확률적 온톨로지와 연구자 네트워크를 이용한 심사자 자동 추천에 관한 연구

Automatic Recommendation of Panel Pool Using a Probabilistic Ontology and Researcher Networks

이정연(JungYeoun Lee)*, 이재윤(Jae-Yun Lee)**,
정한민(Hanmin Jung)***, 강인수(In-Su Kang)****, 신숙경(SukKyung Shin)*****

초 록

심사자 자동추천시스템은 심사 대상에 대한 포괄성, 전문성, 공정성, 타당성을 확보할 수 있도록 설계되어야 한다. 이를 위해 본 연구는 다면적인 학문분야분류표의 각 범주 간 연관성을 자동으로 산출할 수 있는 확률적 온톨로지를 적용하여 포괄적으로 심사자 추천 범위를 넓히고 전문성을 반영한 심사자 랭킹을 가능하도록 한다. 또한 연구자 간의 멘터, 공저역, 공동연구를 포함하는 연구자 네트워크를 구축하고 이를 심사자 배제 규칙으로 활용함으로써 공정한 심사자 추천이 이루어질 수 있도록 한다. 아울러, 전문가들을 통해 상기 방법론과 패널 결과를 검증 받아 타당성 있는 시스템이 갖추어야 할 방향을 제시한다.

ABSTRACT

Automatic recommendation system of panel pool should be designed to support universal, expertness, fairness, and reasonableness in the process of review of proposals. In this research, we apply the theory of probabilistic ontology to measure relatedness between terms in the classification of academic domain, enlarge the number of review candidates , and rank recommendable reviewers according to their expertness. In addition, we construct a researcher network connecting among researchers according to their various relationships like mentor, coauthor, and cooperative research. We use the researcher network to exclude inappropriate reviewers and support fairness of reviewer recommendation process. Our methodology recommending proper reviewers is verified from experts in the field of proposal examination. It propose the proper method for developing a resonable reviewer recommendation system.

키워드 : 온톨로지, 연구자 네트워크, 심사자 추천시스템, 확률적 온톨로지, 과제관리시스템
ontology, probabilistic ontology, researcher networks, panel pool, reviewer

* 한국학술진흥재단 지식정보센터 지식확산팀 (shampoo@krf.or.kr)
** 경기대학교 문헌정보학과 교수 (memexlee@kgu.ac.kr)
*** 교신저자, 한국과학기술정보연구원 정보서비스연구팀 책임연구원 (jhm@kisti.re.kr)
**** 한국과학기술정보연구원 정보서비스연구팀 선임연구원 (dbaisk@kisti.re.kr)
*****한국학술진흥재단 지식정보센터 학술정보팀 팀장 (skshin@krf.or.kr)
■ 논문접수일자 : 2007년 5월 16일
■ 게재확정일자 : 2007년 6월 30일

1. 서 론

1.1 연구의 목적

연구과제를 지원하는 기관에서는 공정하고 전문성 있는 심사 제도를 활용하여 연구 수행 능력을 갖춘 연구자를 올바로 선정하고 그들이 우수한 연구결과를 산출할 수 있는 제도를 마련 해야 한다. 학술진흥재단에서는(이하 학진으로 명명함) 심사 전문성을 제고하기 위하여 프로그램관리자(PM) 제도를 도입하고 있으며, 연구과제를 심사할 때 학문적 우수성 외의 다른 변수에 의하여 평가 결과가 영향을 받지 않도록 상피제도를 도입하여 심사자를 선정하고 있다. 그리고 연구과제 선정의 타당성을 높이기 위하여 동료평가제도(peer review)를 활용하고 있는데 신청된 과제 중에서 유사한 과제들을 클러스터링 하여 패널을 구성하고 심사자들이 이 과제들을 평가하는 패널심사제를 채택하고 있다.

과제관리시스템은 과제신청, 심사, 선정, 사후관리 등 프로세스별로 구성되어 있으며, 심사과정 프로세스 일부에 심사후보자 추천시스템이 지원되고 있다. 현행 심사자 추천시스템에서는 과제 신청자가 심사희망분야를 학진 학문분야 분류표에 근거하여 1순위에서 3순위별로 지정하면, 이에 맞는 각 학문별 심사패널이 구성된다. 그러나 학문 분야에 따라 심사자 풀이 적거나 많게 추천되며, 추천의 우선순위가 나타나지 않는다는 문제점이 있다. 패널심사는 하나의 과제 당 1인 심사자 매칭이 아니라 패널에 할당된 과제 모두를 대상으로 제한된 심사자에 의해 수행되는 심사이다. 패널 심사의 관건은 신청과제의 학문분야의 다양성과 심사단의 구성여부에 따라 심사에 영향을 미칠 수 있다는 데에 있다.

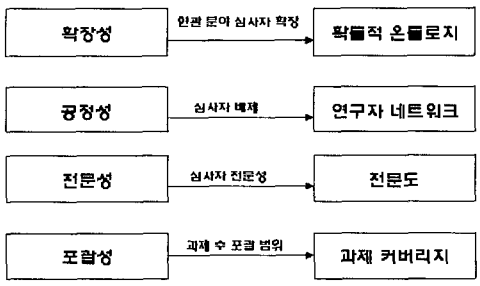
학진의 학문분야분류표는 연구자, 과제, 성과물 등에 부여되는 통일 분류표이므로 과제관리를 위한 여러 업무에서 기본적인 주제 온톨로지가 될 수 있다. 최근에 온톨로지 및 시멘틱 웹 기술이 주목받고 있는 것이 현실이지만 효용적 적용이 가능한 응용분야가 많지 않은 것이 현실이다. 그러나 본 연구에서는 과제, 인력 및 연구성과물 정보는 비교적 명확한 식별자로서 대상을 구분해 낼 수 있으며 시멘틱 웹 기술을 적용한 서비스를 구축할 수 있을 것이라 판단하였다. 또한 열거형인 학진 학문분야 분류표에서 다면적인 범주간 연관성을 자동으로 산출할 수 있는 확률적 온톨로지를 적용해 보고자 하였다.

따라서 본 연구는 현행 시스템을 보완하여 심사자를 보다 객관적이고, 포괄적이며, 전문적인 심사자 풀(pool)을 추천 하는 시스템으로 재설계하고 이를 적용할 수 있는 방안을 제시하는데 그 목적이 있다.

1.2 연구의 방법

현행 심사자 추천시스템의 보완을 위하여 다음과 같은 연구 절차에 의하여 연구를 수행하였다. 첫째, 과제 심사의 기준이 되는 학진의 학문분야분류표를 분석하여 열거형 계층적 분류표로서의 단점을 보완할 수 있는 확률적 주제 온톨로지 모델을 개발하였다. 둘째, 과제를 신청한 연구자들과 심사자들의 상피를 적용시킬 수 있는 방안으로 연구자 네트워크를 구축

하였다. 셋째, 현행 심사자 추천시스템을 분석하여 개선할 수 있는 프로세스를 보완하고 이를 실제로 적용할 수 있는가를 실험적 모델을 통해서 확률적 온톨로지와 연구자 네트워크를 적용하여 심사자 확장과 배제의 효용성을 측정해 보았다. 넷째, 심사 패널을 실험적으로 구성하여 개선 시스템을 구현하고 이를 해당 분야의 전문가 집단에게 검증을 의뢰하고 이를 분석하였다. 다음 〈그림 1〉은 심사자 자동추천에 관한 개념도이다.



〈그림 1〉 심사자 자동추천 개념도

2. 자동추천을 위한 정보자원

2.1 학문분야 분류표

현재 심사자 자동추천을 위한 정보자원으로는 학진 학문분야분류표가 사용되고 있다. 학진 학문분야분류표는 학진의 모든 활동에 있어서 주제상의 기준이 된다. 연구자, 과제, 성과물 등에 부여되는 통일 분류표이기 때문에 과제관리를 위한 여러 업무에서 기본적인 주제 온톨로지가 될 수 있다. 특히 학문분야별 주제 시소러스가 구축되기 이전에는 운용 가능한 유일한 주제 온톨로지라고 할 수 있다.

학문분야분류표는 일곱 범주로 나뉘는 대분류 구분에서부터 시작하여 중분류, 소분류, 세분류의 네 단계로 구성된 계층적 분류표이다. 세분류 항목의 수는 분류코드가 A000000인 인문학(대분류, 중분류, 소분류, 세분류 모두 인문학)에서부터 분류코드가 H990000인 학제간연구(대분류는 복합학, 중분류, 소분류, 세분류 모두 학제간연구)에 이르기까지 총 3,374 범주에 달한다.

이와 같이 전 학문분야를 포괄하는 분류표로서 학문분야분류표는 전형적인 열거형 계층적 주제분류표이다. 일반적으로 열거형 분류표(계층분류표)는 개념간의 다양한 관계를 정확하게 제시하기 어렵고, 주제의 특수한 관점을 충분히 표현하지 못하며, 개념간의 유연한 결합이 불가능하다는 한계를 가지고 있다(김태수 2000).

학진 학문분야분류표도 이런 열거형 분류표의 한계가 그대로 적용되므로 다면적인 주제를 표현하기 어렵고 학문간의 경계를 넘나드는 학제성을 반영하지 못한다. 이 때문에 현행 학문분야분류표에서는 유사한 세분류 범주 명칭이 상이한 중분류나 대분류 범주 이하에 흩어져 있는 경우가 흔하다. 〈표 1〉과 〈그림 2〉-〈그림 4〉에 그와 같은 사례를 몇 가지만 소개하였다.

〈그림 2〉과 같이 동일한 명칭의 세분류 범주가 대분류 범주로는 같으나 중분류 범주부터 달라지는 경우가 있는가 하면, 〈그림 3〉의 '체육교육학'과 '스포츠교육' 처럼 유사한 명칭의 세분류 범주가 대분류 범주부터 '사회과학'과 '예술체육'으로 다른 줄기에 속하는 경우도 흔히 나타난다. '영어교육'과 '영어교육학', '수

〈표 1〉 유사명칭/동일명칭 세분류 범주 사례

대분류	중분류	소분류	세분류
사회과학	경영학	전자상거래	전자상거래
공학	컴퓨터학	인터넷정보처리	전자상거래
인문학	영어와문학	영어교육	영어교육
사회과학	교육학	교과교육학	영어교육학
사회과학	교육학	교과교육학	수학교육학
자연과학	수학	수학일반	수학교육
사회과학	교육학	교과교육학	체육교육학
예술체육	체육	체육일반	스포츠교육
사회과학	경영학	분야별경영	스포츠경영
예술체육	체육	스포츠경영학	스포츠경영학
사회과학	교육학	교육공학	교육공학
공학	컴퓨터학	컴퓨터교육	교육공학
사회과학	경제학	권역경제	지역경제
사회과학	지역개발	지역경제	지역경제

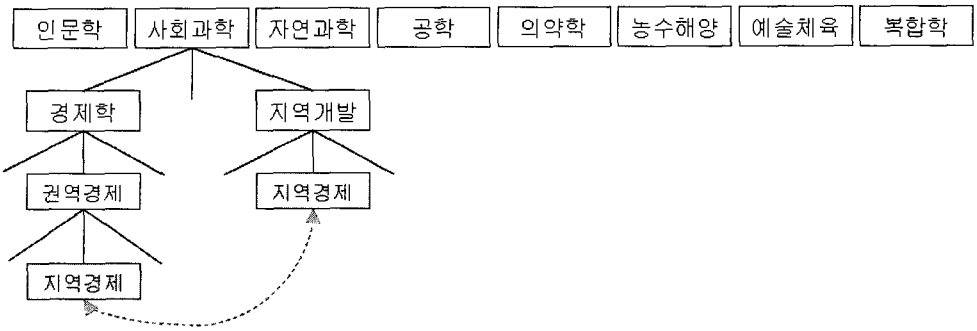
학교육'과 '수학교육학'과 같은 경우는 유사하다기 보다는 의미상 같은 명칭이 상이한 대분류 범주 아래에 중복된 경우이다. 아예 〈그림 4〉의 '전자상거래'와 같이 대분류 범주를 가로질러 동일한 명칭의 세분류 범주가 나타나는 경우도 여러 건이다. 마찬가지 경우로 '교육공학'도 공학과 사회과학에 각각 속하는 것으로 되어 있다.

이와 같이 학제성이 강한 분야를 다루기 위해서 동일하거나 매우 유사한 세분류 범주를 대분류 구분을 넘나들면서 중복해서 설정한 것은 열거형 계층적 분류표로서는 어쩔 수 없는 일이다.

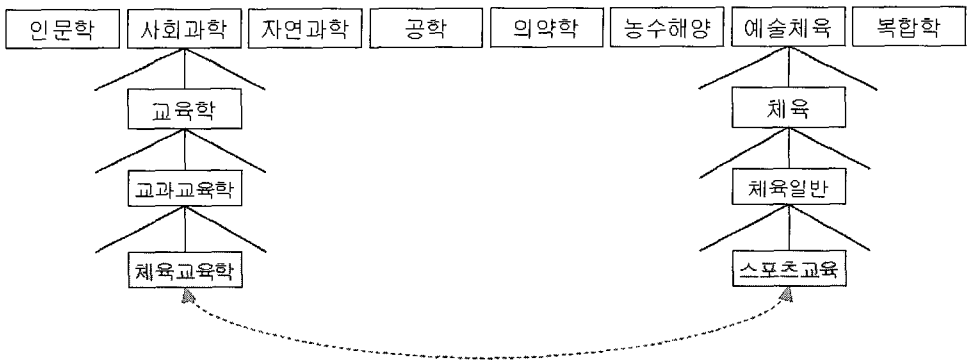
그러나 분류표의 주제 계층을 넘나드는 세분류 범주간의 관계가 현행 학문분야 분류표에서는 구조적으로 전혀 드러나지 않기 때문에 연구과제 관리나 성과물 정보 서비스에 큰 제약이 되고 있다. 그나마 앞의 사례에서처럼 세분류 범주의 명칭이 동일하거나 거의 일치하는 경우에는 분야명 검색을 통해서 다소간 보완할 수 있는 경우도 있지만, 명칭이 다르면서 내용상 관련성이 깊은 세분류 범주간 관계는 이를 통해서도 파악하기가 불가능하다. 이 때문에 과제의 신청이나 심사에서 더 적절한 주제 범주나 심사자를 결정하기 어려운 경우가 발생하거나, 과제물 정보 서비스에서 주제 브라우징을 할 때 관련 자료가 속한 유사 범주를 이용자가 찾지 못하는 경우도 나타난다.

이와 같은 학문분야 분류는 계층 구조에 구속되기 때문에 계층 구조상으로는 멀리 떨어져

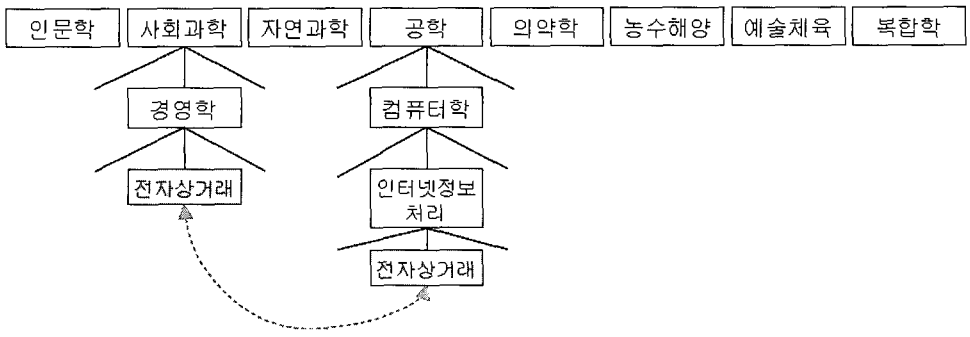
있으나 관련이 깊은 유사분야를 파악하기가 어
렵다.



〈그림 2〉 상이한 중분류 범주에 속한 동일 명칭의 세분류 범주



〈그림 3〉 상이한 대분류 범주에 속한 유사 명칭의 세분류 범주



〈그림 4〉 상이한 대분류 범주에 속한 동일 명칭의 세분류 범주

2.2 확률적 온톨로지

2.2.1 확률적 온톨로지의 개념과 구축법

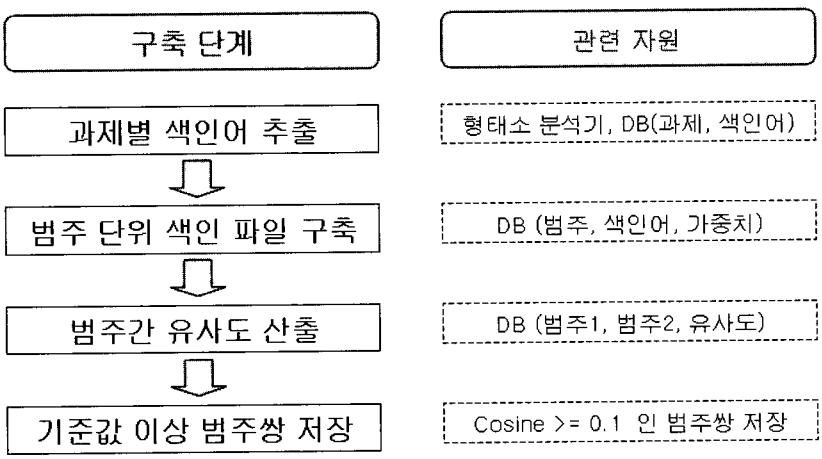
이 연구에서 제안하는 것은 기존에 정보검색이나 데이터마이닝 분야에서 개발되어온 통계적 연관성 측정방식(Chung & Lee 2001)을 이용하여 세분류 범주간 연관성을 통계적, 혹은 확률적으로 파악하여 도출하는 일종의 확률적 온톨로지(probabilistic ontology)이다. 확률적 온톨로지는 불확실성을 온톨로지에 도입하려는 시도의 산물이다. 확률적 온톨로지에 대한 관심이 최근 증가하고는 있으나 아직까지 널리 받아들여지는 정의가 제시되지는 못하고 있다(Costa & Laskey 2006).

여기서는 확률적 온톨로지를 확률적 관계(probabilistic relations)로 구성된 온톨로지라고 정의하기로 한다. 확률적 관계는 기존의 온톨로지에서 개념간의 관계가 확정적(deterministic)인 것과 달리 확률적으로 연결 강도가 표현되는 관계이다.

확정적인 관계는 일반적으로 해당 영역의 이론가나 실무자가 판단하여 온톨로지를 구축하게 된다. 반면에 확률적인 관계는 경험적인 수단을 통해서 두 개념 사이의 관계 확률을 구하게 된다. 확률적인 관계를 자동으로 산출하는 방법은 다음의 두 가지로 나눌 수 있다.

첫째, 두 개념 사이의 직접적인 확률을 구한다. 이 경우에는 두 개념이 동시에 나타난 경우의 확률, 즉 동시출현확률을 두 개념 사이의 관계 확률로 삼을 수 있다. 직접적인 관계에 기반을 두고 있으므로 근거가 확실한 반면에, 개념이 배타적으로 사용되는 경우에는 적용이 불가능하다는 단점이 있다. 즉, 항상 배타적으로 한 개념만 할당되는 상황이라면 동시출현이 발생하지 않으므로 관계 확률을 구할 수가 없다.

둘째, 두 개념 사이의 간접적인 확률을 구한다. 이는 각 개념이 사용되거나 할당된 맥락을 단서로 하여 두 개념의 맥락이 유사한 정도를 개념간의 관계 확률로 삼는 방법이다. 직접 동



〈그림 5〉 연구비 신청 과제 정보를 이용한 확률적 온톨로지 구축 절차

시출현하지 않는 개념끼리의 관계도 출현 맥락의 비교를 통해서 파악할 수 있다. 따라서 직접적인 동시출현 확률을 구하는 것보다 적용할 수 있는 경우가 넓다.

여기서는 학진 학문분야분류표라는 확정적인 온톨로지가 존재하는 상황을 전제로 한다. 따라서 구축하고자 하는 확률적 온톨로지는 기존 온톨로지인 분류체계가 존재하는 상황에서 개념간의 부분적인 일치 정도를 반영하는 보조 온톨로지라고 할 수 있다.

2.2.2 신청과제 정보를 이용한 확률적 온톨로지 구축

이 연구에서는 학진 학문분야분류표의 각 범주간 연관성을 자동으로 산출하여 관계 확률로 삼고자 하였다. 학문분야분류표는 배타적인 계층적 분류표로서 학진의 업무에서는 대부분 단일 분류항목만 할당된다. 따라서 두 분류항목 사이의 직접적인 연관성이 아닌 간접적인 연관성을 구해야 한다.

간접적인 연관성을 구하기 위한 분류항목의 맥락 정보로는 학진의 연구비 지원을 신청한 과제 정보를 이용하기로 한다. 과제 정보를 이용한 확률적 온톨로지 구축과정은 〈그림 5〉와 같다.

학진에 신청된 과제는 모두 학문분야를 세분류 단위까지 지정하도록 되어 있고, 기본적으로 과제명과 연구요약문을 입력하도록 되어 있다. 연구요약문은 '연구목표', '기대효과', '연구요약', '한글키워드', '영문키워드'로 구성된다. 세분류 범주간 연관성은 각 범주로 신청된 과제의 제목과 한글키워드로부터 자동 추출한 색인어의 일치도를 기준으로 산출할 수 있다. 범주간 연관성 산출 과정을 정리하면 다음과 같다.

우선 범주 c_x에서 색인어 t_i의 가중치 $w(t_i, c_x)$는 다음과 같이 산출한다.

$$u(t_i, c_x) = f(t_i, c_x) \times \log\left(\frac{C}{cf(t_i)}\right)$$

이 공식은 정보검색분야에서 흔히 사용하는 TF×IDF 공식을 응용한 것이다. 여기서 $f(t_i, c_x)$는 범주 c_x에 속한 과제에서 색인어 t_i의 출현빈도이고, C는 전체 세분류 범주 수(현재 3,374), $cf(t_i)$는 색인어 t_i가 속한 범주의 종수이다. 따라서 색인어가 해당 세분류 범주에 소속된 과제에서 많이 출현할수록, 그리고 색인어가 출현한 세분류 범주의 종수가 적을수록 높은 가중치를 가진다.

이와 같이 각 세분류 범주를 색인어 가중치 벡터로 표현한 다음, 범주 c_x와 c_y간 연관성 $r(c_x, c_y)$는 다음과 같은 코사인 계수 공식으로 산출할 수 있다. 값의 범위는 최저 0에서 최고 1 사이가 된다.

$$r(c_x, c_y) = \frac{\sum_{i=1}^{n}(u(t_i, c_x) \times u(t_i, c_y))}{\sqrt{\sum_{i=1}^{n}(u(t_i, c_x))^2 \times \sum_{i=1}^{n}(u(t_i, c_y))^2}}$$

실제로 2001년부터 2005년까지 5년간 학진에 신청된 과제의 과제정보를 대상으로 색인어를 추출하고 연관성을 산출해보았다. 색인어 추출은 21세기 세종계획에서 개발하여 공개된 지능형형태소분석기를 사용하였고, 형태소분석 결과 중에서 일반명사와 고유명사로 태깅된 형태소를 색인어로 채택하였다.

신청 과제 중에서 학술지 발간 지원이나 교

과과정 개발 지원, 학생 연수 지원 사업과 같이 주제성이 약한 사업을 제외한 모든 사업에 신청된 80,502건을 대상으로 하였다. 과제의 제목과 한글키워드 필드로부터 추출한 명사 색인어를 단서로 하여 앞에서와 같은 방법으로 세분류 범주간 통계적 연관성을 산출하여 연관성이 높은 쌍을 일부만 제시하면 〈표 2〉와 같다.

〈표 2〉에 제시된 세분류 범주 쌍은 대부분 매우 유사한 주제임을 알 수 있다. 앞에서 언급되었던 사회과학의 '체육교육학'과 예술체육의 '스포츠교육'은 0.907로 세 번째로 연관성이 높게 나타났다. 또한 이 표에는 나타나

〈표 2〉 확률적 연관성이 높은 세분류 범주쌍 사례 (연관성 0.7 이상)

연관성	세분류 범주 명	
	범주 1	범주 2
0.955	인문학_영어와문학_영어교육_영어교육	사회과학_교육학_교과교육학_영어교육학
0.930	사회과학_교육학_교과교육학_수학교육학	자연과학_수학_수학일반_수학교육
0.907	사회과학_교육학_교과교육학_체육교육학	예술체육_체육_체육일반_스포츠교육
0.814	사회과학_교육학_분야교육_음악치료교육	예술체육_음악학_음악교육학_음악교육학
0.802	사회과학_교육학_분야교육_음악치료교육	예술체육_음악학_음악사회학_음악사회학
0.795	사회과학_경영학_물류관리_물류정책	사회과학_무역학_국제운송및물류_국제운송및물류
0.773	예술체육_의상_패션디자인_패션디자인	자연과학_생활과학_의류학_패션디자인
0.766	사회과학_사회복지학_청소년복지_청소년복지	자연과학_생활과학_아동학_청소년학
0.755	사회과학_사회복지학_영유아복지_영유아복지	자연과학_생활과학_아동학_아동보육
0.753	사회과학_경영학_분야별경영_스포츠경영	예술체육_체육_스포츠경영학_스포츠경영학
0.748	인문학_한국어와문학_국어교육_국어교육	사회과학_교육학_교과교육학_국어교육학
0.737	사회과학_사회복지학_노인복지_노인복지	의약학_간호학_노인간호_노인간호
0.726	사회과학_교육학_분야교육_음악치료교육	예술체육_음악학_음악사학_한국음악사
0.722	사회과학_지역개발_도시계획/설계/개발_도시계획/설계/개발	공학_건축공학_건축의장_단지/도시
0.715	인문학_언어학_통사론(언어학)_통사론(언어학)	인문학_영어와문학_영어학_통사론(영어학)
0.714	사회과학_교육학_분야교육_재활치료교육	의약학_재활의학_신체/직업재활_신체/직업재활
0.713	인문학_문학_비교문학_비교문학	인문학_한국어와문학_국문학_문학비평(국문학)
0.712	사회과학_교육학_교육상담_교육상담	복합학_심리과학_상담심리/심리치료_상담심리/심리치료
0.711	사회과학_사회복지학_가족복지_가족복지	자연과학_생활과학_가족학_가족관계
0.709	복합학_뇌과학_지능로봇_지능로봇	공학_제어계측공학_로봇공학/로보틱스_로봇공학/로보틱스
0.707	사회과학_정치외교학_한국정치_북한정치/통일	사회과학_지역학_동아시아_북한
0.705	복합학_뇌과학_지능로봇_지능로봇	공학_기계공학_동역학및제어_로봇공학

※ 각 범주는 학진 학문분야분류표의 대분류-중분류-소분류-세분류 순으로 표기하였으며 실제 분석 단위는 마지막 세분류 범주임.

〈표 3〉 세분류 범주별 연관성(확률 관계) 상위 범주 산출 사례

기준 범주	연관성	연관성 상위 범주
사회과학_경제학_권역경제_지역경제	0.649	사회과학_지역개발_지역경제_지역경제
	0.525	사회과학_경제학_경제학일반_경제성장/발전/개발경제
	0.519	사회과학_지역개발_지역개발일반_지역개발정책
	0.494	사회과학_경제학_국제/세계경제_국제/세계경제
	0.485	사회과학_경제학_권역경제_국가별경제
사회과학_지역개발_지역경제_지역경제	0.649	사회과학_경제학_권역경제_지역경제
	0.589	사회과학_지역개발_지역개발일반_지역개발정책
	0.432	사회과학_지리학_인문지리학_경제지리
	0.417	사회과학_경제학_분야별경제_산업/서비스경제
	0.408	사회과학_경제학_경제학일반_경제성장/발전/개발경제
사회과학_경영학_전자상거래_전자상거래	0.699	사회과학_경영학_경영정보시스템_경영정보시스템
	0.669	공학_컴퓨터학_인터넷정보처리_전자상거래
	0.512	사회과학_경영학_경영정보시스템_정보기술관리
	0.486	공학_산업공학_e-Business_공급체인통합
	0.480	사회과학_무역학_무역통신및전자무역_무역통신및전자무역
공학_컴퓨터학_인터넷정보처리_전자상거래	0.669	사회과학_경영학_전자상거래_전자상거래
	0.574	사회과학_경영학_경영정보시스템_경영정보시스템
	0.429	사회과학_법학_사법_상행위/신용거래/전자거래법
	0.425	사회과학_경영학_경영정보시스템_지능형의사결정시스템
	0.365	사회과학_경영학_경영정보시스템_경영전산처리
사회과학_교육학_교과교육학_체육교육학	0.907	예술체육_체육_체육일반_스포츠교육
	0.549	예술체육_체육_사회/생활체육_사회/생활체육
	0.482	예술체육_체육_특수/장애인체육_특수/장애인체육
	0.436	예술체육_체육_체육일반_체육사
	0.396	사회과학_교육학_교수이론/교육방법/교수법_교수이론/교육방법/교수법
예술체육_체육_체육일반_스포츠교육	0.907	사회과학_교육학_교과교육학_체육교육학
	0.618	예술체육_체육_사회/생활체육_사회/생활체육
	0.502	예술체육_체육_체육일반_체육사
	0.483	예술체육_체육_특수/장애인체육_특수/장애인체육
	0.455	예술체육_체육_체육일반_스포츠철학

※ 각 범주는 학진 학문분야분류표의 대분류-중분류-소분류-세분류 순으로 표기하였으며 실제 분석 단위는 마지막 세분류 범주임.

있지 않지만 경제학의 '지역경제'와 지역개발의 '지역경제'는 0.649, 사회과학의 '전자상거래'와 공학의 '전자상거래'는 0.669로 역시 확률적 연관성이 매우 높게 나타났다.

특정 범주를 중심으로 연관성이 높은 분야를 파악한 사례를 살펴보면 〈표 3〉과 같다. 여기서는 기준 범주별로 연관성 상위 다섯 개 범주만 제시하였다.

〈표 3〉에서 볼 수 있듯이 공학의 '전자상거래'처럼 연관성 상위 다섯 개 범주가 모두 이질적인 대분류인 사회과학에 속하는 범주인 경우가 있는가 하면, 이와 달리 사회과학의 '전자상거래'는 연관성 상위 다섯 개 범주가 이질적인 공학에 둘, 동질적인 사회과학에 셋으로 다양하게 분포함을 볼 수 있다.

이와 같이 통계적 유사도를 이용하여 산출한 범주간 확률적 관계는 확률적 온톨로지를 구성하여 계층적 열거식 분류표에서는 파악할 수 없는 세분류 범주간 관계를 추가로 제공하게 된다. 다만 확률적인 접근을 취하는 만큼, 어느 정도 자료의 양이 확보되는 범주에 대해서만 적용하는 것이 바람직하다. 적어도 1년에 한 과제 이상(5년간 다섯 과제 이상) 신청 과제가 있어야 산출된 확률 관계를 신뢰할 수 있을 것이다.

이상의 과정을 통해 구축된 확률적 온톨로지는 학진(KRF) 학문분야분류표에 기반을 두었으므로 *p*KRF(probabilistic KRF) 분류체계라고 부르기로 한다.

2.3 연구자 네트워크

연구자 네트워크는 사회망(Social Network)의 일종으로 연구자를 노드로 하고 연구자들 간의 사회적 관계를 노드 간 링크로 모델링하여 표현된다[Wasserman and Faust, 1994; Scott, 2000; Newman, 2003]. 연구자 간의 사회적 관계로는 연구 성과물과 관련하여 논문/저역서 공동 저술, 논문/저역서 인용/피인용, 공동 과제 수행, 특허 공동 출원 등이 있으며, 동일 기관/부서 소속이나 학연 측면에서 동문, 지도교수/지도학생 관계 등도 포함된다.

상기의 여러 연구자 간 관계들 중, 공동 저술 관계와 인용/피인용 관계는 특별히 공저망(co-authorship network) [Barabasi et al., 2002; Newman, 2001a; Newman, 2001b]과 인용망(author-citation network) [Redner, 1998; Seglen, 1992; White et al., 2004]으로 모델링하여 연구되고 있다. 공저망으로부터 클러스터링 기법을 통해 분야별 공동 연구 그룹의 수와 규모를 자동 탐색해 낼 수 있다. 인용망은 주로 수작업으로 구축되어 저널 등의 피인용 지수 계산에 활용되어 왔으나, 최근 참고문헌 정보로부터의 정보추출을 통한 자동 인용망 구축 관련 연구도 활발하다. 공저나 인용 관계에 비해 다른 관계들은 상대적으로 그것의 획득이 용이하지 않아 사회망에 적극적으로 적용된 사례를 찾아보기 힘들다.

심사자 추천 측면에서, 연구자 네트워크로 모델링된 연구자들 간의 사회적 관계는 상호간 부적격 심사자를 표현하는 것으로 고려될 수 있다. 예를 들어, 한편의 논문을 공동 저술한 저자 A와 B가 있을 때, A나 B는 각각 B나 A가 제안한 과제의 평가자로 추천되기에 부적절한 정도의 개인적인 친분이 이미 형성되어 있어 편파적인 과제 평가가 이루어질 가능성이

있는 것이다. 연구자 네트워크 내 다른 관계들도 그것들이 사람들 간의 온라인/오프라인 상의 사회적 관계를 표현하고 있다는 관점에서 볼 때, 정도의 차이는 있겠으나 심사자 추천 태스크에서 부적절 심사자를 배제하는 용도로 활용될 수 있을 것이다.

본 연구에서는 이 절의 앞에서 기술한 연구자들 간 모든 사회적 관계들을 연구자 네트워크로 표현하고, 이를 심사자 추천 과정에서 부적격 심사자 배제를 위해 사용하고자 한다.

3. 자동 심사자추천시스템 설계 방향

3.1 자동 심사자 추천시스템 설계 방향

심사자 자동 추천시스템은 타당성, 포괄성, 전문성, 공정성을 고려해서 설계되어야 한다.

심사자 선정절차는, 심사요청분야에 3순위를 지정하여 심사요청분야와 연구계획서 내용을 고려해 구성된 심사자 풀 중에서 심사패널을 구성한다. 이 때 심사자추천시스템을 활용하여 심사후보자를 3배수를 추천하고, 상피여부를 고려한 후 최종 선정하게 된다. 현행 심사자 추천시스템에서 보완되어야 할 부분은 다음과 같다.

첫째, 패널로 구성된 과제를 포괄적으로 심사할 수 있는 심사자들이 추천되어야 한다. 둘째, 심사자 추천 인력의 풀이 많을 경우 우선순위자 별로 심사자를 추천하여 제시될 필요가 있다. 셋째, 해당 학문분야에 따라 심사자 추천 인력의 풀이 부족할 경우가 발생되는데 이러한 경우 관련 학문분야를 매핑 시켜 심사자 풀을 확대하는 방안이 필요하다. 넷째, 상피적용을 확대시켜 심사의 공정성을 확대시켜야 한다.

3.2 심사자 확장

전술한 바와 같이 심사자 확장이란, 심사자 추천 프로세스의 자동화 측면에서, 특정 과제 혹은 패널을 평가할 수 있는 심사자 풀을 넓히는 것을 의미한다. 이를 위해, 본 연구에서는 과제 지원자가 제시한 심사희망분야를 평가할 수 있는 심사자의 수가 적다고 판단될 때, 심사희망분야를 관련된 분야들로 확장함으로써 심사자 집합의 크기를 늘리는 방식을 취한다. 이 절에서는 먼저 단일 과제의 심사자 확장 방식에 대해 기술하고, 이를 통해 패널 내 그룹 심사자를 확장하는 방식을 기술한다.

과제 지원자의 심사희망분야와 심사후보자의 심사가능분야를 직접 매치하는 것은 과제 심사자를 선별하는 공정한 방식이긴 하나, 현실적으로 모든 분야에 고르게 충분한 수의 심사 후보자를 확보하기란 쉽지 않은 일이다. 따라서, 상기의 직접 매칭 방식은 일반적으로 과제 심사자를 충분히 제시하지 못하는 어려움을 안게 된다. 이러한 직접 매칭 방식의 단일 과제 심사자 선정은, 패널 내 그룹 심사자 선정에 적용될 경우 더 심각한 문제를 야기 시킨다. 하나의 패널은 통상 수십 개의 과제로 구성되고, 패널 내 그룹은 평균 5~10개 정도로 구성되므로, 패널 내 그룹의 심사자는 그룹 내에 속한 모든 과제를 심사할 수 있는 사람이어야 한다. 만약, 패널 그룹 심사자 선정을 위해, 그룹 내 모든 개별 과제의 심사희망분야와 직

접 매치되는 심사가능분야를 가져야 한다는 조건을 사용한다면, 대부분의 경우 얻어지는 심사자 집합의 크기는 0에 가까울 것이다.

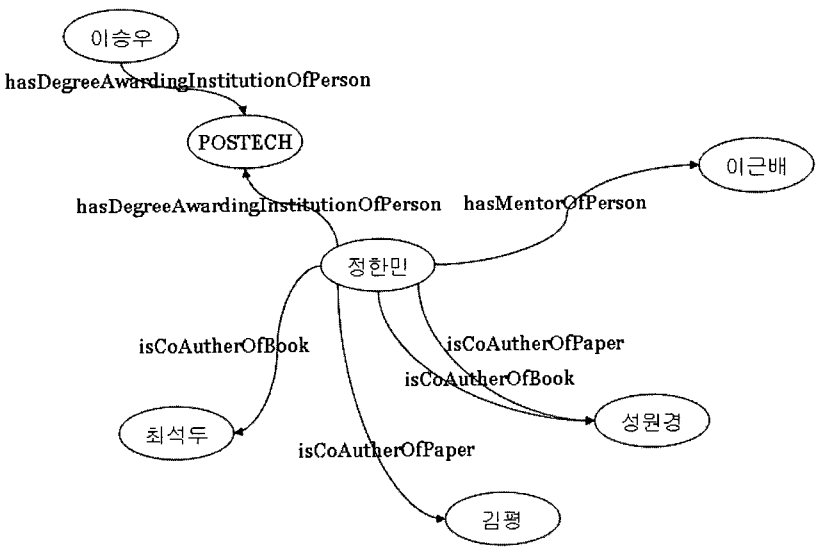
이러한 문제를 완화하기 위한 방편으로, 과제 지원자의 심사희망분야나 심사 후보자의 심사가능분야를 관련 분야(들)로 확장함으로써 과제 심사자의 수를 늘릴 수 있다. 관련 분야 확장을 위해 본 연구에서는, 노드 간 관련성이 확률값으로 표현된 KRF분류체계를 사용한다. 확률값 형식의 노드 간 연관성은 연관성 임계치를 통해 심사자 집합의 확장 범위를 정성적으로 통제하는 수단을 제공한다는 측면에서 특히 유용하다. 예를 들어, 심사자 확장을 위해 노드 간 연관성 임계치 0.8을 적용하면, 과제의 심사희망분야(노드)와 0.8이상의 연관성으로 연결된 KRF분류(노드)들을 심사할 수 있는 모든 심사후보자들이 해당 과제의 심사후보자로 뽑히게 된다.

전술한 단일 과제 심사자 확장은 패널 내 그룹의 심사자 확장에 직접적으로 기여하게 된다. 물론, 연관성 임계치 방식의 적용으로, 단일 과제 심사자 집합의 증가율에는 못 미치겠으나, 연관성임계치를 조정함으로써 하나의 패널 그룹 내의 대부분의 과제를 평가할 있는 심사자 집합의 크기를 통제할 수 있게 된다.

3.3 심사자 배제

심사자 배제를 위한 배제 방식은 분야 매칭, 학연 상피, 연구자 네트워크 적용으로 구성된다. 본 절에서는 단순한 DB 내 정보 검색을 통해 가능한 분야 매칭과 학연 상피를 제외하고 연구자 네트워크 적용에 초점을 맞추어 심사자 배제 시스템을 기술한다.

연구자 네트워크는 심사자를 노드로 하고, 노드 간에 멘터(Mentor) 관계, 공동 연구 관



〈그림 6〉 연구자 네트워크의 예

계, 논문 공저 관계, 저역서 공저 관계가 존재하는 네트워크로 정의한다. 〈그림 6〉은 학진 내에 심사자로 등록되어 있는 '정한민'의 연구자 네트워크를 보여준다.

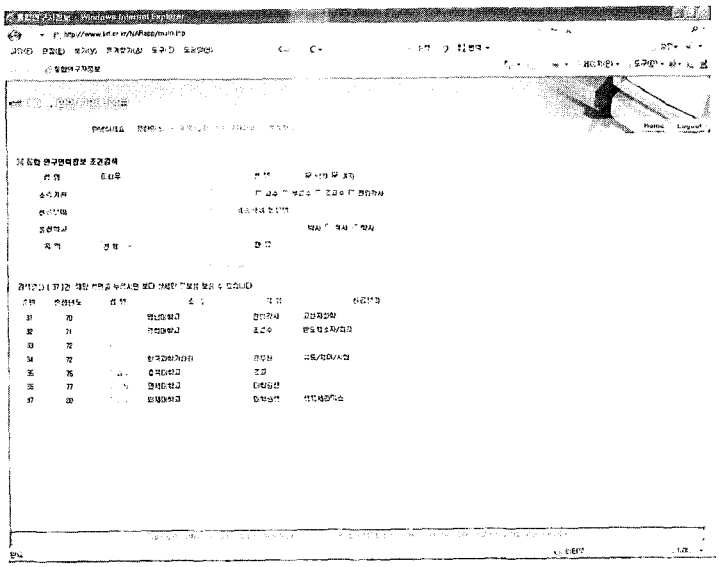
'정한민'의 경우 멘터(Mentor) 관계로서 '이근배'를 가지며, 논문 공저 관계로서 '성원경'과 '김평'을 가지며, 저역서 공저 관계로서 '최석두'를 가지며, 박사 학위 기관 상피 관계로서 '이승우'를 가진다. 특정 과제에 대해, 해당 과제 내의 지원자들 (연구책임자와 공동연구자들)과 연구자 네트워크 관계에 있는 심사자를 배재하는 방식으로 연구자 네트워크를 적용한다.

학진 내의 심사자 정보에는 멘터 관계, 공동연구 관계, 논문 공저 관계, 저역서 공저 관계 등을 직접 획득할 수 있도록 메타데이터가 구축되어 있지 않으므로, 학위 사항, 연구 실적 등으로부터 해당 정보를 추출하여 연구자 네트워크를 구축해야 한다. 데이터 베이스로부터 직접 획득할 수 있는 멘터 관계를 제외한 나머지 관계들의 구축 방안은 다음과 같다.

(1) 공동 연구 관계: 학진 내 과제 내 메타정보에서 연구책임자와 공동연구자들을 추출하고 이들의 식별자를 공동 연구 관계로 맺는다.

(2) 논문 공저 관계: 심사자 간에 같은 논문을 공유하는 경우 논문 공저 관계로 맺는다. 이는 심사자의 연구 실적에서 공저자들이 식별자 기반이 아닌 문자열로 기술되어 있어 동명이인인 문제 등으로 인해 정확한 연구자 네트워크 관계를 바로 맺을 수 없기 때문이다. '이승우'의 경우 심사자로 등록된 사람이 37명이 존재하는 것처럼 동명이인인 문제의 해결 없이 논문 공저 관계를 맺을 수 없다〈그림 7 참조〉.

이 문제를 해결하기 위해 두 명이상의 심사자가 같은 논문을 공유하는 지의 여부를 통해 논문 공저 관계를 파악한다. 즉, 심사자들이



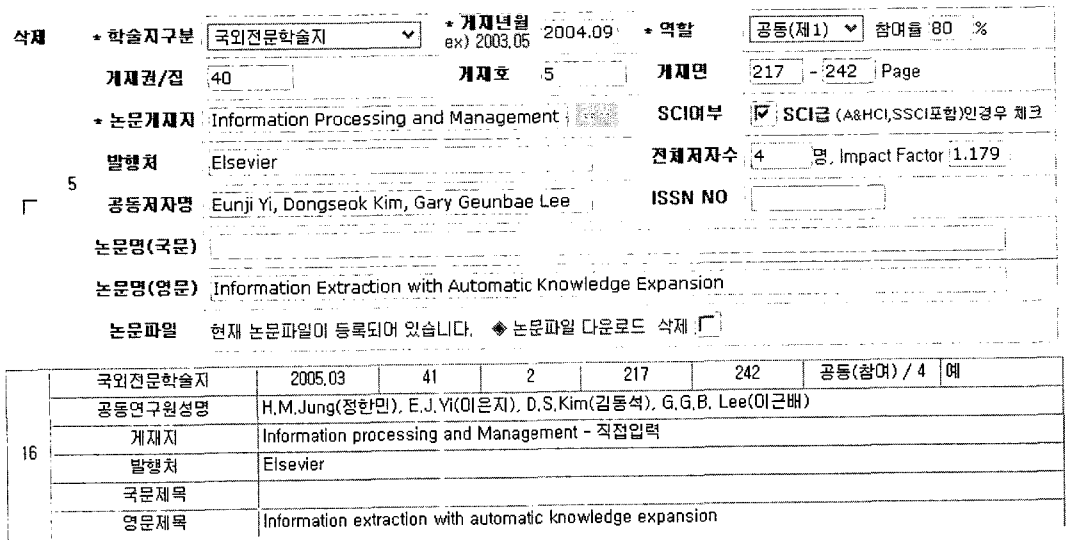
〈그림 7〉 동명이인 문제의 예 (학진 통합인력정보 내의 '이승우' 검색 결과)

같은 논문을 공유한다면 비록 각 심사자의 연구 실적 내 공저자가 문자열로 되어 있더라고 식별자 기반으로 관리되는 심사자 정보를 이용하여 다수의 심사자를 관계 지을 수 있는 것이다. 이 방식에서의 문제점은 동일한 논문에 대해서도 각 심사자가 등록한 논문 정보가 정확히 일치하지 않을 수 있다는 점이다. 〈그림 8〉과 같이 논문 정보는 제목, 출판연도, 공저자, 출처, 페이지 등으로 구성되는데 각 필드 대부분이 문자열로 직접 입력할 수 있으므로 철자 오류, 생략, 이형태, 입력 실수로 인해 동일한 논문인지 판단하기 어려운 경우가 많다. 4장에서 사용된 실험 데이터 중 논문 155,646편을 대상으로 논문 비교 실험을 한 결과를 살펴보면, 제목만으로 매칭하는 경우 1,000편(0.64%)이 동일 논문으로 인식되지만, 제목, 출판연도, 출처, 페이지를 모두 매칭하는 경우 256편(0.16%)만이 동일 논문으로 인식 된다.

본 연구에 기술된 실험에서는 언급된 모든 필드들이 매칭되는 경우로 한정하여 연구자 네트워크를 엄격히 적용하였다.

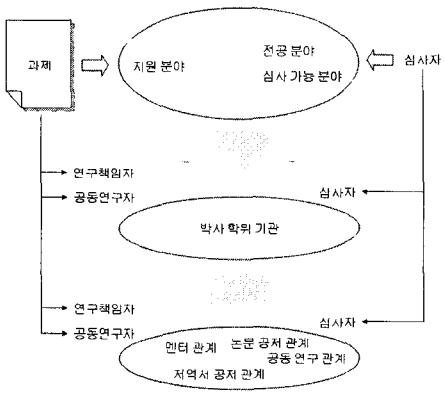
(3) 저역서 공저 관계: 심사자 간에 같은 저역서를 공유하는 경우 저역서 공저 관계로 맺는다. 매칭 대상 필드로는 제목, 출판연도, 출판사가 있다. 4장에서 사용된 실험 데이터 중 저역서 48,692권을 대상으로 논문 비교 실험을 한 결과를 살펴보면, 제목만으로 매칭하는 경우 4,452편(9.14%)이 동일 논문으로 인식되지만, 제목, 출판연도, 출판사를 모두 매칭하는 경우 3,067편(6.3%)만이 동일 논문으로 인식된다.

상기와 같이 연구자 네트워크를 구축한 후 심사자 배제를 적용하는 프로세스는 다음과 같다. 먼저, 과제의 지원 분야와 심사자가 등록한 전공 분야, 심사 가능 분야를 매칭하여 해당 분야로 제한된 심사자를 선정한다. 다음으



〈그림 8〉 논문 정보의 예 ('정한민'과 '이근배'가 입력한 동일한 논문 정보, 출처 내의 게재호와 공저자 표현 형식이 상이함)

로 박사 학위 기관을 공유하는 심사자와 과제 내 연구책임자 및 공동연구자들을 발견함으로써 학연을 배제할 수 있도록 한다. 마지막으로 과거 과제에서 공동 연구를 수행한 적이 있는 지원자 (연구책임자 및 공동연구자들)와 심사자, 논문이나 저역서를 공저한 적이 있는 지원자와 심사자. 멘터 관계에 있는 지원자와 심사자를 연구자 네트워크에서 찾아 배제한다〈그림 9 참조〉.



〈그림 9〉 심사자 배제 프로세스

그러나 이와 같은 배제 프로세스는 항상 일관된 것이 아니라 상황에 따라서 선택적 적용을 할 수 있도록 설계하였다.

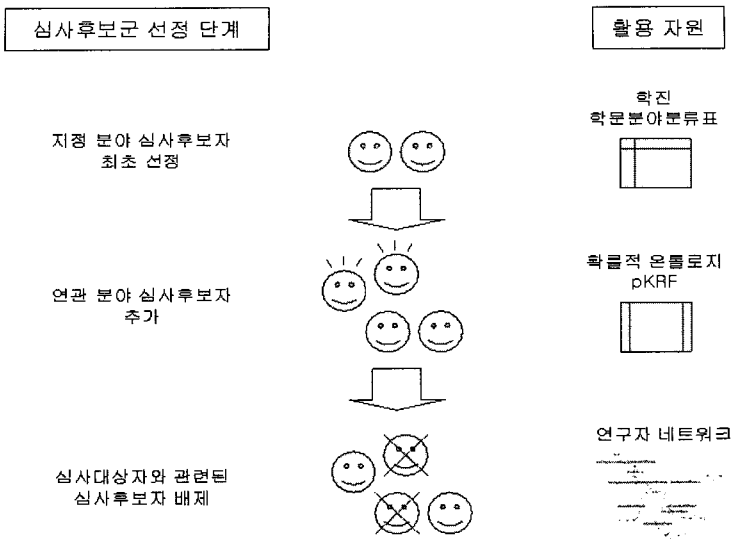
4. 구현 및 검증

4.1 심사자 추천 시스템 구현

이 절에서는 제안한 심사자 추천 방식의 자동화를 평가하기 위해 구현한 심사자 추천 시스템에 대해 간략히 기술한다. 평가용 심사자 추천 시스템은 스크립트 언어인 펄(PERL)로 구현되었으며, 〈그림 10〉과 같이 확률적 온톨로지와 연구자 네트워크를 활용한다.

구현된 시스템은 크게 단일 과제 모드와 패널 모드로 나뉘어 동작한다. 단일 과제 모드는 하나의 과제 정보를 입력으로 받아 해당 과제의 심사자 집합을 출력으로 낸다. 입력되는 과제 정보는 '과제번호', '과제책임자ID', '과제공동연구자(들)ID', '과제책임자기관코드', '과제공동연구자기관(들)코드', '제1~3심사희망분야(들)', '연관성 임계치값'로 구성된다. 단일 과제 모드의 출력으로는 '심사자ID', '제1~3심사희망분야 연관성(들)'로 구성된다.

패널 모드는, 실질적으로 단일 과제 모드를 포함하는데, 패널에 포함된 모든 개별 과제에 대해 단일 모드를 실행하고, 그 결과로 얻어지는 과제별 심사자 집합을 통합하는 과정으로 동작한다. 패널 모드에서 추가로 요구되는 입력으로는, 패널 내 그룹 구성을 위해 필요한 패널 내 과제번호별 그룹 번호에 대한 매핑 테이블이 있다. 패널 모드의 출력은 '심사자ID', '심사자전문도', '심사가능과제수', '심사가능그룹수', '심사가능KRF분야수' 등을 포함하여 구성된다. '심사자 전문도'는 심사자의 논문, 저역서, 특허, 작품 등의 연구성과를 바탕으로 0이상의 실수 값으로 미리 계산된 값으로, 타 조건이 동일할 경우 심사자를 전문도 순으로 정렬하여 검토하기 위해 요구되는 항목이다. 심사자 전문도는 대학교육대학종합평가(한국대학교육협의회 2006) 기준을 준용하여 적용하였으며, 최근 5년간의 연구실적에 한하였다. '심사가능 과제수'는 해당 심사자가 패



〈그림 10〉 확률적 온톨로지와 연구자 네트워크를 활용한 심사자 추천 단계

널 내에서 심사할 수 있는 과제의 총 수를 의미한다. '심사가능 그룹수'는 심사가능과제수의 크고 적음에 상관치 않고 해당 그룹 내의 하나의 과제라도 심사 가능한 경우 해당 그룹을 심사할 수 있는 것으로 고려하여 계산한 수치이다. '심사가능KRF분야수'는 패널 내 모든 과제들의 전체 심사희망분야들 중 해당 심사자가 심사 가능한 분야의 수를 의미한다.

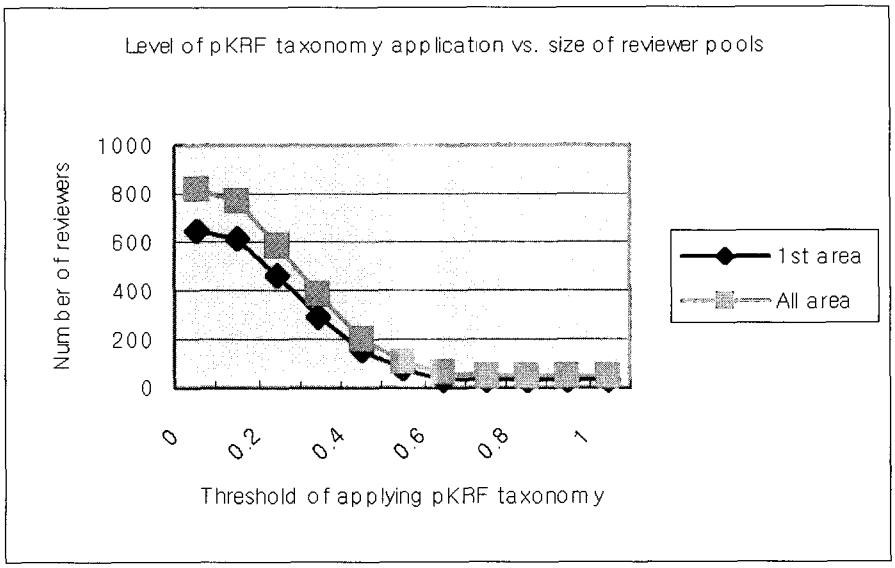
전술한 '심사자ID', '심사자전문도', '심사가능과제수', '심사가능그룹수', '심사가능KRF분야수' 들은 추천된 심사자를 정렬하기 위한 자질로 사용된다.

4.2 실험 데이터

실험을 위한 데이터의 구성은 다음과 같다. 학진에 등록된 인문학 분야를 대상으로 실험 데이터를 구성하였다. 전체 심사자 중 그룹을 통해 심사되는 최소 단위인 심사에 참여 가능한 인문학 분야의 심사자는 5,508명이며, 이들이 성과로 등록한 논문은 155,646편 (1인당 평균 28.2편)이며, 저역서는 48,692권 (1인당 평균 8.8권)이다.

4.3 심사자 확장

심사자 확장을 위해 학진 학문분야분류표의 노드쌍에 관계 확률이 부착된 확률적 온톨로지 *p*KRF분류체계를 사용하였다. 학진 학문분야 분류표는 4,232개의 노드로 구성되므로, 이론적으로 8,952,796(=4,232×4231/2)개의 확률 부착 노드쌍이 존재해야 하나, 확률값 획득을 위해 사용한 신청 과제정보의 부족 및 실제 관련이 없는 노드쌍의 존재 등으로 *p*KRF 분류체계는 총 4,622,974개의 확률 부착 노드쌍으로 구성되어 있다.



〈그림 11〉 *p*KRF분류체계를 통한 심사자 확장 효과

〈그림 11〉은 *p*KRF분류체계 내 노드 간 연관성 임계치값의 변화에 따른 한 과제의 평균 심사자 수의 추이를 보여 주고 있다. 임계치 1은 과제 지원자 심사희망분야와 심사 후보자의 심사가능분야를 직접 매칭한 경우에 해당한다. 임계치 0은 0을 포함하지 않은 0초과를 의미한다. 평균 심사자 수는, 실험에 사용한 전체 3,679개 과제에 대해 얻어지는 3,679개의 심사자 집합의 크기에 대한 평균을 의미한다. 1st area와 All area는 각각 과제 지원자의 심사희망분야 중 제1심사 희망분야만 사용한 경우와 전체 3개의 심사희망분야를 모두 사용한 경우를 가리킨다.

〈그림 11〉을 통해, 연관성 임계치 값이 감소할수록 평균 심사자의 수가 증가하며, 제1심사 희망분야만 확장에 사용한 경우보다 전체 심사희망분야(들)을 모두 사용했을 때의 평균 심사자 수의 증가폭이 더 커짐을 알 수 있다. 또한, 연관성 임계치 0.7 이상의 경우 평균 심사자 수의 변화가 거의 없으며, 0.6 이하부터 심사자 수의 변화를 나타내고 있다. 이는 pKRF분류체계에 부착된 노드 간 확률값에서 0.7이상 확률값의 분포가 거의 출현하지 않기 때문인데, 이는 노드 간 확률값의 획득을 위해 사용한 코퍼스에 의존적인 부분이므로 이 논문의 논의에서 제외한다.

〈그림 11〉의 추이 곡선에서, 임계치 0.4이하의 경우 심사자의 수가 급격히 증가하는데, 이는 *p*KRF분류체계에서 특정값 이상 연관성을 갖는 노드들의 평균 수가 연관성 0.4이하로 갈수록 급격히 늘어난 데서 기인한 결과이다. 이처럼 한 노드의 평균 관련 노드 수가 급격히 증가한 지점은, 그 확률적 연관성의 실질적 적합성을 의심하게 되는 지점이므로 실제 적용에서 신중하게 고려되어야 할 것이다.

이 연구에서는 심사자 확장을 위한 연관성

임계치의 적정 범위에 대해서는 깊이 다루지 않았으나, 경험적으로 심사자 수의 추이 곡선이 완만히 증가하는 범위에 해당하는 임계치를 사용하는 것이 바람직해 보인다.

4.4 심사자 배제

2004년부터 2006년까지의 인문학 분야 과제 3,679건을 살펴보면, 592건 (16.1%)이 과제 내 연구자들이 2명 이상인 과제이며, 나머지는 단독 연구 과제이다. 연구자 네트워크 적용 범위를 살펴보면 다음과 같다〈표 4 참조〉.

〈표 4〉 연구자네트워크 유형 비율

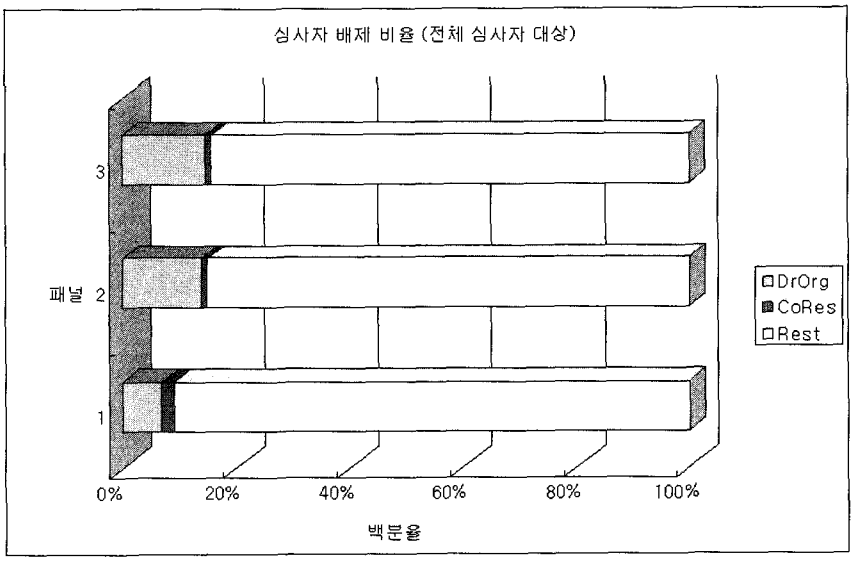
연구자 네트워크 유형	영향받는 과제 수 (3,679건 대비)	비율
멘터 관계	0	0%
공동 연구 관계	1,221	33%
논문 공저 관계	1,292	35%
저역서 공저 관계	1,943	53%

전체 과제의 50% 이상이 연구자 네트워크의 영향을 받아 기존의 분야 매칭이나 학위 기관 상피를 보완할 수 있는 배제 방식이 될 수 있음을 알 수 있다. 멘터 관계가 발견되지 않은 것은, 실험 대상 크기가 전체 심사자 수인 170,000여 명에 비해 상당히 작기 때문으로 추정되며, 전제 심사자를 대상으로 하여 추가적인 실험을 할 예정이다.

심사자 배제의 효용성을 살펴보기 위해 인문학 분야 중 3개 세부 분야를 대상으로 패널을 구성하고, 각 패널을 대상으로 실험을 진행하

〈표 5〉 3개 패널별 과제 구성

패널1 (8개 분야, 21개 과제)	
분야	과제 수
A020109 여성사(역사학)	1
A020110 비교사(역사학)	1
A020301 중국고대사	5
A020302 중국중세사1(위진수당)	2
A020304 중국근세사(명청)	3
A020305 중국근대사	3
A020306 중국현대사	5
A020307 대만사	1
패널2 (10개 분야, 19개 과제)	
분야	과제 수
A030101 형이상학	2
A030104 정치/사회철학	3
A030107 역사철학	1
A030109 문화/기술철학	1
A030405 서양근대독일철학	1
A030406 서양근대프랑스철학	4
A030408 해석학	3
A030501 미학일반	1
A030502 예술철학	2
H050300 성(sexuality)	1
패널3 (8개 분야, 19개 과제)	
분야	과제 수
A040700 비교종교학	1
A120202 중국시	2
A120203 중국희곡	1
A120206 중국소설	3
A120208 중국현대문학	9
A120210 비교문학(중문학)	1
A120211 경학(중문학)	1
A120213 중국서지학	1



〈그림 12〉 전체 심사자 대상 심사자 배제 적용 결과

였다. 다음은 각 패널에 속한 과제들의 1순위 지원 분야를 보여준다〈표 5 참조〉.

심사자 배제는 크게 심사자의 전공 분야 및 심사 가능 분야와 과제 지원 분야를 비교하는 방법, 심사자의 박사 학위 기관과 과제 내 연구자들의 박사 학위 기관을 비교하는 방법, 3.2절에서 기술한 연구자 네트워크를 적용하는 방법으로 나눌 수 있다. 첫 번째는 심사자들이 학진 분야분류체계 내에서 선택한 전공 분야와 심사 가능 분야 (최대 5개)를 과제 지원 분야와 비교하고 매칭이 되지 않는 경우에 배제하는 방식이다 〈그림 12의 'Rest'〉. 두 번째는 심사자와 지원자 간의 박사 학위 기관 기준으로 학연을 배제하는 방식이다 〈그림 12의 'DrOrg'〉. 세 번째는 심사자와 과제 내 연구자들이 멘터(Mentor) 관계, 공동 연구 관계, 논문 공저 관계, 저역서 공저 관계 등 연구자 네트워크로 연결되는 경우 배제하는 방식이다〈그림 12의 'CoRes'〉.

전제 심사자 5,508명에 대한 배제 결과를 살펴보면 그림1과 같은데, 세 개 패널에 대해 평균 35.5%의 심사자가 배제되었다. 심사자 배제 규칙 적용은 분야 매칭, 박사 학위 기관 상피, 멘터 관계, 공동 연구 관계, 논문 공저 관계, 저역서 공저 관계 순으로 진행하였다. 이전 배제 규칙에 적용되는 심사자는 이후 배제 규칙을 적용하지 않고 배제하였다.

각 유형 별로 살펴보면, 분야 매칭 방식을 적용하는 경우 30.5%, 박사 학위 기관 상피 방식을 적용하는 경우 추가로 4.5%, 연구자 네트워크 방식을 적용하는 경우 추가로 0.5%가 배제 되었다. 연구자 네트워크 중 공동 연구 관계만 3개 패널에서 발견되었다. 앞선 실험처럼 전체를 대상으로 하는 경우 다른 관계들에서도 배제가 되리라 예측한다. 연구자 네트워크가 심사자 배제에 미치는 영향이 작지만, 다

른 방식들에서 배제하지 못한 심사자를 추가로 배제하는 것이며, 데이터베이스로부터 SQL 질의를 통해 바로 얻어낼 수 없는 정보라는데 그 효용 가치가 있다.

상기 실험을 통해 여러 심사자 배제 규칙들은 상호 보완적이며 복합적으로 적용할 필요가 있다는 것을 알 수 있다. 추후 실험에서는 연구자 네트워크를 직접 관계 뿐만 아니라 간접 관계까지 적용함으로써 적용 범위 확장에 그에 따른 위험성을 살펴볼 예정이다.

4.5 검증

앞서 실험한 3개 패널의 실제 심사자 풀과 실험 적용한 데이터의 결과에 대한 검토평가를 각 패널별로 전문가 1인에게 의뢰하였다. 패널별 전문가는 전년도 학진의 해당 분야의 프로그램 관리자로 심사자 추천시스템과 해당 분야의 전문가 집단을 판단할 수 있는 전문가이다. 검증 데이터는 3개의 패널별로 지원과제의 심사요청분야를 대상으로 *p*KRF분류 확장도, 전문도, 심사가능 과제 수, 심사가능 소그룹 수, 심사가능 KRF 분야 수 및 배제규칙을 전체 심사자 풀에 적용한 추천심사자들의 리스트 이고 각 항목별로 순위를 점검할 수 있게 하였다. 여기서 *p*KRF분류 확장도는 확률적 연관성 임계치 값을 1에서 0까지를 단위별로 적용하여 제시하였다.

추천 심사자들의 리스트는 실명과 소속기관 그리고 해당 심사가능분야, 전문성 여부가 순위별로 적절한지에 대한 내용검증을 우선으로 하였다. 검증방법은 설문과 인터뷰 형식으로 이루어졌다. 검증 항목은 확률적 온톨로지의 적용성과 전문도 적용 가능성 및 배제규칙 적용에 관한 검증을 실시하였다.

확률적 온톨로지인 *p*KRF 분류체계의 적용 가능성에 관해서는 확률적 연관성 임계치 값을 0.4에서 0.6정도로 설정하는 것이 적절하다는 의견이었고 현재의 심사자 풀을 확장하여 심사 가능 분야를 포괄적으로 측정하여 제시할 수 있다고 하였다. 이는 앞서 이론적으로 언급되었듯이 0.4 이하의 급격한 확장은 연관성이 미비한 연관관계임을 알 수 있다. 실제 적용에 있어서는 패널 그룹의 특성이나 학문분야에 따라 임계치는 다르게 적용 될 수 있도록 설계되어야 한다.

전문도 적용에 관해서는 적절한 가중치 산출법이라는 평가를 받았다. 이는 포괄적인 부분뿐 아니라 전문적인 적합한 심사자 추천에 매우 큰 도움이 될 수 있을 것이라는 의견이었다. 단, 연구실적 산정을 5년에서 10년까지로 확장시키는 방안과 분야별 합리적 설문조사를 통해서 해당 분야 전문가들을 조사하여 계량적인 연구 성과물에 의한 측정 뿐 아니라 정성적인 평가를 겸하여 학계의 중진도 포함될 필요가 있음이 지적되었다.

배제규칙 적용에 있어서는 평가자들이 멘터(지도교수)관계, 공저자 적용은 모두 동의하였으나, 공저(단행본), 공역, 공동연구 등의 적용은 사업 특성이나 심사자 풀 등을 고려하여 상황에 따라 선별적으로 적용해야 한다고 지적하였다. 이는 학문에 따라 공저인 경우가 학문적 관계가 밀접하다고 보는 경우가 있고 공동연구인 경우가 관련성이 높다고 판단하는 측면이 있기 때문이다.

본 자동추천 시스템 개선 방안에 관한 의견

을 묻는 설문항목은 5점 만점으로 점수를 물은 결과 4.85점(백점 만점일 경우 97점)으로 나타나서 매우 바람직한 방안으로 평가되었다.

5. 결 론

심사자 자동추천시스템은 심사자들을 보다 객관적이고 공정하며, 전문성 있게 추천할 수 있도록 설계되고 적용되어야 한다.

본 연구에서 실험하여 분석된 의미 있는 연구결과는 다음과 같다.

첫째, 확률적 온톨로지는 학문분야분류표의 열거형 주제분류표를 다면적으로 표현하고 학문 간의 경계를 넘나드는 학제성을 반영하였다는데 큰 의의가 있다. 또한 실제 패널 심사에서 요구하는 과제의 대상범위 확장과 관련분야 확장에 유의하게 활용될 것으로 보이며, 부족한 심사자 풀은 확장시켜 적용할 수 있을 것이다. 나아가서는 학문 분류법 상의 엄밀성을 확립할 수 있을 뿐 아니라 학문 간의 상호 연계성과 학제적 특성을 분석해 낸 결과로도 의미가 있다. 앞으로 학문분야분류표를 보완한 확률적 온톨로지인 *p*KRF 분류체계는 학제 간 연구의 기초 자료로도 활용될 수 있으리라 본다.

둘째, 심사자들의 전문도 적용 가능성이다. 본 연구에서 적용한 사례는 연구성과물의 유형별로 가중치를 적용하고 최신 연도로 제한하여 각 분야별로 전문도를 산정하고 이를 심사자 추천에 적용하였다. 이론적으로나 실험 결과적으로 이의 활용성은 매우 높아 보이지만, 실제로 성과물로서 측정할 수 없는 정성적인 평가항목도 포함되어야 하고 학문 영역에 따라 최근의 연구성과물 위주로 전문도를 산정할 것인지 혹은 산정기간을 늘리는 방법도 고려해야 할 것이다. 전문도는 심사자 풀이 많을 경우에 적절히 적용하여 균형을 이루는데 활용도가 높다고 볼 수 있다.

셋째, 연구자 네트워크 등의 배제규칙 적용은 현재 문자열 기반으로 세부 분야 필드를 계속해서 매칭해 나가는 방법으로 구현 적용하였으나 이러한 경우에 철자오류, 생략, 이형태, 입력 실수로 인한 오류가 발생된다. 이러한 부분을 보완하기 위해서는 저자, 소속기관, 논문, 저역서 등의 서지데이터에 관한 전거가 마련되어 일정한 식별자로 구분되어야 정확한 정보를 추출해 낼 수 있다. 심사자 배제 규칙을 멘터 관계, 저역서 공저관계, 공동 연구 관계, 논문 공저관계, 학위사항 등으로 적용할 수 있다. 그러나 이와 같은 배제 규칙은 기본적인 배제 규칙을 일괄적으로 적용하기는 어렵고 과제 지원 사업의 형태에 따라서 선택적으로 융통성 있게 적용하는 것이 바람직하다.

마지막으로 학진의 연구인력 풀 시스템에 대한 정확한 데이터의 검증 및 지속적인 업데이트가 기반이 되어야 하며, 심사자추천시스템 개선의 기능들은 과제 성격과 상황에 따라 적용될 수 있도록 설계되어야 한다. 향후 학진의 인력정보시스템은 학술지 평가시스템과 인용분석 시스템과의 연동하여 연구성과물을 정확히 검증하고 전문가 산정방법을 다각적으로 적용할 수 있어야 한다.

이 연구를 기반으로 제안할 수 있는 후속 연구로는 과제 지원 신청자들의 연구계획서에서 내용을 추출하여 유사 패널을 자동 클러스터링하는 방법과 과제 지원 신청자의 연구계획서와

심사자 풀의 연구성과물을 직접 비교하는 연구를 수행하여 내용기반 분석시스템을 구축하는 것이 과제이다.

또한 확률적 온톨로지와 연구자 네트워크는 국내 학술연구의 학제적인 동향과 협력연구 동향을 파악하기 위한 수단이 될 수 있으므로 이런 방향의 응용 연구도 가능성이 클 것으로 기대된다.

참 고 문 헌

김태수. 2000. 『분류의 이해』. 서울: 문헌정보처리연구회.

이재윤. 2006. 『온톨로지 기반 과제관리 및 분석체제 구축 방안 연구』. 서울: 한국학술진흥재단

한국대학교육협의회. 2006. 『대학종합평가 인정제 시행을 위한 대학종합평가 편람』.

Barabasi, A. L., H. Jeong, E. Ravasz, Z. Neda, A. Schuberts, and T. Vicsek. 2002. "Evolution of the social network of scientific collaborations." *Phys.* A. 311: 590-614.

Costa, P., and K. Laskey. 2006. "PR-OWL: A framework for probabilistic ontologies." *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2006).* [online] 〈http://ite.gmu.edu/~klaskey/papers/FOIS2006_CostaLaskey.pdf〉.

Costa, P., K. Laskey, and K. Laskey. 2006. "Probabilistic ontologies for efficient resource sharing in semantic web services." *Proceedings of the Second ISWC Workshop on Uncertainty Reasoning in the Semantic Web.* [online] 〈http://www.iet.com/iswc/2006/ursw/files/papers/URSW06_T5_CostaLaskeyLaskey.pdf〉.

Chung, Y. M., and J. Y. Lee. 2001. "A corpus-based approach to comparative evaluation of statistical term association measures." *Journal of the American Society for Information Science and Technology*, 52(4): 283-296.

Newman, M. E .J. 2001a. "Scientific collaboration networks: I. Network construction and fundamental results." *Phys. Rev. E*, 64. art. no. 016131.

Newman, M. E. J. 2001b. Scientific

collaboration networks: II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E*, 64. art. no. 016132.

Newman, M. E. J. 2003. "The structure and function of complex networks." *SIAM Review*, 45(2): 167-256.

Redner, S. 1998. "How popular is your paper? An empirical study of the citation distribution." *European Physics Journal B*, 4: 131-134.

Scott, J. 2000. *Social Network Analysis: A Handbook*. London: Sage.

Seglen, P. O. 1992. "The skewness of science." *Journal of the American Society for Information Science*, 43: 628-638.

White, H. D., B. Wellman and N. Nazer. 2004. "Does citation reflect social structure? Longitudinal evidence from the Globenet interdisciplinary research group." *Journal of the American Society for Information Science and Technology*, 55(2): 111-126.

Wasserman, S. and K. Faust. 1994. *Social Network Analysis*. Cambridge: Cambridge University Press.