토픽 모델링을 이용한 비정형 데이터 기반 산업간 유사도 분석

*김경원 **박종빈 ***정종진 ****윤경로

*,**,***전자부품연구원 *,***건국대학교

*kwkim@keti.re.kr

Analysis of similarity between industries based on unstructured data using topic modeling

*Kyungwon Kim **Jongbin Park ***Jongjin Jung ****Kyoungro Yoon

*,**,***Korea Electronics Technology Institute *,****Konkuk University

요약

최근 빠르게 변화하는 산업 환경에서 뉴스 기사와 같은 비정형 데이터를 기반으로 산업 트랜드를 분석하기 위한 연구가 진행되고 있다. 뉴스와 같은 비정형 데이터를 기반으로 산업별 트랜드를 분석하기 위해서는 분석 대상 산업에 대한 많은 양의 시계열 데이터가 요구된다. 하지만, 수집된 비정형 데이터를 분류하면 산업별/기간별 일정하지 않은 데이터 분포를 보이거나, 특정 산업에 대해서는 특정 기간에 데이터가 존재하지 않은 경우가 발생하여 산업별 시계열 분석이 어려운 경우가 발생할 수 있다. 이에, 본 논문에서는 산업별/기간별 균일하지 못한 비정형 데이터의 분포를 보정하기 위한 방법으로 비정형 데이터 기반 산업간 유사도를 분석 기법을 제안한다. 산업별 유사도 분석을 위해 각 산업별 주요 키워드를 도출하고 토픽 모델링 기법을 이용하여 산업간 유사도 분석을 통해 산업별/기간별 비정형 데이터 부족현상을 보완하는 방법을 제시한다.

1. 서론

기업 및 산업에 대한 투자 결정 및 특정 산업에 대한 평가 등의 의사결정을 위해 산업 트랜드 분석은 필수적인 요소로 자리잡고 있다. 산업 트랜드 분석을 위해 다양한 기법들이 연구되고 있으며, 대표적인 기존 산업 트랜드 분석 방법은 통계청에서 제공하는 표준산업분류를 기반으로 개별 기업을 하나의 산업 분류로 매칭하고, 각 기업들의 재무재표와 같은 통계적 정형 지표를 합산하여 시계열 분석 기법을 통해 분석하는 방식이다. 하지만, 기업의 통계적 재무 정보를 기반으로 분석하는 기존 방식은 각 기업의 재무지표의 취합이 완료되는데에 오랜 시간이 소요되어 급변하는 최신의 트랜드를 반영하지 못하는 단점이 있었다. 최근에는 빅데이터 및 인공지능 기술의 발전에 힘입어 산업 트랜드 분야에서도 비정형 데이터 기반 산업 트랜드 분석에 대한 연구가 활발하게 진행되고 있으며, 뉴스 기사와 같은 비정형 데이터 분석을 기반으로 시시각각 변화하는 산업 트랜드를 분석하는 기법에 대한 연구과 활발하게 진행되고 있다[1, 2]. 비정형 데이터를 기반으로 산업 트랜드 분석을 위해 뉴스 기사와 같은 비정형 데이터를 수집하고, 수집된 데이터를 산업별로 분류하여 시계열 분석을 수행하게 된다. 이 과정에서 수집된 데이터가 특정 산업이나 기간에 편중되는 현상이 자주 발생하게 되며, 이로 인하여 산업별 시계열 분석이 어려워지는 경우가 빈번하게 발생하게 된다.

본 논문에서는 비정형 데이터 기반 산업 트랜드 분석에서 빈번하게 발생되는 산업별/기간별 데이터 편중현상을 보완하기 위해 산업별로 수집된 뉴스 기사의 키워드를 기반으로 산업간 연관도를 분석하고, 이를 이용하여 산업별 부족한 데이터를 보정하는 방법을 제안한다.

2. 산업간 연관 관계 분석

비정형 데이터 기반 산업 트랜드 분석을 위해 일정기간 동안의 뉴스 기사를 수집하고, 수집된 뉴스 기사를 산업별로 분류한다. 뉴스 기사의 산업별 분류는 산업 평가 기관에서 제공하는 산업별 색인어를 기반으로 진행한다. 각 산업별로 분류된 뉴스 기사들은 시계열 분석을 위해 다시 월별로 분류된다. 이와 같이 산업별/월별 분류된 뉴스 기사를 이용하여 산업별 관심도, 호감도 등의 다양한 트랜드 분석이 이루어지게 된다. 이러한 산업별 트랜드 분석을 위해서는 수집 분류된 뉴스 기사들이 산업별/월별 일정한 수준의 분포를 유지해야만 한다. 하지만, 실제 뉴스 기사들은 산업별/기간별 일정하지 않은 분포로 발생하는 경우가 대부분이며, 이로 인하여 산업별/기간별 분석이 어렵게 된다. 본 논문에서는 뉴스 기사의 산업별 분류를 위해 사용된 산업별 색인어를 기반으로 산업별 연관관계를 도출하고, 산업별 연관도를 가중치로 하여 분석을 위해 부족한 산업별/기간별 데이터를 보완하도록 하였다.

2.1. 산업별 색인어 정제 및 확장

비정형 데이터 기반 산업간 연관관계 분석은 산업별 키워드를 이용하여 산업간 유사도를 이용하여 도출된다. 따라서, 양질의 산업별 키워드 리스트의 구성이 필수적으로 요구된다. 수집된 뉴스 기사를 산업별 색인어를 기반으로 산업간 연관관계 분석을 시도할 수 있다. 본 논문에서는 산업별 연관관계 분석을 위해 토픽 모델링(Topic Modeling) 기법 중 하나인 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 알고리즘을 적용한다[3, 4]. 토픽 모델링 기법을 사용하여 산업간 연관관계를 분석하는 경우, 토픽을 구성하는 말뭉치 크기가 분석 정

확도에 영향을 미치게 된다. 본 논문에서는 기 구축된 산업별 색인어 사전의 키워드 양이 산업간 연관관계 분석의 정확도에 미치는 영향을 최소화하기 위해 산업별 키워드를 확장하여 새로운 산업별 키워드 셋을 구축한다.

확장된 산업별 키워드 셋은 산업별 다중 분류된 뉴스 기사를 기반으로 기사의 주제어 가중치와 산업 분류 가중치를 활용하여 도출한다. 산업별 다중 분류 및 추가 정제 과정을 거친 뉴스 기사는 산업 분류 상위 5개 산업에 대하여 각 산업으로 분류될 확률로 정규화된 값을 가중치로 가진다. 뉴스 기사별 산업 분류 가중치를 해당 기사의 각 주제어 가중치와 곱해서 해당 키워드의 산업별 가중치를 도출하고, 산업별로 키워드 분류한다. 산업별 키워드 리스트가 생성이 완료되면, 산업별 동일 키워드별로 가중치를 합산하여 산업별 키워드 셋을 구성한다.

$$D_i = \{K_{1,d_i}, K_{2,d_i}, \cdots, K_{j,d_i}\}$$

$$D_i = [w_{d_i1}, w_{d_i2}, \cdots, w_{d_iC}]$$

$$K_{d_i,j} = [w_{d_ik_1}, w_{d_ik_2}, \cdots, w_{d_ik_J}]$$

$$K_c = \{k_{1,c}, k_{2,c}, \cdots, k_{N,c}\}$$

$$K_{j,c} = [w_{k_j1}, w_{k_j2}, \cdots, w_{k_jC}]$$

$$w_{k_jc} = \sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{j=1}^{J} w_{d_ic} \times w_{d_ik_j}$$

D_i : 뉴스 기사별 키워드 리스트,
각 뉴스 기사는 키워드별 가중치 w_{d_i,k_j}를 가지고 있음
각 뉴스 기사별 산업분류 가중치 $w_{d_i,c}$를 가지고 있음

$K_{d_i,j}$: 뉴스 기사 d_i 에 포함된 키워드 K_{d_i}의 가중치 벡터

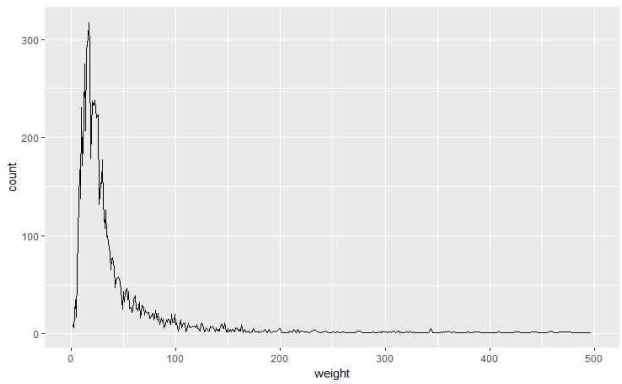
K_c : 산업별 키워드 리스트

$w_{d_i,c}$: 뉴스 기사 d_i 의 c 산업에 대한 분류 가중치

$w_{d_ik_j}$: 뉴스 기사 d_i 에 포함된 키워드 k_j의 가중치

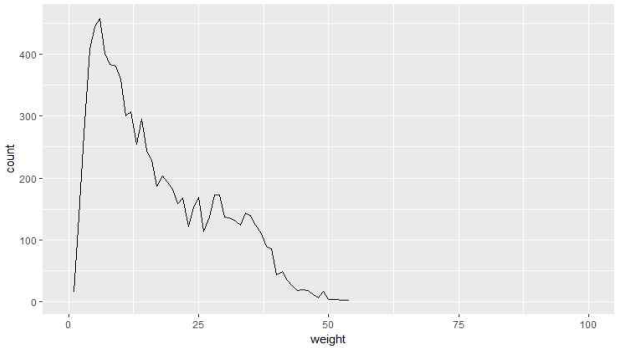
w_{k_jc} : c 산업 분류에 포함되는 키워드 k_j 의 가중치

도출된 산업별 키워드 셋에서 산업별 키워드 가중치 분포를 보면 상위 가중치 값과 하위 가중치 값의 편차가 심하게 나타나게 되는데, 이는 산업별로 상위 랭크되는 키워드가 유사하기 때문에 발생하는 현상이다[5].



<그림 1> 산업별 키워드 가중치 분포도

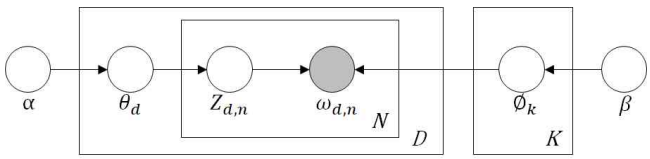
<그림 1>은 산업별 키워드 분포를 보여준다. 본 논문에서는 산업별 키워드 가중치 값의 편차를 줄이면서, 해당 키워드가 포함된 산업에 대한 중요도 반영을 위해 TF-IDF 알고리즘[6]을 적용하여 산업별 키워드 가중치를 보정하였다.



<그림 2> TF-IDF를 통해 보정된 산업별 키워드 가중치 분포도

2.2. 산업 연관도 분석

산업 연관 모형은 키워드 특성을 기반으로 산업 연관도 표현을 위한 분석 모델, 생성된 모델을 이용하여 각 산업 특성을 벡터로 표현하고 분석된 산업 벡터를 이용하여 산업간 유사도를 도출하는 유사도 분석 엔진으로 구성된다. 본 논문에서는 산업 특성을 표현하기 위한 분석 모델링 방법으로 토픽 모델링 기법 중 하나인 LDA 알고리즘을 이용하며, LDA 모델 생성 시에, 기 구축한 산업별 확장 키워드 셋을 활용한다. 또한, 산업별 키워드별 도출된 가중치를 반영하여 분석을 수행한다.



<그림 3> LDA 모델 아키텍처

본 논문에서는 산업별로 다중 분류된 뉴스 기사들의 산업별 분류 가중치와 각 뉴스 기사에 포함된 주제어 가중치 연산을 통해 산업별 키워드를 확장한다. 확장된 산업별 키워드는 키워드가 속하는 산업별 키워드 가중치를 포함하고 있다. 각 산업별 키워드 셋을 LDA 모델의 전체 문서 개수 D로 설정하고, 해당 산업에 해당하는 키워드의 가중치를 반영하여 키워드 $w_{d,n}$을 생성한다. 따라서, 산업 분류 개수 d는 본 논문의 산업 분류 기준인 232개의 소분류가 되며, 해당 산업군의 전체 키워드 개수 n은 해당 산업군에 속하는 키워드와 해당 키워드의 가중치 곱으로 표현된다. Φ_k는 k번째 토픽에 해당하는 벡터로 모든 산업 분류에 포함되는 가중치가 반영된 키워드 개수만큼의 길이를 가지게 되고, 최적 토픽의 개수 K는 Perplexity 지표와 실험을 통해 결정한다. θ_d의 값은 k번째 토픽이 해당 d번째 산업군에서 차지하는 비중을 표현하게 된다.

산업별 분류된 키워드를 이용하여 LDA 모델을 생성한다는 것은 토픽의 산업 키워드 분포와 각 산업별 토픽 분포를 추정하는 과정이다. 따라서, 토픽의 산업 키워드 분포와 각 산업별 토픽의 결합확률이 최대가 되도록 모델링된다.

산업별 키워드와 키워드 가중치를 이용하여 LDA 모델이 완성되면, 모델의 토픽 공간에 각 산업을 사상할 수 있게 된다. i번째 산업군 d_i은 다음 수식과 같이 K차원의 벡터로 표현할 수 있다.

$$d_i = \left[\theta_{d_i1}, \theta_{d_i1}, \cdots, \theta_{d_ik}\right], \quad k \in \{1, 2, \cdots, K\}$$

d_i : i 번째 산업군에 대한 각 토픽의 비중 벡터
(i는 소분류 기준으로, $i \in \{1, 2, \cdots, 232\}$)

θ_{d_ik} : i 번째 산업군d_i의 k번째 토픽의 비중

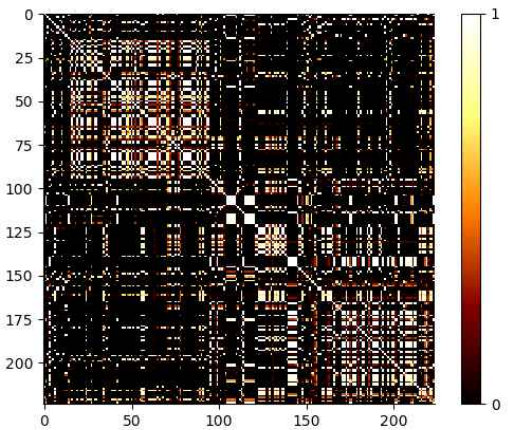
각 산업의 키워드를 기반으로 생성된 K차원으로 사상된 산업 벡터를 이용하여 산업간 유사도를 도출하며, 산업간 유사도는 코사인 유사도 기법을 사용한다[6]. 코사인 유사도는 내적공간에서 두 벡터간의 코사인 값을 기반으로 벡터간의 유사도를 측정하는 용도로 사용되며, 다차원의 양수 공간에서의 벡터간의 유사도를 측정하는 데에 유용한 기법이다. 두 산업 벡터 A, B 의 코사인 유사도는 다음 수식을 이용하여 [0, 1] 사이의 값으로 표현할 수 있다.

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

각 산업 벡터를 이용하여 산업들 간 유사도 매트릭스를 <표 1>과 같이 생성하고, 유사도를 기준으로 산업별 유사도 랭킹을 이용하여 특정 산업군에 대해 타 산업군과의 유사도를 구할 수 있다.

<표 1> 산업간 유사도 매트릭스 (예시)

산업 코드	C10100	C10200	C10300	C10400	C10500	C10600
C10100	1	0.3316	0.9936	0.9999	0.9999	0.1096
C10200	0.3316	1	0.4725	0.4755	0.4755	0.0442
C10300	0.9936	0.4725	1	0.9937	0.9937	0.0924
C10400	0.9999	0.4755	0.9937	1	0.9999	0.0959
C10500	0.9999	0.4755	0.9937	0.9999	1	0.0930
C10600	0.1096	0.0442	0.0924	0.0959	0.0930	1



<그림 4> 산업간 유사도 시각화 결과

<그림 4>는 도출된 산업간 유사도 시각화한 그래프로 산업간 연관도가 높을수록 밝은 색으로 표시된다. 도출된 산업간 유사도는 비정형 데이터 기반으로 산업 트랜드 분석 시에 분석 대상이 되는 각 산업별 평가 지표를 보정하기 위한 가중치로 이용되며, 본 논문에서는 산업별/기간별 데이터 분포의 불균형을 보정을 위해 활용하였다.

3. 결론

본 논문에서는 뉴스 기사와 같은 비정형 데이터를 기반으로 산업 트랜드를 분석시에 산업별/기간별 데이터 분포의 불균형을 보정하기 위한 방법으로 활용 가능한 산업별 키워드 기반 산업 연관도 분석 기법에 대해 제안하였다. 본 논문에서 제안한 산업 유사도는 산업별/기간별 비정형 데이터 분포의 불균형 보정뿐 아니라, 산업간 지표 연산이나, 산업 트랜드 비교 등 다양한 분야에 활용이 가능하다.

향후 연구로 본 논문에서 제안한 키워드 기반의 산업간 연관도 분석 이외에 산업별로 도출된 다양한 지표들을 활용하여 산업간의 연관도를 분석하여 산업별 평가 방법에 대해 연구를 진행할 계획이다.

[ACKNOWLEDGEMENT]

이 연구는 2018년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임('20000195', '중소·중견 가전사의 IoT가전 제품개발 전주기 지원을 위한 빅데이터 상용화 플랫폼 개발')

[참고 문헌]

[1] K. Kim, T. Lim, K. Yoon, "Industry Evaluation Analysis System for Enhanced Industry Analysis Information", Information, vol.20, pp.2537-2542, 2017.

[2] Q. Fang, Y. Chen, "Stock Portfolio Analysis based on Margin of Safety and Competitive Edge Evaluation in the American Wine Liquor Industry", International Conference on Big Data and its Applications, 2016.

[3] Blei, M. David, A. Y Ng, M. I. Jordan, "Latent Dirichlet Allocation", The Journal of machine Learning research, pp.993-1022, 2003.

[4] S. Lee, "News Keyword Extraction for Topic Tracking", Networked Computing and Advanced Information Management, Vol.2, pp.554-559, 2008.

[5] S. Nam, "A Meta-Analysis of the Relationship between Mediator Factors and Purchasing Intention in E-Commerce Studies", JICCE, pp.257 - 262, 2014.

[6] Singhal, Amit "Modern Information Retrieval: A Brief Overview", IEEE Computer Society Technical Committee on Data Engineering, Vol.24, pp.35 - 43, 2001.