*Journal of the Korean Data & Information Science Society*
*2015,* **26**(4), 827–839

http://dx.doi.org/10.7465/jkdi.2015.26.4.827
한국데이터정보과학회지

# 논문 인용에 따른 학술지 군집화 방법의 비교

김진광[1] · 김소형[2] · 오창혁[3]

[13]영남대학교 통계학과
[2]한국연구재단 학술기반진흥팀

접수 2015년 4월 3일, 수정 2015년 4월 28일, 게재확정 2015년 6월 20일

**요 약**

학술지 인용 데이터베이스에서 네트워크 구조분석을 통해 학술지의 공동체를 추출하는 것은 인용관계에 따른 학술지의 집단을 파악하는 유용한 수단이다. 전 세계적으로 널리 활용되는 학술지 인용 데이터베이스인 Thomson Reuters의 SCI나 Elsevier의 SCOPUS가 제공하는 자료를 활용하여 인용관계에 따른 공동체 구조를 파악하는 시도가 이루어진 바 있으나, 국내 학술지 인용 데이터베이스인 KCI에서는 이러한 연구가 현재까지는 이루어지지 않은 것으로 알려져 있다. 따라서 본 연구에서는 기존의 여러 가지 네트워크 군집화 알고리즘을 이용하여 KCI에 등재되어 있는 자연과학 분야 학술지를 대상으로 인용관계에 따른 공동체를 파악하고 KCI에 등록된 학술지 분류와 비교하여 보았다. 적용된 군집화 방법 중 인포맵 알고리즘에 의한 분류가 KCI 등재 자연과학 분야 학술지의 인용관계 구조를 잘 반영하며, 기존의 KCI 분류와 가장 유사한 것으로 나타났다. 본 연구를 통해 얻은 KCI의 기존 분류와 차이점들은 장차 KCI 학술지의 재분류가 이루어질 시 고려의 대상이 될 수도 있을 것이다.

주요용어: 공동체, 네트워크 군집화 알고리즘, 학술지 인용 데이터 베이스, KCI.

## 1. 서론

학술지 인용 네트워크에서 노드인 학술지 간의 인용 형태에 따라 학술지를 군집화하여 분류하는 것은 많은 관심을 받아왔다. Narin 등 (1972)과 Carpenter와 Narin (1973)은 Science Citation Index (SCI) 자료를 대상으로 단일연결 군집화 알고리즘을 이용하여 학술지를 군집화하였다. Leydesdorff (2004)는 Journal Citation Reports (JCR) 2001년 자료로부터 SCI 학술지 인용 행렬을 구성하여 양연결 컴퍼넌트 분석 알고리즘을 이용한 군집화를 실시하였다. Kim과 Lee (2008)는 국내 연구개발 활동의 구조적 특성을 파악하고자 SCI 의 자료를 이용하여 논문의 인용/피인용 관계에 따른 연결정도 중심성을 바탕으로 과학기술 분야 간 융합에 대하여 살펴보았다. Zhang 등 (2010)은 2002년부터 2006년 사이 Thomson Reuters가 제공하는 웹데이터베이스 자료를 수집하여 다단계 군집화 알고리즘을 이용해 SCI의 주제별 분류 데이터 간 인용관계를 분석하는 연구를 하였다. Jeong 등 (2008)은 SCOPUS를 활용한 건설교통분야의 유망 연구영역 추출을 위하여 유사도 계수를 이용해 서지결합분석을 수행하고 학문적 근거가 같은 연구영역을 군집화하였다. Kim (2008)은 SCOPUS 데이터베이스에서 검색한 특정 프로시딩을 대상으로 네트워크분석을 시도하였으며, 완전연결 군집화 기법을 통하여 프로시딩의 지식이 어떠한 주제영역을 중심으로 네트워크 구조를 형성하고 있는지를 확인하였다.

[1] (712-749) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 박사과정.
[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

이러한 군집화는 학술지 인용네트워크에서만 국한되는 문제가 아니어서 일반적인 네트워크에서 노드들의 군집화에 관한 연구도 오랫동안 이루어져 왔다. 네트워크에서 공동체 검출은 관련된 자료들을 분석하는데 있어 핵심적인 문제이다 (Brandes 등, 2008). Girvan과 Newman (2002)은 노드들 간의 연결방향을 고려하지 않는 무향 네트워크에서 군집화 방법을 소개하였다. 공동체 검출에 관한 대다수 논문들은 무향 네트워크에 관한 것이다 (Levorato와 Petermann, 2011). 실제적으로는 유향 네트워크를 가정해야하는 상황에서 방향성을 고려하지 않는 무향 네트워크를 가정한 후 공동체 검출을 위한 향상된 방법만을 모색하는 것은 유향 정보의 손실로 이어져 검출된 공동체의 실질적 의미가 없어지는 경우가 발생할 수 있다. Lancichinetti와 Fortunato (2009a)는 네트워크 공동체 검출 알고리즘으로 알려진 일부 방법들을 대상으로 링크의 방향성과 가중치에 따라 발생할 수 있는 다양한 네트워크 환경 하에서 공동체 검출에 대한 알고리즘 간 성능 비교 실험을 실시하였다. Newman과 Leicht (2007)은 복잡한 형태의 대규모 네트워크의 공동체를 검출하기 위하여 EM (Expectation-Maximization) 알고리즘을 이용하여 유향 네트워크 모형을 가정하고 공동체에 대한 사전 지식없이 공동체의 검출이 가능함을 보였다. 또한, Rosvall과 Bergstrom (2008)은 가중 유향 네트워크에서 임의보행을 적용하여 공동체의 구조를 밝히는 방법을 소개하고 이를 많은 하위구조를 가지는 생물·사회계열의 구조적 특성을 밝히는데 적용하였다. Arenas 등 (2007)은 가중 네트워크의 노드 크기를 줄이는 방법으로 공동체 내의 가중치 합을 하나의 노드로 대체한 네트워크를 구성하여 모듈성 값을 구하는 방법을 제안하였다.

본 연구의 관심은 KCI 등재 자연과학분야 학술지를 대상으로 한 인용관계에 따른 군집화 즉, 공동체 검출에 있다. KCI는 한국연구재단 (National Research Foundation of Korea)에서 구축한 국내 학술지 인용데이터베이스로서 2011년 현재 1,408개의 학술지가 등재되어 있으며, 8개의 대분류 학문 분야와 146개의 중분류 학문분야가 있다. KCI의 학술지의 분류는 인용관계에 의한 분류가 아니므로 학술지 간의 인용관계에 따르는 공동체의 구조에 관심을 가지게 된다. 그러므로 본 연구에서는 KCI 등재 자연과학 분야 학술지 간 인용관계에 따른 공동체 검출 결과와 KCI 분류에 의한 공동체와의 차이가 어떠한지를 알아보고자 한다.

분석결과를 살펴보면 본 연구에 사용한 네트워크 공동체 검출 알고리즘 중 닫힌 길 알고리즘 그리고 인포맵 알고리즘의 검출 결과가 적절한 공동체 수와 제시한 측도를 충족시키는 것으로 나타났다. 특히 인포맵을 사용한 검출결과는 KCI 분류와 상당히 유사한 형태를 보였다.

본 논문의 구성은 다음과 같다. 2절에서는 네트워크의 개념 이해와 네트워크 내 공동체 검출 알고리즘을 소개한다. 3절에서는 KCI의 자연과학분야 등재 학술지를 대상으로 공동체를 검출하고, 군집화 방법에 따른 공동체 구성을 비교한다. 마지막절에서는 토의, 결론 그리고 추후 연구 방향을 제시한다.

## 2. 네트워크와 군집화 방법의 소개

### 2.1. 네트워크 개념의 이해

네트워크는 노드의 집합과 링크의 집합으로 구성된 집합이다. 네트워크는 그래프, 노드는 정점, 링크는 변이라고도 불린다. 노드의 집합을 $V = \{v_1, \ldots, v_n\}$라고 하고, $V$에 속하는 두 노드 사이의 관계 집합을 $E = \{(u, v) : u, v \in V\}$ 라고 하고, 이들로 구성되는 네트워크는 $G = (V, E)$로 표현하자.

네트워크를 그림으로 나타낼 때 노드는 점으로, 링크는 두 점을 연결하는 선으로 표현하는 것이 일반적이다. 네트워크의 각 링크에 0 이상의 값을 부여하고 이를 가중치라고 부른다. 즉, $(u, v) \in E$에 대하여 가중치 $w_{uv}$를 대응한다. 모든 두 노드 $u$와 $v$에 대하여 $w_{uv} = w_{vu}$인 경우에 네트워크를 무향이라고 부르며, 그렇지 않은 경우 유향이라고 부른다. 모든 가중치가 0 또는 1인 경우에 비가중 네트워크, 그렇지 않은 경우에는 가중 네트워크라고 부른다. 네트워크의 모든 노드에 대한 가중치로 이루어지는 행렬 $A = (w_{uv})_{n\times n}$을 인접행렬이라고 부른다. 무향 네트워크인 경우 인접행렬은 $A = A^T$을 만족하는 대칭

행렬이 된다. 본 연구에서는 무향 네트워크는 $G_U$, 유향 네트워크는 $G_D$로 나타내기로 한다. 유향 네트워크 $G_D$에서 두 노드 $u$와 $v$에 대하여 노드 $u$를 중심으로 볼 때 링크 $(u,v)$는 $u$에서 $v$로 나가는 링크, $(v,u)$는 $v$에서 $u$로 들어오는 링크라고 부른다.

네트워크 $G$에 대하여 노드 집합을 $V(G)$, 링크 집합을 $E(G)$로 표현한다. 네트워크 $G$의 노드 크기는 노드의 갯수로 정의되며, $|V| = n$로 나타내기로 하자. 네트워크 $G$의 링크 크기는 링크의 갯수로 정의되며, $|E| = m$ 로 나타내기로 하자. 노드의 크기가 $n$, 링크의 크기가 $m$인 네트워크를 $G(n,m)$이라고 한다.

무향 네트워크에서 노드 $u$의 차수는 그 노드에 연결되어 있는 링크의 갯수이며 $k_u$로 표시한다. 유향 네트워크에서 노드 $u$의 차수는 진입차수와 진출차수로 구분된다. $u$의 진입차수는 노드 $u$에 들어오는 링크의 갯수 $k_u^{\text{in}} = ||\{v|(v,u) \in E\}||$로, $u$의 진출차수는 노드 $u$에서 다른 노드로 나가는 링크의 갯수 $k_u^{\text{out}} = ||\{v|(u,v) \in E\}||$로 정의된다.

네트워크에서 시작 노드가 $v_0 = u$이고 마지막 노드가 $v_{k+1} = v$인 연결 링크의 열 $(u,v_1)$, $(v_1,v_2)$, $\ldots$ , $(v_{k-1},v_k)$, $(v_k,v)$을 노드 $u$에서 부터 노드 $v$에 이르는 경로라고 한다. 만일 두 노드 $u$와 $v$ 사이에 경로가 존재하면, 두 노드 $u$와 $v$는 연결되었다고 한다. 경로의 길이는 경로를 이루는 링크의 갯수이며, 두 노드 $u$와 $v$의 거리는 두 노드를 연결하는 경로 중 최단 길이를 가지는 경로 즉, 최단경로로 정의된다. 또한 네트워크에서 계산된 모든 최단경로 중 가장 긴 최단경로를 그 네트워크의 직경이라고 한다.

네트워크의 밀도는 네트워크 내 전체 노드들이 서로 간에 얼마나 많은 관계를 맺고 있는가를 표현하기 위한 것으로 가능한 총 관계 수 중에서 실제로 맺어진 관계 수의 비율로 정의된다 (Schaeffer, 2007). 네트워크 $G(n,m)$의 밀도는 다음과 같이 정의된다.

$$\delta(G) = \begin{cases} \dfrac{m}{\binom{n}{2}}, & \text{유향 네트워크인 경우,} \\ \dfrac{m}{n(n-1)}, & \text{무향 네트워크인 경우.} \end{cases} \tag{2.1}$$

밀도의 값은 0과 1 사이의 값을 갖는다. 밀도가 0인 경우는 노드간 연결이 전혀 없는 경우이며, 밀도 1은 모든 노드들이 서로 연결된 경우이다. 주어진 네트워크 내에서 링크의 갯수 $m$이 증가할수록 밀도는 증가한다.

본 연구와 관련하여 학술지 인용 망의 노드는 학술지로, 링크는 학술지 간 인용관계로 정의하고 학술지 간 인용 횟수를 링크의 연결강도로 사용하며 여기서는 이를 가중치라고 부른다.

네트워크 상의 공동체는 군집, 그룹, 컴퍼넌트, 응집된 서브그룹, 모듈 등으로 불리며, 네트워크의 노드를 그룹으로 나누는 일을 공동체 검출 또는 군집화라고 한다.

공동체의 정의는 다양하며 보편적으로 받아들여지는 정의는 없다. Radicchi 등 (2004)은 그래프 내의 노드 간 연결이 더 밀접한 노드들의 부분집합을 공동체라고 표현하였다. Malliaros (2013)는 네트워크 내에서 유사한 역할을 수행하거나 공통적인 속성을 공유하고 있는 노드 그룹을 공동체로 정의하였다. Newman과 Girvan (2003)은 각 공동체 내에는 많은 링크를 가지며 공동체 간에는 상대적으로 적은 링크를 갖도록 노드들을 공동체로 묶어가는 작업을 네트워크 군집화라고 정의하였다.

군집화 방법으로 네트워크의 공동체 형태를 추출할 때 하나의 노드가 두 개 이상의 군집에 속할 수 있는 경우에 소프트군집화라고 하며, 그렇지 않은 경우에는 하드군집화라고 한다.

군집화 방법은 계층적 방법과 비계층적 방법으로 분류되기도한다. 계층적 방법은 유사성 척도에 기초해 밀접한 관계를 가진 노드를 단계적으로 묶어 나가면서 공동체를 형성해 나가는 반면, 비계층적 방법은 사전에 정해진 공동체의 숫자에 따라 노드들이 공동체에 할당된다. 계층적 군집화는 군집의 수에 대한 사전지식을 필요로 하지 않지만, 예외적 특이노드가 제거되지 않고 반드시 어느 군집에 속하게 되는 문제가 발생된다. 또한 군집 병합으로 인한 군집간 중복은 허용되지 않는다. 계층적 군집은 각 노드를

하나의 군집으로 보고 가까운 군집을 합하는 상향식 병합 군집화와 전체 노드를 하나의 공동체로 보고 각 군집을 두 개의 군집으로 나누어 나가는 하향식 분할 군집화로 나뉘어진다.

군집화의 정도를 측정하는 측도로서 Newman (2004)의 모듈성지수 $Q$가 흔히 사용된다.

$$Q = \begin{cases} \dfrac{1}{2m} \displaystyle\sum_{u,v \in V} \left[ w_{uv} - \frac{k_u k_v}{2m} \right] \delta(c_u, c_v), & \text{무향 네트워크인 경우,} \\ \dfrac{1}{m} \displaystyle\sum_{u,v \in V} \left[ w_{uv} - \frac{k_u^{\text{out}} k_v^{\text{in}}}{m} \right] \delta(c_u, c_v), & \text{유향 네트워크인 경우.} \end{cases} \tag{2.2}$$

여기서 $c_u$는 노드 $u$가 속한 공동체를 나타내며, $\delta(c_u, c_v)$는 두 노드 $u, v$가 같은 공동체에 속하면 1, 그렇지 않으면 0의 값을 가진다. 식 (2.2)가 의미하는 바는 주어진 네트워크에서 무작위로 균등하게 선택된 두 노드 $u$, $v$의 차수가 $k_u, k_v$ 혹은 $k_u^{\text{out}}, k_v^{\text{in}}$이고 링크의 기댓수가 $k_u k_v/(2m)$ 혹은 $k_u^{\text{out}} k_v^{\text{in}}/m$ 일 때, 같은 공동체에 속하는 두 노드의 순서쌍에 대해 $w_{uv} - k_u k_v/(2m)$의 합을 정규화한 것이다. 따라서 $w_{uv} - k_u k_v/(2m)$이 큰 양의 값을 가질수록 모듈화된 정도가 크다고 볼 수 있다. 모듈성지수 $Q$는 항상 1보다 작으며, 모든 노드가 하나의 공동체에 속하게 되면 0이 된다. 일반적인 네트워크에서 $Q$는 0.3에서 0.7 사이의 값을 가지며, 0.7 이상의 값을 가지는 경우는 드물다고 알려졌다 (Newman과 Girvan, 2004; Lin 등, 2012).

### 2.2. 공동체 검출을 위한 네트워크 군집화 알고리즘의 소개

본 연구에서는 KCI에 등재된 자연과학분야 학술지를 기존에 연구되어 온 몇 가지 군집화방법을 적용하여 공동체를 추출하고 KCI에서 분류한 분류와 비교한다. 공동체 검출에 사용될 방법은 7가지이며, 닫힌 길 알고리즘, 빠른 탐욕 알고리즘, 다단계 알고리즘, 라벨 전파 알고리즘, 선행 고유벡터 알고리즘, 최적 알고리즘, 인포맵 알고리즘이다. 먼저 이들 공동체 검출 알고리즘의 특징을 소개한다.

**닫힌 길 알고리즘 (Walktrap algorithm)**

Pons와 Latapy (2006)가 제시한 방법으로써 네트워크 상의 임의 노드에서 출발한 임의 보행자가 매 시점에서 이동하고, 일정시간 후에 네트워크의 어떤 부분에 머문다면 그 부분을 하나의 공동체로 인식할 수 있다는 생각에서 출발하였다. 즉, 일정한 시간 후의 보행자의 위치는 전이행렬의 거듭 제곱 형태로 나타낼 수 있으므로, 이동 단계를 네 단계로 하여 전이행렬을 네 제곱하여 사용하였다.

닫힌 길 알고리즘은 무향 가중 링크를 가진 네트워크에 적용하며, 알고리즘 초기 단계에서는 개별 노드를 하나의 공동체로 두어 순차적으로 공동체의 크기를 키워나간다. 두 공동체를 병합하여 새로운 공동체를 만드는 과정은 ward (1963) 방법에 의한 거리 계산을 적용하며, 이 거리 값이 가장 작은 두 공동체를 병합하여 새로운 공동체를 만든다. 병합을 통한 반복의 마지막 단계에서는 하나의 공동체가 만들어진다. 최적 공동체 분할은 병합이 이뤄지는 매 단계별로 모듈성지수를 산출하고 그 가운데 가장 큰 값을 갖는 단계에서 구한다.

**빠른 탐욕 알고리즘 (Fast Greedy algorithm)**

최적해를 구하는 데 사용되는 근사적인 방법으로, 여러 경우 중 하나를 결정해야 할 때마다 그 순간에 최적이라고 생각되는 것을 선택해 최종적인 해답에 도달한다 (Lancichinetti와 Fortunato, 2009b).

Clauset 등 (2004)은 모듈성지수 Q의 탐욕 최적화를 수행하는 계층적 병합 군집화 방법으로 빠른 탐욕 알고리즘을 제안하였다. 알고리즘 초기에 각 노드를 하나의 공동체로 두고, 모든 가능한 공동체 쌍에 대하여 모듈성지수를 계산한다. 그 당시 가장 큰 모듈화 지수를 가진 공동체 간 병합이 이뤄지며, 다시 병합된 공동체와 나머지 공동체 간 모듈화 작업과 모듈성지수 산출과정이 반복된다. 만일 병합과정에서 현 단계의 모듈성지수보다 전 단계의 모듈성지수가 더 커 변화량이 감소되면, 그 단계에서 병합은 이뤄지지 않고 병합되지 않은 공동체 간 모듈화 작업이 계속된다. 이러한 공동체 병합 과정은 하나의 공동체

가 생성될 때까지 반복된다. 최적의 공동체 검출은 병합 단계별로 모듈화 지수를 산출하고 그 값이 최대가 되는 단계에서 최적 공동체 분할을 얻어낸다.

**다단계 알고리즘 (Multilevel algorithm)**

모듈성지수 최적화를 기반으로 한 계층적 탐욕 알고리즘으로 Blondel 등 (2008)이 제안하였다. 이 방법은 크게 두 단계로 구성되어 교대로 실행되며, 무향 가중 링크를 가진 네트워크를 대상으로 적용한다.

알고리즘 초기에 각각의 노드에 서로 다른 공동체를 할당하여 노드 개수 만큼의 공동체를 형성한다. 1단계 역할은 국소적으로 모듈화를 최적화하는 작은 공동체들을 형성하는 것이다. 각각의 노드에 대해 이웃노드의 공동체와 군집을 형성할 수 있는지 파악하고자 임의의 노드를 선택하는 것으로부터 시작하여 그 노드의 인접 노드가 속한 공동체에 선택 노드를 삽입하고, 기존 공동체와의 모듈성지수 차이 값을 비교한 뒤 수치가 증가할 경우, 삽입된 현재 공동체를 유지한다. 만일 증가하지 않았다면, 이전의 공동체로 돌아가 다음 이웃 노드에 대해 계속하여 연산을 진행한다. 이러한 계산은 선택 노드의 모든 이웃 노드에 대해 순차적으로 수행되며, 노드의 공동체 간 이동으로 모듈성지수의 값이 더 이상 증가하지 않을 때까지 계속된다. 2단계는 1단계의 결과로 동일 공동체를 갖는 노드를 합쳐 하나의 노드로 재구성하고, 공동체 간 링크를 재조정하여 새로운 네트워크를 구성한다. 위의 두 단계가 모두 끝난 과정을 하나의 패스라고 부르며, 실제 군집화 과정에 소요되는 시간은 첫 번째 패스에서 소비하게된다. 이 단계들은 최대 모듈성지수가 유지되는 한 계속적으로 반복시행된다.

이 알고리즘은 계산과정이 중첩되고, 매 단계마다 전체 데이터를 순차적으로 여러 번 접근하는 구조를 가지고 있어 데이터가 클 경우 계산과정 중 메모리 부족으로 인한 성능 저하가 발생된다.

**라벨 전파 알고리즘 (Label propagation algorithm)**

Raghavan 등 (2007)이 제안한 방법으로 네트워크의 각 노드들이 모듈성지수를 이용하여 공동체를 형성하던 방법과 다르게 단지 이웃한 노드들이 가진 가장 큰 라벨 공동체에 가입함으로써 빠르게 주변으로 라벨을 전파시켜가며 공동체를 형성하는 방식을 사용한다. 초기 알고리즘은 무향 비가중 링크를 갖는 네트워크에 적용되었으나, Zhang 등 (2013)이 무향 가중 링크 네트워크에서도 적용할 수 있는 일반화된 라벨 전파 알고리즘을 발표하였다.

알고리즘 초기에 각 노드들에게 유일한 라벨을 할당한다. 매 반복 단계마다 랜덤하게 각 노드의 라벨 업데이트 순서를 정하고, 각 노드와 연결된 인접 노드의 라벨 빈도가 가장 큰 라벨을 선택하여 자신의 라벨로 업데이트한다. 만일 동일한 크기를 갖는 인접 노드의 라벨이 있는 경우 임의의 한 라벨을 선택하도록 한다. 네트워크 내 모든 노드들의 라벨이 더 이상 변하지 않는 최대 라벨을 가질 경우 알고리즘은 종료되며, 종료 싯점에서 동일 라벨을 갖는 노드 그룹을 공동체로 결정한다.

라벨 전파 알고리즘의 특징으로 초기 노드의 업데이트 순서 결정과 동일 크기를 갖는 인접 노드의 최대 라벨이 주어진 경우 임의로 라벨을 선택함으로써 알고리즘을 반복 실행할 때 마다 매번 달라진 공동체가 생성되거나 혹은 단일 공동체를 형성하는 거대 네트워크가 발생하기도 한다.

모듈성지수가 양의 값을 갖는다는 것은 공동체 구조의 존재 가능성을 암시한다. 따라서 모듈성지수 값이 가능한 한 큰 양의 값을 갖는 네트워크의 분할을 탐색함으로써 정밀하게 공동체 구조를 찾아낼 수 있다.

**선행 고유벡터 알고리즘 (Leading eigenvector algorithm)**

Newman (2004)이 네트워크의 스펙트럴 특성의 관점에서 모듈성지수를 재구성한 차별화된 시도로, 이 방법의 특징은 무향 비가중 링크를 가진 입력 네트워크의 인접행렬 요소와 대응된 두 노드의 기대 링크 수를 뺀 모듈성 행렬을 정의하는 것이다.

선행 고유벡터 알고리즘은 모듈성 행렬에서 가장 큰 양의 고유값과 대응되는 고유벡터를 계산하여 고유벡터 내 해당되는 요소의 부호를 기반으로 노드들을 두 개의 공동체로 분할한다. 만일 고유벡터 내 모

든 요소들이 동일한 부호를 가지면 네트워크는 더 이상 분리될 공동체 구조를 가지지 않는다는 것을 의미하며 해당 서브 네트워크를 남겨둔다. 이런 방법으로 전체 네트워크가 더 이상 나눠질 수 없는 서브 네트워크로 분해될 때 군집화 과정이 종료된다.

**최적 알고리즘 (Optimal algorithm)**

Brandes 등 (2008)은 최대 모듈성지수 문제를 정수선형계획법으로 계산할 수 있는 방법을 제안하였다. 선형계획법에서 이용되는 결정변수는 두 노드 사이의 거리를 나타내며, 두 노드의 순서 쌍에 대해 하나의 변수를 갖는다. 만일 두 노드가 동일 공동체 내에 있는 경우 결정변수는 0의 값을 가지며, 그렇지 않은 경우 1의 정수값을 가진다. 이것으로 식 (2.2)와 동일한 의미를 갖는 목적함수를 구할 수 있다.

최적 알고리즘에서 동일 공동체를 형성하는 노드들은 이들 변수들의 등위성 관계로 해석할 수 있다. 따라서 등위성 관계 제약식 (Agarwal과 Kempe, 2008; Dinh과 Thai, 2013)을 충족하는 결정변수의 해를 구하여 노드 간 쌍방 관계가 있는 0 값을 갖는 동일 공동체 노드들의 결정변수를 목적함수에 대입함으로써 네트워크에서 가장 최적화된 구조의 모듈성지수를 구한다.

알고리즘은 노드 갯수 $n$에 대하여 $\binom{n}{2}$개의 결정변수와 $3\binom{n}{3}$개의 제약식을 가지므로 노드가 늘어나면 계산량이 크게 증가한다.

**인포맵 알고리즘 (Infomap algorithm)**

앞서 소개한 알고리즘의 경우 모듈성지수를 이용하여 공동체 구조를 찾았던 것과 달리 맵 방정식으로 명명된 정보묘사길이를 최소화하는 공동체를 검출하는 방법으로 네트워크의 엔트로피를 이용한다 (Rosvall과 Bergstrom, 2008).

알고리즘을 구현하기 전 무손실 압축 방법으로 널리 사용되는 허프만 코드를 사용하여 네트워크를 구성하고 있는 노드들에 고유한 이름을 부여한다. 이는 공동체 검출 과정에 많은 노드들의 정보가 메모리에 적재되어 무수한 반복을 통해 구조가 결정되는 만큼 압축된 노드 정보는 큰 장점으로 작용한다. 최상의 분할을 찾는 문제는 네트워크 내 임의 보행자의 이동 경로 정보를 최소화하는 것으로 네트워크 간 이동과 네트워크 내 이동 경로 정보를 최소화하는 엔트로피를 구하여 하나의 공동체를 구성하도록 하는 것이다.

검출방법으로는 다단계 알고리즘과 같은 방법이 사용된다. 즉, 먼저 모든 노드가 하나의 공동체라는 가정 하에서 출발하며, 탐욕 알고리즘을 사용하여 임의의 노드로 부터 인접한 이웃 두 공동체를 짝지워 가며 가장 작은 최소 정보량을 나타내는 그룹을 하나의 공동체로 묶어 나간다. 공동체가 형성되면 2단계로 공동체를 형성한 노드들의 차수를 합쳐 단일 노드의 차수로 만들고 인접 노드들과의 관계성을 고려하여 네트워크를 재구성한다. 여러 노드가 묶인 단일 노드와 주변 노드들의 정보는 다시 허프만 코드로 재명명되어 부여한다. 이런 과정들을 반복적으로 수행하다 묶이기 이전 공동체 노드의 정보 묘사 길이가 더 짧으면 공동체로 묶지 않고 대신 새로운 공동체 노드를 기준으로 다시 앞서 언급한 방법을 반복수행한다. 최종적으로 네트워크 내 모든 공동체를 대상으로 구한 정보 묘사 길이가 가장 짧은 공동체가 최적의 공동체로 선별된다. 이때 공동체를 형성하는 노드들은 하나의 거대 노드로 재구성하여 전체 네트워크를 간소하게 나타낼 수 있다.

## 3. KCI 자연과학 분야 학술지 군집 결과

### 3.1. 데이터

본 연구에서 사용한 학술지 인용 데이터는 2011년 한국연구재단 KCI에 등재된 대분류 자연과학 분야 전체 91개 학술지를 대상으로 하였다. 대분류 자연과학 분야 내에는 자연과학, 자연과학일반, 수학, 통계학, 물리학, 천문학, 화학, 생물학, 지구과학, 지질학, 대기과학, 해양학, 생활과학, 기타자연과학의 14개 중분류가 있다. 학문 분야에 관한 분류는 각 학술지의 등재 신청시 분류를 학술지 발간 주체가 선정하게 된다.

### 3.2. KCI 자연과학 분야 학술지 인용 네트워크의 구조적 특성

학술지 네트워크의 공동체 검출에 앞서 네트워크 구조를 시각적으로 나타내기 위해 연결도를 나타내었다.

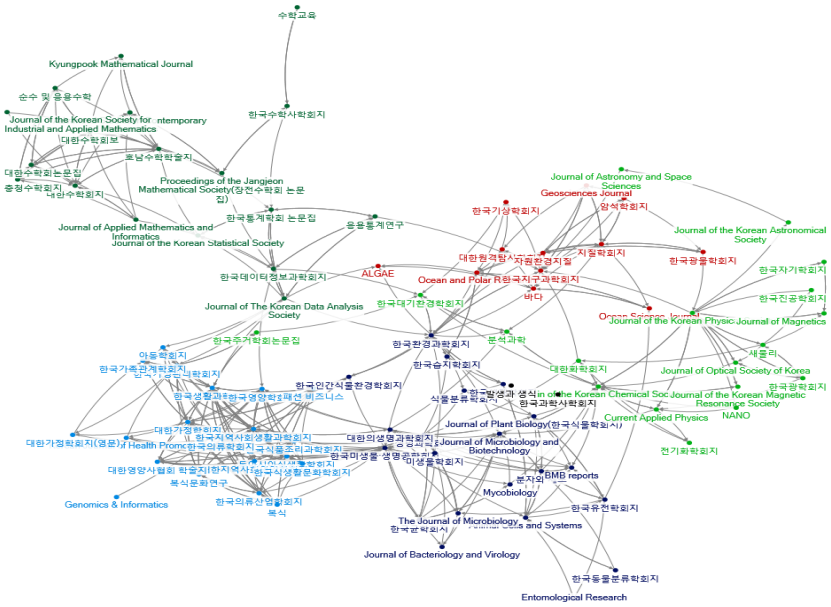


**Figure 3.1** Article citation network map of the journals in the area of natural science, KCI

Figure 3.1에서 전체적인 학술지 인용 네트워크의 구조를 살펴보면, 자연과학 분야 91개 학술지의 인용 횟수는 모두 455회로 조사되었다. 네트워크 밀도는 0.056로 낮아 논문 인용의 상당 수가 소수 학술지 간 집중된 것으로 나타났다. 한 노드와 직접 연결되어 있는 노드들 사이의 모든 가능한 링크와 실제 링크 비율의 평균으로 정의한 네트워크 결속계수는 0.4447로 네트워크 내에서 인접한 노드들과 교류를 맺는 경우는 과반을 조금 넘는 것으로 조사되었으며, 하나의 노드가 다른 노드에 도달할 수 있는 평균 거리는 3.199이고 직경은 7로 나타났다.

### 3.3. 모듈성지수를 이용한 향상된 공동체 추출 알고리즘들의 구현 결과

#### 닫힌 길 알고리즘

공동체의 모듈성지수는 0.66435이며 공동체의 갯수는 17개로 나타났다. 자연과학 분야 학술지 인용 네트워크의 전체 연결링크 455개 중 KCI 분류에 의한 공동체에도 속하고, 닫힌 길 알고리즘에 의한 공동체에도 속하는 링크 (true positive link; TP)의 수는 244개이었고, 두 개의 분류 방법에 의한 공동체의 어느 곳에 속하지 않는 링크 (true negative link; TN)의 수는 109개이어서 KCI에 의한 분류를 옳은 분류로 가정하였을 경우에 닫힌 길 알고리즘의 정분류율 (accuracy)은 (244+109)/455=77.6%로 나타났다. TP와 TN 그리고 accuracy에 대한 정량화 값은 Fawcett (2006) 논문을 참고하였다.

닫힌 길 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 자연과학일반 분야의 ‘Genomics & Informatics’와 생물학 분야의 ‘발생과 생식’, 그리고 기타자연과학 분야의 ‘한국과학사학회지’ 등 3개의 학술지가 고립노드로 분할됨을 알 수 있었다. 생활과학 분야와 수학 분야의 경우 두 개의 공동체로 양분화되어 나타났고, 생물학 분야의 ‘ALGAE’ 학술지의 경우 해양학 분야에 할당되어 군집을 형성하였

다. 대기과학 분야의 '한국기상학회지'와 '한국대기환경학회지' 등도 서로 다른 공동체로 할당되는 현상을 보였다.

빠른 탐욕 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 생물학 분야의 '발생과 생식', 그리고 기타자연과학 분야의 '한국과학사학회지'가 고립노드로 분할되었다. 생활과학 분야의 경우 두 개의 공동체로 양분화되어 나타났고, 생물학, 수학, 통계학 분야는 KCI와 동일하게 분할되었다. 물리학 분야와 화학 분야는 하나의 공동체로 형성하였고, 해양학, 지질학, 지구과학, 대기과학 분야가 하나로 묶여 규모가 큰 공동체를 구성하였다.

**다단계 알고리즘**

네트워크로부터 추출한 공동체의 모듈성지수는 0.60459 이었으며, 공동체의 갯수는 7개로 나타났다. 자연과학 분야 학술지 인용 네트워크의 전체 연결링크 수 455개 중 TP는 277개였고, TN은 47개로 나타나 KCI를 기준으로 한 다단계 알고리즘의 정분류율은 71.2%로 나타났다.

다단계 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 빠른 탐욕 알고리즘에서와 같은 고립노드가 발생하였다. 특히 수학분야와 통계분야가 하나의 공동체를 형성하고, 물리와 화학 그리고 천문학 분야가 각각의 동일 공동체를 이루는 특징을 보였다. 생활과학 분야의 '한국주거학회논문집'을 제외한 생활과학 전 분야와 통계학 학술지 'Journal of the Korean Data Analysis'가 하나의 공동체를 구성하였다.

**라벨 전파 알고리즘**

추출된 공동체에 대한 모듈성지수는 0.60346로 공동체 갯수는 모두 20개로 나타났다. 자연과학 분야 학술지 인용 네트워크의 전체 연결링크 수 455개 중 TP는 188개였고, TN은 117개로 나타나 KCI를 기준으로 한 라벨 전파 알고리즘의 정분류율은 67.0%로 나타났다.

라벨 전파 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 앞서 언급한 공동체 검출 방법에서는 KCI 분류기준과 비교하여 여러 학문 분야의 학술지들이 하나의 공동체 내에 혼합적으로 구성된 것에 반해 라벨 전파 공동체의 결과는 상당부분 단일 분야로 분류된 공동체 구조 형태를 보였다. 군집화 결과로 생물학 분야의 '발생과 생식' 그리고 기타자연과학 분야의 '한국과학사학회지'가 고립노드로 분류되었다. 생활과학은 4개의 공동체로, 생물학과 물리학 그리고 수학 분야는 세 개의 군집으로, 화학 분야는 두 개의 군집으로 나눠져 공동체를 형성하는 것을 알 수 있었다. 또한 지구과학, 지질학, 대기과학 분야가 하나의 군집으로 나타났고 통계학 분야의 경우도 하나의 군집을 형성하였다.

**선행 고유벡터 알고리즘**

모듈성지수는 0.61864이고 공동체 갯수는 모두 10개로 나타났다. 자연과학 분야 학술지 인용 네트워크의 전체 연결링크 수 455개 중 TP는 221개였고, TN은 77개로 나타나 KCI를 기준으로 한 선행 고유벡터 알고리즘의 정분류율은 65.5%로 나타났으며, 7개 알고리즘 중 가장 낮은 수치를 나타내었다.

선행 고유벡터 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 고립노드는 생물학 분야 '발생과 생식', 그리고 기타자연과학 분야 '한국과학사학회지'로 나타났다. 물리학과 화학 전 분야의 학술지와 생물학 일부 학술지들이 동일 공동체를 형성하는 것으로 나타났고, 해양, 지질학, 대기과학 분야의 학술지와 생물학 일부 학술지들이 하나의 공동체를 구성하는 것으로 나타났다.

**최적 알고리즘**

모듈성지수가 0.68595으로 나타났으며, 공동체 갯수는 모두 10개로 나타났다. 자연과학 분야 학술지 인용 네트워크의 전체 연결링크 수 455개 중 TP는 239개였고, TN은 73개로 나타나 KCI를 기준으로 한 최적 알고리즘의 정분류율은 68.6%로 나타났다.

최적 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 생물학 분야의 '발생과 생식', 그리고 기타자연과학 분야의 '한국과학사학회지' 등이 고립노드로 분할되었고, 물리학과 화학 분야의 학술지가 동일

한 공동체를 형성하여 전체적으로 KCI 분류와 비슷한 공동체 형태를 보였다. 생활과학 분야 학술지가 세 개의 군집으로 나눠 공동체를 형성하는 것으로 드러났고, 생물학 분야의 'ALGAE'의 경우 지질학과 대기과학, 천문학, 그리고 해양학 분야의 학술지와 더불어 공통된 공동체 그룹에 포함되어 나타났다.

**인포맵 알고리즘**

모듈성지수가 0.67044로 비교 대상 7개 알고리즘에서 두 번째 큰 값을 가지며, 검출된 네트워크의 공동체 구조 측면에서도 KCI가 분류하는 14개 범주에 가장 근접한 13개의 분할된 공동체 수를 나타내었다. 자연과학 분야 학술지 인용 네트워크의 전체 연결링크 수 455개 중 TP는 253개였고, TN은 102개로 나타나 KCI를 기준으로 한 인포맵 알고리즘의 정분류율은 78.0%로 7개 알고리즘 중 가장 큰 수치를 보였다.

인포맵 알고리즘 추출 방법에 의한 공동체 특징을 살펴보면, 닫힌 길 알고리즘을 제외한 다른 공동체 추출 알고리즘과 동일한 고립노드를 검출하였다. 생물학 분야의 경우 세 개의 군집으로 생활과학 분야와 수학 분야는 각각 양분화된 형태의 공동체 구조를 형성하였다. 이에 반해 화학 전 분야와 대기과학 분야의 경우 하나의 군집을 이루었고, 생물학 분야의 'ALGAE'와 해양학 분야 역시 하나의 군집을 구성하였다.

Table 3.3는 7가지 군집화 알고리즘을 적용하여 추출한 공동체의 특징을 각 알고리즘별로 정리한 것이다.

**Table 3.1** Comparison of 7 clustering algorithms

| Algorithm | Modularity | $NC^*$ | Accuray | Features |
|---|---|---|---|---|
| Walktrap | 0.66435 | 17 | 77.6% | Isolated nodes : Genomics & Informatics, Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : ALGAE, Korean Journal of Remote Sensing and Oceanographic field is formed by the same community |
| Fast greedy | 0.68595 | 10 | 68.6% | Isolated nodes : Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : Mathematics and statistics sectors tied to the same community, Physics and chemistry are to form a community. The field of Biology, mathematics and statistics are formed in the same communities as compared with KCI. |
| Multi level | 0.60459 | 7 | 71.2% | Isolated nodes : Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : Some areas of biology and geology, oceanography is formed the same community, Natural Science in general and life science sectors are tied. |
| Label propagation | 0.60346 | 20 | 67.0% | Isolated nodes : Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : Biology, physics, mathematics is divided into several clusters. Earth science, geology and atmospheric sciences is formed to a single community. |
| Leading eigenvector | 0.61864 | 10 | 65.5% | Isolated nodes : Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : statistics form a single cluster, Chemistry, Physics and several of biology is bound to the same community. Life sciences journals are divided into two clusters. |
| Optimal | 0.68595 | 10 | 68.6% | Isolated nodes : Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : Physics and chemistry are formed by the same community. Life sciences journals are divided into three clusters. The field of Biology, mathematics and statistics are formed in the same communities as compared with KCI. |
| Infomap | 0.67044 | 13 | 78.0% | Isolated nodes : Development & Reproduction, Journal of the Korean History of Science Society.<br>Classification differences : It is very similar compared to the KCI classification. Biology (ALGAE) sector and oceanographic areas are tied as one statistics form a single cluster. The journals of Life sciences and Mathematics are divided into two clusters. |

$NC^*$: Number of the Communities

## 4. 토의 및 결론

본 연구에서는 KCI에 등재되어 있는 자연과학 분야 학술지를 대상으로 네트워크 군집화 방법을 적용하고 인용관계에 따른 공동체를 파악하여 학술지를 분류하였으며, 기존 KCI에 등록된 학술지 분류와의 차이를 알아보았다. 인포맵 알고리즘을 사용한 분류 결과는 KCI의 기존 분류와 유사한 형태를 보였으나 일부 학술지의 경우 차이가 존재하였으며 이러한 차이점들은 KCI 학술지의 재분류시 고려의 대상이 될 수 있을 것이다.

네트워크 군집화의 정도를 측정하는 모듈성지수는 최적 알고리즘과 빠른 탐욕 알고리즘, 인포맵 알고리즘, 닫힌 길 알고리즘 순으로 낮았다. 한편 분류된 공동체의 갯수는 라벨전파 알고리즘, 닫힌 길 알고리즘, 인포맵 알고리즘 순으로 작아졌고, 특히 공동체 수가 가장 많은 라벨전파 알고리즘의 경우 다른 학술지와의 인용관계가 없거나 자기 인용만으로 네트워크를 구성한 고립노드의 수가 가장 많이 검출되었다. 이와 대조적으로 최적 알고리즘과 인포맵 알고리즘의 경우에서는 고립노드의 수가 다른 알고리즘에 비해 가장 작게 나타났다. 생물학 분야의 '발생과 생식' 그리고 기타자연과학분야의 '한국과학사학회지'가 고려한 모든 알고리즘에서 고립노드의 공동체로 판정되었다. 최적 알고리즘과 빠른 탐욕 알고리즘, 인포맵 알고리즘의 경우 적용한 다른 알고리즘에 비해 모듈성지수가 높았으나 인포맵 알고리즘의 추출 공동체 갯수가 KCI의 분류 수 14개와 가장 유사하였다. 최적 알고리즘에 의해 분류된 공동체의 경우 화학분야의 학술지가 KCI 등록 분야와는 달리 물리학 분야의 학술지 공동체와 묶이고, 생활과학분야로 등록된 학술지들이 여러 개의 공동체로 분할된 형태를 보였다. 이에 반해 인포맵 알고리즘의 경우 자연과학 분야, 자연과학일반 분야, 기타자연과학 분야 만이 여러 공동체에 나뉘어져 할당된 것을 볼 수 있었다.

Figure 4.1은 KCI 학술지 네트워크를 대상으로 본 연구에서 언급한 공동체 검출 방법에 따라 추출된 공동체 개체의 수와 군집을 형성한 모듈 내 노드 간 밀집 정도의 측도로 모듈성지수를 구하여 비교한 것이다.

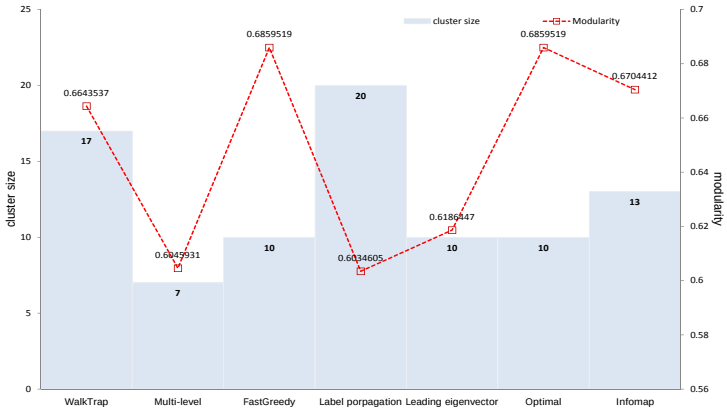


**Figure 4.1** Comparison of 7 comunity detection algorithms

KCI의 학문분야별 분류는 학술지의 등재 신청시 학술지 발간 주체가 선정하므로 이러한 학술지의 분류는 인용관계에 의한 분류가 아니다. 따라서 인용 네트워크의 공동체를 추출하는 알고리즘을 적용하여 KCI 등재 학술지들을 분류하여 보았으며, 인포맵 알고리즘이 다른 방법에 비해 기존의 KCI 분류에 더 유사하고 인용관계를 잘 나타내는 것으로 파악되었다.

본 연구에서는 KCI에 등재된 자연과학 분야 학술지의 인용 네트워크에 대하여 공동체를 파악하였으나, 후속 연구에서는 KCI 학술지 인용 데이터 베이스에 등록된 전체 학술지를 대상으로 확장하는 연구가 진행되어야 할 것이다.

## References


Agarwal, G. and Kempe, D. (2008). Modularity-maximizing graph communities via mathematical programming. *The European Physical Journal B-Condensed Matter and Complex Systems*, **66**, 409-418.

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

Blondel, V., 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.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, **20**, 172-188.

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

Clauset, A., Newman, M. and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, **70**, 66111.

Dinh, T. N. and Thai, M. T. (2013). Towards optimal community detection: From trees to general weighted networks. *Internet Mathematics (accepted pending revision)*.

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

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

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**, 7821-7826.

Jeong, E. S., Cho, D. Y., Suh, I. W. and Yeo, W. D. (2008). Emerging research field selection of construction & transportation sectors using scientometrics. *The Journal of the Korea Contents Association*, **8**, 231-238.

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. A. and Lee, H. S. (2008). A study on network analysis for science and technology activity. *Proceedings of the Autumn Conference of the Korean Operations Research and Management Science Society*, 498-503.

Lancichinetti, A. and Fortunato, S. (2009a). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, **80**, 016118.

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

Levorato, V. and Petermann, C. (2011). Detection of communities in directed networks based on strongly p-connected components. *International Conference on Computational Aspects of Social Networks, CASoN, IEEE*, 211-216.

Leydesdorff, L. (2004). Clusters and Maps of Science Journals Based on Bi-connected Graphs in the Journal Citation Reports. *Journal of Documentation*, **9**, 715-723.

Lin, W., Kong, X., Yu, P. S., Wu, Q., Jia, Y. and Li, C. (2012). Community detection in incomplete information networks. In *Proceedings of the 21st international conference on World Wide Web, ACM*, 341-350.

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports Journal*, **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. and Girvan, M. (2003). Mixing patterns and community structure in networks. *in Statistical Mechanics of Complex Networks*, **625**, 66-87.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**, 26113.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, **69**, 066133.

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.

Park, C. (2013). Simple principle component analysis using Lasso. *Journal of the Korean Data & Information Science Society*, **24**, 533-541.

Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications,*, **10**, 191-218.

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.

Raghavan, U. N., Albert, R. and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, **76**, 036106.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105**, 1118-1123.

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

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, **58**, 236-244.

Zhang, A., Ren, G., Cao, H., zhu Jia, B. and bin Zhang, S. (2013). Generalization of label propagation algorithm in complex networks. In *Control and Decision Conference (CCDC)*, 2013 25th Chinese, IEEE, 1306-1309.

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*
*2015,* **26**(4), 827–839

http://dx.doi.org/10.7465/jkdi.2015.26.4.827
한국데이터정보과학회지

# Comparison of journal clustering methods based on citation structure

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 3 April 2015, revised 28 April 2015, accepted 20 June 2015

**Abstract**

Extraction of communities from a journal citation database by the citation structure is a useful tool to see closely related groups of the journals. SCI of Thomson Reuters or SCOPUS of Elsevier have had tried to grasp community structure of the journals in their indices according to citation relationships, but such a trial has not been made yet with the Korean Citation Index, KCI. Therefore, in this study, we extracted communities of the journals of the natural science area in KCI, using various clustering algorithms for a social network based on citations among the journals and compared the groups obtained with the classfication of KCI. The infomap algorithm, one of the clustering methods applied in this article, showed the best grouping result in the sense that groups obtained by it are closer to the KCI classification than by other algorithms considered and reflect well the citation structure of the journals. The classification results obtained in this study might be taken consideration when reclassification of the KCI journals will be made in the future.

*Keywords*: Community, journal citation database, KCI, network clustering algorithm.

[1] Graduate student, Department of Statistics, Yeungnam University, Gyeongbuk 712-749, Korea.
[2] Researcher, National Research Foundation of Korea, Daejeon 305-754, Korea.
[3] Corresponding author: Professor, Department of Statistics, Yeungnam University, Gyeongbuk 712-749, Korea. E-mail: choh@yu.ac.kr