스터링을 위하여 특허 및 논문의 인용정보를 활용하여 클러스터링을 수행하거나[2] 텍스트 정보를 활용하여 클러스터링을 수행하였다[12, 13]. 특허와 논문은 공통적으로 참고문헌을 포함하고 있으며, 기존의 연구에서는 동시인용(co-citation) 또는 서지결합(bibliographic coupling)관계를 기반으로 문서간의 관계를 정의하는 방법을 취했지만, 문서간의 동시인용 빈도 또는 공유하고 있는 레퍼런스의 빈도수가 많지 않을 경우 유사도 값이 매우 낮게 나와서 관계를 정의하기 어려운 측면이 있다[1]. 특허와 논문을 군집화 하는 다른 방법으로는 텍스트 정보를 기반으로 문서를 벡터화하여 문서간의 유사도를 측정하여 군집화를 수행하는 방법이 있다[6]. 기존의 특허 및 논문의 텍스트정보를 활용한 군집화에는 특허와 논문의 내용을 기반으로 키워드의 특징을 추출하는 방법으로써 특허 및 논문의 전체문서(full-text)를 대상으로 하거나, 주로 제목(title)과 요약 또는 초록(abstract)으로부터 텍스트마이닝을 수행하여 문서간의 유사도를 측정하고 군집화를 수행한다[4, 6, 7].

특허와 논문의 제목과 요약 부분의 기술문서를 나타내는 중요한 핵심부분임에도 불구하고, 작성자가 사용하는 언어가 차이가 있을 수 있다. 유사도 관점에서 보다 나은 정보를 제안하기 위하여 본 연구에서는 논문의 참고문헌 리스트를 텍스트정보로서 활용한다. 논문의 전체 내용을 요약하여 제시하는 부분이 초록이고, 핵심적으로 나타낸 것을 제목이라고 할 때, 논문에서 인용하고 있는 해당 논문과 관련된 참고문헌의 제목들도 해당 논문을 나타내는 중요한 텍스트 정보로써 역할을 할 수 있을 것이다는 가설이 본 연구의 출발점이다. 특허문서에서는 관련 선행기술 문헌이 특허인 경우에는 특허의 번호를 기재하게 되어 있고, 비특허문헌의 경우 논문 등의 형태와 유사하게 기재하게 되어 있지만, 대부분의 관련 선행기술 문헌이 특허이기 때문에 텍스트 정보로써 활용하기는 어려운 측면이 있다. 게다가, 특허의 제목은 대부분 자세하게 표현되지 않는 측면이 있다. 반면에, 논문의 참고문헌은 논문, 학술대회 프로시딩 논문, 저서 등 관련된 기술 지식을 포함하며 제목이 그 참고문헌을 가장 대표할 수 있는 축약적인 언어로 기술되어 있는 특징이 있다. 하나의 학술 논문은 많은 참고문헌을 인용하고 있기 때문에 관련 참고문헌의 리스트를 제목만 나열한 하나의 텍스트 정보로써 활용할 수 있을 것이다.

본 논문은 참고문헌의 제목을 텍스트 정보로써 활용하여 군집화를 수행하는 방법론을 제안하는 것을 목표로 한다. 제안한 방법론의 성능은 기존의 방법론과 비교함으로써 타당성을 제시하고자 한다. 기존의 방법론과 비교를 위하여, 논문의 제목 및 초록의 텍스트 정보를 활용하여 군집화를 수행한 결과와 논문의 인용관계 기반의 유사도정보를 활용하여 군집화를 수행한 결과를 비교하고제안한 접근방법의 타당성을 제시한다. 본 논문의 제2장에는 연구를 수행하는데 관련된 방법론의 이론적 배경에 대해서 설명한다. 제 3장에서는 연구의 개념과 전체적인 연구 프로세스를 단계별로 상세하게 설명한다. 제 4장에서는 연구수행을 통해 분석 결과를 제시하고 결과를 통해 얻을 수 있는 시사점을 제시한다. 마지막으로 제 5장에서 연구내용의 요약과 학계의 공헌하는 점, 한계점 및 향후 연구 주제에 대해서 제시한다.

2. 이론적 배경

2.1 Girvan-Newman 클러스터링

본 연구에서 논문의 유사도 기반의 클러스터링은 Girvan-Newman(G-N) 클러스터링을 통해서 수행된다. G-N 클러스터링은 군집화 대상을 노드(node)로 하고 대상 간의 관계를 엣지(edge)로 하는 네트워크로 구성하고 엣지 매개중심성(Edge Betweenness centrality)과 모듈성(modularity)이라는 개념을 활용하여 최적의 군집을 설정하는 방법론

이다[9]. 엣지 매개 중심성은 기존의 네트워크분석 이론의 노드의 매개 중심성을 엣지에 적용한 개념이다. 노드의 매개 중심성은 노드들 간의 최단 경로를 고려하여 각 노드들 사이에 놓일 수 있는 정도로 정의한다. 노드 매개 중심성이 높은 노드를 정보 흐름의 중요한 역할을 하는 노드라고 생각한다고 할 때, 엣지의 매개 중심성은 역시 정보흐름의 중요한 역할을 하는 엣지로 생각할 수 있다.

Girvan-Newman 클러스터링은 전체의 네트워크에서 정보의 흐름이 가장 많다고 할 수 있는 즉, 엣지 매개 중심성 값이 큰 엣지를 끊어 냄으로써 군집을 나누는 방식을 반복해서 수행하는 개념이라고 생각할 수 있다. 엣지를 끊어내고 새로 생성된 네트워크에서 엣지 매개 중심성을 계산하고 다시 가장 높은 엣지를 끊어내어 네트워크를 나누며 군집화를 하는 과정을 반복하면서 모듈성(modularity) 값이 최대값이 될 때까지 네트워크를 나누며 최종 군집을 생성하는 과정이다. 모듈성 계산을 통해서 최적 군집의 수를 자동적으로 결정할 수 있는 장점 때문에 군집화에 많이 활용되는 방법론이다[8]. G-N 클러스터링은 주로 네트워크에서의 군집을 도출하는 데 활용되고 있으며, 대표적으로 특허 네트워크를 작성하여 유망기술을 탐색하거나[8], 사회연결망 분석을 통해 건축물 공간을 설계하는 연구에 적용되었다[5].

2.2 서지결합법

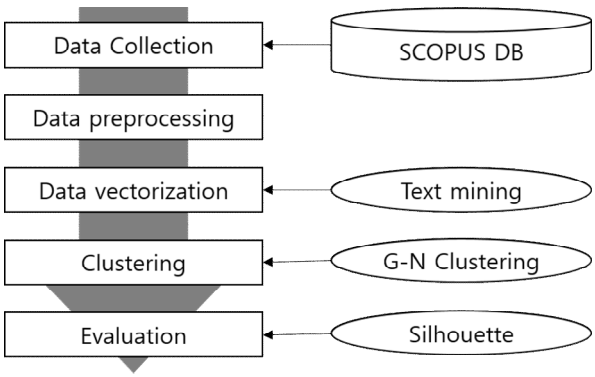
본 연구에서 텍스트마이닝을 통한 군집화 결과와 비교하여 인용기반의 문서 유사도를 기반으로 한 군집화를 위하여 서지결합법(bibliographic coupling)을 활용한다. 서지결합법은 서지결합 관계를 계산하는 것으로 서지결합 관계는 참고문헌을 공유하고 있는 정도를 나타낸다. 즉, 두 개의 논문의 유사도를 비교할 때 공통의 참고문헌이 많을수록 강한 서지결합 관계를 가지고 있다고 할 수 있으며, 논문 간의 유사도가 높다고 할 수 있다. 서지결합관계는 동시인용(co-citation) 관계와 대비되는 개념으로 동시인용 관계는 두 개의 논문이 다른 하나의 논문으로부터 동시에 인용(co-cited)되면 유사하다고 판단하는 개념이다. 동시인용 관계는 미래의 출판되는 논문에 의해서 동시에 인용되는 관계를 판단해야 하기 때문에 향후 미래에 논문이 인용관계가 생길 경우 논문 간의 유사도 값이 변할 수 있는 반면에 서지결합 관계는 각 논문의 인용(citing)하여 참고문헌으로써 공유하는 관계를 계산하는 것이기 때문에 값이 변하지 않는 장점이 있다. 본 연구에서는 정규화 된 서지결합강도를 계산하기 위하여 식 (1)을 활용하여 계산하였다[3]. 식 (1)의 N_A, N_B는 논문 A, B가 각각 인용한 참고문헌 수이고, N_{AB}는 논문 A와 B가 공통으로 인용한 참고문헌 수를 나타낸다.

$$N_{norm} = \frac{N_{AB}}{\sqrt{N_A N_B}} \quad (1)$$

최근에는 서지결합 강도를 향상시키기 위해 TF-IDF와 결합한 새로운 방법론을 제시하거나[14], 저탄소 기술의 프론티어를 탐색하기 위해 서지결합 방법론 적용하는 연구가 수행되었다[16].

3. 연구 프레임워크

본 연구는 과학기술 논문 데이터베이스로부터 논문 데이터를 수집하여 논문의 참고문헌의 제목을 추출하는 과정을 수행하고 참고문헌의 제목을 텍스트로써 활용하여 텍스트마이닝을 거쳐 문서 유사도 기반의 클러스터링을 수행한다. 연구의 자세한 진행과정은 <Figure 1>과 같고, 본 장에서 단계별로 설명한다.



<Figure 1> Research Process

3.1 Step 1 : 데이터 수집과 데이터 전처리

과학기술 논문 데이터는 전 세계 우수 학술논문인용색인 데이터베이스 중에 하나인 Scopus 데이터베이스를 활용하였다. 자동차 산업의 연료전지(fuel cell) 기술분야로 한정하여 자동차분야 도메인 전문가로부터 검토 받은 키워드로 검색식을 작성하고 2010년부터 2019년 사이 10년 동안 자동차 완성차 15개사로부터 Scopus 저널에 등재된 총 7,008건의 논문 및 학술대회 논문 데이터를 수집하였다. 논문의 참고문헌에는 저자, 참고문헌의 제목, 저널명, 출판연도, 저널의 권호 등이 포함되어 있기 때문에 참고문헌의 제목을 텍스트 정보로써 활용하기 위하여 제목만을 추출하기 위한 전처리 과정이 필요하다. 저널 마다 요구하는 참고문헌 스타일이 상이하고 특히, 학술대회 논문의 경우에는 일관되지 않은 포맷으로 정보가 존재하는 경우가 많기 때문에 저널 포맷을 일정한 유형을 나누고, 예외를 처리

하는 등의 과정을 거쳐서 각 논문의 참고문헌리스트에서 제목만 추출하여 텍스트화하는 과정을 수행하였다. 도출된 논문의 참고문헌의 제목 리스트의 텍스트로부터 단어의 어근 및 어미의 추출, 대소문자의 처리, 대명사, 숫자, 기호 등에 해당하는 stop-word 처리 등 R의 tm패키지를 활용하여 텍스트마이닝 수행을 위한 전처리를 수행하였다.

3.2 Step 2 : 데이터의 구조화

논문의 군집화를 수행하기 위하여 논문 간의 유사도를 나타내는 문서 간 유사도 매트릭스(Document-to-Document matrix : D2D matrix)를 생성이 필요하다. 이를 위해 첫 번째로 논문을 벡터화한다. 논문의 벡터화를 위하여, 텍스트 정보로부터 텍스트 마이닝을 수행하여 도출된 대표 키워드로부터 논문을 키워드 벡터의 형태로 표현한 문서-단어 매트릭스(Document-Term matrix : DTM)를 구성한다. 두번째로, 논문 간의 유사도를 측정하여 D2D 매트릭스를 생성한다. 논문 간의 유사도를 측정하기 위하여 본 연구에서는 일반적으로 활용되는 두가지 접근 방법인 코사인(Cosine)과 자카드(Jaccard) 유사도를 활용하였다. 코사인 유사도로 측정하는 경우에는 빈도수 상위 50건 이상의 키워드를 대표 키워드로써 활용하여 DTM을 구성한 후 TF-IDF(Term Frequency-Inverse Document Frequency) 값을 활용하여 논문 간의 코사인 유사도를 측정했고, 자카드 유사도로 측정한 경우에는 빈도수 상위 50건의 키워드를 활용하여 구성한 DTM 내의 키워드의 빈도 값을 활용하여 자카드 유사도를 측정하였다.

3.3 Step 3 : 클러스터링

논문 간의 유사도 매트릭스를 기반으로 Girvan-Newman (G-N) 클러스터링 방법론을 활용하여 논문 간의 클러스터링을 수행한다. 본 연구에서는 데이터 시각화 도구인 Gephi를 활용하여 G-N클러스터링 및 데이터의 시각화를 수행하였다. G-N클러스터링은 엣지 매개 중심성(Edge Betweenness centrality)과 모듈성(modularity)를 활용하여 최적 군집을 설정하는 클러스터링 방법론으로 전체 네트워크를 각각의 분할된 네트워크로 군집화 하는데 용이하다. 논문 간의 유사도는 [0, 1]의 값으로 표현되어 있기 때문에 적절한 절단값(threshold)을 활용하여 논문 간의 관계를 0 또는 1로 표현하여 논문 간의 연결 관계를 표현한 네트워크를 기반으로 G-N클러스터링을 수행할 수 있다. 본 연구에서는 노드(논문)와 엣지(논문 간의 유사도 관계)의 수의 비율이 각각 1:20, 1:50, 1:00, 1:200인 경우를 절단값으로 선정하여 1과 0으로 연결관계를 표현하고 군집화를 수행하였다. 이 비율이 높아질수록 엣지의 수는 증가하여, 시각적으로는 복잡해질 수 있다.

3.4 Step 4 : 군집 결과의 평가

G-N 클러스터링을 기반으로 도출된 군집 결과를 평가하기 위하여 군집을 평가할 수 있는 기존의 다양한 군집 타당성 지표(clustering validity index) 중에 성능이 우수한 Silhouette 지표(S(i))를 활용하였다[1]. S(i)를 계산하는 수식은 (2)와 같다. (2)의 b(i)는 i번째 개체와 다른 군집에 속한 요소들 간 거리들의 평균을 군집마다 각각 구한 뒤, 이 가운데 가장 작은 값을 취한 것이고, a(i)는 i번째 개체와 같은 군집에 속한 요소들 간 거리들의 평균을 의미한다. S(i)값이 가장 큰 경우에 군집화가 우수한 경우로 판단하여 최적 군집수로 판단한다. 본 논문에서는 각각 코사인과 자카드 유사도를 기반으로 구성된 문서 간 유사도 매트릭스 내의 값이 [0, 1] 값으로 구성되었기 때문에 논문 간의 거리는 1-(유사도) 값으로 활용하였다.

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \quad (2)$$

예를 들면, G-N 클러스터링 결과로 각 논문의 클러스터가 할당 된 상황에서 논문 i가 속하지 않은 군집 중 가장 가까운 군집 내의 논문들과 논문 i 간의 거리의 평균 값이 23.78 (b(i))이고, 논문 i가 속한 군집 내의 다른 논문들 간의 거리의 평균 값이 4.63 (a(i))인 경우, Silhouette값은 (23.78-4.63) / 23.78로 0.81이 계산된다.

군집화 성능의 비교를 위하여 논문의 제목 및 초록의 텍스트로부터 Step 1~Step 3의 과정을 동일하게 수행하여 도출한 군집화 결과의 Silhouette 값과 t-test를 통해서 비교하였다. 추가적으로 논문의 인용 관계를 기반으로 유사 관계를 표현하는 방법인 서지결합법(bibliographic coupling : BC)을 활용하여 논문의 군집화를 수행하고 Silhouette 값을 계산하여 참고문헌 논문 제목 기반 군집화, 논문의 제목 및 초록 기반 군집화, 논문의 서지결합법 기반 군집화 결과를 분산분석(ANOVA)을 통해서 비교하였다.

4. 결 과

4.1 군집화 결과

각 유형별로 클러스터링을 수행한 결과와 군집화 성능을 나타내는 Silhouette값은 <Table 1>과 같다. 요약하면, 각 유형은 첫째, 참고문헌의 제목을 텍스트로써 활용하여 텍스트마이닝 후 논문 간의 유사도를 코사인 유사도로 측정한 결과(Ref-title & Cosine similarity), 둘째, 논문의 제목 및 요약부분을 텍스트로써 활용하여 텍스트마이닝 후 논문간의 유사도를 코사인 유사도로 측정한 결과(Title &

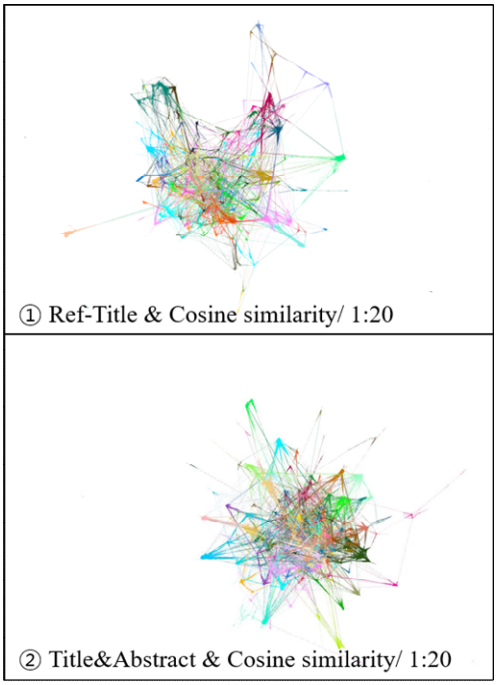
<Table 1> Results of Clusters

Type	Node : Edge Ratio	No. of Clusters	No. of Clusters including over 10 papers	Silhouette score
① Ref-Title & Cosine similarity	1:20	241	55	0.28328
	1:50	25	25	0.01418
	1:100	15	15	0.01028
	1:200	14	14	0.00876
② Title & Abstract & Cosine similarity	1:20	986	49	-0.77597
	1:50	741	27	-0.80707
	1:100	735	21	-0.80579
	1:200	729	15	-0.80579
③ BC	1:20	1199	33	-0.06413
	1:50	1197	33	-0.06522
	1:100	986	33	-0.06524
	1:200	1201	35	-0.06411
④ Ref-Title & Jaccard similarity	1:20	200	33	-0.17816
	1:50	20	15	-0.25097
	1:100	16	19	-0.01095
	1:200	12	12	-0.16472
⑤ Title & Abstract & Jaccard similairy	1:20	1796	54	-0.44129
	1:50	736	22	-0.38679
	1:100	726	12	-0.40794
	1:200	807	30	-0.48467

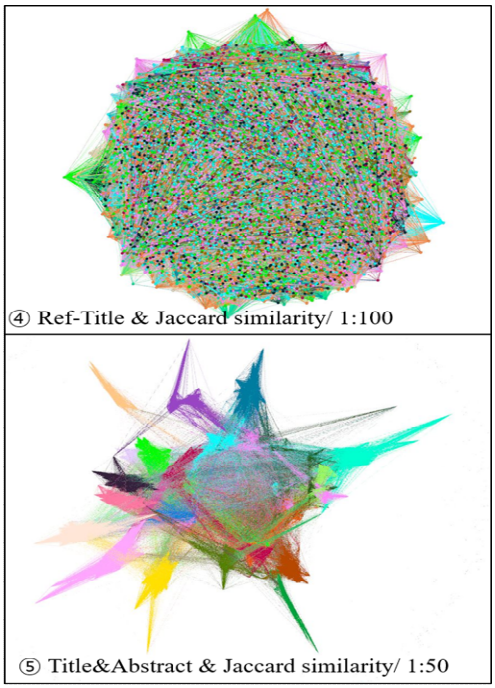
Abstract & Cosine similarity), 셋째, 논문의 서지결합법 관계를 기반으로 유사도를 측정한 결과(BC), 넷째, 참고문헌의 제목을 텍스트로써 활용하여 텍스트마이닝 후 논문간의 유사도를 자카드 유사도로 측정한 결과(Ref-title & Jaccard similarity), 마지막으로, 논문의 제목 및 요약부분을 텍스트로써 활용하여 텍스트마이닝 후 논문간의 유사도를 자카드 유사도로 측정한 결과(Title & Abstract & Jaccard similarity)로 총 다섯 가지 유형의 결과가 존재한다. 클러스터의 결과를 나타낸 <Table 1>의 음영처리 된 부분은 각 유형별로 가장 높은 Silhouette값을 나타내는 것이다. 전체 7,008건의 논문 데이터가 각 노드의 역할을 하며 군집화 결과에 따라 동일 군집 내의 논문 간의 거리와 다른 군집 내의 논문 간의 거리를 코사인 및 자카드 유사도를 기반으로 측정하여 Silhouette값을 계산하였다.

4.2 군집화 성능의 비교

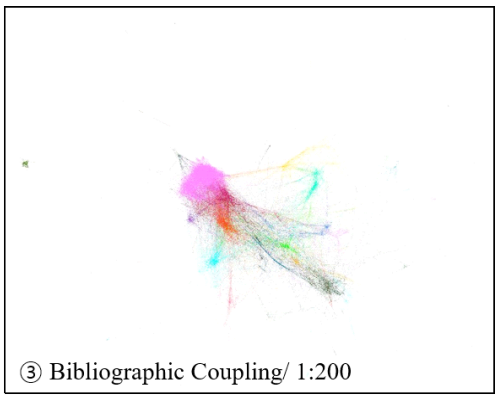
논문 간의 유사도의 측정을 코사인 유사도를 활용하여 측정한 경우, 유형 1과 2에서 가장 높은Silhouette값을 나타낸 G-N클러스터링을 수행한 결과는 <Figure 2>이고, 유사도의 측정을 자카드 유사도를 활용한 경우, 유형 4와 5의 가장 높은 Silhouette 값을 나타낸 G-N클러스터링 결과를 시각화한 것이 <Figure 3>이다. 그리고 서지결합법을 기반으로 유사도를 측정하고 G-N클러스터링의 결과를 시각화한 것이 <Figure 4>이다. Gephi를 활용한 클러스터 결과는 같은 군집에 속한 논문들은 같은 색상으로 구분되어 시각화 된다. 군집내 논문 간의 거리가 가깝고 군집 외 논문간의 거리가 멀게 측정 될수록 시각화 결과의 그림도 명확하게 구분되어 표현된다.



<Figure 2> Visualized Results of Type 1 and 2



<Figure 3> Visualized Results of Type 4 and 5



<Figure 4> Visualized Results of Type 3

코사인과 자카드 유사도 기반으로 데이터를 구성하여 클러스터링을 수행한 Silhouette값의 차이를 알아보기 위하여 t-test를 수행한 결과는 <Table 2>와 같다. 코사인 유사도를 활용한 결과로 Levene의 등분산 검정을 통해 유의확률 0.038로 등분산을 가정하지 않은 결과를 활용하였고, 자카드 유사도를 활용한 결과로 Levene의 등분산 검정을 통해 유의확률 0.283으로 등분산을 가정한 결과를 활용하였다. 코사인 유사도를 기반으로 데이터를 구성하여 클러스터링을 수행한 결과 군집화 성능을 나타내는 Silhouette 값이 평균적으로 참고문헌(Ref)의 제목을 텍스트로 활용한 값이 논문의 제목 및 초록(TiAb)을 텍스트로 활용한 값보다 높으며 유의미한 차이를 나타내었다. 자카드 유사도를 기반으로 데이터를 구성한 경우에도 Silhouette값의 차이가 유의미한 결과를 나타냈고, 참고문헌의 제목을 텍스트로 활용한 유형이 논문의 제목 및 초록을 텍스트로써 활용한 유형보다 높은 평균 값을 나타내었다. 결과적으로 두 가지 경우에서 논문의 참고문헌의 제목들이 기존에 많이 활용되어 오던 논문의 제목 및 초록의 텍스트 정보를 활용한 것보다 더 우수한 클러스터 성능을 나타내었다.

참고문헌의 제목을 텍스트로써 활용하거나 논문의 제목 및 초록을 텍스트로써 활용하는 접근 방법의 차이점을 서지결합법을 기반으로 클러스터링을 한 결과와 비교하기 위하여 분산분석을 수행하였고 그 결과는 <Table 3>과 같다. 코사인 유사도를 활용한 결과 3개의 집단 중 어느 하나가 차이가 있다는 가설이 유의한 결과가 나타났다 (p-value < 0.05). Levene의 등분산 검정 결과 유의확률이 0.01로 등분산 가정이 만족하지 않기 때문에, Dunnett T3 사후분석을 수행하여 결과를 확인하였을 때, 논문의 제목 및 초록을 활용한 결과가 참고문헌의 제목과 서지결합관계를 활용한 결과와 유의미한 차이가 있다는 결과를 얻었다. 자카드 유사도를 활용한 결과 역시 3개의 집단 중 어느 하나가 차이가 있다는 가설이 유의한 결과가 나타났다 (p-value < 0.05). Levene의 등분산 검정 결과 유의확률이 0.70으로 등분산 가정이 만족하기 때문에, Duncan 사후분석을 수행하여 결과를 확인하였을 때, 마찬가지로 논문의 제목 및 초록을 활용한 결과가 참고문헌의 제목과 서지결합관계를 활용한 결과와 유의미한 차이가 있다는 결과를 얻었다.

<Table 2> Results of t-test between Silhouette of Ref. and Title & Abstract

Sim	Mean		S.D.		t	Sig.
	Ref	TiAb	Ref	TiAb		
Cos	0.791	-0.798	0.136	0.015	12.81	0.001*
Jac	-0.151	-0.430	0.100	0.042	5.093	0.002*

<Table 3> Results of ANOVA among Silhouette of Ref., Title & Abstract, and Bibliographic Coupling

Sim	Type	Mean	S.D.	F	Sig.
Cos	Ref a	0.791	0.1361	141.792	0.000*
	BC a	-0.064	0.0006		
	TiAb b	-0.798	0.1513		
Jac	Ref a	-0.151	0.1008	36.478	0.000*
	BC a	-0.064	0.0006		
	TiAb b	-0.430	0.0427		

4.3 결과의 시사점과 활용

본 연구에서 제안하는 참고문헌의 제목들을 하나의 논문과 관련된 텍스트 정보로써 활용하여 문서의 군집화 소스로 활용하는 방안이 기존의 방법과 비교하여 코사인 유사도와 자카드 유사도를 활용하였을 때 모두 통계적으로 유의한 군집화 성능 차이를 보이는 결과를 얻었다. 과학기술 논문의 군집화를 통해서 과학기술의 테마를 정의하기 위하여 활용할 수 있는 적절한 방법으로 고려해 볼 수 있을 것이다. 하지만, 논문의 참고문헌은 저널 또는 학술대회마다 요구하는 참고문헌 스타일이 다르기 때문에 이것을 정규화하는 과정의 자동화가 필요하다. 본 연구에서는 일정한 유형을 구분하고 예외 등을 처리하는 방법으로 논문의 제목을 도출하는 과정을 수행했지만, 논문의 서지정보를 제공하는 DB에서 정제된 포맷의 데이터가 제공된다면 제안된 방법론을 활용하기 더욱 용이할 것으로 생각된다.

군집화 결과의 클러스터 수의 구성을 보면 클러스터수가 과도하게 많거나 적은 경우가 관찰된다. 클러스터의 수가 증가하여 Silhouette값이 높아진다고 하더라도, 군집화 결과를 확인할 때 너무 많은 군집이 생긴다면 활용하기 적절하지 않은 문제가 발생할 수 있다. 일반적으로

Silhouette값이 0.5 이상이면 군집화 결과가 적절하다고 평가될 수 있지만, 제시된 <Table 1>의 결과의 따르면 Silhouette값이 그에 미치지 못하는 값을 나타낸다. 일반적으로 구조화된 데이터(structured data) 보다는 비구조화된 데이터(unstructured data)의 군집화는 노이즈가 많이 포함되어 Silhouette 값이 낮게 산출될 수 있다. 또한, 본 연구는 다양한 클러스터링 방법을 상대적으로 비교하여, 성능이 우수한 클러스터링 결과를 활용하기 위한 것이므로, 0.5 이상이라는 기준은 적용하지 않았다. 본 연구에서는 노드와 엣지 비율을 조정하여 군집화를 수행하였으며, 이를 더 상세하게 나누어 조정하면 높은 Silhouette값을 얻을 수 있을 것이다. 하지만, 하나의 논문으로 구성된 클러스터가 생성되는 등의 데이터의 과적합 현상이 발생할 수 있다. 따라서, 자동차 분야 기술 분석 도메인 전문가의 자문을 통해 <Table 1>과 같이 클러스터 내의 적어도 10개 이상의 논문을 포함한 클러스터만을 추려내어 활용하는 방안을 생각할 수 있다. 추가적으로 활용의 측면에서, 대표적인 클러스터 결과가 논문에서 제시되었지만 시각적으로 표현된 클러스터의 결과 자체로는 판단의 어려움이 있기 때문에 Silhouette 및 도메인 전문가의 정성 평가 등을 통한 클러스터의 구성 등을 종합하여 판단 할 필요가 있다. 또한 Silhouette 지표가 군집 타당성을 평가하는 다양한 지표 중에 우수한 지표로 기존 연구에서 평가되기 때문에 본 연구에서는 활용되었지만, Silhouette 지표만 단독으로 활용하기에는 무리가 있을 수 있다. 따라서, Silhouette 뿐만 아니라 다양한 지표를 부수적으로 활용하는 방안을 생각해 볼 수 있다.

군집화 결과의 ANOVA분석을 통해서 코사인과 자카드 유사도를 활용하였을 때 모두 세 집단(Ref, TiAb, BC) 중에 적어도 하나의 집단이 차이가 있다는 결과를 얻었으며, 사후분석을 통해 논문의 제목 및 초록을 활용한 군집화 결과는 참고문헌의 제목과 서지정보를 활용한 군집 결과와 차이가 나타남을 확인하였다. 이론적으로 참고문헌의 제목을 기반으로 텍스트마이닝을 수행하여 유사도를 측정하는 작업은 제목을 단어수준으로 나눠서 유사도를 측정하는 것이지만, 분석 과정에서 제목에서는 같은 단어가 추출되어 나오기 때문에 같은 참고문헌을 공유하고 있는 문헌을 유사하다고 판단하는 서지결합법의 개념을 포함하고 있는 것이라고 생각할 수 있다. 서지결합법의경우에는 같은 참고문헌을 포함하고 있는 경우에 유사하다고 계산하는 반면에 참고문헌의 제목들을 텍스트 정보로써 활용하여 유사도를 계산한다면, 같은 참고문헌은 아니더라도 유사한 키워드를 사용하고 있는 다른 참고문헌과의 관계도 측정할 수 있기 때문에 기존의 서지결합법에서 참고문헌이 없는 경우 0의 값이 나타나는 점을 보완할 수 있는 방법으로 생각할 수 있을 것이다.

5. 결 론

본 연구는 논문의 참고문헌 제목의 리스트를 텍스트 정보로써 활용하여 클러스터링을 수행하였으며 여러가지 학술적인 공헌점을 제시한다. 첫째, 기존의 논문의 전문(full-text) 또는 제목 및 초록으로부터 텍스트를 추출하여 클러스터링을 수행하는 접근에 비해서 새로운 접근으로 군집화를 위한 새로운 정보소스를 제공했다는데 의미를 가진다. 이를 통해서 과학기술 지식 데이터 중에 특허와 구별되는 특징으로써 기초과학기술 지식을 나타내는 논문데이터를 활용할 수 있는 다양한 활용이 향후에 가능할 것으로 기대할 수 있다. 둘째, 기존 접근방법과 비교하여 통계적으로 유의미한 차이를 나타내며 군집 타당성 지표의 비교를 통해서 우수한 결과를 나타내었다. 셋째, 논문의 참고문헌의 포맷이 기존의 논문 DB에서 일관되게 제시되지 않고 있기 때문에 논문의 제목만을 추출하여 분석을 수행한 시도는 의미를 가진다.

본 연구의 학술적인 공헌에도 불구하고 몇 가지 한계점이 존재한다. 우선, 본 연구에서는 논문을 기반으로 군집화를 수행하고 군집의 결과에 대한 컨텐츠 분석을 수행하지 못하고 정량적인 평가지표인 Silhoutte 지표를 사용하여 군집 타당성을 제시하였다. 향후 연구에서 클러스터링을 활용한 군집화 결과에 대해 정량적 지표뿐만 아니라 도메인 전문가의 참여를 통한 군집의 결과를 정성적으로 평가하여 발전시킬 필요가 있다. 다음으로, 논문의 참고문헌의 제목을 추출하는 과정에서 본 연구에서 발견된 참고문헌의 유형과 예외처리를 통해서 논문의 제목을 도출해 내는 과정을 수행했음에도 불구하고, 일부 논문에서는 참고문헌의 제목을 전혀 제공하지 않는 경우도 존재하였기 때문에 해당 논문은 결측치로 간주하고 분석을 진행했다. 논문의 참고문헌이 향후 중요한 텍스트 정보로써 인식이 되어, 논문 DB에서 표준화된 논문의 참고문헌 포맷을 제공한다면 관련된 연구가 활발하게 행해질 수 있을 것이다.

Acknowledgement

This research was financially supported by Hansung University for Inchae Park. Also, this work was supported by National Research Fund(NRF-2019R1A2C1085388) for Byungun Yoon.

References

[1] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PeRez, J. M., and Perona, I., An extensive comparative study of

cluster validity indices, *Pattern Recognition*, 2013, Vol. 46, No. 1, pp. 243-256.

[2] Fujita, K., Kajikawa, Y., Mori, J., and Sakata, I., Detecting research fronts using different types of weighted citation networks, *Journal of Engineering and Technology Management*, 2014, Vol. 32, pp. 129-146.

[3] Glanzel, W. and Czerwon, H.J., A new methodological approach to bibliographic coupling and its application to research-front and other core documents, *in ISSI'95, Proceedings of the fifth biennial international conference of the International Society for Scientometrics and Infometrics*, River Forest, Illinois, USA, 1995, pp. 167 -176.

[4] Jeong, Y. and Yoon, B., Development of patent roadmap based on technology roadmap by analyzing patterns of patent development, *Technovation*, 2015, Vol. 39, pp. 37-52.

[5] Jeon, Y. and Kim, Y., A study on improvement of the school space through socio-spatial network analysis, *Journal of the Architectural Institute of Korea-Planning*, 2019, Vol. 35, No. 5, pp. 21-30.

[6] Jeong, Y., Park, I., and Yoon, B., Identifying emerging Research and Business Development(R&BD) areas based on topic modeling and visualization with intellectual property right data, *Technological Forecasting and Social Change*, 2019, Vol. 146, pp. 655-672.

[7] Kim, S., Park, I., and Yoon, B., SAO2Vec : Development of an algorithm for embedding the subject-action-object(SAO) structure using Doc2Vec, *PLoS One*, 2020, Vol. 15, No. 2, e0227930.

[8] Lim, C., Yun, D., Park, I., Park, G., Koh, S., and Yoon, B., Exploring prospective research areas in UI/UX through the Analysis of Patents, *Korean Management Science Review*, 2015, Vol. 32, No. 4, pp. 1-18.

[9] Meyer, M., Does science push technology? patents citing scientific literature, *Research policy*, 2000, Vol. 29, No. 3, pp. 409-434.

[10] Newman, M.E. and Girvan, M., Finding and evaluating community structure in networks, *Physical Review E*, 2004, Vol. 69, No. 2, pp. 1-16.

[11] Park, I. and Yoon, B., Identifying promising research frontiers of pattern recognition through bibliometric analysis, *Sustainability*, 2018, Vol. 10, No. 11, pp. 1-32.

[12] Peters, H.P. and van Raan, A.F., Co-word-based science maps of chemical engineering, Part I : Representations by direct multidimensional scaling, *Research Policy*, 1993, Vol. 22, No. 1, pp. 23-45.

[13] Peters, H.P. and van Raan, A.F., Co-word-based science maps of chemical engineering, Part Ⅱ : Representations by combined clustering and multidimensional scaling, *Research Policy*, 1993, Vol. 22, No. 1, pp. 47-71.

[14] Shen, S., Zhu, D., Rousseau, R., Su, X., and Wang, D., A refined method for computing bibliographic coupling strengths, *Journal of Informetrics*, Vol. 13, No. 2, 2019, pp. 605-615.

[15] Shibata, N., Kajikawa, Y., and Sakata, I., Detecting potential technological fronts by comparing scientific papers and patents, *Foresight*, 2011, Vol. 13, No. 5, pp. 51-60.

[16] Wei, Y., Wang, J., Chen, T., Yu, B., and Liao, H., Frontiers of low-carbon technologies : Results from bibliographic coupling with sliding window, *Journal of Cleaner Production*, Vol. 190, No. 20, 2018, pp. 422-431.

ORCID

Inchae Park | http://orcid.org/0000-0002-5108-6786
Songhee Kim | http://orcid.org/0000-0002-9327-7863
Byungun Yoon | http://orcid.org/0000-0002-1110-4011