http://dx.doi.org/10.3745/KIPSTD.2011.18D.6.501

컴퓨터공학 분야 학술 논문 데이터베이스를 이용한 키워드 연관 네트워크 기반 지식지도

정 보 석[†] · 권 영 근[††] · 곽 승 진[†††]

요 약

최근 여러 분야에서 활용되고 있는 지식지도는 대량의 정보 속에 숨겨진 특징을 찾아서 그 의미를 파악할 수 있도록 가시적인 형태의 결과를 보여주는 것을 말한다. 본 논문에서는 2000년부터 2010년까지 컴퓨터 공학 분야의 국내 학술지에 게재된 논문들의 데이터베이스를 활용하여 연구동향 분석을 위한 키워드 연관 네트워크 기반의 지식지도를 제안하였다. 그 지식지도를 통해 키워드 연관 네트워크에서 개별 키워드가 속한 연결 요소의 크기 변화를 살펴봄으로써 관련 연구 주제의 영향력 변화를 추론할 수 있었다. 또한, 랜덤 네트워크와의 비교를 통해 키워드 연관 네트워크에서 최대 연결 요소의 크기가 상대적으로 매우 작으며, 상호 관련성이 높은 키워드 쌍들의 그룹이 밀집되어 있음을 보였다. 이는 최대 연결 요소에 대응하는 연구 분야가 크지 않으며 여러 소규모의 연구 주제들이 느슨한 형태로 연결되어 있음을 암시한다. 이러한 분석 결과들은 단순히 개별 키워드의 사용 빈도수 등을 분석하는 전통적인 방식으로는 얻기 어렵다는 점에서 본 논문에서 제안한 지식지도가 연구 동향 분석의 방법이 될 수 있다.

키워드 : 지식지도, 키워드 연관 네트워크, 랜덤 네트워크, 연구 동향 분석

A Knowledge Map Based on a Keyword-Relation Network by Using a Research Paper Database in the Computer Engineering Field

Bo Seok Jung[†] · Yung-Keun Kwon[††] · Seung Jin Kwak[†††]

ABSTRACT

A knowledge map, which has been recently applied in various fields, is discovering characteristics hidden in a large amount of information and showing a tangible output to understand the meaning of the discovery. In this paper, we suggested a knowledge map for research trend analysis based on keyword-relation networks which are constructed by using a database of the domestic journal articles in the computer engineering field from 2000 through 2010. From that knowledge map, we could infer influential changes of a research topic related a specific keyword through examining the change of sizes of the connected components to which the keyword belongs in the keyword-relation networks. In addition, we observed that the size of the largest connected component in the keyword-relation networks is relatively small and groups of high-similarity keyword pairs are clustered in them by comparison with the random networks. This implies that the research field corresponding to the largest connected component is not so huge and many small-scale topics included in it are highly clustered and loosely-connected to each other. our proposed knowledge map can be considered as a approach for the research trend analysis while it is impossible to obtain those results by conventional approaches such as analyzing the frequency of an individual keyword.

Keywords : Knowledge Map, Keyword Network, Random Network, Research Trend Analysis

1. 서 론

오늘날 정보통신기술의 발전은 정보의 생산, 전달, 교환을 쉽게 할 수 있도록 촉진시켰다. 그러나 급증하는 정보와 지식의 양에 비해 대량 생산된 정보를 분류하여 사용자에게 필요한 정보를 전달하는 기술의 발전은 더딘 상황이다. 이런 상황에서 축적된 정보를 가공하는 작업이 필요한데, 사용자에게 필요한 지식을 효과적으로 전달하고자 하는 방법

※ 이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2010-32A-H00006).
† 준 회 원 : 울산대학교 전기공학부 석사과정
†† 정 회 원 : 울산대학교 전기공학부 조교수
††† 정 회 원 : 충남대학교 문헌정보학과 부교수(교신저자)
논문접수 : 2011년 7월 12일
수 정 일 : 1차 2011년 9월 5일
심사완료 : 2011년 9월 6일

〈표 1〉 지식지도에 관한 기존 연구에서의 분석 방법

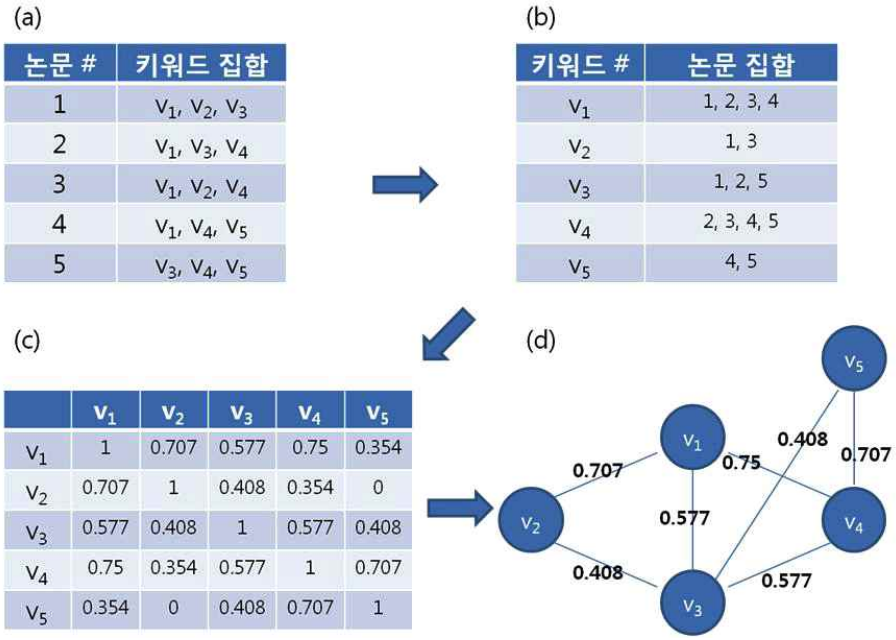
	분석 내용	분석 방법	분석 결과
KRF 선정 연구 과제 [1]	연구현황 분석	키워드 벡터 합산 후 키워드 빈도수 도출	키워드 빈도수, 허핀달 지수
	연구관계 분석	연구 분야간 네트워크 형성	유사도 행렬, 네트워크의 정점 중심성 및 매개 중심성
	연구 분야 재분류 분석	세분야간 유사도 매트릭스 도출 후 계층적 군집분석	세분류간의 유사도 행렬
	선정과제 특성 분석	연구 분야별 선정율 도출 후 선정과제 특성 도출	연구 분야별 선정과제비율, 평균증가율
사회과학 분야 [2]	학술연구 관계분석 지식맵	서로 다른 정보요소(사업명, 중분류, 성별, 세대, 소속기관) 간의 관계를 네트워크로 표현	관계 네트워크
	학술연구 키워드 맵 분석	주요 키워드 간의 관계를 네트워크로 표현	네트워크의 정점 중심성 및 매개 중심성
	학술연구의 협력 연구자 집단분석	협력 연구 집단 발견 및 집단 특성 분석	연구자 관계 네트워크, 정보요소별 연구자수
	학술연구 분야간 연계 분석	사회과학분야에서 내외로의 파급력 분석	학술분야 관계 네트워크, 네트워크의 정점 중심성 및 매개 중심성

중 하나가 지식지도이다. 지식지도는 대량의 정보 속에 숨겨진 특별한 형태와 패턴을 찾아서 그 의미를 파악할 수 있도록 가시적인 형태의 결과를 보여주는 도구이다. 지식지도는 다양한 목적으로 활용되는데, 일반적으로 국가 정책결정이나 사회 현상 연구의 목적으로 개발되고 있으며[1][2], 기업의 업무수행을 위한 지식지도도 존재한다[3]. 기존의 지식지도 관련 연구를 살펴보면 데이터베이스에 기초한 빈도수 조사, 평균 증가율 등 통계적 분석을 실시하거나, 분석 대상을 구성하는 개체들 사이의 관계를 네트워크로 표현하여 이를 시각화한 후 직관적으로 해석하는 방법과 같은 단순한 분석에 그치고 있다[1][2]. 특히, 그러한 접근 방법에 의한 결과물은 지식지도에서 구축된 네트워크의 특징을 거의 반영하지 못한다는 문제점이 있다. 따라서 보다 다양하고 복합적인 새로운 지식을 가공하기 위해서는 네트워크로 표현된 분석대상을 시스템 측면에서 살펴볼 필요가 있다. 최근 이러한 네트워크 기반의 지식지도 구축의 중요성이 강조되고 있으며[4][5] 실제로 복잡계나 사회적 네트워크 등을 연구하는 분야에서 진행되었다. 예를 들면, [6]에서는 범죄자들의 관계를 네트워크로 구성하고 시뮬레이션을 통해 범죄의 전파과정을 설명하는데 활용하였다. [7]에서는 대학교의 구성원들 사이의 이메일을 주고받는 횟수를 가중치로 계산하여 관계 네트워크를 형성하였고, 이를 시간의 흐름에 따라 변화하는 형태를 분석하였다. 또한, [8]에서는 한 마을의 사람들 간의 행복이 전파되는 과정을 네트워크 모델링을 통해 시간에 따라 분석하였다. [9]에서는 학술논문의 저자와 제목을 이용하여 형성한 네트워크에 대해 차수 분포(degree distribution), 평균 분리 정도(average separation), 클러스터링 계수(clustering coefficient) 등을 조사하여 그 특징을 살펴보았으며, [10]에서는 학술논문의 메타데이터에 대해 거리(distance)만 고려한 네트워크와 거리와 차수를 동시에 고려한 네트워크 모델을 만들어 각 모델에 대한 구조 분석을 하기도 하였다. 이러한 예들에서처럼 지식지도는 그것이 적용되는 목적에 따라 생성되는 네트워크의 성격이나 내용이 다르다. 본 연구에서는 어떤 학술 분야의 연구 동향 분석에 도움을 줄 수 있는 지식지도 구현에 초점을 맞추는데, 이러한 목적을 위해 네트워크 시스템 차원에서의 분석을 바탕으로 한 지식지도가 구축된 사례가 아직 없다. 여기에서는 국내 컴퓨터공학 분야의 학술 논문 데이터베이스를 이용하여 키워드 연관 네트워크 기반의 지식지도를 구축하고 이를 네트워크 관점에서 다양한 분석을 시도해 본다.

본 논문은 다음과 같이 구성된다. 2장에서는 학술 연구 분야 분석을 위한 지식지도 구축의 최근 사례를 소개한다. 3장에서는 본 논문에서 컴퓨터공학 분야 키워드 연관 네트워크를 구축하는 방법에 대해 설명한다. 4장에서는 다양한 키워드 연관 네트워크 분석 결과를 보이고 5장에서 결론을 맺는다.

2. 주요 관련 연구

최근 지식지도를 통해 특정 분야의 연구 동향이나 특성을 파악하려는 연구가 시도되고 있는데, 그 중에서 한국학술진흥재단에서 수행한 두 건의 연구 사례가 보고되었다 [1][2]. 먼저, [1]에서는 한국학술진흥재단이 지원하는 학술연구조성사업에 제출된 연구과제 데이터베이스를 통해 선정된 연구과제들의 특성을 파악하기 위해 지식지도를 다음과 같이 구축하였다. 연구 분야는 크게 [대분류]-[중분류]-[소분류]-[세분류]의 네 가지 단계로 나누어지는데, 이 중 거시분석은 대분류 수준에서 중분류 유사도 네트워크 혹은 중분류 수준에서 세분류 유사도 네트워크를 작성함으로써 학문분야의 전체적인 구조를 표현하였다. 반면 미시분석을 통해 중분류 내에서 소분류들을 대상으로 연구현황, 연구관계, 연구 분야 재분류, 선정과제 특성 분석 등 전략적이고 정책적인 시사점을 얻을 수 있는 심화된 분석을 수행하였다. 한편, [2]에서



논문 #	키워드 집합
1	v_1, v_2, v_3
2	v_1, v_3, v_4
3	v_1, v_2, v_4
4	v_1, v_4, v_5
5	v_3, v_4, v_5

키워드 #	논문 집합
v_1	1, 2, 3, 4
v_2	1, 3
v_3	1, 2, 5
v_4	2, 3, 4, 5
v_5	4, 5

	v_1	v_2	v_3	v_4	v_5
v_1	1	0.707	0.577	0.75	0.354
v_2	0.707	1	0.408	0.354	0
v_3	0.577	0.408	1	0.577	0.408
v_4	0.75	0.354	0.577	1	0.707
v_5	0.354	0	0.408	0.707	1

(그림 1) 키워드 연관 네트워크의 예

는 2002년부터 2007년까지 한국학술진흥재단에서 지원한 사회과학 분야 연구과제들에 대한 통계 및 특성을 분석한 지식지도를 구축하였다. 여기에서는 학술연구 관계분석 지식맵, 학술연구 키워드 맵 분석, 학술연구의 협력연구자 집단분석, 학술연구 분야간 연계분석 등 4가지 종류의 분석이 수행되었다.

[1]에서 사용된 미시분석 방법과 [2]에서 사용된 분석 방법을 비교해 보면 매우 유사함을 알 수 있다 <표1>. 그 표에서 보듯이, 구축된 네트워크에 대해 정점 중심성이나 매개 중심성과 같은 네트워크의 단순한 구조적 특징에 대한 분석과 특정 속성값에 대한 평균값 등 통계적 수치를 계산하는 분석이 주를 이룬다. 비록 이러한 연구들이 복잡한 관계를 네트워크로 표현하여 시각화한 점에서 이전의 접근법에 비해 진일보하기는 하였지만, 그러한 지식지도내의 네트워크가 어떤 구조적 혹은 동역학적 특징을 보이는지에 대한 분석이 여전히 미비한 실정이다.

3. 컴퓨터공학 분야 지식지도 구축

본 논문에서는 컴퓨터공학 분야에 대한 연구 동향을 파악하는 데 도움이 될 지식지도 구축을 시도한다. 이 절에서는 국내 학술논문 데이터베이스를 활용하여 키워드 연관 네트워크 기반 지식지도 구축 방법을 설명한다.

3.1 데이터베이스 수집

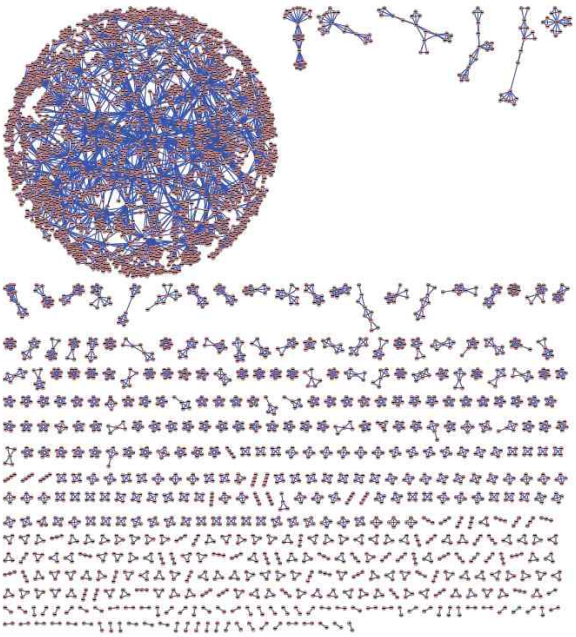
본 연구에서는 한국과학기술정보연구원(KISTI)에서 제공한 컴퓨터공학 분야의 학술 논문 데이터베이스를 활용한다. 2000년부터 2010년까지 컴퓨터공학 분야 53,432건의 국내 학술논문 중에서 분석 대상물의 신뢰성을 높이기 위해 학술지명, 학회명, 저자명, 년도, 소속기관명, 키워드 등 여섯 가지의 메타데이터가 모두 존재하는 18,297건의 논문만을 이용하여 분석하였다. 그 논문 집합에서 서로 다른 키워드를 추출한 결과 총 키워드의 개수는 37,723개였다.

3.2 키워드 연관 네트워크 구축

본 논문의 학술 논문 지식지도는 기본적으로 키워드를 분석하여 가공한 정보를 사용자에게 제공한다. 키워드를 주요 매개체로 선정한 이유는 그것이 학술 논문에 대한 내용과 연구 분야를 효과적으로 축약해 놓은 항목이기 때문이다. 키워드 연관 네트워크는 무방향 가중치 그래프(undirected weighted graph) $G(V, E)$로 표현될 수 있는데, 정점 $v \in V$는 각각의 키워드를 나타내고 v_i, $v_j \in V$인 간선 (v_i, v_j)의 가중치 $w(v_i, v_j)$는 해당 키워드 쌍(v_i, v_j)의 유사도를 나타낸다. 간선의 가중치는 두 키워드가 얼마나 빈번히 논문에 동시에 출현했는지를 상대적으로 측정하기 위해 아래와 같은 코싸인 유사도[1][2] 값으로 정의되는데, 이 식에서 $P(v)$는 키워드 v가 포함된 논문들의 집합을 표현한다.

$$w(v_i, v_j) = \frac{|P(v_i) \cap P(v_j)|}{|P(v_i)| \cdot |P(v_j)|}.$$

(그림 1)은 키워드 연관 네트워크 구축 과정의 예를 보여준다. 이 예에서처럼 다섯 개의 논문과 다섯 개의 키워드가 (그림 1(a))와 같이 존재할 때, 각 키워드에 대해 출현하는 논문들의 리스트를 (그림 1(b))와 같이 작성할 수 있으며, 이를 바탕으로 5×5 유사도 행렬을 구하게 된다(그림



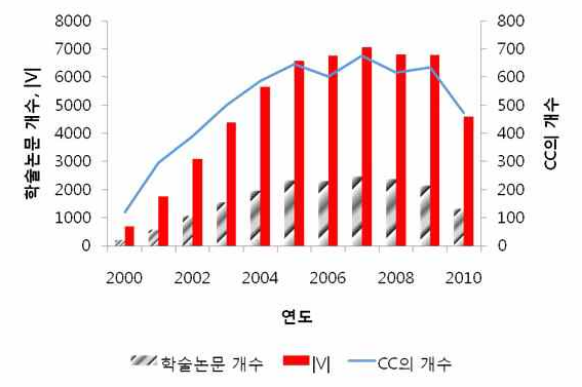
(그림 2) 2003년 키워드 연관 네트워크

1(c)). 한편, 키워드 연관 네트워크의 간선의 수를 조절하기 위해 간선임계값(edge-threshold)을 파라미터로 설정하는데, 가중치가 간선임계값보다 클 때 해당하는 간선이 네트워크에 포함되도록 하였다. (그림 1(d))는 간선임계값이 0.4일 때 구축된 키워드 연관 네트워크 결과를 보여 준다 (예를 들면, $w(v_2,v_4)=0.354$이므로 $(v_2,v_4)\notin E$이 됨을 알 수 있다)[1].

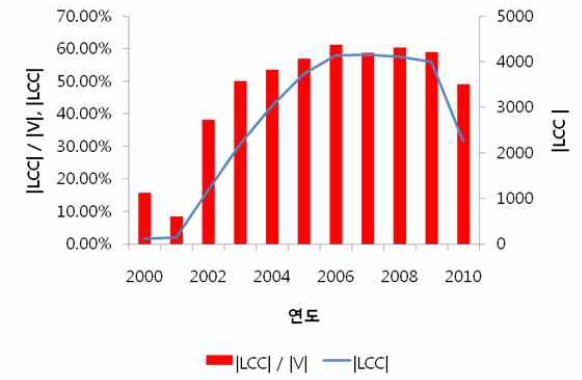
본 논문에서는 연구 동향 분석을 위하여 키워드 연관 네트워크를 연도별(2000년부터 2010년까지 모두 11년)로 구축하여 그 변화를 살펴본다. (그림 2)는 전체 데이터베이스 중 2003년도 학술 논문들만을 이용하여 구축한 2003년도 키워드 연관 네트워크이다. 그 그림에서 보듯이 네트워크는 매우 거대한 연결요소(Connected Component; CC) 한 개와 매우 많은 개수의 작은 CC로 구성됨을 알 수 있다[2]. 다른 연도의 키워드 연관 네트워크들도 비슷한 경향을 보인다. 한편, 가장 많은 수의 정점을 포함한 CC는 가장 폭넓은 연구 분야를 가리키는 것으로 이해할 수 있으며 이러한 가장 큰 CC를 LCC(Largest Connected Component)라 정의한다.

4. 컴퓨터공학 분야 지식지도 분석

이 절에서는 3절에서 구축한 키워드 연관 네트워크를 바탕으로 연구동향에 관한 다양한 분석을 시도한다. 4.1절에서는 키워드 네트워크의 연도별 CC/LCC 크기의 변화에 대해



(a) 연도별 학술논문의 개수, |V|, CC의 개수 변화



(b) 연도별 $\frac{|LCC|}{|V|}$ 및 $|LCC|$ 변화

(그림 3) 연도별 기초 데이터 값의 변화

서 살펴보고, 4.2절에서는 키워드별로 그것이 속한 CC의 영향력 변화에 대해 살펴본다. 4.3절과 4.4절에서는 랜덤 네트워크와의 비교를 통해 키워드 연관 네트워크가 통계적으로 유의한 구조적 특징을 가지고 있는지에 대해 살펴본다.

4.1 키워드 연관 네트워크의 연도별 변화

먼저, 학술 논문의 개수, 서로 다른 키워드의 개수($|V|$), 그리고 키워드 연관 네트워크(간선임계값=0)에서 CC의 개수가 연도별로 어떤 변화를 보이는지를 살펴보았다(그림 3(a)). 그 그림에서 보듯이 학술 논문과 키워드 개수는 모두 2005년까지 계속 증가하다가 그 이후에는 큰 변화가 없음을 알 수 있다. CC의 개수의 경우도 대체로 비슷한 경향을 보이지만, 2006년과 같이 학술 논문이 증가함에도 오히려 감소하는 약간의 경향의 차이를 보이는 경우도 존재한다. 이러한 결과는 논문의 양적 증가와 연구 분야의 다양성 증대가 밀접하게 관련이 있음을 뜻한다. 다음으로 CC중에서 가장 큰 LCC에 대해 좀 더 면밀히 살펴보기 위해 LCC를 구성하는 정점 개수($|LCC|$)와 전체 키워드 개수에서 LCC를 구성하는 키워드 개수가 차지하는 비율($\frac{|LCC|}{|V|}$)에 대해 연

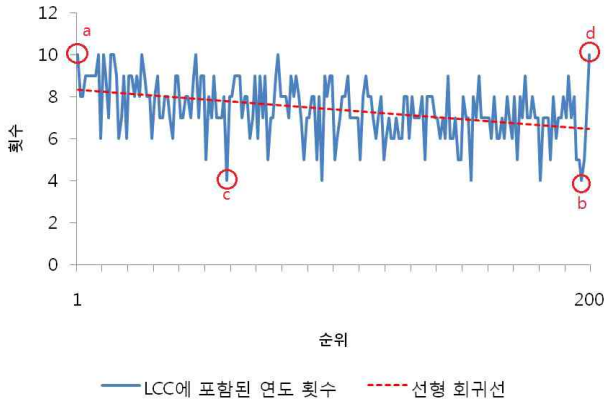
1) $\forall v\in V,\ w(v,v)=1$이지만 자기간선(self-loop)은 키워드네트워크에 항상 포함되지 않는다고 가정한다 (즉, $(v,v)\notin E$).

2) 본 논문에서의 CC는 2개 이상의 정점으로 구성된 경우만을 고려하며 이 그림에서는 어떠한 정점과도 연결되어 있지 않은 고립(isolated) 정점에 대해서는 분석대상에서 제외되었다.

도별 변화를 조사하였다(그림 3(b)). (그림 3(a))의 CC의 개수와 비슷하게 2002년부터 2006년까지 $|LCC|$가 꾸준히 증가하고 2009년까지 그 수가 유지됨을 알 수 있다. 흥미로운 점은 $\frac{|LCC|}{|V|}$도 비슷한 패턴으로 2002년부터 2006년까지 계속 증가한다는데, 이는 논문 수가 양적으로 증가하면서 LCC의 절대 크기가 커졌을 뿐만 아니라 그것의 상대적 비중도 커졌음을 뜻한다. 특히, 2003년부터 LCC에 포함된 키워드의 개수가 전체 키워드 개수의 50%를 넘는다는 점에서 LCC가 연구 동향 분석에 있어 중요한 대상임을 설명한다.

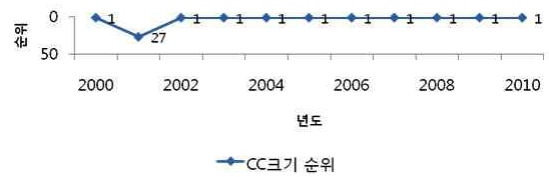
4.2 키워드별 영향력 변화

다음으로 주요 키워드에 대해 LCC포함 관련성을 조사해 보기 위해 2000년부터 2010년까지 전체 데이터베이스에서 가장 출현 빈도가 높은 상위 200개의 키워드를 추출하였다. (그림 4)는 그 키워드들의 출현 빈도에 따른 순위와 연도별 키워드 연관 네트워크에서 LCC내에 속한 횟수의 관련성을 조사한 결과이다[3]. 그 그림에서 보듯이 상위 200개의 키워드들은 모두 11년 중 최소 4회 이상은 LCC에 속해 있었다. 또한, LCC내에 포함된 횟수에 대한 선형 회귀선의 기울기가 음수 값(-0.0095)으로 나오는 것을 볼 때(p-value : 4.48×10^{-09}), 출현 빈도수가 높은 키워드일수록 LCC에 포함된 횟수가 많다는 것을 알 수 있다. 즉, 자주 출현하는 키워드일수록 가장 넓은 연구 분야에서 자주 사용된다고 설명될 수 있다. 이런 현상이 두드러진 예는 (그림 4)에서 a와 b에 해당하는 키워드를 꼽을 수 있다. 반면, (그림 4)의 c와 d는 각각 높은 출현 빈도 순위에도 불구하고 LCC포함 횟수가 적거나 낮은 출현 빈도 순위에도 불구하고 LCC포함 횟수가 많은 경우라 할 수 있다. 조사 결과, (그림 4)의 a, b, c, d는

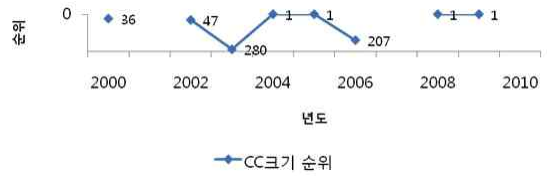


(그림 4) 출현 빈도 상위 200개 키워드의 LCC 포함 횟수

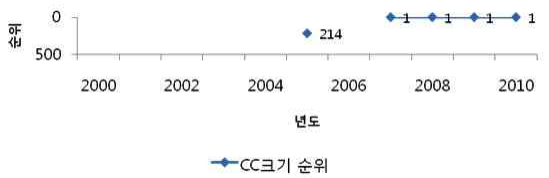
각각 "XML", "DEVS(Discrete Event System Specification)[4]", "IPTV", "Similarity"에 해당하였다. (그림 5)는 그 네 개 키워드 각각에 대해 연도별로 구축된 키워드 연관 네트워크에서 해당 키워드가 포함된 CC크기 순위의 변화를 나타낸 것이다[5]. (그림 5(a))와 (그림 5(b))는 키워드 "XML"(출현 빈도수 1위)과 "DEVS"(출현 빈도수 197위)에 대한 결과로서, "XML"의 경우 2001년도를 제외하고는 항상 LCC에 속했던 반면 "DEVS"은 11회 중 4회만 LCC에 속했음을 알 수 있다. 한편, (그림 5(c))와 (그림 5(d))는 키워드 "IPTV"(출현 빈도수 59위)와 "Similarity"(출현 빈도수 200위)에 대한 연도별 CC크기 순위 변화의 결과로서, (그림 4)의 선형 회귀선의 경향에 잘 맞지 않는 예이다. "IPTV"의 경우 2000년부터 2006년까지는 LCC에 포함되지 않았지만, 그 이후에는 계속 LCC에 포함되고 있는데, 이는 "IPTV" 관련 연구가 최근에 발전하기 시작했기 때문인 것으로 유추할



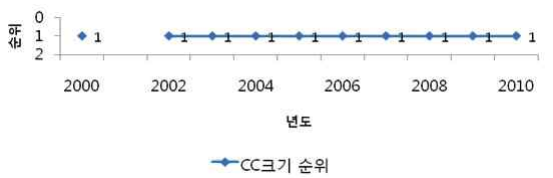
(a) "XML"이 포함된 CC의 크기 순위 변화



(b) "DEVS"가 포함된 CC의 크기 순위 변화



(c) "IPTV"가 포함된 CC의 크기 순위 변화



(d) "Similarity"가 포함된 CC의 크기 순위 변화

(그림 5) 키워드 연관 네트워크에서 키워드가 포함된 CC의 크기 순위 변화

3) 따라서 y값의 범위는 0이상 11이하의 정수이다.

4) 연속 시간 사건을 이산시간 사건으로 표현하는 방법론

5) CC크기는 CC에 포함된 정점의 개수를 뜻하며, 순위가 표시되지 않은 연도는 크기가 2이상인 CC에 포함되지 않았음을 나타낸다.

수 있다. 반면 "Similarity"는 200번째 빈도순위를 가지는 키워드임에도 2001년을 제외하고 항상 LCC에 포함되었음을 알 수 있다. 이는 "Similarity"가 키워드로 포함되는 연구(주로 두 데이터의 유사도에 관한 연구)가 어떤 특정 기술이나 분야에 한정되어 사용되는 것이 아니라 매우 다양한 분야에 두루 관련이 있음을 유추할 수 있다. (그림 5)의 결과처럼 개별 키워드의 연도별 LCC포함 여부나 CC크기 순위를 조사하는 것은 해당 키워드의 성격과 연구 동향을 살필 수 있는 하나의 유용한 방법이 될 수 있다. 즉, 단순히 키워드의 사용 빈도만을 가지고 관련 분야의 연구 활성화 정도를 파악하기보다는 키워드가 속한 CC의 크기를 파악함으로써 연계된 연구 분야 전체의 활성화 정도와 시간에 따른 변화를 알 수 있고, 그러한 연구 동향의 이유를 추론해 볼 수 있다.

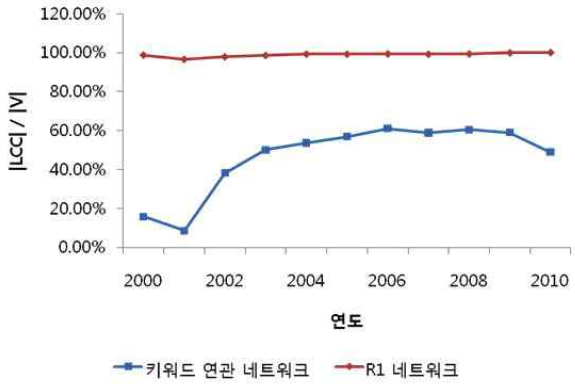
4.3 LCC 크기 분석

4.3절과 4.4절에서는 키워드 연관 네트워크 구조적 특징에 대한 검증을 위해 비교군으로서 기존의 네트워크 비교연구에서처럼 다수의 랜덤 네트워크를 생성한다[9]. 이에 먼저 랜덤 네트워크 생성 방법에 대해서 설명한다. 본 실험을 위해서 두 가지 종류의 랜덤 네트워크를 생성하는 데, 하나는 네트워크의 간선들은 그대로 둔 채 간선들의 가중치를 뒤섞는 가중치섞임 랜덤 네트워크(weight-shuffled random network)이고 다른 하나는 간선들이 연결하고 있는 정점들 대신 임의의 새로운 정점들로 교체하는 방식으로 생성하는 정점섞임 랜덤 네트워크(vertex-shuffled random network)이다. 두 종류 랜덤 네트워크 모두 원래의 네트워크와 같은 정점들의 집합과 같은 개수의 간선으로 구성된다. 또한 랜덤 네트워크를 생성하는 데 있어서 뒤섞임(shuffling)의 범위를 전체 네트워크뿐만 아니라 LCC 부분 네트워크로 한정하기도 하였다. 이렇게 하여 <표 2>와 같은 4가지 종류의 랜덤 네트워크를 시뮬레이션을 위해 고려한다.

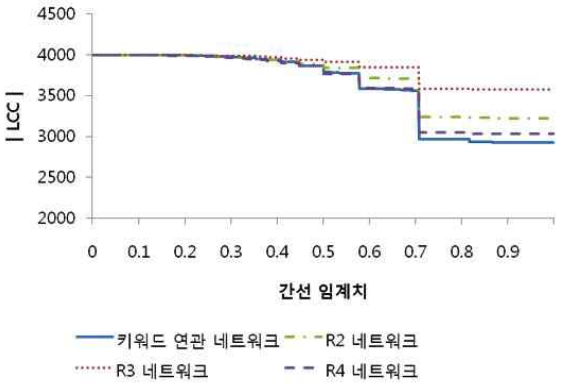
〈표 2〉 랜덤 네트워크의 종류

랜덤 네트워크	뒤섞임 대상	뒤섞임 범위
R_1	정점	전체 네트워크
R_2	정점	LCC
R_3	가중치	전체 네트워크
R_4	가중치	LCC

이제 키워드 연관 네트워크의 LCC크기가 랜덤 네트워크와 어떤 차이가 있는지를 살펴본다. 이를 위해 키워드 연관 네트워크의 LCC의 크기를 R_1 랜덤 네트워크의 LCC크기와 비교하였다. (그림 6)은 키워드 연관 네트워크의 LCC 크기, 100개 R_1 랜덤 네트워크의 LCC 평균 크기 및 표준편차를 연도별로 그린 것이다(간선임계값=0.0). 그 그림에서 보듯이, 키워드 연관 네트워크의 LCC 크기가 랜덤 네트워크의 그것에 비해 매우 작음을 알 수 있다 (모든 연도의 p-value<0.01). 키



(그림 6) 키워드 연관 네트워크와 R1 랜덤 네트워크의 연도별 $\frac{|LCC|}{|V|}$ 변화 비교



(그림 7) 간선임계값 변화에 따른 LCC 크기 변화

워드 연관 네트워크에서 간선의 존재는 해당 키워드 쌍이 적어도 한 편 이상의 연구 논문에 동시에 출현하였음을 의미한다. 따라서, 키워드 연관 네트워크의 LCC 크기가 랜덤 네트워크의 그것에 비해 작다는 사실은 키워드들 사이의 연관성이 무작위적으로 존재하는 것이 아니라, 매우 집약적으로 존재함을 의미한다.

4.4 간선임계값 변화에 따른 LCC 크기 변화 분석

키워드 연관 네트워크에서 LCC는 여러 소규모 분야들에서 사용된 공통 키워드들이 연결되면서 형성된 것이라 볼 수 있다. 그러한 LCC를 구성하는 간선들의 가중치들이 어떤 특징적인 분포를 보이는지를 살펴보았다. 이를 위해, 간선임계값을 0부터 1까지 0.1단위로 변화시키는 동안 LCC의 크기 변화를 조사하였다. (그림 7)은 간선임계값 변화에 따른 2009년도 키워드 연관 네트워크의 LCC의 크기 변화와 100개씩 임의로 생성한 R_2, R_3, R_4의 평균 LCC의 크기 변화를 비교한 것이다. 간선임계값이 0일 때, 키워드 연관 네트워크와 랜덤 네트워크의 LCC 크기는 3993으로 동일하다[6]. 그러나, 간선임계값이 증가함에 따라 키워드 연관 네트워크

6) R_1의 경우는 간선임계치가 0일 때 LCC의 크기가 달라지므로 비교하지 않았다.

의 LCC크기가 R_2, R_3, R_4의 그것에 비해 더 급격하게 감소하는 것을 확인할 수 있다. 특히, 간선임계값이 0.9일 때, 키워드 연관 네트워크의 LCC크기는 2929로서 R_2, R_3, R_4의 평균 LCC크기(각각 3047.39, 3588, 3222.16)보다 통계적으로 유의하게 작았다 (모든 경우의 p-value<0.01). 이러한 결과는 키워드 연관 네트워크의 LCC에서 높은 가중치를 가지는 간선들이 밀집해 있음을 뜻한다. 좀 더 구체적으로 살펴보면, 전체 네트워크에서 가중치에 대해 뒤섞은 R_3의 평균 |LCC|가 가장 큰 것으로 보아 LCC를 구성하는 간선들의 가중치들이 LCC 이외의 CC들을 구성하는 간선들의 가중치보다 값이 평균적으로 작다는 것을 암시한다. 즉, LCC이외의 CC들이 오히려 높은 유사도의 키워드 쌍들로 구성된다는 흥미로운 결과를 암시한다. R_2와 R_4의 평균 |LCC| 크기가 R_2의 그것보다 작은 이유는 뒤섞임의 범위가 LCC로 한정되었기 때문이다 (즉, LCC이외의 CC들은 변화가 없다). 키워드 연관 네트워크의 |LCC| 크기가 R_2의 그것보다 작은 것은 LCC의 구조가 통계적으로 특징적인 분포를 보임을 뜻하며, R_4의 그것보다 작은 것은 그러한 특징적인 구조 내에서도 높은 가중치의 간선들이 어느 정도 클러스터를 이루고 있음을 암시한다. 이처럼 키워드 연관 네트워크는 구조적 측면과 높은 가중치 간선들의 분포 측면에서 랜덤 네트워크와는 다르게 형성되었음을 알 수 있다.

5. 결 론

본 논문에서는 국내 컴퓨터공학 분야 학술 논문에 포함된 키워드 정보를 바탕으로 2000년부터 2010년까지 연도별 키워드 연관 네트워크 기반 지식지도를 구축하였다. 개별 키워드가 포함된 연결 요소의 크기 변화를 살펴봄으로써 관련 연구 분야의 동향을 추론할 수 있는 정보를 제공할 수 있음을 보였다. 또한, 랜덤 네트워크와의 비교를 통해 키워드 연관 네트워크에서 최대 연결 요소가 매우 크기가 작으면서 높은 유사도의 키워드 쌍들이 상당히 밀집되어 있음을 보였다. 이는 소규모 연구 주제들이 클러스터를 이루면서도 서로 간에 유사도가 비교적 작은 키워드 쌍에 의해 느슨하게 연결되어 있음을 추론케 한다. 이처럼 키워드 연관 네트워크를 활용한 지식지도는 키워드 빈도와 같은 분석에 의한 단편적인 연구 동향에 대한 지식보다 다양한 정보를 제공해 줄 수 있다.

앞으로의 연구는 네트워크의 클러스터링 구조를 파악할 수 있는 연구를 통해 LCC의 구조를 구체적으로 분석해 보는 것이 될 것이다. 현재까지 연구된 방법들을 살펴보면 [11]과 같은 네트워크안의 구조를 알아보는 연구가 있다. 그 연구에서는 네트워크를 거리가 가까운 정점들끼리 군집화하는 방법과 네트워크에서 매개 중심성(betweenness centrality)을 계산하여 가장 작은 정점에 연결된 간선을 제거하는 방식의 방법을 통해 군집화한다. 이러한 네트워크의 클러스터링은 본 연구에서 추정했던 LCC의 구조에 대해 보다 정확한 이해를 도울 것이다. 또한, 키워드 연관 네트워크의 특정 연결 요소가 시간적으로 어떻게 형성되었는지를 계량적으로 분석할 수 있는 알고리즘의 개발을 통해 특정 연구 주제나 분야가 시간적으로 어떻게 형성되어 왔는지를 분석할 수 있는 연구가 필요할 것이다. 마지막으로 본 논문의 연구 결과를 어떻게 활용할 수 있을 지에 대한 연구도 필요할 것이다.

참 고 문 헌

[1] 이광희, “지식지도 작성을 위한 기초연구”, 한국학술진흥재단, 2007.

[2] 원동규, “사회과학분야 학술연구 지식지도(knowledge map)의 개발 및 구현”, 한국학술진흥재단, 2009.

[3] 서의호, 유기동, “사례적용을 통한 지식지도 작성방법론 연구”, 한국경영과학회 학술대회논문집, pp.337-340, 2000.

[4] Remko Helms, Kees Buijsrogge, “Knowledge Network Analysis : a technique to analyze knowledge management bottlenecks in organizations”, Proceedings of the 16th International Workshop on Database and Expert Systems Applications, pp.410-414, 2005.

[5] Wang Zhiquan, Liu Jinhao, Hou Dongliang, Miao Rui,. “Knowledge Network System Building and Realization” 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, pp.336-340, 2009.

[6] Richard M. Adler, “A Dynamic Social Network Software Platform for Counter-Terrorism Decision Support”, 2007 IEEE Intelligence and Security Informatics, pp.47-54, 2007.

[7] Gueorgi Kossinets, Duncan J.Watts, “Empirical Analysis of an Evoving Social Network”, Science 6 January, Vol.311, No.5757, pp.88-90, 2006.

[8] Fowler James H, Nicholas A. Christakis, “Dynamic Spread of Happiness in a Large Social Network : Longitudinal Analysis Over 20 Years in the Framingham Heart Study”, British Medical Journal, Vol.337, No.a2338, pp.1-9, 2008.

[9] A.L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek, “Evolution of the social network of scientific collaborations”, Physica A: Statistical Mechanics and its Applications, Vol.311, No.3-4, pp.590-614, 2002.

[10] H. Zhang, B. Qiu, K. Ivanova, C. Lee Giles, H. C, Foley, J. Yen, “Locality and Attachedness-Based Temporal Social Network Growth Dynamics Analysis: A Case Study of Evolving Nanotechnology Scientific Collaboration Networks”, Vol.61, No.5, pp.964-977, 2010.

[11] M. E. J. Newman, M. Girvan, “Finding and evaluating community structure in networks”, Physical Review E, Vol.69, No.2, pp.69-83, 2004.

정 보 석

e-mail : wruwami@mail.ulsan.ac.kr
2009년 울산대학교 컴퓨터공학부(학사)
2009년~현 재 울산대학교 전기공학부 석사과정
관심분야 : 복잡계산시스템, 소셜네트워크 등

권 영 근

e-mail : kwonyk@ulsan.ac.kr
1999년 서울대학교 전산과학과(학사)
2001년 서울대학교 컴퓨터공학부(공학석사)
2006년 서울대학교 컴퓨터공학부(공학박사)
2008년~현 재 울산대학교 전기공학부 조교수
관심분야 : 최적화 이론 및 실제, 복잡계산시스템, 시스템생물학, 소셜네트워크 등

곽 승 진

e-mail : sjkwak@cnu.ac.kr
1990년 성균관대학교 문헌정보학과(학사)
1995년 성균관대학교 문헌정보학과(석사)
2004년 성균관대학교 문헌정보학과(박사)
1994년~2004년 LG연암문화재단 LG상남 도서관 팀장
2009년~2009년 University of South Carolina, Visiting Scholar
2004년~현 재 충남대학교 문헌정보학과 부교수
관심분야 : 디지털도서관, 정보시스템, 메타데이터 등