

# Understanding of Structural Changes of Keyword Networks in the Computer Engineering Field

Yung-Keun Kwon<sup>†</sup>

## ABSTRACT

Recently, there have been many trials to analyze characteristics of research trends through a structural analysis of keyword networks in various fields. However, most previous studies have mainly focused on structural analysis harbored in some static networks and there is a lack of research on changes of such networks structure with time. In this paper, we constructed annual keyword networks by using a database of papers published in the international computer engineering-field journals from 2002 through 2011, and examined the changes of them. As a result, it was shown that most keywords in a network are preserved in the network of the next year, and their degree of connectivity and the average weight of the connections were higher and smaller, respectively, than those of the keywords which are not preserved. In addition, when a keyword network shifted to one of the next year, the connections between keywords were more likely to be removed than preserved, and the average weight of the removal connections was higher than that of the preserved ones. These results imply that the keywords are not changed over time but their connections are very likely to be changed; and there is apparent differences between the preserved and removal groups of keywords/connections with respect to degree and weights of connections. All these results are consistently observed over the ten-year datasets and they can be important principles in understanding the structural changes of the keyword networks.

**Keywords :** Keyword Network, Network Structure Change, Research Trend Analysis

## 컴퓨터공학 분야 키워드네트워크의 구조적 변화 이해

권영근<sup>†</sup>

### 요 약

최근 여러 분야에서 키워드네트워크의 구조 분석을 통해 연구동향의 특징을 분석하는 시도가 있어 왔다. 하지만 대부분의 기존 연구는 주로 정적인 네트워크의 구조 분석에 집중하였으며 시간에 따라 네트워크 구조가 어떻게 변화하는지에 대한 연구는 부족하였다. 본 논문에서는 2002년부터 2011년까지 컴퓨터 공학 분야의 해외 학술지에 게재된 논문들의 데이터베이스를 활용하여 연도별 키워드네트워크를 구축하고, 구조적 변화를 조사하였다. 그 결과, 키워드네트워크에서 대부분의 키워드는 다음 연도에서도 잔존하며 제거되는 키워드에 비해 연결(부속된 간선)의 차수는 크지만 평균 강도는 약한 특징을 보였다. 또한, 다음 연도의 키워드네트워크로 변화할 때, 키워드들 사이의 연결은 잔존되기보다 제거되는 비율이 높았으며 제거되는 연결들의 강도가 더 큰 특징을 보였다. 이러한 결과들은 연구 분야를 대변하는 키워드 자체의 변화는 작지만 그들 사이의 관계는 크게 변화하며 잔존 또는 제거되는 키워드 및 간선 그룹 사이에는 연결의 차수나 강도 측면에서 큰 차이가 존재함을 뜻한다. 본 논문의 분석결과들은 10년 동안의 데이터에 대해서 일관되게 관찰되었으며 이는 컴퓨터공학 분야 키워드네트워크 변화를 이해하는 데 중요한 원리임을 암시한다.

**키워드 :** 키워드네트워크, 네트워크 구조 변화, 연구 동향 분석

## 1. 서 론

오늘날 정보통신기술의 발전으로 정보의 생산, 전달, 교환이 촉진되어 정보와 지식의 양이 급증함에 따라 대량의 정보를 활용하여 사용자에게 필요한 정보를 전달하는 기술이 요구되고 있다. 이러한 정보처리기술은 연구나 특허의 동향 분석에도 적용될 수 있는데 활용 목적이나 방법에 따라 다양한 연구가 진행되어 왔다. 가장 기본적으로 사용된 방법

\* 이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2010-32A-H00006).

<sup>†</sup> 정 회 원 : 울산대학교 전기공학부 조교수

논문접수 : 2012년 11월 14일

수정일 : 1차 2013년 1월 16일

심사완료 : 2013년 1월 16일

\* Corresponding Author : Yung-Keun Kwon(kwonyk@ulsan.ac.kr)

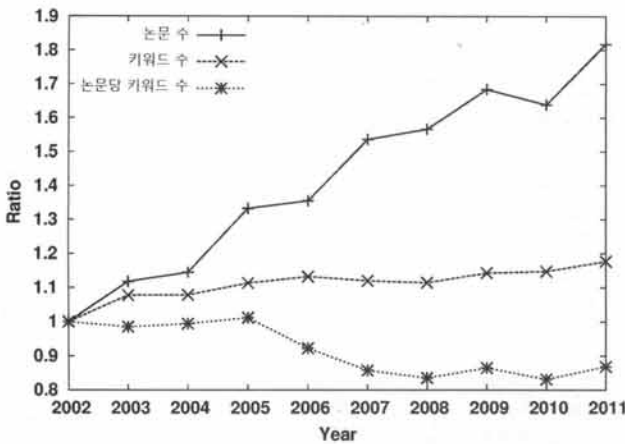


Fig. 1. Ratio changes of the numbers of papers, keywords, and keywords per paper compared over those in 2002

은 연구자, 연구기관, 키워드 등의 조사 대상에 대해 그 사용빈도를 기반으로 유용한 정보를 도출하는 것이다. 예를 들면, 논문의 제목에 자주 등장하는 단어나 키워드의 빈도를 도출하여 최근의 연구 주제[1]나 해당 분야의 특징을 이해하는 방식이다[2]. 비록 그러한 분석이 개별적인 조사 대상의 중요도를 파악하는 데 도움이 되지만 그들 사이의 종합적인 관계를 이해하는 데에는 한계가 있다. 이러한 단점을 보완하기 위하여 최근 네트워크 기반의 분석이 빈번하게 시도되고 있다. 네트워크 기반 분석을 적용하기 위해서는 일반적으로 조사 대상 사이의 관계 여부 혹은 그 세기를 계산하여 네트워크를 구축하여야 한다. 이를 위해 논문이나 특허의 분류 체계에 따른 유사성[3], 상호 참조 관계 여부에 따른 유사성[4-6], 혹은 공통 키워드나 단어의 사용에 따른 유사성[7-13] 등이 자주 활용되었다. 특히 그 중에서 키워드 네트워크는 키워드가 어떤 연구 분야의 내용을 함축적으로 잘 표현한다는 점에서 연구 분야의 특징이나 연구 동향 분석을 위해 자주 사용되고 있다. 예를 들면, 연결이 가장 많은 키워드들 조사를 통해 연구 동향을 분석하거나 키워드의 차수분포 조사를 통해 인기 키워드의 지속성을 분석하였다[8]. 또한, 바이오 네트워크나 소셜 네트워크 등 복잡네트워크의 중심성 분석에 자주 이용되어 온 평균최단거리, 군집계수, betweenness, closeness 등 정량화된 중심성 척도를 이용하여 구조적 특징을 분석하거나 [7, 9, 14], 랜덤네트워크와의 통계적인 비교를 통해 최대연결요소 등의 실제 키워드 네트워크의 구조가 얼마나 특이한지를 설명하기도 하였다[13]. 이러한 기존의 연구들은 정적 키워드네트워크를 구축한 뒤 주로 어떤 구조적 특징이 있는지를 분석하고 그 의미를 추정해 보는 것이었다.

하지만 키워드네트워크는 시간에 따라 변화하게 되는데 그러한 구조 변화는 연구 동향의 변화를 나타낸다는 점에서 구조 변화의 이해는 해당 연구 분야의 특징을 이해하기 위해 중요한 문제이다. 따라서 본 연구에서는 정적 키워드네트워크의 구조 분석에 집중하였던 기존의 연구와는 달리 키워드네트워크의 구조가 어떤 형태로 변화하는 지에 관한 분

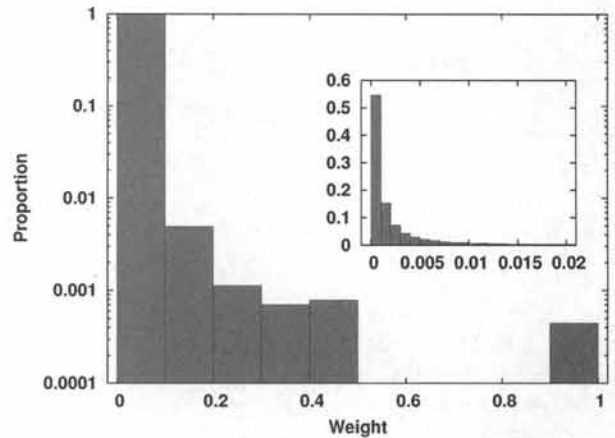


Fig. 2. Distribution of edge weights in the keyword network in 2011

석을 시도한다. 즉, 시간이 흐름에 따라 키워드네트워크에서 어떤 정점과 간선이 소멸되거나 잔존하게 되는지를 파악함으로써 구조 변화의 특징을 살펴보고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존 관련 연구와 본 논문에서 컴퓨터공학 분야 키워드 연관 네트워크를 구축하는 방법 등의 배경 지식에 대해 설명하고, 3장에서는 키워드네트워크 구조 변화에 관한 다양한 분석 결과를 보인다. 그리고 4장에서 결론을 맺는다.

## 2. 배경 지식

본 논문에서는 컴퓨터공학 분야에 대한 키워드네트워크 구축 및 분석을 시도한다. 이 절에서는 기존의 키워드네트워크 관련 연구를 살펴본 뒤, 본 연구에서 해외 학술논문 데이터베이스를 활용하여 키워드네트워크를 구축한 방법을 설명한다.

### 2.1 키워드네트워크관련 최근 연구

논문이나 특허문서에서 사용된 키워드들 사이의 관계를 네트워크의 형태로 표현한 것을 키워드네트워크라고 부르며 이러한 키워드네트워크는 어떤 연구 분야의 특징을 기술하거나 연구 동향을 분석하는 데 활용될 수 있다. 예를 들면, 논문에서 함께 사용된 모든 키워드 쌍에 대해 링크를 연결하여 네트워크를 구축한 뒤, 연결이 가장 많은 키워드들을 연도별로 파악함으로써 연구 동향을 파악하거나 개별 키워드의 차수(degree)가 멱급수 분포를 보임으로써 인기 키워드가 논문에서 계속 사용되는 경향이 있음을 보이기도 하였다[8]. 또한, 키워드네트워크의 구조를 평균최단거리, 군집계수 등 정량화된 척도를 이용하여 평균최단거리가 매우 커서 작은 세상(small-world) 속성은 없지만 계층화된 무척도네트워크임을 보이거나[14], 더 나아가 betweenness, closeness 등 다양한 복잡네트워크 구조 분석 척도들을 해당 분야의 특징 분석의 지표로 사용하여 분석된 연구 분야가 잘 확립된 소연구분야로 잘 구분될 수 있음을 보이기도

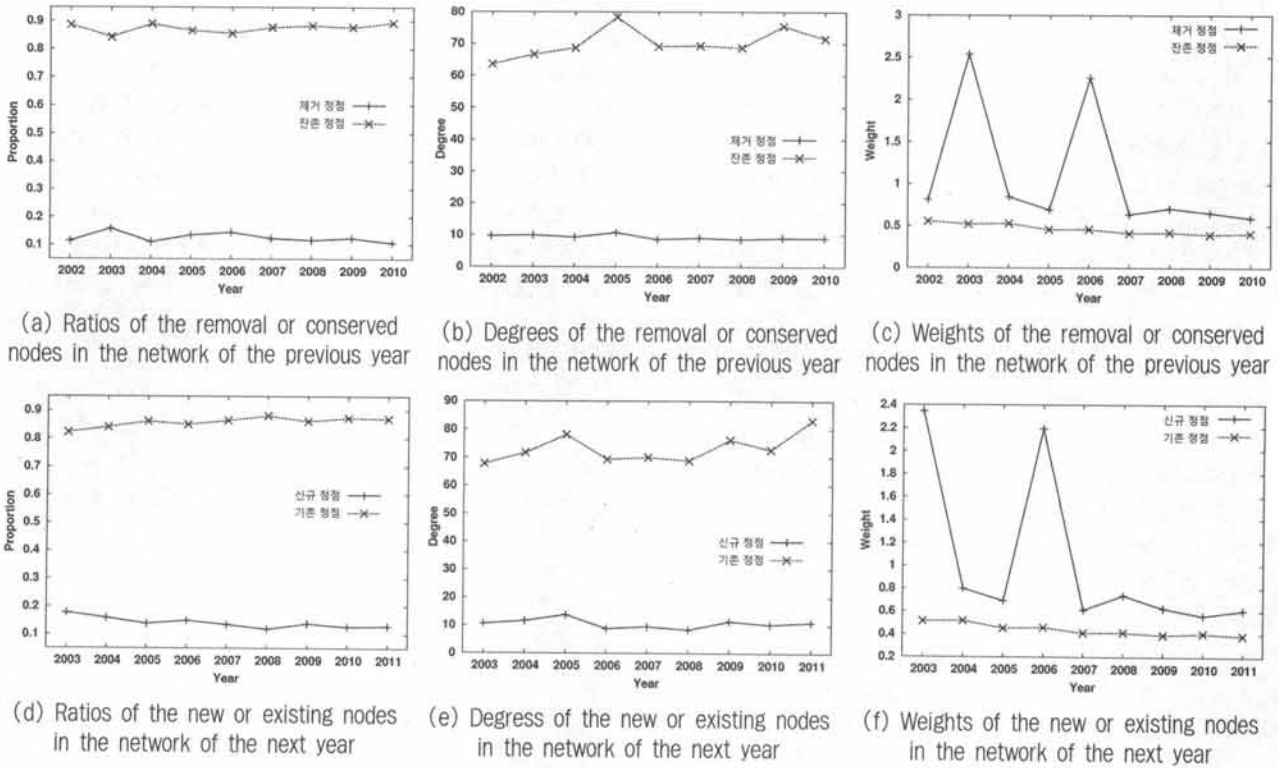


Fig. 3. Structural changes with respect to nodes in the keyword networks

하였다[9]. 혹은 키워드네트워크 내 최대연결요소가 랜덤네트워크와 얼마나 다른지를 비교함으로써 해당 키워드네트워크가 임의로 조직된 것이 아님을 보이기도 하였다[13]. 한편, 원래의 키워드네트워크를 분석하지 않고 그것을 군집화하여 몇 개의 작은 네트워크로 쪼갬 뒤 그들의 중심성과 밀도 등을 측정함으로써 해당 연구 분야를 소분야로 분류하기도 하였다[7].

이러한 기존의 연구들에서는 논문에서 키워드의 동시 출현 여부를 링크로 표현하여 키워드네트워크를 구축한 뒤 차수의 분포, 평균최단거리, betweenness, closeness 등과 같은 구조적 특징을 살펴보고 그 결과의 의미를 해석하였다. 이러한 연구들은 특히 네트워크에서 어떤 키워들 (혹은 키워드 그룹)이 중요한 지를 파악하거나 구분하기 위한 정적 네트워크의 구조 분석에 관한 것이라 할 수 있다. 하지만 키워드네트워크는 시간에 따라 변화하게 되는데 기존의 연구에서는 키워드네트워크의 구조가 어떤 형태로 변화하는 지에 관한 분석이 부족하였다. 이러한 구조의 동적 변화에 관한 이해는 연구 동향을 파악하고 해당 연구 분야의 특징을 이해하는 데 도움이 된다는 점에서 중요하다. 이에 본 연구에서는 기존의 정적 구조 분석과 달리 키워드네트워크 구조 변화 분석에 초점을 둔다.

2.2 키워드네트워크 구축을 위한 데이터베이스

본 연구에서는 컴퓨터공학 분야 키워드네트워크 구축을 위해 IEEE 전자도서관 웹사이트(<http://ieeexplore.ieee.org>)

에서 제공하는 논문들을 활용한다. 이를 위해 2002년부터 2011년까지 “Computing & Processing (Hardware/Software)” 분야로 분류되어 있는 저널들에 게재된 183,936 편의 논문을 이용하였다. 또한, 그 논문집합에서 서로 다른 키워드를 추출한 결과 총 키워드의 개수는 8,905개였다.

2.3 키워드네트워크 생성 방법

본 논문에서 키워드네트워크는 무방향 가중치 그래프 (undirected weighted graph)  $G(V, E)$ 로 표현될 수 있는데, 정점  $v \in V$ 는 각각의 키워드를 나타내고 간선  $(v_i, v_j)$ 의 가중치  $w(v_i, v_j)$ 는 해당 키워드 쌍  $(v_i, v_j)$ 의 유사도를 나타낸다. 본 연구에서 간선의 가중치는 두 키워드가 얼마나 빈번히 논문에 동시에 출현했는지를 상대적으로 측정하기 위해 아래와 같은 코사인 유사도[13] 값으로 정의되는데, 이 식에서  $P(v)$ 는 키워드  $v$ 가 포함된 논문들의 집합을 표현한다.

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

3. 키워드네트워크 분석 결과

3.1 키워드네트워크 구축 데이터 분석

Fig. 1은 논문 수, 키워드 수, 논문당 키워드 수의 2002년 대비 비율의 변화를 연도별로 나타낸다 (2002년 논문 수, 키

워드 수, 논문당 키워드 수는 각각 12,960편, 5,484개, 6.65개/편이었다). 그 그림에서 보이듯이, 논문 수는 지속적으로 상승하여 2011년 논문 수가 2002년 대비 약 82% 증가하였다. 하지만, 2011년 키워드 수는 2002년 대비 약 18% 증가하는 데 그쳤는데 이는 논문 수의 증가에 비해 키워드 종류의 증가는 매우 작음을 알 수 있다. 또한, 키워드의 수가 키워드네트워크에서 정점의 개수( $V$ )를 의미함을 고려할 때 키워드네트워크의 크기는 소폭 증가하고 있음을 뜻한다. 한편, 논문당 키워드 수는 최근들어 다소 감소하고 있음을 알 수 있다 (2011년 논문당 키워드 수는 5.78개/편이었다).

Fig. 2는 2011년 키워드네트워크에 포함된 237,996 개의 간선들에 대해서 가중치의 분포를 나타낸 것이다. 그 그림에서 보이듯이, 대부분의 간선 가중치가 0.1이하였으며 (236,100개의 간선), 매우 드물게 가중치가 0.9~1.0의 범위에 속하였다(106개). 더 나아가서, 가중치가 0.001이하인 간선이 약 55%를 차지했다. 가중치 0.001은 두 키워드가 각각 약 30편의 논문에 등장할 때, 1편의 논문에서 같이 등장하는 정도의 유사도를 의미함을 감안할 때, 키워드 네트워크에서 연결의 세기가 비교적 약한 특징을 보인다고 할 만 하다. 한편, 다른 연도의 키워드네트워크 역시 그림 2와 비슷한 경향을 보였다.

### 3.2 키워드네트워크 구조의 연도별 변화

Fig. 1에서 최근 논문의 수는 증가하고 있지만 키워드 수는 크게 증가하고 있지 않음을 알 수 있었다. 또한 Fig. 2에서는 그러한 키워드 네트워크의 간선들의 연결 강도가 크지 않음을 보였다. 이러한 결과들은 키워드네트워크 구조 변화의 폭과 특징에 대한 궁금증을 야기한다.

그러한 궁금증을 풀기 위해 먼저 키워드네트워크의 정점에 관한 연도별 변화를 분석하였다(Fig. 3). 기준 연도의 키워드네트워크의 정점이 차년도 키워드네트워크에서 사라지는 경우에는 제거 정점, 계속 존재하는 경우에는 잔존 정점이라 부른다. Fig. 3(a)는 기준 연도를 2002년부터 2010년까지 바꿔 가면서 제거 정점과 잔존 정점의 개수 비율을 비교한 결과이다. 그 그림에서 보이듯이, 대부분의 정점은 차년도 키워드네트워크에서 제거되지 않고 잔존함을 알 수 있으며, 이러한 경향은 기준 연도에 따라 크게 다르지 않음을 알 수 있다. Fig. 3(b)와 3(c)는 제거/잔존 정점들의 평균 차수(degree: 정점에 부속된 간선<sup>1)</sup>의 개수로 정의됨)와 부속된 간선들의 평균 가중치를 각각 비교한 결과이다. 흥미롭게도 잔존 정점들에 비해 제거 정점들은 차수는 더 작으면서도 평균 가중치는 더 크다는 사실을 알 수 있으며, 이러한 경향은 모든 연도에서 일관되게 성립한다. 이러한 결과는 현재의 키워드네트워크에서 다른 키워드와의 연결도가 비교적 작지만 높은 가중치를 갖는 간선들에 부속된 정점들이 제거가 될 가능성이 높음을 뜻한다. 한편, Fig. 3(d)-(f)는 기준 연도의 키워드네트워크의 정점이 전년도 키워드네트워크 대비 새롭게 나타난 경우에는 신규 정점, 계속 존재했던 경우

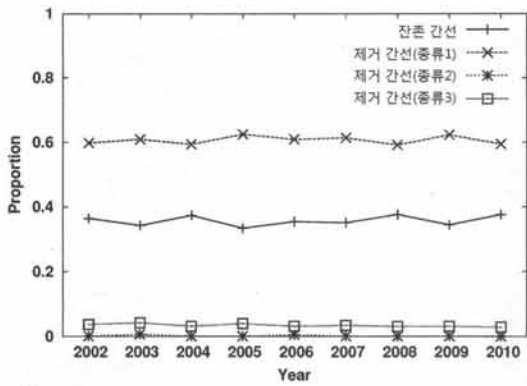
에는 기존 정점이라 분류했을 때, 그 두 가지 정점 그룹에 대한 개수의 비율, 평균 차수, 부속 간선들의 평균 가중치를 보여준다. 앞서 결과와 유사하게 신규 정점은 기존 정점에 비해 그 비율이 낮고 차수는 작으며 부속 간선의 가중치 평균이 크다. 또한, 그러한 특징은 모든 연도에서 관찰된다.

다음으로 키워드네트워크의 간선에 대해 연도별 변화를 분석하였다(Fig. 4). 기준 연도의 키워드네트워크의 간선이 차년도 키워드네트워크에서 사라지는 경우에는 제거 간선, 계속 존재하는 경우에는 잔존 간선이라 부른다. 또한, 제거 간선( $v, w$ )는  $v$ 와  $w$ 가 차년도 키워드네트워크에서 제거 정점인지 여부에 따라 세분화된다. 즉,  $v$ 와  $w$ 가 모두 잔존하는 경우("종류1"),  $v$ 와  $w$ 가 모두 제거되는 경우("종류2"),  $v$ 와  $w$  중 하나만 제거되는 경우("종류3") 등 3가지로 분류된다. Fig. 4(a)는 기준 연도를 2002년부터 2010년까지 바꿔 가면서 세 종류의 제거 간선과 잔존 간선의 개수 비율을 비교한 결과이다. 그 그림에서 보이듯이, 제거 간선(종류1)과 잔존 간선이 대부분을 차지하며 제거 간선(종류2)와 제거 간선(종류3)은 매우 낮은 비율을 차지한다. 또한, 이러한 경향은 연도에 따라 크게 다르지 않음을 알 수 있다. 특히, 제거 간선(종류1)이 약 60%를 차지함으로써 가장 비중이 높는데, Fig. 3(a)와 3(d)의 결과를 고려할 때, 그것은 정점이 매우 높은 확률로 잔존할 수 있지만 간선은 비교적 쉽게 제거됨을 뜻한다. Fig. 4(b)는 세 가지 종류의 제거 간선 그룹과 잔존 간선 그룹에 대해 가중치 평균을 비교한 것이다. 그 그림에서, 잔존 간선의 평균 가중치가 가장 작고 제거 간선(종류1)이 다음으로 작음을 알 수 있다. 반면, 제거 간선(종류2)가 가장 높은 평균 가중치를, 제거 간선(종류3)이 다음으로 높은 평균 가중치를 보여준다. Fig. 4(a)와 4(b)의 결과를 종합해 보면, 현재의 키워드네트워크에서 대부분의 간선은 차년도에 잔존하게 되는 정점들 사이에 존재하지만 잔존하는 간선들이 제거되는 간선들에 비해 가중치가 낮다고 할 수 있다. 또한, 차년도 키워드네트워크에서 제거되는 정점에 부속된 제거 간선들은 그 비중이 낮지만 매우 가중치가 높다는 특징을 보인다. 이는 일반적으로 키워드 쌍의 유사도(즉, 가중치)가 높을 때 그 관계가 매우 강하다고 예상하지만, 그러한 관계가 오히려 오래 지속되지 못하며 해당 키워드의 생명력이 더 중요함을 시사한다. 한편, Fig. 4(c)와 4(d)에서는 기준 연도의 키워드네트워크의 간선이 전년도 키워드네트워크 대비 계속 존재했던 경우에는 기존 간선, 새롭게 나타난 경우에는 신규 간선이라 부른다. 또한, 신규 간선( $v, w$ )는  $v$ 와  $w$ 가 전년도 키워드네트워크로부터 신규 정점인지 여부에 따라 세분화된다. 즉,  $v$ 와  $w$ 가 모두 기존 정점인 경우("종류1"),  $v$ 와  $w$ 가 모두 신규 정점인 경우("종류2"),  $v$ 와  $w$  중 하나만 신규 정점인 경우("종류3") 등 3가지로 분류된다. Fig. 4(a)와 4(b)에서의 결과가 매우 유사한 결과임을 확인할 수 있으며, 모든 연도에서 일관되게 관찰된다.

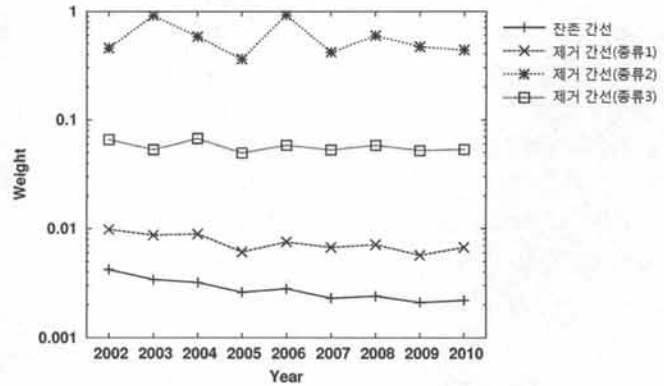
### 3.3 키워드네트워크의 뼈대 분석

앞의 절에서는 매년 키워드네트워크의 구조적 변화에 대해서 살펴보았다. 그러한 구조적 변화는 일종의 섭동

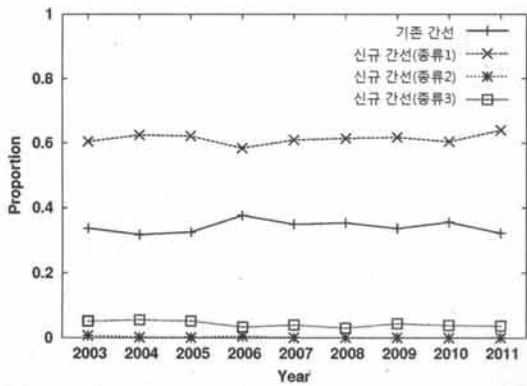
1)  $G(V, E)$ 에서 간선( $v, w$ )는 정점 $v$ 에 부속되어 있다고 부른다.



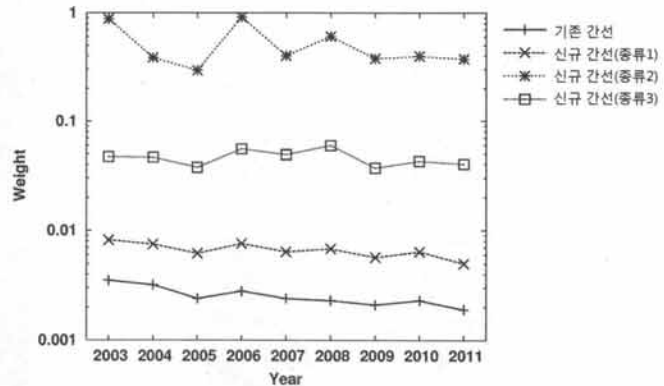
(a) Ratios of the removal or conserved edges in the network of the previous year



(b) Weights of the removal or conserved edges in the network of the previous year

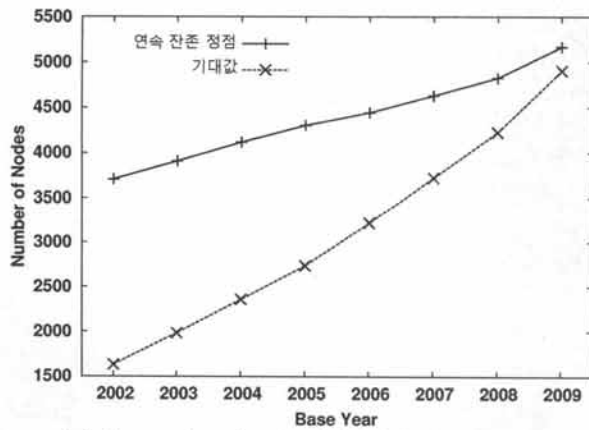


(c) Ratios of the new or existing edges in the network of the next year

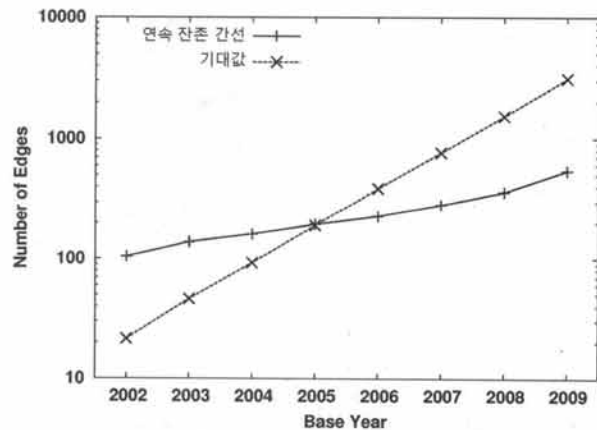


(d) Weights of the new or existing edges in the network of the next year

Fig. 4. Structural changes with respect to edges in the keyword networks



(a) The number of nodes included in the skeleton



(b) The number of edges included in the skeleton

Fig. 5. Numbers of nodes and edges included in the skeleton of the keyword networks

(perturbation)으로 이해할 수 있으며, 연속적인 섭동에 대해서 유지되는 네트워크의 구조를 그 네트워크의 뼈대 (skeleton)로 정의할 수 있다. 좀 더 구체적으로,  $y$ 연도 키워드네트워크를  $G(V_y, E_y)$ 라고 할 때, 기준연도(base year)가  $x$ 인 키워드네트워크의 뼈대는 다음과 같이 정의되는 그래프

$G(V^x, E^x)$ 로 표현된다.

$$V^x = \bigcap_{i=x}^{2011} V_i \quad \text{and} \quad E^x = \bigcap_{i=x}^{2011} E_i.$$



즉, 뼈대는 기준연도부터 최종연도(이 논문에서는 2011년)까지 지속적으로 잔존하게 되는 정점들과 간선들로 구성된 그래프이다. Fig. 5는 기준연도를 2002년부터 2009년까지 변화시키면서 해당 뼈대를 구하여 정점의 개수와 간선의 개수를 조사한 것이다. 비교를 위하여 Fig. 3(a)와 4(a)에서 관찰된 연도별 잔존 정점과 간선의 비율을 이용하여 기준연도가  $x$ 인 키워드네트워크의 뼈대에 대해 정점의 개수 기대값( $\Xi(V^x)$ )과 간선의 개수 기대값( $\Xi(E^x)$ )을 아래와 같이 구하였다.

$$\Xi(V^x) = |V_{2011}| \prod_{i=x}^{2010} \alpha_i \text{ and } \Xi(E^x) = |E_{2011}| \prod_{i=x}^{2010} \beta_i$$

이 식에서  $\alpha_i$ 와  $\beta_i$ 는 연도  $i$ 에서 차년도  $i+1$ 로 키워드네트워크가 변화할 때 전체 정점 중에서 잔존 정점의 비율, 전체 간선 중 잔존 간선의 비율을 각각 나타낸다. 따라서  $\Xi(V^x)$ 과  $\Xi(E^x)$ 는  $i$ 연도에서 각각  $\alpha_i$ 와  $\beta_i$ 의 비율로 정점과 간선이 임의로 잔존할 때  $x$ 연도 키워드네트워크의 뼈대에 포함될 것으로 기대되는 정점 및 간선 개수의 기대값이다. 먼저 Fig. 5(a)의 결과에서 보이듯이, 키워드네트워크의 뼈대에 포함되는 실제 정점의 개수는 기대값에 비해 크다는 것을 알 수 있다. 그러한 차이는 기준연도가 더 오래될수록 빠르게 커진다. 이러한 결과는 키워드네트워크에서 잔존 정점이 임의로 결정되는 게 아니라 매우 강건하게 잔존하게 됨을 함을 시사한다. 즉, 뼈대에 포함되는 키워드는 이후의 키워드네트워크에서도 매우 높은 확률로 생존하게 됨을 뜻한다. 한편, Fig. 5(b)에서 뼈대에 포함되는 간선의 개수에 관한 결과는 Fig. 5(a)의 정점의 결과와는 조금 다르다. 즉, 기준연도가 오래된 경우에는 뼈대에 포함된 간선이 기대값보다 크지만 기준 연도가 오래되지 않은 경우에는 오히려 전자가 후자보다 작다. 이것은 기준연도가 오래될수록 기대값의 감소가 더 빠르기 때문이다. 즉, 얼마 되지 않은 키워드네트워크의 뼈대에는 비록 간선의 개수가 작지만 그러한 간선들은 매우 높은 확률로 생존하게 됨을 뜻한다. Fig. 5(a)와 5(b)의 결과를 종합하면 키워드네트워크의 뼈대는 정점이나 간선이 임의로 제거되면서 유지되는 것이 아니라 핵심적인 정점과 간선이 강건하게 유지되고 있다 할 수 있다.

더 나아가 이러한 키워드네트워크의 뼈대를 비교해 보면 연구 동향을 이해하는 데 도움을 줄 수 있다. Fig. 6은 기준연도가 2008년과 2009년인 뼈대를 구해 차이의 일부를 나타낸 것이다. 그 그림에서 검은 실선은  $E^{2008}$ 에 포함된 간선을 붉은 점선은  $E^{2009}$ 에 포함된 간선을 나타낸다 (뼈대의 정의에 의해  $E^{2009} \subseteq E^{2008}$ 이 성립한다). 이를 통해, "Biology computing-Biological tissues-Biological techniques" 연결요소와 "Biomechanics-Biomedical MRI" 연결 요소 사이에 2009년도부터 새로 연결되며 "Biological organs"와 "Bioinformatics"와도 연결이 됨을 알 수 있다. 이러한 분석은 2009년 연구 동향을 이해하는 데 도움이 된다.

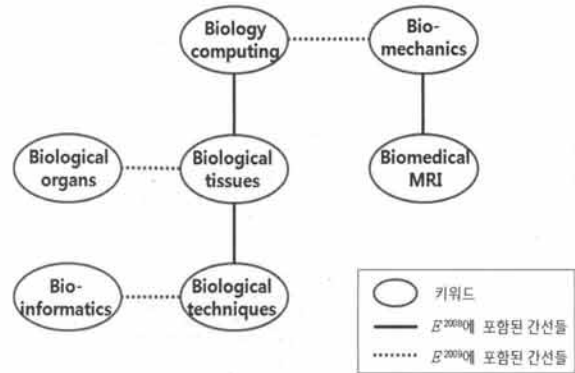


Fig. 6. Partial difference between  $E^{2008}$  and  $E^{2009}$

### 3.4 키워드네트워크의 중심성 분석

네트워크의 구조적 특성을 파악하는 다양한 지표들이 존재하는데 그 중에서 정점들이 얼마나 군집을 잘 이루는지를 나타내는 지표가 군집계수(clustering coefficient)이다[15]. 그래프  $G(V, E)$  군집계수는 아래와 같은 식으로 정의된다.

$$\text{군집계수} = \frac{1}{|V|} \sum_{v \in V} \frac{2e_v}{|N_v|(|N_v| - 1)}$$

이 식에서  $N_v$ 는 정점  $v$ 와 이웃하는 정점들의 집합이고  $e_v$ 는  $N_v$ 에 포함된 정점들 사이의 간선의 개수이다. 즉, 군집계수는 얼마나 많은 세 개의 정점들 사이에 삼각형 형태의 간선집합(즉, clique)을 형성하는지를 나타낸다. 군집계수가 높을수록 지역적으로 정점들이 군집을 이룬다고 할 수 있다. Fig. 7은 키워드네트워크의 연도별 군집계수를 조사한 것으로서 0.48~0.50의 범위에서 일정하게 유지됨을 알 수 있다. 이러한 값의 범위는 2011년 키워드네트워크와 정점/간선의 개수가 동일한 랜덤네트워크를 100개 생성하여 군집계수를 조사한 결과  $0.011 \pm 0.00004$ 였음을 감안할 때, 매우 높은 수준임을 시사한다. 즉, 본 연구에서 생성된 키워드네트워크는 포함된 키워드들 사이에 매우 강한 군집현상을 보이며 높은 간선 제거율(Fig. 4(a))에도 불구하고 그러한 현상이 계속 유지되고 있음을 의미한다.

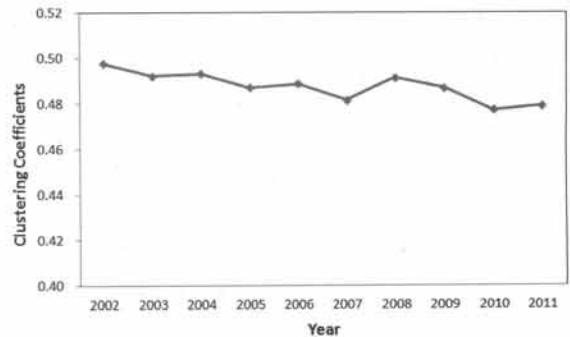


Fig. 7. Change of the clustering coefficients in the keyword networks

#### 4. 결론 및 논의

본 논문에서는 해외 컴퓨터공학 분야 학술 논문에 포함된 키워드 정보를 바탕으로 2001년부터 2011년까지 연도별 키워드네트워크를 구축하고, 연도별 변화를 분석하였다. 네트워크의 정점측면에서 보면, 계속 잔존하는 정점(즉, 키워드)의 비중이 소멸되거나 새로 생성되는 정점에 비해 크게 높았으며 평균 차수도 더 컸다. 하지만 정점에 부족된 간선의 가중치는 오히려 전자의 경우가 더 작았다. 즉, 차수가 높고 가중치가 낮은 간선들과 연결되어 있는 대다수의 키워드들이 잘 보존됨을 뜻한다. 이는 여러 다른 키워드와 골고루 같이 사용되는 키워드가 시간이 지나도 계속 사용되는 경향이 있음을 뜻한다. 다음으로 간선의 측면에서 분석해 보면, 우선 소멸하는 간선의 대부분은 양 끝의 키워드가 계속 잔존함에도 불구하고 소멸하는 경우였다. 또한, 계속 잔존하는 간선의 비중이 양 끝 키워드의 잔존에도 불구하고 소멸하게 되는 간선보다 더 작았으며 전자에 속하는 간선들의 가중치도 후자에 속하는 간선들보다 작았다. 즉, 간선은 보존될 확률이 높지 않으며 특히 가중치가 낮은 간선이 사라질 경우가 높다고 할 수 있다. 이는 두 키워드가 비록 개별적으로 빈번하게 사용되더라도 공통으로 사용되는 빈도가 낮을 경우에는 이후에는 공통으로 사용되기 쉽지 않음을 뜻한다. 이러한 결과들을 토대로 가장 최근의 키워드네트워크까지 계속 보존되는 뼈대를 분석한 결과, 정점이나 간선 모두 임의의 형태로 정점과 간선들이 보존/제거되는 것이 아님을 알 수 있었다. 이는 키워드네트워크에서 특정 키워드와 그들 사이의 연결은 오래 동안 강하게 유지됨을 뜻한다. 추가적으로 키워드네트워크의 군집계수를 조사하였으며 랜덤네트워크의 그것에 비해 매우 높음을 알 수 있었다. 이는 비록 간선의 변화가 심하지만 군집성을 떨어뜨리지는 않음을 시사한다. 이러한 모든 결과는 10년 동안 일관되게 관찰되었으며 이는 키워드네트워크의 구조 변화를 이해하는 데 중요한 원리중의 하나임을 시사한다.

앞으로의 연구는 네트워크의 구조 변화를 잘 설명할 수 있는 통계적 모델을 구하는 것이다. 이를 통해 네트워크 진화에 대한 통찰을 구할 수 있으며 다른 사회관계네트워크의 진화모델과 비교함으로써 보다 일반적인 진화모델로 발전할 수 있을 것이다. 또한, 키워드네트워크 군집에 따른 키워드 연관성 변화를 분석하는 것도 흥미로운 것으로 예상된다. 더 나아가 본 연구의 통찰을 통해 향후 컴퓨터공학 분야의 연구방향을 예측해 보거나 다른 학문 분야에 적용하는 것도 중요한 연구 주제가 될 것이다.

#### 참고 문헌

- [1] Ord, T. J. E. P. Martins, S. Thakur, K. K. Mane, and K. Bömer, "Trends in animal behaviour research (1968.2002): ethoinformatics and the mining of library databases", *Animal Behaviour*, Vol.69, No.6, pp.1399-1413, 2005.
- [2] Huang, C. P. "Bibliometric analysis of obstructive sleep apnea research trends", *Journal of the Chinese Medical Association*, Vol.72, No.3, pp.117-123, 2009.
- [3] Kumaresan, N. and K. Miyazaki, "An integrated network approach to systems of innovation. the case of robotics in Japan", *Research Policy*, Vol.28, pp.563-585, 1999.
- [4] Fujigaki, Y. and A. Nagata, "Concept evolution in science and technology policy: The process of change in relationship among university, industry and government", *Science and Public Policy*, Vol.26, No.6, pp.387.395, 1998.
- [5] Yoon, B. G. and Y. T. Park, "A text-mining-based patent network : Analytical tool for high-technology trend", *Journal of High Technology Management Research*, Vol.15, No.1, pp.37-50, 2004.
- [6] Li, X., H. Chen, Z. Huang, and M. C. Roco, "Patent citation network in nanotechnology(1976~2004)", *Journal of Nanoparticle Research*, Vol.0, pp.337-352, 2007.
- [7] Lee, B. and Y.-I. Jeong, "Mapping Korea's national R&D domain of robot technology by using the co-word analysis", *Scientometrics*, Vol.77, No.1, pp.3-19, 2008.
- [8] Choi, J. S. Yi, and K. C. Lee, "Analysis of keyword networks in MIS research and implications for predicting knowledge evolution", *Information & Management*, Vol.48, pp.371-381, 2011.
- [9] Lee, W. H, "How to identify emerging research fields using scientometrics: An example in the field of Information Security", *Scientometrics*, Vol.77, No.3, pp.503-525, 2008.
- [10] Clara, C. B. Renald, C. C. Valdés, and E. Noyons, "How to identify research groups using publication analysis: an example in the field of nanotechnology", *Scientometrics*, Vol.66, No.2, pp.365-376, 2006.
- [11] Yoon, B. G. and Y. T. Park, "Development of new technology forecasting algorithm : hybrid approach for morphology analysis and conjoint analysis of patent information", *IEEE Transactions on Engineering Management*, Vol.54, No.3, pp.588-599, 2007.
- [12] Lee, S. J., B. G. Yoon, and Y. T. Park, "An approach to discovering new technology opportunities: Keyword-based patent map approach", *Technovation*, Vol.29, No.6/7, pp.481-497, 2009.
- [13] Jung, B. S., Y.-K. Kwon, S. J. Kwak, "A Knowledge Map Based on a Keyword-Relation Network by Using a Research Paper Database in the Computer Engineering Field", *The KIPS Transactions: Part D*, Vol.18, No.6, pp.501-508, 2011.
- [14] Yi, S. and J. Choi, "The organization of scientific knowledge: the structural characteristics of keyword networks", *Scientometrics*, Vol.90, pp.1015-1026, 2012.
- [15] A.L. Barabási and Z.N. Oltvai, "Network biology: understanding the cell's functional organization", *Nature Review Genetics*: 5, 101-113, 2004.



### 권 영 근

e-mail : kwonyk@ulsan.ac.kr

1999년 서울대학교 전산학과(학사)

2001년 서울대학교 컴퓨터공학부  
(공학석사)

2006년 서울대학교 컴퓨터공학부  
(공학박사)

2008년~현재 울산대학교 전기공학부 조교수

관심분야: 최적화 이론 및 실제, 복잡계산시스템, 시스템생물학,  
소셜네트워크 등