Journal of the Korean Institute of Industrial Engineers
Vol. 40, No. 1, pp. 52-59, February 2014.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2014.40.1.052

<Original Research Paper>

텍스트 마이닝을 이용한 산업공학 연구기법의 분석

조근호 · 임시영 · 허 선†

한양대학교 산업경영공학과

An Analysis of the Research Methodologies and Techniques in the Industrial Engineering Using Text Mining

Geun Ho Cho · Si Yeong Lim · Sun Hur

Department of Industrial and Management Engineering, Hanyang University

We survey 3,857 journal articles published on the four domestic academic journals in the industrial engineering field during 1975～2012. Titles, abstracts, and keywords of the papers are searched by means of text mining technique to draw the information on the methodologies and techniques adopted in the papers, and then we aggregate and merge similar ones to obtain final 38 representative methodologies and techniques. Trends of these methodologies and techniques are studied by analyzing frequencies, clustering, and finding association rules among them. Results of the paper can shed a light to choose tools in the future education and research in the industrial engineering related area.

Keywords: Text Mining, Association rule, Clustering, Methodology, Industrial Engineering

1. 서 론

산업공학(Industrial Engineering)은 기업경영의 합리화와 효율성 향상을 위해 인간, 물자, 장비, 에너지 등 다양한 개별분야로 구성된 통합시스템의 설계, 개선 및 실행 등을 연구하는 학문이다. 이 같은 특성상 전통적인 생산, 제조 산업에서 벗어나 IT, 금융, 유통, 물류, 의료 등 매우 다양한 분야에 적용할 수 있으며 이에 따라 다양한 학문분야에서 개발된 많은 방법론과 기법들을 활용하고 있다. 산업공학의 방법론 중 하나인 경영과학(management science)은 지난 수십 년 동안 고유한 학문 분야로 자리잡았으며, 컴퓨터과학도 산업공학만의 차별화된 방법으로 산업공학 학문연구에 적용되고 있다(Park *et al*., 2006).

국내 산업공학은 지난 50년 동안 많은 발전을 이루었으며 현재 60여 개에 달하는 4년제 대학과 많은 2년제 대학에 산업공학 관련학과가 개설되어 있다. 산업공학의 다양한 적용분야에 맞게 대학에서 이루어지는 연구에도 다양하고 이질적인 기법들이 사용되고 있으며 또한 그 추세도 산업공학을 둘러싼 외부환경의 변화에 대응하면서 적절히 변화하고 있다. 한동안 이공계 대학교육의 위기를 우려하는 현상이 있었고 아직도 이공계인력부족과 기피현상은 지속되고 있어 첨단기술을 습득하고 활용할 수 있는 유능한 이공계 인력육성은 필수적이고 시급한 국가과제이다. 특히 교육의 주체인 대학 차원에서 이공계 교육 정상화를 위한 노력이 절실히 요구되고 있음에 따라, 산업공학 분야에서도 각종 방법론들과 기법들의 변화추세를 파악하고 이를 대학교육에 반영하는 노력이 필요하다. 특히 최근 들어 많은 기기들이 디지털화, 네트워크화, 모바일화 되어가면서 제조, 서비스, IT, 의료 등 전 산업 영역에서 방대한 양의 데이터가 쏟아져 나오고 있으며(이른바 빅데이터(Big Data) 시대의 등장) 산업공학 내에서도 이 빅데이터를 활용하여 다양한 시스템을 분석하는 연구기법과 방법론들이 생겨나기 시작했다(Park, 2011). 따라서 이와 같이 새롭게 탄생하는 기법 및 방법론을 대학교육에 반영하는 노력을 기울여야 한다.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012R1A2A2A01005219).
† 연락저자 : 허 선 교수, 426-791 경기도 안산시 상록구 한양대학로 55 한양대학교 산업경영공학과, Tel : 031-400-5265, Fax : 031-400-5265, E-mail : hursun@hanyang.ac.kr
2013년 12월 10일 접수; 2014년 1월 29일 수정본 접수; 2014년 2월 4일 게재 확정.

이에 따라 본 연구에서는 대학에서 이루어지고 있는 산업공학 관련연구에서 사용되는 기법 및 방법론들을 분석하여 연도별 추세와 상호 연관성을 파악하고자 한다. 이를 위해 산업공학 관련학과 교수들이 게재한 국내 논문들을 대상으로 주제어와 초록 데이터를 수집하고 데이터마이닝 기법을 활용하여 분석한다. 이와 유사하게 주제어(keywords)나 특허, SNS 등을 분석하여 특정 학문분야의 연구나 기술개발 동향을 분석하고자 하는 연구가 최근 다양하게 이루어지고 있다. Choi *et al.*(2013)과 Ahn *et al.*(2013) 등은 기술특허를 대상으로 주제어를 분석하여 특정 기술의 동향을 파악하였고, Choi(2013)는 방호복 기술을 예측하는 방법을 제시한 연구를 수행하였으며 Song *et al.*(2013)은 국내경제관련 학술논문의 주제어 분석으로 경제분야에 연구 동향을 분석하였다. 특히 최근에는 해외저널인 IIE Transaction에 게재된 논문들의 주제어들을 분석하여 산업공학 연구분야 동향 파악 및 주제어간에 관계를 분석한 연구(Cho *et al*., 2012)가 있다. 이에 반해 본 연구에서는 국내 산업공학 관련저널들에 게재된 논문들의 주제어들과 초록 내용들을 바탕으로 산업공학의 주요 연구기법들의 동향을 파악한다. 특히 본 연구의 초점은 연구분야에 두지 않고 연구에 사용되는 기법들과 방법론들에 두었으며, 주로 사용되는 기법들의 변화 추세와 주요 기법들간의 상호연관성을 파악함으로써 향후 산업공학 연구나 교육에 참고가 되고자 하는 데에 그 목적이 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구의 대상이 되는 데이터를 선정하고 이를 수집, 정리하는 방법을 제시한다. 제 3장에서는 본 분석에 사용된 방법들에 대해 간략히 소개하고 제 4장에서는 연구기법 출현빈도에 대한 분석과 연구기법 군집화, 그리고 연구기법간 연관관계 분석 결과를 제시한다.

2. 데이터 수집과 정리

2011년~2013년 중앙일보 학과평가에서의 산업공학과 순위 상위 25개 대학의 산업공학 관련학과를 선정하고 소속 전임교수들의 홈페이지를 검색하여 이들이 가장 많이 논문을 게재한 상위 5개의 국내저널을 선택하였다. 그 결과 대한산업공학회지, 산업공학(IE Interface), 한국산업경영시스템학회지, 한국경영과학회지, 한국인간공학회지 등을 선정하였으나 한국인간공학회지에 게재된 논문들의 경우 다른 4개 학회지와 비교하여 주로 사용된 연구기법이 매우 상이하므로 이를 제외하였다. 이들 4개 학회지에 대한 분석을 목적으로 국가과학기술정보센터(www.ndsl.kr)에서 저널명을 검색어로 지정하여 해당 웹페이지 및 링크를 R(버전 3.0.2)을 이용해 추적함으로써 1975년부터 2012년까지 총 3,857개 논문들에서 제목, 초록과 저자가 작성한 주제어를 수집하였다.

본 논문의 목적이 산업공학 연구기법의 분석이기 때문에 이들 주제어 중 기법에 대한 용어들만을 추출하거나 또는 연구에 사용되었음에도 불구하고 주제어에서 제외된 연구기법을 파악해야 한다. 또한 이들 수집된 주제어 가운데에는 동일한 기법이지만 저자들이 상이한 표현이나 용어를 사용한 경우, 또는 띄어쓰기나 단/복수의 차이에 의해 중복 계산된 것들도 포함되어 있는 등의 문제점이 있다. 이를 해결하기 위해서 본 연구에서는 산업공학의 연구기법들을 대표할 수 있는 용어를 우선 선정하고, 이를 수집된 각 논문의 제목, 초록, 주제어 등에서 탐색하는 방법을 적용하였다. 산업공학의 주요 연구기법(이하 대표기법)들을 선정하기 위해서 산업공학용어사전(대한산업공학회, 1992)에서 1차로 총 173개의 기법용어를 추출하였고, 이들에 대한 계층구조를 파악하고 적절한 수준에서 이들을 통합하여 최종적으로 38개의 대표기법을 선정하였다. 본 연구에서 선정한 계층상의 레벨이 다를 수 있고 또 알고리즘과 연구분야, 방법론 등이 혼재되어 있을 수 있지만 산업공학 연구에서의 기법 활용에 대한 현황을 알아보기 위한 본 연구의 목적을 달성하는 데에 충분하다. 38개 기법은 다음과 같다(알파벳 순).

Bayesian analysis, Branch and bound, Case study, Data envelopment analysis, **Database management**, **Data mining, Decision support**, Design of experiment, Diffusion model, Dynamic programming, Engineering economy, **Expert system**, **Factor analysis**, Fault tree analysis, Fuzzy theory, Genetic algorithm, **Heuristic algorithm**, Integer programming, **Inventory control**, Linear programming, **Markov process**, **Mathematical programming**, Motion time analysis, Network flow analysis, **New product design**, Nonlinear programming, Petri net, **Production system**, Queueing theory, Regression analysis, Reliability analysis, **Scheduling**, Simulation, **Statistical inferences**, Statistical quality control, Stochastic processes, Survey/Questionaires, Time series.

위의 대표기법 가운데 굵은 글씨로 강조 된 것은 대표기법에 포함된 기법들이 자명하지 않은 것으로 판단하여 다음 <Table 1>에 해당 대표기법에 포함된 기법 중 자명하지 않은 것들을 추가로 설명하였다.

이제 하나의 대표기법이 포함하는 유사기법들과, 그 대표기법을 나타내는 용어와 유사한 용어들을 이 대표기법이 대표할 수 있도록 각 논문의 초록 단어들과 주제어들을 치환하는 작업을 우선 수행하였다. 예를 들면 대표기법 Queueing theory는 M/M/1, M/G/1, Queueing system, Queueing network 등의 유사기법들을 포함하며 또한 Queueing theories, M/M/1 system, Network of Queues 등의 유사용어들도 포함한다. 따라서 수집된 대상 논문들의 초록과 주제어 중에서 이들 유사기법이나 유사용어로 표현된 단어들은 모두 Queueing theory로 치환하였다. 이 같은 방법으로 38개 대표기법을 수집된 3,875개 논문의 주제어와 초록의 단어들과 비교하면서 출현여부를 판단하여 38×3,857 크기의 대표기법-논문 행렬을 구성하고 이를 본 연구의 각종 분석의 출발점으로 삼았다(제 4.3절의 <Table 3> 참조).

Table 1. Techniques included in the representative techniques(partial)

Rep. Techniques	Included Techniques	Remarks
Database management	Binary search, Database	
Data mining	Association rule, Decision tree, CART, C4.5, LDA, k-NN, Lift chart, Naïve Baysian, SVM, Kernel method, Text mining, Web mining	
Decision support	Utility theory	
Expert system	AI, feature extraction, ANN, SOM	
Factor analysis	FEM, structural equation model, PCA	
Heuristic algorithm	Ant colony, Bootstrap, Immune optimization, Meta heuristic, Particle swarm optimization, Simulated annealing, Tabu search	Genetic algorithm is classified separately
Inventory control	EOQ, EPQ, Inventory system	
Mathematical programming	Bin packing, Cholesky factorization, Convex optimization, Goal programming, Graph theory, Hungarian method, Knapsack problem, Maximal covering	DP, IP, LP, NLP are classified separately
Markov process	Markov decision problem	
New product design	FAST, TRIZ, QFD, Kano model	
Production system	CAD/CAM, CAPP, Concurrent engineering, Group technology, Job shop, Kanban system, MRP, Pokayoke, P-Q analysis, Taguchi method	
Scheduling	Sequencing, Sequential ordering	
Statistical inferences	Various test techniques such as ANOVA, Correlation analysis, Goodness-of-fit, MLE, MVUE, MOM, t-Test and so on are included here.	

3. 연구 방법

3.1 단순 빈도 분석

1975～2012년의 기간 동안 38개 대표기법 빈도수를 계산하여 출현이 잦은 순으로 상위 20개의 대표기법을 비교하였고, 총 기간을 3개의 구간으로 나누어 각 구간마다 대표기법 등장 순위의 변화 추세를 살펴보았다. 분석결과는 제 4.1절과 제 4.2절에 제시한다.

3.2 K-평균군집화 기법(K-Means Clustering Method)

군집분석은 다양한 개체들을 몇 개의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적인 분석 방법이다. K-평균군집화 기법은 비계층적 분석방법에 속하는 기법으로서 사전에 군집 개수 K를 설정하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 구분하는 것이다. K-평균화 군집 분석의 절차는 다음과 같다. 먼저 객체들을 K개의 초기 군집으로 나눈 후, 각 군집의 K개 중심과 각 데이터와의 거리를 계산하여 그 중심과 가까우면 그 군집에 할당하고 중심과 가깝지 않으면 다른 군집에 할당한다. 그 결과 달라진 군집으로 군집의 중심을 다시 계산하며 더 이상 중심의 이동이 없을 때까지 위 단계를 반복한다(MacQueen, 1967). 자료에 이상값(outlier)이 존재하면 군집을 명확히 나누지 못하는 단점이 있으므로 이상값을 탐색, 제거한 후 시행하는 것이 좋으며, 적절한 K값을 정하는 것이 중요하다(Lee, 2010). 분석결과는 제 4.3절에 제시한다.

3.3 연관성분석(Association Rule)

연관성분석은 데이터 안에 존재하는 항목간의 연관규칙을 발견하는 과정이다. 한 항목과 다른 항목 사이에 연관성을 수치화된 값으로 찾아내는 방법으로서 연관관계 “A이면 B이다(A → B)”의 지지도(support), 신뢰도(confidence), 그리고 향상도(lift)를 이용하여 이 연관관계의 정도를 파악하고 해석한다(Agrawal *et al*., 1993). 지지도란 전체 데이터 중 항목 A와 항목 B를 모두 포함할 확률이며, 신뢰도는 항목 A가 나타난 데이터 중 항목 B 또한 나타난 데이터의 비율이고 향상도는 항목 B가 항목 A와 포함되어 나타난 경우와 항목 B만으로 나타난 경우의 비율을 나타낸다.

지지도는 두 항목이 전제 데이터에서 얼마나 자주 출현하는가를 나타내며, 신뢰도는 해당 연관규칙의 유용성을 나타내는 척도이다. 향상도는 항목 A와 B가 연관관계 없이 무작위로 발견될 확률에 비하여 두 항목 간 연관규칙을 알고 관찰되는 경우 얼마나 관찰될 확률이 증가하는지를 나타내는 것으로, 연관규칙의 효용성을 표시하는 것이라 할 수 있다(Jun, 2012). 연관성 분석의 결과는 제 4.4절에 나타나 있다.

4. 연구 결과

4.1 대표기법 순위 분석

1975～2012년의 대표기법 출현 빈도 데이터를 이용하여 출현 빈도가 높은 상위 20개의 대표기법들을 비교한 것이 <Figure

1>이다. Simulation 기법이 논문에 총 355번 등장하면서 가장 많이 사용되었고 그 뒤로 Inventory control, Decision support, Reliability analysis 기법이 각각 223번, 201번, 196번 등장하였다.

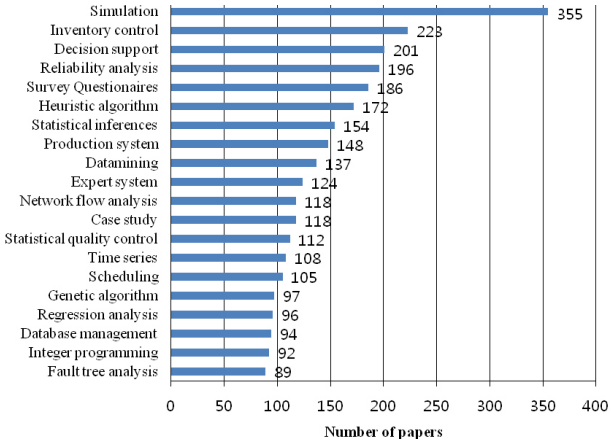

Figure 1. Top 20 techniques in appearance frequencies during 1975～2012

4.2 기간의 구간별 분석

1975～2012년 기간 동안 4대 저널에 게재된 총 논문수의 추이를 보면 <Figure 2>와 같이 제 1구간(1975～1990년), 제 2구간(1991～2000년), 그리고 제 3구간(2001～2012년) 등 세 개의 구간으로 구분할 수 있다. 각 구간별로 논문수가 확연히 구분되는데, 제 1구간은 논문게재 자체가 저조하던 시기, 제 2구간은 각 대학별로 연구업적 평가에 대한 기준이 강화됨에 따라 논문게재가 점차적으로 활발하게 이루어지는 시기, 그리고 제 3구간은 해외저널 논문게재를 중시함에 따라 국내저널 논문게재가 상대적으로 정체상태에 있는 시기와 부합되는 것으로 보인다.

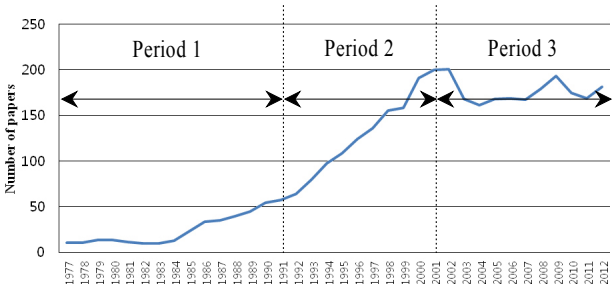

Figure 2. Periods in appearance frequencies of papers during 1975～2012

<Table 2>를 살펴보면, Simulation 기법은 제 2구간에서 순위가 한 단계 상승해서 3구간까지 그 순위를 이어가면서 제 2, 3구간에서 가장 출현 빈도가 높은 기법인 것을 알 수 있다. Inventory control, Reliability analysis, Data mining 등의 기법들은 제 1구간보다 제 2구간에서 순위가 하락했지만 제 3구간에서는 다시 순위가 상승하였다. 반대로 Heuristic algorithm(Genetic algorithm 제외), Production system 기법들은 제 2구간에서 순위가 상승하였으나 제 3구간으로 오면서 다시 하강하였다. 또한 제 1구간 이후 출현 빈도가 20위권 밖으로 사라진 기법으로는 정통적인 산업공학 연구기법이라고 할 수 있는 Branch and bound, Dynamic programming, Mathematical programming 등 있으며, Scheduling, Queueing theory, Fuzzy theory의 경우는 제 3구간에서부터 그 출현빈도가 20위 밖으로 밀려난 기법들이다. 반면에 컴퓨터과학 관련 기법이라고 할 수 있는 Expert system, Database management 등과 정성적 방법인 Survey/Questionaires, Case study 기법과 Markov process의 경우 제 2구간부터 출현빈도가 20위권 안으로 증가했으며, Expert system 기법을 제외한 나머지 4개 기법들은 제 3구간까지도 20위 이내에 등장하였다. 그리고 Genetic algorithm과 Regression analysis 기법들은 비교적 최근인 제 3구간에 오면서 출현 빈도가 상위 20위권 안으로 상승하였다.

<Figure 1>과 <Table 2>에서 본 것은 대표기법들의 절대적 중요성을 나타내는 단순출현빈도 수를 바탕으로 추세와 경향을 파악한 것이다. 대표기법의 상대적 중요성을 파악하고 그 중요성이 시간흐름에 따라 어떻게 변화하는가를 보기 위해서 1975～2012년 동안 출현빈도가 높은 상위 20개 대표기법의 출현 비율(출현비율 = 해당 키워드 출현 횟수/해당 기간의 총 논문 수)의 변화를 다음 <Figure 3>과 같이 이차원 좌표 형태로 표현하였다. 그림에서 X축의 값은 제 1구간에서의 출현 비율과 제 2구간에서의 출현비율의 차이를 나타내고 Y축의 값은 제 2구간에서의 출현 비율과 제 3구간에서의 출현비율의 차이를 표시한다. 따라서 제 1사분면에 있는 대표기법들은 1→2→3구간으로 시기가 변하면서 출현비율이 지속적으로 상승한 기법들이며, 2사분면에 존재하는 대표기법들은 1→2 구간에서 그 출현 비율이 하강했으나 2→3 구간에서 출현비율이 상승한 기법들이다. 각 기법들의 점의 크기는 1975～2012년 동안의 출현비율의 크기를 표현하였다. 그림에서 보면 제 3사분면, 즉 1→2→3 구간으로 오면서 계속 출현비율이 하강한 대표기법들은 하나도 없는 것으로 나타나는데, 이는 출현빈도가 높은 상위 20개 대표기법 만을 대상으로 하였기 때문이다.

<Figure 3>에서 보면 Survey/Questionaires, Case study, Genetic algorithm, Regression analysis 등은 1사분면에 자리하고 있는데 이는 출현비율이 지속적으로 상승해 왔음을 나타내고 있는 반면, Heuristic algorithm, Database management 등은 그 출현비율이 2기에서는 전기에 비해 증가하였으나 3기에 와서 하락하였으며 Expert system, Production system은 그 하락세가 크게 두드러졌다. Data mining, Statistical quality control, Statistical inference, Reliability 등의 통계 관련 기법들과 Fault tree analysis 등은 2기에서는 전기에 비해 다소 감소 내지는 정체상태이었으나 3기에 넘어오면서 상승세를 보이고 있으며 Inventory control의 경우 2기에서는 출현비율이 크게 하락하였으나 공급사슬관리(SCM)에 관한 연구가 활기를 띰에 따라 3기에서는 증가세로 돌아서고 있다. 특히, Simulation 기법의 경우는 <Figure

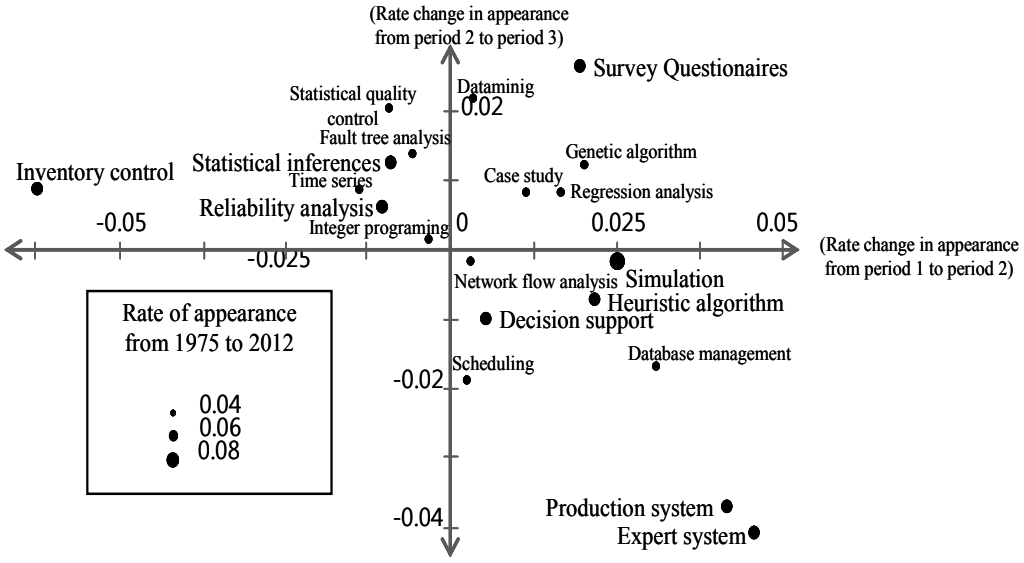

Figure 3. Change in the rates of rep. techniques appearance during 3 periods

Table 2. Change of rep. techniques appearance rankings over 3 periods

Period 1(1975～1990)			Period 2(1991～2000)			Period 3(2001～2012)		
Ranking	Rep Techniques	NUM	Ranking	Rep Techniques	NUM	Ranking	Rep Techniques	NUM
1	Inventory control	42	1	Simulation	127	1	Simulation	201
2	Simulation	27	2	Production system	84	2	Survey Questionaires	133
3	Reliability analysis	22	3	Expert system	79	3	Inventory control	119
4	Decision support	20	4	Decision support	77	4	Reliability analysis	112
5	Linear programming	16	5	Heuristic algorithm	67	5	Decision support	104
6	Statistical inferences	16	6	Inventory control	62	6	Data mining	97
7	Scheduling	14	7	Reliability analysis	62	7	Statistical inferences	95
8	Time series	14	8	Scheduling	51	8	Heuristic algorithm	94
9	Datamining	13	9	Database management	49	9	Statistical quality control	80
10	Heuristic algorithm	11	10	Survey Questionaires	47	10	Case study	76
11	Network flow analysis	11	11	Queueing theory	43	11	Genetic algorithm	70
12	Branch and bound	10	12	Statistical inferences	43	12	Regression analysis	65
13	Dynamic programming	10	13	Network flow analysis	42	13	Network flow analysis	65
14	Integer programming	10	14	Linear programming	37	14	Time series	64
15	Statistical quality control	10	15	Case study	36	15	Fault tree analysis	61
16	Queueing theory	9	16	Fuzzy theory	36	16	Production system	56
17	Fault tree analysis	8	17	Datamining	32	17	Integer programming	52
18	Production system	8	18	Integer programming	30	18	Markov process	51
19	Matematical programming	8	19	Markov process	30	19	Database management	43
20	Fuzzy theory	7	20	Time series	30	20	New product design	43

1>에서 보듯이 단순 출현빈도가 다른 대표기법에 비해 월등히 높지만 출현비율을 볼 때에는 <Figure 3>에서 보듯이 1→2 구간에서는 그 비율이 상승했으나 2→3 구간에서 정체상태에 있음을 볼 수 있다. 단순 출현빈도가 높다는 것은 많은 연구논문에서 이 기법을 자주 사용하고 있어서 절대적 중요성은 계속 증가하며, 연구과정에서 사용하는 기법 중 차지하는 위치를 나타내는 상대적 중요성은 확고한 위치를 차지하고 있으며 그것이 시간 흐름에 따라서 크게 변하지 않고 있다고 볼 수 있다.

4.3 군집화

하나의 논문에 함께 사용되고 있는 기법을 파악하기 위해서 대표기법의 논문 출현여부를 기록한 행렬을 이용하여 K-평균 군집화를 적용하였다. 38×3857차원의 대표기법-논문 행렬을 <Table 3>에서 예시하는 바와 같이 특정 대표기법이 해당 논문에 출현했을 경우 '1' 그렇지 않으면 '0'의 값을 갖도록 구성하였다. 예를 들어 대표기법 Reliability analysis는 Document 1에서는 출현하였지만 Document 2 논문에는 등장하지 않음을 나타내고 있다.

Table 3. Example of matrix of techniques-documents

Keyword	Document 1	Document 2	…	Document 3857
Simulation	0	1	…	0
Reliability analysis	1	0	…	0
…	…	…	…	0
Motion time analysis	0	0	…	0

Table 4. Clustering result of rep techniques

Cluster 1	Genetic algorithm, Heuristic algorithm, Mathematical programming, Network flow analysis, Petri net, Scheduling
Cluster 2	Design of experiment, Fault tree analysis, Fuzzy theory, Queueing theory, Reliability analysis, Simulation, Stochastic processes
Cluster 3	Branch and bound, Dynamic programming, Integer programming, Linear programming, Non linear programming, Production system
Cluster 4	DEA, Database management, Factor analysis, Motion time analysis, Regression analysis, Survey/Questionaires
Cluster 5	Bayesian analysis, Case study, Data mining, Decision support, Diffusion model, Engineering economy, Expert system, Inventory control, Markov process, New product design, Statistical inferences, Statistical quality control, Time series

군집개수는 2～8개로 다양하게 적용한 결과 각 군집간 관측치 개수가 적절히 배분되는 5개로 설정하고 자료 형태가 이진변수인 것을 감안하여 이진거리 계산방식을 적용하였다. 군집 결과는 <Table 4>에 나타나있는데, 군집 1～군집 3에 포함되어 있는 기법들은 기법들 간에 공통적 특성이 보이나 군집 4와 군집 5의 경우는 서로 다른 특성의 여러 기법들이 혼재되어 있는 모습을 보였다. 구체적으로는 군집 1에 포함된 대표기법들을 보면 복잡한 Scheduling 관련모델에 Heuristic algorithm이나 각종 수리계획법을 적용한 것임을 알 수 있으며, 군집 2는 실험계획법, 확률적 모형관련 기법들로 군집이 구성되었고 군집 3은 생산시스템 분석에 정통적인 수리계획기법들을 사용한 것을 알 수 있다.

다음으로 최근 산업공학 연구 기법들의 추세를 파악하기 위해서 2000～2012년 연도별 대표기법 출현 비율 시계열자료를 이용하여 K-평균 군집화를 수행하였다. 우선 제 2장에서 언급한 38×3857차원의 대표기법-논문 행렬을 바탕으로 2000～2012년 연도별 대표기법의 출현 비율을 기록한 38×13차원의 대표기법-연도 행렬을 <Table 5>에서 예시하는 바와 같이 구성하였다. 예를 들어 Simulation의 경우 2000년도에 출현한 대표기법들 중에서 출현 비율이 0.0802이며, 2001년도에는 그 출현 비율이 0.1281로 상승하였다.

Table 5. Change of the rate of rep. techniques

Techniques	2000	2001		2012
Simulation	0.0802	0.1281	…	0.1308
Reliability analysis	0.0420	0.0394	…	0.0841
…	…	…	…	…
Motion time analysis	0.0000	0.0099	…	0.0047

위의 자료를 바탕으로 유사한 시계열 패턴을 갖는 대표기법들을 군집화하기 위하여 상관계수 거리를 적용하여 군집을 형성하였다. 역시 초기 군집 수를 변화시키며 분석하였을 때, K = 4일 때 각 군집간 관측치의 개수가 적절히 배분되었으며 군집의 분류 역시 가장 명확하였다. <Figure 4>는 군집화 결과를 각 군집의 평균으로 표현한 것이다. 군집 1의 경우 그 출현 비율이 점점 하강하고 있는 군집이고, 군집 2의 경우는 그 추세가 비교적 일정한 비율로 출현하는 대표기법들로 구성되어 있다. 군집 3은 군집 2와 비슷하지만 연도에 따른 변동폭이 큰 대표기법들로 분류되어 있고, 군집 4의 경우는 연도별 출현 비율이 다른 군집들에 비해서 상대적으로 증가하고 있는 대표기법들로 모여 있다. <Table 6>은 이 같은 시계열 자료의 군집화 결과를 요약한 표이다.

최근 기법에 대한 추세의 군집화에 따르면, Branch and bound나 Mathematical programming과 같은 정통적인 OR 기법과 생산관련 기법들은 다소 줄어드는 추세이며 Genetic algorithm을 포함하여 Heuristic algorithm이나 Inventory control 등의 대표기법은 상승하는 경향이다. Cho *et al.*(2012)에서 제시된 해외 논문의 최근 상승 대표기법과 비교할 경우 Bayesian, Dynamic programming, Markov Chain, Generic Algorithm, Inventory은 동일하게 상승하고 있음을 확인할 수 있다. 그러나 Integer programming, Linear programming의 경우 해외에서는 하강 대표기법에 속하나 국내에서는 여전히 상승하고 있음을 확인할 수 있다. 또한 Queueing theory, Scheduling 등은 해외 및 국내에서 모두 하강하는 추세이다.

Figure 4. K-means clustering result of the time series of appearance rate for recent research techniques

Table 6. Trends of the clusters of the time series of appearance rate for recent research techniques

Cluster	Rep. Technique	Trend
Cluster1	Branch and bound, Database management, Expert system, Fuzzy theory, Mathematical programming, Production system, Queueing theory, Scheduling, Statistical inferences	Descent
Cluster2	Case study, Design of experiment, Diffusion model, Engineering economy, Fault tree analysis, Motion time analysis, Regression analysis, Reliability analysis, Simulation, Stochastic processes, Survey/Questionaires, Time series	Stable
Cluster3	Decision support, New product design, Non linear programming, Statistical quality control, Data Mining	Fluctuating
Cluster4	Bayesian analysis, Data envelopment analysis, Dynamic programming, Factor analysis, Genetic algorithm, Heuristic algorithm, Integer programming, Inventory control, Linear programming, Markov process, Network flow analysis, Petri net	Rising

Table 7. Association result of rep. techniques

No.	Antecedent	Consequent	Support	Confidence	Lift
1	Linear programming	Scheduling	0.50	0.86	1.43
2	Database management	New product design	0.50	0.86	1.37
3	Markov process	Case study	0.50	0.83	1.31
4	Regression analysis	Statistical quality control	0.61	0.88	1.29
5	Expert system	Regression analysis	0.58	0.88	1.29
6	Production system	Reliability analysis	0.55	0.88	1.28
7	Markov process	Reliability analysis	0.53	0.87	1.27
8	Heuristic algorithm	Network flow analysis	0.53	0.80	1.27
9	New product design	Heuristic algorithm	0.53	0.83	1.27
10	Survey Questionaires	Regression analysis	0.55	0.84	1.23

4.4 대표기법 관계 분석

여기서는 논문에서 특정 기법과 함께 사용되는 산업공학 기법은 무엇이 있는지, 그리고 두 기법간 연관관계가 어떻게 되는지 분석하였다. <Table 7>은 선-후행 대표기법의 연관관계 가운데 향상도(lift)가 높은 순으로 10개를 나열한 것이며, 특히 한 논문에서 함께 사용되는 기법들을 살펴보기 위하여 지지도가 0.5 이상 되는 기법들만을 추출하였다. 예를 들어 대표기법 Linear programming 기법이 사용된 논문에서는 Scheduling 기법이 발견될 확률, 즉 신뢰도는 0.86으로 매우 높고, Linear programming 기법이 사용된 논문에서 Scheduling 기법이 발견될 확률 Pr(Scheduling|Linear programming)은 Scheduling 기법이 발견될 확률 Pr(Scheduling)의 1.4배(향상도)이다. 따라서 Linear programming과 Scheduling 기법은 같이 사용되는 경우가 많다고 유추할 수 있다. 다른 대표기법들도 같은 방법으로 해석이 가능하다.

5. 결 론

본 연구에서는 텍스트 마이닝 방법을 사용하여 대한산업공학회지, IE Interface, 한국산업경영시스템학회지, 한국경영과학회지 등 4개의 학회지에 게재된 논문들에 대해 대표기법 출현빈도 분석, 구간별 분석, 군집화, 연관성 분석을 실시하였다. 이로써 1975～2012년까지 연구에 사용되었던 주요 기법들을 살펴보았고, 전체 기간을 3구간으로 나누어 각 구간별로 유행했던 기법들을 알아보았다. 또한 K-평균 군집화를 사용하여 같은 논문에 함께 사용되는 기법들을 파악하였고, 비교적 최근인 2000～2012년 동안 연구기법의 사용 추세가 어떻게 변하였는지 살펴보았다. 마지막으로 연관성 분석을 이용하여 한 논문에 연계되어 사용되는 기법들간의 관계를 수치적으로 분석하였다.

본 연구에서 도출된 결과를 통해 산업공학 기법의 발전 과정을 알 수 있을 뿐 만 아니라 추가적으로 산업 환경의 발전을 고려하면 변화에 적응하고 있는 산업공학 발전 특성을 확인할 수 있을 것이다. 더불어 해외 연구 동향과 비교를 통해 국내 산업공학 연구기법의 세계적 위상을 점검할 수 있으며 해외 추세를 따라가기 위하여 국내 연구에서 필요한 연구기법을 확인할 수 있는 자료로 활용될 수 있다. 연구자 입장에서는 각광받고 있는 기법을 확인할 수 있는 기회이자, 문제해결을 위해 적용해야 하는 추가적 기법을 고려할 때 판단 자료로 활용할 수 있을 것이다. 추후 연구로, 산업공학의 연구대상 영역 및 연구목적을 조사하고 이를 관련 기법과 연계함으로써 연구대상 영역별, 연구 목적별로 유효한 연구기법을 제안하는 연구가 수행된다면 유용할 것이다.

참고문헌

Agrawal, R., Imielinski, T., and Swami, A. (1993), Mining Association Rules between Sets of Items in Large Databases, *Proc. of the 1993 ACM SIGMOD Conference on Management of Data*, 207-216.

Ahn, G. S. and Hur S. (2013), Analysis of Biometric Technology Trend Using Key words Based Patent Analysis, *Proc. Fall Conference of The Korea BI Datamining Society.*

Cho, S. G. and Kim, S. B. (2012), Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining, *Journal of The Korean Institute of Industrial Engineers*, **38**(1), 67-73.

Choi, D. H., Kim, G. J., Park, S. S., and Jang, D. S. (2013), Forecasting Emerging Technology in AMOLED Using Keyword Quantitative Analysis Based on Textmining, *Proc. Spring Conference on The Korea Contents Association*, 365-366.

Choi, J. S. (2013), Keyword-based Patent Trend Analysis Using Statistical Analysis : The Case of Armor Technology, *The Journal of Intellectual Property*, **8**(1), 223-252.

Jun, C. H. (2012), *Datamining Techniques and Application*, Han Na Rae Academy, Seoul, Korea.

The Korean Institute of Industrial Engineers (1992), A Dictionary of Industrial Engineering, Cheongmoongak, Seoul, Korea

Lee, H. R. (2010), *A Study on Criteria for Determining the Number of Cluster in the K-means Method : Application to Cluster Analysis of Wind Regions in Korea Peninsula*, Master's Thesis, Dongguk University.

MacQueen, J. B. (1967), Some Methods for Classification an Analysis of Multivariate Observations, *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, **1**, 281-297.

Park, J. H. (2011), Data Science and Industrial Engineering, *IE magazine*, **18**(4), 22-24.

Park, Y. B., Lim. S. C., Hong, S. J., Kim, J. H., Yoon, M. H. and Lee, D. J.(2006), Development of Creative Engineering Education System and Curriculum by Case Studies and Task Analysis : *Focused on Industrial Engineering National Research Foundation of Korea*, D00006.

Song, H. J., Park, K. S., Jung, H. E., and Song, M. (2013), Trend Analysis of Korean Economy in the Economic Literature by Text Mining Techniques, *Proc. the* 20th *Conference on The Korean Society for Information Management*, 47-50.