Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 4 pp. 26-33, 2022

https://doi.org/10.5762/KAIS.2022.23.4.26
ISSN 1975-4701 / eISSN 2288-4688

LDA를 활용한 Batch 공정 모니터링 연구의 토픽 모델링 분석

조현우
대구대학교 기계공학부

LDA-based Topic Modeling and Analysis on Batch Process Monitoring

Hyun-Woo Cho
Division of Mechanical Engineering, Daegu University

요 약 이 연구의 목적은 고부가가치 제품 생산에 사용되는 batch 조업에 대한 공정 모니터링 연구 논문들의 주요 연구 토픽과 트렌드를 분석하는 것이다. 1990년부터 2021년 9월까지 해외 저널에 게재된 1,196편의 논문 중 353편을 최종 선별하여 논문의 제목, 키워드, 초록 데이터를 수집하였다. 이 텍스트 데이터에 대하여 파이썬과 텍스트 마이닝의 토픽 모델링 방법인 잠재 디리클레 할당(latent Dirichlet allocation, 이하 LDA) 알고리즘을 활용하여 토픽 모델링 분석을 수행하였다. LDA 모델에서 사용될 최적 토픽 수 선정을 위해 일관성 지표를 계산하고 이 중 최대값(0.332)을 가지는 6개의 토픽을 도출하였다. 이들 토픽은 품질 예측 모니터링(34.8%), 다중 단계 감지(15.6%), 신규 알고리즘 (15.0%), 분광학 데이터 모니터링 (13.6%)의 비중 순서로 구성되었다. 토픽 디스턴스 맵 분석 결과 6개 토픽간 유사성은 크지 않고 명확하게 구분되었다. 향후 batch 공정 모니터링 후속 연구 관점에서는 다양해지며 복잡도가 높아진 batch 공정에 대응할 수 있는 분광학 데이터 활용 연구와 다중 단계 batch 조업에 대한 품질 예측과 이상 감지 연구가 필요할 것이다.

Abstract This research aims to analyze the research topics and their trends in batch process monitoring articles focused on high value-added products. Three hundred fifty-three articles out of the 1,196 articles published from 1990 to September 2021 were collected with titles, keywords, and abstracts. This text data was analyzed using Python and the latent Dirichlet allocation (LDA) algorithm. The consistency scores were calculated to select the optimal number of topics for the LDA model, resulting in 6 topics with the maximum score of 0.332. These 6 topics include quality prediction monitoring (34.8 %), multiphase detection (15.6 %), novel algorithm (15.0 %), spectrum-based monitoring (13.6 %), etc. These six topics have relatively low similarities, as found from the inter-topic distance map. The research concludes that it will be necessary to conduct follow-up studies on spectroscopic data for the various and complex batch operations and quality prediction and detection for multiphase batch processes.

Keywords : Batch Process, LDA, Monitoring, Text Mining, Topic Modeling

1. 서론

Batch 공정은 국내 주력 수출 품목인 반도체, 디스플레이, 고분자, 바이오 등 첨단 제품의 조업 방식으로서 시장 수요 변화에 유연하게 대처가 상대적으로 용이해 다품종 소량 생산에 적합한 장점을 가지고 있다. Batch 공정은 규모의 경제에 맞는 대량생산 방식의 연속 공정(continuous process)과 다르게 사전에 정해진 조건에

*Corresponding Author: Hyun-Woo Cho(Daegu Univ.)
email: hwcho@daegu.ac.kr
Received January 13, 2022 Revised February 22, 2022
Accepted April 1, 2022 Published April 30, 2022

맞는 양과 순서로 원료 물질이 투입된 후 일정 조건에서 생산되어 최종품이 나오는 하나의 batch 사이클이 완성되며 이를 반복적으로 수행하게 된다[1].

이러한 batch 공정의 품질과 생산성을 확보하기 위해서는 공정 내에 발생된 비정상적 상황을 실시간으로 감지(detection)하여 원인을 진단(diagnosis)하는 공정 모니터링(monitoring)이 필수적이라 할 수 있다. 이를 위하여 수학적 모델이나 전문가 지식에 기반한 방법이 제안되었으나 측정 변수 간 상관성이 크며 비선형성과 유한한 조업시간 등 batch 공정의 특성으로 인하여 제한된 결과를 보여주었다[2-4]. 이의 대안으로서 공정에서 측정되는 데이터에 기반한 다변량 통계적 공정 모니터링(multivariate statistical process monitoring)이 1990년대에 등장한 이후 머신 러닝과 인공 지능 분야의 다양한 방법론을 받아들여 batch 공정 모니터링의 성능을 개선하고 그 적용 범위를 확장하고 있다[5].

단변량 SPC(univariate statistical process control, SPC)이 지닌 한계를 극복하기 위해 주성분 분석(principal component analysis, PCA)이나 부분 최소 제곱법(partial least squares, PLS)이 batch 공정 모니터링 분야에서 활용되었다. 이러한 다변량 기법들은 batch 공정의 상관성이 강한 측정 데이터로부터 잠재된 독립 성분(변수)를 인위적으로 생성하고 정상 조업과 비정상 조업 데이터에 대한 감시 및 진단 모델을 구축한 후 현재 조업되는 공정의 실시간 데이터를 독립 변수 공간으로 투영함으로써 모니터링 모델과의 비교를 통해 공정 건강성을 확인하게 된다. 변수선택이나 자세한 모니터링 절차는 본 논문의 범위를 벗어나므로 참고문헌에서 확인할 수 있다[2-4]. 한편, 연속공정의 2차원(변수, 시간) 데이터와 달리 3차원(변수, 시간, batch)인 batch 공정의 데이터에 적합하도록 multiway PCA와 multiway PLS이 제안되었다. 3차원의 데이터를 2차원으로 'unfolding' 하여 연속공정의 방법론을 그대로 batch 공정에 적용할 수 있으며, 이와 더불어 서포트 벡터 머신(support vector machine), 신경망(neural network), 커널 기반 비선형 기법, 웨이블릿(wavelet)이나 동적 시간 워핑(dynamic time warping) 등 다양한 방법론과 결합한 융합 연구가 batch 공정 모니터링 분야에서 현재까지도 진행되고 있다[2-7]. 이러한 다변량 기법이나 신경망 등의 인공지능 방법론의 장점은 모니터링 성능 향상은 물론 공정 센서의 빅데이터 활용이 가능하고 모델 구축과 유지가 용이하다는 점이다[7].

한편, 텍스트 마이닝은 소셜 미디어나 신문 등에서 생산되는 자연어로 이루어진 대용량의 비정형 텍스트에서 패턴이나 관계를 추출하여 가치있고 의미를 가지는 정보를 찾아내는 일종의 비정형 데이터 마이닝 기법이라 할 수 있다[8]. 다양한 비정형 데이터에 텍스트 마이닝을 적용함으로써 이를 군집화(clustering), 분류(classification), 시각화(visualization)등의 작업을 통해 상품이나 서비스의 선호도 파악, 수요자 감성 분석, 마케팅, 연구 주제 분석 등 다양한 분야에서 활용되고 있다[8-13]. 특히 토픽 모델링(topic modeling)은 문서 데이터의 핵심임에도 숨겨져 있으며 관측이 어려운 토픽을 자동으로 추출해내는 모형을 의미한다. 일반적으로 토픽 모델링 구현을 위해서는 지도 학습(supervised learning)이 아닌 비지도 학습(unsupervised learning)이 필요한데 LDA, 잠재 의미 분석(latent semantic analysis, LSA), 상관 토픽 모델(correlated topic model, CTM) 등의 여러 알고리즘이 제안되어 사용되고 있다. 확률 기반의 LDA와 달리 LSA는 행렬분해 기반의 방법론으로 문서의 유사도 계산에는 유리하나 알고리즘의 특성상 추가된 정보의 업데이트가 어려운 단점이 있으며 CTM은 토픽에 포함된 단어 간 상관성이 있을 때 유리하지만 문서의 토픽 분포에 있어서 로지스틱 정규 분포를 가정하는 단점이 있다[14]. 따라서 타 알고리즘 대비 사용의 편의성과 범용성에서 장점을 가지고 있는 LDA 기법이 범용성을 가지고 사용되고 있다. LDA는 대상 문서들에서 관측된 단어 패턴을 분석함으로써 문서에 존재하는 잠재 토픽들을 추출하는 확률적 모형이며 환경, 마케팅, 체육, 특허, 정보통신, 에너지, 음악 분야 등 여러 연구 분야에서 다양한 목적으로 활용되어 왔다[11-16]. 그러나 통계적 공정 모니터링과 batch 공정 관련 논문의 연구 주제 분석에 있어 토픽 모델링을 활용한 연구는 거의 진행된 바가 없다.

본 연구에서는 2021년 9월까지 주요 해외 학술지에 게재된 batch 공정 모니터링 연구 논문의 텍스트 데이터를 분석 대상으로 LDA에 기반한 토픽 모델링 분석을 수행하여 핵심 연구 주제를 파악하고자 한다. 분석 대상 해외 학술지는 batch 공정 모니터링 연구가 게재되는 대중적이며 인지도가 높은 chemometrics and intelligent laboratory systems, AIChE journal, journal of chemometrics, journal of process control, expert systems with applications, journal of quality technology, International journal of reliability, quality and safety Engineering을 포함하였다. 이들 학술지에 게재된 연구 논문들의 제목, 초록, 키워드를 수집하여 LDA에 기반한 토픽 모델링 분석을 통해 연구 토

픽 간 상호 연관성과 비중을 분석하고 토픽의 변화 추이를 분석하고자 한다. 이를 통하여 batch 공정 모니터링 분야에서 연구자들이 관심을 기울였던 연구 토픽은 무엇이며 과거의 연구 토픽 트렌드는 현재까지 어떻게 변화하였는지를 파악할 수 있다. 또한 토픽을 표현하는 핵심어와 도출된 연구 토픽에 대한 분석 결과는 batch 공정의 제조 분야는 물론 이를 응용한 여러 산업 분야에서의 미래 융복합 연구에 기초 분석 자료로 활용될 수 있을 것이다.

2. 방법론

LDA 모델은 수집된 대상 문서들에 존재하는 토픽을 찾아가는 토픽 모델링 알고리즘을 말하는데 여기서 문서는 여러 개의 토픽을 가지며 각 토픽들은 디리클레(Dirichlet) 분포를 따른다고 가정한다[17]. 전체 문서 수를 D, 전체 토픽 수를 K, d번째 문서의 단어 수를 N이라 할 때 LDA에서는 토픽의 단어 가중치와 문서의 토픽 가중치를 고려하여 d번째 문서 내 i번째 단어의 토픽 $z_{d,i}$이 j에 할당될 확률을 아래와 같이 구한다.

$$p(z_{d,i}=j|z_{-i},w) = \left\{\frac{n_{d,k}+\alpha_k}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)}\right\}\left\{\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}v_{k,j}+\beta_j)}\right\} \quad (1)$$

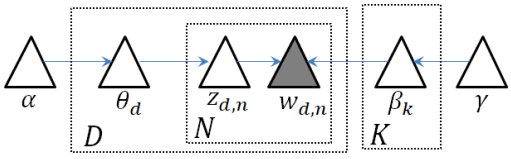

Fig. 1. A diagram of LDA model

LDA 모델의 과정을 단순화하여 그림으로 표현하면 Fig. 1로 나타낼 수 있는데 변수는 원으로 표현되며 화살표가 출발(도착)하는 원은 조건(결과)에 해당하는 변수를 나타내고 있다[17]. 유일하게 관측 가능한 변수는 d번째 문서의 n번째 단어 $w_{d,n}$으로서 음영으로 표현되어 있는데 우리는 이 정보만을 이용하여 α와 γ를 제외한 나머지 잠재 변수들을 추정하게 된다. α는 d번째 문서의 토픽의 비율을 나타내는 θ_d를 결정하는 디리클레 분포의 파라미터이며, γ는 k번째 토픽 단어들의 분포를 나타내는 β_k값을 결정하는 파라미터이다. d번째 문서의 토픽 비율 θ_d에 따라 d번째 문서의 n번째 단어의 토픽 확률 $z_{d,n}$이 결정되고 최종적으로 토픽 내 단어 분포 β_k와 $z_{d,n}$로부터 $w_{d,n}$을 구하게 된다. LDA 모델은 특정 문서가 특정 토픽에 속하는 정도를 표현해주는 일종의 클러스터링이라 할 수 있으며 클러스터링 분석에서처럼 토픽의 수를 사전에 지정해주어야 한다. 최적의 토픽 수 선택을 위한 지표로써 일반적으로 혼란도(perplexity)와 일관성 (coherence) 지표를 사용하고 있다[17].

Table 1. Journals and the number of articles

No.	Journal	Articles	
		1st	2nd
1	Chemometrics and Intelligent Laboratory Systems	220	162
2	AIChE Journal	370	64
3	Journal of Chemometrics	101	60
4	Journal of Process Control	323	51
5	Other three journals	182	18

본 연구의 LDA 모델링에 사용된 데이터는 batch 공정의 모니터링 논문으로서 2021년 9월까지 인지도와 영향력이 높은 7개 저널에 게재된 353편의 제목, 초록, 키워드를 대상으로 하였으며 이를 Table 1에 나타내었다. 각 저널별로 1차 키워드 검색을 통해 수집된 1,196편에 대해 보다 정확한 대상 논문 선정을 위하여 논문을 검증하는 2차 스크린을 거쳐 최종적으로 353편을 수집하였다. Chemometrics Intelligent Laboratory Systems 저널의 경우 관련 논문이 활발하게 게재되는 저널로서 논문 220편중 162편이 분석 대상에 포함되었다. 이와 같이 수집된 분석 대상 논문들을 게재 시기별로 살펴보기 위해 논문 비율을 Fig. 2에 나타내었다.

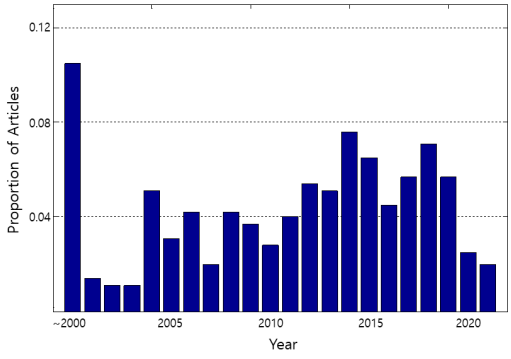

Fig. 2. A plot for number of articles analyzed

Fig. 2에서 '~2000'은 관련 연구 초창기인 1990년부터 2000년 동안 게재된 대상 논문들에 대한 비율을 표시하고 있는데 Fig. 2에서 보이듯이 2000년대 이후 지속적으로 batch 공정의 모니터링 관련 게재 논문이 늘어나고 있다. 특이한 점은 상대적으로 2020년과 2021년에 급격한 게재 논문수의 감소가 관찰된다는 것이다. 2020년 시작된 팬데믹 상황과의 연관성은 확인할 수 없으나 산업현장의 실제 이슈를 주로 다루는 batch 공정 모니터링 분야의 특성을 고려할 때 사회적 거리두기가 논문 게재에 영향을 주는 것으로 판단된다.

이렇게 수집된 문서 데이터를 바탕으로 본격적인 LDA 토픽 분석을 실행하기에 앞서 분석 시간을 줄이고 토픽 모델의 정확성을 높이기 위한 목적으로 텍스트 데이터에 대한 전처리(preprocessing) 작업을 수행하였다. 형태소에 따라 단어 분리가 다르기 때문에 분석에 필요한 단위로 단어를 분리하는 토큰화(tokenization)를 전처리 과정의 첫 단계로서 진행하였으며 파이썬의 Konlpy 한국어 자연어처리 라이브러리 내 형태소분석기를 활용하였다. 또한 분석에 불필요한 단어를 삭제하는 불용어 제거를 수행하였는데 관사나 전치사 대명사 등을 우선적으로 제거하였다. 그리고 논문이라는 대상 문서 데이터의 특성상 자주 나타나지만 토픽 모델링과는 관련성이 낮은 'paper', 'study' 등의 단어를 불용어 리스트에 포함해 제거하였다. 마지막으로는 단어중에서 형태는 다르지만 같은 의미를 가지는 서로 다른 단어들을 하나의 표제어로 정규화시키는 표제어 추출(lemmatization) 과정을 거쳤다.

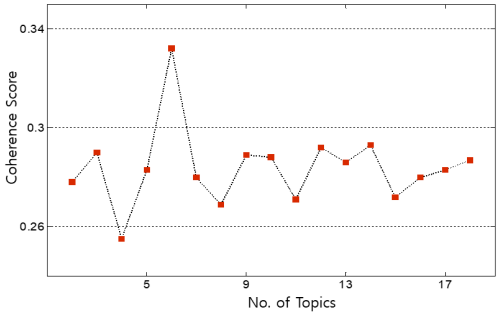

Fig. 3. Coherence score plot

일반적으로 토픽 모델링의 결과는 사전에 설정한 토픽의 수에 따라 다르게 나타날 수 있기 때문에 앞에서 언급하였듯이 LDA 분석 전에 토픽 수를 결정하게 된다. 이를 위하여 본 연구에서는 특정 모델이 대상 문서를 적절히 학습한 사실만을 나타낼 뿐 실제 해석 시 한계를 가진 혼란도 지수보다는 일관성 지수에 기반하여 토픽 수를 결정하였다[7]. 일관성 지수의 장점은 실제 해석에 적합한 척도로서 수치가 높을수록 LDA 분석 결과인 각 토픽이 높은 일관성을 가지며 유사한 단어들로 구성되어 있다는 것이다. 본 연구에서는 최적의 토픽 수 선정을 위해 토픽 수에 따른 일관성 지수를 산출하였다. 토픽의 수에 따른 일관성 지수를 나타낸 Fig. 3을 보면 최대의 일관성 지수 0.332를 달성한 6개로 토픽의 최적 갯수를 선정하였다.

3. 토픽 모델링 결과

2021년 9월까지 대상 저널에 게재된 353편의 논문 제목, 키워드, 초록에 대한 전처리 과정이 완료된 이후 핵심어에 대한 빈도 분석을 수행하였다. Table 2에 그 결과로서 토픽 모델링 대상인 논문들의 핵심어에 대하여 빈도가 높은 순으로 상위 20위까지의 핵심어 단어를 표시하였다. 핵심어 단어를 빈도순으로 보았을 때 'batch', 'method', 'data'가 각각 677회, 641회, 637회로 가장 많으며 다음으로는 'model', 'variable', 'monitoring', 'fault', 'analysis' 등이 사용되었다.

Table 2. Top 20 keywords

No.	Keyword	Freq.	No.	Keyword	Freq.
1	batch	677	11	difference	190
2	method	641	12	information	174
3	data	637	13	time	171
4	model	590	14	component	161
5	variable	386	15	phase	158
6	monitoring	345	16	performance	157
7	fault	319	17	detection	156
8	analysis	271	18	algorithm	149
9	approach	235	19	control	143
10	quality	200	20	PLS	142

Batch 공정의 통계적 공정 모니터링 논문에 대한 2021년까지의 텍스트 데이터로부터 토픽과 핵심 단어를 도출하기 위하여 LDA 토픽 모델링을 수행한 결과를 Table 3에 나타내었다. 토픽 수 선정을 위해 토픽의 수 변화에 따른 일관도 지수에 기반하여 결정된 6개 토픽에 대하여 해당 토픽을 표현하는 고빈도 핵심 단어 10개씩을 추출하였으며 이러한 토픽별 핵심 단어들의 연관성을 바탕으로 토픽의 이름을 결정하였다. 토픽 1은 'fault',

‘analysis’, ‘monitoring’, ‘PLS’, ‘quality‘ 등으로 이루어져 있으며 관심 대상인 품질 특성치를 예측하고 이를 바탕으로 공정의 건강상태를 모니터링하는 주제이다. 공정의 성능 지표로서 중요한 품질 변수를 batch 조업 중에 실시간으로 예측하여 공정의 이상 여부를 감시하게 된다. 물리적 센서 대비 데이터에 기반한 소프트 센서(soft sensor)를 구축하거나 batch 공정에 적합한 선형 결합 및 비선형 PLS 및 예측 기법 등에 관한 연구들로 구성된다.

토픽 2는 ’phase‘, ’monitoring’, ’difference‘, ’variable’, ‘information’ 등으로 이루어져 있으며 다중 단계(multiphase) batch 조업의 이상 감지와 관련한 주제이다. 바이오, 고분자 공정 등에서 자주 나타나는 다중 단계 batch 조업에 대하여 가우시안 혼합 모델 등을 활용하여 각 단계의 거동 및 전이(transition)구간과 단계 간 차이점을 모델링한다. 이를 기반으로 실시간 동기화(synchronization)하거나 품질 예측 또는 이상 감시 연구들로 구성된다.

Table 3. Results of topic modeling

No.	Topic	Ratio	Top 10 words in each topic
1	PLS-based quality monitoring	0.348	fault, analysis, variable, quality monitoring, approach, detection, PLS, component, performance
2	Multiphase detection	0.156	phase, quality, monitoring, difference, variable, analysis, fault, information, prediction, case
3	Novel algorithm	0.150	variable, approach, monitoring, fault, new, time, quality, difference, analysis, parameter
4	Spectrum based monitoring	0.136	monitoring, analysis, phase, time, difference, reaction, prediction, spectrum, quality, approach
5	Nonlinear kernel methods	0.130	variable, monitoring, analysis, control, information, result, component, trajectory, fault, algorithm
6	Fault diagnosis	0.079	monitoring, algorithm, variable, fault, space, approach, local, result, industrial, information

토픽 3은 ‘variable’, ‘fault’, ‘new’, ‘time’, ‘parameter’ 등으로 이루어져 있으며 공정 모니터링 방법론의 새로운 알고리즘과 관련된 주제이다. 제안된 기존 방법론을 개선하는 새로운 알고리즘에 관한 연구 주제로서 SIMCA(soft independent modelling of class analogy), PARAFAC(parallel factor analysis), 동적 시간 워핑 (dynamic time warping), OSC(orthogonal signal correction) 등의 신규 기법들에 관한 연구로 구성된다.

토픽 4는 ‘monitoring’, ‘analysis’, ‘reaction’, ‘prediction’, ‘spectrum’ 등으로 이루어져 있으며 분광학 데이터에 기반한 모니터링과 관련된 주제이다. 근적외선(near infrared), 핵자기 공명(nuclear magnetic resonance), Raman spectroscopy 등의 분광학 데이터를 입력변수로 삼아 batch 공정의 반응 상수를 예측하거나 다중 규격(multi-grade) 또는 다중 제품(multi-product)을 생산하는 semi-batch 공정의 모니터링 등의 연구들로 구성된다.

토픽 5는 ‘variable’, ‘monitoring’, ‘component’, ‘trajectory’, ‘algorithm’ 등으로 이루어져 있으며 비선형 커널 방법론과 관련된 주제이다. 이상 감지 및 진단 성능의 향상을 목표로 다양한 비선형 및 커널 방법론과 공정의 빅데이터 내에 존재하는 불필요한 공정 변수를 제거하는 변수선택(variable selection)도 다루어지고 있다. 또한 이들의 모델 구조를 다중 블록(multi-block)이나 계층모델(hierarchical model)로 구성함으로써 예측적 모니터링(predictive monitoring)을 가능하게 하며 공정 변수간 상관(correlation)관계에 민감하지 않은 알고리즘에 관한 연구로 구성된다.

토픽 6은 ‘monitoring’, ‘algorithm’, ‘variable’, ‘local’, ‘information’ 등으로 이루어져 있으며 공정 이상 진단과 관련된 주제이다. 실시간 이상 감지의 다음 단계로서 이상의 원인이 되는 근본 원인이나 공정 변수 특정을 위해 out-of-control signal에 공정 변수가 영향을 미치는 정도를 MPCA, 선형분류기, 독립성분분석(ICA) 모델 등의 관점에서 기여도(contribution)로 수치화하기도 한다. 또한 과거 수집해 놓은 공정 이상 데이터가 있을 경우에는 서포트 벡터 머신 등의 분류(classification) 기법 적용을 통한 방법 등의 다양한 이상 진단 접근법으로 구성된다.

위와 같이 토픽 분석을 통해 도출된 6개의 토픽에 대하여 토픽 디스턴스 맵 (Inter-topic Distance Map) 분석 결과를 Fig. 4와 같이 얻을 수 있었다. 토픽의 비중과 토픽 간 거리를 나타내는 토픽 디스턴스 맵이 Fig. 4의 상단에 있으며 하단에는 상단에서 선택된 특정 토픽 (그림에서는 토픽 1)에 대한 주요 단어들을 보여준다. 상단의 토픽 디스턴스 맵은 각 토픽이 다른 토픽과 가지는 연관성과 유사도를 시각적으로 보여주는데 원의 중앙 숫자들은 도출된 6개 토픽을 표시하고 있다. 여기서 각 원들의 거리가 가까울수록 (멀어질수록) 토픽의 상호 유사성은 상대적으로 높다고(낮다고) 할 수 있다.

6개 토픽 중 1번과 3번 토픽 간 일부 중첩 영역이 그림에서 확인되나 나머지 토픽들은 겹치는 영역이 없어 유사성이 낮으며 명확하게 구분되어 있음을 알 수 있다. IDM에서 유일한 중첩 토픽은 예측에 기반한 공정 모니터링분야와 신규 알고리즘에 관한 주제인데 공통적으로 품질변수를 예측하거나 측정 데이터를 통한 모니터링 모델을 구성한다는 측면에서 두 토픽의 유사성을 이해할 수 있다. 본 연구에서는 토픽 1과 겹치는 토픽 3의 중첩 영역이 크지 않아 합치지 않았으나 만약 중첩된 공통 영역이 토픽 3 전체인 경우라면 두 토픽을 합쳐 재정의하는 것이 합리적일 것이다.

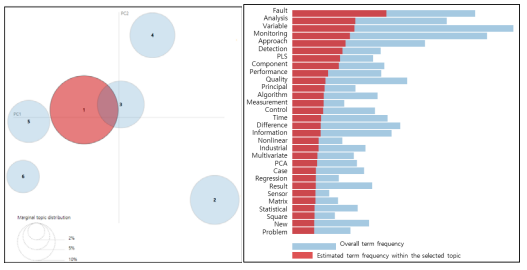

Fig. 4. IDM results

토픽의 시간에 따른 변화 추이를 살펴보기 위하여 6개 토픽별 논문 점유율을 논문이 게재된 연도별로 계산하여 그 결과를 Fig. 5에 나타내었다. LDA 토픽 모델링 결과를 통하여 개별 논문이 어떤 토픽에 소속되는지를 알 수 있으므로 이를 연도별로 모은 후 그 해 게재된 전체 논문 수 대비 각 토픽의 논문 수의 비율을 계산하였다. Fig. 5의 변화 추이를 살펴보면 품질 변수의 예측에 기반한 모니터링 (토픽 1)의 경우 대부분의 기간에 걸쳐 꾸준히 높은 비중으로 연구되고 있는 것을 알 수 있다. 다중 단계 batch 조업의 이상 감지 (토픽 2)의 경우 해당 점유율은 2005년까지는 굉장히 낮았으나 2006년 이후 꾸준히 높아지며 유지되고 있다.

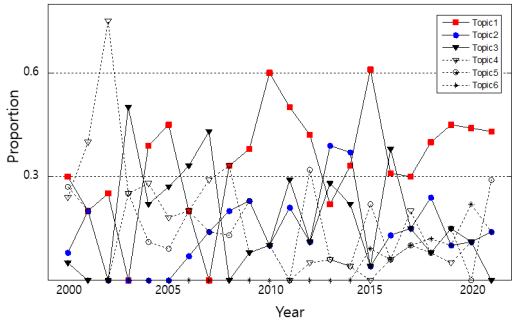

Fig. 5. A topic trend plot

이러한 토픽 1과 토픽 2의 변화 추이는 batch 공정 모니터링 연구 분야의 특성과 현재까지의 변화를 잘 반영해주는 것으로 보인다. 즉, 품질 예측과 이상 감지라는 토픽 1은 batch 공정 모니터링의 필수 요소로서 현업에서 가장 필요하며 다음 단계인 진단 성능에 영향을 준다는 측면에서 과거는 물론 미래에도 비중 있게 다루어질 수밖에 없을 것이다. 토픽 2인 multiphase batch 조업 감시 연구 트렌드는 2000년 중반이후 고부가가치의 다품종 소량생산 방식이 바이오, 제약, 반도체 등의 분야에서 증가함에 따라 batch 공정의 복잡도가 높아진 상황과 이를 해결하기 위한 논문의 게재 증가와 관련된 것으로 판단된다.

한편, 일반적인 추이를 보이는 다른 토픽들과는 다르게 토픽 4인 분광학 데이터 모니터링의 경우에는 2007년까지 높은 비중으로 다루어지다 감소하고 있음을 알 수 있다. 이는 Batch 공정의 복잡도가 크지 않았던 과거 공정에 분광학 데이터 기반 연구를 진행하던 흐름이 복잡도가 높은 나노, 바이오 분야의 batch 공정의 등장으로 기존 방법론의 한계를 드러내 새로운 전환점이 필요한 상황으로 판단된다. 이들 분야에서는 초미세공정에 따른 초정밀도와 대규모 공정장치가 요구되므로 연속공정 대비 batch공정의 중요도는 더욱 높아지게 될 것이다. 따라서 향후 이들 분야의 분광학 데이터 분석에 적합한 신규 방법론에 대한 연구와 나노, 바이오 분야로의 확대 적용이 필수적이다.

4. 결론

본 연구에서는 batch 공정에 적용된 통계적 공정 모니터링 연구에 관한 저널 논문에 대하여 텍스트 마이닝의 LDA 토픽 모델링을 기반으로 분석하였다. 2021년 9월까지 주요 학술지에 게재된 1,196편 중 353편 논문들의 제목, 주제어, 초록 데이터에 LDA 모델링 분석을 통해 일관성 지수 0.332에서 6개의 토픽을 도출하였다. 토픽은 품질 예측 모니터링(34.8%), 다중 단계 감지(15.6%), 신규 알고리즘 (15.0%), 분광학 데이터 모니터링 (13.6%)의 비중 순서로 구성되었다. Batch 조업이 가지는 반복적인 일회성 조업의 특성을 반영하여 34.8% 비중의 품질 예측 모니터링과 15.0%비중의 신규 방법론이 연구에서 많은 비중을 차지하였다. 다중 단계 batch 공정과 분광학 데이터를 활용한 연구의 중요성이 높아질 것으로 예측되는데 이는 고부가가치의 다품종 소량생산

방식의 확산과 비선형성과 복잡도가 커지는 신규 공정의 등장 등 산업적 필요성 관점에서 실제적인 연구의 중요성이 커지고 있는 상황과 깊게 연관되어 있다. 본 연구의 한계점으로는 국내 학술지가 제외되었는데 향후 분석의 범위를 확장시킨 추가 연구가 요구된다. 또한 특정 산업 분야, 예를 들어 반도체나 디스플레이 분야에 국한하여 모니터링은 물론 다양한 목적의 생산성 및 품질 향상 논문에 관한 토픽 모델링 분석 결과가 제시된다면 효율적인 공정 운용과 생산성 향상 관점에서 산업계에 실질적인 도움이 될 것으로 기대된다.

References

[1] B. Lennox, H. G. Hiden, G. A. Montague, G. Kornfeld, P. R. Goulding, "Application of multivariate statistical process control to batch operations", *Computers and Chemical Engineering*, Vol.24, pp.291-296, 2000.
DOI: https://doi.org/10.1016/S0098-1354(00)00480-4

[2] K. Tidriri, N. Chatti, S. Verron, T. Tiplica, "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: a review of researches and future challenges", *Annual Reviews in Control*, Vol.42, pp.63-81, 2016.
DOI: https://doi.org/10.1016/j.arcontrol.2016.09.008

[3] Z. Ge, "Process data analytics via probabilistic latent variable models: a tutorial review", *Industrial and Engineering Chemistry Research*, Vol.57, pp.12646-12661, 2018.
DOI: https://doi.org/10.1021/acs.iecr.8b02913

[4] P. Nomikos, J. F. MacGregor, "Monitoring batch processes using multiway principal component analysis", *AIChE Journal*, Vol.40, pp.1361-1375, 1994.
DOI: https://doi.org/10.1002/aic.690400809

[5] P. Nomikos, J. F. MacGregor, "Multi-way partial least squares in monitoring batch processes", *Chemometrics and Intelligent Laboratory Systems*, Vol.30, pp.97-108, 1995.
DOI: https://doi.org/10.1016/0169-7439(95)00043-7

[6] Y. Hui, X. Zhao, "Multi-phase batch process monitoring based on multiway weighted global neighborhood preserving embedding method", *Journal of Process Control*, Vol.69, pp.44-57, 2018.
DOI: https://doi.org/10.1016/j.jprocont.2018.06.012

[7] Z. Ge, "Review on data-driven modeling and monitoring for plant-wide industrial processes", *Chemometrics and Intelligent Laboratory Systems*, Vol.171, pp.16-25, 2017.
DOI: https://doi.org/10.1016/j.chemolab.2017.09.021

[8] M. Vanhala, C. Lu, J. Peltonen, S. Sundqvist, J. Nummenmaa, K. Järvelin, "The usage of large data sets in online consumer behaviour: a bibliometric and computational text-mining-driven analysis of previous research", *Journal of Business Research*, Vol.106, pp.46-59, 2020.
DOI:https://doi.org/10.1016/j.jbusres.2019.09.009

[9] G. H. Cho, S. Y. Lim, S. Hur, "An analysis of the Research Methodologies and Techniques in the Industrial Engineering Using Text Mining", *Journal of the Korean Institute of Industrial Engineers*, Vol.40, pp.52-59, 2014.
DOI:https://doi.org/10.7232/JKIIE.2014.40.1.052

[10] A. Gupta, V. Dengre, H. A. Kheruwala, M. Shah, "Comprehensive review of text-mining applications in finance. *Financial Innovation*, Vol.6, 39, 2020.
DOI:https://doi.org/10.1186/s40854-020-00205-1

[11] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", *Multimedia Tools and Applications*, Vol.78, pp.15169-15211, 2019.
DOI:https://doi.org/10.1007/s11042-018-6894-4

[12] A. Amado, P. Cortez, P. Rita, S. Moro, "Research trends on big data in marketing: A text mining and topic modeling based literature analysis", *European Research on Management and Business Economics*, Vol.24, No.1, pp.1-7, 2018.
DOI:https://doi.org/10.1016/j.iedeen.2017.06.002

[13] S. Kim, H. Cho, J. Kang, "The status of using text mining in academic research and analysis methods", *Journal of Information Technology and Architecture*, Vol.13. pp.317-329, 2016.
DOI:https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[14] R. Alghamdi, K. Alfalqi, "A survey of topic modeling in text mining", *International Journal of Advanced Computer Science and Applications*, Vol.6, pp.147-153, 2015.
DOI:https://dx.doi.org/10.14569/IJACSA.2015.060121

[15] S. Yoon, M. Kim, "topic modeling on fine dust issues using LDA analysis", *Journal of Energy Engineering*, Vol.29, pp.23-29, 2020.
DOI:https://doi.org/10.5855/ENERGY.2020.29.2.023

[16] J. Jeong, S. H. Kim, "Failure diagnosis using text mining and deep learning: development of prediction algorithm for responsible department", *The Transactions of the Korean Institute of Electrical Engineers*, Vol.69, pp.1225-1236, 2020.
DOI:https://doi.org/10.5370/KIEE.2020.69.8.1225

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

조 현 우(Hyun-Woo Cho) [정회원]

- 2003년 8월 : 포항공과대학교 기계산업공학부 (공학박사)
- 2003년 8월 ~ 2007년 8월 : 포항공과대학교, 조지아텍, 테네시주립대 연구원
- 2007년 9월 ~ 2011년 2월 : 삼성전자, 삼성디스플레이 책임연구원
- 2011년 3월 ~ 현재 : 대구대학교 기계공학부 교수

〈관심분야〉
공정모니터링, 빅데이터, 인공지능