




A sequential pattern mining approach to identifying potential areas for business diversification

Gyumin Lee, Daejin Kim & Changyong Lee

To cite this article: Gyumin Lee, Daejin Kim & Changyong Lee (2019): A sequential pattern mining approach to identifying potential areas for business diversification, Asian Journal of Technology Innovation, DOI: [10.1080/19761597.2019.1693900](https://doi.org/10.1080/19761597.2019.1693900)

To link to this article: <https://doi.org/10.1080/19761597.2019.1693900>




View supplementary material 



Published online: 24 Nov 2019.



Submit your article to this journal 



View related articles 



View Crossmark data 



A sequential pattern mining approach to identifying potential areas for business diversification

Gyumin Lee^a, Daejin Kim^b and Changyong Lee^a

^aSchool of Management Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea; ^bSchool of Business Administration, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

ABSTRACT

Although many quantitative models have been presented to identify potential areas for business diversification, most have focused on the assessment of technological capabilities and/or similarities using patent information. New data sources and scientific methods have thus seldom been addressed. We propose a sequential pattern mining approach to identifying potential areas for business diversification using the historical business segment data. Our approach includes (1) sequential pattern mining to identify potential areas for business diversification by extracting the significant changing patterns of firms' business segments; and (2) index analysis to assess the market and financial characteristics of the areas identified. Taken together, three diversification strategy maps are developed to provide comprehensive views of analysis results. An empirical analysis of 25,126 unique firms with 1320 business segments confirms that the proposed approach enables a wide-ranging search for potential areas for business diversification and the quick assessment of their characteristics.


KEYWORDS

Business diversification; sequential pattern mining; index analysis; diversification strategy map; historical business segment data

1. Introduction

Business diversification is a pivotal strategy for organisations to recreate and enlarge their competencies (Breschi, Lissoni, & Malerba, 2003). While prior studies have identified many firm- and industry-specific success factors such as the spread of risk (Amit & Livnat, 1988), reusing of core competencies (Luo, 2002), and achievement of an economy of scale (Nath, Nachiappan, & Ramanathan, 2010), business diversification has been considered as a complex and risky task (Leten, Belderbos, & Van Looy, 2007). In practice, managers typically depend on task force teams comprising of internal and/or external experts to identify potential areas for business diversification and to assess their characteristics (Larrodé, Moreno-Jiménez, & Muerza, 2012). However, such expert-centric approaches are time-consuming and labour-intensive (Kim, Hong, Kwon, & Lee, 2017), and moreover, have difficulties in defending judgements rationally

CONTACT Changyong Lee ✉ changyong@unist.ac.kr  School of Management Engineering, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulsan 44919, Republic of Korea

 Supplemental data for this article can be accessed <https://doi.org/10.1080/19761597.2019.1693900>.

© KOSIME, ASIALICS, STEPI 2019

when others question analysis results (Shin, Coh, & Lee, 2013). Hence, industrial practitioners demand high-quality and well-organised information based on quantitative data and scientific methods to assist decision making regarding business diversification.

Patent analysis has been widely employed to identify potential areas for business diversification based on organisational technological capabilities and/or technological similarities between the current and potential business areas. Various models and methods have been suggested, such as collaborative filtering-based patent analysis (Lee & Lee, 2017), DEA and text mining-based patent analysis (Seol, Lee, & Kim, 2011), and link analysis of patents and trademark databases (Kim et al., 2017). However, while these models and methods have proved quite useful for different purposes, the results of previous studies are not specific about potential areas for business diversification because patents do not explicitly contain market and industry information (Lee, Yoon, Lee, & Park, 2009). Moreover, previous studies cannot provide insight into dynamic aspects of business diversification because they are limited to the assessment of technological capabilities and/or relationships at a certain time (Kim et al., 2017).

To counter these problems, this study proposes a sequential pattern mining approach to identifying potential areas for business diversification (see Table 1). We use the Compustat historical business segment database for the following reasons. First, this database provides ample information on business segments or product lines of over 70% of the North American firms (Standard and Poor, 2002). Second, this database provides historical information available back to 1976 and is thereby appropriate for identifying significant patterns for business diversification. Finally, this database provides consistent and accurate information across firms and time and can be linked to other types of databases such as patent databases provided by United States Patent and Trademark Office (Hall, Jaffe, & Trajtenberg, 2005) and financial transactions databases (Jagannathan, Stephens, & Weisbach, 2000).

At the heart of the proposed approach are (1) sequential pattern mining to identify potential areas for business diversification by extracting the significant changing patterns of firms' business segments at the industry level; and (2) index analysis to assess the market and financial characteristics of the areas identified. Sequential pattern mining identifies interesting subsequences in a sequence database (Agrawal & Srikant, 1995). This method is considered appropriate for this research since it can model different changing patterns of firms' business segments such as entering new business areas and quitting existing businesses and can consider the direction of diversification. It should be noted that although a company has business area A and can potentially diversify into area B, this situation does

Table 1. Comparisons of previous studies and the current research.

Factor	Previous studies	Current research
Approach	Technology intelligence	Competitor intelligence
Data	Mainly patent information	Historical business segment information
Level of analysis	Technology level (e.g. patent class)	Industry level (4-digit standard industry classification code)
Focus of analysis	Technological capabilities and similarities	Significant changing patterns of firms' business segments
Method	Mainly patent network analysis	Sequential pattern mining and index analysis
Results and implications	Potential areas for business diversification and their technological characteristics	Potential areas for business diversification and their market and financial characteristics

not imply that the company owning business area B could diversify into area A (Kim et al., 2017). Moreover, this method can calculate the time taken for establishing a diversification strategy which provides implications on the feasibility of business diversification (Garcia-Vega, 2006). As for the index analysis, it is important for organisations to evaluate and prioritise potential areas for business diversification to allow more detailed investigation. Ten indexes are defined and measured to examine the market and financial characteristics of the areas identified. Putting together, three diversification strategy maps – a diversification feasibility map, a financial activity map, and an operational strategy map – are developed to provide comprehensive and balanced views of analysis results.

An empirical analysis of 25,126 unique firms with 1320 business segments from 1976 to 2016 confirms that the proposed approach enables a wide-ranging search for potential areas for business diversification and the quick assessment of their characteristics. The proposed approach is expected to be a useful complementary tool for strategic decision making regarding the feasibility of business diversification.

2. Data

The historical segment data of the Compustat database is employed in this study. The segment data provided in the database is self-reported by companies following the issuance of the ‘Financial Reporting for Segments of Business Enterprise 14’.¹ Specifically, the data contains four different types of segment information including business, geographic, operating, and state segments (Standard and Poor, 2002). Among them, this study identifies each company’s diversification information from the business segment data. The business segment data contains 4-digit standard industrial classification (SIC) codes corresponding to the industries in which a company is operating each year, allowing to identify particular industries where the company has entered or withdrawn (Office of Management and Budget, 1987).

3. Methodology

Figure 1 describes the overall process of the proposed approach. Considering the inputs, throughputs, and outputs of analysis, the proposed approach is designed to be executed in four steps: (1) data collection and transformation; (2) identification of potential areas for business diversification; (3) assessment of potential areas for business diversification; and (4) development of diversification strategy maps.

3.1. Data collection and transformation

This step constructs a business trajectory matrix (Table 2(a)) and a firm-financial characteristics matrix (Table 2(b)) using the historical business segment data based on certain search conditions. First, the business trajectory matrix is filled out with all the business segments for each firm and each time and presents the changes of firms’ business segments over time. In the table, firms are distinguished by gvkey which is the unique 6-digit identifier for each firm defined by the Compustat database; time is represented as reporting date; and business segments are classified by SIC codes which are the 4-digit hierarchical codes composed of division, major group, and industry group. Second, the firm-financial

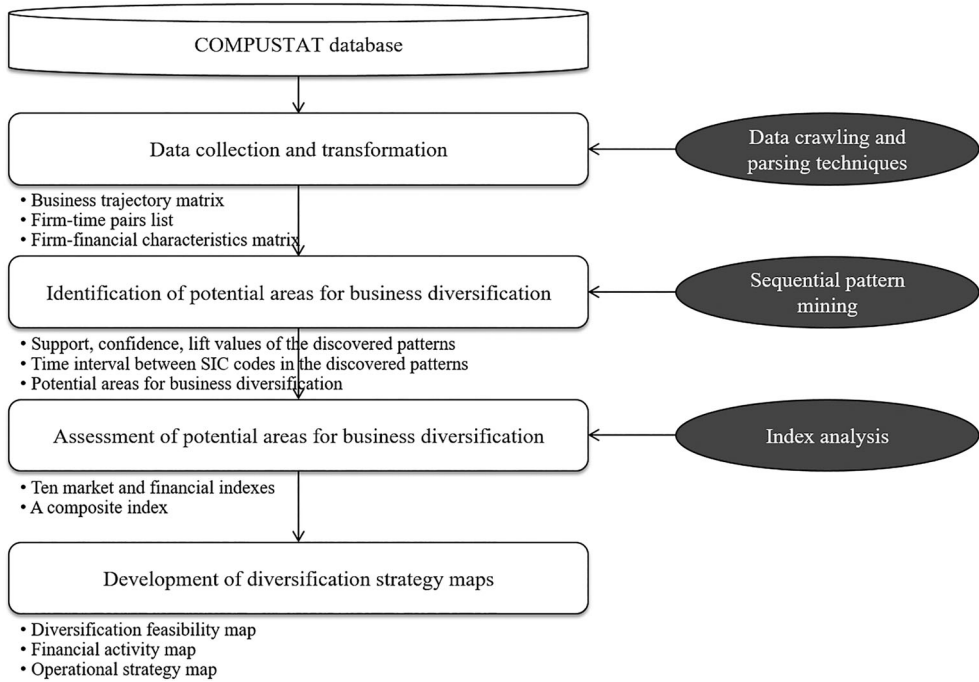


Figure 1. Overall process of the proposed approach.

characteristics matrix includes ten financial indexes regarding profitability, activity, and investment status for each firm for each business segment. The detailed explanations about these indexes are given in Section 3.2.3.

3.2. Identification of potential areas for business diversification via sequential pattern mining

We employ sequential pattern mining to extract the significant changing patterns of firms' business segments from the business trajectory matrix. Sequential pattern mining, which is derived from association rule mining, is an unsupervised data mining technique to find out interesting subsequences in a sequence database (i.e. a list of transactions in order time) (Agrawal, Imieliński, & Swami, 1993; Agrawal & Srikant, 1995). This method generates sequential patterns $X \rightarrow Y$ which indicates 'X is followed by Y', for instance, firms enter business area X and then Y.

Similar to association rule mining, there are three general measures of sequential patterns $X \rightarrow Y$: support, confidence, and lift (Srikant & Agrawal, 1996). First, the support of $X \rightarrow Y$, $\text{sup}(X \rightarrow Y)$, is defined as the fraction of subsequences that satisfy the order that X is followed by Y against the total sequences, as shown in Equation (1), and indicates the significance of the discovered patterns.

$$\text{sup}(X \rightarrow Y) = \frac{|\{s \in S; X, Y \subseteq s; X \text{ appear before } Y\}|}{|S|} \quad (1)$$

where s is a subsequence, S represents the whole set of sequences, and X and Y are items

Table 2. Forms of a business trajectory matrix and a firm-financial characteristics matrix.

Firm	Time		Business segments									
(a) Business trajectory matrix												
Gvkey ₁	Date _{1,1}	SIC _{1,1,1}	SIC _{1,1,2}									
Gvkey ₁	Date _{1,2}	SIC _{1,2,1}	SIC _{1,2,2}									
...										
Gvkey ₂	Date _{2,1}	SIC _{2,1,1}	SIC _{2,1,2}	SIC _{2,1,3}								
Gvkey ₂	Date _{2,2}	SIC _{2,2,1}	SIC _{2,2,2}	SIC _{2,2,3}								
Gvkey ₂	Date _{2,3}	SIC _{2,3,1}	SIC _{2,3,2}	SIC _{2,3,3}	SIC _{2,3,4}							
...										
Gvkey _n	Date _{n,t-1}	SIC _{n,t-1,1}	SIC _{n,t-1,2}	...	SIC _{n,t-1,k}							
Gvkey _n	Date _{n,t}	SIC _{n,t,1}	SIC _{n,t,2}	...	SIC _{n,t,k}							
			Profitability		Activity		Investment					
Firm	Time	Business Segment	ROA1	ROA2	PMR1	PMR2	ATR	FAT	CEX	RD	INV1	INV2
(b) Firm-financial characteristics matrix for each SIC code												
F ₁	Date _{1,1}	SIC _{1,1,1}	ROA _{1,1,1}	ROA _{2,1,1}	PMR _{1,1,1}	PMR _{2,1,1}	ATR _{1,1,1}	FAT _{1,1,1}	CEX _{1,1,1}	RD _{1,1,1}	INV _{1,1,1}	INV _{2,1,1}
		SIC _{1,1,2}	ROA _{1,1,2}	ROA _{2,1,2}	PMR _{1,1,2}	PMR _{2,1,2}	ATR _{1,1,2}	FAT _{1,1,2}	CEX _{1,1,2}	RD _{1,1,2}	INV _{1,1,2}	INV _{2,1,2}
F ₁	Date _{1,2}	SIC _{1,2,1}	ROA _{1,2,1}	ROA _{2,2,1}	PMR _{1,2,1}	PMR _{2,2,1}	ATR _{1,2,1}	FAT _{1,2,1}	CEX _{1,2,1}	RD _{1,2,1}	INV _{1,2,1}	INV _{2,2,1}
		SIC _{1,2,2}	ROA _{1,2,2}	ROA _{2,2,2}	PMR _{1,2,2}	PMR _{2,2,2}	ATR _{1,2,2}	FAT _{1,2,2}	CEX _{1,2,2}	RD _{1,2,2}	INV _{1,2,2}	INV _{2,2,2}
...
F _n	Date _{n,t-1}	SIC _{n,t-1,1}	ROA _{1,n,t-1,1}	ROA _{2,n,t-1,1}	PMR _{1,n,t-1,1}	PMR _{2,n,t-1,1}	ATR _{n,t-1,1}	FAT _{n,t-1,1}	CEX _{n,t-1,1}	RD _{n,t-1,1}	INV _{1,n,t-1,1}	INV _{2,n,t-1,1}
	
		SIC _{n,t-1,k}	ROA _{1,n,t-1,k}	ROA _{2,n,t-1,k}	PMR _{1,n,t-1,k}	PMR _{2,n,t-1,k}	ATR _{n,t-1,k}	FAT _{n,t-1,k}	CEX _{n,t-1,k}	RD _{n,t-1,k}	INV _{1,n,t-1,k}	INV _{2,n,t-1,k}
F _n	Date _{n,t}	SIC _{n,t,1}	ROA _{1,n,t,1}	ROA _{2,n,t,1}	PMR _{1,n,t,1}	PMR _{2,n,t,1}	ATR _{n,t,1}	FAT _{n,t,1}	CEX _{n,t,1}	RD _{n,t,1}	INV _{1,n,t,1}	INV _{2,n,t,1}
	
		SIC _{n,t,k}	ROA _{1,n,t,k}	ROA _{2,n,t,k}	PMR _{1,n,t,k}	PMR _{2,n,t,k}	ATR _{n,t,k}	FAT _{n,t,k}	CEX _{n,t,k}	RD _{n,t,k}	INV _{1,n,t,k}	INV _{2,n,t,k}

or item sets. Second, the confidence of $X \rightarrow Y$, $\text{conf}(X \rightarrow Y)$, is defined as $\text{sup}(X \rightarrow Y)$ divided by the number of subsequences containing the item X , $\text{sup}(X)$, as shown in Equation (2), and indicates the reliability of the discovered patterns.

$$\text{conf}(X \rightarrow Y) = \frac{\text{sup}(X \rightarrow Y)}{\text{sup}(X)} \quad (2)$$

Finally, the lift of $X \rightarrow Y$, $\text{lift}(X \rightarrow Y)$, is defined as the ratio of $\text{sup}(X \rightarrow Y)$ to that expected in the case of X and Y being independent, as shown in Equation (3), and represents the statistical dependence between X and Y in the discovered patterns.

$$\text{lift}(X \rightarrow Y) = \frac{\text{sup}(X \rightarrow Y)}{\text{sup}(X) * \text{sup}(Y)} \quad (3)$$

Based on the three measures in Equations (1–3), the process of discovering significant changing patterns of firms' business segments consists of two steps. In the first step, all the frequent changing patterns over the prescribed threshold values for support and confidence, minsup and minconf , are generated. In the second step, the patterns having a lift value greater than one are selected as significant. Given the significant pattern $X \rightarrow Y$, we consider Y as a potential area for business diversification from X .

Table 3 summarises the types of changing patterns extracted in this study. In Table 3, $(t_{\min}, t_{\text{avg}}, t_{\max})$ represents the minimum, average, and maximum time interval between business segments in the discovered pattern, respectively. Here, A and B can be not only a single segment but also multiple segments (e.g. $(\text{SIC}_1, \text{SIC}_2) \rightarrow (\text{SIC}_1, \text{SIC}_3)$). Moreover, complex patterns with the length greater than two can be identified (e.g. $\text{SIC}_1 \rightarrow \text{SIC}_2 \rightarrow \text{SIC}_3$).

3.3. Assessment of potential areas for diversification via index analysis

We employ ten indexes to examine the market and financial characteristics of the areas identified in the preceding step. Among them, four indexes (ROA1, ROA2, PMR1, PMR2) are related to profitability, two (ATR, FAT) are related to activity, and the remaining four (CEX, RD, INV1, INV2) are related to investment. For a given year, we calculate industry-level (i.e. 4-digit SIC code) index values by computing the equal-weighted averages of indexes. The followings provide the definitions of indexes as well as their implications.

(1) Profitability

- Return on asset 1 (ROA1): The first type of return on assets is defined as net income divided by total assets. The net income represents the income or loss after expenses and losses have been subtracted from all revenues and gains.
- Return on asset 2 (ROA2): The second type of return on assets is defined as operating income divided by total assets. Operating income represents sales of each

Table 3. Types of changing patterns of firms' business segments.

Changing patterns of firms' business segments				Meaning
1	A	$(t_{\min}, t_{\text{avg}}, t_{\max})$	B	Switch from industry A to B
2	(A,B)	$(t_{\min}, t_{\text{avg}}, t_{\max})$	B	Exit industry A
3	A	$(t_{\min}, t_{\text{avg}}, t_{\max})$	(A,B)	Enter industry B

segment minus its allocated share of operating costs and expenses. Since operating income is computed by including only items related to a firm's own business, it accounts for all revenues and expenses necessary to keep the business running.

- Profit margin ratio 1 (PMR1): The third measure of firm profitability is the profit margin ratio defined as net income divided by sales. This measure is similar to the ROA1 but differs in that it changes the denominator from total asset to sales.
- Profit margin ratio 2 (PMR2): The final measure of this category is defined as operating income divided by sales.

(2) Activity

- Asset turnover ratio (ATR): The first type representing firm activity is asset turnover ratio defined as sales divided by total assets.
- Fixed asset turnover (FAT): The second type of firm activity measure is fixed asset turnover ratio defined as sales divided by property, plant, and equipment. The property, plant, and equipment represent the cost of tangible fixed property used in the production of revenue.

(3) Investment

- Capital expenditure (CEX): The first measure measuring investment activity is capital expenditure defined as firm capital expenditures divided by total assets. The capital expenditures are resources used for additions to the firm's property, plant, and equipment, excluding amounts coming from acquisitions. Thus, this measure represents a firm's spending to purchase assets necessary for business.
- R&D intensity (RD): The second investment measure is R&D intensity defined as R&D spending divided by total assets. If a firm does not report R&D expenditure, which is missing in the database, we treat it as zero R&D expenditure. The R&D expenditure represents the estimated costs for the development of new products or services.
- New investment ratio (INV1): The third investment measure is new investment ratio defined as the sum of capital expenditures and R&D expenditures divided by total assets. This measure is simply the sum of capital expenditure ratio and R&D intensity.
- Total investment ratio (INV2): The final investment measure is total investment ratio defined as a new investment ratio plus depreciation and amortisation amounts scaled by total assets. The depreciation and amortisation represent non-cash charges for allocation of the current portion of capitalised expenditures and depletion charges.

Ten industry-level indexes can be compiled into a composite index to assess the overall attractiveness of the areas by examining the relative importance of individual indexes. Moreover, the trends of the market and financial characteristics of the areas can be assessed by examining the computed index values over time.

3.4. Development of diversification strategy maps

Three diversification strategy maps are developed to provide comprehensive and balanced views of analysis results (see [Figure 2](#)). First, the diversification feasibility map explores implications on the feasibility of business diversification. This map utilises the intensity

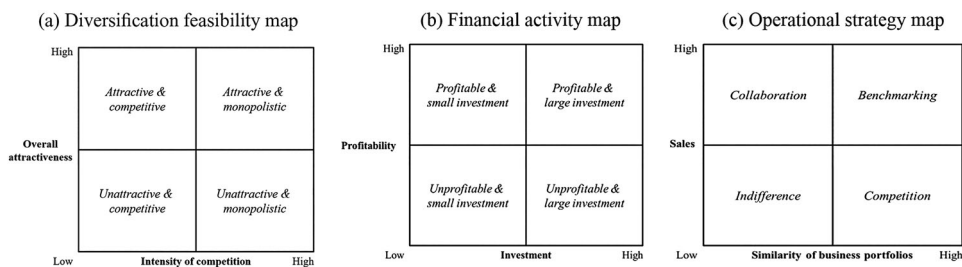


Figure 2. Diversification strategy maps.

of competition and the overall attractiveness of the potential areas. Specifically, we employ one minus the Hirschman-Herfindahl index (HHI) value and the composite index value for each potential area as a measure of the competition intensity and the overall attractiveness, respectively. Figure 2(a) presents an example of the diversification feasibility map which describes four categories of the potential areas: *attractive & competitive*, *attractive & monopolistic*, *unattractive & competitive*, and *unattractive & monopolistic*. Second, the financial activity map provides implications on the market and financial characteristics of the potential areas for business diversification. It employs the profitability and investment index values for the corresponding area and utilises the activity index value as the size of a marker. Figure 2(b) presents an example of the financial activity map which describes four categories of the potential areas: *profitable & small investment*, *profitable & large investment*, *unprofitable & small investment*, and *unprofitable & large investment*. Finally, the operational strategy map offers implications on the actors in a certain potential area. Given a target firm, the map utilises the sales of individual firms and the similarity value of business portfolios between the target firm and other firms that are already involved in the area. Figure 2(c) provides an example of the operational strategy map where each firm is categorised into one of four types: *collaboration*, *benchmarking*, *indifference*, and *competition*.

4. Empirical analysis and results

4.1. Data collection and transformation

We collected the historical business segment data covering 25,126 unique firms with 1320 business segments from 1976 to 2016. There are 421,767 firm-year observations. Then, the business trajectory matrix and firm-financial characteristics matrix were constructed. The business trajectory matrix (Table 4(a)), which was created by gathering business segments reported on the same date, traces the history of business segments that each firm belongs to and identifies the time when the firm changes its business areas. For more efficient computation, the matrix was transformed into a firm-time pairs list for each business segment by collecting pairs of firm identifiers and reporting dates according to each business segment, as shown in Table 4(b). In the firm-time pairs list, the pairs are arranged in order of time according to each business segment. Finally, Table 4(c) reports a part of the firm-financial characteristics matrix used to calculate the industry-level indexes and to develop three diversification strategy maps.

Table 4. Results of data collection and transformation.

Company	Date	Business segments					
		(a) Part of the business trajectory matrix					
1004	19780531	4582	5088				
1004	19790531	4582	5088				
...				
1004	19800531	4582	5088	5084			
...			
1004	19820531	3443	5088				
1004	19830531	3443	4581	5088			
...			
315318	20131231	2821	2899				
315318	20141231	2821	2879	2899			
315318	20151231	2879	2899				
...			
318815	20131231	2836					
318815	20141231	2836					
318815	20151231	2836					
Business segment		Firm-time pairs list					
(b) Part of the firm-time pairs list							
100	001266–1992	001266–1993	001266–1994	...	187285–2015	221127–1998	221127–1999
110	008274–1984	015490–1995	015490–1996	015490–1997			
111	160362–2011	175994–2010	175994–2011	...	179846–2013	179846–2014	179846–2015
112	158133–2007	158133–2008	158133–2010	...	186275–2013	186275–2014	186275–2015
115	002249–1990	002249–1991	002249–1992	...	178795–2015	178795–2016	201336–2001
...
8999	001165–1984	001165–1985	002519–1978	...	165698–2003	165698–2005	165698–2006
9223	110772–1998						
9229	003297–2001	003297–2002	003297–2002	...	026021–2013	026021–2014	026021–2015
9510	011771–1998	011771–1999					
995	001050–1988	001050–1989	001050–1990	...	224796–2010	224796–2013	224796–2014

(Continued)

Table 4. Continued

			Profitability				Activity		Investment			
Firm	Time	Business Segment	ROA1	ROA2	PMR1	PMR2	ATR	FAT	CEX	RD	INV1	INV2
(c) Part of the firm-financial characteristics matrix												
1004	19770531	5088	0.121		0.052	2.321		0.019	0.000	0.0191	0.0330	0.121
		4582	0.154		0.112	1.368		0.041	0.000	0.0406	0.0580	0.154
...
1004	20160531	5088					1.508		0.017	0.000	0.017	0.045
		4582					0.647		0.136	0.000	0.136	0.201
...
29173	19921231	1311	0.173		0.401	0.432		0.358	0.000	0.358	0.553	
...
29173	20151231	1311	−0.705	−1.056	−2.159	−3.233	0.327	0.390	0.366	0.000	0.366	0.558
29173	20161231	1311	−0.285	−0.202	−0.836	−0.593	0.341	0.445	0.318	0.000	0.318	0.451
...
318815	20131231	2836	−0.789	−0.870	−27.742	−30.561	0.028	0.215	0.009	0.372	0.381	0.429
318815	20141231	2836	−0.146	−0.153	−46.939	−49.053	0.003	0.104	0.020	0.072	0.092	0.099
318815	20151231	2836				−7.222						

The descriptive statistics of the diversification cases are summarised as follows: (1) 11,255 firms have experienced diversification at least once, and each firm diversified 1.27 times on average. For example, HRG Group is the most frequently diversified firm that has diversified 24 times; (2) business segments have been diversified a total of 109,350 times; a maximum of 2349; an average of 82.84, and a median of 33 times. Since a firm can enter more than two business segments at the same time, the number of diversification cases at the business segment level exceeds that at the firm level. The most frequently diversified business segment is the crude petroleum and natural gas industry; and (3) each diversification can be divided into two types, i.e. related and unrelated diversification, according to the first two digits of the SIC codes corresponding to business segments (Wernerfelt & Montgomery, 1998). There exist 21,554 and 87,796 cases of related and unrelated diversification. However, these numbers do not consider the cases classified as unrelated but actually related diversification that has advantages of reusing core competencies and achieving an economy of scale. For instance, pattern 1311 (Crude petroleum and natural gas) \rightarrow 4922 (Natural gas transmission) is categorised as unrelated diversification but can also be interpreted as related diversification based on vertical integration, as the latter is the distribution channel of the former (Lemelin, 1982).

4.2. Identification of potential areas for diversification via sequential pattern mining

We discovered the significant changing patterns of firms' business segments via sequential pattern mining and identified the minimum, average, and maximum time taken for each diversification by computing the time intervals between business segments in each extracted pattern. Considering the computational efficiency of different sequential pattern mining algorithms such as GSP, PrefixSpan, and SPAM, we employed the cSPADE algorithm in this study. The distinct characteristic of the cSPADE algorithm, *vis-à-vis* others, is the use of a vertical id-list corresponding to the firm-time pairs list. The use of a vertical id-list enhances the computational efficiency of generating the significant changing patterns and calculating the time intervals between business segments. The detailed process of extracting the significant changing patterns of firms' business segments is as follows:

- (1) All the possible frequent k -sequences are generated. The frequent 1-sequence is the same as each business segment, i.e. individual SIC code forming the firm-time pairs list. The frequent 2-sequences are computed via a temporal join between each of the 1-sequences. The frequent k -sequences can be computed by using $(k-1)$ -sequences in the same way.
- (2) During the first step, infrequent sequences are removed based on the preposition that all subsequences of a frequent sequence are frequent (Mannila, Toivonen, & Verkamo, 1995; Srikant & Agrawal, 1996). When generating a new k -sequence, if all its subsequences of length $(k-1)$ are frequent, the joining is performed. Otherwise, the sequence is dropped from consideration.
- (3) The support, confidence, and lift values of the generated k -sequences are calculated, as represented in Equations (1–3). The subsequences under the specified *minsup* and *minconf* are ruled out from the final pattern set.

During the process described above, determining the cut-off values for support and confidence (i.e. *minsub* and *minconf*) is subject to the context of analysis. For instance, if a company is interested in macro-level analysis, using large cut-off values may create more meaningful results by restricting the number of patterns according to their significance and reliability. In contrast, if a company carries out micro-level analysis, using small cut-off values may give a practical solution by including more diversification cases. Considering these factors, the cut-off values for support and confidence were set to 0.003 and 0.1. Moreover, the maximum length of a pattern was restricted to six since a complex pattern can be divided into smaller patterns, although the proposed approach can allow for more complex analyses.

Another issue is the way of computing the time interval between business segments which is a proxy for the time required for diversification. To calculate the time interval, we focused on the first recorded dates when a firm enters its initial business and diversifies (or exits) its businesses into another business area. For example, the time interval of the pattern 5088 (Transportation equipment and supplies) → 5084 (Industrial machinery and equipment) is computed by subtracting the first reporting date of 5088 (31 May 1978) from the first reporting date of 5084 (31 May 1981). For each pattern, we computed the time intervals of all the significant diversification cases and calculated the descriptive statistics such as the minimum, average, and maximum time intervals.

A total of 301 significant changing patterns were discovered. Table 5 includes different types of changing patterns. For instance, 7373 (Computer integrated systems design) → 7372 (Prepackaged software) corresponds to A → B, 4911 (Electric services), 4924 (Natural gas distribution) → 4924 corresponds to (A, B) → A, and 1311 → 1311, 2911 (Petroleum refining) corresponds to A → (A, B). We divided the changing pattern into sub-patterns comprising of two successive business segments. Contrary to the descriptive

Table 5. Part of the business diversification patterns.

Diversification pattern								
SIC _{source}	Interval _{min, avg, max}				SIC _{target}	support	confidence	Lift
1311, 1381	*[0.6, 1.0, 2.0]				1311	0.00545	0.94483	14.29244
1311, 1381	[0.6, 1.0, 1.4]				1381	0.00521	0.90345	96.5959
1311	[0.75, 4.56, 26.74]				4922	0.00458	0.87121	99.50034
...
7373	[1.0, 5.0, 27.0]				7372	0.00537	0.18698	2.79647
7375	[0.5, 3.0, 21.0]				7372	0.00513	0.12274	1.83570
...
4911	[1.0, 4.8, 33.0]	4911, 4924	[1.0, 1.0, 4.0]	...	4911	0.00334	0.98824	69.94479
4911	[1.0, 4.8, 33.0]	4911, 4924	[1.0, 1.3, 26.0]	...	4924	0.00318	0.9625	125.95716
2911, 1311	[1.0, 1.0, 2.0]	2911	[1.0, 3.5, 22.0]	...	1311	0.00358	0.95745	14.48333
...
7379	[1.0, 3.0, 17.0]				7372	0.00306	0.22449	3.35746
6552	[1.0, 5.1, 33.0]				6512	0.00310	0.16049	9.24901
...
4911, 4924	[1.0, 1.3, 26.0]	4924	[1.0, 3.6, 35.8]	...	4911, 4924	0.00314	0.96341	275.07680
1311	[1.0, 2.2, 25.0]				2911, 1311	0.00354	0.89899	215.124
...

*Time interval of diversification (minimum, average, maximum).

statistics of whole samples, 64% and 36% of the sub-patterns are classified as related and unrelated diversification, respectively. These results confirm that firms attempt to implement related diversification strategies that have been considered relatively low-risk but profitable. Moreover, the extracted time interval can provide information about the feasibility of business diversification. For instance, pattern 1381 (Drilling oil and gas wells) \rightarrow 1311 requires 0.6, 1.6, and 18.0 years as the minimum, average, maximum time interval, respectively. In other words, a firm belonging to 1381 needs to have the preparation period of about one and a half years to diversify into business segment 1311.

The extracted patterns are divided into three types according to the support and confidence values. First, the patterns with high support and confidence values (e.g. 1311, 1381 \rightarrow 1311) are considered as mainstream strategies showing strong correlations between the associated business segments. Second, the patterns with low support and high confidence values (e.g. 7373 \rightarrow 7372) are not frequently occurred but are considered highly reliable since in most cases the firms belonging to the former business segments diversify into the latter business segments. Finally, the patterns with high support and low confidence values (e.g. 4911 \rightarrow 4911, 4924 \rightarrow 4911) represent that weak correlations between the associated business segments exist, although many diversification cases are observed. That is, the firms belonging to the former business segments have diversified into not only the latter segments but also many other segments.

4.3. Assessment of potential areas for diversification via index analysis

To introduce a practical example, we carried out an index analysis on business segment 1311. This segment has been diversified into four areas: 1381, 2911, 4922, and 5172 (Petroleum and petroleum products wholesalers, except bulk stations and terminals). Table 6 (a) provides the summary of the ten industry-level indexes. The analytic hierarchy process (AHP) was conducted to determine the importance weights of the indexes and to calculate the composite index values representing the overall attractiveness (see Table 1 of the Online Appendix A). The indexes pertaining to a firm's profit present higher weights while those related to a firm's activity are relatively lower than others. Table 6(b) presents the composite index values for the four areas over time.

Figure 3 illustrates the results of the index analysis to examine the market and financial characteristics of the potential areas identified. In each graph of Figure 3, the horizontal axis represents the time from 1998 to 2016 and the vertical axis indicates the value of the market and financial indexes in each category. We converted the index values into the normalised moving average within three years to alleviate any effects from fluctuations. Figure 3(a) shows that the profitability remarkably fluctuates in every four or five years for business areas 1381 and 5172 while the rest two business areas present flattening trends except a sudden descent of 4922 in the latest year. Figure 3(b) shows that the financial activities of 2911 and 5172 tend to peak at a certain point while those of 1381 and 4922 are unchanged in most periods. In addition, 5172 shows an increasing trend in recent years while 2911 is declining. Figure 3(c) shows that the investments of all business areas except 2911 present a similar shape of graphs fluctuating in the course of time. Business areas 1381, 4922, and 5172 simultaneously hit the bottom in 2005 and reached a peak at the same time around 2009. Finally, Figure 3(d) shows the overall attractiveness

Table 6. Part of the results of index analysis.

Business segment	Time	Profitability				Activity		Investment			
		ROA	ROA2	PMR	PMR2	ATR	FAT	CEX	RD	INV1	INV2
(a) Market and financial characteristics of the areas											
1381	1976		0.196		0.146	1.344		0.463	0	0.463	0.568
	1977		0.105		0.197	0.537		0.187	0.000	0.187	0.259

2911	2015	0.005	0.083	0.049	0.290	0.316	0.342	0.068	0	0.068	0.123
	2016	−0.055	0.051	−0.177	0.226	0.242	0.249	0.041	0	0.041	0.099
	1976		0.194		0.124	1.555		0.149	0.000	0.149	0.194
	1977		0.186		0.131	1.559		0.097	0.000	0.097	0.141

4922	2015	0.117	0.099	0.077	0.081	1.737	3.305	0.067	0.000	0.067	0.116
	2016	0.083	0.097	0.073	0.114	1.467	2.772	0.052	0.000	0.052	0.104
	1976		0.131		0.191	0.987		0.039	0	0.039	0.084
5172	1977		0.131		0.163	0.937		0.064	0.000	0.064	0.116

	2016	0.009	0.035	0.029	0.209	0.220	0.366	0.060	0	0.060	0.105
	2017	−0.104	−0.059	−1.762	−0.999	0.059	0.076	0.003	0	0.003	0.061
	1976		0.156		0.043	3.883		0.041	0	0.041	0.102
	1977		0.073		0.024	2.460		0.096	0	0.096	0.128

	2017	0.345	0.130	0.065	0.052	2.642		0.036	0	0.036	0.057
	2018		0.121		0.045	2.510		0.051	0	0.051	0.074

Table 6. Continued

Business segment	Time	Profitability	Activity	Investment	Overall
(b) Industry attractiveness of each area					
1381	1998–2000	0.690	0.028	0.350	0.485
	1999–2001	0.701	0.023	0.333	0.490

	2013–2015	0.708	0.019	0.280	0.489
	2014–2016	0.694	0.016	0.199	0.449
2911	1998–2000	0.698	0.106	0.209	0.473
	1999–2001	0.709	0.100	0.184	0.467

	2013–2015	0.650	0.137	0.154	0.433
	2014–2016	0.667	0.113	0.156	0.440
4922	1998–2000	0.709	0.017	0.099	0.424
	1999–2001	0.702	0.019	0.103	0.423

	2013–2015	0.572	0.008	0.087	0.349
	2014–2016	0.503	0.005	0.067	0.309
5172	1998–2000	0.656	0.122	0.091	0.414
	1999–2001	0.653	0.125	0.115	0.419

	2013–2015	0.709	0.156	0.073	0.451
	2014–2016	0.721	0.152	0.071	0.459

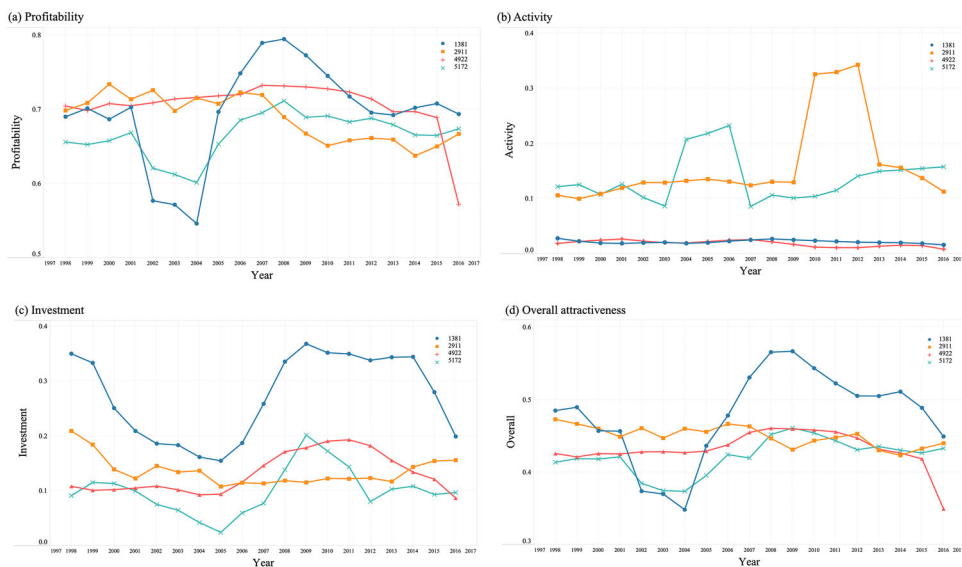


Figure 3. Part of the results of dynamic index analysis.

for each potential area of diversification. Under the influence of the importance weight of profitability, the fluctuating trend of 1381 and the recent sharp fall of 4922 are observed in the graph of the overall attractiveness. There is no perceptible change in the overall index values for 2911 and 5172.

4.4. Development of diversification strategy maps

In Figure 4, we developed the three diversification strategy maps to illustrate a decision making process for diversification. First, the diversification feasibility map was constructed for the four potential business areas identified. In the map, 1381, 2911, and 5172 are classified as *attractive & competitive* while 4922 belongs to *unattractive & monopolistic*.

A firm trying to diversify needs to determine whether to engage in the competition or not under its management strategy. If a firm tries to avoid the competition, it could consider the diversification into 4922. On the other hand, if a firm wants to secure profitability, diversifying its business into one of 1381, 2911, or 5172 could be a better choice. Second, the financial activity map was generated for the four potential business areas identified. In the map, 5172 is classified as *profitable & small investment*,

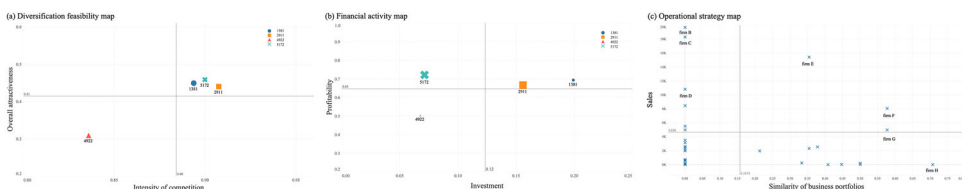


Figure 4. Diversification strategy maps.

4922 belongs to *unprofitable & small investment*, and 1381 and 2911 are placed in the *profitable & large investment* quadrant. Moreover, the size of 5172 represents the financial activity value of the area and is the largest among the four. Thus, the potential areas need to be investigated according to each firm's internal strategy and financial situation. Finally, the operational strategy map was constructed from a position of firm A, which has only been in the business area 1311. To introduce a practical example, firm A is assumed to diversify from 1311 to 5172. We constructed the operational strategy map based on the firms belonging to 5172. Most firms are located in the left-downward quadrant where firm A needs to consider the *indifference* strategy. Firm A needs to consider *collaboration* with firms B, C, and D because of the relatively higher sales and the dissimilarity in business portfolios. Since firms E, F, and G are similar to firm A in business portfolios and have high sales, firm A could benchmark their business activities rather than competing with firms E, F, and G. As for firm H, firm A may compete with firm H since firm A is likely to be in the similar position to firm H after the diversification into 5172. These findings are expected to be useful for strategic decision making regarding the feasibility of business diversification, especially for small and medium-sized companies that consider entering new business areas but have little domain knowledge.

5. Discussion

As Table 3 shows, we identified the diversification pattern of a firm as one of switching, exiting, or entering. Since each pattern can be classified according to the relatedness, the diversification pattern can be further broken down into six cases. In this session, we discuss how the average performance of diversified industries could change before and after the firm diversification based on the above six cases. When a firm enters into or exits from a particular industry, its performance could not be observed directly before or after diversification. Therefore, we use the averages of the median ROA of companies in the entering or exiting industries to assess the success or failure of diversification. When a company switches from one industry to another, the success of the switch is examined through the ROA difference between switching industries.

Figure 5(a) illustrates the diversification into new industries and shows that firms tend to enter a new industry having low-volatile and steady performances after entering rather than expecting a short-term rise in ROA. This effect appears to be more pronounced when entering into a related industry than when diversifying into an unrelated industry. Unlike entering into new industries, Figure 5(b) shows that the industry performance has already declined before the company exits. After the exit, the ROA of the exiting industry

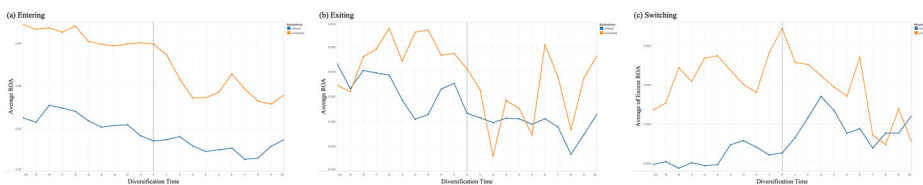


Figure 5. The ROA trends before and after diversification.

continues to decline and seems to be very volatile. Therefore, our methodology appears to provide a clearer guideline for exit strategies than for entry strategies. Figure 5(c) shows the trend of ROA difference between industries A and B before and after the diversification when the business area of a firm switches from industry A to B. In the case of switches, the ROAs of the new industry seem to significantly outperform those of the existing industry. In particular, the related diversification leads to the increase in ROAs which could be considered as a desirable diversification while the unrelated case shows a significant decrease in ROAs after switching. Therefore, the switch to related industries might be considered a successful diversification while unrelated diversification might be not. In summary, our methodology suggests that exiting from poorly performing industries or switching to related industries can be seen as successful diversification. Moreover, entering or switching to unrelated industries might be an undesirable diversification.

6. Conclusion

This study has proposed a sequential pattern mining approach to identifying potential areas for business diversification. The primary contributions of this research are two folds. First, from an academic perspective, this study contributes to business diversification research in the technology and innovation management fields by extending previous technology intelligence approaches to competitor intelligence approaches using the historical business segment data. In contrast to previous technology intelligence approaches, the proposed approach identifies the potential areas for business diversification at the industry level and assesses the market and financial characteristics of the areas identified. The proposed approach enables a wide-ranging search for potential areas for business diversification and the quick assessment of their characteristics. Second, from a practical standpoint, a software system has been developed to automate the proposed approach, allowing those unfamiliar with the data and methods employed in this study to examine the feasibility of business diversification. The proposed approach and software systems could be useful as a complementary tool to assist expert decision making.

However, this study has limitations that should be complemented by future research. First, this study relies solely on the Compustat historical business segment data to identify potential areas for business diversification. The integration with other types of databases such as patent and M&A databases could be helpful for expanding and diversifying the scope of analysis. Second, the proposed approach focuses only on significant changing patterns of firms' business segments. However, different implications can be derived from the diversification cases that are rare but meaningful. Third, with respect to the second issue, the proposed approach cannot provide information about the performance of business diversification. This point should be further investigated by integrating the performance of business diversification into sequential pattern mining. Specifically, such methods as event study analysis (Ding, Lam, Cheng, & Zhou, 2018; Fama, Fisher, Jensen, & Roll, 1969) and difference-in-difference analysis (Abadie, 2005; Card & Krueger, 1994) could be useful for this purpose. Finally, the time taken for diversification (especially for average and maximum time) should be elaborated further since the historical business segment data does not include the exact time when the actual diversification strategy started to be investigated by companies

although the proposed approach computed the time intervals between business segments in each extracted pattern.

Note

1. Since 1976, U.S. firms are required to report segment information about assets, revenues, sales, depreciation, and capital expenditures for each business segment representing over 10% of firm revenues.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIP) (No. 2017R1C1B2011434) and the UNIST2014 research fund (No. 1.140072.01).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [grant number 2017R1C1B2011434] and the UNIST2014 research grant funded by Ulsan National Institute of Science and Technology [grant number 1.140072.01].

Notes on contributors

Gyumin Lee is a PhD student of the School of Management Engineering at Ulsan National Institute of Science and Technology (UNIST). He received a BS in computer science from UNIST. His research interests include data mining and machine learning, recommendation system, and systematic technology intelligence.

Daejin Kim is an assistant professor of finance in the School of Business Administration at UNIST. He received a Bachelor of Business Administration from Korea University, an MS in statistics from Stanford University, and a PhD in Finance from Owen Graduate School of Management at Vanderbilt University. His research lies principally in the area of market microstructure, derivatives, and portfolio management.

Changyong Lee is currently an associate professor of the School of Management Engineering at UNIST. He received a BS in computer science and industrial engineering from Korea Advanced Institute of Science and Technology (KAIST), and a PhD in industrial engineering from Seoul National University. Prior to joining UNIST, he had worked at Korea Institute of Science and Technology Information as a senior researcher and at the Centre for Technology Management (CTM), University of Cambridge, as a visiting scholar. His research interests include data mining and machine learning, technology management, service science, and prognostics and health management of electronics.

It is confirmed that this item has not been published nor is currently being submitted elsewhere.

References

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1), 1–19.
- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207–216.
- Agrawal, R., & Srikant, R. (1995). *Mining sequential patterns*. Proceedings of the Eleventh International Conference on Data Engineering (pp. 3–14).
- Amit, R., & Livnat, J. (1988). Diversification strategies, business cycles and economic performance. *Strategic Management Journal*, 9(2), 99–110.
- Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32(1), 69–87.
- Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4), 772–793.
- Ding, L., Lam, H. K. S., Cheng, T. C. E., & Zhou, H. (2018). A review of short-term event studies in operations and supply chain management. *International Journal of Production Economics*, 200, 329–342.
- Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review*, 10(1), 1–21.
- Garcia-Vega, M. (2006). Does technological diversification promote innovation?: An empirical analysis for European firms. *Research Policy*, 35(2), 230–246.
- Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics*, 36(1), 16–38.
- Jagannathan, M., Stephens, C. P., & Weisbach, M. S. (2000). Financial flexibility and the choice between dividends and stock repurchases. *Journal of Financial Economics*, 57(3), 355–384.
- Kim, H., Hong, S., Kwon, O., & Lee, C. (2017). Concentric diversification based on technological capabilities: Link analysis of products and technologies. *Technological Forecasting and Social Change*, 118, 246–257.
- Larrode, E., Moreno-Jiménez, J. M., & Muerza, M. V. (2012). An AHP-multicriteria suitability evaluation of technological diversification in the automotive industry. *International Journal of Production Research*, 50(17), 4889–4907.
- Lee, M., & Lee, S. (2017). Identifying new business opportunities from competitor intelligence: An integrated use of patent and trademark databases. *Technological Forecasting and Social Change*, 119, 170–183.
- Lee, S., Yoon, B., Lee, C., & Park, J. (2009). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change*, 76(6), 769–786.
- Lemelin, A. (1982). Relatedness in the patterns of interindustry diversification. *The Review of Economics and Statistics*, 64(4), 646–657.
- Leten, B., Belderbos, R., & Van Looy, B. (2007). Technological diversification, coherence, and performance of firms. *Journal of Product Innovation Management*, 24(6), 567–579.
- Luo, Y. (2002). Product diversification in international joint ventures: Performance implications in an emerging market. *Strategic Management Journal*, 23(1), 1–20.
- Mannila, H., Toivonen, H., & Verkamo, A. I. (1995). *Discovering frequent episodes in sequences*. Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD' 95) (pp. 210–215).
- Nath, P., Nachiappan, S., & Ramanathan, R. (2010). The impact of marketing capability, operations capability and diversification strategy on performance: A resource-based view. *Industrial Marketing Management*, 39(2), 317–329.
- Office of Management and Budget. (1987). *Standard industrial classification manual*.
- Seol, H., Lee, S., & Kim, C. (2011). Identifying new business areas using patent information: A DEA and text mining approach. *Expert Systems with Applications*, 38(4), 2933–2941.
- Shin, J., Coh, B. Y., & Lee, C. (2013). Robust future-oriented technology portfolios: Black-Litterman approach. *R&D Management*, 43(5), 409–419.

- Srikant, R., & Agrawal, R. (1996, March). *Mining sequential patterns: Generalizations and performance improvements*. International Conference on Extending Database Technology (pp. 1–17).
- Standard and Poor. (2002). *Compustat data guide*.
- Wernerfelt, B., & Montgomery, C. A. (1998). Tobin's q and the importance of focus in firm performance. *The American Economic Review*, 78(1), 246–250.