토픽 모델링을 활용한 국내 지질학 연구 동향 분석

임송준, 김태용, 천정희, 박혜민

부경대학교, 지구환경과학과, 지하수 환경 연구실

빅데이터 연구팀

Environmental Hydrogeology Laboratory

연구목적 및 배경

지질학은 19세기 후반에 시작된 이래로 과학 학문으로서 상당히 발전했다. 지난 수십 년 동안 이 발전에 대한 분석들 중 상당수는 저명한 전문가들에 의해 예견되었으나, 그들의 분석은 실질적으로 정량화가 부족했다. 본 연구는 과거의 연구 흐름 분석을 통해 과거 연구의 성과와 한계점을 파악하고 향후의 지질학과 관련 연구에 대한 방향을 모색하는데 그 목적이 있다.

이론적 배경

LDA 분석

Topic Modeling은 Text Mining기법 중 하나로 비구조화된 문서집합에서 잠재된 Topic들을 추출해주는 확률적 모델 알고리즘이다(Blei et al. 2003).

그 중에서도 LDA(Latent Dirichlet Allocation)는 Topic modeling의 가장 대표적인 방법이다. 이는 이산 자료들에 대한 확률적 생성 모델로, 단어들의 확률을 이용하여 문서 집합 내의 잠재된 topic들을 찾아내는 기법이다(김태경, 최회련, 이홍철, 2016).

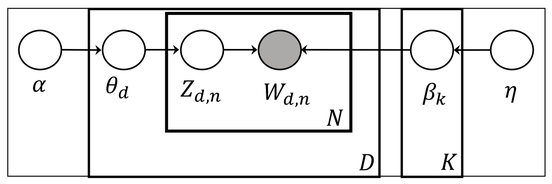

Fig. 1. LDA graph model.

시각화 도구

LDAvis는 기존의 Topic modeling 시각화의 한계점을 보완하고 topic과 단어의 관계를 전반적으로 살펴볼 수 있다. 각 topic과 topic 내 단어를 중요도에 따라 순위화 할 수 있으며, 해당 데이터베이스에서 주요 topic, 단어를 쉽게 파악할 수 있는 웹 기반 시각화 도구이다(Sivert, Shirely, 2014).

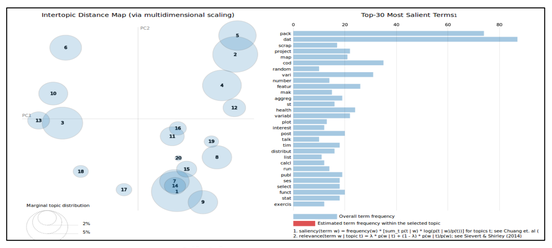

Fig. 2. LDAvis 예시.

연구방법

연구 모델

본 연구는 지질학 관련 학회에 등재된 논문의 제목에서 언급된 연도별 단어 빈도수를 통해 연구 동향을 파악할 수 있는데 전제를 두고 다음 그림과 같이 진행되었다. 전 처리 작업을 거친 데이터를 R을 이용하여 도출한 Wordcloud와 LDA 분석 결과값을 접목시켜 각각의 주제를 분류하고, 각 주제를 대표할 수 있는 단어를 선정하였다. 선정한 단어에 대한 연도별 빈도수의 LOWESS함수를 통해 각 연구의 시간에 따른 연구 흐름을 파악하고, 동시에 각 연구의 네트워크 분석을 통해 궁극적으로 국내 지질학 연구의 전반적인 동향을 분석하고자 한다.

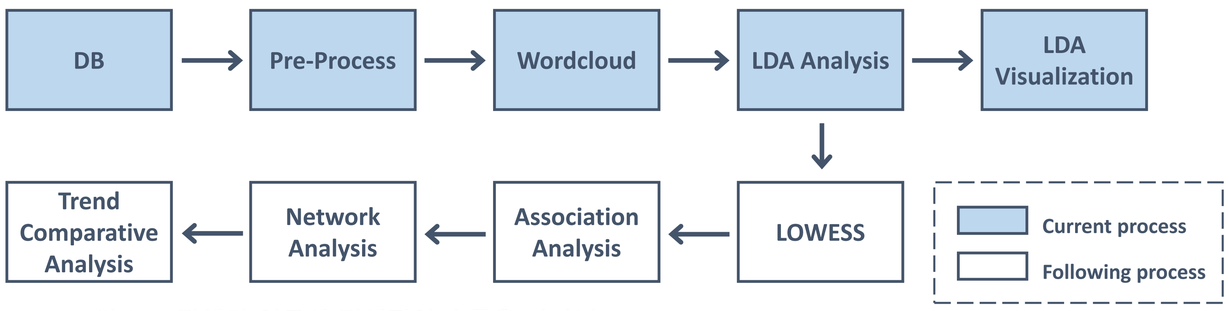

Fig. 3. 앞으로 진행될 연구의 전체적인 흐름을 나타낸 flow chart.

데이터 수집

본 연구는 지질학 관련 국내 학술지 9권을 대상으로 연구를 진행하였다. 국내 학술 DB 사이트 RISS에서 초기 발행 연도부터 2018년도까지 총 13,080개의 논문 제목 데이터를 수집하였다. 수집한 데이터의 종합 통계는 다음 표와 같다.

Table. 1. 학회 별 수집 데이터의 출간 연도와 논문 개수]

학회지 명	출간 연도	논문 개수	비율
한국지하수토양학회	1996	2,187	16.72%
대한지질학회	1964	2,014	15.40%
대한자원·환경지질학회	1968	1,980	15.14%
한국지반공학회	1985	1,889	14.44%
한국지반환경공학회	1997	1,022	7.81%
대한지질공학회	1990	949	7.26%
한국광물학회	1988	893	6.83%
Geoscience	1997	800	6.12%
한국지구물리·물리탐사학회	2004	799	6.11%
한국암석학회	1992	547	4.17%

전처리 과정

수집된 데이터는 Topic Modeling 분석을 수행하기 전에 적절한 전처리 과정이 수반되어야 한다. 연구 내용분석을 위한 정확한 용어 추출 과정은 Topic Modeling 수행 시 분석의 정확도 및 성능에 많은 영향을 주기 때문이다. 또한 수집한 데이터는 Topic Modeling을 수행하기에 적합한 형식으로 변경해야 한다. 다음은 본 연구에서 진행하는 전처리 과정을 도식화한 그림이다.

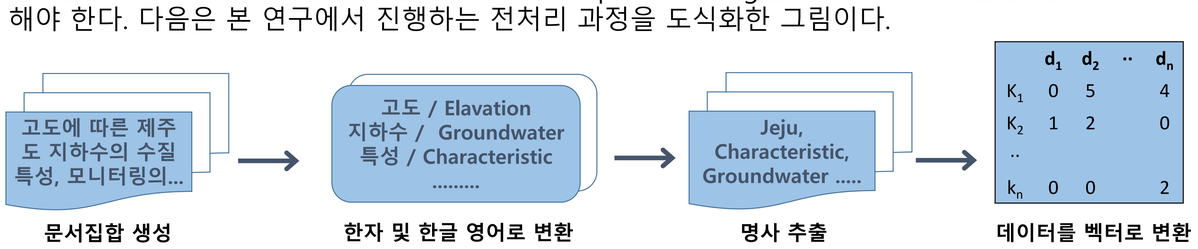

Fig. 4. 분석을 위한 전처리과정 도시화 Flow chart.

분석 결과

Wordcloud

Wordcloud는 논문 제목에 최다 빈출 100개의 단어를 나타낸다. 글자의 크기는 높은 빈도수, 색깔은 유사 빈도수를 나타낸다. Soil, Groundwater, Seismic의 단어 빈도수가 다른 단어들에 비해 높다. 그 다음으로 Rock, Basin, Model 등의 단어가 논문 제목에서 높은 빈도수로 사용되었고 Shear, Geochemical 등의 단어가 그 뒤를 이었다.

Wordcloud를 통해 산출된 결과를 미루어 보아 LDA의 결과 값은 Soil, Groundwater, Seismic 등 높은 빈도수를 나타낸 단어의 수가 많을 것이고, 그에 따른 연관된 단어들도 각 topic 내에 형성될 것이라고 판단된다.

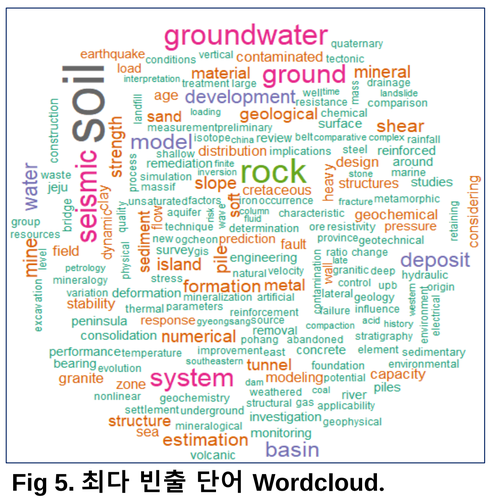

Fig 5. 최다 빈출 단어 Wordcloud.

LDA 결과

LDA Topic Modeling은 사전에 적절한 topic 수를 설정해야 한다. topic 수를 너무 높게 설정하면 특별한 키워드가 없어 의미 없는 topic이 도출될 수 있으며, topic 수를 적게 설정하면 한 topic에 많은 키워드가 뭉쳐 topic을 구분하기 어렵다. 이에 본 연구에서는 topic 수를 5~15개까지 설정한 후 각각 Topic Modeling을 수행한 결과, 8개가 각 topic을 적절하게 표현하는 것을 확인할 수 있었다.

좌측 그림의 topic 간의 거리는 연관성을, topic의 크기는 단어량을 의미하며, 우측의 그래프는 해당 topic 내 단어의 빈도수를 의미한다.

Table 2는 각 topic의 상위 5개 키워드와 각 지질학 분야 전문가의 의견에 따라 topic의 주제를 나타낸다.

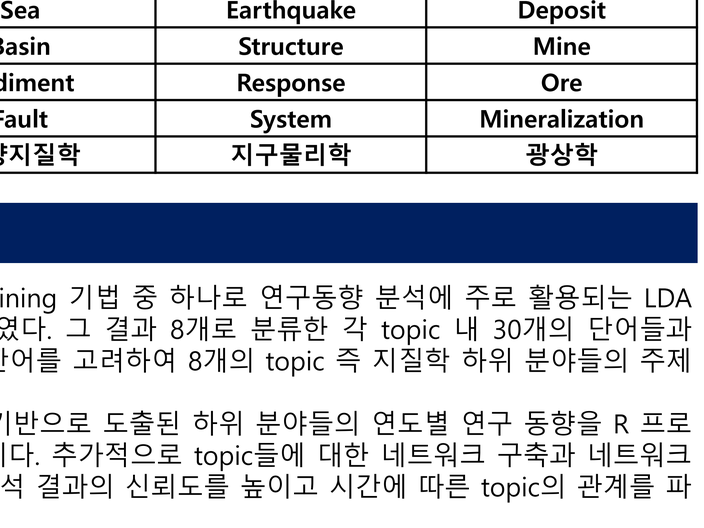

Fig 6. 8개로 분류한 topic과 상위 단어를 나타낸 LDAvis.

Table 2. 각 Topic의 주제와 상위 5개의 keyword

순위	Topic 1 (16.5%) 키워드	Topic 2 (15.2%) 키워드	Topic 3 (13.2%) 키워드	Topic 4 (13.2%) 키워드
1	Groundwater	Pile	Soil	Rock
2	System	Ground	Strength	Basin
3	Flow	Soil	Shear	Formation
4	Water	Tunnel	Material	Cretaceous
5	Model	Reinforcement	Clay	Age
Topic	지하수학	지반공학	지질공학	퇴적학
순위	Topic 5 (13%) 키워드	Topic 6 (10.6%) 키워드	Topic 7 (9.8%) 키워드	Topic 8 (8.4%) 키워드
1	Soil	Island	Seismic	Mineral
2	Metal	Sea	Earthquake	Deposit
3	Heavy	Basin	Structure	Mine
4	Contaminated	Sediment	Response	Ore
5	Groundwater	Fault	System	Mineralization
Topic	오염 및 정화	해양지질학	지구물리학	광상학

결론 및 연구 방향

국내 지질학 연구동향 분석을 위해 Text Mining 기법 중 하나로 연구동향 분석에 주로 활용되는 LDA Topic Modeling을 적용, 그 결과를 분석하였다. 그 결과 8개로 분류한 각 topic 내 30개의 단어들과 Wordcloud 결과에서 도출된 빈도수 상위 단어를 고려하여 8개의 topic 즉 지질학 하위 분야들의 주제를 선정하였다.

본 연구의 향후 방향은 Topic Modeling을 기반으로 도출된 하위 분야들의 연도별 연구 동향을 R 프로그램의 LOWESS 함수를 통해 파악하는 것이다. 추가적으로 topic들에 대한 네트워크 구축과 네트워크 간의 연관성 연구를 진행하여, 연구 동향 분석 결과의 신뢰도를 높이고 시간에 따른 topic의 관계를 파악하는 연구가 필요하다.

본 연구는 이전에 제시되지 않았던 지질학 관련 연구 동향을 파악하여 기존의 지질학 전문가, 새로이 연구를 시작하는 지질학 관련 종사자들에게 제공되는 하나의 지표가 될 것이다.

참고 문헌

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993-1022.

Carson Sievert and Kenneth E. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. proceedings of workshop on interactive language learning, visualization, and interfaces, Baltimore, Maryland.

김태경, 최회련, 이홍철. 2016. 토픽 모델링을 이용한 핀테크 기술 동향 분석. 한국산학기술학회 논문지, 17(11): 670-681.

박준형, 오효정. 2017. 국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교. 한국도서관-정보학회지 48 (4), 235-258.