Journal of the Korean Data & Information Science Society
2016, **27**(4), 947–957

http://dx.doi.org/10.7465/jkdi.2016.27.4.947
한국데이터정보과학회지

KCI 등재 학술지의 분류를 위한 네트워크 군집화 방법의 비교

김진광[1] · 김소형[2] · 오창혁[3]

[13]영남대학교 통계학과 · [2]한국연구재단 학술기반진흥팀

접수 2016년 6월 24일, 수정 2016년 7월 22일, 게재확정 2016년 7월 22일

요 약

KCI는 국내 학술지 및 게재 논문과 인용에 대한 데이터베이스이며, 이를 이용하여 국내 학술지 간의 인용 관계를 파악할 수 있다. 현재 사용 중인 KCI의 학술지 분류는 각 학술지의 등재 신청 시 학술지 발간 주체가 선정한 분류로 인용 관계에 의한 분류가 아니다. 이로 인해 같은 분류에 속하는 학술지 사이의 인용관계가 없거나 낮은 현상이 발생하기도 하여 인용관계가 많은 학술지끼리 같이 묶여야 한다는 기준에 부합하지 않는 문제점이 발생하고 있다. 따라서 학술지 분류가 학술지 간의 인용 정도를 잘 대표하지 못하는 것으로 알려져 있다. 본 연구에서는 KCI에 등재된 학술지 분류와 KCI 인용망에 네트워크 군집화 알고리즘을 적용한 군집 결과를 토대로 어떠한 차이가 있는지 살펴보았다. 이를 위해 최근 논문에서 대표적으로 다뤄지는 네트워크 알고리즘을 제시하고, 인용관계에 따른 각 알고리즘의 군집 결과 차이를 비교하였다. 그 결과 '인포맵' 알고리즘이 기존 KCI 분류망과 모듈화 구조 측면에서 유사성이 가장 높은 것으로 나타났다.

주요용어: 군집, 네트워크 군집화, 모듈화, 인포맵, 학술지 분류, KCI.

1. 서론

최근 다양한 학문 분야에서 네트워크 분석을 이용한 군집화 연구 시도가 진행되고 있다 (Chun과 Leem, 2014; Won과 Choi, 2014). 상호 연관성에 의해 결합된 방대한 양의 데이터들이 네트워크 군집화를 통하여 유사성이 높은 몇 개의 군집으로 분할되면, 이들 세분화된 군집을 통해 새로운 유형의 정보를 추출하거나 군집 간에 소통을 이끌어 내는 핵심 역할의 개체를 파악할 수 있게 된다. 따라서 이런 네트워크 군집화를 이용하여 데이터가 가진 정보를 재해석하려는 새로운 시도들이 활발하게 진행되고 있다. 네트워크의 군집화는 두 개 이상의 군집을 갖도록 단순히 반복적으로 분할하여 노드를 할당하는 형태로부터 추출된 군집의 밀도를 최대화하는 방법, 통계적 추론에 의한 검출에 이르기까지 다양한 연구가 있어왔다. Kernighan과 Lin (1970)은 군집의 크기가 동일하도록 제한을 갖는 이분법을 이용한 군집방법을 제안하였다. Suaris와 Kedem (1988)은 Kernighan-Lin (1970)을 일반화하여 군집의 크기를 제한하지 않는 방법을 제시하였다. 또한 Barnes (1982)는 스펙트럼 속성을 기반으로 하는 이분법을 소개했다. Donath와 Hoffman (1973)은 인접 행렬의 고유벡터를 이용하는 군집화 방법을 발표하였다. Flake 등 (2000)은 웹에서 최대 흐름 정보량을 이용한 네트워크 군집 식별 방법을 제시하였다. Newman과 Girvan (2004)은 군집 간 연결선의 중심 역할을 측정하여 그 값이 가장 큰 연결선을 제거해

[1] (712-749) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 겸임교수.
[2] (305-754) 대전광역시 유성구 가정로 201, 한국연구재단 학술기반진흥팀, 연구원.
[3] 교신저자: (712-749) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 교수. E-mail: choh@yu.ac.kr

야 될 대상으로 선정하는 방법을 제안하였다. Radicchi 등 (2004)은 Watts와 Strogatz (1998)의 군집화 계수를 일반화한 연결선 결속계수라는 군집화를 위한 새로운 측도를 제시하였다.

2000년도 초기에는 네트워크의 모듈화 지수 값을 이용한 군집 추출 방법도 소개되었다 (Fortunato와 Barthelemy, 2007). 모듈화 지수를 극대화하기 위해 고안된 최초의 알고리즘으로는 Newman (2004)의 탐욕 (Greedy) 알고리즘이 있다. Clauset 등 (2004)의 탐욕 최적화 방법은 거대 네트워크 상에서 모듈화 지수 최대화를 추구하는 우수한 알고리즘으로 평가받고 있다. Danon 등 (2006)은 작은 군집을 선호하여 두 군집 합병으로 만들어진 모듈화 지수 변화량을 두 군집 중 어느 하나에 링크 가중치의 비율로써 정규화시키는 방법을 제시하였다. Arenas 등 (2007)은 가중 네트워크에서 군집 내 가중치 합을 하나의 노드로 대체하여 노드 크기를 줄이는 방법으로 네트워크를 재구성하고 모듈화 지수를 적용하는 방법을 제안하였다. 또 다른 탐욕 접근법으로는 가중치를 갖는 일반적인 그래프를 대상으로 Blondel 등 (2008)이 제시한 다단계 군집화 알고리즘이 있다.

2000년 후반에는 통계적 추론 방법을 이용한 군집 검출 알고리즘 방법이 소개되기도 하였다 (Newman과 Leicht, 2007; Copic 등, 2009). 또한, Rosvall과 Bergstrom (2008)은 가중 유향 네트워크에서 임의보행을 적용한 군집 구조식을 제시하고 네트워크 최적 분할을 이끄는 정보 흐름을 찾아 군집 구조를 탐색하는 방법을 제시하였다. 한편, Scott (2012)은 노드들 간의 연결 밀도가 높은 네트워크에서 군집을 파악하기 위해 사이클이라는 개념을 도입하였다.

트위터나 페이스북 등 모바일 앱을 통해 구축된 사회네트워크 환경 하에서 특정 주제로 연관된 사람들의 공동체를 추출하거나 군집을 통해 특정한 사회현상을 밝히려는 노력이 있었다 (Fortunato, 2010).

이러한 연구방법은 학술지들의 특성을 연구하는데 마찬가지로 적용될 수 있으며, 학술지 인용네트워크를 활용하여 학술지의 집단을 분류하고자 하는 연구 노력이 있었다. 대표적인 학술지 인용 데이터베이스인 Thomson Reuters 사의 SCI 나 Elsevier 사의 SCOPUS에 대하여 인용관계를 이용한 학술지 분류가 시도되었다 (Narin 등, 1972; Carpenter와 Narin, 1973; Leydesdorff, 2004; Zhang 등, 2010; Kim, 2008). 또한 국내 학술지 인용 데이터베이스인 KCI (Korea Citation Index)의 자연과학 분야를 대상으로 인용관계를 이용한 군집화 분석 연구도 있었다 (Kim 등, 2015). 이처럼 학술지 인용 데이터베이스를 대상으로 다양한 분야에서의 연구가 이루어지고 있으나, 아직까지 KCI의 논문 인용 정보망을 통한 군집 도출 연구는 아직 미미한 실정이다. 따라서 KCI가 제공하는 국내 학술지의 인용관계 네트워크를 활용하여 학술지 군집화에 대한 문제점들을 살펴볼 필요가 있다.

현재 KCI의 학술지 분류는 학술지의 등재 신청 시 발간 주체가 선정한 분류로 인용 관계에 의한 분류가 아니다. 따라서 같은 분류에 속하는 학술지 사이의 인용관계가 없거나 낮은 현상이 발생하기도 하여 인용관계가 많은 학술지끼리 같이 묶여야 한다는 기준에는 맞지 않는 문제점이 발생하고 있다.

본 연구에서는 KCI에 등재된 학술지 분류와 KCI 인용망에 네트워크 군집화 알고리즘을 적용한 군집 결과를 토대로 어떠한 차이가 있는지 살펴보고자 한다. 이를 위해 최근 논문에서 대표적으로 다뤄지는 네트워크 알고리즘을 제시하고, 인용관계에 따른 각 알고리즘의 군집 결과 차이를 비교한다. 또한, 그 결과를 바탕으로 KCI 분류와 가장 유사성이 뛰어난 알고리즘을 찾고 군집 결과에 나타난 특성을 살펴본다.

본 논문의 구성은 다음과 같다. 2절에서는 연구대상과 연구절차를 소개하고, 군집 검출을 위한 여러 가지 네트워크 군집화 알고리즘 및 군집화 정도를 평가할 수 있는 판단 기법을 제시한다. 그리고 3절에서는 KCI 분류 학술지 인용 네트워크의 구조적 특성을 기술하고 군집 추출 알고리즘들의 적용 결과를 비교한다. 마지막 절에서는 요약 및 향후 추가로 필요한 연구 방향을 제시하고 결론을 맺는다.

2. 연구방법

사용 데이터는 2008년부터 2010년까지 3년간의 KCI 인용 자료이다. 이 자료는 8개의 대분류 분야 (인문학, 사회과학, 자연과학, 공학, 의약학, 농수해양, 예술체육, 복합학)와 대분류 분야 내 146개의 중분류 분야 총 1,408개의 학술지에 관한 인용정보로 구성된다. Table 2.1은 KCI 학술지 인용 네트워크의 구조를 학문 분야별로 분석 요약한 것이다.

Table 2.1 Comparison of characteristics of classification of KCI

classification	Number of journals	Number of edges	Density	Clustering coefficient	Average distance	Diameter
Humanities	350	5,362	0.044	0.1979	2.766	7
Social Science	434	9,848	0.052	0.2428	2.541	6
Natural Science	91	455	0.056	0.4447	3.199	7
Engineering	201	2,053	0.051	0.2673	2.845	7
Medicine and Pharmacy	161	1,465	0.057	0.2405	2.959	8
Marine Agriculture, Fishery	62	509	0.135	0.3945	2.647	8
Arts and Kinesiology	71	508	0.102	0.4230	3.092	7
Interdisciplinary Studies	38	79	0.056	0.4906	3.156	8

KCI 학술지 인용 네트워크에서 규모가 가장 큰 학문 분야는 사회과학 분야로 434개의 학술지를 보유하였으며, 자기 인용을 제외한 학술지 인용 횟수가 9,848회로 가장 활발한 교류를 보였다. 네트워크 내 전체 노드들의 관계 정도를 나타내는 밀도 (Schaeffer, 2007; Wasserman, 1994)는 대다수 대분류 분야에서 10% 안팎이며, 자기인용을 제외한 다른 학술지의 인용이 크지 않다는 것을 보여준다. 네트워크 결속계수 (Soffer와 Vazquez, 2005)도 50%를 넘지 않은 것으로 나타났다. 그러나 학술지 수가 상대적으로 적은 복합학, 예술체육 분야에서는 결속계수가 다른 분야에 비해 높게 나타났다. 이들 두 측도의 상관 정도는 0.427이었고, 유의확률이 0.292로 나타나 유의수준 5% 하에서 두 측도 간 관련성은 없는 것으로 나타났다. 하나의 학술지가 다른 학술지에 도달할 수 있는 평균 거리는 2∼3 정도였고, 두 학술지 간에 인용된 복수의 경로가 있을 경우 최단경로 중 가장 긴 거리를 뜻하는 직경은 6∼8로 조사되었다 (O'Malley와 Marsden, 2008; Wasserman, 1994).

본 연구에서는 KCI 인용 네트워크의 군집화를 위해 기초 분석을 적용한 11개의 알고리즘 중에서 메모리의 확장 문제를 발생시키는 최적 알고리즘과 지나치게 적거나 많은 군집을 가져오는 약한 연결 컴퍼넌트 알고리즘, 강한 연결 컴퍼넌트 알고리즘을 제외한 재귀적 연결 (recursively connected components)알고리즘, 연결선 중개성 (edge betweenness)알고리즘, 빠른 탐욕 (fast greedy)알고리즘, 닫힌 길 (walktrap)알고리즘, 선행 고유벡터 (leading eigenvector)알고리즘, 라벨 전파 (label propagation)알고리즘, 다단계 (multilevel)알고리즘, 인포맵 (infomap)알고리즘에 대하여 분석을 실시하였다. 분석 알고리즘에 관하여는 Kim 등 (2015), Lancichinetti와 Fortunato (2009b), Malliaros와 Vazirgiannis (2013), Orman 등 (2011)을 참조하여라.

분석 과정에서는 첫째, KCI 학술지 인용 네트워크가 가진 구조적 특성을 기술한다. 여기에는 대분류 분야를 기준으로 분류된 저널 보유수, 논문 간 인용관계 횟수, 네트워크 밀도 및 결속계수, 평균거리와 직경을 표기한다. 둘째, 각 군집 추출 알고리즘에 의해 도출된 군집 결과들의 크기와 성능평가 지표로 모듈화 정도, 정분류율을 산출하여 기술한다. 마지막 단계에서는 기존 KCI에 등록된 학술지 분류와 비교해 유사성이 가장 뛰어난 네트워크 알고리즘을 선별하고 군집 결과의 세부적 특성을 살펴본다. 즉, 생성된 각 군집 내부에서 다수를 차지하는 중분류 학술지의 점유율을 조사하고 군집 내 인용 정도를 파악하기 위해 군집 내 밀도를 산출한다.

본 논문에서는 군집화 우수성을 평가하는 판단 기준으로 모듈화 지수 (Newman, 2004; Malliaros와 Vazirgiannis, 2013)와 혼동행렬을 이용한 정분류율을 산출한다 (Fawcett, 2006; Tang과 Liu, 2010). 모듈화 지수는 네트워크 내 동일한 군집에 속한 모든 두 노드의 순서 쌍에 대하여 이들 노드 사이에 실제 존재하는 인용 관계 수에서 동일 노드들이 전체 네트워크에 있는 노드들을 대상으로 무작위로 가질 수 있는 인용 관계 수의 기댓값을 뺀 값이다. 모듈화 지수는 0과 1 사이의 값을 가지며 모든 노드가 하나의 군집에 속하게 되면 0이 된다. 정분류율은 먼저 주어진 네트워크 내 연결된 모든 노드의 순서 쌍을 대상으로 동일 군집 범주에 속하는지 유무를 점검한다. 이때 두 노드가 추출 알고리즘의 동일 군집 내에 할당되고 원시 네트워크의 군집에도 포함될 경우의 빈도수를 '참인 긍정 (True Positive, TP)'이라 부르며, 두 노드들이 추출된 군집과 원시 네트워크 내 동일 군집에 할당되지 않은 경우의 빈도수를 '참인 부정 (True Negative, TN)'이라 부른다. 또한 순서 쌍의 두 노드가 추출된 군집과 원시 네트워크 내 군집 어느 한 곳에만 소속되었다면 오류로 간주한다. 정분류율은 이들 네 가지 범주의 총 빈도수 즉, 네트워크 내 연결된 모든 가능한 노드 쌍들의 수에서 TP와 TF가 차지하는 상대도수 값이다.

3. 결과 분석

3.1. 군집 추출 알고리즘에 의한 분석 결과

• **재귀적 연결 컴포넌트 알고리즘**: '재귀적' 연결 컴포넌트 알고리즘은 127개의 군집과 0.004의 모듈화 지수 값을 가지는 것으로 측정되었다. 컴포넌트 알고리즘의 정분류율은 27.6%로 나타났다. '재귀적' 연결 컴포넌트는 단일 거대 네트워크와 군집 내 노드 수가 7개를 넘지 않는 소규모 네트워크 일곱 개가 추출되었을 뿐 나머지의 경우 모두 고립 노드들로 나타났다. 따라서 인용 관계에 의한 컴포넌트 알고리즘의 군집을 해석하는 것은 큰 의미가 없어 보인다.

• **연결선 중개성 알고리즘**: 군집의 모듈화 지수는 0.062이며, 군집 수는 508개로 나타났다. 연결선 중개성 알고리즘의 정분류율은 66.2%로 측정되었다. 전체 508개의 군집 중 KCI 대분류 8개 분야의 노드들이 총체적으로 연계된 거대 단일 네트워크가 생성되었고, 인문학 분야의 '인도철학'과 복합학 분야의 '인도연구'가 결합한 하위 네트워크 하나와 사회과학 분야의 '라틴아메리카연구'와 복합학 분야의 '이베로아메리카' 학술지가 결합된 소규모 네트워크 두 개를 제외한 나머지 505개 군집들이 모두 고립 노드 형태로 추출되었다. 군집 검출 결과는 KCI 분류와 비교해 상당한 차이를 보였다.

• **빠른 탐욕 알고리즘**: 군집화 정도를 나타내는 모듈화 지수는 0.464로 나타났으며, 군집 수는 모두 12개가 생성되었다. 빠른 탐욕 알고리즘의 정분류율은 57.5%로 나타났다. 군집의 특징으로 고립 노드가 전혀 발생하지 않았다는 것이며, 최소 3∼4개의 대분류 분야들이 서로 연계된 군집 네트워크를 형성하였다. 일례로 대분류 사회과학 분야의 '스포츠와 법' 학술지의 경우, 중분류 교육학 분야의 '한국체육교육학회'학회지와 연계를 이뤄 대분류 의약학 분야 내 중분류 물리치료학의 'Physical Therapy Korea'와 대분류 예술체육 분야 내 중분류 체육과 무용 학술지들과 더불어 하나의 군집 네트워크를 구성하였다. 가장 규모가 큰 군집은 인문학, 사회과학, 자연과학, 예술체육, 복합학 분야의 335개의 학술지가 연계된 인용 네트워크였고, 가장 작은 규모의 군집은 7개의 학술지로 구성된 대분류 복합학 분야의 중분류 복합학, 문헌정보학 분야 학술지의 인용 네트워크로 나타났다.

• **닫힌 길 알고리즘**: 추출된 군집의 모듈화 지수는 0.456이며, 군집 수는 37개로 나타났다. 정분류율은 55.9%로 측정되었고, 37개의 군집 중 27.0%가 고립 노드로 밝혀졌다. 고립 노드는 인문학 분야 2개 학술지 (19세기 영어권 문학, 중세르네상스영문학)와 사회과학 분야 5개 (International Economic Journal, Journal of Economic Integration, Seoul Journal of Economics, Journal of Economic Development, Journal of East Asian Studies), 공학 2개 (Industrial Engineering & Management Systems, Geosystem Engineering), 의약학 분야 1개 (Archives of Aesthetic Plastic Surgery)로 나타났다. KCI

분류범주와 비교하면 대분류의 다수 학문 분야들이 인용 관계에 의해 하나의 군집을 구성하는 경우가 많았다. 8개 대분류 분야 272개 학술지가 연계된 하위 네트워크가 단일 네트워크 규모로 가장 큰 군집을 형성하였다. 한편, 중분류에서 타 학문 분야와 중첩을 보이지 않은 분할 군집으로는 자연과학 분야의 수학, 공학 분야의 토목공학, 의약학 분야의 정형외과학과 안과학, 예술체육 분야의 음악학 정도로 나타났다.

• **선행 고유벡터 알고리즘**: 네트워크로부터 추출한 군집 모듈화 지수는 0.326이고, 군집 수는 14개로 나타났다. 알고리즘의 정분류율은 60.6%로 나타났다. 추출된 군집 결과에서는 고립 노드 없이 비교적 규모가 큰 군집 분할이 발생하여 작은 수의 군집이 형성되었다. 또한 다른 알고리즘과 비교해 모듈화 지수 값이 상대적으로 낮은 특징을 보였다. 추출된 군집의 삼분의 일 이상이 100여 개가 넘는 노드 수를 가진 중·대형 공동체 형태를 띄었고, 대분류인 인문학 분야와 사회과학, 예술체육 및 복합학 분야가 함께 결합된 322개의 학술지 군집이 가장 큰 규모의 군집 집단으로 분류되었다. 이와 함께 가장 작은 군집으로는 17개 학술지 노드로 구성된 대분류 사회과학 분야 내 회계, 세무, 조세법 관련 학술지로 나타났다. 대분류 자연과학 분야 내 14개 중분류에서는 물리학, 화학, 지구과학, 지질학, 천문학, 대기과학 등 6개 분야 학술지가 단일 군집을 형성하였다. 통계학 분야의 학술지 경우 여러 학문 분야와의 연계를 통해 다수의 군집을 이루었다. 자연과학 분야 내 중분류인 생활과학 학술지는 대분류 공학, 예술체육 및 복합학 분야와 군집을 형성하였고 농수해양 분야 내 중분류인 식품과학과 축산학, 수산학 분야의 학술지들과도 연계되어 군집을 이루었다.

• **라벨 전파 알고리즘**: 추출된 군집 모듈화 지수는 0.336로 나타나며, 군집 수는 76개로 나타났다. 정분류율은 71.6%로 나타났으며, 각 노드와 연결된 인접 노드의 라벨 빈도가 가장 큰 라벨을 선택하여 자신의 라벨을 업데이트하는 라벨 전파 알고리즘의 특성으로 인해 고립 노드 없이 대분류 내 연구분야들이 다양한 크기의 군집으로 나누어졌다. 이 중에서 인문학 분야와 사회과학, 예술체육 및 복합학 분야 200개 학술지로 구성된 하위 네트워크가 가장 큰 군집을 형성하였다. 그러나 초기 노드의 업데이트 순서 결정과 인접 노드의 최대 라벨이 동일할 경우 임의의 한 라벨을 선택함으로써 반복 실행 때 마다 달라진 군집을 형성하는 것이 큰 문제점으로 남는다. 대분류로 구분한 분야별 군집 크기를 살펴보면, 사회과학 (30개), 공학 (24개), 인문학 (21개), 자연과학 (20개), 의약학 (19개), 복합학 (17개), 예술체육 (11개), 농수해양 (8개) 순으로 분할이 발생하였다. 또한 2개의 노드로 구성된 소규모 공동체의 경우 자연과학 분야가 3개, 공학과 의약학 분야가 2개, 그리고 인문학과 복합학 분야에서 각각 1개씩 발생하였다.

• **다단계 알고리즘**: 네트워크로부터 추출한 군집 모듈화 지수는 0.504로 나타나며, 군집 수는 모두 6개로 나타났다. 알고리즘의 정분류율은 48.4%로 측정되었다. 추출된 군집의 특징을 살펴보면, 빠른 탐욕 알고리즘의 결과에서 처럼 고립 노드가 발생하지 않았고 생성된 군집 대부분이 200개가 넘는 학술지로 구성된 규모가 큰 하위 네트워크를 형성하였다. 가장 큰 규모의 군집은 대분류 인문학, 사회과학, 공학, 예술체육, 복합학으로 구성된 인용 네트워크로 426개의 학술지로 구성되어있으며, 가장 작은 규모의 군집은 대분류 사회과학 내 사회과학, 무역학, 지역학, 교육학, 법학, 행정학의 중분류로 구성된 76개 학술지 인용망으로 나타났다. 본 연구에서 구현한 군집화 알고리즘 중 가장 작은 군집 수를 나타내 추출된 각 군집 구성 멤버들의 대부분이 모든 대분류 분야 내 중분류 학문 분야들을 포함한 복잡한 인용구조 형태를 띠었다.

• **인포맵 알고리즘**: 추출된 군집은 모두 69개로 나타났으며, 군집화 정도를 나타내는 모듈화 지수는 0.410으로 측정되었다. 인포맵 알고리즘의 정분류율은 71.0%로 시행 알고리즘 중에서도 최상위를 차지하였다. 인포맵 알고리즘 시행 결과, 전체 학술지 인용 네트워크에서 대분류 인문학 분야 내 중분류인 영어와 문학의 '19세기 영어권 문학'과 대분류 공학 내 중분류인 자원공학의 'Geosystem Engineering'

학술지가 고립 노드로 분류되었다. 라벨 전파 알고리즘과 더불어 제시된 8가지 네트워크 추출 알고리즘 중 대분류 분야 내 의미 있는 중분류 집합 형태를 세분화하여 학문 간 분류 및 인용 관계에 의한 특징 파악이 용이하다는 긍정적 평가를 내릴 수 있었다. 반면에 KCI 분류 결과와 비교해서는 여전히 추출된 군집의 규모가 크고 이로 인해 모듈화 지수가 작게 나타난다. 특히 규모가 큰 인문학 분야를 비롯해 의약학 분야에서 세분화된 군집 추출의 필요성이 요구된다.

Table 3.1은 8가지 군집화 알고리즘을 적용하여 추출한 군집 특징을 정리한 것이다. 표의 결과를 살펴보면, KCI 중분류 수 146개와 가장 유사한 결과를 보인 군집화 알고리즘으로 '재귀적 컴포넌트' 방법과 '라벨 전파' 알고리즘, 그리고 '인포맵' 방법이 군집 수에서 근사한 것으로 파악되었다. 네트워크의 분할 정도를 나타내는 모듈화 지수에서는 '다단계' 알고리즘, '빠른 탐욕' 알고리즘, '닫힌 길' 알고리즘, '인포맵' 알고리즘 순으로 나타났으며, 정분류율에서는 '라벨 전파' 알고리즘, '인포맵' 알고리즘, '연결선 중개성' 알고리즘 순으로 분류의 정확성을 보였다.

Table 3.1 Comparison for the characteristics of detecting community structure based on the eight clustering algorithms

Algorithm	NC^*	Modularity	Accuray(%)
recursively connected component	127	0.004	27.6
edge betweenness	508	0.062	66.2
fast greedy	12	0.464	57.5
walktrap	37	0.456	55.9
leading eigenvector	14	0.326	60.6
label propagation	76	0.336	71.6
multilevel	6	0.504	48.4
infomap	69	0.410	71.0

NC^*: Number of Communities

Figure 3.1은 KCI 학술지망을 본 연구에서 언급한 군집화 방법을 이용하여 추출한 군집 수와 모듈화 지수를 비교한 것이다.

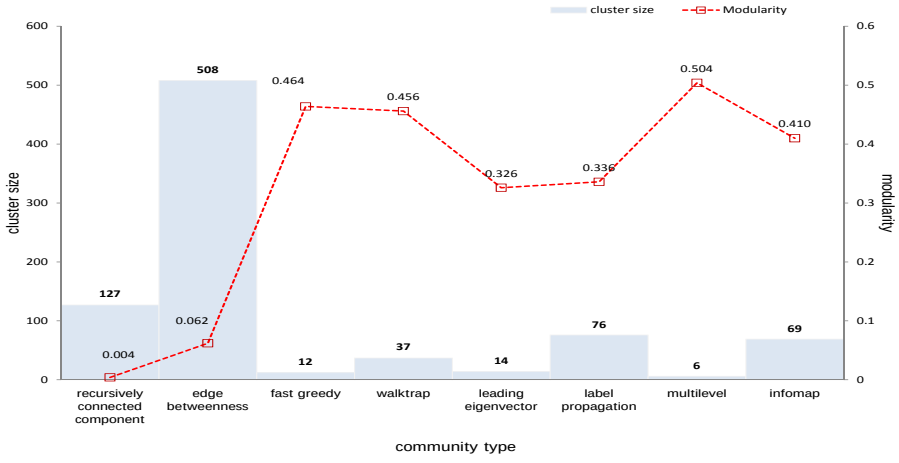

Figure 3.1 Comparison of the eight community detection algorithms

분류된 군집 내 노드 특성을 고려하지 않고 단순히 모듈화 지수의 크기 만을 비교하는 것은 의미가 없다. '컴포넌트' 방법과 '연결선 중개성' 알고리즘의 경우, 단일화된 거대 네트워크를 구성하거나 혹은 너무 세분화된 군집으로 분류되는 특징이 있어 올바른 학술지 군집화 방법으로 볼 수 없다. 또한 '라벨 전

파' 알고리즘의 경우, 군집 형성 과정에서 임의성 문제로 일관된 결과를 얻을 수 없다는 단점이 발생한다.

이러한 이유에서 8가지 군집화 알고리즘 방법 중 '인포맵' 군집화 알고리즘 방법이 학술지 인용 네트워크 군집화 방법으로 가장 우수하다고 판단된다. 이에 연구에서는 인포맵 알고리즘을 적용한 군집화 결과를 토대로 군집 내 상위 누적비율 50% 부근 중분류 분야와 점유율을 산출하는 추가 분석을 실시하였다.

Table 3.2의 군집 중 단일 중분류로 구성된 경우는 전체 클러스터의 17.4% (=12/69)를 차지하였고, 나머지의 경우 학술지 인용 관계가 빈번한 타 분야 학문들과 연계된 군집화 네트워크를 구성하였다. 가장 많은 중분류 분야를 포함한 군집화 네트워크의 경우 농수해양 (18.1%), 생물학 (16.9%), 농학 (12.0%)이 주축을 이루는 군집으로 총 83개의 저널을 보유하는 것으로 나타났다. 특히 가장 많은 저널을 보유한 군집의 경우는 한국어와문학 (60.2%)이 주축을 이루는 군집으로 98개의 학술지를 가지는 것으로 파악되었다. 군집 결과를 종합해 보면, 아직까지 군집 내 대표 학문 분야의 점유율이 높지 않고, 주류 학문 분야 역시 여러 분야가 복합적으로 결합된 군집을 형성하고 있음을 알 수 있었다.

Table 3.2 Results of clustering using Infomap algorithm based on the journal citation network

Cluster's ID	number of middle classification	number of Journals	middle classification level and shares with close to upper cumulative rate 50% by cluster(%)	density	Total Shares (%)
1	22	86	Education(14.0), Social Welfare(14.0), Nursing Science(11.6), Psychological Science(10.5)	0.230	50.0
2	13	98	Korean Language and Literature(60.2)	0.194	60.2
3	23	83	Marine Agriculture, Fishery(18.1), Biology(16.9), Agriculture(12.0)	0.161	47.0
4	15	88	Business Management(44.3)	0.183	44.3
5	22	64	Internal Medicine(20.3), Medicine and Pharmacy(10.9), Radiology(7.8), Pediatrics(7.8)	0.130	46.9
6	5	63	Law(79.4)	0.297	79.4
7	9	73	History(78.1)	0.165	78.1
8	13	63	Political Science(31.7), Area Studies(17.5)	0.159	49.2
9	6	56	Education(85.7)	0.257	85.7
10	14	54	Engineering(31.5), Computer Science(24.1)	0.188	55.6
11	11	39	Philosophy(64.1)	0.179	64.1
12	5	12	Tourism(58.3)	0.773	58.3
13	5	27	Kinesiology(55.6)	0.469	55.6
14	7	25	Public Administration(52.0)	0.492	52.0
15	10	31	Mechanical Engineering(48.4)	0.218	48.4
16	6	16	Journalism and Broadcasting(68.8)	0.504	68.8
17	3	27	English Language and Literature(48.1)	0.306	48.1
18	13	21	Architectural Engineering(19.0), Regional Development(19.0), Social Science(9.5)	0.386	47.6
19	11	27	Environmental Engineering(37.0), Civil Engineering(14.8)	0.251	51.9
20	8	23	Chemical Engineering(30.4), Polymer Engineering(17.4)	0.345	47.8
21	9	25	Design(24.0), Life Sciences(20.0)	0.345	44.0
22	3	22	Economics(86.4)	0.197	86.4
23	4	16	Japanese Language and Literature(68.8)	0.592	68.8
24	5	12	Civil Engineering(41.7)	0.485	41.7
25	4	9	Business Management(55.6)	0.472	55.6
26	6	17	Geology(29.4), Civil Engineering(23.5)	0.305	52.9
27	3	11	Geography(63.6)	0.400	63.6
28	4	12	Education(75.0)	0.758	75.0
29	9	16	Physics(43.8)	0.200	43.5
30	2	11	German Language and Literature(90.9)	0.527	90.9
31	3	12	Mathematics(83.3)	0.379	83.3
32	8	12	Religious Studies(33.3), Buddhist Studies(16.7)	0.394	50.0
33	3	10	French Language and Literature(60.0)	0.500	60.0
34	8	16	History(43.8)	0.250	43.8
35	1	14	Chinese Language and Literature(100.0)	0.462	100.0
36	4	18	Korean Medicine(83.3)	0.552	83.3
37	2	17	English Language and Literature(94.1)	0.125	94.1
38	2	7	Library and Information Science(85.7)	0.929	85.7
39	6	10	Biology(20.0), Forestry(20.0), Landscape Architecture(20.0)	0.400	60.0
40	2	13	Dentistry(92.3)	0.224	92.3
41	5	12	Fishery science(41.7), Oceanography(25.0)	0.356	66.7
42	4	8	Metallurgical Engineering(37.5), Mechanical Engineering(25.0)	0.500	62.5
43	3	11	Art(81.8)	0.309	81.8
44	5	6	Safety Engineering(33.3), Engineering(16.7)	0.567	50.0
45	1	6	Civil Engineering(100.0)	0.400	100.0
46	1	7	Electrical engineering(100.0)	0.524	100.0
47	3	7	Agricultural Economics(71.4)	0.548	71.4
48	3	13	Christian Theology(84.6)	0.246	84.6
49	2	6	Statistics(83.3)	0.567	83.3
50	5	7	Other Engineering(28.6), Civil Engineering(28.6)	0.714	57.1
51	3	7	Psychiatry(57.1)	0.286	57.1
52	2	5	Education(60.0)	0.900	60.0
53	3	7	Russian Language and Literature(57.1)	0.357	57.1
54	2	3	Korean Language and Literature(66.7)	1.000	66.7
55	1	3	Orthopedic Surgery(100.0)	1.000	100.0
56	2	3	Ophthalmology(66.7)	0.500	66.7
57	5	5	Other Oriental Languages and Literature(20.0), Literature(20.0), Religious Studies(20.0)	0.600	60.0
58	2	4	Law(75.0)	0.667	75.0
59	4	5	Area Studies(40.0), Spanish Language and Literature(20.0)	0.700	60.0
60	5	6	Physical Therapy(33.3), Other Medicine and Pharmacy(16.7)	0.367	50.0
61	1	2	Business Management(100.0)	1.000	100.0
62	2	4	English Language and Literature(75.0)	0.417	75.0
63	1	5	Musicology(100.0)	0.450	100.0
64	1	2	Dentistry(100.0)	1.000	100.0
65	1	2	English Language and Literature(100.0)	0.500	100.0
66	1	2	Economics(100.0)	0.500	100.0
67	1	2	Industrial Engineering(100.0)	0.500	100.0
68	1	1	English Language and Literature(100.0)	-	100.0
69	1	1	Resources Engineering(100.0)	-	100.0

4. 결론 및 토의

본 연구에서는 종래의 KCI의 학문 분야별 분류 방식이 학술지의 등재 신청 시 학술지 발간 주체가 학술지의 분류를 선정하는 방식이므로 인용 관계를 가진 학술지의 네트워크 구조를 이용한 학문 분야별 군집 검출 방법을 시도하였다.

KCI 학술지 인용 데이터 베이스에 등록된 전체 학술지를 대상으로 학술지 간 인용 정도를 인접행렬로 변환한 뒤 여덟 개의 네트워크 군집화 방법을 적용하여 군집을 추출하였다. 이 군집 결과를 바탕으로 기존 KCI에 등록된 학술지 분류와의 차이를 살펴보았다.

KCI 학술지 인용 네트워크의 구조를 학문 분야별로 분석한 결과를 요약하면 다음과 같다. KCI 학술지 인용 네트워크에서 규모가 가장 큰 학문 분야는 사회과학 분야로 조사되었으며, 자기 인용을 제외한 학술지 인용 횟수가 가장 커 가장 활발한 교류를 보였다. 또한 대다수 대분류 분야에서 10% 안팎의 밀도를 가진 것으로 나타나 자기인용을 제외한 다른 학술지의 인용이 생각보다 크지 않다는 것을 보여 주었다.

본 연구에서 적용한 군집 검출 결과를 토대로 군집 수와 모듈화 지수를 비교한 분석 결과를 정리하면 다음과 같다. '컴포넌트' 방법과 '연결선 중개성' 알고리즘의 경우, 단일화된 거대 네트워크를 구성하거나 혹은 너무 세분화된 군집으로 분류되는 특징이 있어 올바른 학술지 군집화 방법으로 볼 수 없었다. 또한 '라벨 전파' 알고리즘의 경우, 군집 형성 과정에서 임의성 문제로 일관된 결과를 얻을 수 없다는 단점이 발생하였다. 이러한 이유에서 8가지 군집화 알고리즘 방법 중 '인포맵' 군집화 알고리즘 방법이 학술지 인용 네트워크 군집화 방법으로 가장 우수하다고 판단하였다.

분류된 군집 내 노드 특성을 고려하지 않고 단순히 모듈화 지수의 크기 만을 비교하는 것은 의미가 없다. 이에 연구에서는 인포맵 알고리즘을 적용한 군집화 결과를 토대로 군집 내 상위 누적비율 50% 부근 중분류 분야와 점유율을 산출하는 추가 분석을 실시하였다. 분석 결과에서는 아직까지 군집 내 대표 학문 분야의 점유율이 높지 않고, 주류 학문 분야 역시 여러 분야가 복합적으로 결합된 군집을 형성하고 있음을 알 수 있었다.

네트워크 알고리즘을 적용할 경우에 발생할 수 있는 문제점으로는 KCI 분류 기준과 비교할 때 여전히 추출된 군집의 규모가 크고, 모듈화 지수가 작게 나타나는 특징들을 보였다. 특히 대분류 분야의 인문학과 의약학 분야의 경우 세분화된 군집의 추출이 요구되었다. 이러한 특성은 의미 있는 분야별 군집화를 위해 앞으로 개선해야 할 여지를 남긴다. 제한된 알고리즘을 사용한 한계는 발생될 수 있으며, 추후 연구에서 최신 데이터로 갱신된 군집 결과를 파악하거나 시간 변화에 따른 군집화 경향을 파악할 수도 있을 것이다.

본 연구는 KCI가 제공하는 논문 인용 정보망을 통해 국내 학술지의 군집화를 시도해 보았다. KCI 학술지 분류 방법을 네트워크 군집화 알고리즘을 적용한 결과와 비교해 봄으로써 KCI 학술지 망의 분류에 의한 구조적 특성을 파악할 수 있었고, 추출된 군집 결과는 KCI 분류기준과 최적화된 알고리즘을 선별하는 수단을 제공하였다.

References

Arenas, A., Duch, J., Fernández, A. and Gómez, S. (2007). Size reduction of complex networks preserving modularity. *New Journal of Physics*, **9**, 176.

Barnes, E. R. (1982). An algorithm for partitioning the nodes of a graph. *SIAM Journal on Algebraic Discrete Methods*, **3**, 541-550.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**, P10008.

Carpenter, M. P. and Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science*, **24**, 425-436.

Chun, H., and Leem, B. (2014). Face/non-face channel fit comparison of life insurance company and non-life insurance company using social network analysis. *Journal of the Korean Data & Information Science Society*, **25**, 1207-1219.

Clauset, A., Newman, M. and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, **70**, 066111.

Copic, J., Jackson, M. O. and Kirman, A. (2009). Identifying community structures from network data via maximum likelihood methods, *The BE Journal of Theoretical Economics*, **9**.

Danon, L., Díaz-Guilera, A. and Arenas, A. (2006). The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2006**, P11010.

Donath, W. E. and Hoffman, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, **17**, 420-425.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, **27**, 861-874.

Flake, G. W., Lawrence, S. and Giles, C. L. (2000). Efficient identification of web communities. In *Proceedings of the sxth ACM SIGKDD international conference on knowledge discovery and data mining*, 150-160, ACM.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**, 74-174.

Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, **104**, 36-41.

Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, **49**, 291-307.

Kim, H. (2008). Citation flow of the ASIST proceedings using pathfinder network analysis. *Journal of the Korean Society for Information Management*, **25**, 157-166.

Kim, J. K., Kim, S. H. and Oh, C. H. (2015). Comparison of journal clustering methods based on citation structure. *Journal of the Korean Data & Information Science Society*, **26**, 827-839.

Lancichinetti, A. and Fortunato, S. (2009b). Community detection algorithms: A comparative analysis. *Physical Review E*, **80**, 056117.

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. *Journal of Documentation*, **60**, 371-427.

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, **533**, 95-142.

Narin, F., Carpenter, M. and Berlt, N. (1972). Interrelationships of scientific journals. *Journal of the American Society for Information Science*, **23**, 323-331.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**, 026113.

Newman, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, **38**, 321-330.

Newman, M. E. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, **104**, 9564-9569.

O'Malley, A. J. and Marsden, P. V. (2008). The analysis of social networks. *Health Services and Outcomes Research Methodology*, **8**, 222-269.

Orman, G. K., Labatut, V. and Cherifi, H. (2011). *On accuracy of community structure discovery algorithms*, arXiv preprint arXiv:1112.4134.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2658-2663.

Rosvall, M. and Bergstrom, C. T. (2008). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, **104**, 7327-7331.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, **1**, 27-64.

Scott, J. (2012). *Social network analysis*, Sage.

Soffer, S. N. and Vazquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E*, **71**, 057101.

Suaris, P. R. and Kedem, G. (1988). An algorithm for quadrisection and its application to standard cell placement. *IEEE Transactions on Circuits and Systems*, **35**, 294-303.

Tang, L. and Liu, H. (2010). Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, **2**, 1-137.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge university press, **8**.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, **393**, 440-442.

Won, D., and Choi, K. (2014). Network analysis and comparing citation index of statistics journals. *Journal of the Korean Data & Information Science Society*, **25**, 317-325.

Zhang, L., Liu, X., Janssens, F., Liang, L. and Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, **4**, 185-193.

Journal of the Korean Data & Information Science Society
2016, **27**(4), 947–957

http://dx.doi.org/10.7465/jkdi.2016.27.4.947
한국데이터정보과학회지

A classification of the journals in KCI using network clustering methods

Jinkwang Kim[1] · Sohyung Kim[2] · Changhyuck Oh[3]

[13]Department of Statistics, Yeungnam University
[2]Academic Infrastructure Promotion Team, National Research Foundation of Korea

Received 24 June 2016, revised 22 July 2016, accepted 22 July 2016

Abstract

KCI is a database for the citations of journals and papers published in Korea. Classification of a journal listed in KCI was mainly determined by the publisher who registered the journal at the time of application for the journal. However, journal classification in KCI was known for not properly representing the quoting rate between journals. In this study, we extracted communities of the journals registerd in KCI based on quoting relationship using various network clustering algorithms. Among them, the infomap algorithm turned out to give a classification more being alike to the current KCI's in the aspect of the modular structure.

Keywords: Community, infomap algorithm, journal classification, KCI, modular, network clustering.

[1] Adjunct professor, Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk 712-749, Korea.
[2] Researcher, National Research Foundation of Korea, Gajeong-ro, Yuseong-gu, Daejeon 305-754, Korea.
[3] Corresponding author: Professor, Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk 712-749, Korea. E-mail: choh@yu.ac.kr