Journal of the Korean Institute of Industrial Engineers
Vol. 38, No. 1, pp. 67-73, March 2012.

텍스트마이닝을 활용한 산업공학 학술지의 논문 주제어간 연관관계 연구

조수곤 · 김성범†

고려대학교 산업경영공학과

Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining

Su Gon Cho · Seoung Bum Kim

School of Industrial Management Engineering, Korea University

Identification of meaningful patterns and trends in large volumes of text data is an important task in various research areas. In the present study we crawled the keywords from the abstracts in IIE Transactions, one of the representative journals in the field of Industrial Engineering from 1969 to 2011. We applied low-dimensional embedding method, clustering analysis, association rule, and social network analysis to find meaningful associative patterns of key words frequently appeared in the paper.

Keywords: Data Mining, Text Mining, Clustering, Association Rule, Social Network Analysis

1. 서 론

데이터마이닝(Data Mining)은 방대하고 복잡한 데이터 내부에 존재하는 유용하고 의미 있는 정보를 이끌어 내는 방법을 연구하는 학문이다(Clifton, 2010). 주로 숫자형태의 일정한 데이터 구조로 정형화된 데이터(structured data) 를 분석해오던 데이터 마이닝의 연구자들은 최근 들어 텍스트, 이미지, 동영상, 음성 등과 같이 구조화되지 않은 비정형 데이터에 관심을 기울이고 있다(Chakrabartij, 2002). 특히 비정형 데이터 중에서 최근 인터넷 사용자의 폭발적인 증가에 힘입어, 웹 마이닝(Web Mining) 과 텍스트마이닝(Text Mining) 의 중요성이 더욱 부각되고 있다 (Choi *et al*., 2002). 이러한 관심은 대량의 텍스트에 대한 분석 및 연구로 이어져 생물학의 유전자 정보, 기술경영의 특허정보 등의 영역에서 적극적으로 연구되고 있다. 특히 특허문서로부터 기술동향의 패턴을 알아보는 연구가 행해 졌다. 이 연구는 특허문서의 제목과 초록으로부터 주제어를 추출한 후, 주제어의 출현 빈도수로 기술 동향을 파악하고 주제어간의 관계를 파악해 기술과 제품의 상관관계에 대한 규명을 하였으며 나아가 특정 제품과 기술에 관한 특허의 진화패턴을 발견하여 차세대 개발 분야의 방향을 제시하였다(Lee *et al*., 2008). 이 외에도 특허문서를 텍스트마이닝 기법을 이용해 분석한 연구가 행해졌다(Chen and Chen, 2011; Breitzman and Mogee, 2002).

본 논문에서는 산업공학 분야의 대표적인 국제학술지인 IIE Transactions에 출판된 논문의 내용을 저널 홈페이지로부터 추출하여, 유용한 정보를 이끌어내는 웹 컨텐츠 마이닝(Web Content Mining) 연구를 수행하였다. IIE Transactions 저널홈페이지(www.tandf.co.uk/journals/titles/0740817x.asp)에서 수집한 초록을 관찰하여 주제어의 출현횟수를 기록한 고차원의 데이터를 생성하고, 주제어간 상관관계의 분석을 위해 시각화, 군집분석, 연관성분석을 실시하였다. 또한, 1969부터 2011년 현재까지 산업공학의 연구동향을 논문의 주제어의 시계열 패턴 분석을 통하여 살펴보았다.

본 연구는 BK21(Network Enterprise)의 지원으로 수행된 연구임.
† 연락저자 : 김성범, 136-701 서울시 성북구 안암동 5가 1번지 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2011년 10월 26일 접수; 2012년 2월 15일 게재 확정.

2. 데이터

본 논문의 연구대상으로 선정된 IIE Transactions는 1969년 이후 현재까지, Design and Manufacturing, Operations Engineering and Analysis, Quality and Reliability Engineering, 그리고 Scheduling and Logistics 4가지 연구 분야의 논문을 출판하고 있는 산업공학의 대표적인 국제저널이다.

본 연구에서는 IIE Transactions에서 출판된 모든 아티클 중, 서평 또는 일반기고문 등을 제외하고 제목과 초록이 존재하는 2,527개의 연구논문을 분석대상으로 삼았다. 또한 관찰대상이 되는 주제어는 모든 논문에 기록된 주제어 중, 총 10회 이상 출현하면서, 중의어, 일반동사, 범용어를 제외한 48개의 단어를 선정하여 분석했다. 선정된 48개의 주제어를 알파벳 순서로 나열하면 다음과 같다. assembly, Bayesian, branch and bound, classification, cluster, control chart, curve, CUSUM, decision-making, decomposition, distribution, dynamic programming, EWMA, flexibility, forecasting, genetic algorithm, Inspection, integer programming, inventory, job shop, lead time, linear programming, location, maintenance, makespan, manufacturing, Markov chain, monitoring, network, optimization, pricing, priority, process control, quality, queueing, regression, reliability, sampling, scheduling, simulation, statistical process control, stochastic, storage, supply chain, transportation, uncertainty, warehouse, work-in-process. 수집한 항목과 이를 위해 활용된 프로그램의 요약은 <Table 1>과 <Table 2>에서 보여주고 있다.

Table 1. A summary of articles and keywords in IIE Transactions

구 분	수집 대상	연구 대상
Article	3,049	2,527
Keywords	1,266	48

Table 2. Items and software for data collection

항목	내용
수집 정의항목	제목, 출판년도, 저자, 초록, 주제어, 소속
수집 프로그램	Microsoft IIS 6.0, Active Server Pages Crawling Engine
데이터베이스	Microsoft SQL Server 2000

이와 같이 수집된 IIE Transactions 논문의 초록에서 선정된 주제어의 출현빈도를 관찰하여 48×2,527차원의 행렬(Matrix)을 구성하고, 연도별 주제어의 관측치를 기록한 48×43차원의 행렬 또한 생성하였다.

<Table 3>을 살펴보면, 일련번호 'D0001' 논문의 초록에서 주제어 'assembly'는 0회, 'D0002' 논문에서는 1회 사용되었음을 알 수 있다. 그리고 <Table 4>로부터 주제어 'assembly'는 1969년에 2회, 1970년도에는 0회가 관찰됨을 알 수 있다. 각각의 데이터는 다음 장에서 소개하는 다양한 방법을 이용하여 분석하게 되는데, <Table 3>의 데이터는 주제어간의 관계를 살펴보기 위한 용도로 사용되었고, <Table 4>의 데이터는 주제어의 연도별 출현 패턴을 알아보는데 활용되었다.

Table 3. The dataset containing the frequency of keywords from 2,527 abstracts

	D0001	D0002	…	D2527
assembly	0	1	…	0
Bayesian	2	0	…	0
…	…	…	…	…
work-in-process	0	0	…	0

Table 4. The dataset containing the frequency of keywords from 1969 to 2011

	Y1969	Y1970	…	Y2011
assembly	2	0	…	6
Bayesian	0	4	…	1
…	…	…	…	…
work-in-process	1	2	…	1

3. 방 법

3.1 지역선형사상(Locally Linear Embedding-LLE)

지역선형사상은 고차원 데이터의 차원을 축소하여, 이를 시각화하고 해석하기 위해 최근 사용되는 다변량 통계적 분석기법이다(Rowie *et al*., 2000). 지역선형사상은 기존에 많이 쓰이고 있는 차원축소방법인 주성분분석과는 목적은 같으나, 차원축소 시 인접한 점들의 특징을 고려한 지역적인 방법이라는 데 차이가 있다.

지역선형사상의 절차는 다음과 같다. D차원의 실수 벡터 X_i가 n개가 있다고 가정하자. 먼저 각각의 데이터 점 X_i와 가장 가까운 점들 X_j (j = 1,… k)를 k개 구한다. 그리고 X_j 와 가중벡터 W_{ij}의 선형방정식($\sum_j W_{ij}X_j$)과 X_i의 거리를 최소로 만드는 가중치 W_{ij}를 구한다. 다시 말해, 식 (1)으로 정의된 오차 E의 값을 최소로 하는 W_{ij}를 구하기 위하여, 최소제곱기법을 사용한다.

$$\min_W E(W) = \sum_i \left| X_i - \sum_j W_{ij} X_i \right|^2 \qquad (1)$$

마지막으로, 고차원의 점들과 동일한 가중치를 갖도록 구성된 저차원의 벡터 Y_i를 구한다. 이 과정에서는 식 (2)의 Φ를 최소화하는 Y_i를 구하며, Y_i는 고유치 문제를 해결하여 구하게 된다(Saul *et al*., 2000).

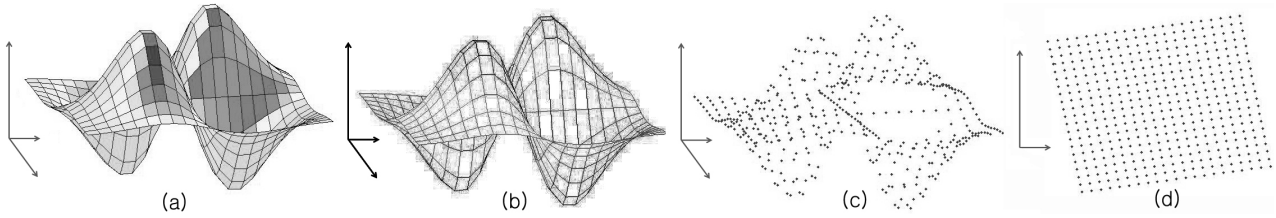

Figure 1. An example of dimensionality reduction by LLE. (a) 3-dimension manifold. (b) sampled line on manifold. (c) sampled points on manifold, (d) points in the reduced dimension by LLE

$$\min_W \Phi(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2 \quad (2)$$

<Figure 1>은 지역선형사상의 효과를 보여주는 예이다. (a)는 3차원의 다양체의 실제 모습이며, (b)와 (c)는 선과 점으로 샘플링 된 데이터들이다. 점으로 샘플링 된 (c)에 대하여 지역선형사상을 수행한 결과는 (d)에 나타나고 있다. 3차원 상 데이터들의 인접관계가 그대로 유지되고 있음을 확인할 수 있다.

3.2 k-평균 군집화 기법(k-Means Clustering Method)

k-평균 군집화는 각 관측치의 최소평균거리를 이용하여, 전체 데이터를 k개의 군집으로 분할하는 군집기법의 하나이다(Dubes *et al*., 1988). k-평균 군집화 기법을 간단히 요약하자면, 먼저 k개의 초기점(seed point)을 임의로 선택하고, 각각의 관측치와 k개의 초기점과의 거리를 계산하여 해당 관측치를 가장 가까운 초기점에 배정하는 k개의 군집을 형성한다. k개의 군집이 생성되면 각 군집의 평균점을 새로 계산하고 새로이 결정된 평균점과 각각의 관측치와의 거리를 계산하여 새로운 군집을 형성한다. 위 단계를 반복하여 더 이상 평균점이 바뀌지 않아 군집형태가 동일하게 유지 될 때 최종 군집결과가 결정된다(Hartigan, 1975). k-평균 군집화 기법을 사용하기 위해서는 먼저 군집수(k)와 거리척도를 결정해야 한다. 군집수를 결정하는 방법들은 몇몇 존재하고 있으나(Gorden, 1999) 어느 하나의 방식이 좋다고 말하기는 어려우며 보통 문제에 대한 배경지식을 기반으로 방법을 결정한다. 거리계산방식은 유클리드(Euclidean), 맨하튼(Manhattan), 상관관계(Correlation) 등 다양한 방식이 있으나 데이터의 특성과 분석 목적에 맞게 결정을 한다.

3.3 연관규칙(Association Rule)

연관 규칙이란 동시에 발생하는 사건들의 규칙을 수치화한 것으로, 한 항목과 다른 항목 사이에 연관성을 찾아내는 방법이다. 데이터간의 연관성을 찾기 위해 사용되는 대표적인 지표로는 지지도(support), 신뢰도(confidence), 그리고 향상도(lift)가 있으며, 연관정도를 파악하고 해석하는데 유용한 역할을 한다(Tan *et al*., 2006).

<Table 5>에서 요약하고 있는 연관 규칙의 지표들을 이용하여 두 개의 관찰 대상 간 연관 정도를 수치로 보여줄 수 있다. 예를 들어 관찰 대상 'A'와 'B'가 있다고 가정할 때, 지지도는 'A'와 'B'가 동시에 출현할 확률로 $P(A \cap B)$로 표시할 수 있다. 신뢰도는 조건부확률로써 'A'를 포함하고 있는 관측 중에 'B'가 포함되어 있을 확률을 나타내며 즉, 선행사건 'A'가 출현했다는 가정 하에 후행사건 'B'가 포함되어 있을 조건부 확률 P(B|A)로써 표현한다. 향상도는 규칙을 모를 때에 비하여 규칙을 알 때 얼마나 주제어의 출현이 향상되는가를 나타낸다. 즉, 주제어 'B'를 연관규칙과 관계없이 관찰하는 것에 비하여 연관 규칙을 알고 'A'를 발견한 경우에 대하여 'B'가 관찰되는 경우 얼마나 관찰이 증가하는가를 나타낸다(Lee *et al*., 2005).

Table 5. Performance measures in an association rule algorithm

지표	수식
지지도 (support)	$P(A \cap B)$
신뢰도 (confidence)	$P(B\|A) = \frac{P(A \cap B)}{P(A)}$
향상도 (lift)	$\frac{P(B\|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$

3.4 사회연결망 분석 내 모듈성분석(Modularity in Social Network Analysis)

사회연결망(social network)은 하나 이상의 노드(node)들을 서로의 연관성을 고려하여 선 (links, edges, ties)으로 연결한 그림이다. 이러한 사회연결망은 복잡하고 다양한 관계의 의미를 효율적으로 보여줄 수 있으며 일반적으로 밀도(density), 중심성(centrality) 및 모듈성(modularity) 등으로 특성화할 수 있다(Hanneman *et al*., 2011).

본 논문에서 활용하는 모듈성분석(modularity analysis)은, 연결망을 구성하는 노드의 군집구조를 파악하는데 유용하게 사용되는 방법이다. 그룹 내의 링크가 그룹간의 링크보다 많도록 사회연결망을 관찰하여 연결관계가 많은 노드끼리 군집화하는 분석방법이며, CNM(Clauset-Newman-Moore) 알고리즘이 대표적으로 활용된다(Clauset *et al*., 2004, Bansal *et al*., 2010). CNM 알고리즘은 각 노드가 1개의 군집을 이루는 것으로부터 시작하여, 작은 군집을 하나씩 합쳐나가는 응집(Agglomerative) 알고리즘 중의 하나이다(Theodoridis *et al*., 2008).

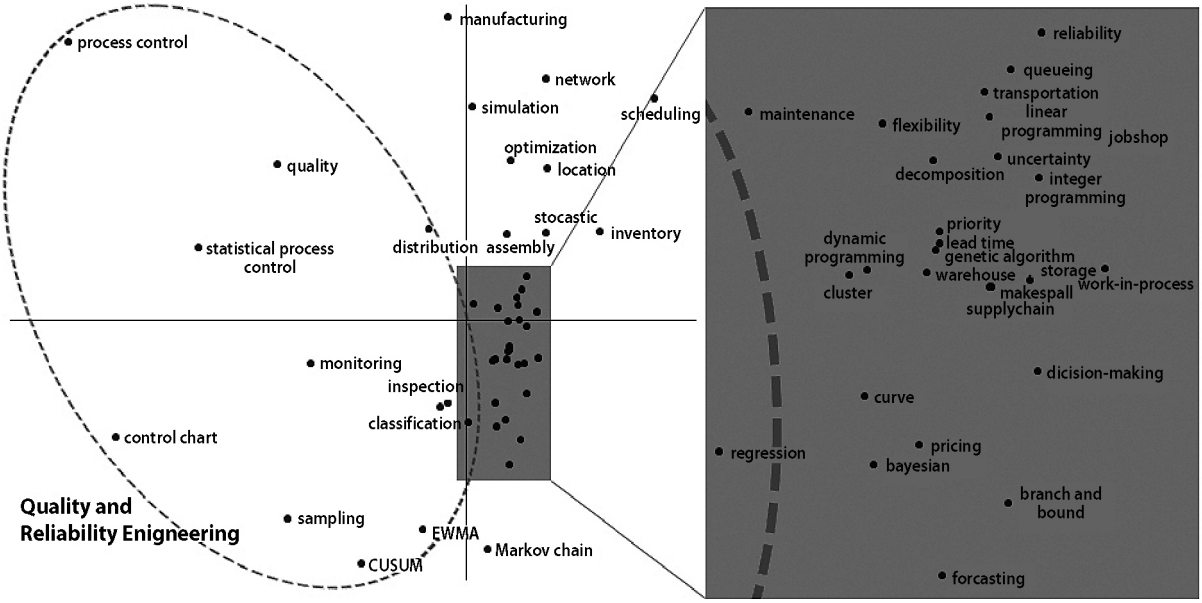

Figure 2. Visualization of 48keywords in the reduced dimensions by LLE

4. 결 과

4.1 시각화(Visualization)

앞서 제 3장에서 설명한 분석 방법 중, 지역선형사상은 고차원 데이터의 차원축소를 통한 시각화 기법이다. IIE Transactions의 논문 중, 관찰대상으로 선정된 2,527개의 논문의 초록에서 관찰되는 주제어의 출현빈도 데이터를 이용하였다. 지역선형사상(LLE)를 이용한 시각화의 결과는 <Figure 2>와 같다. 점선 영역 내에는 Quality and Reliability Engineering 분야의 주제어들이 주로 포함되어 있음을 볼 수 있으며 그 외의 분야(회색으로 확대된 부분)에도 어느 정도 그룹이 있음을 확인할 수 있다. 이는 지역성을 고려하는 지역선형사상의 특성이 반영된 것으로 볼 수 있다. 하지만 이 그룹들이 각각의 분야를 의미하는지에 대해서는 다소 이견이 있을 수 있는데, 그 이유는 Design and Manufacturing, Operations Engineering and Analysis 그리고 Scheduling and Logistics 분야에서는 주제어들이 한 분야에서만 독립적으로 사용되기보다, 분야 간 공통적으로 사용되는 경우가 많기 때문이라고 사료된다.

4.2 군집화(Grouping)

위에서 보여준 시각화기법은 고차원의 데이터를 저차원의 그림으로 보여준다는 점에서 의의가 있으나 대부분의 시각화 기법이 그렇듯, 결과에 대한 최종 해석이 주관적일 수밖에 없다는 단점이 있다. 시각화 기법과 더불어 객관적으로 데이터의 그룹 패턴을 살펴보기 위해 군집화 기법을 사용하였다. 모든 논문에 대한 주제어의 출현빈도 데이터를 이용한 k-평균 군집화의 결과는 <Table 6>과 같다. 군집수(k)는 IIE Transactions에 분류한 4가지 논문 분야 따라 군집수는 4개로 설정하고, 그룹

Table 6. A clustering result of 48 keywords by a k-means clustering algorithm

구 분	주제어	분류
군집 1 (Cluster 1)	classification, control chart, CUSUM, EWMA, inspection, Markov chain, monitoring, process control, quality, regression, sampling, statistical process control	Quality and Reliability Engineering
군집 2 (Cluster 2)	assembly, decision-making, decomposition, genetic algorithm, job shop, maintenance, manufacturing, network, optimization, queuing, reliability, simulation, work-in-process	Design and Manufacturing
군집 3 (Cluster 3)	branch and bound, cluster, curve, dynamic programming, flexibility, integer programming, linear programming, location, makespan, scheduling, storage, warehouse	Operations Engineering and Analysis
군집 4 (Cluster 4)	Bayesian, distribution, forecasting, Inventory, lead time, pricing, priority, stochastic, supply chain, transportation, uncertainty	Scheduling and Logistics

간 중심점을 찾기 위한 거리는 상관관계거리를 사용하였다. <Table 6>은 각 군집 내에 속해있는 주제어들을 보여주고 있으며, 마지막 열의 분류는 주제어들의 성향에 따라 임의로 정해 보았다. <Table 6>의 <군집 1>~<군집 4>에 포함되어 있는 주제어들을 보면 연구영역에 따른 분류가 어느 정도 명확하게 이루어졌음을 볼 수 있다. <군집 1>의 경우 'control chart', 'CUSUM', 'EWMA' 등 대부분의 주제가 품질공학/관리 관련 주제어임을 볼 수 있으며, <군집 2>의 경우에는 대다수의 주제어들이 제조와 설계에 관련된 주제임을 볼 수 있다. <군집 3>-'Scheduling and Logistics'과 <군집 4>-'Operation Engineering and Analysis'는 연구분야 자체가 명확히 구분이 되어 있지 않아 정확한 분류는 어렵지만, <군집 3>의 주제어들은 주로 기법에 관련이 되어있고 <군집 4>는 <군집 3>에서 명시된 기법을 응용하는 분야와 관 계가 있음을 볼 수 있다.

또 다른 군집분석으로 '논문 주제어의 연도별 빈도 데이터'를 이용한 k-평균 군집화를 수행하였다. 이는 논문이 처음 출간된 1969년부터 현재 2011년까지의 각 주제어들의 출현패턴을 알아가 보기 위함이다. 군집개수(k)는 3개, 4개, 5개로 나누어 분석해 보았고, 그 결과 3개로 할 때 가장 의미 있는 결과를 얻을 수 있었다. 거리방식으로는 주제어 출현빈도의 시계열 패턴을 반영하기 위해 상관관계거리를 이용하였다.

<Figure 3>는 군집화의 결과를 이용하여, 각 군집의 해당하는 주제들의 평균 출현정도를 연도별로 보여주고 있다. <군집 1>은 출현양의 추세가 완만하게 증가하고 있는 주제어들의 평균 시계열 패턴을 보여주고 있으며, <군집 2>는 2000년대 들어 급격히 증가하는 주제어들의 평균 시계열 패턴을 보여주고 있다. <군집 3>는 1987년 이후 관찰 정도가 감소하고 있는 연구 주제어의 평균 시계열 패턴을 보인다.

<Table 7>은 k-평균 군집화를 통해 구분된 3개 군집 내 주제어의 목록을 보여주고 있다. 각각의 군집은 <Figure 3>에 보여주고 있는 패턴인 '상승 주제어', '최근 상승 주제어', '하강 주제어'로 분류하였다. <군집 1>과 <군집 3>에 속한 주제어들은 각각 빈도수가 꾸준히 증가 또는 감소하는 것을 볼 수 있으며 <군집 2>에 속해 있는 주제어들은 2000년대 들어 빈도수가 급격히 증가하고 있음을 보여주고 있다. 논문의 초록이 연구의

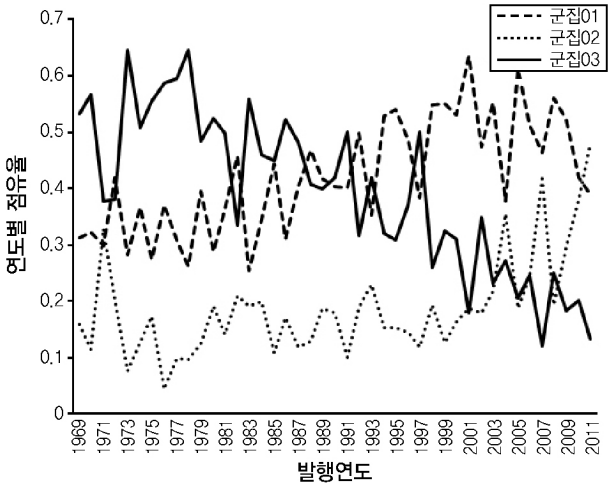

Figure 3. The frequency of keywords in each of 3 clusters over the time(1969~2011)

요약정보를 담고 있다는 점을 감안했을 때, <Table 7>은 연구자들에게 꾸준한 사랑을 받아온 주제어들, 또는 서서히 관심을 잃어가거나 최근 뜨겁게 탐구되는 연구 분야를 보여주고 있다. 이와 같이 시간에 따른 주제어 군집결과는 산업공학의 연구추이의 알아볼 수 있는 매우 흥미로운 자료로 사용될 수 있을 것이다.

4.3 주제어의 연관규칙

<Table 8>은 논문초록에서 동시에 출현하는 가능한 모든 쌍의 주제어간 상관관계를 발췌하여 지면상 그 일부를 정리한 표이다. 예를 들면, 논문의 초록에서 주제어 'CUSUM'과 'control chart'가 동시에 발견될 확률인 지지도는 0.0082로 작게 나왔지만, 선행주제어 'CUSUM'이 포함된 논문초록에서 후행주제어 'control chart'가 발견될 확률인 신뢰도는 0.7037로 매우 높았다. 또한 이 두 단어는 무작위로 문서를 관측하여 발견될 확률에 비해, 동시에 발견될 확률의 비율이 17.2482배로 상당히 큼을 알 수 있다. 다른 주제어의 쌍들도 위와 같은 방식으로 해석할 수 있다.

Table 7. A list of keywords in each of 3 clusters based on their patterns(increasing, recent increasing, and decreasing)

구 분	주제어	분류
군집 1 (Cluster 1)	Bayesian, classification, decomposition, dynamic programming, flexibility, forecasting, lead time, Markov chain, monitoring, network, process control, regression, reliability, statistical process control, uncertainty	상승 주제어
군집 2 (Cluster 2)	assembly, cluster, control chart, CUSUM, decision-making, distribution, EWMA, genetic algorithm, Inventory, location, maintenance, makespan, manufacturing, optimization, pricing, priority, quality, simulation, stochastic, supply chain, transportation	최근 상승 주제어
군집 3 (Cluster 3)	warehouse, branch and bound, curve, inspection, integer programming, job shop, linear programming, queueing, sampling, scheduling, storage	하강 주제어

Table 8. Support, confidence, and lift values from a part of the keyword pairs

선행 주제어	후행 주제어	지지도	신뢰도	향상도
CUSUM	control chart	0.0082	0.7037	17.2482
makespan	scheduling	0.0169	0.6842	4.6365
EWMA	control chart	0.0056	0.5417	13.2766
work-in-process	inventory	0.0074	0.5	3.4595
…	…	…	…	…

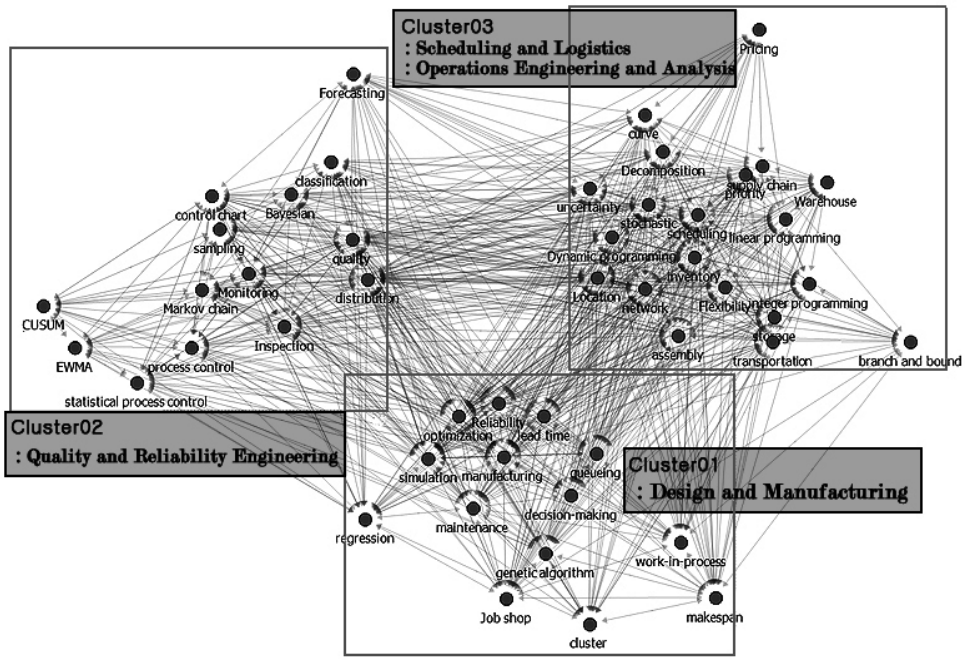

Figure 4. Social network and modularity analyses for 48 keywords

4.4 사회연결망 분석

<Figure 4>는 서로 다른 한 쌍의 주제어가 논문의 초록에서 동시에 관찰되는 횟수를 활용하여 구성된 사회연결망이다. 이 결과는 사회연결망 분석 소프트웨어인 넷마이너 4.0(www. netminier.com)을 활용하였으며, 응집(cohesion)분석 중 모듈성(modularity)분석을 실행하였다. 분석에 앞서, 소프트웨어의 활용을 위해 사전 선택된 옵션은 주제어 쌍의 동시관찰 횟수가 5회를 초과하는 데이터를 대상으로 하였으며, 모듈성 분석으로는 앞서 설명한 CNM 알고리즘을 사용하였다.

<Figure 4>에서 보여주고 있는 3개의 군집은 각각 군집 내의 링크가 서로 다른 군집간의 링크보다 많도록 형성되었다. 다시 말해, 그룹 내 포함된 주제어들은 다른 그룹의 주제어들보다 더 자주 관찰된다는 것을 의미한다. Quality, Distribution, Decision making, Location, Network 등의 주제어는 다른 군집의 주제어들과 서로 강한 상호 연관이 있음을 관찰할 수 있고 이와는 반대로 Statistical process control, Cluster, Branch and bound 등의 주제어들은 해당 군집의 주제어들과만 연결이 되어있는 고립된 주제어라고 할 수 있다.

5. 결 론

본 연구에서는 산업공학의 대표적 국제학술지인 IIE Transactions를 이용하여 지난 43년간 출판된 논문초록의 주제어간의 상관관계를 살펴보았다. 지역선형사상을 이용하여 고차원 데이터의 관계들을 효과적으로 시각화하고, k-평균 군집화를 활용하여 유사한 특성을 보이는 주제어들을 군집화 하였다. 또한 연관 규칙과 사회연결망분석을 이용하여 각 주제어들 사이의 관계를 수치화하고 시각화함으로써 정량적 분석을 위한 기초 연구 자료를 도출했다.

다양한 데이터마이닝 방법으로 논문초록에 포함된 키워드를 분석함으로써, 과거에서 현재에 이르는 산업공학 내 연구의 현황과 추이를 관찰할 수 있었다. 본 연구는 특정 분야 학술지를 대상으로 분석이 수행되었지만, 본 논문에서 제시한 분석절차 및 기법은 다른 분야에도 얼마든지 적용이 가능하다. 예를 들면, 대량의 정치기사에 포함된 정치인들의 상호관계 또는 특정 제품의 후기에 포함된 주제어 및 주제어 간의 관계를 살펴 볼 수 있을 것으로 판단된다. 본 연구가 웹 마이닝과 텍스트마이닝 영역의 다양한 연구를 활성화 시킬 수 있기를 기대해 본다.

참고문헌

Bansal, S., Bhowmick, S., and Paymal, P. (2010), Fast Community Detection For Dynamic Complex Networks, *Proceedings of the Second Workshop on Complex Networks*, CompleNet.

Breitzman, A. F. and Mogee, M. E. (2002), The many applications of patent analysis, **28**, 187-205.

Chakrabartij, S. (2002), *Mining The Web : Discovering Knowledge from Hypertext Data*, Morgan kaufmann publishers, 2-5, San Francisco, CA, USA.

Chen, Y.-S. and Chen, B.-Y. (2011), Utilizing patent analysis to explore the cooperative competition relationship of the two LED compnanies : Nichia and Osram, *Technological Forecasting and Social Change*, 78, 294-302.

Clauset, A., Clauset, Newman, M. E. J., and Moore, C. (2004), Finding community structure in very large networks, *Physical Review E*, **70**, 066111.

Clifton, Christopher. (2010), Definition of Data Mining, *Encyclopedia Britannica*.

Choi, Y. J. and Park, S. S. (2002), Interplay of Text Mining and Data Mining for Classifying Web Contents, *The Korea Society for Cognitive Science*, **13**, 1-60.

Dubes, R. C. and Jain, A. K. (1988), *Algorithm for Clustering Data*, Prentice Hall College Div, New Jersey, USA.

Gordon, A. D. (1999), *Classification*, Champman and Hall, New York, USA.

Hanneman. R. A. and Riddle, M. (2011), *Introduction to social network methods*, http://faculty.ucr.edu/~hanneman/.

Hartigan, J. A. (1975), *Clustering Algorithms*, John Wiley and Sons, New York, USA.

Hotelling, H. (1933), Analysis of a Complex of Statistical Variables into Principal Components, *Journal of Educational Psychology*, **24**, 417-441.

Jolliffe, I. T. (2002), *Principal Component Analysis*, *Second Edition*, Springer, New York, USA.

Lee, S., Lee, S., Seol, H., and Park, Y. (2008), Using patent information for designing new product and technology : keyword based technology roadmapping, R&D Management, **38**, 169-188.

Lee, T. R., Koo, J. Y., Park, H. J., Lee, K. H., and Choi, D.W. (2005), *Data Mining*, **2**, 162-163, Seoul, Korea.

Liu, B. (2011), *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*, Springe, 6-7, Berlin, Germany.

Theodoridis, S. and Koutroumbas, K. (2008), Pattern Recognition, **4**, 654-658.

Rowie, S. T. and Saul, L. K. (2000), Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, **290**.

Saul, L. K. and Roweis, S. T. (2000), An Intorduction to Locally Linear Embedding, http://cs.nyu.edu/~roweis/lle/publications.html.

Tan, P. and Steinbach, M. (2006), Introduction to Data Mining, Addison Wesley, Boston, USA.