가중치 정보를 가진 연구자 네트워크 기반의 연구자 클러스터링 기법*

문 현 정** · 이 상 민*** · 우 용 태****

Researcher Clustering Technique based on Weighted Researcher Network

Mun, Hyeon Jeong · Lee, Sang Min · Woo, Yong Tae

〈Abstract〉

This study presents HCWS algorithm for researcher grouping on a weighted researcher network. The weights represent intensity of connections among researchers based on the number of co-authors and the number of co-authored research papers. To confirm the validity of the proposed technique, this study conducted an experimentation on about 80 research papers. As a consequence, it is proved that HCWS algorithm is able to bring about more realistic clustering compared with HCS algorithm which presents semantic relations among researchers in simple connections. In addition, it is found that HCWS algorithm can address the problems of existing HCS algorithm; researchers are disconnected since their connections are classified as weak even though they are strong, and vise versa. The technique described in this research paper can be applied to efficiently establish social networks of researchers considering relations such as collaboration histories among researchers or to create communities of researchers.

Key Words : Researcher network, Researcher clustering, HCS, HCWS

Ⅰ. 서론

사회연결망은 사람들이 일련의 관계에 의해 연결된 관계망으로 '지인네트워크' 등의 의미로 사용되고 있다. 최근에 사회연결망은 사회학분야에서 사회연결망 이론(social network theory)을 통해 활발하게 연구가 진행되고 있다[1, 2]. 이러한 사회연결망은 연구자간의 지식을 공유하거나 연구자 검색 그리고 협업 연구를 지원하기 위한 연구자 네트워크 구성에 응용할 수 있다.

연구자 네트워크는 공동 연구 성과물, 공동 과제, 학연 등을 기반으로 연구자를 연결하여 구성한 모델로서 국가적 차원에서의 인적 자원 관리, 과제 심사를 위한 심사자 선정, 협업 연구 지원을 위해 사용될 수 있다.

하지만 '케빈 베이컨 게임'에서 모든 배우들이 2~3단계만 거치면 케빈 베이컨과 지인 관계가 형성되는 것처럼 연구자 네트워크에서도 몇단계만 거치면 모든 연구자가 연결될 수 있다. 또한 1단계인 직접 관계에 있다고 하더라도 친밀도가 낮을 수도 있다.

* 이 논문은 2004년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임. (KRF-2004-050-D00016)
** 하이브레인넷 수석연구원
*** 대림I&S SM사업부 IS운영팀
**** 교신저자, 창원대학교 교수

이러한 부분을 개선하기 위하여 연구자 네트워크에서 연구자간 관계 친밀도를 가중치로 표현하기 위한 연구가 진행되어 왔다. Junichiro Mori 등은 문서에서 연구자 이름이 동시에 나타난 빈도수를 기반으로 연구자간 가중치를 부여하였고[3], 이승우 등은 공저 논문수와 피인용수에 기반하여 연구자간 가중치를 부여하였다[4]. 그리고 전인하는 공저 논문과 공저 논문 수에 기반하여 연구자간 친밀도를 부여하였다[5].

또한 연구자들은 전공이나 관심 분야와 아주 밀접한 관계를 가지고 이를 기반으로 연구나 협업을 진행하므로 연구자를 클러스터링하기 위한 기법이 필요하다. 지금까지 연구자를 클러스터링하기 위한 기법은 마이닝 기법이나 관심 속성 정보나 사용자 질의어의 유사성에 기반하였다[6-8]. 하지만 사회 연결망과 같은 네트워크 구조에서 기존의 클러스터링 기법을 적용하여 연구자를 클러스터링하기에는 어려운 부분이 있다.

본 논문에서는 가중치 정보를 가진 연구자 네트워크에서 연구자를 클러스터링하기 위한 기법을 제안한다. 제안하는 클러스터링 기법은 HCS(Highly Connected Subgraphs) 알고리즘을 확장한 HCWS(Highly Connection Weighted Subgraphs) 알고리즘이다. HCS 알고리즘은 그래프상에서 강한 연결 관계인 노드의 집합을 클러스터하기 위한 것으로, 약한 연결 관계에 있는 최소의 간선 제거로 그래프를 분할해 가는 방식이다. HCWS 알고리즘은 연구자간의 친밀도를 기반으로 친밀도가 약한 연구자를 분할해 가는 방식이다. 연구자간 친밀도는 전인하 등이 제안한 친밀도를 이용한다[5].

본 논문에서 제안한 클러스터링 기법의 타당성을 검증하기 위하여 컴퓨터공학과 교수들이 최근 5년간 발표한 80여편의 논문 실적을 대상으로 실험을 하였다. 연구자간의 가중치 정보는 논문의 공저자 관계를 이용하여 구성하였다. 그리고 HCS 알고리즘과 HCWS 알고리즘을 이용한 클러스터링 결과를 비교 분석하였다.

II. 관련 연구

2.1 사회연결망 구성에 관한 연구

최근 사회연결망은 시맨틱 웹의 발전으로 사용자간 관계를 추출하거나 사용자들의 협업을 효과적으로 지원하기 위한 방향으로 연구되고 있다. 사용자간의 관계를 추론하거나 동적으로 구성하기 위해 온톨로지 기반으로 구성하기 위한 연구도 진행되고 있다. 일반 사용자들의 대표적인 네트워크 구성을 위한 FOAF는 사용자 정보와 사용자간의 관계를 관리하기 위해 제안한 RDF 기반의 사용자 모델이다[9, 10]. FOAF에서 사용자간 관계는 사용자가 명시적으로 입력하거나 웹마이닝 기법을 통해 추출할 수 있다.

사용자들간의 관계에 가중치를 부여하여 네트워크를 구성한 연구들도 진행되고 있다. Junichiro Mori 등은 웹 마이닝 기법을 사용하여 사용자간의 연관성을 분석하고 사회연결망을 구성하였다. 사용자간 연관성 정도는 웹 페이지에서 사용자 이름이 동시에 많이 나타날수록 연관성이 높다고 판단하였다. 성원경 등은 연구자들의 인적자원을 의미적으로 관리·공유하기 위해 시맨틱 웹 기술과 온톨로지를 이용한 OntoFrame-K를 제안하였다[6]. 이 모델에서 연구자 네트워크는 논문의 공저자 수, 피인용수를 이용하여 구성하였고, 친밀도는 공저한 논문의 수에 비례하여 높다고 계산하였다. POLYPHONET 시스템에서는 연구자명과 키워드의 동시 출현 빈도수를 분석하고 키워드 집합이 유사한 연구자를 네트워크로 연결하였다[11]. 국내에서는 조직내 사용자간의 지식 공유를 위해 지식 네트워크(knowledge network)를 구성하려는 연구도 있었다[12].

2.2 사용자 클러스터링에 관한 연구

사용자 클러스터링은 비슷한 관심사를 가진 사용자를 그룹핑하는 것으로, 웹 마이닝(web mining), LSA(Latent Semantic Analysis), 온톨로지 등을 이용한 연구들이 진

행되어 왔다[7, 8, 13]. 사용자 클러스터링의 목적은 개인의 관심사를 반영하여 같은 그룹의 사용자에게 컨텐츠나 아이템을 추천하기 위한 것이다. 대부분의 사용자 클러스터링 기법은 사용자의 명시적인 관심 정보나 사용자들의 묵시적인 행동을 포함한 로그 정보를 기반으로 한다.

연구자 클러스터링은 주로 관심 전공이나 키워드를 기반으로 클러스터링하기 위한 연구가 진행되고 있다. 성원경 등은 문서에 동시에 나타난 사용자명간의 관계에 기반하여 사용자를 클러스터링하였다[6]. Smirnov 등은 온톨로지를 이용하여 사용자 질의어와의 유사도를 기반으로 사용자를 클러스터링하였다. 즉, 질의어와 온톨로지 명과의 유사도를 그래프 형태로 표현하고 그래프에서 거리가 짧은 사용자들을 대상으로 계층적 클러스터링 기법으로 사용자를 클러스터링하였다[8].

2.3 기존 연구의 문제점

연구자 협업을 위한 연구자 네트워크의 구성이 필요하지만, '케빈 베이컨 게임'에서처럼 20만명이 넘는 배우들이 2-3단계만 거치면 연결되는 형태의 네트워크 구성은 실제적으로는 무의미하다. 따라서 연구자 네트워크에서 연관성이나 친밀도에 따라 유사 연구자들을 클러스터링하는 것이 필요하다. 하지만 현재까지 연구자를 대상으로 한 네트워크에서는 논문의 공저자 관계를 이용한 네트워크 구성과 공저자 수를 기준으로 한 가중치 부여에 대한 연구가 대부분 진행되어 왔다.

기존의 사용자 클러스터링 기법은 사용자의 행위나 검색 키워드를 분석하여 관심 내용을 기준으로 비슷한 성향의 사람을 그룹핑하는 것에 초점을 맞추었다. 사회연결망은 사용자의 관심 내용보다는 사용자간의 지인 관계나 공동 연결 속성으로 연결되므로 기존의 클러스터링 기법으로는 사용자를 그룹핑하기가 힘들다. 또한 가중치 그래프에서 사용자를 클러스터링하기 위한 연구도 미흡하다. 기존의 그래프를 이용한 클러스터링 기법은 온톨로지를 기준으로 사용자 질의어와의 유사도를 그래프로 표현한 것으로, 사용자간 관계를 표현한 사회연결망과는 근본적인 차이가 있다.

Ⅲ. 연구자 네트워크 기반의 연구자 클러스터링 기법

본 논문에서는 가중치 정보를 가지는 연구자 네트워크에서 연구자간의 협업 집단을 효과적으로 구성하기 위한 새로운 형태의 연구자 클러스터링 기법을 제시하였다. 이 기법은 서브그래프로 분할하기 위한 HCS 알고리즘을 확장하여 가중치 그래프에서 그래프를 분할하기 위한 HCWS 알고리즘을 사용하였다. 본 연구에서 사용한 연구자간 가중치는 전인하 등이 제시한 사회연결망상에서 연구자간 연결 강도를 반영한 친밀도를 이용하였다[5].

3.1 HCS(Highly Connected Subgraphs) 알고리즘을 이용한 클러스터링

3.1.1 HCS 알고리즘

HCS 알고리즘은 그래프 이론에 기초한 그래프 분할 알고리즘의 하나이다[14]. 이 알고리즘은 최소의 간선 제거로 그래프를 분할하기 위한 것으로 분할된 서브그래프에서는 강한 연결 관계(highly connected)를 가지는 노드로 구성된다. HCS 알고리즘은 분자 생물학에서 유전자의 기능과 상호관련성을 분석하기 위하여 제안된 알고리즘이다. <그림 1>은 HCS 알고리즘을 나타낸 것이다.

```
HCS( G(V,E) )
begin
   (H, H̄, C ) ← MINCUT(G )
   if k(G) > n/2
      then return(G )
   else
      HCS(H )
      HCS(H̄ )
   end if
end
```

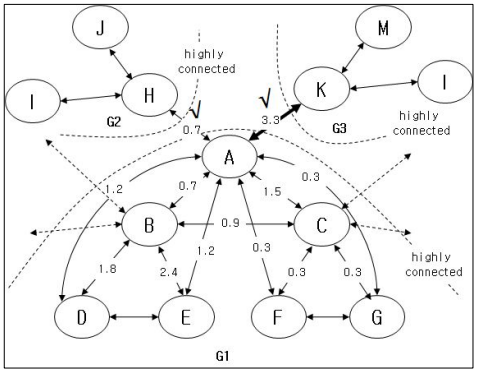
<그림 1> HCS 알고리즘

<그림 1>의 HCS 알고리즘에서 그래프 G는 자원간의 의미적인 유사성에 의해 연결된 그래프이다. H는 강한 연결 그래프이고, $\overline{H}$는 강한 연결 그래프의 여집합이다. C는 그래프 G를 서브그래프 H와 $\overline{H}$로 분할하기 위한 최소분할조건(MINCUT)이다. G는 그래프가 정지 조건을 만족할때까지 반복하여 최소분할조건을 통해 두 개의 서브그래프(H, $\overline{H}$)로 순환적으로 분할된다. 최소분할조건은 그래프를 서브그래프로 분할하기 위한 간선의 최소 수이다. 즉, 그래프를 서브그래프로 분할하기 위해서 제거될 간선의 최소 수를 k(G)라 하고, 노드의 총수를 n이라 할 때, 강한 연결(highly connected)은 k(G) > n/2 조건을 만족해야 한다.

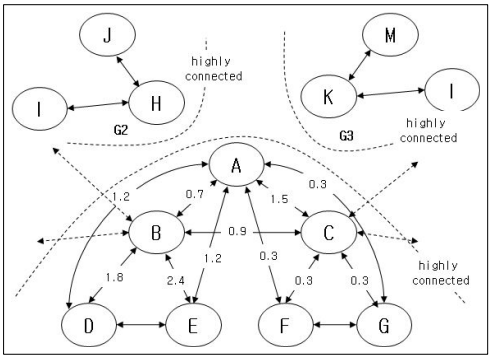
3.1.2 HCS 알고리즘을 이용한 연구자 클러스터링

가중치 정보를 가진 연구자 네트워크에서 HCS 알고리즘의 문제점을 분석하기 위하여 연구자를 클러스터링 하였다. <그림 2>는 논문의 공저 관계에 있는 연구자끼리 연결한 연구자 네트워크에 대해 HCS 알고리즘이 수행되는 과정을 보여주는 그림이다. 그래프 G에서 노드는 사용자를 나타내고, 간선은 논문을 공저한 사람끼리 연결된 관계를 의미한다. 간선위의 숫자값은 연구자간 친밀도를 의미한다. <그림 2>에서 전체 간선 수는 18이므로 n/2는 9가 되고 그래프 G1과 G2를 분할하기 위해 필요한 k(G)는 1이다. 강한 연결이 되기 위한 k(G) > n/2 조건을 만족하지 못하므로 분할된다. G3도 G2와 같은 계산에 의해 강한 연결 조건을 만족하지 못하므로 분할된다.

<그림 3>은 HCS 알고리즘에 의해 하나의 연구자 네트워크가 3개의 연구자 집단으로 분할된 결과를 보여주는 그림이다. HCS 알고리즘의 실행 결과, 사용자 A, H, K간의 공저 관계는 약한 연결(weakly connected)로 분리되어 세 개의 그룹으로 나뉘어진다. 그리고 G1의 {A, B, C, D, E, F, G}, G2의 {I, J, H}, G3의 {K, L, M}은 강한 연결에 의해 클러스터링되었다.



<그림 2> HCS 알고리즘의 수행 과정

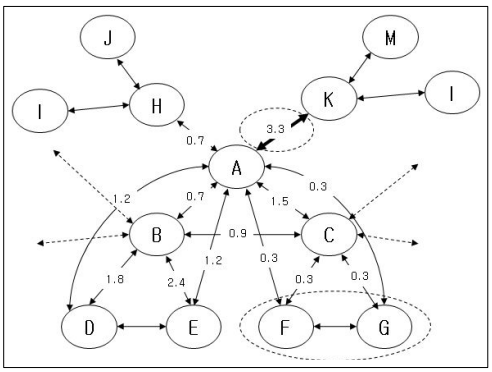


<그림 3> HCS 알고리즘의 수행 결과

3.1.3 HCS 알고리즘의 문제점

HCS 알고리즘은 연구자간의 단순 연결 정보를 기준으로 연구자를 클러스터링 하는 관계로 다음과 같은 문제점이 발생한다. 첫째, 연구자간의 의미적인 연결강도가 평균치보다 높음에도 불구하고 약한 연결로 분류되는 문제점이다. 예를 들어, <그림 4>에서 연구자 A와 K의 논문 공저 가중치는 3.3으로 전체 평균 1.15보다 높지만 HCS 알고리즘에서는 약한 연결로 분류되어 분리되었다. 둘째, HCS 알고리즘에서는 연구자간의 의미적인 연결강도가 평균치보다 낮음에도 불구하고 강한 연결로 분류되는 문제점이다. 예를 들어, 연구자 F-G와 연구자 A간의 논문 공저 가중치가 0.3으로 평균보다 낮지만, HCS 알고

리즘에서는 간선 수가 강한 연결 조건에 만족하기 때문에 분리되지 않는 문제점이 발생한다.



<그림 4> HCS 알고리즘의 문제점

3.2 HCWS(Highly Connection Weighted Subgraphs) 알고리즘

본 연구에서는 의미적인 친밀도를 가중치로 표현한 연구자 네트워크상에서 연구자를 효과적으로 그룹핑하기 위해 HCWS 알고리즘을 제안한다. HCWS 알고리즘은 연구자간의 관계를 단순 연결로 표현하는 HCS 알고리즘에 비해 친밀도에 의해 보다 사실적인 클러스터링이 가능하다. 특히, HCWS 알고리즘은 친밀도를 클러스터링하는 과정에 반영하여 연구자간의 관계가 강함에도 불구하고 분리되거나 연구자간의 관계가 약함에도 불구하고 분리되지 않는 기존 HCS 알고리즘의 문제점을 개선할 수 있다.

연구자간의 친밀도는 연구자 네트워크의 유형에 따라 다양하게 정의할 수 있다. 예를 들어, 전공, 취미, 학력 등과 같이 연구자간 관계의 정도를 표현할 수 있는 의미적인 자원은 친밀도를 정의하는 요소로 사용할 수 있다. 본 연구에서는 연구자간의 연구자 네트워크 구성을 위하여 전인하가 제안한 논문의 공저자 수와 공저 논문 편수에 의해 연구자간의 친밀도를 정의한다[5].

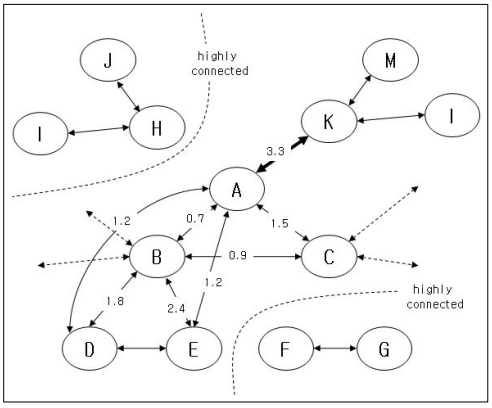
<그림 5>는 본 논문에서 제시한 HCWS 알고리즘이다.

```
HCWS( G(V,WE) )
begin
   (H, H̄, C ) ← MINCUT(G )
   if k(G) > n/2
         and w(k(G)) > avg(w)
      then return(G )
   else
      HCWS(H )
      HCWS(H̄ )
   end if
end
```

<그림 5> HCWS 알고리즘

HCWS 알고리즘에서는 서브그래프로 분할하기 위하여 제거될 간선의 친밀도 평균값이 그래프 전체의 친밀도 평균값보다 크면 강한 연결로 정의한다. 그리고 전체 그래프의 친밀도 평균값보다 작은 서브그래프는 약한 연결로 분류되어 분할된다. C는 서브그래프를 분할하기 위한 최소분할조건으로, 서브그래프를 분할하기 위한 간선의 최소수를 만족하는 간선의 친밀도값이 평균 이하인 것이다. 즉, 그래프를 서브그래프로 분할하기 위해서 제거될 간선의 최소 수를 k(G), 전체 노드의 수 n, k(G)의 친밀도 평균값을 w(k(G)), 전체 친밀도 평균값 avg(w)라 할 때, 강한 연결(highly connected)은 k(G) > n/2 와 w(k(G)) > avg(W) 조건을 만족해야 한다.

<그림 6>은 A와 K노드간의 강한 친밀도에 의해 유지되는 간선과 F-G를 서브그래프로 분할하는 간선을 표시한 그림이다. 그래프에서 전체 간선의 수는 18개이고 전체 간선의 가중치 합은 20.7이다. 따라서 그래프 전체의 친밀도 평균값은 1.15이다. A와 K노드간의 친밀도는 3.3으로 평균치보다 높아 간선이 그대로 유지된다. 연구자 F-G와 연결된 4개의 간선에 대한 친밀도 평균은 0.3으로 전체 친밀도 평균 1.15보다 낮아서 서브그래프로 분할된다. <그림 3>의 HCS 알고리즘에서는 A와 K노드가 약한 연결로 분류되어 분리되고, F-G 서브그래프는 강한 연결로 분류되어 연결된다.



<그림 6> HCWS 알고리즘의 결과

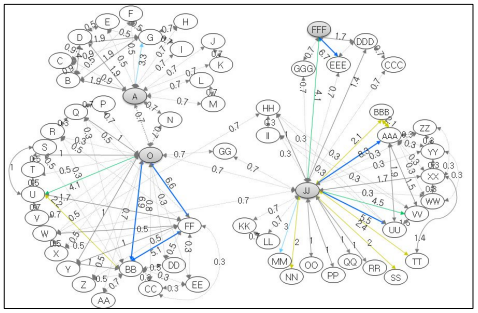
IV. 실험 및 결과

4.1 실험 데이터 구성

본 논문에서 제안한 기법의 효율성을 검증하기 위하여 저자가 소속된 컴퓨터공학과 교수들이 2002년부터 2006년까지 5년간 발표한 80여편의 논문 실적을 대상으로 실험하였다. 각 논문에서 공저자를 추출하여 연구자 네트워크를 구성하였다. 그리고 각 논문에 대한 공저자 수와 공저 논문 수를 이용하여 연구자간의 친밀도를 정의하였다. 실험에 사용된 그래프의 전체 간선의 수는 126개이고, 전체 간선의 친밀도 합은 144.6이다. 따라서 그래프 전체의 친밀도 평균값은 1.2이다. <그림 7>은 실험 데이터로 구성한 친밀도를 가지는 연구자 네트워크의 예이다.

4.2 실험 결과 비교 및 분석

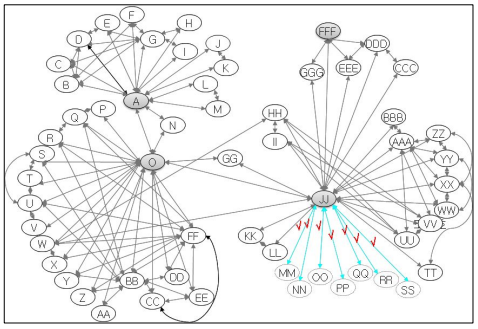
실험에서는 <그림 7>의 연구자 네트워크를 대상으로 HCS 알고리즘과 본 연구에서 제안한 HCWS 알고리즘을 이용한 연구자 클러스터링 결과를 비교하였다.



<그림 7> 연구자간 친밀도를 반영한 연구자 네트워크의 예

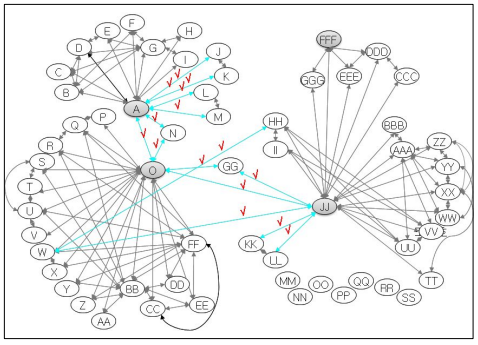
4.2.1 HCS 알고리즘을 이용한 연구자 클러스터링 실험

먼저, HCS 알고리즘을 이용한 연구자 클러스터링에 대해 실험하였다. <그림 8>에서 서브그래프로 분할하기 위한 간선의 최소수가 1 이기 때문에 연결된 간선이 1개인 노드를 분리한다. 서브그래프 조건에서 노드가 하나의 경로만 존재할 경우에는 약한 연결로 분류된다.



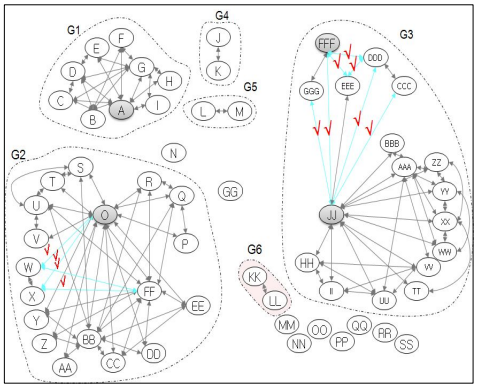
<그림 8> HCS를 이용한 연구자 클러스터링 1단계

그 결과, <그림 9>와 같이 JJ를 기준으로 간선이 하나인 SS, RR, QQ, PP, OO, NN, MM은 약한 관계로 분류되어 서브그래프로 분리된다. HCS 알고리즘은 강한 연결 조건 K(G) > n/2 를 만족할 때까지 반복 수행한다. 다음 단계의 서브그래프로 분할하기 위한 간선의 최소수는 2개이다. <그림 9>는 <그림 8>에서 간선이 1개 이하인 간선을 제거한 결과와 간선이 2개 이하인 간선을 분리하는 과정이다.



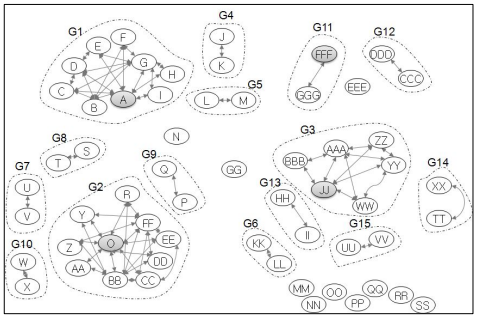
<그림 9> HCS를 이용한 연구자 클러스터링 2단계

<그림 9>에서 간선이 2개인 간선을 제거한 결과 <그림 10>과 같이 A, O, JJ를 기준으로 그룹이 형성되었고, 약한 연결로 인해 분리된 작은 그룹이 따로 형성되었다. <그림 10>은 <그림 9>에서 간선이 2개인 노드를 분리한 결과이다. 그리고 JJ, O를 기준으로 서브그래프로 분할하기 위한 간선의 최소수 4개 이하인 노드를 분리한다.



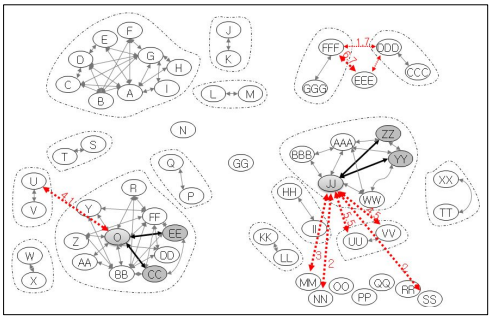
<그림 10> HCS를 이용한 연구자 클러스터링 3단계

이와 같은 과정을 강한 연결 관계를 가진 서브그래프로 분류될까지 반복한다. 최종적으로 실험 데이타에서는 <그림 11>과 같이 15개의 그룹이 형성된다. 하지만 HCS 알고리즘의 클러스터링 결과에서 두 가지의 문제점이 발생한다. 첫째, 특정 연구자간의 논문 공저 편수가 전체의 평균치보다 많음에도 불구하고 약한 연결로 분류되어 분리되는 점이다. 그리고 둘째는 연구자간 논문 공저 편수가 평균치보다 낮음에도 불구하고 강한 연결로 분류되어 더 이상 분리되지 않는 점이다.



<그림 11> HCS를 이용한 연구자 클러스터링 최종 결과

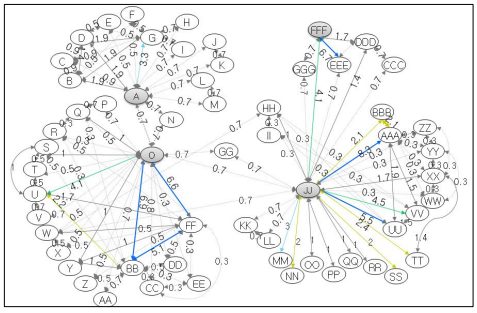
<그림 12>는 HCS 알고리즘에서 연구자 클러스터링 과정의 문제점을 보여주는 예이다. 점선의 화살표로 표현한 간선은 친밀도 평균값보다 높지만 약한 연결 관계로 분류되어 서브그래프로 분리된 것이다. 그림에서 O와 U, JJ와 MM, NN, UU, SS, VV는 친밀도값이 높지만 약한 연결 관계로 분류되어 분리된다. 회색 노드는 친밀도값은 평균 친밀도값보다 낮지만 강한 연결 관계로 분류되어 분리되지 않은 것이다. EE와 CC는 O와 강한 연결 관계이고, ZZ와 YY는 JJ를 기준으로 친밀도는 낮지만 강한 연결관계로 분리되지 않은 것이다.



<그림 12> HCS를 이용한 연구자 클러스터링의 문제점

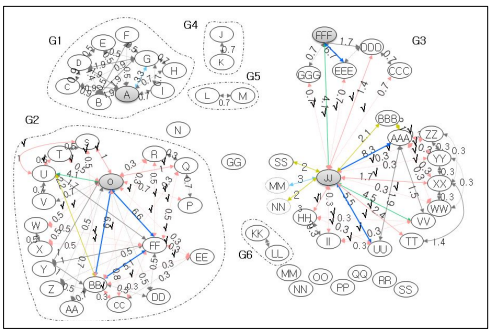
4.2.2 HCWS 알고리즘을 이용한 연구자 클러스터링 실험

실험에 사용된 그래프 전체의 친밀도 평균값은 1.2이므로 서브그래프의 간선 평균값이 1.2 보다 작으면 서브그래프로 분할된다. <그림 13>은 HCWS 알고리즘을 이용한 연구자 클러스터링 과정이다. 먼저, 간선의 수가 1개인 간선의 평균값 중 전체의 친밀도 평균값 보다 작은 간선은 제거되어 서브그래프로 분할된다. 그리고 간선의 수가 2개 이하인 간선에 대하여 반복적으로 비교하여 서브그래프로 분할하는 과정이다.



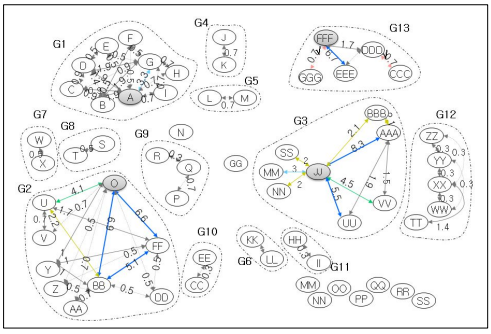
<그림 13> HCWS을 이용한 연구자 클러스터링 1단계

<그림 14>는 전체 그래프 친밀도 평균값보다 작은 간선을 제거한 결과이다. 그림에서처럼 그래프 G는 두 번의 과정을 거쳐 G1, G2, G3, G4, G5, G6로 6개의 서브그래프로 분할되었다. 그리고 분할된 그래프내에서 다시 서브그래프로 분할하기 위해 제거될 간선을 표시한다. 분할된 6개의 서브그래프 중에서 G1, G2, G3의 친밀도 평균값은 각 1.09, 1.08, 1.56이다. <그림 14>에서 G2, G3 서브그래프는 HCWS 알고리즘의 조건에 의해 다시 서브그래프로 분할된다. G2의 전체의 친밀도 평균값이 1.08이기 때문에 서브그래프로 분할하기 위하여 제거될 간선의 평균값이 1.08 이하이면 서브그래프로 분할된다. G2에서 W-X의 간선의 평균값 0.5, CC-EE의 간선의 평균값 0.4, S-T의 간선의 평균값 0.7, P-Q-R의 간선의 평균값 0.6으로 G2의 전체 친밀도 평균값보다 작은 간선을 차례대로 제거하여 서브그래프로 분할한다.



<그림 14> HCWS을 이용한 연구자 클러스터링 2단계

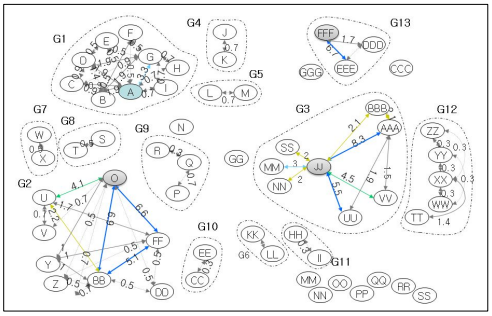
<그림 15>는 G2, G3를 각 그래프의 친밀도 평균값보다 작은 간선을 차례대로 제거하여 서브그래프로 재분할한 결과이다.



<그림 15> HCWS을 이용한 연구자 클러스터링 3단계

<그림 16>은 HCWS 알고리즘을 이용하여 최종적으로 13개의 서브그래프로 분할된 결과이다. 이 13개의 서브그래프는 강한 연결로 분류되어 더 이상 분할되지 않는 친밀도가 높은 그룹을 의미한다.

<그림 16>에서처럼 HCWS 알고리즘에서는 JJ를 기준으로 MM, NN, UU, VV, SS 그리고 O를 기준으로 U 그리고 FFF을 기준으로 EEE, DDD은 강한 연결로 하나의 그룹에 포함되었다. 하지만 <그림 12>에서처럼 HCS 알고리즘에서는 약한 연결로 분리되었다. 그리고 <그림 12>의 HCS 알고리즘에서 강한 연결로 분리되지 않는 R, CC, EE 그리고 ZZ, YY는 HCWS 알고리즘에서는 다른



<그림 16> HCWS을 이용한 연구자 클러스터링 결과

그룹으로 분리됨을 알 수 있다. HCWS 알고리즘에서는 공저 관계가 약한 연결로 분류되어 연결이 끊기는 문제점을 개선하였다. 그리고 연구자간 논문 공저에 대한 친밀도가 평균 수치보다 낮음에도 불구하고 강한 연결로 분류되어 더 이상 클러스터링이 되지 않는 문제점도 함께 개선하였다.

V. 결론 및 향후 연구 방향

본 연구에서는 연구자 네트워크상에서 개인간의 의미적인 친밀도를 반영하여 연구자 집단을 그룹핑하기 위한 HCWS 알고리즘을 제안하였다. 본 논문에서는 연구자간 관계의 정도를 표현할 수 있도록 연구자간의 친밀도를 사용하였다. 친밀도는 논문의 공저자 수와 공저 논문 편수에 의해 연구자간의 친밀도를 정의하였다.

본 논문에서 제안한 기법의 타당성을 검증하기 위하여 컴퓨터공학과 교수들이 최근 5년간 발표한 80여편의 논문 실적을 대상으로 실험을 하였다. 연구자간 관계는 논문에서 공저자를 추출하여 구성하였다. 그리고 각 논문에 대한 공저자 수를 가중치로 반영하여 연구자간의 친밀도를 정의하였다. 그리고 HCS 알고리즘과 HCWS 알고리즘을 이용한 결과를 비교 분석하였다.

실험 결과, HCWS 알고리즘은 연구자간의 의미적인 관계를 단순 연결로 표현하는 HCS 알고리즘에 비해 친밀도에 의해 연결하여 보다 사실적인 클러스터링이 가능함을 보였다. 그리고 HCWS 알고리즘은 연구자간 관계가 강함에도 불구하고 약한 연결로 분류되어 분리되거나 연구자간의 관계가 약함에도 불구하고 강한 연결로 분류되는 기존 HCS 알고리즘의 문제점을 개선할 수 있음을 보였다.

본 논문에서 제안한 기법은 연구자간의 협업이력이나 친밀도 등과 같은 의미적인 관계를 반영한 연구자 네트워크를 효율적으로 구성하는 분야에 응용할 수 있다. 또한 전문가 집단 검색을 통한 협업 모델 구성이나 연구자 커뮤니티에서 친밀도 분석을 통한 효과적인 연구자 커뮤니티 구성에 응용할 수 있다.

향후 연구 방향으로, 연구자 네트워크는 본 논문에서 사용한 공동 연구 실적 데이터를 기반으로 구성하는 경우가 많다. 하지만 연구 실적 데이터의 수집이 현실상 어려운 점을 감안하여 연구 실적이외에 온라인상에서 지식 공유와 같은 행동을 분석하여 연구자 네트워크를 구성하고 이를 기반으로 연구자를 클러스터링하기 위한 기법에 대한 연구를 진행할 것이다.

참고문헌

[1] S. E. Seibert, M. L. Kraimer, R. C. Liden, “A Social Capital Theory of Career Success,” The Academy of Management Journal, Vol. 44, No. 2, 2001, pp. 219-237.

[2] 김지수, “디지털 인맥과 인간 관계망(human network),” 정보통신정책, Vol. 16, No. 16, 2004, pp. 1-19.

[3] J. Mori, Y. Matsuo, K. Hasida, and M. Ishizuka, “Web Mining Approach for a User-centered Semantic Web,” Workshop on End User Aspects of the Semantic Web, 2005.

[4] 이승우, 김평, 정한민, 구회관, 성원경, “학술 연구

분야에서의 추론 기반 연구자네트워크 생성," 한국컨텐츠학회 추계종합학술대회 논문집, Vol. 4, No. 2, 2006, pp. 90-94.

[5] 전인하, "사회연결망 구성을 위한 OWL 기반의 사용자 모델링 기법," 창원대학교 석사학위 논문, 2006.

[6] 성원경, 정한민, 박동인, "OntoFrame-K: 협업 연구 지원을 위한 시멘틱 웹 기반 지식정보 공유·유통 플랫폼," 한국정보과학회논문지, Vol. 33, No. 4, 2006, pp. 65-74.

[7] Y. Zhang, G. Xu, X Zhou, "A Latent Usage Approach for Clustering Web Transaction and Building User Profile," Lecture Notes in Computer Science, Vol. 3584, 2005, pp. 31-42.

[8] A. Smirnov, M. Pashkin, N. Chilov, T. Levashova, "Ontology-Based Users & Request Clustering in Customer Service Management System," Lecture Notes in Computer Science, Vol. 3305, 2005, pp. 231-246.

[9] D. L. McGuinness, "Ontological Issues for Knowledge-Enhanced Search," Proc. of the Formal Ontology in Information Systems, 1998, pp. 302-316.

[10] 최옥경, 한상용, "자동화된 통합 프레임워크를 위한 시맨틱 웹 기반의 정보 검색 시스템," 한국정보처리학회 논문지, Vol. 13, No. 1, 2006, pp. 129-136.

[11] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, M. Ishizuka, "POLYPHONET: An advanced social network extraction system from the Web, "Proc. of the 15th Int'l conference on World Wide Web, 2006, pp. 397-406.

[12] 서치종, 서의호, "Knowledge Network 구축 수단으로서의 Social Network Analysis에 관한 연구," 한국경영과학회 학술대회논문집 논문집, 2002, pp. 317-322.

[13] Y. Fu, K. Sandhu, M. Shih, "Clustering of Web Users Based on Access Patterns," Proc. of the WEBKDD, Vol. 5, 1999, pp. 560-567.

[14] E. Hartuv, R. Shamir, "A Clustering Algorithm based on Graph Connectivity," Information processing letters, Vol. 76, 2000, pp. 175-18.

■ 저자소개 ■

문 현 정
Mun, Hyeon Jeong

2007년~현재 하이브레인넷 수석연구원
2004년~2007년 창원대학교 연구교수
2003년 창원대학교 컴퓨터공학과 (공학박사)
1996년 창원대학교 전자계산학과 (이학석사)
1994년 한국방송대학교 전자계산학과(이학사)

관심분야 : KDD, 온톨로지마이닝, 시맨틱웹
E-mail : hjmun@hibrain.net

이 상 민
Lee, Sang Min

2008년~현재 대림아이앤에스(주) 대리
2007년 창원대학교 컴퓨터공학과 (공학석사)
2003년 창원대학교 경영학과 (경영학사)

관심분야 : 사용자클러스터링, 쇼셜네트워크
E-mail : leemet@dmc.co.kr

우 용 태
Woo, Yong Tae

1987년~현재창원대학교 컴퓨터공학과 교수
1995년 경북대학교 전자공학과(공학박사)
1984년 경북대학교 전자공학과(공학석사)
1982년 경북대학교 전자공학과(공학사)

관심분야 : 데이터마이닝, 온톨로지마이닝, 시맨틱웹
E-mail : ytwoo@sarim.changwon.ac.kr

논문접수일 : 2009년 3월 16일
수 정 일 : 2009년 4월 20일
게재확정일 : 2009년 4월 27일