학술 논문 데이터베이스를 이용한 컴퓨터 분야 지식지도에 관한 연구

정보석*, 권영근*
*울산대학교 전기공학부
wruwami@mail.ulsan.ac.kr

A study on the knowledge map of the computer engineering field by using a research paper database

Bo-Seok Jung*, Yung-Keun Kwon*
*School of Electrical Engineering,
University of Ulsan

요 약

지식지도는 대량의 정보 속에 숨겨진 특별한 형태와 패턴을 찾아서 그 의미를 파악할 수 있도록 가시적인 형태의 결과를 보여주는 것이다. 기존의 지식지도를 살펴보면, 대부분 빈도수나 평균 증가율과 같은 단순한 정보 분석에 그친다는 것을 알 수 있다. 그러나 사용자에게 보다 고차원적인 정보 분석을 통해 주어진 분석대상에 관한 유용한 지식 전달이 필요하다. 본 논문에서는 2000년부터 2009년까지의 컴퓨터 공학 분야의 학술 논문 데이터베이스를 활용하여 지식지도 구축을 시도한다. 이를 위해 학술 논문 데이터베이스를 네트워크의 관점에서 분석한다. 그 결과, 네트워크 내에서 패턴을 발견할 수 있었다. 네트워크에서 이러한 분석방법은 학술 분야의 연구동향 패턴을 이해할 수 있게 한다.

1. 배경지식

오늘날 정보통신기술의 발전은 정보의 생산, 전달, 교환을 쉽게 할 수 있도록 촉진시켰다. 그러나 정보와 지식의 양은 4년마다 2배씩 증가하고 있지만, 대량 생산된 정보를 분류하여 사용자에게 필요한 정보를 전달하는 기술의 발전은 미비한 상황이다. 이런 상황에서 축적된 정보를 가공하는 작업이 필요한데, 사용자에게 필요한 지식으로 전달하고자 하는 방법 중 하나가 지식지도이다. 지식지도는 대량의 정보 속에 숨겨진 특별한 형태와 패턴을 찾아서 그 의미를 파악할 수 있도록 가시적인 형태의 결과를 보여주는 것이다. 지식지도는 다양한 목적으로 활용될 수 있지만, 일반적으로 국가 정책적, 연구적 목적으로 개발, 연구되어 지고 있다. 이러한 목적에 의해 최근 이러한 지식지도에 대한 연구가 이루어지고 있다. 하지만 이에 대한 연구는 아직까지 단순한 분석에 그치는 경우가 많다. 빈도수 조사, 평균 증가율 등 분석에서는 사용자에게 알려주는 정보의 범위가 많지 않고, 의미 해석에 있어서도 다양한 정보를 전해주는데 어려움이 있다. 그래서 보다 다양하고 필요한 지식을 가공하기 위해 분석대상을 네트워크로 표현하여 시스템 측면에서 살펴볼 필요가 있다.

본 논문에서는 컴퓨터공학 분야의 학술 논문에 대한 메타데이터를 통해 지식지도를 구축하고 이를 네트워크의 관점에서 분석해본다. 이를 본 논문에서는 네트워크에서 연결요소(Connected Component; CC)를 조사하여 연도별 변화를 조사한다. 이것은 사용자에게 연구동향에 관한 다양하고 의미 있는 정보를 제공할 수 있고, 연구자에게 새로운 방법을 제시한다는 측면에서 매우 의미 있는 작업이 될 것이다.

2. 분석 데이터

본 논문에서는 한국과학기술정보연구원(KISTI)에서 제공한 컴퓨터공학 분야의 학술 논문 데이터베이스를 활용한다. 제공되는 메타데이터는 논문의 제목, 제출학회명, 제출학회지명, 제출년도, 저자정보, 키워드, 초록 등 여러 정보가 주어진다. 이러한 정보 중 지식지도에 사용될 정보는 주어진 메타데이터에서 지식지도 구축에 용이한 정보로 가공하여 사용되어 진다.

본 논문에서 구현하고자 하는 지식지도는 학술 논문 메타데이터를 이용한 지식지도이다. 먼저 학술 논문의 메타데이터는 2000년에서 2009년까지 컴퓨터 분야의 53432건의 학술논문 중에서 학술지명, 학회명, 저자명, 년도, 소속기관명, 키워드 등 6가지의 메타데이터를 다 가지지 못한 학술 논문을 제외하여 18297건의 논문을 이용하여 분석하였다.

3. 컴퓨터 공학 분야 지식지도 구축 및 분석

본 논문에서 지식지도를 구축하는 방법으로 먼저 학술 논문 데이터베이스에 대한 키워드 네트워크 구축이 이루어진다. 그리고 구축된 키워드 네트워크에 대한 분석이 이루어진다.

(1) 키워드 네트워크 구축

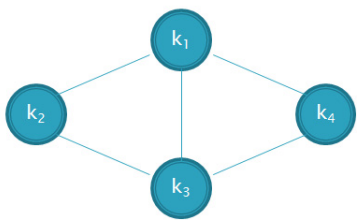
본 논문의 학술 논문 지식지도는 기본적으로 키워드를 분석, 가공한 정보를 사용자에게 제공한다. 키워드를 주요 매개체로 선정한 이유는 학술 논문에 대한 내용을 효과적으로 축약해 놓은 항목이기 때문이다. 키워드 네트워크는 논문에 명시되어 있는 키워드를 이용하여 만든 네트워크이다. 키워드 네트워크를 만들기 위해서 먼저 각 키워드가 명시되어 있는 논문의 리스트를 만든다. 각 키워드 간의 논문의 리스트를 벡터로 생각하여 키워드 쌍 k_1, k_2에 대한 유사도 $\varpi(k_1, k_2)$를 아래와 같은 코사인 유사도를 구하게 된다.

$$\omega(k_1, k_2) = \frac{|P(k_1) \cap P(k_2)|}{|P(k_1)| \cdot |P(k_2)|} \quad \left(P(k) : k\text{를 포함하는 논문의 집합} \right)$$

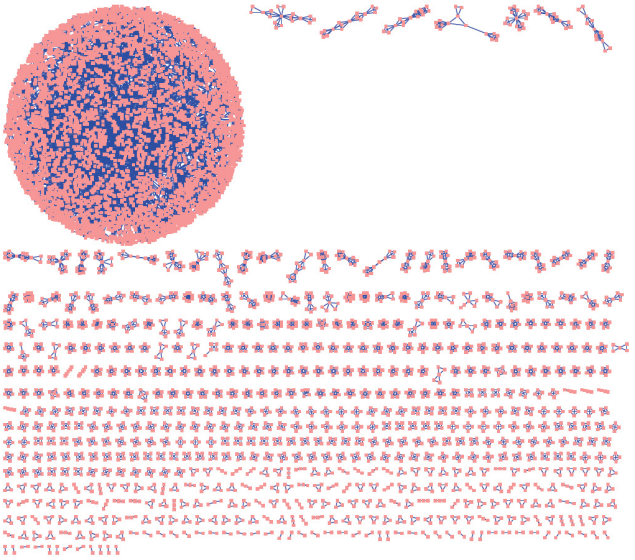
이렇게 구한 코사인 유사도를 바탕으로 키워드 네트워크를 그릴 수 있다. 아래의 (그림 1)은 네트워크 구축 과정을 설명한다. 가중치 ϖ값의 유무에 따라 정점간의 간선(edge)의 유무가 결정된다. 네트워크의 각 정점은 키워드로 이루어져 있고, 간선은 키워드간의 유사도가 0이 아닌 값을 가질 경우에 간선이 생성된다. 자신과의 키워드 유사도는 1을 가지지만, 간선은 그리지 않았다.

(그림 2)는 위의 구축과정에 따라 만들어진 네트워크 중 2007년 학술 논문을 이용하여 만든 키워드 네트워크이다. (그림 2)에서 보듯이 네트워크는 여러 개의 CC로 구성되며 그 중 가장 큰 CC를 LCC(The largest connected component)라 정의한다.

	k_1	k_2	k3	k_4
k_1	1	0.5	0.3	0.2
k_2	0.5	1	0.5	0
k_3	0.3	0.5	1	0.5
k_4	0.2	0	0.5	1



(그림 1) 네트워크 구축 과정

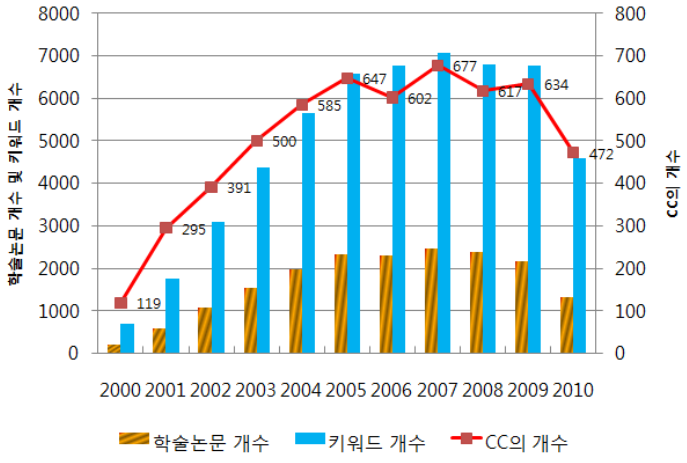


(그림 2) 2009년 키워드 네트워크 그래프

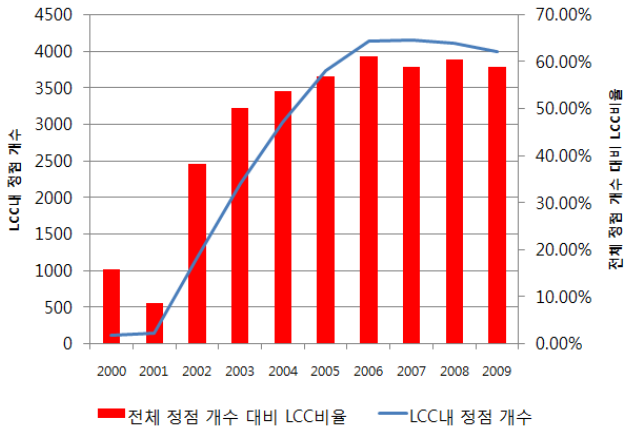
(2) 키워드 네트워크 분석

먼저, 연도별 학술 논문의 수, 학술 논문에 나와 있는 키워드의 개수, 그리고 키워드 네트워크를 형성했을 때 나오는 CC의 개수를 분석하였다.

(그림 3)에서 보듯이 학술 논문이 2005년까지 계속 증가하다가 그 이후에는 큰 변화가 없음을 알 수 있다. 키워드 개수가 학술 논문의 개수에 따라 어느 정도 비슷한 양상을 보인다는 것을 알 수 있다. CC의 개수의 경우, 학술 논문의 개수와 비슷한 경향을 보이지만, 2006년과 2008년과 같이 약간의 차이를 보이는 경우도 존재한다.



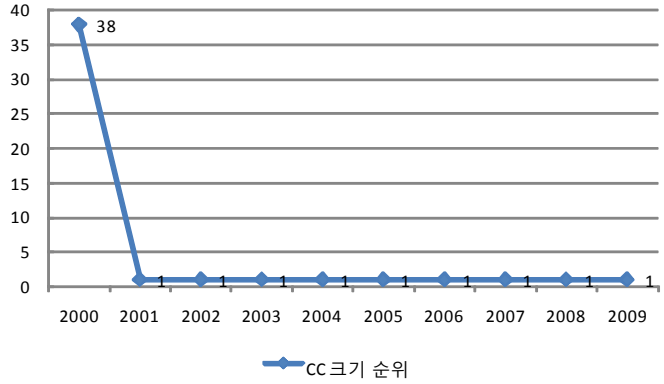
(그림 3) 연도별 CC, 학술 논문, 키워드의 개수 변화



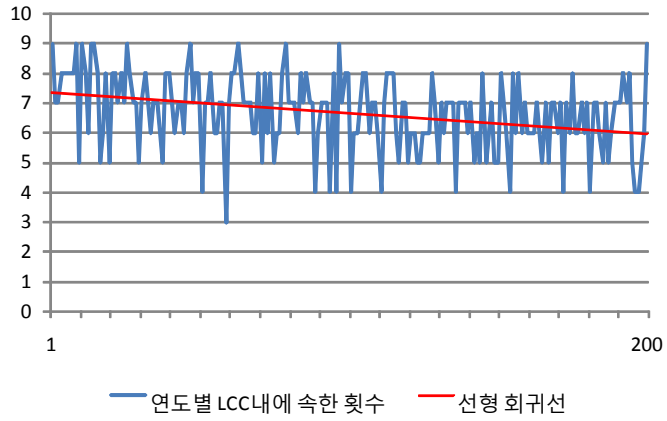
(그림 4) 전체 키워드에서 LCC내의 키워드가 차지하는 백분율 및 연도별 LCC의 정점수의 변화

다음으로 LCC에 관해 분석을 하였다. 연도별 가장 큰 CC의 정점 개수의 변화와 전체 키워드에서 LCC내의 키워드가 차지하는 비율을 (그림 4)에서 확인할 수 있다. (그림 4)에서 CC의 개수처럼 2005년까지 LCC 역시 그 크기(LCC내 정점 개수)가 증가함을 알 수 있다. 흥미로운 점은 LCC의 상대적 비율도 2001년을 제외하면 대체적으로 증가한다는 점이다.

한편, 키워드 중에서 가장 높은 빈도를 차지하는 상위 200개의 키워드를 추출하여 각 키워드들이 연도별로 위치한 CC를 분석하였다. (그림 5)는 모든 키워드 중 10번째로 높은 빈도를 가지는 키워드인 "Neural Network"에 대한 연도별 속한 CC 크기 순위 차트이다. 이런 CC 크기 순위는 연도별 CC내의 정점 개수에 따라 나열되어 있고, CC 크기 순위가 0인 경우는 키워드가 어떠한 CC에도 속해있지 않다는 것을 의미한다. (그림 6)는 상위 200개 키워드의 연도별 LCC내에 속한 횟수이다. 인덱스가 1에 가까울수록 빈도수가 높은 키워드를 의미한다. 이를 분석한 결과 상위 200개의 키워드 중 한번도 LCC에 속하지 않은 키워드는 단 한 건도 존재하지 않았다. 연도별 LCC내에 속한 횟수에 따라 그린 선형 회귀선이 음수 값으로 나오는 것을 볼 때, 대체적으로 빈도수가 높은 키워드일수록 LCC에 속한 횟수가 높다는 것을 알 수 있다. 이는 상당히 의미 있는 결과로서, 연도별 키워드들 중 LCC내에 위치하고 있는 키워드의 숫자는 60%가 되지 않는다는 점을 (그림 4)에서 확인할 수 있다. 학술 논문에서 많은 빈도로 사용된 키워드들이 대부분 LCC에 속해있다는 점은 연도별 네트워크에서 LCC를 분석하는 것이 매우 의미 있는 연구가 될 것이라는 점을 의미하기 때문이다.



(그림 5) 연도별 "Neural Network" 키워드가 속한 CC의 크기 순위



(그림 6) 상위 200개 키워드의 연도별 LCC내에 속한 횟수

4. 결론

키워드 네트워크 분석 결과를 통해 알 수 있었던 점은 키워드 네트워크를 구축하였을 때, 연도별 네트워크에서 가장 큰 CC이 많은 의미를 내포하고 있다는 점이다. 가장 많은 빈도수를 가지는 키워드들이 연도별 키워드 중 60%가 안 되는 키워드 수를 가지는 CC내에 포함되어 있다는 사실은 앞으로도 연구 가치가 높다는 것을 입증한다.

앞으로의 연구 진행은 LCC를 분석하여 내부적인 구조를 파악하는 것을 중심으로 진행될 것이다. 위의 네트워크에선 가중치에 의한 분석이 나와 있지 않은 상태이지만, 가중치를 고려한 분석이 이루어진다면, CC의 구조를 더 알아볼 수 있을 것이다. 이러한 분석은 LCC를 이해한다는 점에서 나아가 키워드의 시간에 따른 패턴 분석 및 키워드의 순위 변화 예측을 할 수 있는 발판이 될 수 있다는 점에서 의미 있는 연구가 될 것이다.

참고문헌

[1] 원동규 "사회과학분야 학술연구 지식지도(knowledge map)의 개발 및 구현" 한국학술진흥재단, 2007

[2] 이광희 "지식지도 작성을 위한 기초연구" 한국학술진흥재단 , 2009

[3] 이종훈, 이경순, 최기선 "분야 간 유사도와 통계기법을 이용한 전문용어의 자동 추출" 정보과학회논문지, 2002