Technological Forecasting & Social Change 90 (2015) 355–365

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Novelty-focused patent mapping for technology opportunity analysis

Changyong Lee [a,1], Bokyoung Kang [b,2], Juneseuk Shin [c,*]

[a] School of Business Administration, Ulsan National Institute of Science and Technology, UNIST-gil 50, Ulsan 689-798, Republic of Korea
[b] Department of Industrial Engineering, School of Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea
[c] Department of Systems Management Engineering, School of Engineering, Sungkyunkwan University, 300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do 440-746, Republic of Korea

ARTICLE INFO

Article history:
Received 16 July 2013
Received in revised form 4 May 2014
Accepted 6 May 2014
Available online 12 June 2014

Keywords:
Technology opportunity analysis
Novel patents
Novelty-focused patent identification map
Text mining
Local outlier factor

ABSTRACT

Patent maps are an effective means of discovering potential technology opportunities. However, this method has been of limited use in practice since defining and interpreting patent vacancies, as surrogates for potential technology opportunities, tend to be intuitive and ambiguous. As a remedy, we propose an approach to detecting novel patents based on systematic processes and quantitative outcomes. At the heart of the proposed approach is the text mining to extract the patterns of word usage and the local outlier factor to measure the degree of novelty in a numerical scale. The meanings of potential technology opportunities become more explicit by identifying novel patents rather than patent vacancies that are usually represented as a simple set of keywords. Finally, a novelty-focused patent identification map is developed to explore the implications on novel patents. A case study of the patents about thermal management technology of light emitting diode (LED) is exemplified. We believe the proposed approach could be employed in various research areas, serving as a starting point for developing more general models.

1. Introduction

The strategic importance of technology opportunity analysis (TOA) has become more apparent due to the risks inherent in launching and growing new businesses. Companies are focusing increasing attention on future key technologies by keeping recent developments of technologies under surveillance, often via organising task force teams. However, previous expert-centric approaches have become extremely time-consuming and labour-intensive as markets shift rapidly, technologies proliferate unceasingly, and thus innovation cycles become shorter [1]. Consequently, industrial practitioners call for concrete ways to reduce time, cost, and effort associated with TOA. In academia, recent years have witnessed a significant increase in attempts to devise appropriate models, methods, and tools for systematic TOA.

The ways of systematic TOA strongly depend on data sources. In this respect, the sources of technological information can be divided into four categories: patents, scientific and technical publications, people, and products and processes [2]. Among others, patents, as direct outputs of R&D projects, have been recognised as a valuable source for TOA in that they are collected, screened, and published according to the international standards. It is also noteworthy that almost 80% of all technological information can be found in patent publication [3]. Considering these, patent analysis has long been employed as a useful analytical tool for TOA, and significantly benefited from the use of computerised methods such as text mining and bibliometric analysis [4].

The results of patent analysis can be represented as patent maps in the form of charts, tables, graphs, and networks, which allow the complex information to be understood easily

* Corresponding author. Tel.: +82 31 290 7607; fax: +82 31 290 7610.
E-mail address: jsshin@skku.edu (J. Shin).
[1] Tel.: +82 52 217 3125; fax: +82 52 217 3101.
[2] Tel./fax: +82 2 878 3511.

http://dx.doi.org/10.1016/j.techfore.2014.05.010
0040-1625/

and effectively [5]. Recently, there have been growing interests in integrating data mining techniques (e.g. text mining and dimension reduction methods) into patent analysis for systematic TOA. Several methods and tools have been proposed for this purpose — patent vacuum maps [6], patent vacancy maps [7], GTM-based patent maps [8], and semantic patent maps [9]. At its most basic, patents are mapped on a two-dimensional display according to their similarity in technological contents. On the map, patent vacancies[3], as surrogates for potential technology opportunities, are defined as relatively large areas where the density of patents is extremely low. However, while all of the previous studies have proved quite useful for reducing the burden of manual work that is required to analyse unstructured, lengthy, and rich textual data, the salient problems and deficiencies of previous methods can be summarised along with the development and application process as follows:

- *Patent mapping*: Previous studies have focused only on the ways of developing visual expressions [10]. The major methods for patent mapping have been principal component analysis (PCA) and self-organising feature-map (SOFM). Despite their strengths in reducing the number of dimensions of keywords to acceptable levels, for instance for a two-dimensional map, their utility is limited since multi-dimensional information is decomposed into two "unclear" dimensions [6]. Even though the principal components are generated by a set of observations of possibly correlated variables, these are hard to interpret in practice [7]. This leads to difficulties in defining and assessing patent vacancies, as stated below.
- *Vacancy definition*: The discovery and interpretation of potential technology opportunities tend to be intuitive since this step relies solely on experts' judgements [8]. Note that there is no prior information for the use of identifying patent vacancies at this step. How sparse and how large should the area be on the map to be considered as a patent vacancy? Patent vacancies are detected differently depending on researchers' knowledge and experience, even in a single patent map. Consequently, the meanings of patent vacancies can be interpreted differently since they are defined by investigating the keywords of patents located in the border of patent vacancies. In this regard, some of the previous studies have attempted to define patent vacancies systematically by employing the inverse mapping techniques such as generative topographic mapping (GTM) [8]. Yet, this approach basically identifies the sets of keywords that have not been co-occurred so far. In many cases, the patent vacancies identified by this process are interpreted as infeasible areas, and cannot be easily linked to technology opportunities in practice. Therefore, more specific and complete information, rather than a set of keywords, should be provided to facilitate the process of vacancy definition.
- *Vacancy assessment*: Due to the ambiguity of meanings of patent vacancies, the assessment of potential technology opportunities is likewise vague, and heavily relies on the adjacent patents that are defined subjectively by experts or even that may not be relevant to the potential technology opportunities [10]. Recent literature on TOA has also emphasised the usefulness of patent-level analysis because this way more closely captures actual technological components or elements [11]. Moreover, despite the fact that successful innovation requires sources of novelty [12], how to measure the novelty of patents has hardly been addressed in the literature.

[3] The area is called in many different ways including patent vacancy and patent vacuum.

These shortcomings necessitate the development of a new way to define and assess potential technology opportunities. As a solution, we propose an approach to detecting novel patents based on systematic processes and quantitative outcomes. The meanings of potential technology opportunities become more explicit by identifying novel patents rather than patent vacancies that are usually represented as a simple set of keywords. At the heart of the proposed approach is text mining and local outer factor (LOF). Text mining is employed to extract the patterns of word usage, while the LOF is adopted to measure the degree of novelty in a numerical scale. Unlike other novelty indicators based on knowledge flows and linkages [13,14], the novelty is regarded as the degree of newness of technological information compared to prior art in this study [15,16]. Specifically, the novelty of patents is measured based on the degree to which patents resemble or differ in patterns of keyword usages.

By combining the merits of text mining and the strengths of the LOF and interpreting novel patents instead of patent vacancies as potential technology opportunities, the meanings of potential technology opportunities become more explicit. Furthermore, a software system is developed to implement our method more simply and efficiently, reducing the burden of manual work and therefore allowing even those who are unfamiliar with the complex algorithms to benefit from the research results. It is expected that the systematic processes and quantitative outcomes offered by the proposed approach can facilitate consensus-building on potential technology opportunities and serve as a starting point for developing more general models.

The remainder of this paper is organised as follows. A general background of text mining techniques and the LOF is presented in Section 2. The proposed approach is explained in Section 3, and illustrated with a case study of the thermal management technology of light emitting diode (LED) in Section 4. Finally, Section 5 offers our conclusions.

2. Background

2.1. Text mining

The main objective of text mining is to discover previously unknown knowledge from a large collection of texts [17]. It employs various methods from the research fields of information extraction, information retrieval, and data mining [18]. Specifically, text mining puts a set of labels on each document by attaching them to a keyword list that represents domain knowledge. As a result, documents are distinguished according to the keywords, which allow discovery operations to be performed [19]. This method considerably reduces human efforts needed to analyse unstructured, lengthy and rich textual data [20]. Recent years therefore have seen a significant increase in the use of

text mining techniques in a wide array of research areas such as new product development [21,22], new service development [10], and new technology creation [6–8].

The procedure of text mining is composed of four steps. The first step is data collection and pre-processing. Second, the structural elements are identified by linguistic analysis of domain- and situation-related elements. A variety of structures can be employed for different purposes. Markoff et al. [23], for instance, developed a two-place structure for grievances, which contained one syntactic component for the object of grievance and the other for the action that should be taken toward this grievance. Similarly, Bergmann et al. [9] encoded blocks of text as subject-action-object (SAO) triplets while Lee et al. [24] and Yoon and Park [25] employed a morphological structure in order to represent the patterns of word usage. The structural elements extracted at this step should be rearranged to consider abbreviations, synonyms, and singular and plural forms as several different words may represent the same meaning and some common words are of little value in texts [26]. Third, sampled texts are mapped as syntactic components within this template. Finally, the results are evaluated by experts in the relevant domains.

2.2. Local outlier factor (LOF)

The LOF is a density-based anomaly detection method [27]. The distinct strengths of this method lie in its ability to calculate the degree of novelty (or being an outlier) in a numerical scale, enabling quantitative and objective interpretation to be performed [28]. Many researchers have empirically demonstrated that the LOF outperforms other existing detection algorithms including SOFM [29]. Specifically, the LOF can detect natural clusters with arbitrary shapes as well as filter out local outliers, in contrast with the distance-based clustering algorithms that cannot find incoherent patterns from data due to the limitation of preserving the topology based on fixed shapes [30].

The LOF of an object is measured by the ratio of the average density of its surrounding objects to the local density of itself. The procedure of LOF calculation is composed of four steps, as follows. Firstly, for each object p, the $k\text{-distance}(p)$ is computed as the Euclidean distance between p and its k nearest neighbours, where k is the user-defined parameter for the minimum cluster size. Secondly, for each object q, the reachability distance to p, $reachDist_k(p,q)$, is derived via $\max\{d(p,q), k\text{-distance}(p)\}$, where $d(p,q)$ is the Euclidean distance between p and q. Thirdly, when $N_k(p)$ is defined as the set of p's k nearest neighbours, the local reachability density, $lrd_k(p)$, is calculated as:

$$lrd_k(p) = \frac{k}{\sum_{q \in N_k(p)} reachDist_k(p,q)} \quad (1)$$

Finally, the LOF of p with respect to k surrounding objects is derived as:

$$\mathrm{LOF}(p) = \frac{1}{k} \sum_{q \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)} \quad (2)$$

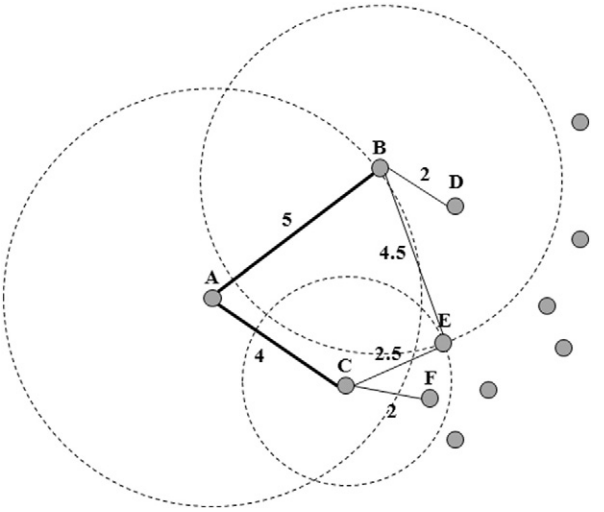

Fig. 1. Example of LOF calculation.

Fig. 1 illustrates a simple example of LOF when k is set to 2. The LOF of A can be derived by comparing the local reachability density of A to the local reachability densities of B and C. The local reachability densities of B and C are also calculated with respect to two surrounding objects $\{D, E\}$ and $\{E, F\}$. In this example, the LOF of A is 1.56, since the local densities of A, B, and C are 0.20, 0.22, and 0.40, respectively. As a consequence, one can identify the regions of similar density, and objects that have a substantially lower density than their neighbours. If an object corresponds to one of the frequent patterns given the k, its density is similar to those of its neighbours, so that the LOF approaches one. Otherwise, the LOF is greater than one and increases as the object is located farther from the normal patterns, since its density is relatively lower than that of normal patterns. Such characteristics based on the relative density offer effectiveness in detecting local outliers as well as global outliers.

3. Research framework

3.1. Concept

Considering that forthcoming technological changes are foreshadowed by current developments [31], the cornerstone for TOA is identification of the current technologies that will drive technological changes [32]. In this respect, patent vacancies have been suggested as surrogates for potential technology opportunities. However, this method has been of limited use in practice since defining and interpreting patent vacancies tend to be intuitive and ambiguous. Moreover, given that there are many technologies to be examined, the amount of time, cost, and effort associated with previous approaches is unrealistic. Such processes need to be supported by good-quality and well-organised information.

This research is initiated in this context, and is based on the premise that analysis using large amounts of objective data and scientific methods enables TOA to be more efficient and successful. Because of these considerations, we propose an instrument for systematic TOA by focusing on novel

Table 1
Comparisons of previous and current research.

Factor	Previous research	Current research
Approach	Mapping patents by reducing the number of dimensions of keywords to acceptable levels	Detecting novel patents based on distribution of patents
Method	Distance-based clustering methods such as PCA and SOFM	LOF which is a density-based anomaly detection method
Output	Patent vacancies that are usually represented as a simple set of keywords	Novel patents with quantitative novelty indicators

patents, instead of patent vacancies. Specifically, given that a set of keywords in a document represents the topics of the document [1,4,6–8,10,20,21], novel patents are identified by analysing the patterns of word usage. The meanings of potential technology opportunities become more explicit by identifying novel patents rather than patent vacancies that are usually represented as a simple set of keywords. By combining the strengths of text mining and merits of LOF, the proposed approach finds novel patents in a systematic and quantitative manner. Table 1 summarises the differences between previous and the current research.

It is important to understand that the objective of the proposed approach is not to produce a definitive set of novel patents, but rather to screen patents having relatively high possibility of being novel. The role of computational methods should be limited to automating experts' routine work and offering information that cannot easily be produced by humans. The communication between experts from different domains and functions still remains critical after this process to discover and crystallise potential technology opportunities.

3.2. Data

The primary source of patents employed in this study is the United States Patent and Trademark Office (USPTO) database. The USPTO database is one of the most representative systems because patents submitted in the US are often simultaneously submitted in other countries, and the US has the largest commercial market in the world [11]. Its database is also well-organised in terms of search conditions and reliability, holding historical data back to 1976 in electronic format, which can be retrieved through screens of 50 titles, each hyperlinked to the full patent text. Of various items included in patents, patent description/specification data is employed to capture the technological information the patents offer while the remaining data is used for interpretation and assessment of novel patents.

3.3. Process

We examine the overall process of the proposed approach, giving a brief explanation of each stage at the same time. As Fig. 2 depicts, the suggested approach employs various methods such as data parsing techniques, text mining, morphological analysis (MA), and the LOF to discover novel patents. As the involvement of many methods and complex algorithms may lead to conceptual misunderstanding and imprecise use in practice, the proposed approach is designed to be executed in four discrete steps: data collection and pre-processing via data parsing techniques; constructing morphological patent contexts via text mining; identifying novel patents via LOF; and finally developing of novelty-focused patent identification map.

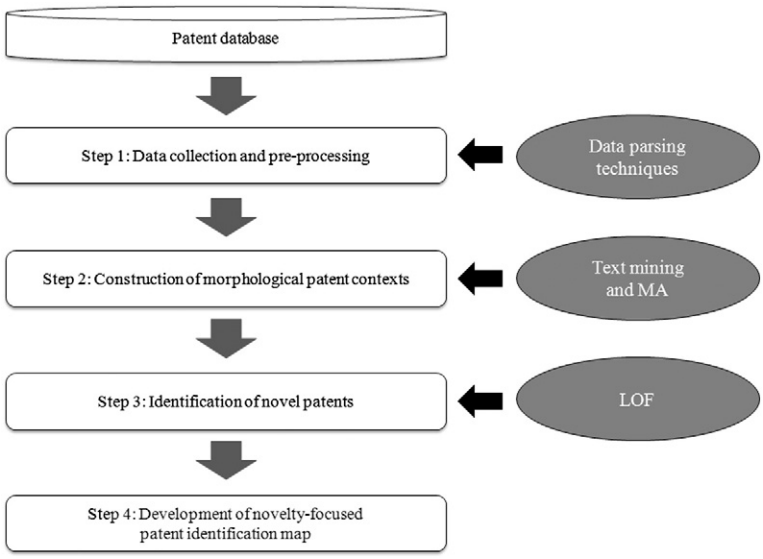

Fig. 2. Overall process of the proposed approach.

3.3.1. Step 1: data collection and pre-processing

Once the focal technology field has been selected, patents of interest are collected from the USPTO database according to relevant search conditions. At this stage, patents are a mixture of structured and unstructured data expressed only as text, and need to be pre-processed. Patents are therefore parsed based on the structure of documents to be transformed into a patent database. The patent database includes structured items (e.g. inventors, assignees, and citations) as well as unstructured ones (e.g. descriptions and claims).

3.3.2. Step 2: construction of morphological patent contexts

The conventional morphological matrix decomposes a system into several dimensions which are mutually exclusive and collectively exhaustive and shapes which each dimension can take. This method allows the alternatives of the system to be derived systematically by combining the shapes of each dimension. In this respect, a morphological patent context is constructed to be utilised as an input of the LOF. The morphological patent context consists of three parts: issued date, patent number, and keyword vector. Among others, taking a morphological perspective, the field of keyword vector is divided into dimensions and shapes. On the contrary to the conventional morphological matrix, multiple shapes are allowed or shapes can be left empty in a dimension to consider the different scope of patents. How to define the morphological structure relies firstly on text mining techniques and secondly on experts' judgements [24]. Specifically, the recent text analysis software yields the importance of keywords and their relationships based on various quantitative indices such as TF-IDF[4] (term frequency-inverse document frequency) and Salton index[5]. However, using only a keyword list extracted by text mining technique is difficult to describe technological characteristics due to a lack of domain-specific technology dictionary. After the keywords of high importance are identified from the dataset, they should be refined based on experts' judgements [4]. Here, repetitive trials between experts and computer-based approach are required to define appropriate technological keyword sets. The dimensions and shapes are then defined based on relationships of keywords and experts' judgements to manifest the properties of different patents. Analytical tools, such as factor analysis and clustering analysis, which can group similar structural elements, could be helpful for this purpose. However, this process should be supported by domain experts, since such automated methods have a limitation in that it may fail to reflect the intrinsic features of patents [25]. A morphological patent context is exemplified in Table 2. In the table, the issued date and patent number are represented in the text format while the keyword vectors are arranged by binary value; '1' means the patent is related to the corresponding shapes, while '0' means the patent does not. For instance, P_5 was issued in Y_2 and related to such shapes as S_{1i} in D_1 and S_{21} in D_2. For more detailed information on the morphological patent context, see Lee et al. [24].

Table 2
Example of morphological patent context.

Patent number	Issued date	D_1			D_2			...	D_n		
		S_{11}	...	S_{1i}	S_{21}	...	S_{2j}		S_{n1}	...	S_{nm}
P_1	YMD_1	1	...	0	0	...	0	...	0	...	0
P_2	YMD_2	1	...	0	1	...	0	...	0	...	0
P_3	YMD_3	1	...	0	0	...	0	...	0	...	0
P_4	YMD_4	0	...	0	1	...	1	...	1	...	0
P_5	YMD_5	0	...	1	1	...	0	...	0	...	0
P_6	YMD_6	0	...	0	0	...	1	...	1	...	1
P_7	YMD_7	0	...	0	0	...	1	...	1	...	1
P_8	YMD_8	0	...	0	1	...	1	...	1	...	1

3.3.3. Step 3: identification of novel patents

The procedure of measuring the novelty of patents is composed of two sub-steps: (1) LOF computation and (2) LOF standardisation.

(1) *LOF computation*: Suppose that PS_j is defined as a set of patents published until year j, in which each patent is represented by the keyword vector $(S_{11}, S_{12}, \ldots S_{\mathrm{nm}})$ in the morphological patent context. For a patent p_i, we can compute $\mathrm{LOF}_j(p_i)$, defined as the LOF of patent p_i, along with the calculation steps explained in Section 2.2. In our approach, the value of k is considered as the amount of patents, which corresponds to the majority with respect to the morphological structures of patents. In this respect, some quantitative methods, such as clustering analysis, could be helpful in determining the k, but qualitative judgements are more flexible in practice. Moreover, this process is of necessity conducted manually in that the criteria may be subjective to the context of TOA and the technology areas. For instance, if a company carries out explorative research to discover novel patents, using a large value of k may create more meaningful results by including more adjacent patents. In contrast, if a company is interested in minor innovation, restricting the scope of analysis to a small number of adjacent patents will give a practical solution. Put together, qualitative judgements by domain experts are employed to determine the value of k, and this process is supported by the results of pilot test using a manageable number of patents to promote consensus-building. In this way, we can compute the LOF values of all the patents for all years.

(2) *LOF standardisation*: By comparing the LOF values, we can analyse the novelty of patents at the year of interest. However, since there exist differences in the range of LOF values across different years, it is difficult to compare one patent's LOF values over time. Even though p_i has the same LOF value at the year t_1 and t_2, its novelties may differ from each other due to the different patent sets. To solve this problem, Kernel Density Estimation (KDE) is adopted to standardise the range of LOF values. KDE is one of the widely used nonparametric estimation methods for determining probabilistic distribution functions from discrete

[4] TF-IDF is calculated as $\mathrm{TF\text{-}IDF}(t,d,D) = \mathrm{TF}(t,d) \cdot \mathrm{IDF}(t,D) = \mathrm{TF}(t,d) \cdot \log \frac{N}{|\{d \in D : t \in d\}|}$, where $\mathrm{TF}(t,d)$, N, and $|\{d \in D : t \in d\}|$ represent the frequency of term t in a document d, the total number of documents in the corpus D, and the number of documents where the term t appears, respectively.

[5] Salton index is defined as $\mathrm{Salton}(x,y) = \frac{Cxy}{\sqrt{CxCy}}$, where C_x and C_y denote the frequency of the keywords x and y, while C_{xy} is the frequency of co-occurrences.

samples [33,34]. Specifically, after constructing a kernel function at the point of each sample, it estimates the probabilistic distribution function of variables by accumulating all kernels. The probabilistic distribution function of LOFs of patents at year j, defined as $f_j(\mathrm{LOF})$, is estimated by KDE, as formulated in Eq. (3).

$$f_j(\mathrm{LOF}) = \frac{1}{n\left(PS_j\right)h}\sum_{i=1}^{n(PSj)} K\left[\frac{\mathrm{LOF}-\mathrm{LOF}_j(p_i)}{h}\right] \tag{3}$$

where $\mathrm{LOF}_j(p_i)$ for $i = 1, \ldots, n(PS_j)$ is the sample point, $n(PS_j)$ is the number of patents in PS_j, K is the Gaussian kernel function, and h is the smoothing factor of the kernel. Finally, we can estimate $R_j(p_i)$ for each $\mathrm{LOF}_j(p_i)$, defined as the rate of relative novelty, as shown in Eq. (4).

$$R_j(p_i) = F(\mathrm{LOF}(p_i)) = \int_{-\infty}^{\mathrm{LOF}(p_i)} f_j(\mathrm{LOF})d\mathrm{LOF} \tag{4}$$

The rate of relative novelty, $R_j(p_i)$, represents the rate of patents having lower LOF values than p_i among all of the patents published until year j. Hence, p_i can be considered to be relatively novel as much as $R_j(p_i)$. It allows the comparison of LOFs of patents at a year of interest as well as the dynamic analysis of one specific patent's novelties to be performed.

3.3.4. Step 4: development of novelty-focused patent identification map

Although novel patents are a good starting point for TOA, they become more explicit when integrated with other information. In this respect, patent citation and patent claim information have long been employed for assessment of significance of technological opportunities. It has been validated by many empirical studies that more frequently cited patents have higher technological and economic impacts [35,36]. The number of claims has also been found to affect the profitability and value of a patent in that the broader the property rights protection, the lower the probability that others may imitate the patent. Because of these considerations, a novelty-focused patent identification map was developed by combining the novelty indicator together with the number of patent citations and the number of patent claims. Specifically, the novelty-focused patent identification map uses the value of novelty, the normalised number of patent citations, and the normalised number of patent claims to determine the size of nodes, the horizontal coordinate, and the vertical coordinate, respectively. The equations for normalisation are shown as follows:

$$\text{Normalised } Cla_i = \frac{Cl_i - \min(Cl)}{\max(Cl) - \min(Cl)} \tag{6}$$

$$\text{Normalised } Cit_i = \frac{Acit_i - \min(Acit)}{\max(Acit) - \min(Acit)} \tag{7}$$

Regarding the number of claims, the value was normalised to make its range from zero to one, as shown in Eq. (6). In the equation, Cl_i represents the number of claims for the ith patent. As for the number of citations, the value was normalised to separate the effects of age and to make its range from zero to one, as shown in Eq. (7). In the equation, $Acit_i$ refers to the number of citations for the ith patent which is divided by the patent's age.

4. Case study

A case study of the patents about thermal management technology of LED is presented to illustrate the proposed approach for the following two reasons. Firstly, LED has received much attention as a substitute for traditional light sources with remarkable advantages in terms of energy efficiency, lifetime, size, and reliability [37,38]. The number of relevant patents has likewise been steeply increased since the year 2000. Secondly, the thermal management technology is a major issue in implementing the advanced LED because the junction temperature of LED is critical to its performance in terms of lifetime, lumen outputs, and stability. As such, identifying novel patents about thermal management technology is essential to research and development.

4.1. Step 1: data collection and pre-processing

Since the number of patents is so huge that we cannot collect all of them in manual. The own-developed Java-based web mining program was used for downloading patents automatically. The search formulas are summarised in Table 3. A total of 649 patents about thermal management technology of LED were collected from the USPTO database after the overlapped patents were removed. Finally, Microsoft Office Access was utilised to construct the patent database based on data parsing techniques. The constructed database included a variety of information such as assignee, citation, and claims. Of these, the data fields of patent number, issue date, and description/specification were employed to develop the morphological patent contexts.

4.2. Step 2: construction of morphological patent contexts

A total of 10 dimensions and 33 shapes that can describe the characteristics of thermal management technology of LED were identified with the aid of domain experts and text mining software (TextAnalysts 2.1) which finds important keywords based on TF-IDF index. We also identified the co-occurrence relationships among keywords based on Salton index via Java-based program, therefore facilitating the keyword selection process that is executed by domain experts.

The thermal management technology of LED was described by nine dimensions: principle of cooling, cooling method, heat management element, type of thermal interface material, substrate material, chip type, encapsulant material, packaging material, and package type. A total of 33 shapes and corresponding keywords were also identified to manifest the characteristics of thermal management technology of LED. Table 4 presents the morphological structure of thermal management technology of LED in terms of dimensions, shapes, and keywords. Finally, the morphological patent context was constructed, which is not reported in its entirety due to a lack of space, as shown in Table 5.

Table 3
Search formulas for patent collection.

No.	Search formula	Number of patents
1	(TTL/(LED OR ((Light AND emitting) AND diode)) AND ABST/(((heat$ OR thermal$) AND resist$) OR ((heat$ OR thermal$) AND conduct$)))	277
2	(TTL/(LED OR ((Light AND emitting) AND diode)) AND ABST/(((heat$ OR thermal$) AND absorb$) OR ((heat$ OR thermal$) AND cool$)))	68
3	(TTL/(LED OR ((Light AND emitting) AND diode)) AND ABST/(((heat$ OR thermal$) AND sink$) OR ((heat$ OR thermal$) AND dissipat$)))	372
4	(TTL/(LED OR ((Light AND emitting) AND diode)) AND ABST/(((heat$ OR thermal$) AND diffus$) OR ((heat$ OR thermal$) AND radiat$)))	84

4.3. Step 3: identification of novel patents

The amount of major patents was investigated by domain experts to determine the value of k using the following two analyses. Firstly, the experts conducted a pilot test by counting the number of patents recognised as majorities. Secondly, the cosine similarities between keyword vectors of the morphological patent context were measured to support the qualitative expert judgements. Cosine similarity is the most frequently adopted indicator in calculating similarities between two unstructured documents [39], and is defined as Eq. (5).

$$\cos\theta = \frac{A \cdot B}{|A||B|} \tag{5}$$

where A and B are keyword vectors of documents. The similarity ranges from 0 to 1, and the greater the similarity, the more similar the documents. Using these results, all the experts reached a consensus that k should be set to 10. The novelty indicator for each patent was derived via the procedure of LOF computation and LOF standardisation. To this end, a MATLAB-based program was developed to automate the calculations, since the number of possible comparisons was very large, and each comparison was so complex that manual comparison was unrealistic. The results are not reported in their entirety due to lack of space. Table 6 depicts the part of rate of relative novelty that is derived from the part of LOF of patents.

4.4. Step 4: development of novelty-focused patent identification map

For the top 10% of novel patents, a novelty-focused patent identification map was developed as shown in Fig. 3. Different colours were assigned to the patents according to

Table 4
Morphological structure of thermal management technology of LED.

Dimension	Shape	Keyword
Principle of cooling (D_1)	Convection (S_{11})	Convection, convector, …
	Conduction (S_{12})	Conduction, conductor, …
	Radiation (S_{13})	Radiation, emissivity, …
Cooling method (D_2)	Natural (S_{21})	Natural air, natural ventilation, …
	Water-cooling (S_{22})	Cooling fluid, water pump, …
	Air cooling (S_{23})	Air cooling, ventilation, …
	Hybrid (S_{24})	Air-water loop, …
Heat management elements (D_3)	Heatsink (S_{31})	Passive cooling, absorption, …
	Thermal via (S_{32})	Thermal via, printed circuit boards, …
	Cooling fins (S_{33})	Cooling fin, fin, …
Type of thermal interface material (D_4)	Epoxy (S_{41})	Thermal interface, epoxy, …
	Thermal grease (S_{42})	Thermal grease, …
	Pressure sensitive adhesive (S_{43})	Pressure sensitive adhesive, PSA, …
	Solder (S_{44})	Solder, solderable, …
Substrate material (D_5)	GaAS (S_{51})	Gallium arsenide substrate, …
	Si (S_{52})	Si substrate, silicon substrate, …
	SiC (S_{53})	Silicon carbon substrate, …
	Al2O3 (S_{54})	Aluminium oxide substrate, …
	ALN (S_{55})	Aluminium nitride substrate, …
	GaN (S_{56})	Gallium nitride substrate, …
	ZnO (S_{57})	Zinc oxide substrate, …
Chip type (D_6)	Non-flip chip (S_{61})	Epi-up, normal posture, …
	Flip chip (S_{62})	Epi-down, inverse posture, …
	Vertical chip (S_{63})	Vertical cylinder, …
Encapsulant material (D_7)	Epoxy (S_{71})	Epoxy encapsulant, …
	Silicone (S_{72})	Silicone encapsulant, …
Packaging material (D_8)	Plastic (S_{81})	Plastic package, PPA, LCP, …
	Ceramic (S_{82})	Ceramic package, glass package, …
	Metal (S_{83})	Metal package, …
Package type (D_9)	Lamp (S_{91})	Liquid resin, transparent mold, …
	SMD (S_{92})	Surface mount device type, …
	COB (S_{93})	Chip on board, COB, …
	BLU (S_{94})	Backlight unit, BLU, …

Table 5
Part of morphological patent context.

Patent number	Issued date	Principal of cooling			...	Package type			
		Convection	Conduction	Radiation	...	Lamp	SMD	COB	BLU
3932761	19760113	0	1	0	...	0	0	0	0
3940846	19760302	0	1	0	...	0	0	0	0
4032945	19770628	0	1	0	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...
7855396	20101221	0	1	0	...	0	0	0	0
7857483	20101228	0	1	0	...	0	0	0	0
7857486	20101228	0	1	0	...	0	0	0	0

Table 6
Part of rate of relative novelty.

Patent number	Issued date	Rate of relative novelty										
		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
3932761	19760113	0.1212	0.2112	0.2909	0.1656	0.2529	0.2411	0.2335	0.2236	0.2475	0.2569	0.2568
3940846	19760302	0.1212	0.1358	0.1468	0.8125	0.8325	0.8531	0.8109	0.7241	0.7983	0.7886	0.7847
4032945	19770628	0.1212	0.2112	0.2909	0.1656	0.2529	0.2411	0.2335	0.2236	0.2475	0.2569	0.2568
4267559	19810512	0.0444	0.1730	0.2909	0.1656	0.2529	0.2411	0.2335	0.2236	0.2475	0.2569	0.2568
4374390	19830215	0.0444	0.1730	0.2909	0.1656	0.2529	0.2411	0.2335	0.2236	0.2475	0.2569	0.2568
...	...	...	...	...	...	...	...	...	...	...	...	...
7854534	20101221	–	–	–	–	–	–	–	–	–	–	0.7359
7855394	20101221	–	–	–	–	–	–	–	–	–	–	0.5063
7855396	20101221	–	–	–	–	–	–	–	–	–	–	0.5458
7857483	20101228	–	–	–	–	–	–	–	–	–	–	0.2568
7857486	20101228	–	–	–	–	–	–	–	–	–	–	0.2568

their morphological structures. The categorisation lines represent the average value of normalised number of forward citations and normalised number of patent claims. On the one hand, the patents having higher citation counts are classified as *influential*. The value and potential of these patents need to be thoroughly investigated. The assignees' products and the

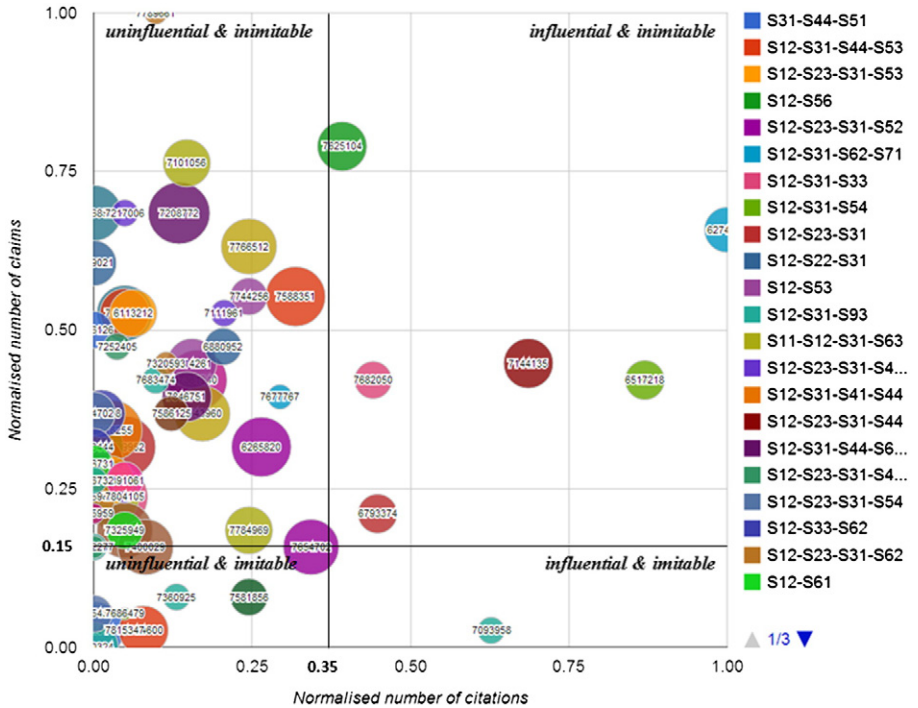

Fig. 3. Novelty-focused patent identification map.

citing patents of these patents could be useful to identify the feasibility and applicability of patents. On the other hand, the patents having higher number of patent claims are categorised *inimitable*. Organisations need to give careful attention to their property rights if they are to explore technology opportunities based on these patents.

The constructed novelty-focused patent identification map is a cross-sectional evaluation at a specific time and needs to be updated continuously. If specific potential opportunities are passed on to the next stage, they need to be removed from the map. If new patents are considered, the LOF procedure needs to be conducted again to position them on the map. Here, it is noteworthy that once the initial morphological patent contexts have been established, they are totally reusable, and new data can be added and analysed through support from a systematic methodology. Moreover, the structure of novelty-focused patent identification maps employed in this study is by no means fixed and exhaustive. Other factors can be employed to diversify the scope of analysis and enhance the richness of potential implications. Structural adjustments are also possible for customised purposes depending on the context of TOA.

5. Discussions

This study proposed an approach to identifying novel patents based on systematic processes and scientific methods. The proposed approach has several advantages over previous methods as follows. Firstly, the meanings of potential technology opportunities become more explicit by identifying novel patents instead of patent vacancies. Note that patent vacancies are represented as a simple set of keywords while a patent contains explicit information. Secondly, the novelty of patents is measured in a numerical scale, and thus allowing its comparison to be performed. Dynamic changes of one specific patent's novelties can also be investigated by standardising the range of LOF values. An average number of 150,000 patents are being issued every year, and therefore the KDE technique is critical to handle different patent sets over time. Finally, regarding to the first issue, the interpretation and assessment of potential technology opportunities become more clear and directed. There is a body of literature on the assessment of value and potential of individual patents and technologies, and these studies could be integrated together with the proposed approach.

Nevertheless, it should be noted that the purpose of the proposed approach is not to produce a definitive set of novel patents, but rather to screen patents having relatively high possibility of being novel. Comparing patents in terms of novelty is a complex task since patents are basically novel according to the international standards. Moreover, the external validity of the proposed approach cannot be easily gained due to the complexity of real world feedback. In this respect, considering that the publication dates of patents generally have positive relationships with their novelty because of the conditions of patentability, we conducted the *t*-test to statistically compare the mean value of publication year in two different patent sets (the top 10% of novel patents vs. the other patents). Specifically, a two-tailed *t*-test for unequal sample size and unequal variance was carried out; the null hypothesis was $Y_1 = Y_2$ while the alternative hypothesis was $Y_1 \neq Y_2$, where Y_1 and Y_2 denote the mean value of publication year for the top 10% of novel patents and the other patents. As summarised in Table 7, the result indicated significance differences between two patent sets, roughly supporting our contention that the proposed approach finds novel patents.

6. Conclusions

The notions of TOA has become strategically more important as a means for generating effective intelligence on emerging technologies [40,41]. This study proposed an instrument for TOA based on quantitative data and systematic processes. The contribution and potential utilities of this study are three-fold. First of all, this study theoretically contributes to the research area by proposing a systematic approach to exploring potential technology opportunities based on existing technologies. An integration of text mining techniques and the LOF makes it possible to identify novel patents, instead of patent vacancies. The strongest features of the proposed approach lie in its systematic processes and quantitative outcomes. Our approach can overcome the limitations of previous methods that heavily relied on experts' judgements to define and assess patent vacancies. Second, from a methodological perspective, the LOF has been used to identify outliers and abnormalities from data sets in the research fields including process control, fault detection, and handwriting recognition. This study extended the application areas of the LOF by integrating it with text mining techniques and interpreting outliers as novel patents. We also emphasised the systematic process of our method in terms of inputs, throughputs, and outputs. Finally, with regard to the practical implementation, the operational efficiency has also been enhanced by software systems, giving specific practical help to staff in charge of speedy and continuous investigation. The analytical results can be updated easily with minimal involvement of experts, since the data are totally reusable and new data can be added and analysed through support from the software system.

Despite contributions, this study is only at the explorative stage and is subject to certain limitations as outlined below. Firstly, regarding types of technology opportunities, this

Table 7
Summary of *t*-test results.

t-statistic	Degree of freedom	Significance (two-tailed)	Mean difference	Std. error difference	95% confidence interval of difference	
					Lower	Upper
3.248	75.111	0.002	2.203	0.678	3.248	75.111

Note: Mean year for the top 10% of novel patents 2007.215; Mean year for the other patents: 2005.012.

study mainly focused on incremental innovation based on existing patents. The breakthrough ideas cannot be easily identified by the proposed approach although existing patents are regarded as a valuable source for TOA. The proposed approach will be more powerful if carefully integrated with other methods such as Delphi analysis. Secondly, in terms of the scope of analysis, the proposed approach needs to be elaborated further by integrating other factors including organisational expertises and resources. Our analysis does not yet encompass such factors explicitly despite their importance in practice. Thirdly, with respect to the performance of our approach, many issues still remain as to how to improve the performance of the proposed approach. Other methods such as semantic text analysis and clustering analysis could be useful for improving its accuracy. Also, generative statistical models such as T distributed stochastic neighbour embedding could be incorporated to improve its accuracy and gain statistical validity. Fourthly, with respect to the automation and evaluation, the proposed approach reduces experts' burden associated with vacancy definition and assessment, but still relies at critical points on experts (e.g. *determination of the value of k*). In particular, the underlying method, LOF, lacks quality metrics for *k*, and thus the quality evaluation of results depends on experts in this research. Integrating it other methods such as factor analysis and clustering analysis as well as developing guidelines for quality evaluation will be helpful for this, although experts' judgements should be incorporated for practicality. Finally, in terms of validity issue, a newly proposed method needs to be carefully deployed in practice. Further testing on a wide range of technologies in different patent databases could help establish the external validity of our approach. For this, experts' judgements on novelty of patents need to be secured and employed in the future research. Nevertheless, we argue that the systematic processes and quantitative outcomes the proposed approach offers make a substantial initial contribution, both to research and practice.

Acknowledgments

This work was supported by the Future Strategic Fund (1.140010.01) of UNIST (Ulsan National Institute of Science and Technology).

References

[1] C. Lee, B. Song, Y. Park, How to assess patent infringement risks: a semantic patent claim analysis approach, Technol. Anal. Strateg. 25 (1) (2013) 23–38.
[2] O. Granstand, The Economics and Management of Intellectual Property: Toward Intellectual Capitalism, Edward Elgar Publishing, Cheltenham, 1999.
[3] M. Blackman, Provision of patent information: a national patent office perspective, World Patent Inf. 17 (2) (1995) 115–123.
[4] C. Lee, J. Jeon, Y. Park, Monitoring trends of technological changes based on the dynamic patent lattice: a modified formal concept analysis approach, Technol. Forecast. Soc. 78 (4) (2011) 690–702.
[5] B. Yoon, Strategic visualisation tools for managing technological information, Technol. Anal. Strateg. 22 (3) (2010) 377–397.
[6] B. Yoon, C. Yoon, Y. Park, On the development and application of a self-organizing feature map-based patent map, R&D Manag. 32 (4) (2002) 291–300.
[7] S. Lee, B. Yoon, Y. Park, An approach to discovering new technology opportunities: keyword-based patent map approach, Technovation 29 (6/7) (2009) 481–497.
[8] C. Son, Y. Suh, J. Jeon, Y. Park, Development of a GTM-based patent map for identifying patent vacuums, Expert Syst. Appl. 39 (3) (2012) 2489–2500.
[9] I. Bergmann, D. Butzke, L. Walrter, J.P. Fuerste, M.G. Moehrle, V.A. Erdmann, Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips, R&D Manag. 38 (5) (2008) 550–562.
[10] C. Lee, C. Son, B. Yoon, Y. Park, An instrument for discovering new mobile service opportunities, Int. J. Mob. Commun. 11 (4) (2014) 374–392.
[11] C. Lee, Y. Cho, H. Seol, Y. Park, A stochastic patent citation analysis approach to assess future technological impacts, Technol. Forecast. Soc. Chang. 79 (1) (2012) 16–29.
[12] S. Cozzens, S. Gatchair, J. Kang, K.S. Kim, H.J. Lee, G. Ordóñez, A. Porter, Emerging technologies: quantitative identification and measurement, Technol. Anal. Strateg. 22 (3) (2010) 361–376.
[13] M. Reitzig, What determines patent value? Insights from the semiconductor industry, Res. Policy 32 (1) (2003) 13–26.
[14] S.W. Cunningham, Analysis for radical design, Technol. Forecast. Soc. Chang. 76 (9) (2009) 1138–1149.
[15] K.B. Dahlin, D. Oard, R. Kostoff, When is an invention really radical?: defining and measuring technological radicalness, Res. Policy 34 (5) (2005) 717–737.
[16] J.M. Gerken, M.G. Moehrle, A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis, Scientometrics 91 (3) (2012) 645–670.
[17] P. Losiewicz, D. Oard, R. Kostoff, Textual data mining to support science and technology management, J. Intellect. Inf. Syst. 15 (2) (2000) 99–119.
[18] W. Fan, L. Wallace, S. Rich, Z. Zhang, Trapping the power of text mining, Commun. ACM 49 (9) (2006) 76–82.
[19] S. Weiss, N. Indurkhya, T. Zhang, F. Damerau, Text Mining Predictive Methods for Analyzing Unstructured Information, Springer, Berlin, 2005.
[20] H. Smith, Automation of patent classification, World Patent Inf. 24 (4) (2002) 269–271.
[21] G. Cascini, P. Rissone, Plastics design: integrating TRIZ creativity and semantic knowledge portals, J. Eng. Des. 15 (4) (2004) 405–424.
[22] C. Lee, B. Song, Y. Park, Design of convergent product concepts based on functionality: an association rule mining and decision tree approach, Expert Syst. Appl. 39 (10) (2012) 9534-9542.
[23] J. Markoff, G. Shapiro, S. Weitman, Toward the integration of content analysis and general methodology, in: D.R. Heise (Ed.), Sociological Methodology, Jossey-Bass, San Francisco, CA, 1974.
[24] C. Lee, H. Park, Y. Park, Keeping abreast of technology-driven business model evolution: a dynamic patent analysis approach, Technol. Anal. Strateg. 25 (5) (2013) 487–505.
[25] B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: keyword-based morphology analysis, Technol. Forecast. Soc. Chang. 72 (2) (2005) 145–160.
[26] C.D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge Univ. Press, Cambridge, 2008.
[27] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: identifying density based local outliers, Proceedings of the ACM SIGMOID Conference, Dallas, Texas, 2000.
[28] D. Pokrajac, A. Lazarevic, L.J. Latecki, Incremental local outlier detection for data streams, Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Honolulu, Hawaii, 2007.
[29] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J. Srivastava, A comparative study of anomaly detection schemes in network intrusion detection, Proceedings of the 3rd SIAM International Conference on Data Mining, San Francisco, CA, 2003.
[30] N.A. Yousri, M.S. Kamel, M.A. Ismail, Pattern cores and connectedness in cancer gene expression, Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE), Boston, MA, 2007.
[31] R.J. Watts, A.L. Porter, Innovation forecasting, Technol. Forecast. Soc. Chang. 56 (1) (1997) 25–47.
[32] A.L. Porter, A.T. Roper, T.W. Mason, F.A. Rossini, J. Banks, Forecasting and Management of Technology, Wiley, NY, 1991.
[33] C.C. Hsu, L.S. Chen, C.H. Liu, A process monitoring scheme based on independent component analysis and adjusted outliers, Int. J. Prod. Res. 48 (6) (2010) 1727–1743.
[34] J.M. Lee, C. Yoo, I.B. Lee, Statistical process monitoring with independent component analysis, J. Process Control 14 (5) (2004) 467–485.
[35] M. Reitzig, Improving patent valuations for management purposes-validating new indicators by analyzing application rationales, Res. Policy 33 (6) (2004) 939–957.
[36] T. Fischer, J. Leidinger, Testing patent value indicators on directly observed patent value-an empirical analysis of Ocean Tomo patent auctions, Res. Policy 43 (3) (2014) 519–529.

[37] T. Nishida, T. Ban, N. Kobayashi, High-color-rendering light sources consisting of a 350-nm ultraviolet light-emitting diode and three-basal-color phosphors, Appl. Phys. Lett. 82 (22) (2003) 3817–3819.
[38] E.F. Schubert, J.K. Kim, Solid-state light sources getting smart, Science 308 (5726) (2005) 1274–1278.
[39] C. Sternitzke, I. Bergmann, Similarity measures for document mapping: a comparative study on the level of an individual scientist, Scientometrics 78 (1) (2009) 113–130.
[40] A.L. Porter, X.-Y. Jin, J.E. Gilmour, S. Cunningham, H. Xu, C. Stanard, L. Wang, Technology opportunity analysis: integrating technology monitoring, forecasting, and assessment with strategic planning, SRA-J. Soc. Res. Admin. 26 (2) (1994) 21–31.
[41] A.L. Porter, M.J. Detampel, Technology opportunity analysis, Technol. Forecast. Soc. Chang. 49 (3) (1995) 237–255.

Changyong Lee is an assistant professor of the School of Business Administration at Ulsan National Institute of Science and Technology (UNIST). He holds 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 (SNU). His research interests lie in the areas of future-oriented technology analysis, systematic technology intelligence, robust technology planning, intellectual property management, and service science.

Bokyoung Kang is a senior engineer at Samsung Electronics Co., Ltd. He holds a BS in industrial engineering from KAIST, and an MS and PhD in industrial engineering, both from SNU. His research interests include the areas of business process management system, real-time business process monitoring, and multivariate statistical process control.

Juneseuk Shin is an assistant professor of Systems Management Engineering in Sungkyunkwan University. He holds a BS, MS and PhD from SNU. His research interests include corporate foresight, technology strategy, and business model. He has published several articles in Technovation, Technology Forecasting & Social Change, Information Economics and Policy, and others.