「정보시스템연구」 제27권 제4호
한국정보시스템학회
2018년 12월, pp. 71~96

http://dx.doi.org/10.5859/KAIS.2018.27.4.71

텍스트 마이닝 방법론과 메신저UI를 활용한 융합연구 촉진을 위한 연구자 및 연구 분야 추천 시스템의 제안*

양낙영** · 김성근*** · 강주영****

〈목 차〉

Ⅰ. 서 론
Ⅱ. 관련 연구
2.1 융합 연구
2.2 연구자 협업 서비스
2.3 단어 임베딩과 토픽모델링
2.4 비즈니스 메신저
Ⅲ. 연구 방법
3.1 데이터 수집
3.2 데이터 전처리
3.3 워드투벡 모델 학습
3.4 LDA 토픽모델링 학습
3.5 비즈니스 메신저 연동
Ⅳ. 연구자 추천 시스템
Ⅴ. 연구 모델 평가
Ⅵ. 결론 및 시사점
6.1 결과 및 토의
6.2 시사점 및 연구의 한계
참고문헌
<Abstract>

Ⅰ. 서 론

현대 과학기술과 정보 통신 기술이 급격히 발전하면서 대중들의 기술 활용에 대한 진입 장벽은 매우 낮아졌으며 기술의 활용 폭 또한 폭발적으로 넓어졌는데, 이는 곧 연구 분야들 사이의 경계를 흐릿하게 만들었고, 특정 분야에 한정되어 수행되었던 기존의 전문 연구와는 다른, 여러 분야를 망라하는 새로운 형태의 연구를 탄생시켰다. 이렇게 탄생한 새로운 연구 패러다임이 바로 융합 연구이다.

융합 연구가 하나의 연구 형태로 정착하면서, 미국에서는 NBIC(NT, BT, IT, CS)를 전면에 내세워 기술을 중심으로 한 융합 연구 전략을 수립해 실행해왔고, 유럽은 더 나아가 인문 및 사회과학과 공학 기술과의 융합을 모두 포함하는 포괄적인 연구 전략을 수행해왔다(조양래

* 본 연구는 2017년 아주대학교 일반연구비 지원에 의해 연구되었음.
** 아주대학교 e-비즈니스학과, yny0506@ajou.ac.kr
*** 아주대학교 연구팀, kimsungguen@ajou.ac.kr
**** 아주대학교 e-비즈니스학과, jykang@ajou.ac.kr(교신저자)

등, 2015). 국내의 경우에는 2008년 교육과학기술부 주관으로 국가 융합기술발전 기본계획이 수립되었고, 현재는 과학기술정보통신부 산하에서 융합 연구 관련 개발 사업들이 관리되고 있다. 최근 국내에서는 4차 산업 혁명 등의 키워드와 함께 융합 연구에 대한 관심이 절정에 이르렀고, 이러한 사회적 분위기와 더불어 과학기술정보통신부는 2018년 제3차 융합연구개발 활성화 기본계획을 수립함으로써 융합 연구가 국내 R&D 발전의 촉매제 역할을 할 수 있도록 그 발판을 마련하였다. 이렇게 세계 각국 정부들은 융합 연구 분야의 성장을 위해 지속적인 관심을 가지고 지원하는 모습을 보여주었는데, 융합 연구 분야에 대한 국내외 정부들의 막대한 예산 산정 또한 이를 뒷받침해준다. 예를 들어 미국의 NNI(National Nano-technology Initiative)는 총 22개의 정부 부처로 구성된 대규모 융합 연구 사업인데, 2006년도 예산 할당이 이미 10억 달러를 초과하였고(박기범, 황정태, 2007) 2017년도에는 14.7억 달러의 예산을 할당 받는 등 미국 정부의 강력한 지원 속에서 성장하고 있다. 국내의 경우도 2010년 융합 부문 사업에 투자되었던 R&D 예산 금액이 1조 9,250억이었던 데에 비해 2017년의 예산 금액은 3조 5,495억으로 책정되어 무려 2배 가까이 증가하였는데, 이는 국내에서도 융합 연구 분야가 정부차원에서 지대한 관심을 받고 있음을 시사한다. 또한, 정부뿐만 아닌 각 대학의 경우에도 융합 연구에 대한 정부차원의 커다란 관심에 발맞추어 융합 관련 학과를 신설하고 운영하면서 융합 연구에 능통한 인력을 양성하고 있고, 기업차원에서도 융합 연구에 대한 사내 지원을 아끼지 않고 있으며, 실제로 융합기술을 활용한 스타트업도 우후죽순으로 생겨나고 있는 추세이다. 이처럼 최근 융합 연구에 대한 사회적 관심은 매우 폭발적이며, 정부 및 기업 관계자와 기타 연구자들도 융합 연구의 가치를 매우 높게 인식하고 융합 연구에 주목하고 있는 상황이다. 실제로 과학기술정보통신부에서 다양한 분야의 연구자들을 대상으로 2017년에 실시한 융합연구개발 활성화 설문조사에 의하면, 전체 응답자(3,873명) 중 “연구개발의 성공적 수행 또는 지속 가능한 혁신 활동을 위해 융합 연구가 필요하다”고 답한 의견은 91.4%(3,539명)로 대부분의 연구자가 융합 연구의 가치를 매우 높게 인식하고 있는 것으로 나타났다.

앞서 언급했던 내용과 같이 융합 연구 산업은 지대한 관심의 중심에 서있다. 하지만 기존 연구는 한 명의 연구자에 의해 단독으로 수행되는 경향이 있었고, 타 연구자와 협업을 진행하더라도 평소 가깝게 지내던 지인 또는 지인의 소개로 조우한 연구자에 한정되어 수행되었다. 대학이나 학회 등에서 타 분야 연구자와 우연한 만남으로 협업할 기회를 얻을 수도 있으나, 알맞은 협업 상대를 찾을 확신이 없는 상황에서 학업과 연구로 바쁜 연구자들이 학회 등을 직접 오가며 연구 인력을 발 벗고 찾아 나서기는 매우 어려운 실정이다. 실제로 융합 연구 입문 및 수행 과정을 서술한 질적 연구에 의하면, 타 분야 연구자와 접촉할 수 있는 주요 통로는 연구실 단위의 연구 내용 확인, 그리고 학회나 대학 내에서 이루어지는 사적이고 우연한 만남인 것으로 나타났다(오헌석 등, 2012). 즉, 타 분야 연구자와의 협업을 위해 구체적인 협업 인프라를 활용하기 보다는 연구실 단위의

대략적인 연구 내용 확인, 해당 분야 연구자와의 우연한 만남 등으로 융합 연구가 수행된 다는 것이다. 융합 연구 수행에 있어서 이와 같은 난관은 융합 연구를 기획하고 수행하려는 연구자들이 타 분야 연구 인력을 쉽게 모집할 인프라가 제대로 구축되어 있지 않아 발생한 것으로 추측된다. 실제로 융합연구개발 활성화 설문조사(2017)에 의하면, “융합 연구자 및 연구 성과 검색을 위한 정보공유 플랫폼이 필요하다”라고 답한 응답이 76.1%(2,949명)을 기록했다. 따라서 타 분야 연구 인력과 쉽게 접촉할 수 있도록 돕고 연구자들 사이의 즉각적인 정보 교류 또한 가능하게 하는 인프라만 확실히 구축해낸다면 융합 연구를 위한 타 분야 인력 수급에 박차를 가함으로써 융합 연구 분야의 급격한 성장을 이끌어 낼 것으로 여겨진다.

이에 본 연구에서는 협업을 원하는 분야의 키워드를 입력하면 해당 분야의 연구자 목록을 출력하고 연락처 및 소속 등을 추가적으로 게재해 그들과 쉽게 접촉할 수 있도록 돕는 연구자 추천 플랫폼을 구현하였다. 또한 사용자들이 추천 시스템을 더 손쉽게 사용할 수 있도록 분야별로 대표적인 연구 키워드를 제안하는 기능을 구현하였다. 연구자 추천 플랫폼을 통해 융합 연구를 계획하고 있는 연구자들이 타 분야의 연구 인력을 쉽게 모집하도록 돕는 인프라를 구축할 것이고, 나아가 융합 연구 분야의 성장을 이끌어 낼 기반을 다지려 한다. 연구자 추천 시스템 구현에는 워드투벡 알고리즘을 활용하였고, 특정분야의 연구 키워드 제안을 위해서는 다수의 문서에서 주제를 추출하는 텍스트 마이닝 기법인 LDA 토픽모델링(LDA Topic Modeling) 알고리즘을 활용하였다. 워드투벡과 LDA 토픽 모델링의 학습 데이터로는 하버드 대학교(Harvard University)의 교수논문 제목을 수집하여 활용하였다. 또, 연구자 추천 시스템을 활용하는 사용자들에게 이용 편의성을 제공하기 위해 시스템의 UI로 비즈니스 메신저의 일종인 슬랙(Slack)을 활용하였다.

본 연구의 구성은 다음과 같다. 우선 제2장에서는 본 연구에 대한 관련 연구들을 소개한다. 제3장은 구체적인 연구 방법에 대해 서술한다. 제4장은 본 연구에서 구현한 연구자 추천 시스템의 구성과 추천 프로세스에 대해 설명한다. 제5장에서는 구현된 추천시스템을 평가할 것이며, 마지막으로 제6장에서는 연구로부터 도출된 결론과 수행한 연구로부터 얻을 수 있는 시사점, 후속 연구 제안, 그리고 본 논문의 한계에 대해 논하고자 한다.

Ⅱ. 관련 연구

2.1 융합 연구

융합 연구라는 개념과 관점은 정부 부처・연구자・국가별로 큰 차이를 보이며, 아직 학술적으로 명확히 정의되어 있지 못하다(이광호 등, 2013). 이에 이광호 등(2013)은 융합 연구에 대해 서술했던 여러 문서들을 검토하고 융합 연구의 개념을 보다 체계화하고자 시도하였다. 이들이 체계화한 결과, 융합 연구는 세 가지 유형으로 분류될 수 있었는데, 다학제적 연구(multidisciplinary research), 학제간 연구(interdisciplinary research), 초학제적 융합 연구 (trans- disciplinary research)가 바로 그 것이

다.

다학제적 연구는 특정 연구를 위해 다수의 학문 분야가 협업하는 것을 의미하며, 이 때 각각 분야에 해당하는 세부적인 문제를 해결한 뒤 단순히 합하기 때문에 연구가 종료된 이후 기존 학문 분야에는 변화가 없다는 것이 특징이다(Mallon and Bunton, 2005; 이광호 등, 2013). 다학제적 연구의 대표 사례로는 미국의 맨하튼 프로젝트가 있다.

학제간 연구는 특정 연구를 위해 다수의 학문 분야가 협업한다는 점에서는 다학제적 연구와 비슷하나, 서로 다른 분야의 지식과 기술이 결합되어 새로운 접근 방식으로 문제를 해결한다는 점과 연구가 종료된 이후 기존의 학문분야에 변화가 발생하고 학문분야 간 교류가 활성화된다는 점에서 차별성이 있다(Schummer, 2004; 이광호 등, 2013).

마지막으로 초학제적 연구는 단일 학문 분야나 과학기술의 영역을 초월한 새로운 문제해결을 위해 다수의 학문분야가 개념, 이론 등을 결합하여 새로운 학문 분야를 창출하는 것이다(이광호 등, 2013). 초학제적 연구의 예로는 합성생물학 (syntehtic biology), 인지과학 (cognitive science), 인공지능 (artificial intelligence) 등이 제시되고 있다.

초기에는 학제간 융합 연구가 매우 활발했으나, 그 불확실성이 커지면서 오히려 다학제적 융합 연구로 추세가 전환되는 양상을 보였었다(Raasch et al., 2013). 하지만, 최근 과학기술 및 정보통신기술이 매우 빠른 속도로 발전하고 있으므로 학제간 연구 및 초학제적 연구들이 수행되는 빈도도 빠르게 증가할 것이며, 그와 더불어 융합 연구 분야는 급격히 성장하게 될 것으로 기대되고 있다.

융합 연구의 필요성과 그 성과에 대한 연구는 꽤 오랜 기간에 걸쳐 끊임없이 수행되어왔다. 송위진(2010)은 IT, BT, NT 등 기반기술(generic technology)은 모든 활동에 활용될 수 있는 기술로 인간의 경제 및 사회 활동 전반을 변화시키고 있으며, 이러한 기반기술을 근간으로 한 신 기술의 등장과 확산은 과학기술과 사회의 동시구성에 대한 융합적 시각을 요구한다고 주장하였다.

배형준(2014)은 2단계 최소제곱모형(2SLS)을 활용하여 융합 연구 수행이 연구 성과에 미치는 효과를 규명하려 하였는데, 융합 연구 수행이 양적 지표인 논문 편수와 질과 양의 통합지표인 IF합계에 큰 영향을 끼친다는 결과를 도출했다. 최종화 등(2017)은 정량적 및 정성적 종속변수에 협력적 융합 연구가 미치는 영향을 다중회귀분석을 활용해 실증 규명하였다. 분석결과, 최종화 등(2017)은 협력적 융합 연구가 다양한 분야의 지식 창출, 새로운 지식 창출, 학술적 연구 성과 창출, 연구자 간 지속적 교류 모두에 긍정적인 영향을 끼친다고 결론을 내렸다. 이들 연구는 융합 연구가 성과의 측면에서 그 가치가 충분함을 보여주었고, 연구 성과와 같은 명시적인 측면뿐만 아니라 연구자 간 교류 확대와 같은 잠재적인 측면에서도 가치가 있음을 보여 융합 연구의 필요성을 입증하였다(이무순, 손달호, 2017; 김진완 등, 2008).

2.2 연구자 협업 서비스

연구자 협업 서비스에 대한 연구는 종종 시도되었는데, 정한민 등(2005)은 연구자 간 협업

지원 서비스를 위한 지식 베이스를 설계하려 하였고, 정한민 등(2006)은 시맨틱 웹(semantic web)을 활용하여 연구자 네트워크, 연구자 추적 등의 서비스를 제공하는 플랫폼 환경을 구현하려 하였다. 이 때, 이들 연구에서는 연구자 협업 서비스를 온톨로지(ontology)를 활용한 의미 기반 검색 시스템으로 구현하려 시도하였다. 이렇듯 온톨로지를 중심으로 연구자 협업 서비스를 구현하려는 시도가 종종 존재하였던 것에 반해, 기계학습 또는 딥 러닝 기반의 연구자 협업 서비스에 대한 연구는 국내외로 매우 미미한 실정이다. 높은 성능으로 최근 주목받고 있는 기계학습 및 딥 러닝 기술을 활용해 연구를 수행한다면 분명 연구자 협업 서비스 분야의 기술적 성장을 도모할 수 있을 것이다.

현재 연구자들의 협업을 지원하는 대표적인 상용 서비스로는 Clarivate Analytics사의 InCites와 Elsevier사의 SciVal을 들 수 있다. 이들은 Web of Science나 Scopus등의 학술전문 DB를 보유하고 있기 때문에 연구자가 소속된 연구기관, 공동저자에 대한 정보 등을 바탕으로 추천 컨텐츠를 제공하며, 협업 가능한 연구기관이나 해당 연구기관의 연구자를 추천해주는 기능을 탑재하고 있다. 하지만, 이들은 과거 공동연구가 이뤄졌던 기관이나 연구자들에 대한 이력을 바탕으로 추천을 할 뿐, 논문의 제목을 구성하고 있는 단어들의 의미까지 활용하여 연구자나 기관을 추천하지는 않는다. 또한 수 천불에 달하는 서비스 이용료를 매년 지불해야 하기 때문에 중소연구기관의 경우 이러한 공동연구를 지원하는 도구를 도입하는 것이 쉽지 않은 실정이다.

2.3 단어 임베딩과 토픽모델링

Cambria and White(2014)는 자연어 처리(NLP, Natural Language Processing)를 인간 언어 분석과 표현의 자동화를 위해 고안된 계산 기법으로 정의했다. 이는 컴퓨터가 이해할 수 있도록 인간의 언어를 수치화하여 표현하는 것이라고 말할 수 있는데, 이를 위해서는 언어의 문법, 의미, 추론 모두의 상호작용을 섬세하게 구성해 결합한 프로그램이 필요하다(Winograd, 1972). 자연어처리 분야에서는 이러한 모델링을 위해 다양한 가정들이 제시되었고, 분산 가정도 그 중의 하나로써 논의되었다. 분산 가정(distributional hypothesis)은 의미가 유사한 단어는 비슷한 맥락상에 위치하는 경향이 있을 것이라는 내용을 기반으로 하는데(Rubenstein and Goodenough, 1965), 단어를 다차원 공간상의 벡터로 변환하는 단어 임베딩(word embedding)은 이러한 분산 가정을 근거로 고안되었다(Young et al., 2017).

단어를 벡터로 변환하기 위해 초기에 고안되고 사용되었던 One-hot Encoding 은 각 단어에 번호를 매긴 후 그 번호에 해당하는 요소에만 1을 부여하고 나머지 요소에는 0을 부여하는 방식으로, 총 n개의 단어가 있다면 각 단어들을 n차원의 공간의 벡터들이 되게 만드는 방식이다. 하지만, One-hot Encoding 방식은 단어와 단어 사이의 관계를 전혀 설명하지 못한다는 단점이 있었는데, 그러한 이유로 One-hot Encoding이 인간의 언어 형태를 온전히 수치화하였다고는 단언할 수 없었고 그 사용 범위도 협소한 범위로 한정될 수밖에 없었다. 이러한 문제점을 해결한 단어 임베딩 방식이 바로 워

드투벡이다.

워드투벡은 Mikolov et al.(2013)가 제안한 모델로 NN(Neural Net)을 전신으로 탄생하였는데, Mikolov et al.(2013)는 기존의 NN 기반의 학습 방법들과는 차별화된 두 가지 모델을 제시하였다. 첫 번째 모델은 CBOW (Continuous Bag of Words)이다. CBOW 모델은 단어를 선택한 뒤 선택된 단어의 주변 단어를 토대로 선택되었던 단어를 예측해내는 방식이다. '나는 ___에 간다.' 라는 문장에서 주변 단어들을 보고 빈 칸에 들어갈 중심단어를 예측하는 방식을 예로 들 수가 있다. 두 번째 모델은 Skip-gram이다. Skip-gram은 CBOW의 접근 방식과 반대인 방식을 사용하는데, 즉, Skip-gram은 선택된 단어를 통해 주변 단어를 예측해낸다. '___ 외나무다리에서 ___.'라는 문장에서 중심단어를 토대로 주변의 단어들을 예측하는 방식을 예로 들 수가 있다. 이러한 방법들을 사용하면 각 단어들은 컴퓨터가 이해할 수 있는 숫자 형태로 임베딩이 되고, 단어 벡터들 사이의 유사도 계산을 통한 의미론적 해석이 가능해진다. 결과적으로, 단어와 단어 사이의 관계를 설명할 수 있게 된다. Ghannay et al.(2016)에 의하면 일반적으로 Skip-gram 방식이 CBOW 방식보다 문법적, 의미론적 성능이 더 뛰어나다고 판단되는데, Ghannay et al. (2016)이 문법적, 의미론적 성능을 평가하기 위해 사용한 방법은 다음과 같다. 먼저 워드투벡 모델의 의미론적 성능을 평가하기 위해서는 'Greece:Athens → Iran:?'의 비례식에서 '?'에 들어가게 될 단어를 예측하는 방식이 사용된다. 즉, 'Iran'이라는 단어에 대하여 'Greece'와 'Athens' 사이의 관계와 비슷한 단어를 도출해 보고 해당 단어가 정확히 도출되었는지를 확인하는 방식인데, 이 경우 '?'에 들어가야 할 정확한 정답은 'Iran'의 수도인 'Tehran'이 된다. 문법적 성능에 대한 문제도 이와 비슷하다. 'amazing:amazingly → calm:?'의 비례식에서 '?'에 들어갈 단어를 예측하는 방식인데, 이 경우 정답은 'calm'의 부사형인 'calmly'가 된다.

Young et al.(2017)에 의하면, 워드투벡의 한계는 워드투벡의 알고리즘이 전후로 살펴보게 하는 단어의 수(window) 내에서만 학습한다는 점에서 비롯되는데, 감성적으로 상반된 극성을 가지는 good과 bad와 같은 단어들이 비슷한 단어로 임베딩될 수 있다는 것이다(Socher et al., 2011). 이러한 이유로 워드투벡은 감정 분석과 같은 작업에서 문제가 발생할 수 있다고 평가되었다(Wang et al., 2015). 이러한 문제점을 해결하기 위해 Tang et al. (2014)는 임베딩 학습 동안 손실 함수에 직접 감정 극성을 포함시키는 방법론인 Sentiment specific word embedding(SSWE)을 제시하기도 하였다.

LDA(Latent Dirichlet Allocation Topic Modeling) 토픽모델링은 다수의 구조화되지 않은 문서에서 각 문서의 주제와 관련된 정보들을 토대로 문서를 분류하고 분류별로 공통적으로 내재되어 있는 주제들을 추출하기 위한 확률론적인 모델링 방법이다(Blei et al., 2003). LDA 토픽모델링은 문서의 저자들이 주제를 정하고 그 주제를 바탕으로 문장을 구성한다는 것을 전제로 하여 주제에서 파생된 단어들이 이루는 확률 분포를 토대로 해당 문서의 주제를 추정한다. LDA 토픽모델링은 효과적으로 문서들의 키워드를 추출하고 분류할 수 있어 국내외 다양한 분야에서 활용되었다. 국외의 경

우 Xianghua et al.(2013)은 중국 내의 소셜 리뷰들을 분석하기 위한 방법으로 LDA 토픽모델링을 활용하였으며, Grimmer(2010)는 미국의 상원의원들이 언급하는 정책들 중 우선순위가 높은 것이 무엇인지를 효과적으로 분석하기 위해 LDA 토픽 모델링을 활용하였다. 국내의 경우에는 채승훈 등(2015)이 소셜 커머스와 온라인 오픈마켓의 이용경험을 비교 분석하기 위해 LDA 토픽 모델링을 활용하였고, 조혁준 등(2017)은 토픽모델링과 토픽트렌드를 활용하여 브랜드 이미지의 변화에 대해 연구를 수행하였다. 본 연구에서는 사용자들이 연구자 추천시스템을 더 효과적으로 이용할 수 있도록 LDA 토픽모델링 기법을 활용해 다양한 분야의 키워드를 도출해보았고, 추천 시스템 내에 분야 별로 키워드를 제안해주는 기능을 구현하였다.

2.4. 비즈니스 메신저

최근 회사 구성원들간의 커뮤니케이션 향상뿐만 아니라 업무 자동화까지 지원되는 비즈니스 메신저가 UI의 새로운 대안으로 떠오르면서 많은 분야에서 활용되고 있다. 이 중 슬랙(Slack)은 기업 내 협업을 위해 고안된 기업용 비즈니스 메신저이다. 슬랙은 사용자들이 쉽고 간편하게 사용할 수 있는 깔끔한 UI를 제공한다는 특징을 가지고 있는데, 슬랙은 해당 집단 내의 모든 구성원들이 어떠한 대화 내용이던 손쉽게 공유할 수 있도록 돕기 때문에 빠른 커뮤니케이션이 필요한 집단에서 효율적으로 사용될 수 있다고 평가된다. 슬랙의 가장 큰 장점은 슬랙에서 제공하는 API(Application Programming Interface)를 활용해 사용자 맞춤형 메신저를 구축할 수 있다는 것이다. 실제로 챗봇(chatbot) 기능을 직접 구현할 수 있고, 화상통화 서비스인 구글 플러스 행아웃(hangouts), 개발자 커뮤니티인 깃허브(github)와 같은 어플리케이션과 연동을 지원하는 등 다양한 어플리케이션을 직접 제작 또는 연동할 수 있어 엄청난 확장성을 지니고 있다고 평가된다.

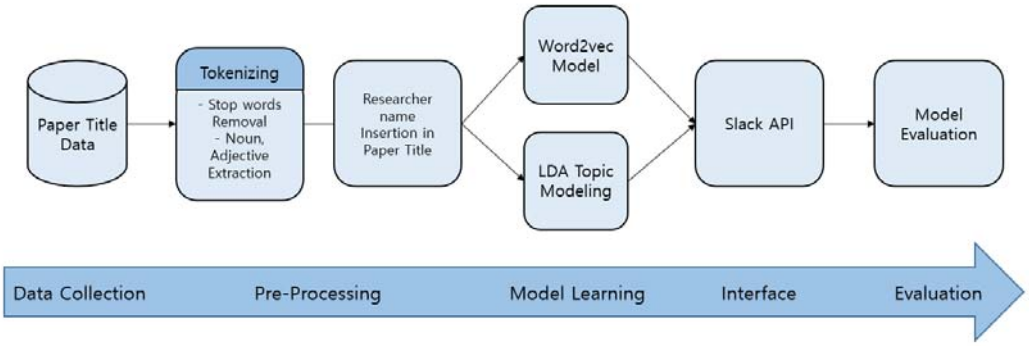
본 연구에서는 연구자들이 효율적이고 간편한 방식으로 추천 시스템을 사용하고 나아가 협업 연구자와 즉각적인 정보 교류를 할 수 있도록 슬랙 API를 활용하여 추천시스템을 구현하였는데, 연구 키워드를 입력하고 연구자를 추천받는 등의 모든 행위를 슬랙 내에서 작업할 수 있도록 설계하였다.

Ⅲ. 연구 방법

제 3장에서는 연구가 수행된 구체적인 과정에 대해 소개한다. 데이터 수집, 데이터 전처리, 워드투벡 및 LDA 모델 학습, 슬랙과 연동, 구현된 모델 평가로 구분하여 열거한 순서대로 서술할 것이며, 이는 <그림 1>과 같다.

3.1 데이터의 수집

본 연구에서는 연구자 협업 플랫폼을 개발하기 위해 워드투벡 알고리즘을 활용하였는데, 워드투벡의 학습을 위해서는 하버드 대학교의 교수들을 중심으로 하여 선택된 연구자들이 투고한 논문 제목 데이터를 활용하였다. 등록된 교수의 수가 다른 분야들에 비해 상대적으로 많



<그림 1> 워드투벡을 활용한 연구자 추천 시스템의 연구 모델

은 의학 분야를 제외한 나머지 분야들의 교수 리스트를 수집하였고, 그 중 구글 스칼라(scholar.google.com)에 프로필이 명시되어 있는 총 723명의 교수에 대해 54771개의 논문 제목을 크롤링(crawling)하였다. 이 때 논문 제목들은 동적 크롤링 방식으로 크롤링되었는데, 이를 위해서는 파이썬의 모듈인 셀레니엄(selenium)을 활용하였다.

3.2 데이터 전처리

워드투벡 알고리즘을 학습하기 위해 앞서 크롤링한 데이터를 워드투벡 알고리즘의 데이터 입력 형태에 알맞게 전처리할 필요가 있다. 워드투벡은 2장의 관련 연구에서 언급한 내용과 같이 단어 임베딩 알고리즘의 일종으로 단어들을 벡터 형태로 변환하고 관련이 깊은 단어들은 거리를 가깝게, 관련이 없는 단어들은 거리를 멀게 배치하는 알고리즘이다. 따라서 우선 각 논문 제목을 단어 수준으로 분할해 워드투벡이 각 단어를 하나의 점으로써 공간상에 배치할 수 있게 해야 한다. 단어 수준의 토큰화를 위해서는 파이썬의 NLTK 모듈을 활용하였다.

분할된 문장들은 이제 단어수준에서 의미를 가지게 되는데, 일반적으로 논문 제목들에는 특수문자, 관사, 전치사, 부사 등(e.g., “in”, “the”, and “a”)의 무의미한 단어들이 포함되어 있으므로 이를 제거해야 한다(Mikolov et al., 2013). 예를 들어, “Methodological fit in management field research” 라는 논문 제목에서 “in” 은 불필요한 단어이므로 제거해야 한다. 또, 특수문자, 관사, 전치사, 부사 외에도 문장이 담고 있는 의미에 크게 영향을 끼치지 않는 단어 역시 제거해야 하는데, 예를 들어, “research” 라는 단어는 많은 논문에서 맥락에 큰 영향 없이 무의미하게 쓰일 수 있는 단어이므로 이 역시 제거한다. 이렇게 문장의 핵심 맥락에 큰 영향을 끼치지 않아 제거할 필요가 있는 단어들을 불용어라고 한다. 본 연구에서는 연구에 활용한 논문 제목들에 맞춤화 된 불용어 사전을 직접 제작하였는데, 총 불용어의 수는 1154개로 논문 제목에서 추출한 단어들을 직접 확인하여 “effect”, “benefit”, “method” 와 같이 대부분의 논문에서 사용될 수 있어 논문 제목의 핵심 맥락에서 벗어난다고 판단되는 단어들을 불용어 사전에 추가하였으며, 불용어 사전에 수록된 단

	0	1	2	3	...	296	297	298	299
economy	0.101404	-0.492307	0.078771	0.423236	...	0.366169	-0.434861	-0.525001	-1.142190
anesthesia	0.054705	0.052901	0.091266	-0.527054	...	0.237396	-0.193276	-0.144751	0.084401
statistics	-0.497513	-1.058034	-1.345432	-0.177121	...	-0.597439	0.644290	0.141737	0.952999
perovskite	-0.196312	-0.040248	0.226870	-0.202999	...	0.196432	-0.100108	0.015573	0.376689

4 rows × 300 columns

<그림 2> 워드투벡의 학습 결과 예시

어들을 토큰화한 논문 제목 목록에서 제거해주었다.

연구자 추천 시스템의 기본 골격은 특정 키워드를 입력했을 때 해당 키워드와 관련성이 높은 연구자 목록을 출력하는 것이다. 이를 구현하기 위해서는 단어수준으로 토큰화된 논문 제목들이 각각의 저자와 매핑이 되어야 한다. 따라서 본 연구에서는 각 논문 제목들 마다 저자의 이름을 삽입하는 방식으로 워드투벡 알고리즘이 해당 논문의 각 단어들과 저자의 이름을 관련이 깊은 단어 관계로 이해할 수 있도록 설계하였다. 앞서 설명하였듯이 워드투벡 알고리즘은 특정 단어의 전후로 살펴보게 하는 단어 수(window) 안에 속하는 단어들에 대해서만 해당 단어와 관련이 높은 단어로 학습한다. 즉, 논문 제목 중간에 저자의 이름을 삽입하게 되면, 저자의 이름이 논문 제목에 수록된 단어들의 window 크기 안에 속하게 되고, 결과적으로, 논문 제목에 수록된 단어들과 저자 이름은 문법적으로 또, 의미상으로 연관성이 높은 단어로 학습된다.

3.3 워드투벡 모델 학습

앞서 3.2장에서는 하버드 대학교의 교수들을 주축으로 선택된 연구자들이 저술한 총 54771개의 논문 제목들을 전처리하였다. 본 연구에서는 이렇게 전처리한 데이터를 바탕으로 워드투벡 알고리즘을 학습시켰고, 이 때 학습 및 분석 도구로는 파이썬의 모듈인 gensim 패키지를 활용하였다. 학습에 있어서 각 단어를 300차원 공간의 벡터로 바꾸어 주었고, 각 단어에 대해 주변 단어는 전후로 6개까지 보게 하였다. 또, 학습 방식의 경우, Skip-Gram 방식이 문법적, 의미론적 해석의 성능에서 CBOW 방식보다 더 좋은 성능을 내는 것으로 판단되므로 Skip-Gram 방식을 채택하였다. 각 단어들을 학습한 결과물은 <그림 2>와 유사한 형태이다.

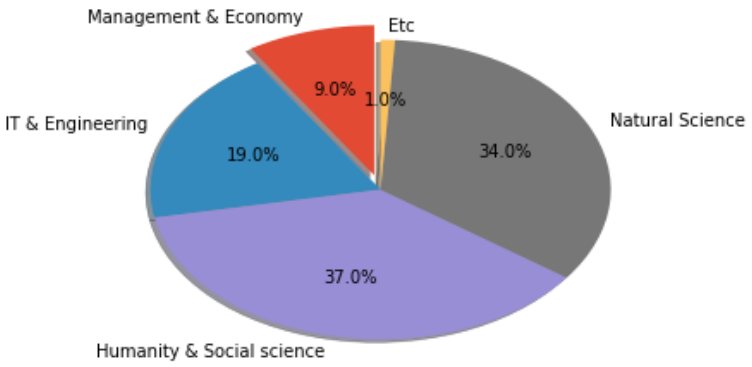
3.4 LDA 토픽 모델링 학습

타 분야 연구자와 융합연구를 수행해본 경험이 없는 대부분의 연구자들은 본인의 전문 분야 외에 다른 분야에 대해서는 문외한일 가능성이 있다. 그러한 이유로 융합 연구를 원하는 연구자들이 연구자 추천시스템을 이용할 때 검색하고자 하는 연구 분야에 대한 지식이 부족하여 검색을 제대로 수행하는데 어려움을 겪을

수 있을 것이다. 따라서 본 연구에서는 앞서 크롤링한 데이터를 토대로 LDA 토픽 모델링을 실시하였고, LDA 토픽 모델링을 통해 다양한 분야에서 자주 쓰이는 연구 키워드들을 도출함으로써 추후에 연구자 추천 시스템을 이용할 연구자들에게 편의성을 제공하려 하였다.

LDA 토픽모델링 알고리즘을 학습하는 과정은 다음과 같다. 먼저 분야별 키워드를 파악하고 LDA 토픽모델링의 성능을 향상시키기 위해 모든 연구자들을 분야별로 분류한다. 본 논문에서는 경영 및 경제, IT 및 공학, 인문 및 사회, 자연과학 총 4개의 분야로 연구자들을 분류하였다. 분류된 결과는 <그림 3>과 같다. 그 다음 모든 논문 제목 데이터들을 각 논문 제목을 한 단위로 하여 bag-of-words 형태로 변환한다. 본 연구에서는 논문 데이터들을 bag-of-words 형태로 변환하기 위해 카운터 벡터(counter vector), tf-idf 의 방식을 사용해보았는데, tf-idf 방식의 성능이 뛰어나다고 판단되어 tf-idf를 활용하여 논문 제목들을 bag-of-words 형태로 변환하였다. 데이터는 bag-of-words 형태로 변환되면서 단어의 출현 빈도를 나타내는 숫자의 형태로 표현된다. 따라서 bag-of-words로 변환된 데이터를 LDA 토픽 모델링 알고리즘에 입력 값으로 넣어 데이터에 존재하는 주제를 추출할 수 있게 된다. 본 연구에서는 앞서 전처리한 데이터를 파이썬 모듈인 Sklearn에 포함된 LDA 토픽모델링 알고리즘의 입력 값으로 넣어주었고, 각 분야별로 총 20개의 토픽으로 분류하였다. 아래 <Table1>은 LDA를 통해 도출된 분야별 주제와 주제별 키워드들의 일부를 담고 있다.

<표 1>에서 두 번째 열은 LDA 토픽모델링 알고리즘이 분류한 주제들인데, LDA 토픽모델링이 수행됨으로써 도출된 결과물에는 각 주제에 대해 라벨링이 되어있지 않다. 따라서 LDA 토픽모델링의 사용자가 도출된 키워드들을 직접 확인하고 그 키워드들에 알맞은 라벨을 붙여주어야 한다. 예를 들어, ‘Management& Economy’ 분야의 첫 번째 주제의 경우 ‘business’, ‘policy’, ‘strategic’, ‘adoption’, ‘differentiation’ 등의 단어들을 키워드로 가지



<그림 3> 분야별 연구자 비율

<표 1> LDA 토픽모델링 결과

연구분야	주제	주제별 키워드
Management & Economy	Business Strategy (Topic #1)	business, policy, product, learning, case, strategic, team, knowledge, internet, wage, entrepreneurship, asset, political, adoption, differentiation, politics, moral, statistic, valuation, job ...
	Heath Economics (Topic #2)	health, system, care, economics, price, patient, alcohol, medical, competition, selection, consumption, survey, taxation, income, production, woman, accounting, veteran, seasonal, trading, childhood ...
	Organization (Topic #3)	organizational, economic, behavior, data, stock, return, economy, open, finance, relationship, competitive, ceo, versus, gender, environmental, welfare, identity, community, corruption ...
	Venture Business (Topic #4)	capital, venture, cost, equity, customer, productivity, future, clinical, institutional, statistical, processing, systematic, critical, pricing, structural, saving, domain, specialization ...
	Enterprise (Topic #5)	structure, corporate, new, investment, investor, interest, fund, negotiation, trade, entrepreneurial, governance, world, right, debt, corporation, channel, earnings, dividend, aid ...
	...	...
IT & Engineering	Computer Science (Topic #1)	network, algorithm, interface, parallel, neural, communication, visualization, computational, direct, iterative, carbon, electronic, artificial, structured, circuit ...
	Mechatronics (Topic #2)	soft, crystal, silicon, photonic, mechanical, stress, dielectric, elastomer, deformation, digital, actuator, laser, protocol, ion, fracture, wireless, programmable, robotics, electrode ...
	Biochemical (Topic #3)	biomaterial, combustion, biology, calcium, cation, care, amorphous, conductive, coleoptera, catalyst, cerebellar, conductivity, cascade, biogeochemical, cell ...
	...	...
Humanity & Social science	Education (Topic #1)	school, urban, future, adolescent, teacher, knowledge, community, educational, lesson, measure, youth, potential, behavioral, mathematics, communication, psychology ...
	Politics (Topic #2)	policy, system, law, economic, management, trade, foreign, value, leadership, competition, election, cost, price, difference, success, search, general, idea, versus, preference, emotional, strategic, adaptive, optimal ...
	Cultural Problem (Topic #3)	social, child, health, race, united, care, racial, poverty, war, moral, trust, black, poor, origin, civic, civil, india, african, sexual, freedom, home ...
	Immigrant Problem (Topic #4)	america, labor, protein, latin, resistance, immigrant, adult, party, investment, immigration, injury, infection, violence, multi-drug, welfare, brazil, signal, regime, crime, news, employment, brasil ...
	Globalization (Topic #5)	globalization, network, american, world, behavior, regulation, culture, corporate, china, role, identity, visual, discrimination, private, game, recognition, chinese, organization ...
	...	...
Natural Science	Biology (Topic #1)	cell, molecular, mutation, natural, zebrafish, phylogeny, cancer, cognitive, adaptation, allergy, child, behavioral, morphology, family, food, versus, milky, mitochondrial, germ, membrane, lens ...
	Physics (Topic #2)	quantum, optical, graphene, visual, plasticity, knowledge, value, phylogenetic, spectrum, fluid, hall, glacial, dopamine, vitro, silicon, mechanic, deformation ...
	Earth Science & Astronomy(Topic #3)	global, supernova, signal, plant, stress, atmospheric, solid, earth, animal, isotope, elastic, ozone, religion, polarization, north, satellite, tropospheric, reionization, signaling, lipid, flux, amazon ...
	Chemistry (Topic #4)	structure, particle, temperature, genome, energy, genomic, organic, emission, climate, collision, cascade, phospholipase, aerosol, spectroscopy, order, nanoscale ...
	...	...

고 있으므로 ‘Business Strategy’ 라는 라벨을 붙여줄 수 있다. 각 주제들과 도출된 키워드들을 살펴보면, 먼저 ‘Management & Economy’ 분야의 경우, ‘Business Strategy’, ‘Health Economics’, ‘Organization’ 등 상당히 세부적인 주제들이 도출되었다. 이는 'Management & Economy' 분야가 애초에 세분화된 분야인데 그에 대해 추가적인 분류를 시도하게 되니 이러한 결과가 도출된 것으로 보인다. ‘Humanity & Social Science’ 분야의 경우, ‘Education’, ‘Politics’, ‘Globalization’ 등 다양한 주제들이 도출되었다. 학습에 사용된 데이터의 양이 많고, 인문사회 분야의 연구 특성상 포괄적이고 다양한 주제를 다루기 때문에 이러한 결과가 나타난 것으로 여겨진다. ‘Natural Science’ 분야의 경우, ‘Biology’, ‘Physics’, ‘Earth Science & Astronomy’ 와 같은 포괄적인 주제들이 도출되었다. ‘IT & Engineering’ 분야의 경우 ‘Natural Science’ 분야와 비슷하게 ‘Computer Science’, ‘Material’, ‘Mechatronics’ 와 같은 포괄적인 주제들이 도출되었다. 본 연구에서는 학습에 이용한 데이터의 크기가 작아 수행할 수는 없었지만, ‘Natural Science’와 ‘IT & Engineering’ 분야의 경우 세부 분야별로 데이터를 더 상세히 분류하여 LDA 토픽모델링을 실시한다면 더 명확한 결과물을 얻을 수 있을 것이라 판단된다.

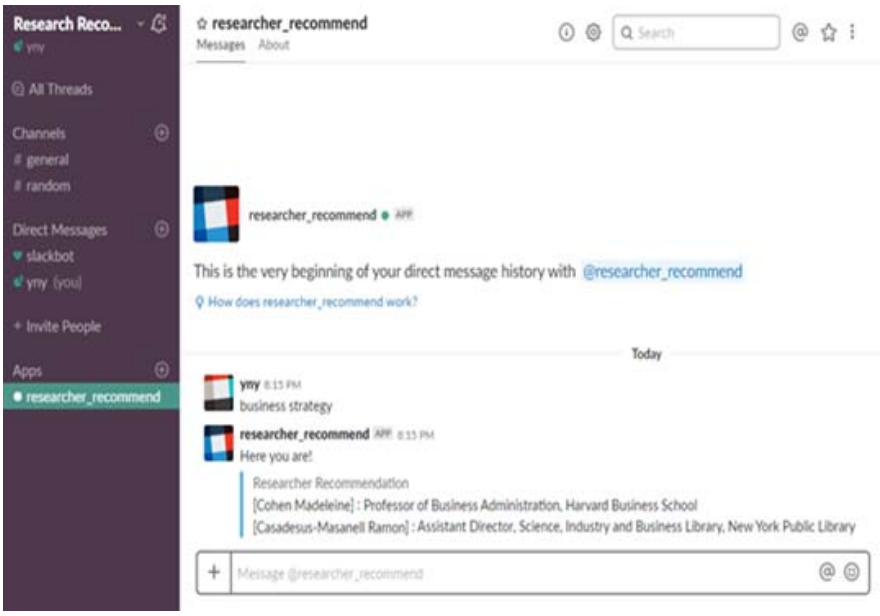
3.5 비즈니스 메신저 연동

추천시스템을 기본 실행 파일로만 이용가능하게 배포한다면 불친절한 사용 환경과 어플리케이션의 확장성 부족 등의 문제로 사용자들에게 외면당할 것이다. 따라서 본 연구에서는 유려한 디자인을 갖춘 메신저 형태의 UI에서 사용자들이 쉽고 간편하게 연구자 추천 시스템을 이용할 수 있도록 슬랙 API를 활용해 챗봇 형태의 어플리케이션을 제작하였다.

슬랙과 연동하여 연구자 추천 시스템을 사용하는 과정은 다음과 같다. 사용자는 연구자 추천 시스템이 구현되어 있는 Workspace에 접근한다. Workspace는 일반적인 메신저 어플리케이션들의 채팅방과 유사하다. 사용자는 Workspace에서 챗봇 어플리케이션에 접근해 메시지를 보낼 수 있는데, 어플리케이션에서 보낸 메시지는 슬랙 API를 활용하고 있는 서버로 전달되고 해당 서버에서 구동 중인 연구자 추천 시스템이 작동한다. 마지막으로, 연구자 추천시스템의 결과물은 다시 메시지의 형태로 사용자들에게 전달된다. 사용자들은 슬랙을 활용함으로써 키워드 검색 및 연구자 검색 기능 모두를 슬랙 메신저 내에서 이용할 수 있으며, 검색된 연구자와 화상통화 또는 화면 공유를 통해 정보를 빠르게 주고받을 수도 있다. 또한 추가적인 기능이 필요하다면 필요한 기능을 직접 구현하거나 타 어플리케이션과 연동해 슬랙 내에서 즉시 사용할 수도 있다. 슬랙을 활용해 연구자 추천을 받는 모습은 <그림 4>와 같다.

Ⅳ. 연구자 추천 시스템 구성

본 논문에서 제안하는 연구자 추천 시스템은 인터페이스, 자연어 처리 시스템, 학습을 완료한 워드투벡 모델, 정규표현식 추출 시스템으로 구성된다. <그림 5>는 연구자 추천 시스템의



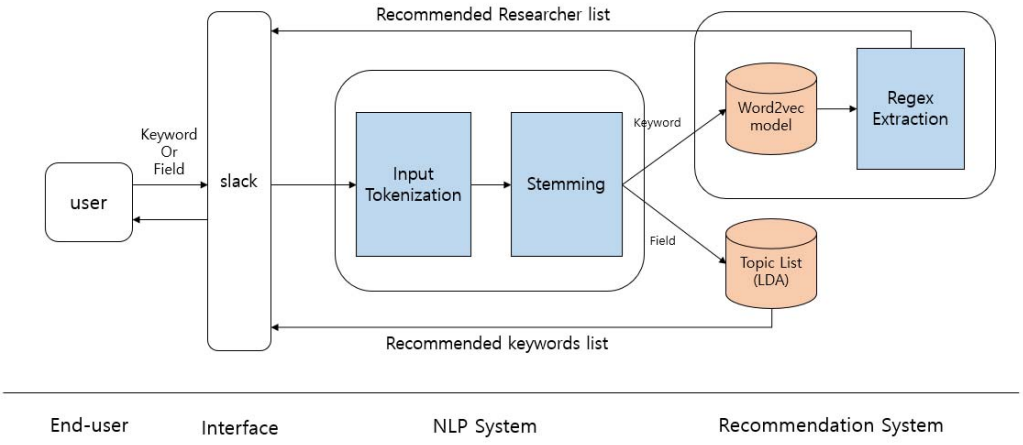
<그림 4> 연구자 추천시스템 프로토타입

구조를 보여준다. 그림에서 인터페이스는 사용자로부터 검색에 필요한 키워드를 받아 이를 자연어 처리 시스템에 전달하고, 연구자 추천 시스템이 모두 동작한 후 추천 결과를 사용자에게 보여주는 역할을 수행한다.

자연어 처리 시스템은 토큰화(tokenization)와 스테밍(stemming) 두 가지 프로세스로 이루어진다. 사용자가 두 단어 이상의 단어 집합을 키워드로 입력하는 경우, 워드투벡 모델이 단어 간 유사도를 도출할 수 있도록 토큰화 과정을 통해 키워드를 단어 수준으로 분할해야 한다. 이 때, 토큰화된 단어는 같은 의미라도 다양한 형태를 지닐 수 있기 때문에 이를 하나의 형태로 통일해야 할 필요가 발생한다. 예를 들어, “argue”, “argues”, ”argued” 세 단어는 같은 의미를 지니고 있음에도 다양한 형태를 지니고 있고, 이들 단어를 하나의 형태로 통일해야만 한다. 이렇게 어형이 변형된 단어에서 접사 등을 제거하고 그 단어의 어간만을 도출하는 것을 스테밍(어간 추출) 이라고 한다.

토큰화와 스테밍 과정을 거쳐 자연어 처리가 완료된 키워드는 학습을 완료한 워드투벡 모델에 입력되어 유사 단어 및 추천 연구자를 도출해낸다. 이 때 연구자 추천을 위해서 gensim 모듈의 most_similar 함수를 활용하는데, most_similar 함수는 인수로 할당된 단어 벡터와 코사인 유사도가 높은 단어들의 목록을 뽑아낸다. 코사인 유사도를 계산하는 식은 <그림 6>과 같다.

본 연구의 모델에서는 연구자들의 이름을 논문 제목 사이에 삽입한 뒤 워드투벡을 학습했기 때문에 연구자들의 이름도 단어 벡터로 워드투벡이 도출한 공간상에 위치한다. 따라서 most_similar 함수는 인수로 할당된 단어 벡터와 코사인 유사도가 높은 단어 벡터들과 연구자 이름 벡터들을 리스트의 형태로 도출해 낼



<그림 5> 연구자 추천 시스템 구성

$$Similarity(A,B) = \cos(\Theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

<그림 6> 코사인 유사도의 계산 식

수 있다. 하지만, 본 시스템에서 필요한 것은 유사도가 높은 단어들이 아닌 유사도가 높은 연구자 이름들뿐이다. 따라서 본 연구에서는 정규표현식을 활용해 연구자 이름만을 선택적으로 도출해냈다. 이때, 정규표현식을 사용하기 위해서는 파이썬 모듈인 re를 활용했다. 마지막으로, 추천된 연구자 목록은 슬랙 인터페이스의 챗봇 어플리케이션을 거쳐 메시지의 형태로 사용자들에게 전달된다.

Ⅴ. 연구 모델 평가

본 장에서는 본 연구에서 구현한 연구자 추천시스템이 제대로 동작하는지를 검증하기 위한 실험을 진행하였다. 검증 방법으로는 총 두 가지의 평가 방식을 활용했는데, 첫 번째 평가 방식의 프로세스는 다음과 같다. 먼저 각 분야의 전문가들을 초빙해 총 50개의 분야별 연구 키워드 리스트를 수집하였다. 분야는 “경영 및 경제”, “정보통신 및 공학”, “인문 및 사회”, “의/약학”, “자연과학” 총 다섯 분야로 나누었으며, 각 분야별로 10개의 키워드를 수집하였다. 수집된 키워드들은 본 연구자 추천시스템에 입력되었으며, 추천시스템으로부터 추천된 연구자들을 구글 스칼라에 검색한 뒤 분야별 전문가들에게 추천된 연구자의 실제 연구 분야와 키워드 사이의 관련성을 평가하도록 하였다. 특히 연구 분야들은 동의어 관계에 있거나 또는 계층 구조가 있는 경우가 상당하기 때문에 추

<표 2> 키워드 별 연구자 추천 정확도 예시

연구 분야	키워드 예제	추천수	맞는 추천수	정확도
Management & Economy	organizational behavior, finance ...	103	90	87.37%
IT & Engineering	ai, semiconductor ...	49	40	81.63%
Humanity & Social science	cultural psychology, labor ...	102	90	88.23%
Natural Science	organic chemistry, neuroscience ...	70	58	82.85%
Medical Science & Pharmacy	bone fracture, cancer ...	80	69	86.25%
	total	404	347	85.89%

천 시스템의 성능 평가는 반드시 분야별 전문가들에게 맡기는 것이 바람직하다고 판단된다.

실제 연구 분야와 키워드 간의 상관관계가 명확하다고 판단되는 경우는 다음과 같은데, “구글 스칼라 프로필에 등록된 연구 분야와 키워드가 일치 또는 동의어 관계에 있거나 서로 포함관계에 있다고 판단되는 경우”와 “연구자가 작성한 논문 제목에 키워드가 속해 있지 않더라도 연구자가 해당 키워드에 대한 연구를 수행했다고 판단되는 경우” 이다. 예를 들어, 검색에 사용된 키워드가 “machine learning”이고 추천된 연구자가 “자연어처리” 분야의 전문가로 확인된다면, 검색 키워드와 추천된 연구자의 연구 분야가 서로 포함 관계에 있기 때문에 해당 연구자는 올바르게 추천된 것으로 판단된다.

평가 결과, 50개의 키워드에 대해 총 404명의 연구자들이 검색되었고 그 중 347명의 연구자들이 키워드와 관련성이 높다고 판단되었다. 즉, 50개의 키워드 표본에 대해 총 85.89%의 적중률을 보여주었다. 키워드들이 연구자들의 구글 스칼라 프로필의 연구 분야 목록에 수록이 되어 있어 대부분의 키워드가 연구자들과 관련성이 있다고 판단되었는데, 키워드와 추천된 연구자의 연구 분야가 일치하지 않는 경우도 있었다. 전혀 연관성이 없는 경우도 간혹 있었으나, 완벽히 일치하지는 않았지만 비슷하다고 판단되는 연구 분야의 연구자를 추천한 경우가 많았다. “의/약학 분야”의 연구 분야인 “anatomy”를 키워드로 입력한 경우 자연과학 분야인 “biology”의 연구자가 추천되는 것을 예로 들 수가 있다. 추천 정확도를 분야별로 통계낸 결과와 평가에 사용된 키워드의 일부는 <표 2>에서 확인할 수 있다.

모델에 대한 두 번째 평가 방식의 프로세스는 다음과 같다. 먼저 워드투벡 모델 학습에 사용된 723명의 하버드 대학교 교수들 중 50명을 무작위로 선택했다. 그 뒤, 교수들의 이름을 워드투벡 모델의 most_similar 함수에 인수로 할당해 결과 값을 뽑아내었다. 앞서 설명했듯이 본 연구의 모델에서는 연구자들의 이름을 논문 제목 사이에 삽입한 뒤 워드투벡을 학습했기

<표 3> 연구자와 연관성이 높은 단어들 및 실제 연구 분야

연구자 성명	연구 분야	키워드 예제
Viceira Luis	Finance, Asset Pricing, Capital Markets, Investment	asset, macro-market, investor, finance, investment, econometrics
Vlassak Joost	Mechanics, Materials	micro-mechanics, battery, plastic, thermomechanical, anisotropic, material
Gersen Jeannie	Crimnal Law, Family Law, Copyright, Sexual Assault&Harassment	discriminability, cyberlaw, violent, copyright, homosexuality, harassment, assault
Zhou Qiao	Regenative Medicine, Neurodegeneration, Stem Cell Biology	stemcell, neuroendocrine, biokinetics, neuroinflammation, antibody, medicine
Rebbeck Timothy	cancer, genetics	cancer, neurogenetics, geneenvironment, pharmacogenetics, mastectomy

때문에 연구자들의 이름도 단어 벡터로 워드투벡이 도출한 공간상에 위치한다. 따라서 교수들의 이름을 단어 벡터로써 most_similar 함수의 인수로 사용할 수 있다. 마지막으로, 단어 목록과 해당 교수의 실제 연구 분야를 비교해 관련성이 깊다고 판단되는 지를 각 분야의 전문가들이 확인하게 하였다. 분석에 사용된 교수 목록과 코사인 유사도가 높은 단어 목록, 그리고 실제 연구 분야의 예시는 <표 3>과 같다.

분석 결과, 도출된 키워드들과 연구자의 실제 연구 분야가 대부분 일치하는 것으로 확인되었다. 도출된 키워드들이 완벽하게 일치하는 경우가 많았고, 완벽하게 일치하지는 않더라도 유사 단어 또는 유사 연구 분야의 단어들을 뽑아내는 경우도 빈번했다.

위 분석에서 워드투벡의 좋은 성능을 확인할 수 있는데, 특정 연구 분야를 나타내는 단어가 해당 연구자가 작성한 논문 제목들에 포함되어 있지 않더라도 해당 연구 분야가 유사한 단어로 도출되는 결과를 보여주었다. 즉, 이 부분에서 워드투벡은 특정 키워드가 논문 제목에 포함되어 있지 않더라도 논문 제목에 속한 단어들과 비슷한 단어로 판단된다면 해당 키워드를 논문을 저술한 연구자와 관련이 깊은 단어로 학습한다는 것을 알 수 있다.

Ⅵ. 결론 및 시사점

6.1 결과 및 토의

본 연구에서는 융합연구를 지원하기 위하여, 특정 연구 키워드를 입력하면 관련성이 높은 연구자리스트를 제공하는 연구자 추천 시스템을 구현하였다. 또한 사용자들에게 검색의 편의를 제공하기 위하여 분야별로 주요 연구 키워드를 제안하는 기능을 추가 구현하였다. 이를 위해 하버드 대학교의 교수들을 중심으로 수집한 연구자 목록을 구글 스칼라에 검색하여 각 연구자들이 집필한 논문들의 제목을 수집하였다. 수집된 데이터는 토큰화하였고 토큰화한 논문 제목에 논문 저자의 이름을 삽입하는 방식의 전처리 과정을 거친 뒤 그 결과물로 워드투벡 알고리즘을 학습시켜 연구자 추천 시스템을

완성하였다. 키워드 제안 기능을 구현하기 위해서는 LDA 토픽모델링을 활용해 전처리된 데이터에서 각 분야별 키워드들을 도출해내었고 키워드들에 알맞은 주제들을 라벨링하였다. 마지막으로, 연구자들이 편리하게 추천 시스템을 사용하고 나아가 협업 연구자와 즉각적인 정보교류를 할 수 있도록 슬랙 API를 활용하여 추천시스템을 구현하였다.

구현된 연구자 추천 시스템의 정확도는 두 가지 방법으로 검증하였다. 첫 번째 방법으로, 연구자 추천 시스템에 다양한 분야의 키워드 50개를 연구자 추천 시스템의 입력한 후 그 결과로 추천된 연구자 목록이 타당한지 전문가집단을 통해 검증하였다. 전문가 검증 결과 추천 콘텐츠의 85.89%가 타당한 결과로 판정 되었다.

두 번째 방법으로, 무작위로 선정한 50명의 교수 이름과 코사인 유사도가 높다고 판단되는 단어들을 도출한 뒤 실제 연구 분야와 일치하는지를 확인해보았다. 검증 결과 연구자와 관련성이 높은 단어들은 실제 해당 연구 분야의 단어와 일치하였고, 완벽하게 일치하지는 않더라도 유사 단어 또는 유사 연구 분야의 단어들을 도출해냈다.

6.2 시사점 및 연구의 한계

본 연구의 시사점은 다음과 같다. 첫째, 융합연구를 계획하고 수행하려는 연구자들이 쉽게 타 분야 연구 인력을 모집할 발판을 제공했다. 기존 연구 환경에서는 연구자들이 타 분야 연구자들과 교류하기 어렵고, 상용 연구자 추천 서비스를 사용하기에도 금전적으로 큰 부담이 된다. 워드투벡 알고리즘은 다양한 프로그램언어로 패키지가 공개되어 있어 보다 쉽게 연구자 추천시스템을 구현할 수 있다는 장점이 있다. 또한 워드투벡 알고리즘을 활용해 연구자 추천 시스템을 구현할 경우 논문 제목, 초록 등의 연구내용과 관련된 비정형 데이터를 그대로 활용할 수 있기 때문에 보다 세부 연구 분야에 대한 연구자 추천이 가능하다는 장점이 있다. 둘째, 워드투벡 알고리즘을 기존의 활용 방식에서 벗어나 새로운 주제에 적용할 수 있게 하는 방법론을 제시함으로써 그 활용 범위를 확대하고 추후의 다른 연구 과제에서 참고할 수 있도록 하는 발판을 제공했다. 기존 연구에서 워드투벡 알고리즘은 문장 및 문서 분류 등의 한정된 주제에서 사용되었다. 본 연구에서는 이런 기존의 틀에 박힌 활용 방법에서 벗어나 워드투벡을 새로운 분야에 창의적으로 적용하는 탐험적인 방법론을 제시하였다. 셋째, 비즈니스 메신저인 슬랙을 연구자 추천 시스템의 인터페이스로 활용함으로써 시스템의 UI를 손쉽게 개발하는 방법을 제안하였다. 또한, 챗봇, 화상통화 등 협업에 도움이 될 추가적인 어플리케이션을 구현 또는 연동할 수 있음을 보여 확장적인 협업 인프라를 간편하게 구축할 수 있게 하는 가이드라인 역시 제시하였다.

본 연구가 지닌 한계점과 더 효과적인 연구자 추천시스템의 구현을 위한 향후 연구 방향은 다음과 같다. 첫째, 연구 분야에 따라서 동일한 단어가 다른 의미로 사용될 경우 이를 구분할 수 없었다. 예를 들어 기계공학 분야에서 “machine”이라는 단어와 기계학습 분야에서 “machine”이라는 단어는 다른 의미로 사용된다. 하지만 두 분야 모두 동일한 단어가 많이

사용되기 때문에 워드투벡을 활용한 추천시스템의 경우 연구자를 잘못 추천할 가능성이 있다. 이를 해결하기 위해서는 어떤 분야의 논문인지를 미리 파악하여 데이터 전처리 과정에서 분야에 따라 동일한 단어를 구분할 수 있도록 하는 전처리 과정이 필요할 것으로 생각된다.

둘째, 워드투벡 학습 시 두 개 이상의 단어로 구성된 합성어를 하나의 단어로 인식하여 학습하지 못하였다. 예를 들어, “machine learning”과 “organizational behavior”는 두 개 이상의 단어로 이루어져 있음에도 불구하고 해당 학문 분야 내에서 개별 단어처럼 인식된다. 따라서 실제 연구를 수행할 때, 이러한 전문 용어들은 개별 단어로써 토큰화되어야 한다. 하지만 본 연구에서는 공백을 기준으로 토큰화를 시도했기 때문에 위와 같은 단어들을 하나의 단어로 인식하고 연구를 수행하지는 못하였다. 본 연구에서는 이에 대한 대안으로 파이썬의 gensim 모듈에서 제공하는 “두 개 이상의 단어와 코사인 유사도가 높은 단어를 뽑아주는 함수”를 활용해 연구자 추천 시스템을 구현했다. 이 방식은 하나의 단어와 유사도가 높은 단어를 뽑아줄 때와 비슷한 성능을 보여주지만, 예를 들어, “machine learning”의 키워드를 검색하는 경우 “machine”과 “learning”이 서로 관련 없이 독립적으로 출현하는 논문들을 작성한 연구자가 존재할 경우에는 해당 연구자가 추천될 위험이 있다. Mikolov et al.(2013)는 이러한 문제점을 해결하기 위해 동시에 등장하는 단어를 포함하는 구문을 식별하고 이들을 별도로 학습 시키는 방법을 제시하였는데, 만약 향후 연구에서 이와 같이 두 개 이상의 단어로 이루어져 있음에도 하나의 단어처럼 사용되는 단어들의 사전을 분야 별로 직접 제작해 사용함으로써 완벽하게 토큰화한다면, 더 높은 성능을 보여주는 모델을 구현할 수 있을 것으로 보인다.

셋째, 본 연구에서는 워드투벡 학습에 있어 논문 제목에서 추출한 단어들만을 사용하였기 때문에 워드투벡 학습 모델의 성능에 한계가 있다. 추후 연구로서 데이터 수집시 각 논문에 수록된 저자가 명시한 키워드를 추가하여 분석을 수행한다면 좀 더 나은 성능을 가진 모델을 구현할 수 있을 것으로 보인다.

넷째, 기존에 제시되었던 연구자 검색 또는 추천 시스템과의 면밀한 성능 비교 분석이 이루어지지 못하였다. 연구자 협업 서비스에 관한 기존 연구들은 시맨틱 웹과 온톨로지를 활용하여 수행된 것이 대부분이고, 기계학습 기반의 연구자 협업 서비스에 대한 연구는 국내외로 상당히 미미한 실정이다. 따라서, 본 연구에서는 제안하는 연구자 추천 시스템에 대하여 타 시스템과의 성능 비교 분석을 통한 정량적인 우월성을 입증하지는 못하였지만, 연구자 추천 시스템을 기존에 수행되었던 시맨틱 웹과 온톨로지를 활용한 방법이 아니라 딥 러닝 알고리즘을 활용하여 데이터를 기반으로 한 접근 방법으로 구현하였다는 점에서 차별화된 가치를 가질 수 있다고 판단된다.

참고문헌

김진완, 김유일, 홍태호, “전자적 파트너십에서 공급자의 전략적 혜택 창출을 위한 협업의 효과에 관한 연구”, 정보시스템 연구, 제17권 제4호, 2008, pp. 341-367.

박기범, 황정태, “융합 연구의 형성과 발전 과정의 고찰을 통한 국내 연구 현황 분석”, 조사연구, 2007. 11, pp. 1-125.

배형준. *이공계 융합연구 수행의 영향요인 및 연구성과 영향 분석*. 서울대학교 대학원, 2014.

송위진, “‘과학기술과 인문사회 융합연구’의 필요성과 과제”, STEPI Insight, 제47호, 2010, pp. 1-22.

오헌석, 배형준, 김도연, “과학기술분야 융합연구자의 융합연구 입문과 과정에 관한 연구”, 아시아교육연구, 제13권 제4호, 2012, pp. 297-335.

이무순, 손달호, “ICT 기업 융합성과의 결정요인에 관한 연구”, 정보시스템연구, 제26권 제3호, 2017, pp. 1-23.

이광호, 김승현, 최종화, 서지영, 강지훈, 이아정, “융합연구사업의 실태조사와 연구개발 특성분석”, 정책연구, 9월호, 2013, pp. 1-393.

정한민, 성원경, 박동인, “연구자 간 협업 지원 서비스를 위한 지식 베이스 설계”, 제17회 한글 및 한국어 정보처리 학술대회, 2005.10, pp. 173-178.

정한민, 이미경, 성원경, 박동인, “Ontoframe-K: 연구자 간 협업 지원 서비스를 위한 시멘틱웹 기반 정보 유통 플랫폼”, 한국정보과학회 2006 한국컴퓨터종합학술대회 논문집, 2006. 6, pp. 100-102.

조양래, 양이석, 서용윤, 전정환, “국가 융합연구사업의 현황 및 연계성 분석”, 대한산업공학회지, 제41권 제3호, 2015, pp. 305-323.

조혁준, 김성근, 강주영, “도플갱어 브랜드 이미지 효과에 대한 실증적 분석: 인터넷 커뮤니티를 중심으로”, 정보시스템연구, 제26권 제1호, 2017, pp. 21-51.

채승훈, 임재익, 강주영, “사용자 리뷰를 통한 소셜커머스와 오픈마켓의 이용경험 비교분석”, 지능정보연구, 제21권 제4호, 2015, pp. 53-77.

최종화, 이광호, 조용래, 우청원, 이충현, 김은아, “정부출연연구기관의 협력적 융합연구 촉진방안”, 정책연구, 12월호, 2017, pp. 1-270.

Blei, D. M., Ng, A. Y. and Jordan, M. I., “Latent Dirichlet Allocation”, *Journal of machine Learning research*, Vol. 3, Jan, 2003, pp. 993-1022.

Cambria, E. and White, B., “Jumping Nlp Curves: A Review of Natural Language Processing Research”, *IEEE Computational intelligence magazine*, Vol. 9, No. 2, 2014, pp. 48-57.

Ghannay, S., Favre, B., Esteve, Y. and Camelin, N., “Word Embedding Evaluation and Combination”, Proceedings of LREC, 2016.

Grimmer, J., “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases”, *Political Analysis*, Vol. 18, No. 1, 2010, pp. 1-35.

Mallon, W. T. and Bunton, S. A., “The Functions of Centers and Institutes in Academic Biomedical Research”,

Analysis in brief, Vol. 5, No. 1, 2005, pp. 2-3.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J., "Distributed Representations of Words and Phrases and Their Compositionality", Proceedings of Advances in neural information processing systems, 2013.

Raasch, C., Lee, V., Spaeth, S. and Herstatt, C., "The Rise and Fall of Interdisciplinary Research: The Case of Open Source Innovation", *Research policy*, Vol. 42, No. 5, 2013, pp. 1138-1151.

Rubenstein, H. and Goodenough, J. B., "Contextual Correlates of Synonymy", *Communications of the ACM*, Vol. 8, No. 10, 1965, pp. 627-633.

Schummer, J., "Multidisciplinarity, Interdisciplinarity, and Patterns of Research Collaboration in Nanoscience and Nanotechnology", *Scientometrics*, Vol. 59, No. 3, 2004, pp. 425-465.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y. and Manning, C. D., "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions", Proceedings of Proceedings of the conference on empirical methods in natural language processing, 2011.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B., "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification", Proceedings of Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

Wang, X., Liu, Y., Chengjie, S., Wang, B. and Wang, X., "Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory", Proceedings of Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015.

Winograd, T., "Understanding Natural Language", *Cognitive psychology*, Vol. 3, No. 1, 1972, pp. 1-191.

Xianghua, F., Guo, L., Yanyan, G. and Zhiqiang, W., "Multi-Aspect Sentiment Analysis for Chinese Online Social Reviews Based on Topic Modeling and Hownet Lexicon", *Knowledge-Based Systems*, Vol. 37, 2013, pp. 186-195.

Young, T., Hazarika, D., Poria, S. and Cambria, E., "Recent Trends in Deep Learning Based Natural Language Processing", *IEEE Computational Intelligence Magazine*, Vol. 13, No. 3, 2018, pp. 55-75.

양 낙 영 (Yang, Nak-Yeong)

아주대학교 e-비즈니스학과에서 재학 중에 있다. 주요 관심분야는 텍스트마이닝, 머신러닝, 딥러닝 등 이다.

김 성 근 (Kim, Sung-Guen)

가톨릭대학교에서 심리학을 전공하였고, 아주대학교 경영정보학과에서 석사를 취득하였다. 현재는 아주대학교 연구팀 교직원으로 재직중에 있으며, 주요 관심분야는 온라인 광고, 머신러닝, 딥러닝 등 이다.

강 주 영 (Kang, Ju-Young)

현재 아주대학교 경영대학 e-비즈니스학과 교수로 재직중이며, 포항공과대학교 컴퓨터공학과에서 학사, 서울대학교 컴퓨터공학과에서 석사, 한국과학기술원 경영공학전공에서 공학박사학위를 취득하였다. 주요 관심분야는 빅데이터, 텍스트마이닝, 시맨틱 웹, 지능형 정보시스템 등이다.

<Abstract>

Researcher and Research Area Recommendation System for Promoting Convergence Research Using Text Mining and Messenger UI

Yang, Nak-Yeong · Kim, Sung-Geun · Kang, Ju-Young

Purpose

Recently, social interest in the convergence research is at its peak. However, contrary to the keen interest in convergence research, an infrastructure that makes it easier to recruit researchers from other fields is not yet well established, which is why researchers are having considerable difficulty in carrying out real convergence research. In this study, we implemented a researcher recommendation system that helps researchers who want to collaborate easily recruit researchers from other fields, and we expect it to serve as a springboard for growth in the convergence research field.

Design/methodology/approach

In this study, we implemented a system that recommends proper researchers when users enter keyword in the field of research that they want to collaborate using word embedding techniques, word2vec. In addition, we also implemented function of keyword suggestions by using keywords drawn from LDA Topicmodeling Algorithm. Finally, the UI of the researcher recommendation system was completed by utilizing the collaborative messenger Slack to facilitate immediate exchange of information with the recommended researchers and to accommodate various applications for collaboration.

Findings

In this study, we validated the completed researcher recommendation system by ensuring that the list of researchers recommended by entering a specific keyword is accurate and that words learned as a similar word with a particular researcher match the researcher's field of research. The results showed 85.89% accuracy in the former, and in the latter case, mostly, the words drawn

as similar words were found to match the researcher's field of research, leading to excellent performance of the researcher recommendation system.

Keyword: Recommendation System, Convergence Research, Slack, Word2Vec, LDA Topic Modeling, Messenger UI

* 이 논문은 2018년 11월 28일 접수, 2018년 12월 11일 1차 심사, 2018년 12월 26일 게재 확정되었습니다.

Appendix A. 분석에 사용된 키워드와 키워드 별 연구자 추천 정확도

Keywords	Field	Total recommend	Correct recommend	Ratio
machine learning	IT & Engineering	5	5	100%
database	IT & Engineering	9	7	77.78%
parallel processing	IT & Engineering	2	2	100%
ai	IT & Engineering	6	5	83.33%
neural network	IT & Engineering	3	2	66.67%
robotics	IT & Engineering	9	8	88.89%
oled	IT & Engineering	2	2	100%
finfet	IT & Engineering	3	1	33.33%
semiconductor	IT & Engineering	7	6	85.71%
combustion engine	IT & Engineering	3	2	66.67%
social network	Humanity & Social science	5	4	80%
cultural psychology	Humanity & Social science	5	3	60%
family	Humanity & Social science	8	8	100%
gender	Humanity & Social science	19	17	89.47%
feminism	Humanity & Social science	5	4	80%
labor	Humanity & Social science	17	14	82.35%
politics	Humanity & Social science	8	8	100%
culture	Humanity & Social science	9	9	100%
education	Humanity & Social science	12	11	91.67%
philosophy	Humanity & Social science	14	12	85.71%
economy	Management & Economy	9	9	100%
market	Management & Economy	4	4	100%
behavior economics	Management & Economy	5	5	100%
customer behavior	Management & Economy	9	8	88.89%
econometrics	Management & Economy	13	9	69.23%
accounting	Management & Economy	18	15	83.33%
finance	Management & Economy	14	13	92.86%
marketing	Management & Economy	20	19	95%
organizational behavior	Management & Economy	9	6	66.67%
business strategy	Management & Economy	2	2	100%
anesthesia	Medical Science & Pharmacy	3	2	66.67%
asthma	Medical Science & Pharmacy	9	9	100%
anatomy	Medical Science & Pharmacy	20	17	85%
poison	Medical Science & Pharmacy	3	2	66.67%
emergency medicine	Medical Science & Pharmacy	7	6	85.71%
urology	Medical Science & Pharmacy	8	7	87.5%
bone fracture	Medical Science & Pharmacy	4	3	75%
cancer	Medical Science & Pharmacy	5	5	100%
hemostasis	Medical Science & Pharmacy	3	2	66.67%
hypertension	Medical Science & Pharmacy	18	16	88.89%
neuroscience	Natural Science	19	17	89.47%
quantum physics	Natural Science	6	6	100%
biophysics	Natural Science	10	8	80%
organic chemistry	Natural Science	2	1	50%
biochemistry	Natural Science	7	6	85.71%
meteorology	Natural Science	7	5	71.43%
paleontology	Natural Science	3	2	66.67%
astronomy	Natural Science	7	5	71.43%
brain science	Natural Science	6	6	100%
cytology	Natural Science	3	2	66.67%
	Total	404	347	85.89%

Appendix B. 연구자와 연관성이 높은 단어들 및 실제 연구 분야

name	keywords	areas of study
Ahmed Amr	dbms, b+tree, hash, topic-models	machine learning, Big data, user modeling
Armitage David	prehistory, history, historia, internationalism, atlantic	intellectual history, international history, Atlantic history
Bebchuk Lucian	financialstability, economy, corporate	Law and Economics, Corporate Finance, Financial Economics, Corporate Law
Borjas George	labormarket, labor, macroeconomy, microeconomics	Labor Economics, Economics
Cammett Melani	developmentalism, lebanon, political	Comparative Politics, Development
Carpenter Daniel	politicized, bureaucratic, regulating	American Political Development, Bureaucraic Politics, regulation
Chandra Amitabh	healthsystem, healthier, diseconomies, cyberculture, second-trimester	economics, healthcare, technology
Cikara Mina	neurogenetics, metacognition, social, psychopath, empathy	social psychology, cognitive neuroscience, empathy
Clarke David	ceramic, polycrystalline, crystallins, material, micromechanics, microrobots	Materials Science and Mechanical engineering
Dave Rahul	gravitation, astronomy, cosmology	Astronomy, Astrophysics, Cosmology
Deshpande Rohit	marketing, marketplace, consumerbrand, brand, crosscultural, enculturation	Marketing, Branding, Cultural Entrepreneurship
Esty Benjamin	finance, macromarkets, debt, financialstability	Finance
Gabaix Xavier	refinancing, finance, market, stock-price	Economics, Finance
Garcia Pena Lorgia	ethnic, romantic, race, blackness, black-lives-matter	Literature, Ethnic Studies
Gaudet Rachelle	biochemistry, proteome, transmembrane, trp, cadherins	Structural biology, membrane proteins, TRP channels, cadherins
Gersen Jeannie	discriminability, cyberlaw, violent, copyright, homosexuality, assault, harassment	Crimnal Law, Family Law, Copyright, Sexual Assault&Harassment
Gershman Samuel	metacognition, cognitiva, neurogenetics, neurobiology, embedding, nlp	computational neuroscience, cognitive science, machine learning
Hensch Takao	interneuron, neuroprotection, neuroleptic, gaba	Neuroscience, Critical Period, GABA
Ivashina Victoria	financialstability, banking, bank, bankruptcy, debt, indebtedness	Financial Intermediation, Banking, Corporate Debt Markets
Jacob Daniel	atmosphere, photochemistry, biogeochemistry	Atmospheric Chemistry
Kaban Leonard	maxilar, maxilla, jaw, pectoralis	Pediatric Oral and Maxillofacial surgery, Jaw Tumors
Kertzer Joshua	internationalism, psychoanalysis, geopolitical, politicized	International Relations, Political Psychology
Long Bridget	macroeconomist, educativas, productmarket, macromarkets, economy	Economics of Education
Malloy Christopher	asset, finance, investor, macromarkets	Asset Pricing, Investor Behavior, Market Efficiency
Manoharan Vinothan	selfassembly, bio, biogenic, biopolymer, colloid	Self-Assembly, Colloid and Interface Science, Biophysics
McCormick Melissa M.	endosymbionts, fungi, wood, seed, eggplant	plant-fungus interactions, endosymbionts

Mickens James	web, datacenter, javascript, dbms, cyberinfrastructure	Web Services, Datacenters, Distributed Systems
Mitragotri Samir	phagocytosis, biomaterials, biosensors, biomedicine, biochip, nanomaterials	Pharmaceutics, Biomaterials, Bioengineering, nanomedicine
Nanda Ramana	entrepreneurship, entrepreneur, microentrepreneurs, finance, venture	Entrepreneurship, Entrepreneurial Finance, Venture Capital
Neafsey Daniel	phylogenomics, coinfection, evolvability, pseudogene, geneexpression	Genomics, Infectious Disease, Evolution, Population Genetics
Nelson Charles	infancy, erp, autism, infant	Infancy, ERP, face processing, autism
Pande Rohini	macroeconomy, politician, policymaking, microfinance	Development Economics, Politcal Economy
Parmigiani Giovanni	epigenome, intragenomic, statistical	Statistical genomics, Bayesian statistics
Pearson Ann	microorganism, organophosphate, microsample, microbe, microorganism, biogeochemistry	Organic Geochemistry, Mcrobial Biogeochemistry
Rao Gautam	macroeconomy, crossmarket, productmarket, marketplace	Development Economics, Behavioral Economy
Reimers Fernando	international, educativa, education	International Education, Global Education, 21st Century Education
Rice James	solid, perfect-fluid, geologic, geodesy, earthquake	Solid and Fluid Mechanics, Geophysics, EarthQuake Science
Rubin Donald	psychometrics, statistics, biofeedback, biomedical	Statistics, Psychology, Biostatiscs
Sima Corneliu	diabetic, periodontal, periodontitis	diabetes, periodontitis, Immunology
Snedeker Jesse	language, crosslanguage, neuropsychology, psychoanalysis, semantics	Language Development, Psycholinguistics, Semantics
Srinivasan Suraj	financialstability, corporation, accounting, finance, internationalisierung	Financial accounting, corporate governance, international
Star Jon	mathematics, mathematician, educativa, education	Mathematics Education
Stern Ariel	healthsystem, nanotechnology, innovation	Health Care Economics, Technology Management, Innovation
Toffel Michael	environmentalism, environment, selfregulation, selfregulatory	Environment, Sustainabillity, Regulation
Viceira Luis	asset, macromarkets, investor, investment, finance, econometrics	Finance, Asset Pricing, Capital Markets, Investment
Vlassak Joost	micromechanics, thermomechanical, battery, plastic, anisotropic, material	Mechanics, Materials
Zhou Qiao	stemcell, neuroendocrine, neuroinflammation, biokinetics, antibody, medicinal	Regenative Medicine, Stem Cell Biology, Neurodegeneration
Zimkus Breda	amphibian, biogeography, biosignatures, phylogenesis, phylogeography, phylogenetics	Amphilbians, Biodiversity, Biogeography, Phylogenetics
Colella Nicholas	metalorganic, organophosphate, semiconductor, Metallodielectric	organic semiconductors
Rebbeck Timothy	cancer, neurogenetics, geneenvironment, pharmacogenetics, mastectomy	cancer, genetics