https://doi.org/10.5392/JKCA.2021.21.12.001

학술 문헌 기반 효율적인 전문가 판별 기법

An Efficient Expert Discrimination Scheme Based on Academic Documents

최도진[*], 오영호[*], 편도웅[*], 방민주[**], 전종우[*], 이현병[*], 박득배[*], 임종태[*], 복경수[***], 유효근[****], 유재수[*]
충북대학교[*], 한양대학교[**], 원광대학교[***], 현대엔지비[****]

**Do-Jin Choi(mycdj91@cbnu.ac.kr)[*], Young-Ho Oh(ohy5268@cbnu.ac.kr)[*],
Do-Woong Pyun(pyun19@naver.com)[*], Min-Ju Bang(minj2357@naver.com)[**],
Jong-Woo Jeon(junjongwoo30@naver.com)[*], Hyeon-Byeong Lee(lhb@cbnu.ac.kr)[*],
Deukbae Park(pdbdbp@naver.com)[*], Jong-Tae Lim(jtlim@cbnu.ac.kr)[*],
Kyoung-Soo Bok(ksbok@wku.ac.kr)[***], Hyo-Keun Yoo(hgyoo@hyundai-ngv.com)[****],
Jae-Soo Yoo(yjs@cbnu.ac.kr)[*]**

요약

특정 연구 분야에 대한 전문성을 가진 연구자를 찾기 위해서는 객관적인 전문가 판별 방법이 필요하다. 기존에는 전문가 판별을 위해 인용 그래프 기반의 판별 기법과 수식 기반의 판별 기법이 존재한다. 본 논문에서는 기존 수식 기반 판별에서 고려하지 못하였던 다양한 특성들을 반영한 효율적인 전문가 판별 기법을 제안한다. 연구자의 전문성을 판별하기 위해 품질, 생산성, 기여도, 최신성, 정확성, 지속성을 고려한 전문성 지수를 제안한다. 또한, 학술 검색 사이트만의 특성을 반영하기 위해 소셜 인용 수를 추가로 고려한다. 다양한 학술 사이트 기반의 논문 수집 및 성능 평가 결과를 제시하고 제안하는 기법의 타당성과 실현성을 입증한다.

■ 중심어 : | 전문가 판별 | 빅데이터 | 랭킹 | 검색 | 데이터베이스 |

Abstract

An objective expert discrimination scheme is needed for finding researchers who have insight and knowledge about a particular field of research. There are two types of expert discrimination schemes such as a citation graph based method and a formula based method. In this paper, we propose an efficient expert discrimination scheme considering various characteristics that have not been considered in the existing formula based methods. In order to discriminate the expertise of researchers, we present six expertise indices such as quality, productivity, contributiveness, recentness, accuracy, and durability. We also consider the number of social citations to apply the characteristics of academic search sites. Finally, we conduct various experiments to prove the validity and feasibility of the proposed scheme.

■ keyword : | Expert Discrimination | Big Data | Ranking | Search | Database |

* 이 논문은 2020년도 현대엔지비㈜의 지원(과제관리번호 : 20T3757)과 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산학연협회(AURI)의 지원(2021년 기업연계형연구개발인력양성사업, 과제번호 : S3047889)과 농촌진흥청 연구사업 (세부과제번호: PJ01624701)의 지원에 의해 이루어진 것임

접수일자 : 2021년 08월 31일
수정일자 : 2021년 09월 23일
심사완료일 : 2021년 09월 23일
교신저자 : 유재수, e-mail : yjs@cbnu.ac.kr

I. 서 론

전문가란 특정 분야에 능통한 사람을 일컫는 말이다. 국내외에서는 다양한 니즈에 따라 전문가를 필요로 하는 경우가 많다. R&D 과제에 대한 심사위원을 위촉하거나 특정 분야의 전문지식에 관한 자문을 구하고자 할 때, 논문에 대한 심사를 요청할 때와 같이 전문가를 필요로 하는 경우가 종종 발생한다. 이러한 전문가에 대한 검색 요구가 지속해서 발생하며, 효과적인 전문가 판별 및 분석 방법이 필요하다.

많은 연구자는 논문이나 특허, 연구 보고서 등 해당 연구 분야의 결과물을 남긴다. 따라서 연구자들의 연구 결과물들을 분석하면 연구자의 분야에 대한 전문성을 판단 할 수 있다. 이러한 정보를 바탕으로 연구자의 전문성을 판단하여 전문가를 추천하는 연구들이 활발하게 연구되고 있다. 또한, 기업이나 연구자들은 다양한 학술 검색사이트를 이용하여 자신의 연구 분야에 해당하는 연구 결과물을 검색함으로써, 전문가를 자체적으로 판단하는 경우도 있다. 그러나 이러한 방식은 해당 분야의 전문가를 판별하는데 다음과 같은 문제점이 존재한다.

첫째, 학술 검색사이트를 이용하는 경우 연구 주제와 관련된 여러 논문을 제공하지만, 해당 분야의 논문의 저자 정보만으로는 전문가를 판별하기 힘들다. 둘째, 전문가를 찾더라도 최근까지 지속해서 연구를 해온 전문가인지, 혹은 최근에는 연구를 이어가지 않은 연구자인지 판별하기 어렵다. 셋째, 논문의 품질을 인용 수와 IF(Impact Factor)와 같은 수치만으로는 분별력이 떨어진다. 넷째, 상대적으로 최근에 논문을 많이 작성하여 인용 수가 적은 신진 전문가를 찾는 어려움이 있다. 마지막으로 학술 검색 사이트의 특성을 반영한 전문가 판별 방법이 존재하지 않는다.

기존 전문가 분석 연구들은 크게 두 가지로 나눌 수 있다. 첫 번째는 모든 논문에 대한 인용 관계를 수집한 후, 인용 관계를 활용하여 그래프로 데이터로 표현하고, 그래프 분석 알고리즘을 수행하여 전문가를 찾는 판별 기법들이 존재한다[1-6]. 두 번째는 전문성 지수를 정의하여 연구자에 대한 전문성 지수를 계산하여 전문가를 찾는 방법이다[7-14]. 그래프 기반 판별 방법은 모든 인용 관계를 수집할 수만 있다면, 매우 효과적일 수 있으나, 이러한 접근은 현실적으로 구현하기 어려운 방법이다. 전문성 지수 접근 방법은 전문성 지수의 수식을 확립하는 것이 어렵지만, 구현이 매우 용이하다. 본 논문에서는 전문성 지수 접근 방법을 통해 분야별 전문가를 판별하는 연구를 수행할 예정이다. 또한, 기존에 다루지 못하였던 추가적인 지표를 제시한다.

본 논문에서는 온라인 학술 사이트의 논문에서 얻을 수 있는 다양한 정보를 활용한 전문가 판별 방법을 제안한다. 전문가 판별은 학술 검색 사이트에서 사용자가 검색하고자 하는 분야와 관련된 논문을 먼저 수집하고, 수집된 논문을 기반으로 제안하는 전문성 지수를 계산한다. 계산된 전문성 지수를 기반으로 최종적인 전문가 점수를 계산하고, 전문가 점수가 높은 순으로 전문가 명단을 제시한다. 본 논문에서는 전문가를 판별하기 위해 6가지의 전문성 지수를 제안하며, 온라인 학술 사이트의 특성을 반영하기 위한 소셜 인용 수를 제안한다. 다양한 학술 사이트에 대한 논문 수집과 분석을 통하여 통합 전문가를 분석하는 새로운 분석 결과를 제시한다. 성능 평가를 통해 제안하는 기법의 타당성과 실현성을 입증한다.

본 논문의 구성은 다음과 같다. 2절에 기존 기법들을 설명한다. 3절에서는 제안하는 기법의 특징과 지수 계산 방법을 설명하고, 4절에서는 제안하는 기법의 우수성을 입증하기 위해 성능 평가를 수행한다. 마지막으로 5절에서 본 논문의 결론과 향후 연구를 제시한다.

II. 관련 연구

기존 전문가 판별 연구들은 크게 두 가지로 나눌 수 있다. 첫 번째는 모든 논문에 대한 인용 관계를 수집한 후, 인용 관계를 활용하여 그래프로 데이터로 표현하고, 그래프 분석 알고리즘을 수행하여 전문가를 찾는 분석 기법들이 존재한다. Dunaiski, Zhao[2][6]는 인용 정보를 활용하여 PageRank를 변형한 CiteRank 기법을 수행하여 전문가를 판별한다. 인용이 높은 논문을 쓴 저자는 전문가일 확률이 높다는 가정이 동반된다. Liao[4]는 온라인 커뮤니티의 품질, 평판, 신용 정보를

활용하여 HITS(Hyperlink-Induced Topic Search) 알고리즘을 수행한다. 여기서는 사용자뿐만 아니라 논문에 대한 랭킹 정보도 제공한다. Lin, Jardine[3][5]는 저자 별 Topic 혹은 LDA(Latent Dirichlet Allocation) 기법을 통해 논문별 토픽을 추출하여 토픽 기반의 PageRank를 수행하여 분야별 전문가를 탐색할 수 있다. 앞서 설명하였듯이, 현재 시점에서는 논문의 모든 인용 정보를 수집하는 것은 불가능하고, 중요하지 않은 정보를 모두 수집해야 하기 때문에 인용 그래프 기반의 분석을 수행하는 것은 실현성이 떨어지고 매우 비효율적이다.

두 번째는 객관적인 전문성 지수를 정의하여 모든 연구자에 대하여 전문가 점수를 계산하여 점수가 높은 연구자를 전문가로 판별하는 방법이다. 전문가 점수 접근 방법은 전문가 점수의 객관성을 제공하기는 어렵지만, 구현이 용이한 장점이 있으며, 모든 정보를 수집하지 않더라도 준수한 성능의 전문가 판별을 수행할 수 있다는 장점이 존재한다. Li[7]은 저널에서 논문 심사 위원에게 논문을 할당하기 위해서 심사위원의 전문성과 연관성을 수치화하여 계산하는 방법을 제안한다. 전문성은 저자가 쓴 논문의 인용 수와 저널의 IF를 활용하여 품질 점수를 계산하고, 발행 연도를 기준으로 최신성 지수를 계산한다. 두 수치와 생산성 지수(논문의 수)를 결합하여 최종적인 전문가 점수를 계산한다. 새로운 논문이 저널에 제출될 때, 논문의 연관성과 전문성이 높은 심사 위원에게 논문을 할당한다.

Bilir, Lin[15][16]은 전문 기관 분석을 수행하는 연구를 제안한다. 전문 기관이란 전문가 집단이 많이 포함된 대학, 연구소 등과 같은 기관에 대한 랭킹을 수식으로 계산하여 분석하는 방법을 의미한다. 인용 수, 발행 연도, 저자 기여도를 기반으로 기관의 영향력 지수를 계산한다.

Wang[12]은 논문, 특허, 연구과제 정보를 고려한 연구자 추천 시스템을 제안한다. 논문 이외에도 중요한 연구 결과들을 활용하지만, 단순한 카운팅 방법을 수행하기 때문에 전문성이 있는 전문가를 찾기가 어렵기 때문에 좀 더 정밀한 계산 방법이 요구된다.

Choi[14]는 논문의 품질, 저자의 기여도, 키워드의 희소성, 최신성을 제안하여 전문가 판별을 수행한다. 희소성을 판별하기 위해서 사전에 정의된 키워드와 관련된 정보를 모두 수집해야 하므로 그래프 기반 분석의 단점을 일부 가지고 있다. 또한, 신진 연구자에 대한 가중치가 존재하지 않아 인용 수가 상대적으로 낮은 전문가에 대한 추천을 수행하지 못하는 경우가 빈번히 발생한다.

h-index, g2 index, I-10 index 등 논문 출간 수와 인용 횟수를 기반으로 저자의 전문성 지수를 계산하는 다양한 수치적 분석 방법이 존재한다[9][10]. Amjad[13]는 사용자가 입력한 질의에 따라 h-index를 실시간적으로 계산하는 ad-hoc h-index를 제안하였다. 이러한 지수 기법들은 연구자에 관한 연구 역량을 직관적이고 명확하게 표현되어 실생활에 자주 활용된다. 다만 이러한 수치는 특정 분야에 대한 전문가를 판별하기 매우 어렵고, 더불어서 논문에 대한 기여도 및 인용도 특성에 대한 정보를 반영하지 못하기 때문에 분별력을 갖추기가 어렵다. 예를 들어, 1편을 쓰고 100개의 인용을 받은 논문의 저자의 h-index는 1이고, 논문 2편을 쓰고 각 2개의 인용을 받은 저자는 2를 부여받는다. 이러한 상황에서는 논문을 1편 쓴 저자더라도 인용을 많이 받았기 때문에 높은 가중치를 부여해야 하지만, 제안된 지수의 특성상 높은 점수를 받지 못하는 문제점을 가지고 있다.

Wang[1]은 그래프 분석 기법과 수식 기반 기법을 혼용한 기법을 제안한다. 먼저 TF-IDF(Term Frequency-Inverse Document Frequency) 분석을 통해 시간에 따른 연구 주제 분석을 수행한다. 또한, 공저자와 인용 정보, 시간 속성을 고려하여 미래의 전문가를 예측한다. 랭킹 매트릭스를 관리하여 시간에 따른 전문가 지수를 지속해서 갱신한다. 기존 그래프 분석 기법과 달리 인용 정보가 일부 부족하더라도 합리적인 전문가 지수 계산이 가능하고, 시간 속성을 고려한 계산을 통해 신진 연구자에 대한 분석이 가능함을 입증하고 있다.

[표 1]은 기존 전문가 판별 방법과 제안하는 기법과의 비교를 나타낸다. 전문가 판별 방법의 분석 방법으로는 그래프 데이터를 기반으로 하는 그래프 분석과 논문에서 제안하는 지수를 기반으로 하는 지수 분석 방법으로 나뉜다. 그래프 기반 분석 기법은 그래프 모델링

을 수행하기 위해서 인용 그래프 데이터가 필요하다. 지수 분석은 논문의 인용 수, 저널 명, 저널 IF, 발행연도, 저자 정보의 데이터만 필요하기 때문에 인용 그래프를 구축할 필요는 없다. Wang [1]은 그래프 기법과 지수 기법을 혼용한 하이브리드 기법을 제안하고 있다. 하이브리드 기법의 수행을 위해서는 인용 그래프 구축과 더불어 논문의 필수 정보 또한 필요하다. 제안하는 기법은 지수 기반의 분석을 통해 전문가를 판별하는 기법이다. 논문의 필수 정보만을 요구하고, 6가지의 객관적인 지표를 제안함으로써 기존 기법들보다 많은 요소를 고려하고 있다.

표 1. 전문가 판별 방법 비교

기법	분석 방법	사용 데이터	특징
Dunaiski, Zhao [2,6]	그래프 분석	인용 그래프	CiteRank
Liao [4]			HITS
Lin, Jardine [3, 5]			PageRank
Li [7]	지수 분석	논문	3가지 지수
Bilir, Lin [15, 16]			전문 기관 분석
Choi [14]			4가지 지수
Amjad [13]			질의 기반 h-index
Wang [12]	단순 카운팅	논문, 특허, 연구과제	-
Wang [1]	하이브리드 (그래프 + 지수)	논문, 인용 그래프	TF-IDF, MRFRank
제안하는 기법	지수 분석	논문	6가지 지수

III. 제안하는 전문가 판별 방법

1. 전체 구조도

본 논문에서는 온라인 학술 사이트의 논문에서 얻을 수 있는 다양한 정보를 활용한 전문가 판별 방법을 제안한다. 전문가를 판별하기 위해 6가지의 전문성 지수를 제안한다. 제안하는 지표를 계산하기 위해서 논문 수집기와 수집된 논문에서의 특정한 정보를 추출하는 데이터 전처리기가 필요하다. 최종 분석된 결과는 가중치 합을 통해 랭킹 형태로 결과를 제공한다.

그림 1. 제안하는 기법의 구조도

[그림 1]은 제안하는 전문가 판별 기법의 전체 구조도를 나타낸다. 먼저, 사용자는 원하는 분야의 키워드 정보를 입력하면, 키워드와 관련된 논문을 수집한다. 논문은 학술 사이트에서 제공하는 Open API 혹은 Crawler를 통해 수집된다. 수집된 논문에서 불완전한 데이터들은 전처리기를 통해 제외된다. 예를 들어, 저자의 정보(소속, 이메일 등)가 불완전하거나 중복된 데이터들을 제거한다. 전문가 분석기는 수집된 데이터를 기반으로 모든 논문 저자에 대하여 6가지 전문성 지수를 계산한다. 전문가 분석기를 통해 계산된 지수는 값의 범위가 다르기 때문에 정규화를 통해 동일한 값의 범위로 표현한다. 사용자는 6가지 전문성 지수에 대한 가중치를 입력으로 부여하면, 각 가중치에 따라 최종적인 전문가 지수가 계산되고, 전문가 지수를 기반으로 랭킹 순으로 결과를 제공한다.

2. 데이터 수집 및 전처리

전문가 판별을 수행하기 위해서는 분석에 필요한 기반 데이터 수집이 수반되어야 한다. 제안하는 기법은 온라인 학술 사이트에서 제공하는 Open API 혹은 자체 개발한 Crawler를 통해 기반 데이터를 수집한다. 기반 데이터 수집은 사용자가 질의를 요청할 때 실시간적으로 수행한다.

[그림 2]는 제안하는 기법의 데이터 수집 및 전처리 과정을 나타낸다. 데이터 수집기는 앞서 설명한 것과 같이 Open API 또는 Crawler를 통해 논문 데이터를

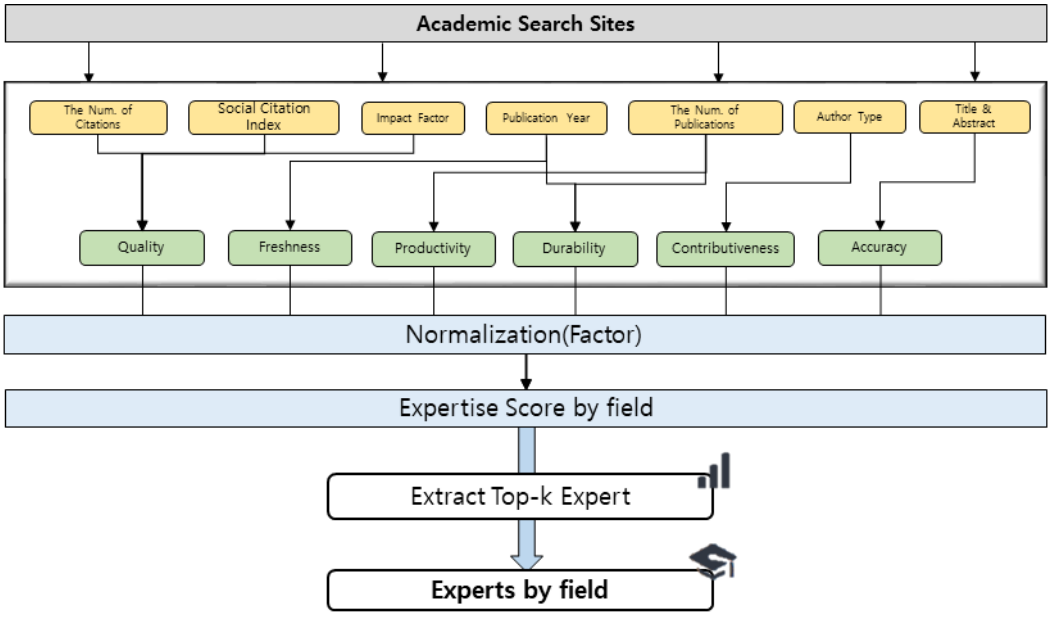


그림 3. 전문가 분석 과정

수집한다. 수집기를 통해 실시간으로 수집된 데이터는 데이터 전처리기에서 데이터의 불용성을 확인하고 데이터의 일관성이 떨어지는 데이터는 삭제한다. 제안하는 기법은 논문의 모든 저자에 대한 전문가 분석을 수행하는데, 전문가를 식별하기 위한 ID 정보가 누락되었거나, 소속 정보가 누락된 데이터는 전처리기에 의해 삭제된다. 중복 데이터 또한 마찬가지로 중복적으로 수집될 필요가 없기 때문에 이를 제거한다. 저자 정보 추출은 논문에서 저자를 식별하기 위한 정보 (소속, 이름, 저자 ID)를 추출한다. 마지막으로, 제안하는 전문성 지수를 계산하기 위한 기반 데이터를 추출하고 추출된 정보들은 데이터 저장소에 저장한다.

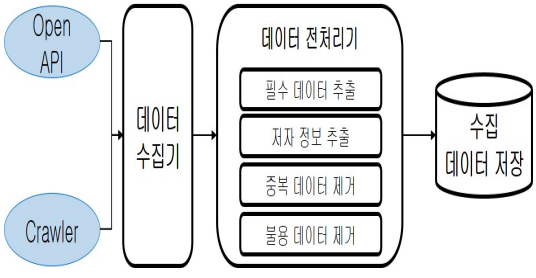


그림 2. 데이터 수집 및 전처리 과정

3. 전문성 지수 계산 및 전문가 랭킹

본 논문에서는 전문가 분석을 위해 6가지 전문성 지수를 제안한다. 6가지 전문성 지수는 품질, 생산성, 기여도, 최신성, 정확성, 지속성이다. 각 지수는 독립적으로 계산이 가능하다. 품질은 논문의 인용 정도와 가능성을 기반으로 한 정보를 바탕으로 계산한다. 인용이 많은 논문과 IF가 높은 저널에 제출된 논문일수록 인용이 될 가능성이 높기 때문에 이러한 지수를 활용한다. 생산성은 전문가가 논문을 작성하는 양에 따라 계산된다. 상대적으로 해당 분야에 많은 논문을 작성한 연구자일수록 전문가로 보는 것이 타당하기 때문이다. 기여도는 논문에 대한 기여도를 판별하기 위해서 사용된다. 논문의 공저자 보다는 주 저자, 교신 저자로 참여하는 것이 논문에 기여한 바가 높기 때문에 저자의 유형에 따라 다른 가중치를 부여하기 위해 사용한다. 최신성은 최근에 관련된 연구를 많이 한 저자에게 가중치를 부여하기 위해 사용한다. 실세계에서는 지속적으로 새로운 기술이 나오기 때문에 해당 분야에 대한 최신 연구를 수행한 연구자일수록 높은 가중치를 주는 것이 타당하기 때문이다. 정확성은 사용자가 찾고자 하는 분야 정보와의 유사성을 판별하기 위해 사용된다. 사용자가 입

력한 질의와 유사한 연구를 수행한 전문가를 찾기 위함이다. 지속성은 연구자가 얼마나 오랜 기간 연구해왔는지를 판별하기 위해 사용된다. 해당 분야에 대해 오랜 기간 동안 지속적으로 연구를 한 연구자일수록 전문가일 가능성이 높기 때문에 높은 가중치를 주는 것이 타당하다.

[그림 3]은 전문가 판별 과정을 나타낸다. 전문성 지수 계산을 위해서 앞서 수집 저장된 데이터에서 인용수, 소셜 인용 수, 저널의 IF, 논문의 출간 연도, 논문 출간 수, 저자 유형(제1저자, 공저자, 교신 저자), 논문 제목 및 요약 정보를 활용한다. 계산된 전문성 지수는 정규화를 통해 0~1 사이로 값을 변환하고, 최종 전문가 지수를 계산한다. 최종 전문가 지수는 사용자의 입력 가중치에 의해 결정되고, 전문가 지수가 높은 순으로 결과를 제공한다. 한 번 분석 되어 계산된 전문성 지수는 지속적으로 사용 가능하다. 이를 통해 사용자는 입력 가중치를 지속해서 변형하면서, 다양한 시각의 전문가 랭킹 정보를 받는다.

식 (1)은 특정 전문가 i의 전문가 지수를 나타낸다. 전문가 지수는 사용자 전문성을 판별하기 위한 값으로 6가지 요소를 이용하여 계산한다. Q는 품질을 나타내고, P는 생산성을 의미한다. C와 F는 저자의 기여도와 최신성을 의미한다. 마지막으로 A와 D는 정확성과 지속성을 의미한다.

$$ES_i = \alpha Q_i + \beta P_i + \gamma C_i + \delta F_i + \epsilon A_i + \theta D_i \quad (1)$$

식 (2)는 연구자에 대한 품질 전문성 지수를 계산하는 수식을 나타낸다. 집합 $Papers$는 특정 연구자가 저자로 포함된 논문의 목록을 나타낸다. c_p는 논문 p에 대한 인용 수를 의미한다. ds_p는 논문 p에 대한 국내외 구분 점수를 의미한다. 국외 저널/학술대회에 대한 가중치를 부여하기 위한 값이다. 본 논문에서는 국내 저널/학술대회에는 1의 가중치를 부여하고, 국외 저널/학술대회에는 3의 가중치를 부여한다. IF_p는 논문 p에 대한 IF를 나타낸다. 특정 저널은 굉장히 높은 IF가 나타나기 때문에 해당 값에 의해 다른 저자의 논문이 고려되지 않는 상황이 있을 수 있기 때문에, IF를 정규화한다. sc_p는 논문 p에 대한 소셜 인용 수를 나타낸다. 소셜 인용 수란, 특정 논문이 온라인 학술 검색 사이트에서 클릭, 공유, 다운로드 횟수를 의미한다. WoS[17] 사이트 같은 경우 이러한 수치를 사용자에게 제공하고 있다. 사용자들이 웹상에서 관심 있게 보는 논문들은 간접적으로 인용되었다고 보고, 이러한 논문들에 가중치를 부여한다.

$$Q_i = \sum_{p \in Papers} [\ln(c_p * ds_p + 1) * IF_p + \ln(sc_p)] \quad (2)$$

식 (3)은 생산성을 계산하는 수식이다. np_i는 저자 i가 저자로 포함된 논문의 수를 의미한다. 저자로 포함된 논문의 수를 log를 취하여 생산성을 계산한다. 논문의 수는 학술대회, 논문지 등 모든 논문의 수를 계산하는데, 이 때 논문의 수를 단순히 합산만한다면 생산성 지수로 인해 다른 지수가 부각되지 않을 가능성이 높다. 즉, 학술대회에서 많은 논문을 발표한 저자가 상대적으로 게재하기 어려운 논문지에 논문을 게재한 저자보다 매우 높은 가중치를 받을 수 있는 상황을 최대한 방지하기 위해 사용된다.

$$P_i = \ln(np_i) \quad (3)$$

식 (4)는 연구자의 기여도를 계산하는 수식이다. 저자가 제1저자 혹은 교신 저자이면 높은 가중치를 부여하고, 공저자라면 공저자의 순서에 따라 기여도를 계산한다. $Order_{p,i}$는 특정 연구자 i가 논문 p에 대해서 몇 번째 저자인지를 값으로 나타낸다. 즉, 공저자여도 순서가 높은 공저자일수록 논문에 많은 기여를 하였다고 보고 좀 더 높은 가중치를 부여한다.

$$C_i = \sum_{p \in Papers} \begin{cases} 1 & , First/Corresponding\ Author \\ \frac{1}{Order_{p,i}} & , Co-Author \end{cases} \quad (4)$$

식 (5)는 연구자의 최신성을 계산하는 수식이다. 최신성은 신진 연구자들에 대한 가중치를 부여하기 위해 사용된다. 최신이라고 생각하는 기준 년도 th를 기반으로

계산한다. py_p는 논문 p에 대한 출판 연도이다. 즉, 출판 연도가 기준 연도 보다 최근에 게재된 논문이라면 1을 부여하고, 그 이전 논문들은 수식에 따라 차등 된 점수를 부여한다. 예를 들어, 기준 연도 th가 2015년인 상황에서 출판 연도가 2020, 2015, 2005, 2000인 4개의 논문이 입력된 상황에서는 다음과 같이 계산한다. 2020년은 기준 연도 보다 높기 때문에 1점을 부여한다. 2015년은 기준 연도와 동일하여 수식에 의해 계산된다. 1-(2015-2015)*0.1을 계산하면 1이고, 0과 비교하여 1이 높기 때문에 마찬가지로 1점을 부여한다. 2005년은 1-(2015-2005)*0.1을 계산하면 0이기 때문에 0 점을 부여한다. 2000년은 1-(2015-2000)*0.1을 계산하면 -0.5이고, max(-0.5, 0)에 의해 0점을 부여한다. 즉, 기준 년도 보다 10년 이상 된 논문은 0점이 부여되는 수식이다.

$$F_i = \sum_{p \in Papers} \begin{cases} 1 & , py_p > th \\ \max(1-(th-py_p)*0.1, 0) & , py_p \leq th \end{cases} \quad (5)$$

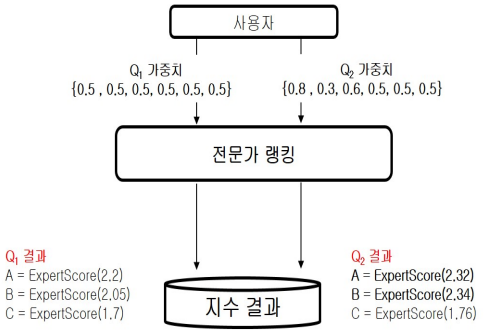
식 (6)은 연구자의 정확성을 계산하는 수식이다. sim은 두 벡터 간의 유사성을 계산하는 함수이다. 본 논문에서는 코사인 유사도를 활용한다. k는 사용자가 입력한 분야 정보를 나타낸다. $TF\text{-}IDF_k$는 사용자가 입력한 질의 k에 대한 TF-IDF 값을 의미한다. $TF\text{-}IDF_p$는 논문 p의 제목, 요약, 키워드 기반의 TF-IDF 값이다. 정확성은 사용자가 질의한 분야 정보에 대한 TF-IDF와 논문 p의 TF-IDF 값 간의 유사도 값이다. 여기서 유사도 값은 코사인 유사도 함수를 통해 계산한다. 즉, 사용자가 질의한 분야 정보와 얼마나 유사한 논문들을 작성했는지를 나타낸다.

$$A_i = \sum_{p \in Papers} sim(TF-IDF_k, TF-IDF_p) \quad (6)$$

식 (7)은 연구자의 지속성을 계산하는 수식이다. 지속성은 연구자가 얼마나 해당 분야에 대한 연구를 지속해 오는지를 나타낸다. 해당 분야에 대한 연구를 전체 연구 경력 대비 얼마만큼의 지속성을 가졌는지를 계산한다. $py_{max,i}$는 저자로 포함된 논문 중 가장 최근 년도 값을 가져오고, $py_{min,i}$는 반대로 가장 오래된 년도 값을 가져온다. 두 값의 차이를 통해 해당 분야에 대한 총 연구 기간을 계산 할 수 있다. ml_i는 저자의 지속적인 연구 기간을 계산한다. 예를 들어, 2011, 2012, 2013, 2014 총 4년간 연속적으로 논문을 출간하였다면 4라는 값을 갖는다. 만약, 2011, 2012, 2014 세 개의 논문을 출간하였다면 2라는 값이 계산됨으로써, 지속적인 연구를 한 연구자에게 높은 가중치를 부여한다.

$$D_i = \frac{ml_i}{py_{\max,i} - py_{\min,i}} \quad (7)$$

[그림 4]는 전문가 랭킹 예시이다. 사용자는 키워드에 전문가 요청을 수행하고, 결과에 대해서 다양한 가중치를 입력할 수 있다. 가중치란 제안하는 6개의 전문성 지수에 대해 어느 지표를 조금 더 의미 있게 볼 건지를 의미한다. 한 번 계산된 전문성 지수는 지표별로 정규화되어 데이터 저장소에 저장된다. 사용자는 다음에 입력 가중치만 부여하면 해당 가중치에 따라 최종 전문가 지수를 계산하고 전문가 지수가 높은 순서대로 결과를 제공한다. 예를 들어, [그림 4]와 같이 A,B,C 전문가에 대해 전문성 지수가 계산되어 있다고 했을 때, 6개의 지표에 대해 동일한 가중치를 부여한 첫 요청 (Q_1) 이 들어오면, A 전문가는 2.2의 전문가 점수를, B 전문가는 2.05의 전문가 점수를, C 전문가는 1.7이라는 전문가 점수가 계산되어 A,B,C 순서대로 랭킹을 제공한다. 차후에 사용자가 품질과 기여도에 대한 가중치를 조금 더 상승시키고, 생산성에 대한 가중치를 감소시킨 요청 (Q_2)을 수행하면, B 전문가가 A 전문가보다 높은 전문가 점수로 계산되어 B,A,C 순서대로 랭킹을 제공한다. 추가로 랭킹에서 사용자가 상위 몇 명의 전문가를 볼 건지에 대한 인자 k를 입력으로 줄 수 있다.



ID	Quality	Productivity	Cont.	Freshness	Acc.	Durability
A	0.8	0.9	0.6	0.5	0.6	1
B	1	0.4	0.7	0.4	0.8	0.8
C	0.6	0.8	0.4	0.4	0.5	0.7

그림 4. 전문가 랭킹 예제

IV. 성능 평가

제안하는 전문가 판별 방법의 우수성을 증명하기 위해서 기존 기법과의 성능 비교 평가를 수행한다. [표 2]는 본 논문에서 수행한 성능 평가 파라미터를 나타낸다. 분석 기반이 되는 논문은 국내 논문 검색 사이트와 해외 논문 검색 사이트를 기반으로 데이터를 수집하였다. 실험 데이터로는 논문의 제목, 키워드, 요약, 인용수, 출간 연도, 저자명, 국내외 구분 정보를 수집하였다. 또한, 분석에 활용될 국내외 IF 지수를 별도로 수집하였다. 수집은 2010~2020년간의 특정 키워드에 대한 정보를 수집하였으며, 총 2만 4천여 건의 논문과 5만여 명의 분석 대상 전문가가 수집되었다. 연구자들의 연구 프로젝트 참여 정보와 논문 출판 정보를 기반으로 전문가 집단을 구성하였다. 해당 분야에서 평가위원으로 활발히 활동 중이신 전문가의 조언에 따라 20여 명의 정답 셋을 구축하였다. 전문가 점수를 기반으로 랭킹을 수행하고, 상위 k명과 정답 셋을 비교하여 정밀도(Precision), 재현율(Recall), F1 measure를 계산하여 비교한다. k는 1~20 사이로 변화하면서 성능 평가를 수행하였다. 성능 평가 비교 대상은 본 연구팀이 제안한 기존 전문가 분석 시스템[14]과 Amjad[13]가 제안한 ad-hoc h-index와의 비교를 수행한다.

표 2. 성능 평가 파라미터

파라미터	값
수집 출처	국내 논문 및 국외 논문
보조 지표	KCIIF[18], SCIIF[17]
수집 논문(건)	24,017
수집 전문가(명)	51,416
정답 셋(명)	20
k	1,2,5,10,20
구현 언어	Python 3.6.8

[그림 5]는 국내 논문을 기반으로 한 전문가 추천의 정밀도 결과를 나타낸다. k는 전문가 점수를 기반으로 상위 k명을 선택한 값으로 최대 전문가 풀인 20까지 늘려가며 성능 평가를 수행하였다. 제안하는 기법은 1명을 추천할 때 정확한 결과를 나타냈기 때문에 정밀도가 1로 나온 것을 확인 할 수 있다. 여기서 k가 1일 때, 기존 기법[14]이 0.33이라는 수치를 나타내는데, 그 이유는 국내 논문 검색 사이트별 전문가 검색을 수행하고, 각 사이트별로 계산한 정밀도 값을 평균 한 값이기 때문이다. 다시 말해, 제안하는 기법은 모든 검색 사이트에서 1명의 전문가를 모두 찾아낼 수 있다는 얘기이다. ad-hoc h-index 기법은 사용자가 요청한 질의와 관련된 논문에 대해서만 h-index를 계산하고 이를 기반으로 랭킹을 제공한다. 비교적 국내 논문은 인용 수가 높지 않기 때문에 낮은 k에서는 정확한 전문가를 찾기 어렵지만, 높은 k(20명)를 추천했을 때는 평균 3명의 전문가를 찾음으로써, 본 연구팀의 기존 기법[14]보다는 좋은 성능을 나타내고 있다.

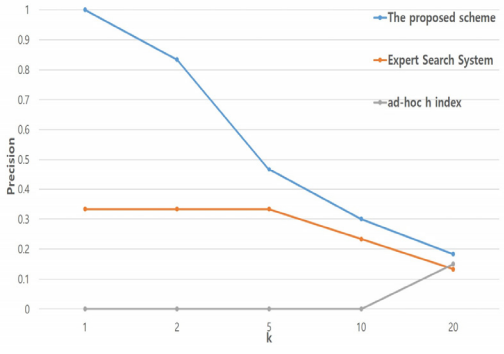


그림 5. k에 따른 전문가 분석 정밀도 평가 (국내)

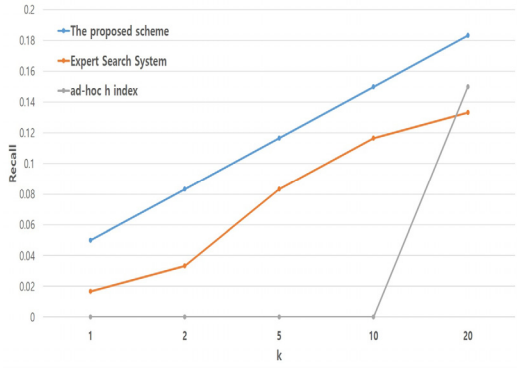


그림 6. k에 따른 전문가 분석 재현율 평가 (국내)

[그림 6]은 국내 논문을 기반으로 한 전문가 추천의 재현율 결과를 나타낸다. 세 기법 모두 k가 늘어남에 따라 재현율 수치가 상승하는 것을 볼 수 있다. 특히 k가 20인 경우 가장 좋은 성능을 나타낸다. 추천을 많이 하면 할 수록 전문가 집단을 찾을 확률이 높아진다는 것이다. 제안하는 기법은 k가 20일 때 평균 0.18이라는 수치를 나타내며, 수치적으로 해석하게 되면, 20명의 정답 데이터 중 평균 3~4명 정도의 전문가를 찾을 수 있으며, 10명을 추천할 경우 3~4명이 정답에 포함되는 것으로 해석 할 수 있다. 기존 기법은 k가 20일 때 대략 0.13 정도의 재현율 값을 가진다. 이는 대략 2~3명 정도의 전문가를 찾는다는 것으로 해석 할 수 있다. ad-hoc h-index는 정밀도 평가와 마찬가지로 k가 20일 때 가장 좋은 성능을 나타낸다.

[그림 7]은 국내 논문을 기반으로 한 전문가 추천의 F1 measure 결과를 나타낸다. F1 measure는 정밀도와 재현율의 조화 평균이기 때문에, 앞선 두 가지 성능 평가에서 가장 높은 성능을 낸 제안하는 기법이 가장 좋은 결과를 나타낸다. 제안하는 기법과 기존 기법[13]은 k가 10일 때 가장 좋은 성능을 나타내며, ad-hoc h-index는 k가 20일 때 가장 좋은 성능을 나타낸다. 국내 논문 기반의 성능 평가를 통해 제안하는 전문가 분석 기법의 타당성을 입증하였다.

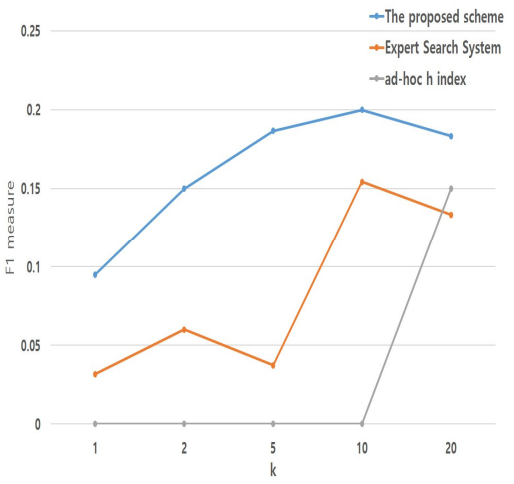


그림 7. k에 따른 전문가 분석 F1 measure 평가 (국내)

[그림 8]은 국외 논문을 기반으로 한 전문가 추천의 정밀도와 재현율 결과를 나타낸다. 국외 논문에서 ad-hoc h -index 기법은 구축한 정답 집합의 전문가를 찾아 주지 못하여 해당 성능 평가에서는 제외되었다. 앞선 국내 논문 성능 평가와 마찬가지로 국외 논문 기반의 성능 평가에서도 제안하는 기법이 기존 기법보다 우수한 성능을 나타낸다. 특히 k가 1~2인 경우 1의 정밀도 값을 보임으로써 정확한 결과를 생성하는 것을 볼 수 있다. 기존 기법은 k가 5일 때 준수한 성능을 나타낸다.

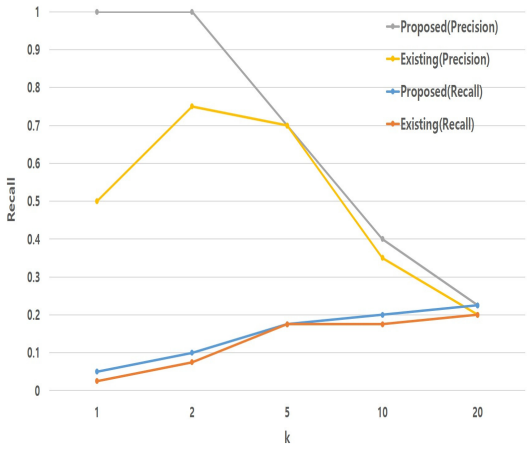


그림 8. k에 따른 전문가 분석 정밀도/재현율 평가 (국외)

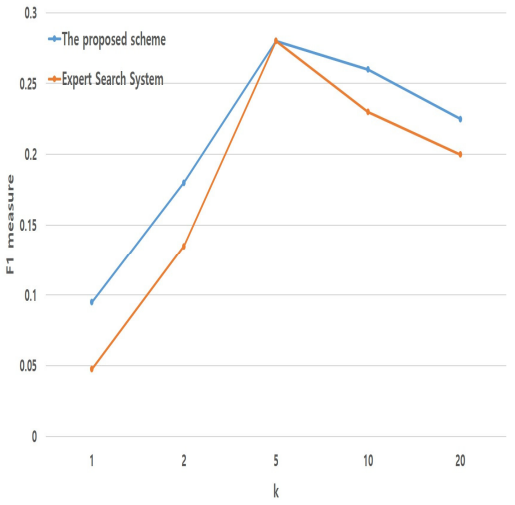


그림 9. k에 따른 전문가 분석 F1 measure 평가 (국외)

[그림 9]는 국외 논문을 기반으로 한 전문가 추천의 F1 measure 결과를 나타낸다. 정밀도와 재현율의 조화 평균값이기 때문에 k가 5일 때를 제외하고 모든 상황에서 기존 기법보다 좋은 성능을 나타낸다. 국외 논문 기반 성능 평가에서도 제안하는 전문가 분석 기법의 타당성을 입증함으로써, 실제 논문 검색 시스템 기반으로 한 전문가 분석이 유효함을 입증하였다.

V. 결 론

본 논문에서는 온라인 학술 사이트의 논문에서 얻을 수 있는 다양한 정보를 활용한 전문가 판별 방법을 제안하였다. 전문가를 판별하기 위해 품질, 생산성, 기여도, 최신성, 정확성, 지속성이라는 6가지의 전문성 지수를 제안하였다. 기존에는 반영하지 못하였던 온라인 학술 사이트의 특성을 반영하기 위해 클릭/다운로드와 같은 소셜 인용 수를 품질 지수에 반영하였다. 국내외 논문 기반의 성능 평가를 통해 제안하는 기법의 타당성과 실현성을 입증하였다. 본 논문에서는 제안하는 객관적인 6개의 전문성 지수의 타당성을 입증하기 위해서 다양한 성능 평가를 수행하였다. 다만 분야에 따라 논문의 특성이 다르고 이러한 특성을 반영하기 위해서는 추가적인 실험과 지수 분석이 필요하다. 또한, 정적으로 고정 된 지수 분석보다는 지속적으로 변화가 가능한 형태의 지수를 제안하는 것이 필요하다. 이를 위해서 향후에는 딥 러닝 기반의 가중치 학습이 가능한 전문성 지수 기법으로 발전시킬 예정이다. 또한 제안하는 기법의 분석 결과를 시각화하여 사용자에게 유의미한 정보를 제공할 예정이다.

참 고 문 헌

[1] S. Wang, S. Xie, X. Zhang, Z. Li, P. S. Yu, and X. Shu, "Future Influence Ranking of Scientific Literature," Proc. SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, pp.749-757, 2014.

[2] M. Dunaiski and W. Visser, "Comparing Paper Ranking Algorithms," Proc. South African Institute for Computer Scientists and Information Technologists Conference, pp.21-30, 2012.

[3] L. Lin, Z. Xu, Y. Ding, and X. Liu, "Finding Topic-Level Experts in Scholarly Networks," Scientometrics, Vol.97, No.3, pp.797-819, 2013.

[4] H. Liao, R. Xiao, G. Cimini, and M. Medo, "Ranking Users, Papers and Authors in Online Scientific Communities," arXiv preprint arXiv:1311.3064, 2013.

[5] J. Jardine and S. Teufel, "Topical PageRank: A Model of Scientific Expertise for Bibliographic Search," Proc. European Chapter of the Association for Computational Linguistics, pp.501-510, 2014.

[6] F. Zhao, Y. Zhang, J. Lu, and O. Shai, "Measuring Academic Influence using Heterogeneous Author-citation Networks," Scientometrics, Vol.118, No.3, pp.1119-1140, 2019.

[7] X. Li and T. Watanabe, "Automatic Paper-to-Reviewer Assignment, based on the Matching Degree of the Reviewers," Proc.

Computer Science, Vol.22, pp.633-642, 2013.
[8] M. T. Afzal and H. A. Maurer, "Expertise Recommender System for Scientific Community," Journal of Universal Computer Science, Vol.17, No.11, pp.1529-1549, 2011.
[9] M. Raheel, S. Ayaz, and M. T. Afzal, "Evaluation of h-Index, Its Variants and Extensions based on Publication Age & Citation Intensity in Civil Engineering," Scientometrics, Vol.114, No.3, pp.1107-1127, 2018.
[10] M. Ameer and M. T. Afzal, "Evaluation of h-Index and Its Qualitative and Quantitative Variants in Neuroscience," Scientometrics, Vol.121, No.2, pp.653-673, 2019.
[11] N. Crespo and N. Simoes, "Publication Performance Through the Lens of the h-index: How Can We Solve the Problem of the Ties?," Social Science Quarterly, Vol.100, No.6, pp.2495-2506, 2019.
[12] Q. Wang, J. Ma, X. Liao and W. Du, "A Context-Aware Researcher Recommendation System for University-Industry Collaboration on R&D Projects," Decision Support Systems, Vol.103, pp.46-57, 2017.
[13] T. Amjad and A. Daud, "Indexing of Authors According to Their Domain of Expertise," Malaysian Journal of Library and Information Science, Vol.22, No.1, pp.69-82, 2017.
[14] D. Choi, H. Lee, K. Bok, and J. Yoo, "Design and Implementation of an Academic Expert System Through Big Data Analysis," The Journal of Supercomputing, pp.1-25, 2021.
[15] S. Bilir, E. Gogus, O. Tas, and T. Yontan, "A New Ranking Scheme for the Institutional Scientific Performance," arXiv preprint arXiv:1508.03713, 2015.
[16] C. S. Lin, M. H. Huang, and D. Z. Chen, "The Influences of Counting Methods on University Rankings based on Paper Count and Citation Count," Journal of Informetrics, Vol.7, No.3, pp.611-621, 2013.
[17] https://www.webofknowledge.com/, 2021.09.17
[18] https://www.kci.go.kr/, 2021.09.17

저 자 소 개

최 도 진(Do-Jin Choi) 정회원

- 2014년 2월 : 한국교통대학교 컴퓨터공학과(공학사)
- 2016년 2월 : 한국교통대학교 컴퓨터공학과(공학석사)
- 2020년 2월 : 충북대학교 정보통신공학과(공학박사)
- 2020년 3월 ~ 2020년 8월 : 충북대학교 정보통신공학과 박사후연구원(Postdoc)

〈관심분야〉 : 연속 질의 처리, 그래프 스트림, 빅데이터

오 영 호(Young-Ho Oh) 준회원

- 2015년 8월 : 충북대학교 공업화학과(공학사)
- 2020년 3월 ~ 현재 : 충북대학교 빅데이터 협동과정(석사과정)

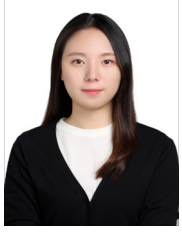
〈관심분야〉 : 연속 질의 처리, 그래프 스트림, 빅데이터

편 도 웅(Do-Woong Pyun) 준회원

- 2020년 2월 : 가천대학교 산업경영공학과(공학사)
- 2020년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

〈관심분야〉 : 소셜 네트워크, 빅데이터

방 민 주(Min-Ju Bang) 준회원



- 2021년 2월 : 충북대학교 정보통신공학부(공학사)
- 2021년 3월 ~ 현재 : 한양대학교 인공지능학과 석사과정

〈관심분야〉 : 추천시스템, 인공지능, XAI, 빅데이터

전 종 우(Jong-Woo Jeon) 준회원

■ 2016년 3월 ~ 현재 : 충북대학교 정보통신공학부 학사과정

〈관심분야〉 : 소셜 네트워크, 빅데이터

이 현 병(Hyeon-Byeong Lee) 준회원

■ 2016년 8월 : 한국교통대학교 컴퓨터공학과(공학사)
■ 2018년 8월 : 한국교통대학교 컴퓨터공학과(공학석사)
■ 2019년 3월 ~ 현재 : 충북대학교 정보통신공학과(박사과정)

〈관심분야〉 : 그래프 스트림, 빅데이터, 데이터베이스 시스템

박 득 배(Deukbae Park) 정회원

■ 2011년 2월 : 충북대학교 정보통신공학과(공학석사)
■ 2013년 8월 : 충북대학교 정보통신공학과(공학박사 수료)

〈관심분야〉 : 빅데이터, 네트워크, 데이터베이스, 5G 소셜 네트워크

임 종 태(Jong-Tea Lim) 정회원

■ 2009년 2월 : 충북대학교 정보통신공학과(공학사)
■ 2011년 2월 : 충북대학교 정보통신공학과(공학석사)
■ 2015년 8월 : 충북대학교 정보통신공학과(공학박사)
■ 2015년 9월 ~ 2019년 8월 : 충북대학교 정보통신공학과 Postdoc.
■ 2019년 10월 ~ 현재 : 충북대학교 전자정보대학 정보통신 공학부 초빙 조교수

〈관심분야〉 : 소셜 미디어, 빅데이터, 시공간 데이터베이스, 위치기반 서비스 등

복 경 수(Kyoung-Soo Bok) 종신회원

■ 1998년 2월 : 충북대학교 수학과(이학사)
■ 2000년 2월 : 충북대학교 정보통신공학과(공학석사)
■ 2005년 8월 : 충북대학교 정보통신공학과(공학박사)
■ 2005년 3월 ~ 2008년 2월 : 한국과학기술원 정보전자연구소 Postdoc
■ 2008년 3월 ~ 2011년 2월 : 가인정보기술 연구소 차장
■ 2011년 3월 ~ 2019년 8월 : 충북대학교 전자정보대학 정보통신공학부 초빙교수
■ 2019년 9월 ~ 현재 : 원광대학교 SW 융합학과 조교수

〈관심분야〉 : 데이터베이스 시스템, 이동 객체 데이터베이스, 이동 P2P 네트워크, 소셜 네트워크 서비스, 빅데이터 처리 등

유 효 근(Hyo-Keun Yoo) 정회원

■ 2015년 2월 : 광운대학교 전자융합공학과(공학사)
■ 2017년 2월 : 한국과학기술원 전기및전자공학과(공학석사)
■ 2018년 3월 ~ 현재 : 현대엔지비(주) 기술협력팀 매니저

〈관심분야〉 : 데이터 기반 전문가 추천

유 재 수(Jae-Soo Yoo) 종신회원

■ 1995년 2월 : 한국과학기술원 전산학과(공학박사)
■ 1995년 2월 ~ 1996년 8월 : 목포대학교 전산통계학과 전임강사
■ 1996년 8월 ~ 현재 : 충북대학교 전자정보대학 정보통신공학부정교수
■ 2009년 3월 ~ 2010년 2월 : California State University, 방문교수
■ 2019년 9월 ~ 2020년 8월 : California State University, 방문교수

〈관심분야〉 : 데이터베이스 시스템, 멀티미디어 데이터베이스, 빅데이터 등