

# 텍스트 마이닝 기법을 활용한 한국의 경제연구 동향 분석

## Trend Analysis of Korean Economy in the Economic Literature by text mining techniques

송혜지, 연세대학교 문헌정보학과, hj89song@yonsei.ac.kr  
박경수, 연세대학교 문헌정보학과, pksda@yonsei.ac.kr  
정혜은, 연세대학교 문헌정보학과, sunsilver61@yonsei.ac.kr  
송민, 연세대학교 문헌정보학과, min.song@yonsei.ac.kr  
Hye-Ji, Song, Dept. of Library & Information Science, Yonsei University  
Kyoung-Soo, Park, Dept. of Library & Information Science, Yonsei University  
Hye-Eun, Jung, Dept. of Library & Information Science, Yonsei University  
Min-Song, Dept. of Library & Information Science, Yonsei University

빅데이터를 활용한 데이터 분석 기법 중 비정형 데이터 분석의 하나인 텍스트 마이닝 기법을 활용하여, 외국 학술지에 나타난 한국의 경제 분야 트렌드를 분석한다. 데이터베이스로 Web of Knowledge의 연구논문을 활용하였으며, 키워드 분석, 네트워크 분석, 토픽모델링 분석을 통해 연구 동향 및 지적구조를 파악하는 데 그 목적이 있다.

## 1. 서론

### 1.1 연구의 필요성

학문의 연구자들이 생산한 문헌은 그 분야의 세부 주제영역, 학파, 연구자들의 정보이용행태 등이 반영된다. 문헌을 통하여 해당 학문분야의 지식속성을 파악할 수 있으며 학문분야의 지적구조에 대한 정보를 얻을 수 있다(곽선영, 2012). 학술논문을 분석함으로써 최근 연구자들 사이에서 관심을 갖고 있는 전문주제의 동향을 파악할 수 있으며 앞으로의 연구에 대해서도 시사점을 나타낼 수 있다. 이러한 주제 동향과 연구의 시사점은 연구대상, 방법, 자료수집 방법 등에 따라 다양한 관점에서 접근할 수 있으며, 그 중 텍스트 마이닝 기법을 통하여 방대한 데이터를 객관적으

로 분석할 수 있다.

본 연구는 이 기법을 적용하여 외국 경제학 학술지에서 나타난 한국 경제학에 대한 시각과 지적구조를 알아보고자 한다. 또한, 문헌클러스터링, 토픽 모델링과의 통합적 분석을 통해 경제학 분야의 연구주제의 변화와 그 양상을 종합적으로 분석한다. 궁극적으로는 연구 주제에 대한 추세 분석을 기반으로 미래의 경제 상황 및 연구주제 변화에 대한 예측 모델 제안의 초석이 될 수 있을 것이다.

## 2. 연구설계

### 2.1 자료 수집 및 설계

세계적인 학술데이터베이스인 Web of Knowledge에서 Korea economy로 검색한 결과 중

2000년부터 2013년까지의 논문을 대상으로 영문 초록을 추출하였다. 전처리 과정을 통해 분석시 불필요한 데이터를 삭제하고 데이터 형식을 통일하였다.

### 3. 연구 방법

#### 3.1 단순빈도 분석

2000년부터 2013년까지 연도별로 키워드 빈도수를 비교하였다. 빈도수별 상위 10개를 비교하였고, 기간별로 4개의 시기로 나누어서 각각의 추이를 비교하였다.

#### 3.2 문헌클러스터링

텍스트 마이닝 기법의 하나인 문헌 클러스터링을 통해 이슈가 되는 주제를 도출하고, 경제학 주제분류표(JEL Classification)와도 비교하였다.

#### 3.3 토픽모델링

토픽모델링은 대량의 텍스트 데이터에 내재되어 있는 Semantic 패턴을 찾아내는 것이다. 이 기법을 통해 연구 주제 변화의 추이를 알아보고, 문헌 클러스터링한 주제와도 비교해 봄으로써 다양한 시각으로 데이터를 분석하여 외국 학술지에서의 한국 경제 연구의 동향에 대해 알아보았다.

### 4. 연구 결과

#### 4.1 연도별 분석

2000년부터 2013년경제이슈와의 연관성까지 연도별 keyword frequency를 비교해 본 결과, 2003년과 2011년에 주목할 만한 결과가 나왔다. 2003년 키워드 1순위로 중국이 급

부상한 것을 알 수 있다. 실제로 2003년은 중국이 8.2%의 고성장을 달성하고, 세계경제성장의 중요한 동력원 역할을 강화한 시기였다.

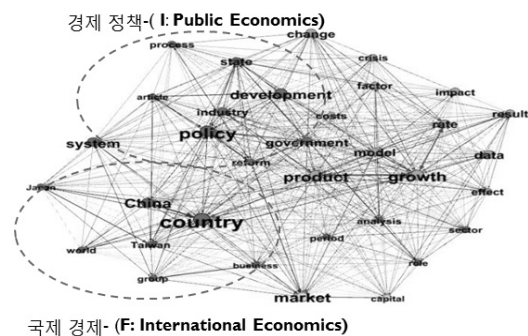
2011년의 키워드로는 Energy가 새롭게 등장하였는데, 이 시기는 한국 경제에서 환경을 고려하는 녹색성장의 노력이 결실을 맺는 시기와 동일하였다. 이러한 결과를 통해 연구논문의 내용을 통해 경제분야 이슈를 도출할 수 있음을 확인 하였다.

<표 1> 연도별 키워드 순위

순위	2003	순위	2011
1	China	1	Country
2	Policy	2	Energy
3	Taiwan	3	Policy
4	System	4	China
5	Result	5	Market
6	Participation	6	Research
7	Singapore	7	Development
8	Government	8	Japan
9	Country	9	Government
10	States	10	Growth

#### 4.2 기간별 분석

시기별로 키워드를 도출하고 클러스터링 하여 경제학 분야의 연구동향이 어떻게 변화되는지를 확인하였다. 사용한 자료의 수는 총 871개로, 데이터의 양과 연도별 키워드의 유사성 및 추세에 따라 총 4개의 기간으로 나누었다.



<그림 1> Period 1 키워드 네트워크

Period 1은 2000년 ~ 2006년의 226개의 논문초록을 데이터로 분석한 결과이다. 노트

는 243개, 엽지의 수는 14,758개 였으며, 총 5개의 그룹으로 클러스터 되었다. 파란색 부분은 Country, World, China, Japan 등 국제 경제와 관련된 그룹으로 JEL Classification F: International Economics로 분류하였고, 보라색 그룹은 Government, Policy, Development, Industry를 키워드로 한 I항목의 Public Economics로 분류 하였다.

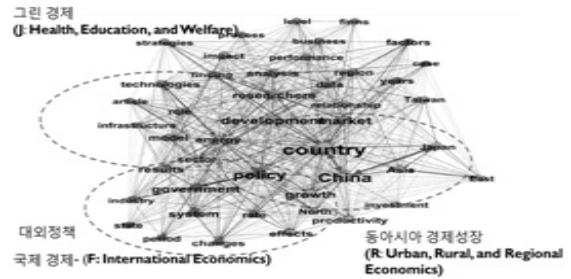
Period 2는 2007~2009년까지 205편의 논문 초록을 데이터로 사용하였으며, 노드의 수는 274개, 엽지의 수는 17,310개, 총 5개의 그룹으로 클러스터되었다.



<그림 2> Period 2 키워드 네트워크

빨간색으로 보이는 그룹은 상위 키워드로 Industry, Product, Market을 나타내며, JEL Classification F: International Economics로 분류했다. 연두색 그룹은 정부 경제 정책으로 선정하고, I: Public Economics로 분류하였다.

Period 3은 2010년부터 2011년까지 총 236개의 논문을 분석하였으며, 총 노드수 221개, 엽지 10,610개, 5개의 그룹으로 클러스터 되었다. 주목할 만한 특징은 환경과 에너지 기술에 관심을 가지는 그린경제(J: Health, Education, and Welfare)분야의 출현이다. 과거 성장위주의 개발중심에서 분배와 균형발전 및 삶의 질에 관심을 가지기 시작하면서 새롭게 연구되기 시작한 것으로 판단된다.



<그림 3> Period 3 키워드 네트워크

Period 4는 2012년과 2013년의 총 159개의 논문을 대상으로 분석하였으며, 노드수 214개, 엽지 9,688개, 총 4개의 그룹으로 클러스터 되었다.



<그림 4> Period 4 키워드 네트워크

분홍색 그룹은 Period 4에서 처음 등장한 분야로써 생산 및 경영 관리 등 산업분야와 관련된 주제로 판단하여 JEL Classification의 L 항목인 Industrial Organization으로 분류하였다.

기간별 분석결과를 종합적으로 확인한 결과, 키워드가 Growth, Market, Strategy, Change 쪽으로 변화된 것을 볼 때, 세계적인 경기 침체로 인해 시장 활성화와 경기부양 전략수립 등의 분야가 중점적으로 연구되는 것으로 판단할 수 있다.

4.3 토픽모델링 분석

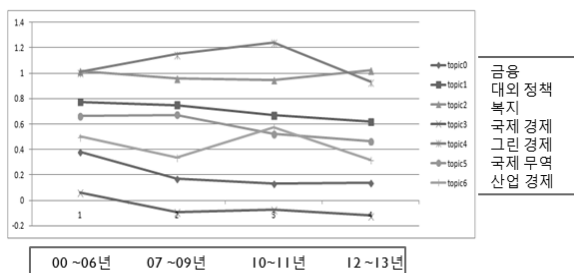
자료의 양을 고려하여 7개의 토픽을 추출한 결과 ‘금융’, ‘대외정책’, ‘복지’, ‘국제 경제’, ‘그린

경제’, ‘국제 무역’, ‘산업 경제’의 결과를 얻었다.

<표 2> 토픽모델링 결과

Topic 0	Topic 1	Topic 2	Topic 3
금융	대외정책	복지	국제경제
growth	north	health	developm
model	state	education	governme
crisis	political	care	market
rate	reform	social	policy
financial	social	population	policies
results	policy	national	articles
effects	relations	medical	role
capital	crisis	women	country
effect	reforms	urban	national
Topic 4	Topic 5	Topic 6	
그린경제	국제무역	산업경제	
energy	china	firms	
hydrogen	countries	business	
environme	asian	performance	
power	east	knowledge	
emissions	japan	technological	
technology	asia	technology	
fuel	taiwan	innovation	
nuclear	trade	research	
rights	singapore	based	

또한, 시기별 연구 동향 시각화를 통해 과거에 비해 활발한 연구 분야는 ‘그린 경제’, ‘복지’, ‘산업경제’이고 나머지는 크게 변화가 없거나 다소 하향하는 경향을 도출하였다.



<그림 5> 토픽모델링 결과: 연구동향 시각화

#### 4.4 경제학 주제분류 비교

단순 빈도수 기반 주제 분류와 토픽모델링과의 유사성을 비교하였다. 공통으로 출현하는 주제는 ‘국제 경제’, ‘그린 경제’, ‘국제 무역’, ‘산업 경제’ 분야이다. 반면에 ‘금융’과

‘복지’ 분야는 토픽모델링의 분류에만 출현하는데, 토픽모델링의 경우 언어의 의미론적 표현법을 기반으로 하는 추론 모델이기 때문에 좀 더 Semantic한 주제 분류가 가능하기 때문으로 해석 가능하다.

### 5. 결론

텍스트 마이닝 기법을 활용해 외국 학술지에 나타난 한국 경제 동향에 대해 분석하였다. 그 결과 연구주제가 한국의 경제 이슈와 연관성이 있음을 살펴 볼 수 있었을 뿐만 아니라, 토픽모델링과의 결합을 통해 시기별 연구동향 변화를 확인할 수 있었다. 또한 토픽모델링과 경제학 주제분류와의 유사성에 대해 분석하였다.

향후에는 연구를 더욱 발전시킬 수 있는 시간을 가지고, 많은 경로를 통해 한국 경제에 대한 연구 데이터 수를 늘려야 할 필요가 있다. 뿐만 아니라, 계량서지학-저자서지결합분석을 이용한 경제학 분야의 지적 구조 분석과의 비교를 통해 연구를 더욱 발전시켜 나갈 수 있을 것이다.

### 참고문헌

Kao, A & Potee, S. R.(Eds) (2007). Natural language processing and text mining. London:Springer-Verlag.

박자현, 송민. 2013. 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석. 『한국정보관리학회』, 30(1), 7-32.

이재영 외 2. 2007. 텍스트 마이닝을 이용한 국내 기록관리학 분야 지적구조 분석. 『한국문헌정보학회』, 41(1), 345-372

곽선영. 2012. 『경제학 분야의 복수저자동시 인용분석』, 석사학위논문. 이화여자대학교 대학원. 문헌정보학과.