Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2021.44.4.113

Vol. 44, No. 4 : 113-127, December 2021
ISSN : 2005-0461 / EISSN : 2287-7975

Topic Modeling Analysis Comparison for Research Topic in Korean Society of Industrial and Systems Engineering: Concentrated on Research Papers from 1978～1999

Dong Joon Park* · Hyung Sool Oh** · Ho Gyun Kim*** · Min Yoon****†

*Department of Statistics, Pukyong National University
**Department of Industrial and Management Engineering, Kangwon National University
***Department of Industrial and Management Engineering, Dongeui University
****Department of Applied Mathematics, Pukyong National University

한국산업경영시스템학회지 연구 주제의 토픽모델링 분석 비교: 1978년～99년 논문을 중심으로

박동준* · 오형술** · 김호균*** · 윤 민****†

*부경대학교 통계학과
**강원대학교 산업경영공학과
***동의대학교 산업경영공학과
****부경대학교 응용수학과

Topic modeling has been receiving much attention in academic disciplines in recent years. Topic modeling is one of the applications in machine learning and natural language processing. It is a statistical modeling procedure to discover topics in the collection of documents. Recently, there have been many attempts to find out topics in diverse fields of academic research. Although the first Department of Industrial Engineering (I.E.) was established in Hanyang university in 1958, Korean Institute of Industrial Engineers (KIIE) which is truly the most academic society was first founded to contribute to research for I.E. and promote industrial techniques in 1974. Korean Society of Industrial and Systems Engineering (KSIE) was established four years later. However, the research topics for KSIE journal have not been deeply examined up until now. Using topic modeling algorithms, we cautiously aim to detect the research topics of KSIE journal for the first half of the society history, from 1978 to 1999. We made use of titles and abstracts in research papers to find out topics in KSIE journal by conducting four algorithms, LSA, HDP, LDA, and LDA Mallet. Topic analysis results obtained by the algorithms were compared. We tried to show the whole procedure of topic analysis in detail for further practical use in future. We employed visualization techniques by using analysis result obtained from LDA. As a result of thorough analysis of topic modeling, eight major research topics were discovered including Production/Logistics/Inventory, Reliability, Quality, Probability/Statistics, Management Engineering/Industry, Engineering Economy, Human Factor/Safety/Computer/Information Technology, and Heuristics/Optimization.

Keywords : Topic modeling, Visualization, Industrial engineering, Latent Dirichlet Allocation

Received 14 November 2021; Finally Revised 9 December 2021;
Accepted 10 December 2021
† Corresponding Author : myoon@pknu.ac.kr

1. 서 론

사회가 디지털화되면서 처리하는 데이터의 양이 급격히 증가하게 되었고 이러한 대용량의 데이터를 1990년대부터 “빅 데이터”라고 부르기 시작하였다[29]. 빅 데이터는 구체적으로 데이터를 양(volume), 입출력 속도(velocity), 다양성(variety)의 세 요소로 분류하는데 학문 분야에서도 학술연구 활동이 왕성해지고 텍스트 형태의 연구 성과들이 누적되면서 방대한 양의 학술 자료들이 데이터베이스에 축적되었다.

1958년에 국내 최초로 한양대학교에 신설된 공업경영학과는 인간, 설비, 자재를 통합한 시스템에 과학적 원리를 도입하여 체계적인 경영을 연구하는 전공으로서 1970년대 후반에는 학과명이 산업공학과로 변경되었다. 최근에는 4차 산업혁명과 대학 전공의 융합 움직임에 따라 전공 영역이 전통적인 산업공학 분야에 국방, 서비스, 금융, 정보통신, 공공 행정 등으로 꾸준히 확장되고 있다. 1965년 산업공학과 관련한 최초의 학회인 한국공업경영학회를 설립, 한양대 중심으로 산업공학 초창기의 발전기반을 다져가다 법인 등록한 한국공업경영학회는 2000년에 한국산업경영시스템학회로 명칭 변경되었다. 1974년 대한산업공학회가 설립되어 산업공학 분야의 주력 연구학회 역할을 하고 있으며, 1982년 대한인간공학회가 설립되어 인간공학에 관한 학술과 기술 진흥을 도모하고 있다. 이 밖에 산업공학 관련 학회로는 품질관리학회, 경영과학회, 군사OR 학회 등이 있어, 기존의 학회와 보완관계를 유지하면서 산업공학의 발전 및 실용화를 꾀하고 있다.

한국산업경영시스템학회는 43년의 역사가 되었으나 학회지에 게재된 연구 논문 주제를 구체적으로 분석한 연구는 현재까지 진행되지 않았다. 학회지를 탐색한 결과 1999년까지 게재된 논문은 형식 측면에서 논문 제목, 초록, 키워드, 본문, 참고문헌 등의 완벽한 학술연구지의 형식에 미흡한 것으로 나타났다. 본 연구는 우선 1999년 이전에 게재된 본 학회지의 연구 논문들에 대하여 논문 제목과 초록을 대상으로 각각 토픽모델링 알고리즘을 활용하여 심도 있는 주제 분석을 함으로써 향후 학회지의 위상과 연구 발전에 기여하고자 한다.

제2절에서 이론적 배경인 토픽모델링의 의미와 알고리즘을 비교하고 토픽모델링을 적용한 선행연구에 대하여 살펴본다. 제3절에서는 본 학회지에 게재된 논문의 주제 분석을 위한 연구개요와 구체적인 절차에 대하여 서술한다. 제4절에서는 본 학회지에 게재된 논문의 제목과 초록을 대상으로 토픽모델링 알고리즘을 적용하여 시각화 기법을 포함한 주제 분석을 실행하며, 분석 결과를 비교하고 요약한다. 마지막으로 제5절은 본 연구의 의미와 한계 그리고 향후 연구 방향에 대하여 서술한다.

2. 이론적 배경

2.1 토픽모델링

텍스트마이닝(text mining)이란 비정형 자료(unstructured data)인 텍스트 데이터로부터 특징을 추출하여 가치 있는 정보를 발견하는 기술이고, 토픽모델링은 여러 문서들로부터 주제를 찾기 위하여 텍스트마이닝을 구체적으로 실현한 기법이다. 즉, 토픽모델링은 확률적 개념을 도입하여 여러 문서에서 많이 나타나고 유의미한 단어들의 연관성을 찾아내고, 단어들 조합의 결과를 제시하여 그 분석 결과인 단어의 조합이 함의하는 잠재적인(latent) 의미인 주제(topic)를 찾는 컴퓨팅 기법이다.

토픽모델링은 자연어(natural language) 연구에서 출발하였는데 주로 언어, 심리, 사회, 컴퓨터학자들이 인간이 사용하는 단어와 단어의 구성체인 문장과의 의미를 인식(acquisition of recognition)하는 과정을 연구하면서 시작되었다[15, 26].

2.2 토픽모델링 알고리즘의 비교

많은 토픽모델링 기법 가운데 대표적인 알고리즘을 살펴본다. 문헌의 주제를 찾는 기법인 토픽모델링의 기본적인 가정은 문헌은 여러 주제로 구성되고, 각 주제는 여러 단어로 구성된다는 점이다.

문헌들의 공통 주제들을 찾기 위하여 문헌의 개수와 주요 단어의 개수를 행 또는 열로 정하고, 각 문헌에서 나타나는 주요 단어의 빈도를 행렬의 원소로 하여 행렬을 구성한다. 이렇게 구성된 행렬의 차원을 축소하기 위하여 특이값분해(singular value decomposition)를 한 다음, 주제를 결정하는 단어들의 조합으로서 축소된 행렬의 행 또는 열의 원소들을 제시하는 알고리즘이 잠재의미분석(LSA: Latent Semantic Analysis or LSI: Latent Semantic Indexing)이다[25, 38]. LSA의 장점은 선형대수의 특이값분해를 사용함으로써 행렬의 차원을 줄이고 계산비용을 절약할 수 있으나, 주제의 수가 늘어날 경우에는 각 주제에 중요 단어들이 중복으로 나타나서 주제 의미부여가 어려운 단점이 있다. LSA를 개선하기 위하여 확률적 추론의 개념을 도입한 알고리즘이 확률적 잠재의미분석(Probabilistic Latent Semantic Analysis)이다[16]. 이 방법은 LSA 보다 주제 검색에는 우수하지만 다중 차원의 벡터생성으로 말미암아 연계된 네트워크 모델과의 추가적인 해석이 필요하다.

토픽모델링에 가장 많이 활용되는 알고리즘은 잠재디리클레할당(LDA: Latent Dirichlet Allocation)이다[4]. LDA는 인간이 문장을 구성할 때 주제를 결정하고, 주제에 포함될 단어들을 선택하는 과정을 모방하여 확률분포로 모

형화한다. 문헌에 나타나는 어떤 주제들이 디리클레(Dirichlet) 분포를 하고, 선정된 주제에 나타나는 단어들은 다항(Multinomial)분포를 하는 것으로 가정한다. 그러면 여러 문헌 가운데 선택된 한 문헌에서 나타난 단어의 확률은 선택된 그 문헌이 주어졌을 때 나타나는 주제의 조건부 확률과 그 문헌과 주제가 주어졌을 때 나타나는 단어의 조건부 확률의 결합 확률로 표현된다. 이 결합 확률분포는 켤레사전(conjugate prior)분포로서 주제들의 분포인 디리클레분포와 가능도(likelihood)함수로서 단어들의 분포인 다항분포를 곱하는 베이스정리(Bayes' theorem)를 적용하여 구할 수 있고, 이 결합 확률분포의 모수들의 최대값은 깁스샘플링(Gibbs sampling)으로 추정할 수 있다. 이와 같이 추정하여 선택한 단어들의 조합을 제시하는 알고리즘이 잠재디리클레할당이다. 그리고 위와 같은 문헌과 주제와 단어의 분포를 계층 구조로 만든 알고리즘이 계층적 디리클레 프로세스(Hierarchical Dirichlet Process)이다[42]. 토픽모델링에 활용되는 대표적인 알고리즘과 알고리즘을 실행하는 도구를 <Table 1>에 요약하였다[1, 3, 32, 43].

<Table 1> Typical Techniques for Topic Modeling Algorithms and Tools

Algorithms		
Techniques	Features	Major Participants
LSA	• LSA uses "singular value decomposition" to reduce a matrix containing word counts per document.	Deerwester et al.[12]
pLSA	• pLSA models co-occurrence information under a probabilistic framework to discover the underlying semantic structure of data.	Hofmann[16]
LDA	• LDA calculates the probability that certain words will be included in each topic, assuming that multiple words can be grouped under different topics.	Blei et al.[4]
HDP	• HDP is a Dirichlet process mixture model with multilevel form that is a nonparametric Bayesian approach to clustering grouped data.	Teh et al.[42]
Tools		
Techniques	Features	Major Participants
Gensim Python	• Gensim Python provides modules that generate automatic extraction of topics.	Rehurek and Sojka[39]
Stanford Topic	• Stanford Topic develops a toolbox for social scientists and non-engineers without background in text processing.	Ramage et al.[38]
Mallet	• Mallet provides a Java-based toolkit which is open source software.	McCallum [30]
Fathom	• Fathom provides a real-time computational framework and a mixed-initiative paradigm to train coherent topic and word distributions.	Dinakar et al.[14]

* Parts of this table are cited from Barde and Bainwad(2017)[3] and Mulunda et al.(2018)[32].

2.3 학술 분야의 활용

텍스트마이닝 또는 토픽모델링을 적용하여 연구 문헌의 중요한 정보를 추출하는 연구는 2010년에 들어서면서 다양한 학문 분야에서 활발히 진행되었다. 우선 산업공학 분야와 관련된 선행연구부터 살펴본다.

텍스트마이닝을 이용한 산업공학의 논문 주제어들의 상관성 연구로서 Cho and Kim[8]은 1969년 이후 43년간 국제학술지 IIE Transactions의 2,527개 논문에서 10회 이상 출현하는 주제어들 가운데 48개 단어를 선정하여 K-평균군집분석(K-means clustering algorithm)을 수행한 결과, "Quality and Reliability Engineering", "Design and Manufacturing", "Operations Engineering and Analysis", "Scheduling and Logistics"의 4개 군집으로 분류하였다. 그리고 넷마이너를 활용하여 관련성이 큰 주제어를 연결한 사회연결망(social network and modularity analysis)을 제시하였다. Cho et al.[6]은 2000년부터 2012년까지 대한산업공학회지, IE Interfaces, 한국산업경영시스템학회지, 한국경영과학회지의 3,875편 논문에서 저자가 작성한 주제어를 수집한 다음, 산업공학용어사전을 기준으로 최종 선정한 38개의 주제어들을 대상으로 단순 빈도분석, K-평균 군집 분석, 연관성 분석을 하고, 연구기법의 변화 추이를 보였다.

토픽모델링을 활용한 산업공학 분야의 실질적인 연구로서 Jeong and Lee[18]는 2001년부터 2015년까지 대한산업공학회지와 IE Interfaces의 논문 1,242편의 영문 제목 및 초록을 자료로 활용하여 LDA 분석을 하였다. 출현확률이 높은 단어들로 구성된 상위 50개 토픽을 도출하고 5년 단위로 나눈 각 구간에 대한 상위 10개 토픽을 제시하였는데 최근 주목받는 유망한 주제들은 "Technology management", "Financial engineering", "Data mining: supervised learning", "Efficiency analysis"임을 보였다. 광범위한 산업공학의 토픽모델링 연구로서 Kim and Jang[22]은 2004년부터 2015년까지 Industrial Engineering & Management Systems, 대한산업공학회지, IE Interfaces, 한국SCM학회지, 한국경영공학회지, 지능정보연구, 한국산업경영시스템학회지, 한국품질경영학회지의 3,251개 논문 초록을 대상으로 R을 이용한 LDA 분석을 하였다. 도출한 20개 연구 주제 가운데 "헬스케어", "금융공학", "기업성과", "텍스트마이닝", "의사결정 시스템", "데이터마이닝"의 활발한 연구가 진행되고 있음을 보였다.

산업공학의 세부 분야로서 제품서비스시스템(PSS: product

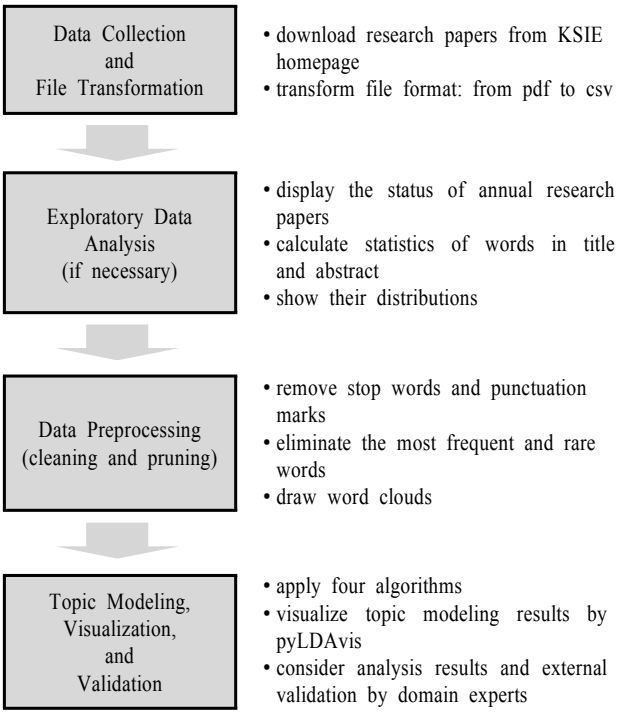
service system)의 토픽모델링 연구도 진행되었다. Seo and Lee[40]는 2017년에 데이터베이스로부터 제품서비스시스템에 관한 1,229편의 논문을 수집하여 관련성이 큰 주제어들의 네트워크를 작성하고, LDA 분석으로 10개의 주제를 도출하였다. 주제 가운데 “PSS for sutainability”는 쇠퇴하는 반면, “PSS business model for value co-creation”, “Industrial PSS”, “PSS framework and methodology”는 부상하는 주제임을 보였다. 품질경영 분야에 대한 토픽모델링 연구는 아직까지 활발하지 못한 실정이다. 한국품질경영학회 창립 50주년을 맞이하여 품질경영과 관련된 논문 106편을 수집하고 “품질경영 이론”, “품질경영 실증연구”, “품질경영상 연구”, “품질경영 기타연구”, “ISO 인증시스템”, “품질보증”의 6개 분야로 나누어 리뷰하거나, 논문 초록에 나타난 최다출현 단어들의 단순 빈도분석만 수행되었다[11, 28].

산업공학 외에 다양한 학술 분야에서도 토픽모델링 연구가 활발히 진행되었다. 수학, 문헌정보학, 기록 관리학, 언어치료, 재난 및 안전관리, 한국문화정책, 대학구조개혁평가의 쟁점 분석, 청소년 문제, 게임 및 뷰티, 컨설팅, 과학 논문 등의 분야가 연구되었다[7, 17, 19, 20, 21, 23, 34, 37, 44]. 위의 연구를 수행하기 위한 자료들은 직접 크롤링(crawling)하거나 웹 스크래핑(web scraping) 프로그램으로 논문 제목, 초록, 주제어 등을 수집하였다. 대표되는 주제를 찾기 위해서는 LDA 분석이, 세부 주제별 미시적 핵심 키워드 도출에는 HDP 분석이 효과적임을 보였다[36]. 주제 발견에 이어서 연도별 추이, 활성화된 연구주제, 주제-방법 간의 네트워크를 제시하고, R, Mallet, R-Studio 프로그램 등을 사용하여 분석하였다[9, 27, 35].

3. 연구 프레임워크

3.1 연구개요

연구 프레임워크는 <Figure 1>과 같이 네 단계로 구성된다: 자료의 수집과 파일변환, 탐색적 자료 분석, 전처리, 토픽모델링과 시각화. 한국산업경영시스템학회지에 게재된 논문들을 수집하여 파일을 변환한 다음, 논문 제목과 초록을 대상으로 전처리 작업을 하여 코퍼스(corpus)를 완성한다. 여기서 코퍼스란 수집한 논문의 제목과 초록에 수록된 단어들이 전처리 작업이 끝난 상태일 때 이들의 집합을 의미한다. 코퍼스를 입력하여 토픽모델링 알고리즘으로 분석하고, 분석 결과의 시각화를 통하여 출현확률이 높은 단어들의 의미를 종합적으로 검토하여 가장 적합한 연구 주제를 부여한다.



<Figure 1> Topic Modeling Procedure

3.2 자료의 수집과 파일변환

주제 분석을 위해 1차적으로 학회 창립 이후 논문 제목과 초록만을 갖춘 전반기 22년간의 논문을 수집한다. 2절에서 살펴본 바와 같이 타 연구의 경우에는 “사용자 리뷰”나 “쉽게 수집 가능한 텍스트”들은 “크롤링” 또는 “웹 스크래핑” 프로그램을 실행하여 수집하였다. 그러나 본 연구의 경우에는 그러한 실행이 불가능하여 학회 홈페이지의 “학술 논문” 내의 “학회지 서비스”에 접속하여 연도별 게재 논문을 각각 수집하였다.

1978년부터 1999년까지 수집한 총 971편의 논문은 pdf 파일 형식이었으나 파이썬에서 읽어 들일 수 없었다. text 파일 형식도 토픽모델링 작업이 가능하지만, 데이터 프레임(data frame) 형식을 갖추고 행과 열로 구성되어 컴퓨팅 작업에 편리한 csv(comma separated values) 파일 형식으로 변환하였다. 그러나 파일변환 과정에서 단어들의 깨짐 현상이 많이 발생하여 단어들을 일일이 확인하며 파일변환을 완성하였다.

3.3 탐색적 자료 분석

탐색적 자료 분석은 토픽모델링에 필수적인 부분은 아니지만, 학회지의 전반기 역사를 확인한다는 관점에서 전반적인 학회지의 연도별 논문게재 현황을 제시한다. 그리고 분석할 자료에 대한 기본적 이해를 위하여. 게재 논문

의 제목과 초록에 나타난 문장, 단어, 문자들과 관련된 값들의 기술 통계량 값을 계산하고, 그 값들의 분포를 그래프로 작성한다.

3.4 전처리

토픽모델링은 전처리가 완성된 코퍼스로부터 통계적 방법으로 찾아낸 단어들의 조합에 잠재적 의미를 갖는 주제를 유추하는 문제이기 때문에 주제 발견에 방해가 되는 요소들을 제거하고 결과에 결정적인 영향을 미치는 전처리(preprocess)과정을 세밀하게 진행해야 한다. 자연어 처리를 위한 구체 적인 전처리 과정이 있으나, 논문의 주제를 찾기 위한 일반적인 전처리 과정은 다음과 같다[31]:

- 토큰화(tokenization): 문서를 단어로 나누는 것 (breaking documents into term components);
  파일변환과정을 통하여 논문의 주제와 초록의 문장에서 단어들을 수정하고 보완한다.
- 구두점 제거(discarding punctuation):
  . , : ; / ‘ ’ ′ & % < > ( ) [ ] * _ - 등을 제거한다.
- 불용어 제거(filtering out stop words):
  논문에 관용적으로 사용되며 주제 의미와 관련 없는 단어로서 주어, 서술어, 일반 동사, 명사, 형용사, 접속사, 전치사, 아라비아 숫자들은 불용어로 취급하여 제거한다. 예를 들면 다음과 같다;
  a, an, the, I, my, we, us, this, that, then, can, be, would, do, where, which, and, or, for, in, under, among, over, firstly, very, so, simple, general, part, main, same, current, suggest, develop, perform, conduct 등
- 최다 출현단어 제거(removing highly frequent terms):
  산업공학의 모든 주제에 공통으로 나타나는 단어로서 주제 결정에 혼란스런 단어들을 제거한다. 예를 들면 다음과 같다;
  system, model, analysis, method, design, performance, technique, management 등
- 희소 출현단어 제거(removing infrequent terms (relative pruning)):
  토픽모델링 분석과 pyLDAvis 시각화 결과, 제시되는 핵심 30개 단어에 가끔 출현하지만, 주제 결정에 기여하지 않으며 빈도수가 상대적으로 적은(약 3~4개 이하) 단어들을 제거한다. 예를 들면 다음과 같다;
  tendency, progressive, grip, nausea, brain, overload, satisfactory, double, seoul 등
- 어간(stemming)과 표제어추출(lemmatization):
  본 연구에서 어간과 표제어추출은 해당이 없어 생략하였다. 예를 들면 “organized”의 어간은 “organ”, 표제어는 “organize”가 되는데 모든 단어에 어간과 표제어를 생성하면 기존의 단어와 중복되어 오히려 주제 결정에 혼란을 일으키므로 생략한다. 그러나 분석 결과, 제시되는 단어들의 조합에서 같은 단어의 단수와 복수형태가 나타나는 것을 방지하기 위하여 정제화(cleaning)한다. 예를 들면 “decisions”을 “decision”으로 변환한다.

실제 알고리즘 실행 후에도 주제 결정에 방해되는 무의미한 단어들의 조합이 반복적으로 나타나는 상황이 발생할 수 있으므로 전처리 과정은 알고리즘 실행과 함께 충분히 반복하여 완성한다. 참고로 코퍼스의 단어들에 대한 bar chart를 작성하고 빈도를 확인하여 제거할 단어를 판단할 수 있다. 전처리 과정이 모두 끝나면 코퍼스 단어들의 word cloud를 작성하여 전체 단어들의 구성을 확인할 수 있다.

3.5 토픽모델링과 시각화

오픈소스 기반의 웹 애플리케이션인 쥬피터 노트북(Jupyter Notebook version 6.0.1)에서 파이썬(python 3.7.4)언어로 토픽모델링을 한다. 쥬피터 노트북의 강점은 코드 작성과 실행 결과를 즉시 확인할 수 있는 상호작용(interactive)기능과 결과에 대한 시각화가 매우 편리하기 때문이다. 전처리가 완성된 코퍼스를 입력 자료로 사용하여 2절에서 살펴본 LSA, HDP, LDA, LDA Mallet(Mallet으로 실행한 LDA) 알고리즘을 실행한다.

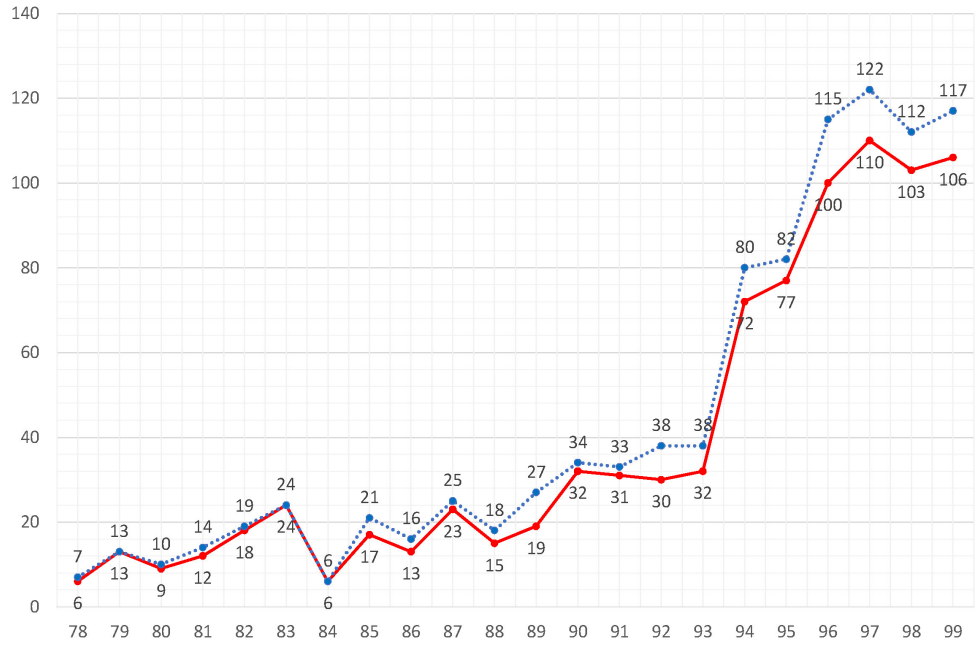
시각화에 가장 많이 활용되고, 토픽과 토픽의 구성 단어들을 쉽게 파악할 수 있는 pyLDAvis는 오픈소스의 파이썬 라이브러리로서 쥬피터 노트북에서 LDA 분석 결과를 보여주는 도구이다[41]. 그 밖에 각 토픽의 단어와 토픽을 각각 행과 열로 구성한 표를 제시하고, 주제에 속한 단어의 중요도를 다양한 원의 크기로 나타내는 Termite가 있다[10].

토픽모델링 분석 결과로부터 나오는 단어들의 조합은 결정적(deterministic)이 아니라, 통계적 모형에 근거한 확률적(stochastic) 계산의 결과이므로 주제를 정확히 결정하기 위해서는 산업공학 변화의 흐름에 대하여 경륜을 가진 전문가의 지식과 경험이 필수적이다[31].

4. 토픽모델링

4.1 게재 논문의 현황

22년간 수집한 논문의 연도별 현황을 <Figure 2>에 선



* Straight lines with numbers and dotted lines with numbers represent the number of research papers converted to csv file format and total number of research papers, respectively.

<Figure 2> The Number of Annual Research Papers

그래프로 작성하였다. 그림에서 점선과 숫자는 학회 홈페이지에서 수집한 연도별 게재 논문 수로서 총 971편을 의미한다.

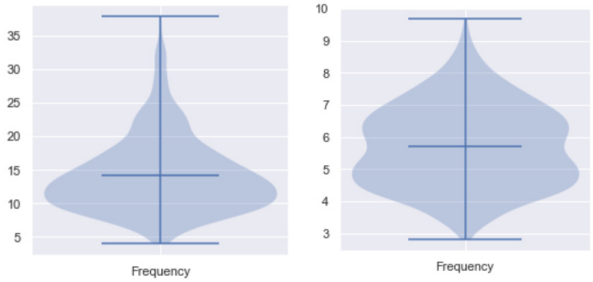
총 971편 가운데 초록이 없는 논문과 csv 파일로 파일변환이 불가능한 논문을 제외한 총 868편의 논문은 실선과 숫자로 나타내었다. 파일변환과정에서 깨짐 현상이 발생한 문자들을 모두 수정하였고, 토픽모델링을 위한 코퍼스의 입력 자료로 논문 제목과 초록을 사용하였다.

1978년부터 1993년까지는 연 1~2회 논문을 출간하여 논문 편수가 적은 편이지만 점진적으로 증가하고 있고, 1994년부터는 연 4~5회 출간함에 따라 논문 편수가 급격히 증가함을 볼 수 있다. 22년간 게재 논문 편수는 전체적으로 증가추세에 있음을 알 수 있다.

4.2 탐색적 자료 분석

868편의 논문 제목과 초록을 구성하는 문장, 단어, 문자 개수의 구체적인 분포를 알기 위하여 <Figure 3>과 <Figure 4>와 같은 바이올린 플롯(violin plot)으로 제시하였다. 바이올린 플롯은 상자 그림(box plot)과 비교할 때 자료 전체 분포를 확률밀도 형태로 보여줌으로써 최빈값을 포함하여 자료들이 몰려있는 봉우리의 위치와 각 자료들의 흩어진 상태를 쉽게 확인할 수 있다. <Figure 3>과 <Figure 4>에서 수평으로 보이는 세 개의 선분은 각각 최대값, 평균값, 최소값을 나타낸다.

<Figure 3>의 (a)는 논문 제목을 구성하는 단어(vocabulary) 개수의 분포로서 평균 단어의 수는 약 14개이고, (b)는 제목을 구성하는 단어들의 문자(character) 개수의 분포로서 봉우리가 2개이고 평균 문자의 수는 약 6개임을 보여준다. <Figure 4>의 (a)는 논문 초록을 구성하는 문장(sentence) 개수의 분포로서 각 논문의 초록에 포함된 평균 문장의 수는 약 6개이고, (b)는 논문 초록을 구성하는 단어(vocabulary) 개수의 분포로서 각 논문의 초록을 구성하는 평균 단어의 수는 약 143개임을 알 수 있다.

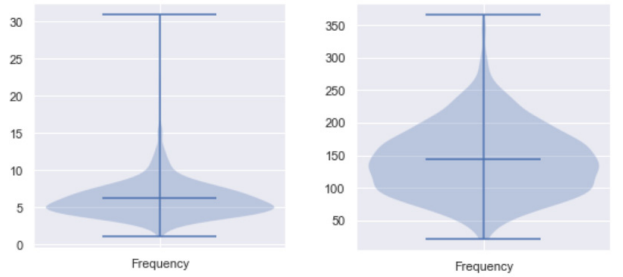


(a) The Number of Words (b) The Number of Characters in Title in a Word of Title

<Figure 3> The Distribution of Words and Characters in Title

전처리가 완료된 코퍼스내 단어들을 word cloud로 시각화함으로써 주제 분석에 활용될 단어들을 사전에 확인할 수 있다. <Figure 5>의 (a)와 (b)는 각각 논문 제목과 초록

에 대하여 전처리 과정을 마친 코퍼스내의 471개, 957개 단어로 생성된 word cloud의 형상이다. 그림에 나타난 단어의 크기는 코퍼스내 단어의 출현 빈도에 비례한다.



(a) The Number of Sentences in Abstract

(b) The Number of Words in Abstract

<Figure 4> Distributions of Sentences and Words in Abstract

(a) Vocabularies within Title

(b) Vocabularies within Abstract

<Figure 5> Word Clouds by Vocabularies within Title and Abstract

논문 제목의 코퍼스에서 가장 많이 나타난 상위 5개 단어는 quality(58개), algorithm(54개), production(44개), scheduling(41개), cost(41개)임을 <Figure 5>의 (a)에서 확인할 수 있고, 논문 초록의 코퍼스에서 가장 많이 나타난 상위 5개 단어는 cost(435개), quality(320개), algorithm(284개), product(234개), production(227개)임을 그림 (b)에서 볼 수 있다.

<Figure 5>의 (a)와 (b)로부터 단어와 크기와 색상이 서로 상이하기 때문에 육안으로 각 단어들을 대조하여 비교하는 것은 어렵지만, 크게 보이는 일부 단어들을 살펴볼 때 두 word cloud안에서 동일한 단어들이 동시에 나타나는 것을 확인할 수 있다.

4.3 주제의 개수 및 의미 결정 시 고려사항

전처리 다음으로 중요한 과정은 주제의 개수와 의미를 결정하는 문제이다. 이때 고려해야 할 기준을 우선순위에 따라 열거하면 다음과 같다:

① 대한산업공학회, 한국경영과학회, 한국연구재단 학술연구의 아래의 분류기준에 따라 결정한다;
확률/통계분야, 품질/신뢰성/품질경영분야, 최적화/휴리스틱스분야, 생산/물류/재고분야, 인간/안전/인지공학분야, 컴퓨터/ICT분야, 경영공학/금융공학/경제성공학분야, 기술경영/서비스경영/R&D분야, 4차 산업혁명/인공지능관련분야, 산업/공공/국방분야, 기타 응용분야 등

② 한 주제의 핵심단어들이 여러 주제에 출현할 때 중첩되는 단어가 최소가 되도록 결정한다;
주제 개수가 작으면 각 분야의 키워드들이 한 주제에 몰려서 나타나고, 주제 개수가 많으면 한 주제의 주요 키워드가 의미 없는 단어와 함께 여러 주제에 나타날 수 있다.

③ LDA분석의 경우 pyLDAvis 시각화 결과에서 각 주제를 의미하는 여러 개의 원들의 중첩이 최소가 되도록 결정한다.

④ 토픽모델링 결과, 평가 측도를 참고하여 결정한다;

- coherence(주제 일관성): 토픽모델링에 의해서 결정된 상위 단어들의 조합이 실제 문서들의 의미와 일치하는 정도를 점수화한 값으로서 높을수록 바람직하다[33].
- perplexity(혼란도): 토픽모델링에 의해서 계산된 가능도(likelihood)가 주제 개수에 따라 학습이 잘 되었는가를 보여주는 측도로서 작을수록 바람직하다[5].
  그러나 여러 토픽모델링 실험 결과, 위 측도들의 계산 결과 값들이 좋게 나오더라도 반드시 주제 개수와 의미부여 시 바람직한 결과를 주는 것이 아니므로 휴리스틱 접근법이 필요하다[2, 43].

따라서 학회의 분류 기준을 충분히 만족하도록 주제 개수를 5~15개로 정하고, 2.3절의 학술분야의 활용에서 주제를 결정하는 단어의 수를 통상 상위 5개 이상을 취하므로 본 연구에서는 상위 10개로 결정하였다. 여러 차례 알고리즘을 수행하고 비교한 결과, 주제의 개수가 8개 일 때

가장 적절하게 주제를 표현할 수 있었다. 2절로부터 LDA가 가장 바람직하다고 알려져 있으나 알고리즘 비교를 위하여 파이썬에서 지원하는 LSA, HDP, LDA, LDA Mallet의 Gensim 라이브러리를 사용하여 논문 초록과 제목의 코퍼스에 대하여 분석하였다. 초록의 코퍼스 단어가 957개로서 제목의 코퍼스 단어 471개 보다 많으므로 먼저 초록에 대하여 분석하고 제목의 분석 결과와 비교한다.

4.4 초록에 의한 토픽모델링 분석 결과

4.4.1 LSA 분석

LSA 알고리즘은 868편의 문헌을 행, 957개의 단어를 열로 하는 행렬을 특이값분해로써 차원을 축소하여 가중치가 큰 상위 단어들의 조합으로 주제를 결정하는 과정이다. 분석 결과 상위 10개 단어로 구성된 8개 주제는 <Table 2>와 같고, Topic 1의 10개 단어들이 다른 Topic에도 나타나는 경우에는 진하게 표시하였다. 8개 주제에서 공통으로 중복하여 발견되는 단어들이 지나치게 많으므로 각 주제에 적절하고 안정된 의미를 부여할 수 없었고, 주제 분석 방법으로 LSA가 적절하지 않음을 확인할 수 있다.

<Table 2> Topic Modeling Results Using Abstracts by LSA

Topic 1	Topic 2	Topic 3	Topic 4
cost	quality	**cost**	**process**
time	**time**	**process**	**quality**
process	**process**	**time**	**algorithm**
quality	**control**	policy	**control**
control	**algorithm**	**quality**	**time**
product	**cost**	**algorithm**	index
production	**product**	**machine**	**cost**
algorithm	**machine**	warranty	scheduling
manufacturing	scheduling	work	capability
machine	heuristic	failure	job
Topic 5	Topic 6	Topic 7	Topic 8
process	**time**	**control**	**product**
algorithm	**control**	**quality**	maintenance
factor	**algorithm**	**time**	distribution
product	**product**	chart	failure
cost	factor	**product**	**machine**
rate	chart	line	repair
accident	failure	distribution	service
human	**manufacturing**	decision	demand
industrial	function	work	center
information	distribution	assembly	accident

4.4.2 HDP 분석

HDP는 상위 Dirichlet process와 하위의 Dirichlet process를 구성하여 동시에 출현확률이 높은 상위의 단어의 조합으로 주제를 구성하는 방법으로서 분석 결과는 <Table 3>과 같다.

알고리즘들을 비교할 때 다른 알고리즘에서는 coherence값이 약 0.30～0.39인 반면, HDP에서는 0.60～0.69로 매우 큰 값을 보였다. 그리고 LDA 또는 LDA Mallet 분석에는 잘 나타나지 않았으나 HDP 분석에서는 주제 의미 결정에 매우 구체적인 소수의 단어들이 등장하여 <Table 3>에 진하게 표시하였다. 예를 들면 Topic 1의 capability (품질), Topic 2의 metal(생산/물류), Topic 3의 echelon(생산/물류), Topic 4의 depreciation(경제성공학), Topic 5의 ks(산업), fortran(컴퓨터/ICT), makespan(생산/물류), Topic 7의 carlo(컴퓨터/ICT), squared(확률/통계), Topic 8의 dea (휴리스틱스), scrap(생산/물류) 등이다.

즉, 주제의 의미를 반영할 수 있는 지나치게 구체적인 소수의 단어들이 주제 의미를 잘 결정할 수 없는 다수의 단어들과 함께 나타나므로, 각 Topic의 주제의 의미를 결정하기가 어려웠다. 따라서 본 연구의 주제 분석에는 적절하지 않은 방법으로 판단된다.

<Table 3> Topic Modeling Results Using Abstracts by HDP

Topic 1	Topic 2	Topic 3	Topic 4
sector	machine	shift	**depreciation**
capability	price	**echelon**	alternative
range	union	interpretation	cooperative
loss	compensation	motion	availability
net	number	solve	check
minimize	cooperation	large	uniform
basis	health	minimal	flexibility
demand	marketing	stable	progress
weight	metal	personal	approach
regional	dimension	origin	deal
Topic 5	Topic 6	Topic 7	Topic 8
market	progress	culture	**dea**
ks	rd	promotion	inverse
mu	configuration	**carlo**	business
fortran	enterprise	factor	fund
condition	forecasting	jit	reliability
environmental	manufacture	destination	map
forecast	additional	rate	max
makespan	optimization	estimator	added
reengineering	direction	**squared**	**scrap**
constraint	deterministic	retrieval	membership

4.4.3 LDA와 LDA Mallet 분석

문헌과 주제가 주어질 때 나타날 단어의 확률을 최대로 하는 LDA와 Java 기반의 LDA Mallet으로 분석한 결과를 <Table 4>에 정리하였다. LDA를 비롯한 알고리즘은 확률적(probabilistic) 모형이므로 알고리즘을 실행할 때마다 출현단어와 확률이 약간씩 차이가 난다. LDA와 LDA Mallet의 coherence 값은 각각 0.323과 0.356으로 매우 유사하였다.

<Table 4>의 LDA 분석에서 출현확률에 따라 나타난 상위 10개 단어(Top 10 key words)를 주제의 의미를 결정하기 위한 변별력을 갖춘 키워드로 간주하여, 대한산업공학회, 한국경영과학회, 한국연구재단 학술연구의 분류기준에 따

라 단어의 연관성을 충분히 검토하여 주제의 이름(Topic name)을 결정하였다.

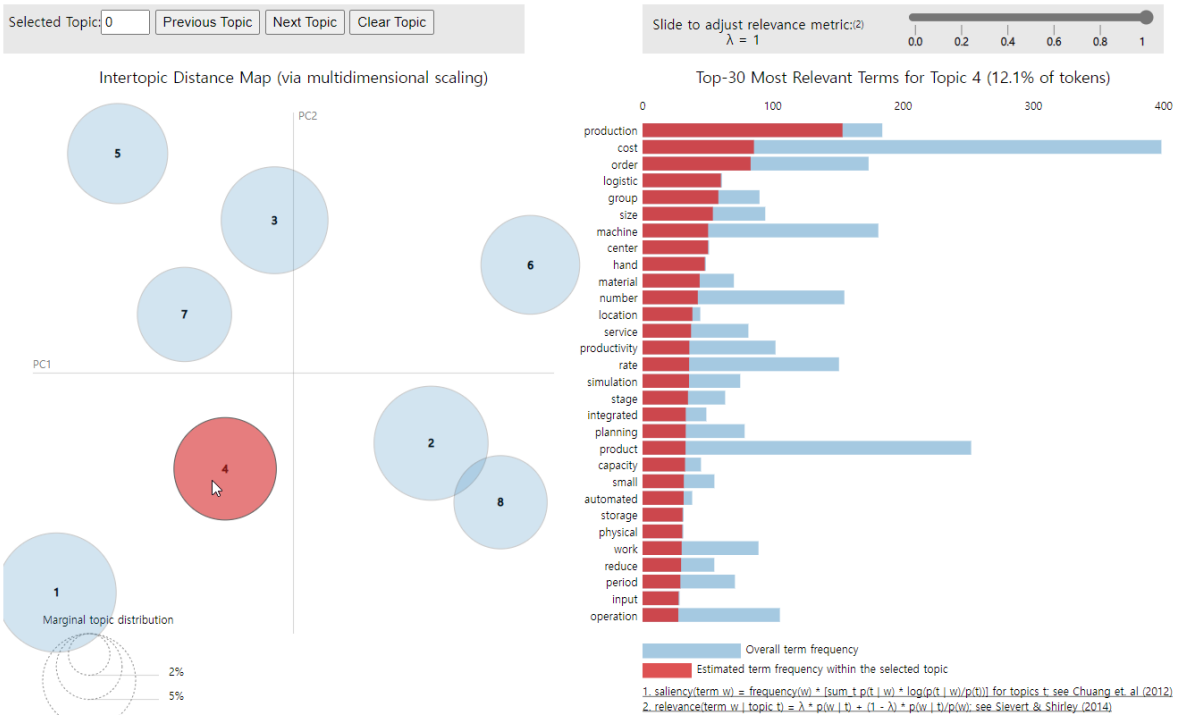
<Table 4>의 LDA 분석에서 상위 10개 단어의 출현확률의 크기에 따라 <Figure 6>에 word cloud를 작성하였다.

<Table 4> Topic Modeling Results Using Abstracts by LDA and LDA Mallet

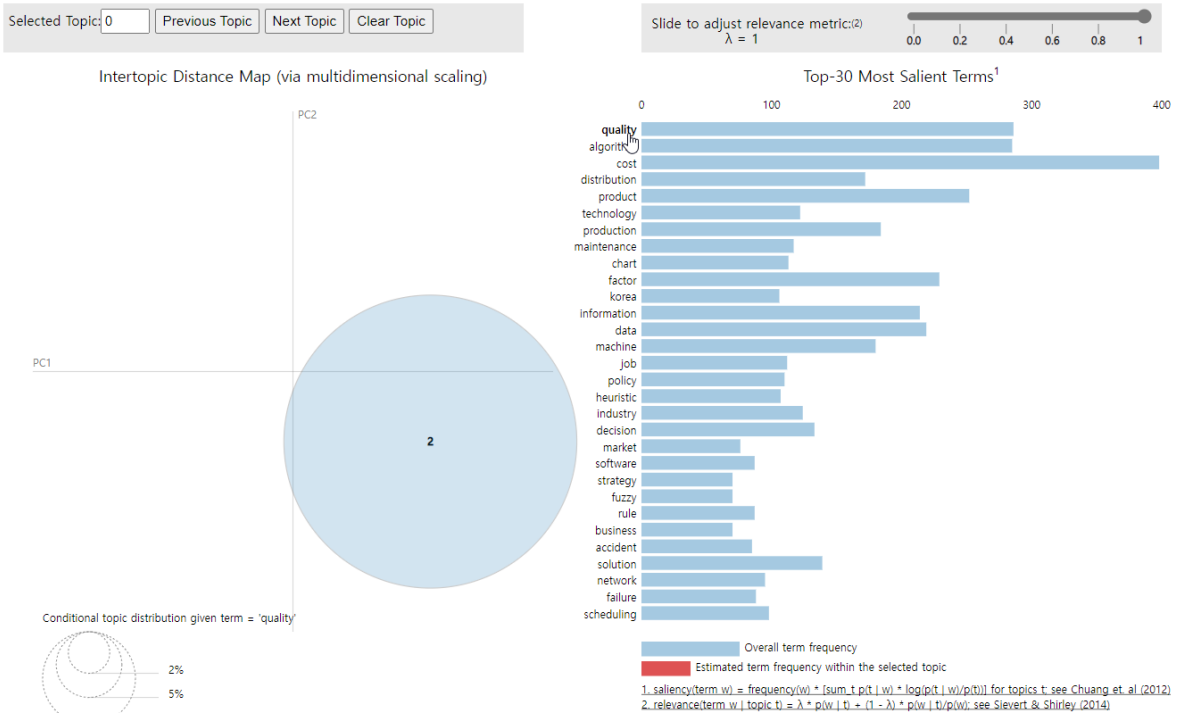
LDA								
No	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Top 10 key words	production cost order logistic group size machine center hand material	cost distribution chart policy total failure demand inspection warranty repair	quality product information factor software human safety index data work	data characteristics parameter tool test allocation database error experiment plan	korea industry market business information structure firm certification modeling domestic	technology factor strategy relationship decision enterprise resource investment change alternative	maintenance fuzzy decision project equipment programming rd solution knowledge arc	algorithm machine job heuristic network rule scheduling accident solution vehicle
Topic name	**Production/ Logistics, Inventory**	**Reliability**	**Quality, Computer/ICT, Human factor/ Safety**	**Probability/ Statistics**	**Management Engineering, Industry**	**Engineering Economy**	**Reliability, Probability/ Statistics, Optimization**	**Heuristics, Production/ Logistics**
LDA Mallet								
No	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Top 10 key words	production decision solution procedure number approach line size stage material	industry change industrial environment safety software accident strategy structure small	cost policy total maintenance failure minimize set tool unit period	order distribution rate variable service demand fuzzy programming probability customer	product information function data human characteristics vehicle application experiment estimate	factor work economic computer plan standard test program chart index	quality technology productivity group business enterprise activity korea situation relationship	algorithm machine job operation scheduling heuristic network processing rule simulation
Topic name	**Production/ Logistics, Optimization**	**Management Engineering, Human factor/ Safety**	**Reliability**	**Probability/ Statistics, Inventory**	**Human factor/ Safety, Industry**	**Engineering Economy, Computer/ICT**	**Quality, Industry**	**Heuristics, Production/ Logistics**

<Figure 6> Word Clouds Using Abstracts by LDA Analysis Result

파이썬에서 숫자의 연산은 0부터 시작하므로 <Figure 6>의 Topic 0은 <Table 4>의 Topic 1을 의미한다. 예를 들면, <Table 4>의 Topic 1에서 최상위 단어인 “production”은 <Figure 6>의 Topic 0의 word cloud에서 가장 크게 나타났다. <Figure 7>은 토픽과 토픽의 구성 단어들을 쉽게 파악할 수 있도록 <Table 4>의 LDA결과를 pyLDAvis를 이용하여 시각화한 그림이다. LDA 분석에서 찾은 주제들이 그림 (a)의 왼편에 보이는 주제거리지도(Intertopic Distance Map)에



(a) Bubble 4 and its Related Top-30 Most Salient Terms



(b) Bubble 2 Represented by Term “Quality”

<Figure 7> pyLDAvis Visualization

8개의 원(Bubble)으로 표시되었다.

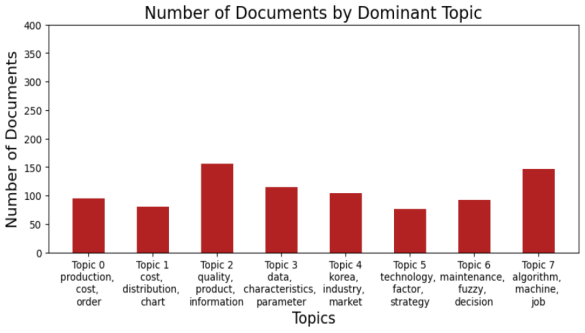
원이 클수록 주제의 비율이 크고, 원들이 겹치지 않고 떨어져 있을수록 주제 의미가 서로 상이함을 의미한다. 현재는 마우스를 그림 (a)의 Bubble 4(<Table 4>의 LDA결과에서 Topic 1을 의미)에 위치하였으므로 <Table 4>의 Topic 1을 구성하는 10개의 단어를 포함한 상위 30개 단어(Top-30 Most Salient Term)가 빈도순으로 붉은 색의 Bar chart로 표시된 것을 확인할 수 있다. 이와 같이 마우스를 그림 (a)의 다른 Bubble로 옮기면 주제를 구성하는 30개 단어들이 변화하며 새로운 Bar chart가 작성되는 것을 볼 수 있다.

<Figure 7>의 그림 (b)에서 오른편 가장 위쪽에 있는 단어 “quality”에 마우스를 위치하면 왼편의 Bubble 2(<Table 4>의 Topic 3)만 크게 나기 때문에 단어 “quality”가 속한 Topic 3의 주제는 “품질”을 반영함을 알 수 있다. 즉, 오른편의 여러 개의 핵심단어들에 마우스를 이동해 보면 그 단어가 함의하는 주제들이 왼편의 하나 또는 여러 개의 원으로 나타나서 단어와 관련된 주제의 의미를 유추할 수 있다.

<Table 4>의 LDA 분석의 결과에 따라 Topic 1부터 8까지(파이썬에서는 0부터 7까지) 상위 3개의 단어들만 포함한 각 주제들을 <Figure 8>에 Bar chart로 작성하였다. Bar chart에 나타난 8개 주제에 의미를 부여하면 생산/물류와 재고(95개, 11%), 신뢰성(81개, 9%)과 품질과 컴퓨터/ICT와 인간공학/안전(156개, 18%), 확률/통계(115개, 13%), 경영공학과 산업(104개, 12%), 경제성공학(77개, 9%), 신뢰성과 확률/통계와 최적화(92개, 11%), 휴리스틱스와 생산/물류(148개, 17%)로 요약할 수 있다. 4.3절의 학회 및 학술연구의 분류 기준에서 산업공학의 주제는 많으나 주제 개수를 8개로 제한하였으므로 분석 결과, Topic안에 몇 개의 주제가 중복되었다.

여러 차례 LDA 알고리즘을 실행한 결과, 단어의 수를 10개, 주제의 개수를 8개로 하였을 때가 주제 의미를 부여하기가 가장 적합하였다. 주제의 수를 7개 이하, 9개 이상으로 했을 때는 특정 주제를 의미하는 핵심단어들이 그 주제와 관련이 적은 평이한 단어들과 함께 여러 주제에 중복 발견되어 학술연구 기준에 적합한 주제를 찾는데 어려움이 있었다.

<Table 4>의 LDA Mallet의 주제 분석 결과에서 보듯이 Java 기반으로 계산되는 알고리즘일 뿐, LDA 결과와 비교에서 큰 차이가 없음을 알 수 있다. LDA Mallet은 Jupyter Notebook에 설치도 어려우므로 온라인과 GitHub에서 오픈소스로 활용할 수 있는 라이브러리가 풍부한 LDA의 활용을 더 권장한다.



<Figure 8> Bar Chart by Topic Number Using Abstracts by LDA

4.5 제목에 의한 LDA와 LDA Mallet 분석 결과

논문 제목의 코퍼스 단어 471개를 사용하여 앞과 같은 방법으로 LDA와 LDA Mallet으로 분석한 결과는 <Table 5>와 같다. LSA와 HDP 분석은 초록의 분석과 유사한 결과가 나타나서 생략하였다. LDA와 LDA Mallet의 분석에서 coherence값은 각각 0.625과 0.670으로서 소수점 둘째 자리에서 근소한 차이가 나타났다. 이 값은 논문 초록을 이용하여 분석한 <Table 4>에서 계산된 0.323과 0.356보다 월등히 큰 값이지만 분석 결과 나타나는 상위 10개의 단어로부터 주제의 의미를 찾는데 특별한 장점을 발견할 수 없었다. 이는 4.3절에서 언급한 coherence값이 크더라도 반드시 바람직한 결과는 아니라는 연구의 내용과 일치하였다[2, 43].

논문 제목을 활용하여 LDA와 LDA Mallet로 분석한 경우에도 주제의 수를 7개 이하 또는 9개 이상으로 했을 때 적절한 결과를 얻지 못하였고, 8개일 때 가장 적합한 주제들을 결정할 수 있었다.

주제 분석의 타당한 근거를 찾기 위하여 868편의 논문 제목을 한 개씩 읽고, 주제를 결정하였다. 그리고 논문 주제(Topic)의 이름과 논문 편수(Frequency)를 <Table 6>에 정리하였다. 4.4.3절의 논문 초록을 활용하여 LDA와 LDA Mallet으로 분석한 결과인 <Table 4>와 4.5절의 논문 제목을 활용하여 LDA와 LDA Mallet으로 분석한 결과인 <Table 5>에서 주제의 이름(topic name)이 있는 경우 <Table 6>의 해당 주제에 부호로 표시하였다. 여기서 부호 ●, ◎, ○, ×는 분석 결과, 주제의 이름이 각각 세 번, 두 번, 한 번, 0번 나타났음을 의미한다.

<Table 6>으로부터 본 연구에서 결정한 바와 같이 단어의 개수를 10개, 주제의 개수를 8개로 정했을 때 토픽모델링을 위해 사용한 자료가 초록 이든 제목이든 주제 이름을 결정하는데 있어서 큰 영향을 받지 않으며, 사용한 알고리즘인 LDA와 LDA Mallet에도 큰 영향을 받지 않는 것을 확인할 수 있다.

<Table 5> Topic Modeling Results Using Titles by LDA and LDA Mallet

LDA								
No	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Top 10 key words	fuzzy data effect environment strategy flexible implementation business structure measure	vehicle number rate traveling computer enterprise project driving salesman resource	algorithm scheduling shop heuristic flow reliability procedure production network dynamic	human application industrial safety approach inventory market technique allocation logistic	quality economic information distribution chart factor automated plan iso industry	product technology machine maintenance cell job software construction fms size	cost production case ergonomic estimation simulation planning accident value organizational	decision industry korea determination small function policy item certification medium
Topic name	**Probability/ Statistics, Management Engineering**	**Engineering Economy, Computer/ICT**	**Heuristics, Reliability, Optimization**	**Human factor/ Safety, Inventory**	**Quality**	**Production/ Logistics**	**Human factor/ Safety**	**Industry**
LDA Mallet								
No	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Top 10 key words	safety factor business planning fms dynamic rate project logistic small	production distribution application sampling data computer human comparison estimation alternative	economic flexible item productivity selection constraint procedure measure expert empirical	quality cost policy information product technique replacement construction effective iso	strategy effect maintenance level industrial plan chart inspection automated failure	korea industry inventory technology software cell vehicle determining case stock	algorithm machine determination heuristic function programming network reliability linear assembly	scheduling decision fuzzy shop flow job approach line activity material
Topic name	**Human factor/ Safety, Production/ Logistics**	**Production/ Logistics, Computer/ICT**	**Engineering Economy, Production/ Logistics**	**Quality, Inventory**	**Reliability, Quality**	**Industry, Inventory**	**Heuristics, Reliability, Optimization**	**Production/ Logistics, Probability/ Statistics**

<Table 6> Topics Sorted by Research Paper Title

Topic	Frequency	Abstract		Title	
		LDA	Mallet	LDA	Mallet
Production/Logistics	175(20%)	◎	◎	○	●
Quality	107(12%)	○	○	○	◎
Human factor/Safety	92(11%)	○	◎	◎	○
Others	88(10%)	×	×	×	×
Management Engineering	72(8%)	○	○	○	×
Computer/ICT	69(8%)	○	○	○	○
Reliability	62(7%)	◎	○	○	◎
Probability/Statistics	55(6%)	◎	○	○	○
Optimization	45(5%)	○	○	○	○
Engineering Economy	36(4%)	○	○	○	○
Heuristics	33(4%)	○	○	○	○
Inventory	16(2%)	○	○	○	○
Industry	14(2%)	○	◎	○	○
Military	4(1%)	×	×	×	×
Total	868(100%)				

* The symbols, ●, ◎, ○, and ×, represent that topic name appeared three times, twice, once, and never appeared, respectively, as a result of each algorithm analysis.

5. 결 론

5.1 분석 결과의 요약

4절의 토픽모델링에서 논문의 초록과 제목을 주제 분석을 위한 입력 자료로 사용하여 LSA, HDP, LDA, LDA Mallet으로 분석하였다.

LSA 알고리즘은 첫 번째 주제에 해당하는 10개 단어의 조합이 나머지 주제에 지나치게 많이 중복되어 주제의 의미를 결정하기가 어려워서 주제 분석에 적절하지 않았다. HDP 알고리즘은 분석 결과의 평가 측도인 coherence 값은 크게 나왔으나, 특정한 주제를 정확하게 반영하는 단어들이 매우 적게 나타나는 반면, 산업공학에서 언급되는 일반적인 다수의 단어들과 함께 나타나서 최적의 방법은 아니라고 판단된다.

<Table 4>와 <Table 5>의 분석 결과를 바탕으로 작성한 <Table 6>으로부터 LDA와 LDA Mallet이 논문 초록 또는 제목을 활용한 토픽모델링에 적합한 방법으로 판단된다. 주제의 정확하고 합리적인 의미를 부여하기 위하여 4.3절에서 서술한 기존의 산업공학 관련학회와 한국연구재단 학술연구의 11개 기준을 근거로 하였다. 분석 결과 적절하게 나타난 8개 주제를 요약하면 “생산,물류,재고” / “신뢰성” / “품질” / “확률,통계” / “경영공학,산업” / “경제성공학” / “인간,안전,컴퓨터,ICT” / “휴리스틱스,최적화”로 정리할 수 있었다.

그리고 <Figure 7>의 pyLDAvis는 LDA 분석 결과를 평가할 수 있는 효용성이 매우 큰 기법임을 확인하였다. <Figure 6>의 주제별 word cloud와 <Figure 8>의 주제별 Bar chart도 토픽모델링의 결과를 요약할 수 있는 적절한 시각화 방법으로 보인다. 그러므로 주제 분석에 적절한 알고리즘으로 온라인 상에서 오픈소스로 활용할 수 있는 라이브러리가 풍부한 LDA를 추천한다.

5.2 연구의 의의 및 제한점

본 연구는 그동안 한국산업경영시스템학회지에 게재된 연구 논문들의 주제를 살펴보기 위한 기초 연구로서 학회의 전반기인 78년부터 99년까지 22년간 논문의 초록과 제목을 이용하여 네 개의 토픽모델링 알고리즘을 여러 차례 수행하고 최종적인 주제 분석 결과를 제시하였다. 특히, LDA 방법이 여러 논문에서 사용되고 있으나 분석 시작부터 마지막 요약까지 구체적인 절차를 세밀하게 서술한 연구는 찾을 수 없었다. 본 연구를 통하여 토픽모델링의 연구자들이 다양한 실무 분야에 활용할 수 있으리라 생각된다.

본 연구의 제한점으로는 학회의 전반기 22년간의 주제 분석만 진행하였다. 최근 인공지능을 포함한 4차 산업의 발전에 따라 산업공학 분야도 다양한 학문과 융합하고 있다. 향후 연구로서 2000년 이후 현재까지 한국산업경영시스템학회에 게재한 연구 논문도 본 연구에 추가하여 주제별 추이, 주제별 클러스터링, 활발한 연구 주제와 쇠퇴하는 연구주제의 변화 등에 관한 종합적인 연구가 진행되어야 할 것으로 판단한다. 그리고 본 연구에서 추천한 LDA 알고리즘의 파라미터 값, 주제 발견을 위한 단어 조합의 개수, 주제의 개수의 변화를 더욱 정교하게 조절하여 학회지의 주제 분석을 제시한다면 본 학회지의 위상을 높이는데 일조해야 할 것으로 사료된다.

Acknowledgement

This work was supported by a Research Grant of Pukyong National University(2021). We appreciate anonymous referees in commenting to improve the quality of our paper.

References

[1] Albalawi, Rania, Yeap, Tet. H., and Benyoucef, Morad., Using Topic Modeling Methods for Short-text Data: A Comparative Analysis, *Frontiers in Artificial Intelligence*, 2020, 3, pp. 1-14.

[2] Arun, R. Suresh, V., Mdahavan, C. E. V, and Murty, M. N., On Finding the Natural Number of Topics with Latent Dirichlet Allocations: Some Observation, *PAKDD*, Springer-Verlag, 2010, pp. 391-402.

[3] Barde, B. V. and Bainwad, A. M., An Overview of Topic Modeling Methods and Tools, *International Conference on Intelligent Computing and Control Systems*, 2017, ICICCS, pp. 745-750.

[4] Blei, D. M., Ng, A. Y., and Jordan, Michael, I., Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 2003, Vol. 3, pp. 993-1022.

[5] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M., Reading Tea Leaves: How Humans Interpret Topic Models, *In Advances in Neural Information Processing Systems*, 2009, pp. 288-296.

[6] Cho, G. H., Lim, S. Y., and Hur, S., An Analysis of the Research Methodologies and Techniques in the Industrial Engineering Using Text Mining, *Journal of the Korean Institute of Industrial Engineers*, 2014, Vol. 40, No. 1, pp. 52-59.

[7] Cho, J. Y. and Cho, K. W., Topic Modeling on the

Adolescent Problem Using Text Mining, *Journal of the Korea Institute of Information and Communication Engineering*, 2018, Vol. 22, No. 12, pp. 1589-1595.

[8] Cho, S. G. and Kim, S. B., Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining, *Journal of the Korean Institute of Industrial Engineers*, 2012, Vol. 38, No. 1, pp. 67-73.

[9] Choi, J. W., Jang, J. J., Kim, D. H., and Yoon, J. H., Identifying Interdisciplinary Trends of Humanities, Sociology, Science and Technology Research in Korea Using Topic Modeling and Network Analysis, *Journal of Society of Korea Industrial and Systems Engineering*, 2019, Vol. 42, No. 1, pp. 74-86.

[10] Chuang, J., Manning, C. D., and Heer, J, Termite: Visualization Techniques for Assessing Textual Topic Models, http://www.researchgate.net/publication/254004974.

[11] Chung, K. S., Sin W. S., Baek, D. H., and Ju, Y. J., Review on the TQM Literature Appeared in KSQM, *Journal of Korean Society for Quality Management*, 2016, Vol. 44, No. 1, pp. 43-60.

[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 1990, Vol. 41, No. 6, pp. 391-407.

[13] Deisenroth, M. P., Faisal, A. A., and Ong, C. S., *Mathematics for Machine Learning*, Cambridge University Press, 2020.

[14] Dinakar, K., Chen, J., Lieberman, H., Picard, R., and Filbin, R., Mixed Initiative Real-Time Topic Modeling & Visualization for Crisis Counseling, *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015, pp. 417–426.

[15] Hearst, M., What is Text Mining?, SIMA, https://www.jaist.ac.jp/~bao/MOT-Ishikawa/Furt-herReadingNo1.pdf.

[16] Hofmann, T., Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, 2001, 42, pp. 177-196.

[17] Hong, J. L., Yu, M. R., and Choi, B. R., An Analysis of Mobile Augmented Reality App Reviews Using Topic Modeling, *Journal of Digital Contents Society*, 2019, Vol. 20, No. 7, pp. 1417-1427.

[18] Jeong, B. K. and Lee, H. Y., Research Topics in Industrial Engineering 2001~2015, *Journal of the Korean Institute of Industrial Engineers*, 2016, Vol. 42, No. 6, pp. 421-431.

[19] Jin, M. and Ko, H. K., Analysis of Trends in Mathematics Education Research Using Text Mining, *Journal of Korean Society Mathematical Education Series E*, 2019, Vol. 33, No. 3, pp. 275-294.

[20] Kim, J. E. and Baek, S. G., Analysis of Issues on the College and University Structural Reform Evaluation Using Text Big Data Analytics, *Asian Journal of Education*, 2016, Vol. 17, No. 3 pp. 409-436.

[21] Kim, M. K., Lee, Y., and Han, C. H., Analysis of Consulting Research Trends Using Topic Modeling, *Journal of Society of Korea Industrial and Systems Engineering*, 2017, Vol. 40, No. 4, pp. 46-54.

[22] Kim, S. K. and Jang, S. Y., A Study on the Research Trends in Domestic Industrial and Management Engineering Using Topic Modeling, *Journal of the Korea Management Engineers Society*, 2016, Vol. 21, No. 3, pp. 71-95.

[23] Kim, S. Y., Analysis of Research Trends in SIAM Journal on Applied Mathematics Using Topic Modeling, *Journal of the Korea Academia-Industrial Cooperation Society*, 2020, Vol. 21, No. 7, pp. 607-615.

[24] Korean Society of Industrial and Systems Engineering, http://www.ksie.or.kr.

[25] Landauer, T. K., Foltz, P. W., and Laham, D., An Introduction to Latent Semantic Analysis, *Discourse Processes*, 1998, Vol. 25:2-3, pp. 259-284.

[26] Langley, P., Selection of Relevant Features in Machine Learning, *AAAI Technical Report FS-94-02*, 1994, pp. 127-131.

[27] Lee, K. H., Jung, H. J., and Song, M., Weighted Subject - Method Network Analysis of Library and Information Science Studies, *Journal of the Korean Society for Library and Information Science*, 2015, Vol. 49, No. 3, pp. 457-488.

[28] Lee, S. B., Analysis of Research Trends in Journal of Korean Society for Quality Management by Text Mining Processing, *Journal of Korean Society for Quality Management*, 2019, Vol. 47, No. 3, pp. 597-613.

[29] Mashey, J., Big Dat and the Next Wave of Infrastress, https://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

[30] McCallum, A. K., MALLET: A Machine Learning for Language Toolkit, 2002, http://mallet.cs.umass.edu/about.php.

[31] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Jaussler, T., Schmid-Petri, H., and Adam, S., Applying LDA Topic Modeling in Communication Research:

Toward a Vaid and Reliable Methodology, *Communication Methods and Measures*, 2018, Vol. 12, No. 2-3, pp. 93-118.

[32] Mulunda, C. K., Wagacha, P. W., and Muchemi, L., Review of Trends in Topic Modeling Techniques, Tools, Inference Algorithms and Applications, *The 5th International Conference on Soft Computing and Machine Intelligence*, 2018, pp. 28-37.

[33] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T., Automatic Evaluation of Topic Coherence, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2010, June, pp. 100-108.

[34] Park, I. C., Kim, S. H., and Yoon, B. U., Technology Clustering Using Textual Information of Reference Titles in Scientific Paper, *Journal of Society of Korea Industrial and Systems Engineering*, 2020, Vol. 43, No. 2, pp. 25-32.

[35] Park, J. H. and Song, M., A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling, *Journal of the Korean Society for Information Management*, 2013, Vol. 30, No. 1, pp. 7-32.

[36] Park, J. H. and Oh, H. J., Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: Focused on LDA and HDP, *Korean Library And Information Science Society*, 2017, Vol. 48, No. 4, pp. 235-258.

[37] Park, S. U. and Lee, B. R., Trend Analysis of Korean Cultural Policy Studies Using Text Mining, *The Korean Governance Review*, 2017, Vol. 24, No. 3, pp. 95-119.

[38] Ramage, D., Rosen, E., Chuang, J., Manning, C. D., and McFarland, D. A., Topic Modeling for the Social Sciences, *NIPS Workshop*, 2009, pp. 1-4.

[39] Rehurek, R. and Sojka, P., Software Framework for Topic Modelling with Large Corpora, *The LREC Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45-50.

[40] Seo, H. B. and Lee, H. Y., PSS Research Trend, *Proceeding of Spring Conference in the Korea Society for Simulation*, 2017, pp. 997-1017.

[41] Siever, C. and Shirley, K. E., LDAvis: A Method for Visualizing and Interpreting Topics, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63-70.

[42] Teh, Y., Whye, J., Michaerl, B., Mattew. J., and Blei, D. M., Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, 2006, Vol. 101, pp. 1566-1581.

[43] Vayansky, I. and Kumar, S. A. P., A Review of Topic Modeling Methods, *Information Systems*, 2020, Vol. 94, pp. 1-15.

[44] Yoon, S. Y. and Yoon, D. K., A Trend Analysis on Disaster and Safety Management Using Topic Modeling, *Journal of the Korean Society for Geospatial Information Science*, 2017, Vol. 25, No. 3, pp. 75-85.

[45] Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W., A Heuristic Approach to Determine an Appropriate Number of Topics in Topic Modeling, *BMC Bioinformatics*, 2015, Vol. 16, No. S8, pp. 1-10.

ORCID

Dong Joon Park | https://orcid.org/0000-0003-0554-1378
Hyung Sool Oh | http://orcid.org/0000-0001-6341-8007
Ho Gyun Kim | http://orcid.org/0000-0002-7695-3348
Min Yoon | https://orcid.org/0000-0002-6124-9163