다수 학술 정보 분석을 통한 전문가 검색 시스템

한건희, 김현, 임종태, 최도진, 이현병, 오영호, 편도웅, 방민주, 전종우, *복경수, 유재수
충북대학교, *원광대학교

dkalzk951@cbnu.ac.kr, kimh0701@cbnu.ac.kr, jtlim@chungbuk.ac.kr, mycdj91@cbnu.ac.kr, lhb@cbnu.ac.kr, ohy5268@kakao.com, pyun19@naver.com, minj2357@naver.com, junjongwoo30@naver.com, *ksbok@wku.ac.kr, yjs@cbnu.ac.kr

Expert Search System Through Analysis of Multiple Academic Information

Han Geon Hee, Kim Hyeon, Lim Jong Tae, Choi Do jin, Lee Hyeon Byeong, Oh Young Ho, Pyun Do Woong, Bang Min Ju, Jeon Jong Woo, *Bok Kyoung Soo, Yoo Jae Soo
Chungbuk National Univ, *Wonkwang Univ.

요 약

융·복합 산업의 중요성이 대두되면서 선도 연구를 진행하는 기관 및 기업들은 특정 분야 전문가들과의 협업을 위해 해당 분야의 전문가를 필요로 한다. 본 논문에서는 다수 학술 정보 분석을 통해 사용자 요구에 적합한 전문가를 검색하기 위한 시스템을 제안한다. 본 시스템에서는 다양한 학술정보 사이트에서 제공하는 학술적 정보를 수집하고 분석하여 전문가 검색 결과를 웹 애플리케이션 형태로 제공한다. 제안하는 시스템은 데이터 처리 속도의 증가를 위해 분산 데이터 스트리밍 플랫폼인 카프카(Kafka)를 사용한다. 또한 전문가들의 실적을 평가하기 위해 IF, 인용 수, 최신성, 연구 품질 등의 다양한 전문가 지수를 평가하기 위한 요소들을 이용하여 전문가 검색을 수행한다.

Ⅰ. 서 론

융·복합 산업이 4차 산업혁명 시대에 크게 주목받으면서 기업들은 다양한 분야의 전문가들과 협업이 필수사항이 되었다. 기업은 논문, 특허 등의 실적을 바탕으로 협업자를 선정하는데, 주로 학술정보 사이트나 전문가 추천 검색 기법들을 사용하고 있다.[1-3] 기업에서 업무를 진행 하는데 있어 해당 분야에 전문성이 보장된 전문가를 선정하여 협업하는 것은 매우 중요한 부분이다.

국내에 논문이 등재되는 학회는 약 2,000개가 존재하며 평균적으로 10건의 논문이 매월 학회에서 발표된다고 가정하면, 매년 약 24만 건의 논문이 발표된다. 추가로 국외 저널까지 고려한다면 24만 건 이상의 논문이 매해 출판된다고 볼 수 있다. 방대한 양의 정보로 인해 사용자가 학술정보 사이트에서 제공되는 정보로 직접 전문가를 찾기에는 시간적으로 소요가 클 뿐만 아니라 결과의 정확성을 기대하기 힘들다. 또한 기존의 전문가 추천 기법들은 다양한 학술정보 사이트로부터 제공되는 논문, 특허, 프로젝트 실적들을 통합한 분석 결과를 제공하지 않을 뿐만 아니라 한두 개 요소들로만 전문가를 검색하기 때문에 사용자가 요구 사항에 대한 신뢰도 있는 결과를 얻기 어렵다[4-5]. 이러한 문제점들은 기업들로 하여금 자동화된 분석 기법을 활용하여 전문가를 검색하고 추천해 주는 시스템에 대한 필요성이 증가되었다.

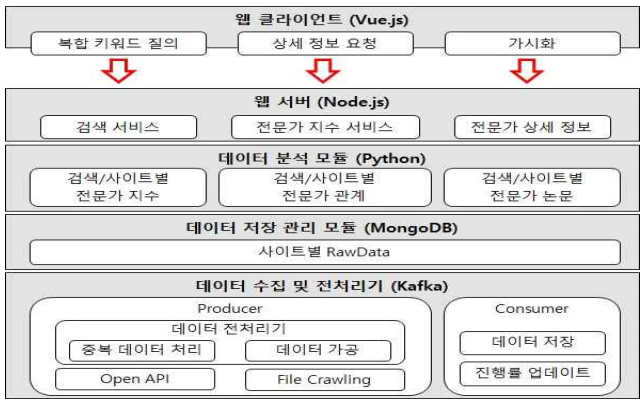
본 논문에서는 이러한 문제점을 해결하기 위해 여러 학술정보 사이트로부터 수집한 데이터를 기반으로 복합적인 요소를 고려한 전문가를 검색할 수 있는 자동화 검색 시스템을 제안한다.

Ⅱ. 제안하는 전문가 검색 시스템

1. 시스템 구조

본 논문에서 제안하는 시스템은 다양한 학술정보 사이트들에서 제공되는 연구자의 실적 정보들을 통합하여 분석한 후 사용자 질의에 적합한 전문가를 검색해 준다. 다양한 학술정보 사이트들로부터 데이터를 수집하기 때문에 풍부한 데이터 셋을 유지할 수 있다. 이때, 대용량의 데이터들을 처리할 때 발생할 수 있는 성능 저하를 방지하기 위하여 분산 데이터 스트리밍 플랫폼 카프카(Kafka)[6]를 통해 대량의 데이터를 빠르고 체계적으로 처리하도록 설계하였다. 또한 수집된 데이터들을 기반으로 도출된 검색/사이트별 전문가 지수와 전문가 관계를 가시화하여 웹 애플리케이션을 통해 사용자에게 제공하기 때문에 사용자는 직관적이고 명확한 결과를 확인할 수 있다.

[그림 1]은 제안하는 전문가 검색 시스템 구조를 보여준다. 제안하는 시스템은 웹, 데이터 수집 및 전처리, 데이터 저장관리, 데이터 분석 모듈로 이루어진다. 웹을 통해 사용자의 요청을 받아 데이터를 수집, 전 처리하여 저장하고 이를 분석한 결과를 다시 웹을 통해 사용자에게 가시화해주는 구조이다. 웹 클라이언트에서 사용자가 원하는 질의에 대한 상세 설정을 하면 웹 서버로 시스템에 필요한 서비스들을 처리한다. 데이터 수집 및 전처리기에서는 사용자로부터 입력받은 질의에 대한 데이터를 학술정보 사이트에서 수집하고 전처리한 후 데이터를 DB에 저장한다. 이때, 데이터를 저장하기 위해 Mongo DB를 사용한다[7]. DB에 저장된 데이터를 분석하여 결과를 도출한 후 사용자에게 웹 클라이언트를 통해 가시화하여 제공한다.

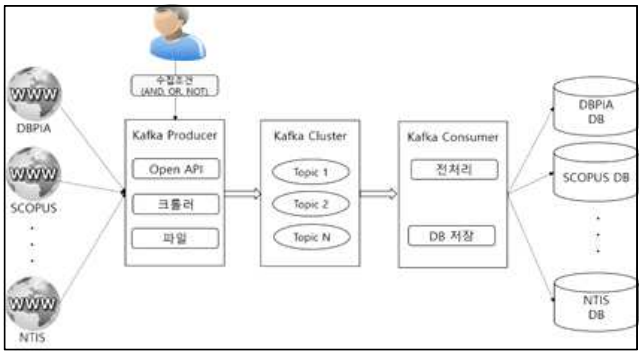


〔그림 1〕 전문가 검색 시스템 구조

2. 데이터 수집 및 전처리기 및 저장관리 모듈

사용자가 검색한 질의를 바탕으로 학술정보 사이트에서 API와 CRAWLER를 이용하여 데이터를 수집한다. 수집된 데이터는 중복 처리 및 데이터 가공 등의 전처리 과정을 거쳐 카프카(Kafka)를 통해 사이트별 RawData로 DB에 적재된다.

[그림 2]는 제안하는 시스템의 데이터 수집 및 전처리를 나타낸다. 먼저 Kafka Producer에서는 Open API와 크롤러, 파일 다운로드의 방식으로 사용자가 설정한 수집 조건에 맞는 정보들을 각 사이트에서 수집한다. 수집된 데이터는 Kafka Cluster에 사이트별로 해당되는 Topic에 분산 저장되어 처리된다. 이 데이터들은 Kafka Consumer에 의해 전처리 과정을 거친 후 각 사이트별 DB에 Raw Data로 저장되게 된다.

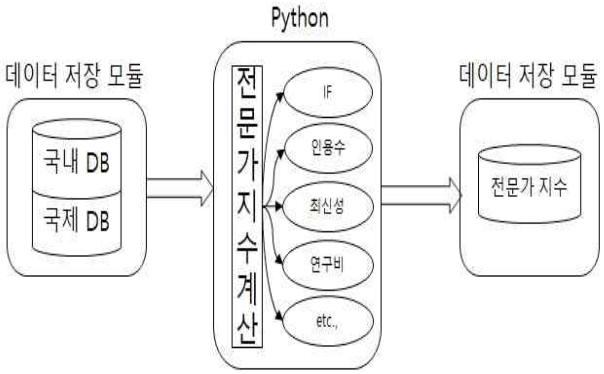


〔그림 2〕 데이터 수집 및 전처리

3. 데이터 분석 모듈

제안하는 시스템의 분석 모듈에서는 DB에 저장된 데이터를 기반으로 전문가 지수와 전문가 관계를 분석한다. 검색/사이트별 전문가 지수는 IF, 인용 수, 최신성, 연구 품질 등의 요소들을 기반으로 계산되며 전문가 관계는 연구자들이 특정 분야 별로 함께 진행한 연구 또는 작성한 논문의 수를 합산해 구성한다.

[그림 3]은 제안하는 시스템의 데이터 분석 과정을 나타낸다. DB에 저장된 데이터를 기반으로 전문가 지수와 전문가 관계를 분석한다. 전문가 지수 분석 시 논문의 IF는 국내 논문은 KCI , 해외 논문은 SCI의 IF를 사용한다. 인용 수는 Scopus, Web of science, KCI에서 제공하는 정보를 기반으로 한다. 최신성과 연구의 품질은 각각 실적의 시작 연도, 연구비 등을 고려하여 계산한다. 사이트별로 전문가 검색에 필요한 지수들이 계산되면 지수들에 사용자가 상세 검색에서 설정하였던 가중치를 부여하고 종합해 최종 전문가 지수를 산정하게 된다. 전문가 관계는 논문 또는 프로젝트에 참여한 각 연구자들 간 1:1 관계를 구성하며, 함께 참여한 횟수를 저장한다.



〔그림 3〕 데이터 분석

4. 웹 클라이언트

웹 클라이언트의 기능은 상세 검색 설정과 검색 제공 부분으로 나눌 수 있다. 상세 검색 설정 기능은 사용자가 웹페이지를 통해 질의를 입력받는다. 질의에 AND, OR 등의 연산을 지원하기 때문에 사용자는 복합 질의로 상세한 검색 조건을 설정할 수 있다. 또한 본인이 원하는 산정 요소에 가중치를 둘 수 있고 결과 산출에 있어 원하는 학술정보 사이트만을 선택하여 검색할 수 있기 때문에 사용자의 요구 사항에 적합한 검색을 지원한다.

결과 제공 부분은 전문가 시스템에서 도출된 결과를 가시화하여 사용자에게 제공한다. 결과 제공 부분은 사용자에게 검색 결과 선정된 전문가들의 순위와 점수, 전문가 관계를 가시화하여 사용자에게 제공한다.

Ⅲ. 결론

본 논문에서는 다양한 학술정보 사이트들로부터 수집한 대량의 데이터를 기반으로 전문가를 검색하는 시스템을 제안하였다. 제안하는 시스템은 다양한 학술정보 사이트들로부터 저자의 논문, 특허, 프로젝트 등과 관련된 실적 데이터를 수집하여 풍부한 데이터 셋을 유지할 수 있다. 데이터 셋을 기반으로 전문가 지수와 전문가 관계를 자동화 기법을 통해 분석하기 때문에 사용자에게 편의성을 동반한 신뢰도 있는 결과를 제공할 수 있다. 그리고 분산 데이터 스트리밍 플랫폼 카프카(Kafka)를 통해 대량의 데이터를 빠르고 체계적으로 처리하도록 설계하였다. 또한 전문가 지수는 논문의 IF, 인용 수, 최신성, 연구 품질 등을 고려하여 계산된다. 이를 바탕으로 전문가를 찾기 위한 기업, 연구자를 검색해보기 위한 일반 사용자들에게 신속하고 정확하며 직관적인 전문가 검색 결과를 제공할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원과(No. 2019R1A2C2084257) 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원과(No. NRF-2017M3C4A7069432) 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산업기술진흥원(KIAT)의 지원을 받아 수행된 연구임. (2019년 기업연계형연구개발인력양성사업, 과제번호 : S2755555)

참 고 문 헌

〔1〕 https://www.researchgate.net/
〔2〕 https://scholar.google.co.kr/
〔3〕 https://www.ntis.go.kr/
〔4〕 https://www.kci.go.kr/
〔5〕 Zhao, F., Zhang, Y., Lu, J., and Shai, O. "Measuring academic influence using heterogeneous author-citation networks", Scientometrics, Vol. 118 No. 3, pp. 1119-1140, 2019.
〔6〕 Kreps, J., Narkhede, N., and Rao, J. "Kafka: A distributed messaging system for log processing," In Proceedings of the NetDB, Vol. 11, pp. 1-7, 2011.
〔7〕 Jiang, W., Zhang, L., Liao, X., Jin, H., and Peng, Y. "A novel clustered MongoDB-based storage system for unstructured data with high availability," Computing, Vol. 96, No. 6, pp. 455-478, 2014.