# Identification of Research Patterns and Trends through Text Mining

Su Gon Cho and Seoung Bum Kim

***Abstract*—Identification of meaningful patterns and trends in large volumes of text data is an important task in various fields. 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 a low-dimensional embedding method, clustering analysis, association rule, and social network analysis to find meaningful associative patterns of the keywords frequently appeared in the paper.**

***Index Terms*—Data mining, text mining, clustering, association rule, social network analysis**

## I. Introduction

Identification of meaningful patterns and trends and the extraction of potential knowledge in large volumes of text data is an important task in various fields (Kao *et al*., 2007). In particular, the advent of high-speed internet generates large amounts of textual data in a variety of forms. However, it is difficult to analyse text data because of their irregularity and complexity. Text mining is the process of deriving useful and meaningful information from text. Unlike conventional data mining tasks that extract the pattern from structured databases, text mining is intended to explore relationship among the objects stored in unstructured databases. In general, text mining involves several steps: (1) construction of the structured database from unstructured text inputs, (2) extraction of patterns and trends from the structured data, and (3) evaluation and interpretation of the patterns and trends. A number of studies have been conducted to extract implicit information from large volumes of text data (Feldman and Sanger, 2007; Miller, 2005). Recently, Lee *et al*. (2008) attempted to identify emerging technology trend from a large volumes of patent documents.

In the present study we attempt to identify research trends and patterns of the Industrial Engineering area from an academic journal using text and data mining. We collected thousands of textual information from the journals through crawling and constructed the databases for analysis

The rest of the paper is organized as follows. Section 2 presents the data collection process and a list of keywords selected. Section 2 presents the methods used in the study. Section 3 presents the results. Finally, Section 4 gives some concluding remarks.

Manuscript received March 20, 2012; revised April 12, 2012.

The authors are with the School of Industrial Management Engineering, Korea University, Seoul, Korea (e-mail: sbkim1@korea.ac.kr.)

## II. Data

IIE Transactions is the one of the prestigious international journals in the Industrial Engineering field. IIE Transactions consists of 4 categories – Design and Manufacturing, Operations Engineering and Analysis, Quality and Reliability Engineering, and Scheduling and Logistics.

We extracted the titles, authors, abstracts, published years, affiliation, and keywords from each of 2,527 papers published from January 1969 to March 2011 through the journal website (www.tandf.co.uk/journals/titles/0740817x.asp). Moreover, we selected 48 important keywords frequently appeared in the abstract of the paper. Here is a list of important key words selected: 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, and work-in-process.

TABLE I: The Data Formats Extracted From the Abstracts of IIE Transactions Documents

(A) The Dataset Containing the Frequency of the Keywords From 2,527 Abstracts.

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

(B) The Dataset Containing the Frequency of the Keywords From 1969 To 2011.

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

Based on these keywords selected the high dimensional datasets were created. Table I show that the data formats for the subsequent analyses. Table I (a) shows the portion of data containing the frequency of the keywords from each of 2,527 abstracts of IIE Transactions papers. For example, ‘D0002’ contains the keywords ‘assembly’ once and the keyword ‘Bayesian’ appeared twice in the abstracts of paper, ‘D0001’.

Table I (b) shows the portion of data representing the frequency of the key words over time period from 1969 to 2011. For example, we could find the keywords 'assembly' twice in all abstracts of papers published in 1969. We utilized the dataset of Table I (a) to discover relationship between among the keywords, and the dataset of Table I (b) was used to identify research trends over the past 40 years.

## III. Methods

### A. Locally Linear Embedding (LLE)

Locally linear embedding (LLE) is the one of the dimensionality reduction methods to facilitate the visualization of high-dimensional data (Rowie *et al*, 2000). Unlike principal component analysis, based on linear projection, LLE can recover global nonlinear structure from locally linear fit. LLE is especially useful for nonlinear manifolds such as documents of text and images of faces (Rowie *et al*., 2000).

The LLE algorithm has two steps. Suppose the dataset consists of $n$ real-valued vector $X_i$, each of dimensionality $D$, sampled from an underlying manifold. In the first step, we determine the neighbours to each data point $X_i$ and compute the weights $W_{ij}$ that best linearly reconstruct $X_i$ from its neighbours by solving the constrained least-squares problem in Eq. (1).

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

In the second step, we obtain the low-dimensional embedding vectors $Y_i$, best reconstructed by $W_{ij}$, by minimizing Eq. (2).

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

Although the weights $W_{ij}$ and vectors $Y_i$ are computed by methods in linear algebra, the constraint that points are only reconstructed from neighbours can result in highly nonlinear embeddings (Saul *et al*., 2000).

### B. k-Means Clustering

*k*-Means clustering partitions *n* observations into *k* clusters in which each observation belong to cluster with the nearest mean (Dubes *et al*., 1988). The brief summary of the *k*-means clustering method is as follows: Given *k* initial points (that can arbitrarily determine), each observation is assigned to one of the *k* initial points close to the observation, creating *k* clusters. These initial points are then replaced with the mean of the clusters currently assigned. This procedure is repeated with updated with initial points until assignments do not change (Hartigan, 1975). The outcome of the *k*-means clustering depends on the number of clusters (*k*) and the distance metrics.

### C. Modularity Analysis in Social Network

Social network analysis provides a nice graphical representation to visualize relationships among the objects in terms of nodes and links. The graphical display resulting from social network analysis enables us to understand the whole relationship among the objects interested. The social network can be characterized by the following measures: degree, density, centrality, modularity, and many others (Hanneman *et al*, 2011). In the present study, we conducted modularity analysis in social network to divide the whole network into subgroups for the purpose of summarization and improved understanding (Bansal *et al*., 2010). The groups are formed based on the strength of associations between the objects. Modularity is often used in optimisation methods for detecting community structure in networks.

## IV. Results

### A. Visualization

We applied LLE to mapping high-dimensional original data (2,527 dimensions) into low-dimensional embedding. Two-dimensional embedding space of 48 keywords discovered by LLE is shown in Fig. 1. The result shows that the keywords related to "Quality and Reliability Engineering" are grouped together although they are somewhat scattered. Other keywords are also grouped based on the similarity of research topics.

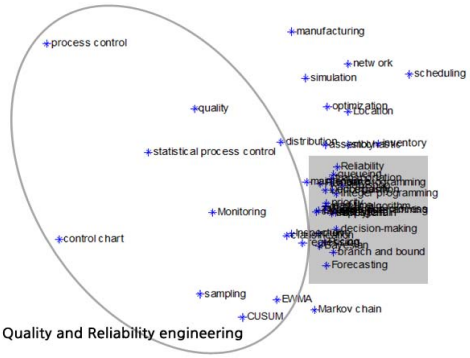


Fig. 1. Visualization of 48 keywords in the reduced dimensions by LLE

### B. Clustering

*k*-means clustering was performed on the dataset containing the frequency of the keywords over time. We set $k$=3 and used the correlation distance that measures the similarity in patterns between the two time series profiles. Fig. 2 shows the mean profiles of three clusters from *k*-means clustering analysis. The result shows the three distinct patterns: cluster 1: steady increase, cluster 2: recent rapid increase, and cluster 3: continuing decrease. A list of the keywords that belongs to each cluster is shown in Table II. This result can provide an understanding of research trends in the Industrial Engineering field over the past 40 years.

### C. Modularity Analysis Result in Social Network

Fig. 3 is the result for social network analysis using 48 important keywords. NetMininer 4 (www.netminier.com) was used to generate social network and to perform modularity analysis for detecting network groups. Three community structures were observed in the social network analysed. If we take a closer look at inside of clusters in Fig. 3, the role of each keyword and relationship among the keywords could be identified by the location. For examples, keywords 'distribution' and 'quality' in Cluster 2 are related

with the keywords in other clusters. Note that the keywords 'CUSUM' and 'EWMA' are isolated in 'Quality and Reliability Engineering' group because these terms are specifically used in the Quality area.

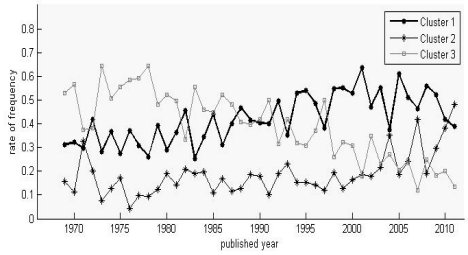


Fig. 2. The mean profiles of three clusters of the keywords observed over the past 40 years (1969-2011)

TABLE II: A LIST OF KEYWORDS IN EACH OF THREE CLUSTERS BASED ON THEIR PATTERNS (INCREASING, RECENT INCREASING, AND DECREASING)

| Cluster | Keywords | Category |
|---|---|---|
| Cluster 1 | Bayesian, classification, decomposition, dynamic programming, flexibility, forecasting, lead time, Markov chain, monitoring, network, process control, regression, reliability, statistical process control, uncertainty | Steady Increase |
| 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 | Recent Rapid Increase |
| Cluster 3 | warehouse, branch and bound, curve, inspection, integer programming, job shop, linear programming, queueing, | Continuing Decrease |

## V. CONCLUSIONS

In the present paper, we revealed research trends and patterns of the Industrial Engineering field from one of the representative journals in Industrial Engineering by using text mining. We employed the dimensional reduction method, clustering analysis, and social network analysis to draw out meaningful patterns of the keywords and research trends over time. Social network analysis visualizes and summarizes association patterns of the keywords. We hope this paper will stimulate further investigation in applying appropriate text and data mining tools to various applications in both academia and industry.

## ACKNOWLEDGMENT

This work was supported by Brain Korea 21 (Networked Enterprise).


Fig. 3. Social network and modularity analyses for 48 keywords

## REFERENCES

[1] A. Kao and S. R. Poteet (Eds), "Natural Language Processing and Text Mining," Springer, London, UK. 2007.
[2] S. Lee, S. Lee, H. Seol, and Y. Park, "Using patent information for designing new product and technology: keyword based technology roadmapping," *R and D Management*, pp. 169-188, 2008.
[3] S. T. Rowie and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, 2000.
[4] L. K. Saul and S. T. Roweis. (2000). An Intorduction to Locally Linear Embedding. [Online]. Available: http://cs.nyu.edu/~roweis/lle/publications.html
[5] R. C. Dubes and A. K. Jain, *Algorithm for Clustering Data*, Prentice Hall College Div, 1988.
[6] J. A. Hartigan, *Clustering Algorithms*, John Wiley and Sons, 1975, New York, USA.
[7] R. A. Hanneman and M. Riddle. (2011). Introduction to social network methods. [Online]. Available: http://faculty.ucr.edu/~hanneman/
[8] S. Bansal, S. Bhowmick, and P. Paymal, "Fast Community Detection For Dynamic Complex Networks," in *Proc. of the Second Workshop on Complex Networks*, CompleNet, 2010.
[9] T. W. Miller, *Data and Text Mining: A Business Applications Approach*, Prentice Hall, 2005.
[10] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approach in Analyzing Unstructured Data*, Cambridge Univ. Press, 2007.