

Technology opportunity analysis based on recombinant search: patent landscape analysis for idea generation

Changyong Lee & Gyumin Lee

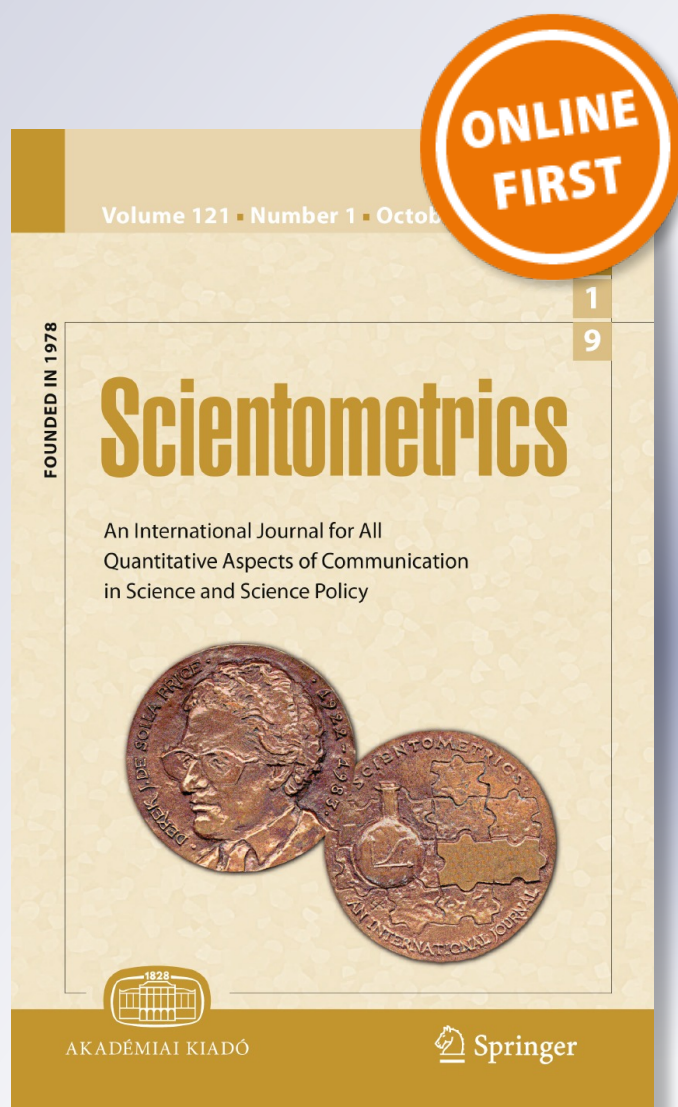
Scientometrics

An International Journal for all
Quantitative Aspects of the Science of
Science, Communication in Science and
Science Policy

ISSN 0138-9130

Scientometrics

DOI 10.1007/s11192-019-03224-7



Your article is protected by copyright and all rights are held exclusively by Akadémiai Kiadó, Budapest, Hungary. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".



Technology opportunity analysis based on recombinant search: patent landscape analysis for idea generation

Changyong Lee¹ · Gyumin Lee¹

Received: 26 December 2017

© Akadémiai Kiadó, Budapest, Hungary 2019

Abstract

This research responds to the need for the use of quantitative data and scientific methods for technology opportunity analysis by focusing on idea generation. Interpreting innovation as a process of recombinant search, we propose a patent landscape analysis to generate ideas which are likely to have more novelty and value than others. For this, first, a patent landscape is constructed from patent classification information as a vector space model, where each position represents a configuration of technological components and corresponds to an idea and, if they exist, relevant patented inventions. Second, the novelty of ideas is assessed via the modified local outlier factor based on the distribution of existing patented inventions on the landscape. Finally, the value of ideas is estimated via naïve Bayes models based on the forward citations of existing patented inventions. In addition, this study also investigates the recombinant synergies between different technological components and the relationships between novelty and value of ideas. A case study of pharmaceutical technology shows that our approach can guide organisations towards setting up effective search strategies for new technology development.

Keywords Technology opportunity analysis · Recombinant search · Patent landscape analysis · Idea generation · Novelty · Value · Synergy

Introduction

The use of quantitative data and scientific methods for technology opportunity analysis (TOA) has become more important as the complexity of technological knowledge and the uncertainty and risk of new businesses increase. Arguably, the most scientific approaches to TOA utilise patent information to discover ideas for new technology development which have not yet been developed or are still emerging (Lee and Lee 2017; Park and Yoon 2017). Existing patent-based approaches to TOA can largely be classified into two categories:

✉ Changyong Lee
changyong@unist.ac.kr

Gyumin Lee
optimist@unist.ac.kr

¹ School of Management Engineering, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulsan 44919, Republic of Korea

patent mapping and morphological approaches. The patent mapping approach maps individual patents that are represented as n -dimensional keyword vectors onto a two dimensional map by reducing the number of dimensions of keywords to acceptable levels; and identifies patent vacancies, as a surrogate of new ideas, which are defined as blank areas that are sparse in patent density but large in size on the map (Lee et al. 2009b; Yoon et al. 2002). Another approach, the morphological approach, breaks down a complex technology system into simpler subsystems, and generates ideas for alternative technologies by modifying and/or combining the shapes of subsystems (Yoon and Park 2005).

However, while these approaches have proved quite useful for reducing the time and effort required to analyse unstructured, lengthy, and rich patent data, concerns have been raised about the subjectivity involved in analytical procedures and the quality and practicality of analysis results. Specifically, although the patent mapping approach offers a strong capability for visualising complex technological relationships between existing patented inventions, patent vacancies can be detected and interpreted differently on a single patent map, depending on researchers' knowledge and experience (Lee et al. 2015). Moreover, the assessment of patent vacancies remains subjective and conceptual due to the ambiguity of the meaning of patent vacancies (Lee et al. 2009b; Yoon et al. 2002). With respect to the morphological approach, although the meaning of ideas derived via this approach is clear and the efficiency of constructing morphological matrices has been significantly improved by text mining techniques, the number of ideas that can be generated from even a simple morphological matrix is too large, leading to difficulties in further investigation to screen and select satisfactory ones (Yoon and Park 2005; Lee et al. 2009a).

These drawbacks necessitate the development of new methods of idea generation for TOA. Three issues are central to this problem, and need to be addressed. First of all, any approach that is proposed should provide a systematic procedure of idea generation and offer a clear meaning and a manageable number of ideas. In this respect, innovation has long been viewed as a process of recombinant search based on a mechanism of coupling prior and/or current technologies (Basalla 1998; Schumpeter 1939; Usher 1954; Rosenberg 1979). Technology landscapes—the realm made up of potential technological innovation that is represented as a configuration of technological components—have proved valuable for conceptualising the notion of recombinant search as a spatial metaphor (Aharonson and Schilling 2016; Fleming 2001; Fleming and Sorenson 2001; Schilling and Green 2011). We suggest that this framework, if integrated with engineering knowledge, could be a good solution for guiding organisations towards generating ideas systematically. Second, past studies using *ex post* approaches are only able to identify valuable ideas after they have succeeded commercially (Kim et al. 2017). It would be far more useful for managers and policy makers to identify ideas with the potential to start a radical change at the early stage of technology development. In this respect, any approach that is proposed should measure the novelty and value of ideas based on *ex ante* approaches (Dahlin and Behrens 2005). Finally, validation of the previous methods has usually been omitted in the literature, and what validation there has been was qualitative and case-specific. Thus any approach that is proposed should establish external validity to give practical assistance.

Considering these issues, we propose a patent landscape analysis to generate ideas which are likely to have more novelty and value than others. For this, first, a patent landscape is constructed from patent classification information as a vector space model, where each position represents a configuration of technological components and corresponds to an idea and, if they exist, relevant patented inventions. Second, the novelty of ideas is assessed by the modified local outlier factor (LOF) based on the distribution of existing patented inventions on the landscape. This method produces arbitrarily shaped decision

boundaries and measures the degree of the novelty of ideas on a numerical scale, allowing comparison to be made (Lee et al. 2015; Kim and Lee 2017). Finally, the value of ideas is estimated by naïve Bayes models based on the forward citations of existing patented inventions. This method provides the probability distribution of the value of ideas, alleviating the complexity and dimensionality problems of the landscape (Bishop 2006). The approach we propose therefore incorporates the issues noted above into TOA, based on which, this study also investigates the recombinant synergies between different technological components and the relationships between novelty and value of ideas.

We applied the proposed approach to pharmaceutical technology to support Korean high-tech companies in discovering ideas for new technology development. We adopted the United States Patent and Trademark Office (USPTO) database for this research, since it contains the most representative data for analysing international technology (Kim and Lee 2015). Our case study showed that the systematic process and quantitative outcomes the proposed approach offers can guide organisations towards setting up effective search strategies for TOA. Our method also enabled a wide-ranging search for ideas and the quick analysis of the novelty and value of ideas, thereby supporting decision making within acceptable limits of time and cost. Moreover, the case study identified ways to improve the proposed approach, which we expect to be a useful complementary tool to facilitate expert-centric idea generation processes.

Background

Recombinant search on technology landscapes

Innovation can be viewed as a process of recombinant search. Schumpeter (1939) stated that “innovation combines components in a new way or ... consists in carrying out new combinations.” Nelson and Winter (1982) also stated that “innovation in the economic system—and indeed the creation of any sort of novelty in art, science, or practical life—consists to a substantial extent of a recombination of conceptual and physical materials that were previously in existence.” In real life, evidence of this phenomenon has been observed in many industries. For example, sodium-ion batteries, which are a type of rechargeable metal-ion battery that uses sodium-ions, instead of lithium-ions as charge carriers, represent a potential power technology for the future because of sodium’s low cost and natural abundance (Wang et al. 2013). Other examples of recombination include: petrol/diesel car–electric car; trackball mouse–optical mouse; film camera–digital camera; inkjet printer–laser printer, and hard disk drive–solid-state drive.

Technology landscapes are regarded as a useful framework for conceptualising recombinant search as a spatial metaphor. This is defined as the realm made up of potential technological innovation, where each position corresponds to an idea and relevant technologies as a particular configuration of technological components (Fleming 2001). The notion of technology landscapes originated from Kauffman’s (1993) *NK* model, where *N* represents the number of components that an inventor recombines; and *K* indicates the interdependence among these components. Figure 1 exemplifies a technology landscape with $N=3$ and $K=0$. In the figure, each component can take 0 or 1, and different configurations of these three components locate on different positions of the technology landscape. Although different topographies of the technology landscapes can be created according to the value of *N* and *K*, at a certain abstraction level,

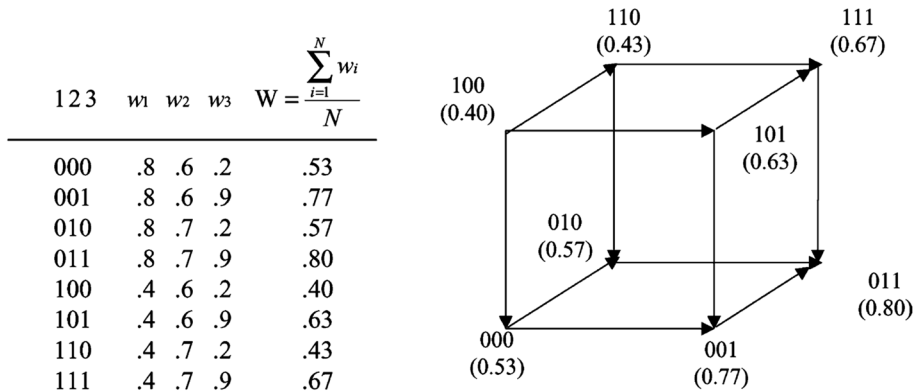


Fig. 1 Example of technology landscapes with $N=3$ and $K=0$ (Fleming and Sorenson 2001)

some positions may be occupied by existing technologies; some positions may not yet be occupied but potentially feasible, and some positions may not be feasible. Moreover, some positions, of course, may have more novelty and value than others.

Recombinant search on technology landscapes have served as an underlying philosophy of many empirical studies in management and economics research. For instance, Fleming (2001) found that experimentation with new components and new combinations leads to less useful inventions on average, but increases the variability that can result in both failure and breakthrough. Arts and Veugelers (2015) found that new combinations of components not only increase average usefulness, but also lead to a higher likelihood of breakthroughs, while reducing the probability of failure. In particular, they noted that the use of familiar components in unprecedented ways can foster the creation of more useful and breakthrough inventions while lowering failure. Similarly, Strumsky and Lobo (2015) identified four distinct sources of technological novelty (i.e., origination, novel combination, combination, and refinement) and assessed the relative importance of refining existing technologies, combining existing and new technologies, and de novo creation of technological capabilities as sources of new inventions. Schilling and Green (2011) found that search scope, search depth, and atypical connections between components influence an invention's impact. Aharonson and Schilling (2016) developed a set of measures that enable a fine-grained characterisation of a firm's technological capabilities on technology landscapes. Arts and Fleming (2018) demonstrated positive correlations between exploring new fields and increased novelty (but decreased value) in an inventor-firm fixed effects panel. They also found that the negative effect of exploring new fields on value is muted via collaboration with experts and/or the use of scientific literature in the new field.

However, while these studies have proved quite useful for providing insight into the mechanism of idea generation, they cannot guide organisations towards setting up effective and practical search strategies for new technology development since the notion of technology landscapes remain conceptual and rely solely on spatial metaphor. Highlighting possible avenues for methodological adaptation, this study develops a patent landscape analysis to generate ideas which are likely to have more novelty and value than others.

Patent analysis for technology opportunity analysis

The methods of TOA strongly depend on data sources. Among various sources for technological information (such as scientific and technical publications, people, and products and processes) (Ernst 2003), patents, as direct outputs of R&D projects, have been recognised as a valuable source for TOA in that they are screened according to international standards (Harhoff and Wagner 2009). It is also noteworthy that patents can empirically explain most aspects of technological innovation activities in developed countries (Trajtenberg 1990). Hence, patent analysis has long been employed as a useful analytical tool for TOA and has significantly benefited from the use of computerised methods such as text mining and bibliometric analysis (Aharonson and Schilling 2016; Daim et al. 2006; Kim et al. 2017; Kim and Lee 2015; Lee et al. 2012; Lee et al. 2009b).

Existing patent-based approaches to TOA can largely be classified into two categories: patent mapping and morphological approaches. First, the patent mapping approach constructs a two dimensional patent map by using dimension reduction techniques after constructing a set of n -dimensional keyword vectors from patents by using text mining techniques. This approach identifies the configurations of technological components that have not yet been occupied by existing patented inventions on the map, i.e., patent vacancies. Previous studies have presented several mapping processes using different dimension reduction techniques such as principal component analysis (Lee et al. 2009b), self-organising feature maps (Yoon et al. 2002), and generative topographic mapping (Son et al. 2012). Specifically, Yoon et al. (2002) developed a self-organising feature map (SOFM)-based patent map to identify potential opportunities of dynamic random access memory (DRMA) technologies using 193 patents. Similarly, Lee et al. (2009a, b) created a principal component analysis (PCA)-based patent map to identify potential opportunities of personal digital assistant (PDA) technologies using 193 patents.

Second, the morphological approach constructs a morphological matrix derived from patents by using text mining techniques, and decomposes a technology system into several dimensions which are mutually exclusive and collectively exhaustive (Wang et al. 2017; Yoon and Park 2005). This approach generates new ideas by modifying and/or combining the shapes of each dimension of the morphological matrix. Yoon and Park (2005) and Yoon and Park (2007) developed a morphological matrix for wide-viewing-angle technology of thin film transistor-liquid crystal display (TFT-LCD) using 137 and 100 patents, respectively. Yoon et al. (2008) developed a morphological matrix for antenna technology of mobile phones using 77 patents. Lee et al. (2013a, b) constructed a modified morphological matrix for electronic shopping technology using 137 patents.

However, while all these previous studies have proved valuable in using quantitative data and scientific methods for TOA, they are subject to certain limitations, as follows. First, although conventional approaches have proved quite useful for idea generation in a single narrow domain at the micro level (e.g., wide-viewing-angle technology of TFT and antenna technology of mobile phones), they are not effective in idea generation for new technology development at the macro level (Aharonson and Schilling 2016), and therefore cannot effectively guide organisations towards reacting to challenges posed by increasingly permeable technology boundaries. Second, previous approaches are not effective in analysing large amounts of data covering multiple domains. Specifically, due to the complexity and information loss problems associated with dimension reduction techniques, large amounts of patents covering multiple domains cannot be effectively

mapped onto a patent map with two “unclear” dimensions, leading to difficulty in defining and interpreting potential technology opportunities (Lee et al. 2015). The number of potential ideas that can be derived from morphological matrices increases exponentially as the scope of analysis is extended and data sample size increases (Lee et al. 2009). Finally, the results of previous approaches strongly depend on experts’ judgments about selection of technological keywords and identification and assessment of potential technological opportunities (i.e., patent vacancies of patent mapping approaches and shapes of technological sub-systems of morphological approaches) (Lee et al. 2009a, 2015). The drawbacks provide our underlying motivation and are fully addressed in this study, as summarised in Table 1.

Methodology

Overall research framework

We propose a patent landscape analysis to generate ideas which are likely to have more novelty and value than others. The premise of this study is four-fold: (1) analysis of large-scale patents can provide objective and reliable information on idea generation (Fleming 2001; Fleming and Sorenson 2001); (2) patent classification systems allow for a precise identification of micro-level knowledge embodied in ideas and existing patented inventions (Aharonson and Schilling 2016); (3) the distribution of existing patented inventions can provide clues to the novelty of ideas (Lee et al. 2015; Yoon and Kim 2012); and finally, (4) the number of forward citations of existing patented inventions can provide clues to the value of ideas (Lerner 1994; Narin et al. 1987).

The overall process of the proposed approach is shown in Fig. 2. Given the complexities involved, the proposed approach is designed to be executed in four discrete steps. First, we collect the relevant patents for a technology field of interest and construct a structured patent database. Second, we create a patent landscape from patent classification information as a vector space model, where each position depicts potential technological innovation with relevant patented inventions. Third, the novelty of ideas is measured by comparing the local densities of the ideas with existing patented inventions on the landscape. For this, the conventional LOF is modified to take into account the characteristics of the patent landscape and TOA. Finally, the value of ideas is estimated via naïve Bayes models based on the forward citation information about existing patented inventions. This method provides class-conditional probability distribution on the value of ideas based on Bayes’ theorem, alleviating the complexity and dimensionality problems of the patent landscape.

Detailed procedures

Data collection and pre-processing

Once a technology field of interest is chosen, the relevant patents (Set 1) are collected based on certain search conditions. The USPTO provides various search commands on the

Table 1 Comparison of previous methods and the proposed approach

Factor	Patent mapping approach	Morphological approach	Proposed approach
Approach	Hybrid (focused more on qualitative approach)	Hybrid (focused more on qualitative approach)	Hybrid (focused more on quantitative approach)
Level of analysis	Micro level (e.g., dynamic random access memory)	Micro level (e.g., antenna technology of mobile phones)	Macro level (e.g., many technologies related to drug, bio-affecting and body treating compositions)
Data sample size and coverage of analysis	Small amounts of data in a single narrow domain	Small amounts of data in a single narrow domain	Large amounts of data covering multiple domains
Focus of analysis	Individual patent	Configuration of technological components	Configuration of technological components with individual patent information
Input	Keyword vector	Morphological matrix	Configuration-value matrix constructed from classification and forward citation information
Methods	Dimension reduction techniques (i.e., principal component analysis and self-organising feature maps)	Concept combination techniques	Modified local outlier factor and naïve Bayes models
Results and implications	Ideas represented as patent vacancies	Ideas generated via combining and/or modifying technological components	Novel and valuable ideas
Characteristics of ideas (number and meaning)	Few unclear ideas	Too many clear ideas	Manageable number of clear ideas

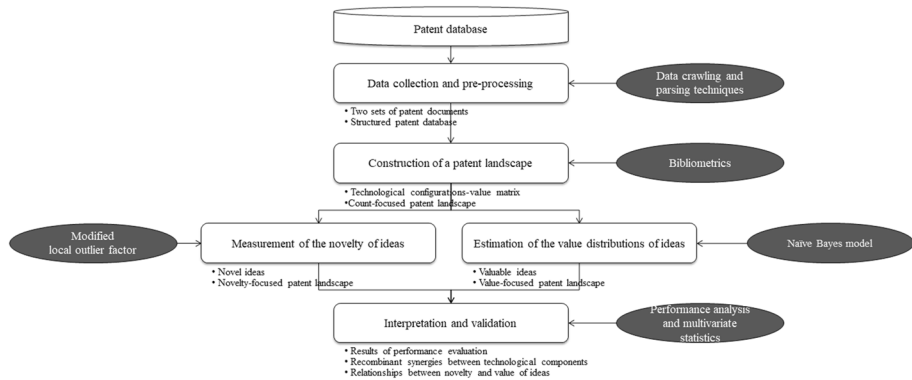


Fig. 2 Overall process of the proposed approach

web (<http://patft.uspto.gov/netahtml/PTO/search-adv.htm>), for instance, CCL for search by current US classification and ACLM for search by claims.¹ The patent documents collected at this stage are a mixture of both structured (e.g., patent numbers and classes) and unstructured (e.g., descriptions and claims) data in either HTML or XML formats. The documents are thus parsed according to the types of information and stored in a structured patent database. In addition to this, the patents that cite Set 1 patents (Set 2) are collected, parsed, and stored in the same manner to analyse the patent forward citations made by later patents. The resulting patent database thereby includes information on both citing and cited patents.

Construction of a patent landscape

This step constructs a patent landscape using the patent database constructed in the preceding step, and consists of two sub-steps: (1) construction of a technological configuration-value matrix and (2) construction of a patent landscape. On the one hand, the technological configuration-value matrix consists of three parts: (1) basic information; (2) technological configuration; and (3) technological value. First, the field of basic information includes patent numbers and publication dates, which define the scope of analysis and time period of interest. Second, the field of technological configuration represents existing patented inventions' technological components and is constructed from patent classification information. Although this field can be constructed in many different ways, i.e., use of patent citations (Alcacer and Gittelman 2006; Benner and Waldfoegel 2008) and word structures (Gerken and Moehrle 2012; Lee et al. 2013a, b, 2014; Wang et al. 2017; Yoon et al. 2013), this study employs patent classification information since it provides rich and reliable picture of a patented invention's technological components (Aharonson and Schilling 2016). Note that the use of patent citations suffers from limitations that stem from the fact that the citation of prior art is discretionary and strategic and that the citation process was never designed to represent a taxonomy (Alcacer and Gittelman 2006; Benner and Waldfoegel 2008). Also, while the use of word structures provides better performance in identifying the contents of a patent (Arts et al. 2018), it requires understanding of the technological

¹ For more detailed information, please see the web-based tutorial provided by USPTO (<https://www.uspto.gov/video/cbt/ptrcsearching/>).

domains and their terms, synonyms, and acronyms (Aharonson and Schilling 2016). Specifically, the existence of different spelling variants and synonyms and the difference in the number of keywords among patents should be considered in the process of selecting keywords and measuring similarity, since these directly influence the accuracy and reliability of analysis (Arts et al. 2018). Moreover, although there are several pre-processing methods (e.g., TF-IDF and entropy-based indicators) for keyword extraction, the repetitive trials between experts and computer-based approach are usually required to define the form and elements of keyword lists from patent documents, which may be time-consuming and labour intensive (Lee et al. 2011a). The domain-specific word (concept) hierarchies are necessary in many cases to measure accurate structural and semantic similarities among patents (Yoon and Kim 2012; Yoon et al. 2013). For this reason, this approach is usually deployed in a single narrow domain rather than across multiple technology domains (Aharonson and Schilling 2016), although advanced pre-processing and text mining techniques such as latent semantic analysis and probabilistic topic modelling can help overcome the above limitations. Here, considering that patents are assigned to one or more classes to delineate the technological fields they cover, and that the primary (i.e., the class in the forefront in bold font) best represents the field where the patented invention can be applied,² we distinguish a primary class from other classes. As a consequence of this, the field of technological configuration is represented as ternary vectors, where 0 represents that the patent does not belong to the corresponding class; 1 represents that the patent belongs to the corresponding class as a non-primary class; and 2 represents that the patent belongs to the corresponding class as a primary class. For instance, with five different classes (i.e., C1, C2, C3, C4, and C5), if a patent belongs to C1 as a primary class, and belongs to C2 and C5 as non-primary classes, the technological configuration of the patent is (2, 1, 0, 0, 1). This information is employed to map existing patented inventions onto specific positions on the landscape and to measure the novelty of ideas. Finally, the field of technological value contains patent forward citation information as a proxy for the technological impact or economic value of patented inventions (Lerner 1994; Narin et al. 1987). Here, we use time-series forward citations so as to aid decision making in short- and mid-term technology planning (Lee et al. 2012). This field is utilised as an input of naïve Bayes models to estimate the probability distribution of the value of ideas.

On the other hand, the patent landscape proposed in this study is a vector space model and depicts all the possible combinations of technological components. Each position on the patent landscape corresponds to an idea and, if they exist, relevant patented inventions, both of which are represented as a specific configuration of technological components.

Measurement of the novelty of ideas using the modified LOF

We modify the LOF to measure the novelty of ideas. The LOF—which is a density-based novelty detection method—measures the degree of novelty on a numerical scale (Breunig et al. 2000). This method is considered more appropriate than other novelty detection approaches for measuring the novelty of ideas for the following reasons: First,

² According to the US Patent Classification System-Class Type (<http://www.acclaimip.com/the-us-patent-classification-system-class-types/>), every US patent has one and only one primary class. It is the class that best describes the invention of a patent. It is double-vetted and reliable since the primary class is used for routing the application through the patent office. If there is a mistake in primary classification, the examiner will reject the patent, and it will be reclassified and routed to a different examiner.

unlike statistical approaches such as Gaussian density estimation and Parzen window density estimation, this method does not require a parametric model describing the distribution of existing patented inventions that are heterogeneous across technology fields and which are difficult to identify due to blurring technology boundaries (Kim and Lee 2017; Lee et al. 2015). Second, this method outperforms distance-based methods (i.e., self-organising feature maps) in that distance-based methods cannot identify outliers from non-uniform and incoherent patterns of data (Lazarevic et al. 2003), thereby having difficulties in finding novel ideas (Lee and Lee 2015).

However, the conventional LOF cannot be directly applied to measure the novelty of ideas and should be modified to take into account the characteristics of the patent landscape and TOA. First, the usual L_2 norm or Euclidean distance measure cannot be employed to calculate the distance between ideas and existing patented inventions; and between existing patented inventions since positions on the patent landscape are represented by ternary vectors (i.e., 0, 1, and 2). Hence, the L_1 norm or Manhattan distance measure is used as shown in Eq. (1) (Lee et al. 2011b).

$$L_1(p_{1:C}, q_{1:C}) = \sum_{c=1}^C |p_c - q_c| \quad (1)$$

where C represents the number of technological components, $p_{1:C}$ and $q_{1:C}$ indicate the positions of patented inventions (or ideas) p and q on the landscape, and p_c and q_c denote the values of c th technological component of patented inventions (or ideas) p and q . Second, the conventional LOF may not work well when there exist many patented inventions on a position of the landscape since the distance between these patented inventions is zero. For this reason, we add a constant to the original distance measured by the L_1 norm or Manhattan indicator.

Using the modified LOF, the novelty of an idea is measured by the ratio of the average density of its surrounding patented inventions to the local density of itself on the patent landscape. The procedure of the modified LOF is composed of four steps, as follows. Firstly, for each idea p , the k -distance(p) is computed as the Manhattan distance between p and its k th nearest patented invention on the patent landscape, where k is the user-defined parameter for the minimum cluster size. Secondly, for each patented invention q , the reachability distance to p , $reachDist_k(p, q)$, is derived via $\max\{d(p, q) + c, k - distance(p) + c\}$, where $d(p, q)$ is the Manhattan distance between p and q and c is the constant to be added to the original distance. Thirdly, when $N_k(p)$ is defined as the set of p 's k -nearest patented inventions, the local reachability density, $lrd_k(p)$, is calculated as:

$$lrd_k(p) = \frac{|N_k(p)|}{\sum_{q \in N_k(p)} reachDist_k(p, q)} \quad (2)$$

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

$$LOF(p) = \frac{1}{|N_k(p)|} \sum_{q \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)} \quad (3)$$

Here, the local reachability density of p 's k -nearest patented inventions, i.e., $lrd_k(q)$, are calculated through the Manhattan distance between q and q 's k -nearest patented inventions.

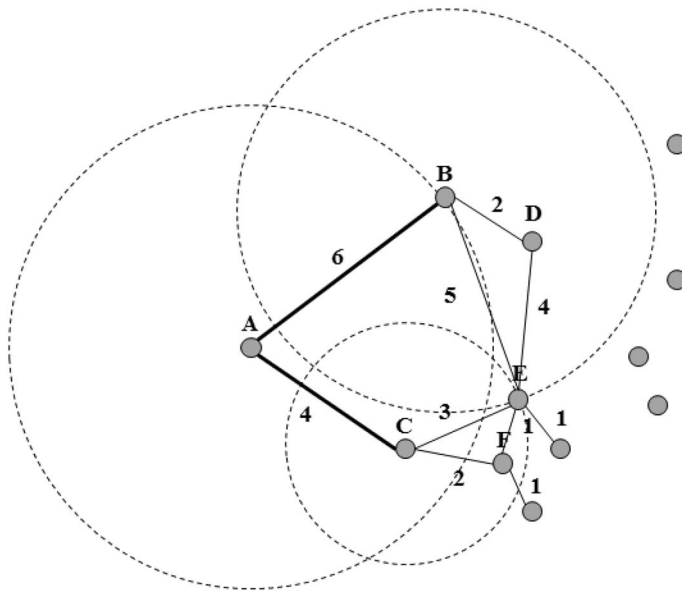


Fig. 3 Example of LOF calculation

Given the k , if the technological configuration of an idea is similar to those of existing patented inventions on the patent landscape, the LOF approaches one. Otherwise, the LOF is greater than one and increases as the technological configuration of an idea is located farther from those of existing patented inventions on the patent landscape, since its density is lower than that of normal patterns. Such characteristics based on the relative density offer effectiveness in detecting the degree of novelty of ideas.

Figure 3 illustrates a simple example of LOF when k is set to 2. The LOF of an idea A can be derived by comparing the local reachability density of A to the local reachability densities of patented inventions B and C . The local reachability densities of B and C are also calculated with respect to two surrounding patented inventions $\{D, E\}$ and $\{E, F\}$. In this example, when c is set to 1, the LOF of A is 1.40, since the local reachability densities of A , B , and C are 0.167, 0.182, and 0.286, respectively.

Estimation of the value of ideas using naïve Bayes models

We employ naïve Bayes models based on the forward citations of existing patented inventions to estimate the probability distribution on the value of ideas. Many empirical studies have found that there is a significant positive relationship between the number of forward citations and the value of the patented inventions (Lerner 1994; Narin et al. 1987). Following this convention, we use the number of forward citations of a patent as a proxy for the value of the patented invention, which is used as a target variable of naïve Bayes models. Specifically, the proposed approach utilises the number of forward citations of patents over the next three and 5 years after the patents are issued so as to aid decision making in short- and mid-term technology planning. As a consequence of this, the value of ideas—that is derived based on the linkages between the technological configurations and future citation counts of existing patented inventions—becomes forward-looking (Lee et al. 2017a, b).

It should also be noted that the value of ideas and patented inventions cannot be ascertained fully from forward citation data. For instance, can we say that a patent that has received 100 forward citations is more valuable than a patent that has received 90 citations? For this reason, recent studies employed the range of the number of forward citations of a patent as a proxy for the value of the patent, instead of the number of forward citations itself (Lee et al. 2017a, b; Woo et al. 2018). In line of these studies, we group the value of ideas and patented inventions into categories according to the (expected) number of forward citations. This transformation is also considered reasonable and practical in that industrial professionals in many cases assess the value of patented inventions using ordinal scales (e.g., A, B, and C).

After the transformation process, measuring the value of ideas corresponds to a classification problem. Naïve Bayes models estimate the probability distribution on the value of ideas by assuming the conditional independence—that is the technological components of ideas are conditionally independent, given the value class v_i . With the conditional independence assumption, instead of computing the class-conditional probability for every combination of technological components of an idea $\mathbf{X} = (X_1, X_2, \dots, X_C)$, we only estimate the conditional probability of each technological component X_j , given the value class v_i . Therefore, this method can measure the value of an idea if the technological components of the idea have been used in classifying existing patented inventions. Specifically, given a technological configuration of an idea $\mathbf{X} = (X_1, X_2, \dots, X_C)$, naïve Bayes models calculate the posterior probability for each value class v_i , as shown in Eq. (4).

$$P(V = v_i | \mathbf{X}) = \frac{P(v_i)P(\mathbf{X}|V = v_i)}{P(\mathbf{X})} = \frac{P(v_i) \prod_{j=1}^C P(X_j|V = v_i)}{P(\mathbf{X})} \quad (4)$$

Naïve Bayes models are considered appropriate for classifying ideas according to their (expected) value on the patent landscape for the following reasons. First, this method performs well even if the conditional independence assumption does not hold, since classification does not need accurate probability estimates so long as the greatest probability is assigned to the correct class (Lewis 1998). Second, this method can handle the patent landscape effectively, since naïve Bayes models are known to be less sensitive to high-dimensionality and sparsity patterns (Bishop 2006). Finally, this method makes classification using Bayes' theorem, which derives the class-conditional probability distribution on the value of ideas from the underlying evidence on the value of existing patented inventions (Tan 2006).

Case study

Overview

A case study of pharmaceutical technology is presented for three reasons. First, a patent normally equals a product in the pharmaceutical industry, so that new ideas for technology development derived from the analysis of large amounts of patents are directly related to new ideas for new product development (Chen and Chang 2010). Second, patent management activities such as valuation and protection are especially important in the pharmaceutical industry, vis-à-vis those of other industries, since the manufacturing process

is relatively easy to replicate and can be implemented with a fraction of the investment (Chaudhuri 2005). Finally, industrial practitioners demand objective information based on scientific methods to identify ideas with the potential to start a radical change, as pharmaceutical technology usually involves substantial investment costs and risks (Chen and Chang 2010). It is therefore worthwhile analysing the extensive information that patents pertaining to pharmaceutical technology provide, so as to assist idea generation for TOA.

Patent landscape analysis for idea generation

Data collection and pre-processing

The USPTO (<http://www.uspto.gov>) serves as our data collection source, since the US is the world's largest patent market—the majority of patents submitted to the USPTO are also submitted in other countries—and so is considered appropriate for analysing international technologies (Kim and Lee 2015). The database is also well-organised and holds historical information back to 1976.

A total of 24,109 patents (Set 1) that belongs to 424 class (entitled “drug, bio-affecting and body treating compositions”) as a primary class were collected over the reference period 2001–2010. Here, a Python-based web mining program was developed to download patents in HTML formats automatically, since the number was sufficiently large that we could not collect them all manually. These documents were then parsed based on their structures, distinguishing each document by its content, and details about patent numbers, publication dates, classes, citations, and other information were stored in a patent database using Microsoft Office Access. In addition to this, the patents that cite the Set 1 patents (Set 2) were collected and transformed in the same manner.

Construction of a patent landscape

The USPTO has categorised patents into the corresponding technology classes defined by the USPC (Unites States Patent Classification). A class has several subclasses with an indent level as a shorthand notation for illustrating dependency (USPTO 2006). A subclass with an indent level of zero is called a mainline subclass, while subclasses with one or more dots are the child of a mainline subclass. In this context, there exist a trade-off between complexities and tractability of analysis. On the one hand, the class-level analysis is simple but has difficulties in the precise identification of technological knowledge embodied in ideas and patented inventions. Note that a class includes countless patents that can be categorised into a number of subgroups. On the other hand, the subclass-level analysis can identify micro-level knowledge embodied in ideas and patented inventions, but the complexity associated with the huge number of subclasses is unrealistic. Following Aharonson and Schilling (2016) and Lee et al. (2012), mainline subclasses served as a unit of analysis in this study to avoid both extremes.

Using a total of 41 mainline subclasses that are related to pharmaceutical technology, the technological configuration-value matrix was constructed for 24,109 patents issued from 2001 to 2010. The matrix is not reported here in its entirety owing to lack of space, but a part of the matrix is shown in Table 2. As stated earlier, our approach distinguishes the primary class from other classes for a more precise identification of micro-level knowledge embodied in ideas and existing patented inventions. For instance, patents 7,655,261

Table 2 Part of the technological configuration-value matrix

Patent number	Issued date	Technological configuration							Value				
		602/41	530/350	424/600	...	424/400	424/184.1	435/4	2001	2002	...	2014	2015
6167888	01/2001	0	0	0	...	0	0	0	0	0	...	0	0
6168028	01/2001	0	0	0	...	1	0	0	0	0	...	0	1
6168776	01/2001	0	0	2	...	0	0	1	0	2	...	1	4
6168777	01/2001	0	0	2	...	0	0	0	0	0	...	1	0
6168778	01/2001	0	0	2	...	0	0	1	0	2	...	8	7
6168779	01/2001	0	0	0	...	0	0	0	1	3	...	1	2
6168780	01/2001	0	0	0	...	0	0	0	0	0	...	0	0
6168781	01/2001	0	0	0	...	0	0	0	0	0	...	0	0
...
7582308	09/2009	0	0	1	...	2	0	0	0	0	...	1	0
...
7655261	02/2010	0	0	2	...	1	0	0	0	0	...	0	0
...
7858679	12/2010	0	0	0	...	0	0	0	0	0	...	3	1
7858753	12/2010	0	2	0	...	0	0	0	0	0	...	0	0
7858756	12/2010	0	2	0	...	0	0	1	0	0	...	1	1
7858762	12/2010	0	0	0	...	0	1	0	0	0	...	0	0
7858765	12/2010	0	0	0	...	0	1	0	0	0	...	0	2
7858837	12/2010	2	0	0	...	1	0	0	0	0	...	0	0

and 7,582,308 belong to the same mainline subclasses, their technological configurations are different.

The count-focused patent landscape was constructed to represent all the possible configurations of 41 mainline subclasses with the patents mapped onto specific positions of the landscape according to their technological configurations. The constructed patent landscape is not reported here in its entirety owing to lack of space, but a part of the count-focused patent landscape ($N=2$) is shown in “Appendix 1(a)”. On the landscape, x and y axis represent the primary and non-primary mainline subclasses, and z axis denotes the number of patents for a specific x – y position of the landscape. For example, the x – y positions such as 424/93.1–424/78.02, 424/184.1–424/63, and 424/130.1–424/78.08 correspond to the ideas that have not yet been occupied on the landscape. The x – y positions such as 424/400–424/70.1, 424/400–424/600, and 424/59–424/400 correspond to the ideas that have been exploited by existing patented inventions, 7,824,705, 7,850,988, and 7,829,068, respectively. It is found that existing patented inventions are concentrated on certain x – y positions such as 424/195.17–424/400, 424/195.18–424.43, and 414/130.1–424/195.16.

Measurement of the novelty of ideas using the modified LOF

Different ideas have different technological complexity and scope in terms of the number of technological components. Moreover, the number of possible configurations of technological components (i.e., positions on the patent landscape) increases exponentially as the number of dimensions of the patent landscape (i.e., the number of subclasses for pharmaceutical technology) increases. In our case study, the number of possible combinations is too large ($\sum_{i=1}^{41} C_{41,i} \times 2$), which makes measuring the novelty of all the possible ideas on the patent landscape unrealistic. Moreover, given the number of technological components for existing patented inventions, the ideas that have more than three mainline subclasses are rare. For this reason, only the ideas with up to three technological components were taken into account. Furthermore, multi-core programming techniques were employed so that many calculations were carried out simultaneously.

Another issue to be considered was the determination of value of k for the modified LOF. In this respect, some quantitative methods, such as cluster analysis, are helpful in determining the value of k , but qualitative judgments are more flexible in practice (Kim and Lee 2017; Lee and Lee 2015; Lee et al. 2015). Moreover, this process is of necessity conducted manually in that the criteria may be subjective to the context of TOA and the technology area of interest. For instance, if a company carries out explorative research to discover novel ideas, using a large value of k may create more meaningful results by including more adjacent patented inventions. In contrast, if a company is interested in minor innovation, restricting the scope of analysis to a small number of adjacent patented inventions will give a practical solution (Lee et al. 2015; Lee and Lee 2015). Experts set the value of k to 30 given the descriptive statistics on the numbers of existing patented inventions for possible combinations.

Considering these issues, a Python-based program was developed to run the modified LOF to measure the novelty of ideas on the patent landscape. A part of the novelty of ideas and the novelty-focused patent landscape are reported in Table 3 and “Appendix 1(b)”. On the landscape, x and y axis represent the primary and non-primary mainline subclasses, and z axis denotes the novelty score for a specific x – y position of the landscape measured via the modified LOF. The x – y positions pertaining to such mainline classes as 424/77 and 424/126 have higher novelty scores, whereas the x – y positions related to such mainline

Table 3 Part of the novelty of ideas

Technological configuration		Novelty score
Primary class	Non-primary classes	
424/77	424/93.1, 424/78.08	5.000
424/77	424/184.1, 424/600	5.000
424/77	424/130.1, 424/600	5.000
424/126	424/93.1, 424/780	5.000
424/126	424/9.1, 424/725	5.000
...
424/77	424/130.1, 424/195.15	4.875
424/126	424/184.1, 424/195.15	4.875
...
424/40	424/93.1, 424/49	4.813
424/40	424/184.1, 424/70.1	4.813
...
424/278.1	424/9.1	1.000
424/278.1	424/78.01	1.000
...
424/62	424/400, 424/175	0.971
424/62	424/400, 424/125	0.971
424/62	424/400, 424/126	0.971
424/59	424/725	0.970
424/278.1	424/93.1	0.968
424/62	424/725, 424/70.1	0.964
424/195.17	424/400	0.958

Table 4 Two categories of value of ideas and existing patented inventions

Category	(Expected) number of forward citations	Number of existing patented inventions
(a) 3-year forward citations (TC3)		
L1	Above 1	1998
L2	0–1	22,111
(b) 5-year forward citations (TC5)		
L1	Above 5	1318
L2	0–5	22,791

classes as 424/62 and 424/195.17 have lower novelty scores than others. The areas having higher novelty scores are found to have little patented inventions (unexplored), while the areas having lower novelty scores are found to have many patented inventions (exploited) on the count-focused patent landscape, which roughly support our contention that the proposed approach finds novel ideas.

Table 5 Part of the value distributions of ideas

Technological configuration		Probability distribution on value of ideas			
Primary class	Non-primary classes	TC3		TC5	
		L1	L2	L1	L2
424/93.1		0.0384	0.9616	0.1727	0.8273
424/400		0.0912	0.9088	0.2768	0.7232
424/184.1		0.0655	0.9345	0.1951	0.8049
...
424/93.1	424/400	0.0404	0.9596	0.1860	0.8140
424/93.1	424/184.1	0.0340	0.9660	0.1539	0.8461
...
424/400	424/93.1	0.0625	0.9375	0.2363	0.7638
424/400	424/184.1	0.0812	0.9188	0.2500	0.7500
...
424/43	424/49, 424/48	0.7776	0.2225	0.6970	0.3030
424/43	424/48, 424/125	0.7667	0.2333	0.6837	0.3163
424/43	424/48, 424/114	0.7667	0.2333	0.6162	0.3838
...
424/43	424/78.08, 424/48	0.6534	0.3466	0.7329	0.2671
424/43	424/130.1, 424/48	0.6831	0.3169	0.7135	0.2865
424/43	424/130.1, 424/78.08	0.5344	0.4656	0.7020	0.2980
...
424/62	424/725, 424/195.16	0.0012	0.9988	0.0449	0.9550
424/195.16	424/725, 424/62	0.0012	0.9988	0.0319	0.9680
...
424/195.17	424/195.16, 424/84	0.0066	0.9934	0.0141	0.9859
424/195.17	424/195.16, 424/61	0.0038	0.9962	0.0124	0.9876

Estimation of the value distributions of ideas using naïve Bayes models

We employed the number of forward citations of a patent over the three (TC3) and 5 years (TC5) after the patent was issued so as to aid decision making in short- and mid-term technology planning. As stated earlier, patents were grouped into two categories according to the (expected) number of forward citations of patents (ideas), as summarised in Table 4. For this, we tried to find comparable breakdowns in TC3 and TC5. The resulting thresholds are similar to those in previous studies. For instance, Hall et al. (2005) reported that patents with less than seven citations during its lifetime do not have nearly any impacts on market values. In our grouping thresholds, L1 patents in TC5 received six or more citations, which is not far from the findings from prior literature. Moreover, given the fact that a small number of patents receive multiple citations while the vast majority are not cited in their lifetime, it is reasonable that most patents are classified as L₂ patents.

The value distributions of the ideas were estimated via naïve Bayes models, as shown in Table 5. In the table, an idea, which is represented by a combination of primary and non-primary classes, can be classified into two categories, L1 and L2, according to the prescribed cut-off value, 0.5. For instance, 424/93.1–424/400 is found less likely to be

Table 6 Recombinant synergies between different technological components

Technological configuration		Synergy	
Primary class	Non-primary classes	TC3	TC5
424/61	424/59, 424/69	0.9540	0.9885
424/93.1	424/400, 424/195.16	0.9476	−0.0385
424/93.1	424/195.15, 424/195.16	0.9476	−0.0385
424/520	424/1.11	0.9467	−0.0376
424/600	424/1.11, 424/9.1	0.9414	0.9676
424/195.15	424/70.1	0.9359	−0.0128
424/195.15	424/184.1, 424/780	0.9359	0.9872
...
424/195.15	424/184.1, 424/195.17	0.9359	0.9872
424/780	424/195.16	0.0000	0.9677
424/780	424/195.15, 424/195.16	0.0000	0.9677
424/520	424/130.1, 424/725	−0.0533	0.9624
424/115	424/78.02, 424/780	0.8800	0.9600
424/69	424/78.08	0.0000	0.9565
424/184.1	424/178.1, 424/94.1	0.4259	0.9533
...
424/43	424/780, 424/195.18	−0.2181	−0.1988
424/43	424/780, 424/75	−0.2181	−0.1988
424/43	424/780, 424/125	−0.2181	−0.1988
424/114	424/61, 424/195.18	−0.3333	0.0000
424/114	424/780, 424/75	−0.3333	0.0000
424/114	424/195.18, 424/125	−0.3333	0.0000
424/114	424/75, 424/126	−0.3333	0.0000
...
424/43	424/400, 424/78.02	−0.1831	−0.1098
424/43	424/78.02	−0.0877	−0.1119
424/43	424/93.1	−0.1147	−0.1299
424/43	424/195.18, 424/126	−0.2181	−0.1988

valuable (classified as L2), while 424/43–424/49, 424/48 is found likely to be valuable (classified as L1). Moreover, the value-focused patent landscapes for TC3 and TC5 are shown in “Appendix 1(c) and 1(d)”. On the landscape, x and y axis represent the primary and non-primary mainline subclasses, and z axis denotes the probability of a x – y position being valuable measured via naïve Bayes models. The value-focused patent landscapes show similar tomographic maps, although there are slight differences in the value distribution for different time periods. These differences may be caused by such factors as different patterns of technology life cycles and technological uncertainties.

One of the most integral questions in setting up effective search strategies for TOA is *what combinations of technological components create more value?* In this context, we measured the synergies between primary and non-primary mainline subclasses by comparing the probability distributions of the value of ideas, as shown in Eq. (5) where C_p and C_{np} represent the primary and non-primary mainline subclasses.

$$Syn(C_p, C_{np}) = P(V = v_1 | C_p, C_{np}) - P(V = v_1 | C_p) \quad (5)$$

Table 7 Performance of the proposed approach

(a) Summary of *t*-test results (number of patents on novel and other positions)

	Top 10% of novel positions	The other inventions
Mean	0.005	1.364
Variance	0.005	2942.600
Observations	3367	30,294
<i>t</i> stat	4.363	
<i>p</i> value	0.000	
<i>t</i> critical two-tail	1.96	

(b) Summary of performance metrics

Time period of interest	Class	Accuracy	Precision	Recall	F_1 score
TC3	L1	0.781	0.125	0.272	0.171
	L2		0.926	0.828	0.874
TC5	L1	0.536	0.282	0.590	0.381
	L2		0.799	0.519	0.629

As Table 6 reports, the following pairs of technologies create more synergies than others: for TC3, 424/61–(424/59, 424/69), 424/93.1–(424/400, 424/195.16), 424/93.1–(424/195.15, 424/195.6), 424/520–424/1.11, 424/600–(424/1.11, 424/9.1), 424/195.15–424/70.1, and 424/195.15–(424/184.1, 424/780); and for TC5, 424/195.15–(424/184.1, 424/195.17), 424/780–424/195.16, 424/520–(424/130.1, 424/725), 424/115–(424/78.02, 424/780), 424/69–424/78.08, and 424/184.1–(424/178.1, 424/94.1).

We also conducted correlation analysis to investigate the relationships between novelty and value of ideas. Although there are differences in the Pearson's correlation coefficients across the time periods of forecasts, the novelty score of an idea is positively correlated with the probability of the idea being valuable (classified as L1) (0.231 and 0.130 for TC3 and TC5, statistically significant at the level of 0.01), supporting the results of previous studies (Arts and Veugelers 2015; Schilling and Green 2011).

Validation

The performance and utility of the proposed approach is directly related to its ability to find novel and valuable positions on the patent landscape. As such, we scrutinised the accuracy and reliability of the underlying methods (i.e., modified LOF and naïve Bayes models) by using the *t* test and several performance metrics. First, in terms of the novelty of ideas, we conducted the *t*-test to statistically compare the mean value of the number of patents on two different sets of positions on the patent landscape (the top 10% of novel positions vs. the other positions). Specifically, a two-tailed *t*-test for unequal sample size and unequal variance was carried out; the null hypothesis was $Y1 = Y2$ while the alternative hypothesis was $Y1 \neq Y2$, where $Y1$ and $Y2$ denote the mean value of the number of patents on the top 10% novel positions and the other positions on the patent landscape. As summarised in Table 7(a), the results indicate the significant difference between two sets, $Y1$ and $Y2$, supporting our contention that the proposed approach finds novel ideas.

Second, with respect to the value of ideas, we examined the accuracy and reliability of the value of ideas derived by naïve Bayes models. However, the value of ideas cannot be verified since they have not yet been exploited. Instead, the value of existing patented inventions was employed to verify if our approach finds valuable ideas. Given that measuring the value of existing patented inventions corresponds to a classification problem, several performance metrics using fivefold cross validation techniques³ were examined to assess the performance of our approach after a confusion matrix was constructed. Here, we used a stratified sampling technique since the number of patents for each category is imbalanced. In other words, after patents are divided into homogeneous strata according to their citation counts (i.e., L1, L2, and L3), the random sampling was applied within each stratum with the sampling fraction that is proportional to the population of a stratum. We measured the accuracy, as shown in Eq. (6).

$$\text{Accuracy}_i = \frac{tp_i + tn_i}{fp_i + fn_i + tp_i + tn_i} \quad (6)$$

Here, true positive (tp_i), true negative (tn_i), false positive (fp_i), and false negative (fn_i) for class i represent the number of positive examples correctly classified, the number of negative examples correctly classified, the number of negative examples wrongly classified as positive, and the number of positive examples wrongly classified as negative; while l is the number of classes. As Table 7(b) reports, although there are differences in the degree of accuracy across different classes, the proposed approach was found to be effective in assessing the value of ideas. We added the precision (positive predictive value), recall (true positive rate or sensitivity), and F_1 score to compensate for the classes having different weights in our case study and because of possible imbalances in the data set (Kim et al. 2019). Precision is the number of true positive results divided by the number of all positive results, whereas recall is the number of true positive results divided by the number of positive results that should have been returned (Bishop 2006). The F_1 score is a measure of the overall effectiveness of a classifier, and is defined as a harmonic average of the precision and recall where an F_1 score reaches its best value at 1 and worst at 0, as shown in Eq. (7) (Bishop 2006).

$$F_1 \text{ score} = 2 \times \frac{1}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

Precision and recall show that the proposed approach is effective in screening existing patents according to value of ideas implied in patented inventions. In particular, the proposed approach shows the highest performance for the least valuable ideas. Moreover, given that recall values are greater than precision values, our method presents liberal behaviours in finding valuable ideas, which is expected to be useful in reducing the size of idea pools and enabling experts to focus more on the small set of ideas. The F_1 score also shows that the proposed approach is effective in screening the least valuable existing patented inventions, although there are differences in the degree of effectiveness of

³ k -fold cross-validation is a statistical technique for assessing how the results of analysis will generalise to an independent data set and how accurately a predictive model will perform in practice. This technique partitions data into k nearly equally sized folds. Subsequently k iterations of training and validation are performed such that, in each iteration, a different fold of the data is held-out for validation while the remaining $k-1$ folds are used for learning a model. Upon completion, k samples of the performance metric are available and they are combined to derive a more accurate estimate of model performance.

classification across different classes. Based on the results of the t -test and performance metrics, we are reasonably confident that the proposed approach will be a useful complementary tool to facilitate expert-centric idea generation processes.

Discussion

Characteristics of the proposed approach

A number of considerations should be made before applying and deploying a novel method in practice. For this, the characteristics of the proposed approach are summarised, as follows. First, the proposed approach measures the novelty and value of ideas and, if they exist, existing patented inventions on the patent landscape, while prior studies have focused only on existing patented inventions. Second, the primary class is distinguished from other classes in constructing a technological configuration-value matrix. Although such treatment increases the time- and space-complexity of analysis, the proposed patent landscape enables a more precise identification of micro-level knowledge embodied in ideas and existing patented inventions and improves the reliability of the proposed approach. Third, the proposed approach examines the distribution of existing patented inventions on the patent landscape to measure the novelty of ideas. Specifically, the novelty score of an idea varies according to (1) the degree of similarity between the technological configuration of the idea and those of existing patented inventions and (2) the number of patents having the same and/or similar combinations of technological components as the idea. Fourth, with regard to value indicators, the value of ideas is not described as a single deterministic value, but as a range of values; and the probability distribution across that range is provided. Specifically, the proposed approach estimates the probability distribution on the value of ideas (and existing patented inventions) after grouping the (expected) number of forward citations into categories, whereas prior studies on recombinant search employed the number of forward citations as a proxy for value of patented inventions. Fifth, the proposed approach is more appropriate for technology opportunity analysis in short-term contexts rather than that of long-term contexts. Many other factors such as discount rates and market information should be incorporated to improve the performance and reliability of the proposed approach. Moreover, the novelty and value of ideas are the results of cross-sectional evaluation at a specific point in time and thus need to be updated continuously. Finally, although a large number of calculations are needed to conduct the proposed approach, the proposed approach can be easily implemented by modifying existing open source packages. We relied on the scikit-learn package for Python to implement and validate the proposed approach. The modified LOF was implemented based on two modules (i.e., `neighbors.LocalOutlierFactor` and `metrics.pairwise`); and naïve Bayes models were implemented based on another two modules (i.e., `naive_bayes.MultinomialNB` and `model_selection.StratifiedKFold`). The pseudo code of the proposed approach is reported in “Appendix 2”.

Comparison of previous methods and the proposed approach

Previous studies have presented different approaches to identifying novel and valuable patents on the landscape. For instance, Strumsky and Lobo (2015) identified novel patents according to the mechanism of recombination, such as origination, novel

Table 8 Summary of *t* test results (modified LOF scores of the novel and the other patents)

	Novel patents	Other patents
Mean	2.392	1.426
Variance	0.848	0.456
Observations	828	23,281
<i>t</i> stat	29.906	
<i>p</i> value	0.000	
<i>t</i> critical two-tail	1.96	

combination, combination, and refinement. Arts and Fleming (2018) identified novel patents via counting the number of pairwise class combinations that appear for the first time in the US patent database. Aharonson and Schilling (2016) measured the technology distance between patents after transforming the patents into binary string vectors based on their mainline subclasses. With respect to valuable patents, most studies have employed the number of forward citations of a patent as the proxy for the value of the patent (Arts and Fleming 2018; Fleming and Sorenson 2001; Fleming 2001).

In this context, we compare the results of previous methods and the proposed approach. First, in terms of novel patents, following Strumsky and Lobo (2015) and Arts and Fleming (2018), a total of 828 novel patents were first identified in our data set by examining if the technological configurations of patents include new pairwise combinations of technological components. We then conducted the *t*-test to statistically compare the mean values of the modified LOF scores of two different sets (i.e., the 828 novel and the remaining 23,181 patents). Specifically, a two-tailed *t*-test for unequal sample size and unequal variance was carried out; the null hypothesis was $Y1 = Y2$ while the alternative hypothesis was $Y1 \neq Y2$, where $Y1$ and $Y2$ denote the mean values of the modified LOF scores of the novel and the other patents. As summarised in Table 8, the results indicate the significant difference between $Y1$ and $Y2$, supporting our contention that the proposed approach discerns novel patents. However, previous methods and the proposed approach differ in that previous methods focus on breakthrough inventions and thus are based on the number of new pairwise subclass combinations that appear for the first time on the landscape, whereas the proposed approach is based on the number of existing patented inventions with technological configurations similar to that of a patent of interest. That is, although a patent does not have a new combination of technological components, the patent can have a relatively high LOF score, if the technological configuration of the patent has rarely been exploited on the landscape (e.g., patent 620713).

Second, with respect to valuable patents, Fleming (2001) and Fleming and Sorenson (2001) measured the value of a patent as the number of forward citations of the patent. Arts and Fleming (2018) measured the value of a patent as the logarithmic transformation of one plus the number of forward citations of the patent. Building upon the work of Fleming (2001) and Fleming and Sorenson (2001), we conducted correlation analysis to examine the relationship between the value of patents measured by previous methods and the probability of the patents being valuable (classified as L1) derived by the proposed approach (using fivefold cross validation techniques). Specifically, we conducted Spearman's rank correlation analysis, instead of Pearson's correlation analysis, since the numbers of forward citations in our data set are highly skewed to the right and do not present linearity and homoscedasticity. The results confirm that the probability of patents being valuable derived by the proposed ex ante approach is positively correlated with the value of patents measured by the actual numbers of forward citations of the patents (ex post measure) (0.771 for TC3 and 0.345 for TC5).

Table 9 Summary of *t* test results (modified LOF scores using different data sets)

	Primary	Primary or non-primary
Mean	2.442	2.436
Variance	0.899	0.823
Observations	33,661	33,661
<i>t</i> stat	0.836	
p-value	0.403	
<i>t</i> critical two-tail	1.960	

Scope of analysis and data collection strategy

Focusing on ‘drug, bio-affecting and body treating compositions’ technology, this study employed the patents that belong to 424 class as a primary class. Of course, the proposed approach can employ all the patents that belong to 424 class regardless of whether it is the primary or non-primary class. The results of such analysis may differ from our findings owing to the difference in distributions of existing patented inventions on the landscape.

We performed additional analyses to compare the novelty scores measured from the patents belonging to 424 class as a primary class and the scores measured from all the patents belonging to 424 class regardless of whether it is the primary or non-primary class. First, we conducted correlation analysis to examine the relationship between the novelty scores derived from the two different data sets. The Pearson’s correlation coefficient (0.98) revealed that the novelty scores are highly correlated with each other, statistically significant at the level of 0.01. In addition, we conducted a dependent *t*-test for paired samples; the null hypothesis was $Y1 = Y2$ while the alternative hypothesis was $Y1 \neq Y2$, where $Y1$ and $Y2$ denote the novelty scores derived from the patents belonging to 424 class a primary class and the scores derived from the patents belonging to 424 class as a primary or non-primary class. We do not find statistically significant difference in the mean values, as summarised in Table 9. This is because that the modified LOF considers the *k* most similar patented inventions to measure the novelty of ideas. From the results of correlation and *t*-test analysis, the proposed approach is considered robust across different data sets.

However, the scope of analysis and data collection strategy should be carefully determined according to the context of analysis. For instance, if a company is interested in identifying potential areas that may be diversifiable based on the company’s core technological capability, restricting the scope of analysis via the primary class will give a practical solution. Note that even though a company owing technology A has the possibility to diversify to technology B, it does not guarantee the company which owns technology B will be able to diversify to technology A (Kim et al. 2017). In contrast, if a company carries out explorative research to identify technology trends, using all the patents that belong to a certain class regardless of whether it is the primary or non-primary class may create more meaningful results.

Conclusions

Innovation is beyond any other concern for organisations to survive in an uncertain and volatile environment. This study has proposed a patent landscape analysis to generate ideas for TOA which are likely to have more novelty and value than others. The

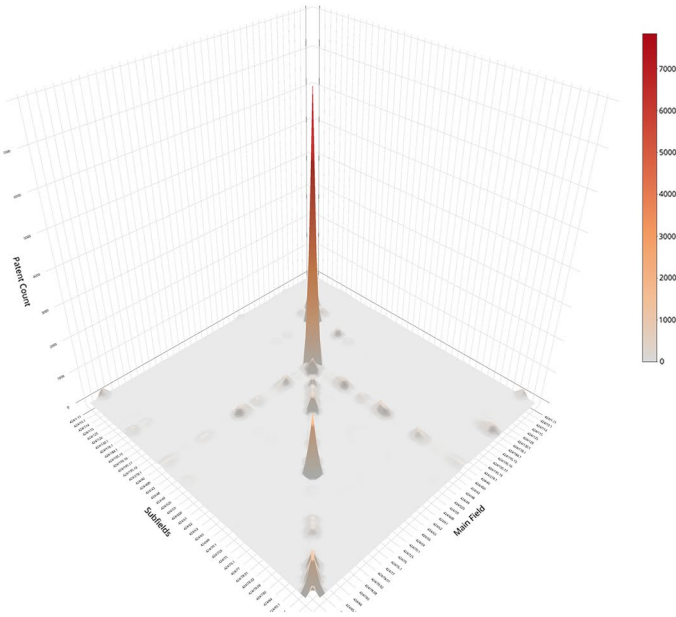
contribution and potential utilities of this study are three-fold. First of all, from a theoretical perspective, this study contributes to the research area by developing an engineering-centric approach to idea generation for TOA. An integration of technology landscapes constructed from patents and machine learning techniques make this possible to identify ideas for TOA and measure their novelty and value. It is expected that the proposed approach can facilitate idea generation and serve as a starting point for developing more general models. Second, from a methodological perspective, this study is not limited to the application of the LOF and naïve Bayes models to TOA, as here we emphasised the systematic process of our approach in terms of inputs, throughputs, and outputs. Moreover, the LOF has been modified to consider the characteristics of the patent landscape for TOA, i.e., a large scale vector space model comprising ternary vectors and having a few spot positions with many patents. Although this study focused on TOA, the proposed approach could be employed in various research areas including weak signal detection and new product/service development. Finally, from a practical standpoint, our approach guides organisations towards setting up effective search strategies for new technology development. Especially with regard to TOA, the proposed method and the software system enable a wide-ranging search for ideas and the quick analysis of the novelty and value of ideas, thereby supporting decision making within acceptable limits of time and cost.

However, this study is subject to certain limitations, which should be complemented by future research. First, although naïve Bayes models provide the probability distribution of the value of ideas and alleviate the complexity and dimensionality problems of the landscape, the proposed approach cannot measure the value of an idea when the idea includes a new technological component (i.e., new combinations of technological components with at least one new technological component). That is, breakthrough ideas cannot be easily identified by the proposed approach. In this respect, the proposed approach will be more powerful if carefully integrated with other methods such as emerging clusters models (Breitzman and Thomas 2015) and other machine learning approaches using multiple patent indicators (Lee et al. 2018; Woo et al. 2018). Second, many issues remain as to how to improve the performance of the proposed approach. Other indicators (e.g., patent families and backward citations) and more advanced methods (e.g., support vector machines and deep learning techniques) could be helpful in finding novel and valuable ideas. Moreover, the integrated use of patent text data and classification information could be helpful in identifying the content of a patent and characterising the technological configuration of the patent. Third, the proposed approach cannot consider non-technological factors affecting technology development and commercialisation. The proposed approach needs to be elaborated further by employing other types of databases such as patent-product databases (Kim et al. 2017), futuristic databases (Kim and Lee 2017), and 10-K product databases (Hoberg and Phillips 2010). Fourth, this study focused on the patent landscape with $N=3$ and $K=0$, although the proposed method is not limited to this level and can allow for more complex analyses. Finally, this study conducted a single case study on pharmaceutical technology. Further testing on diverse technologies is essential to confirm the feasibility and validity of the proposed approach. Nevertheless, the systematic processes and quantitative outcomes of our approach offer a substantial contribution to both current research and future practice.

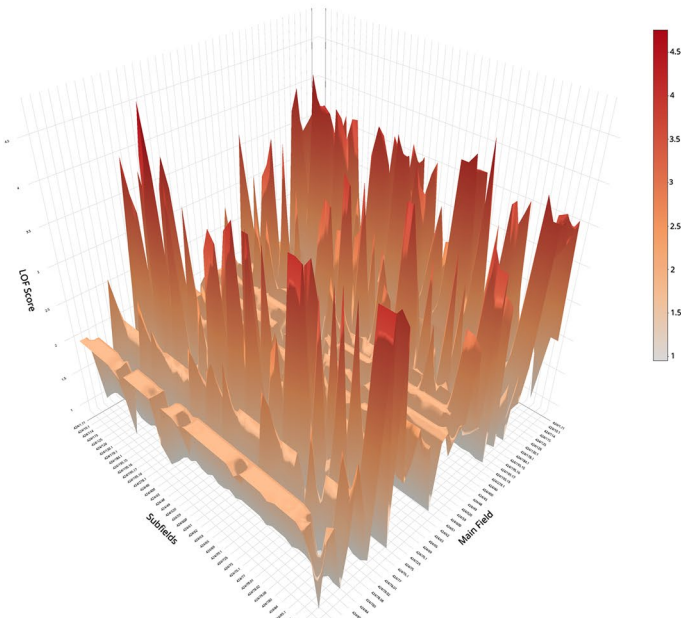
Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) Grants funded by the Korea Government (MSIP) (No. 2017R1C1B2011434).

Appendix 1: Patent landscapes for pharmaceutical technology

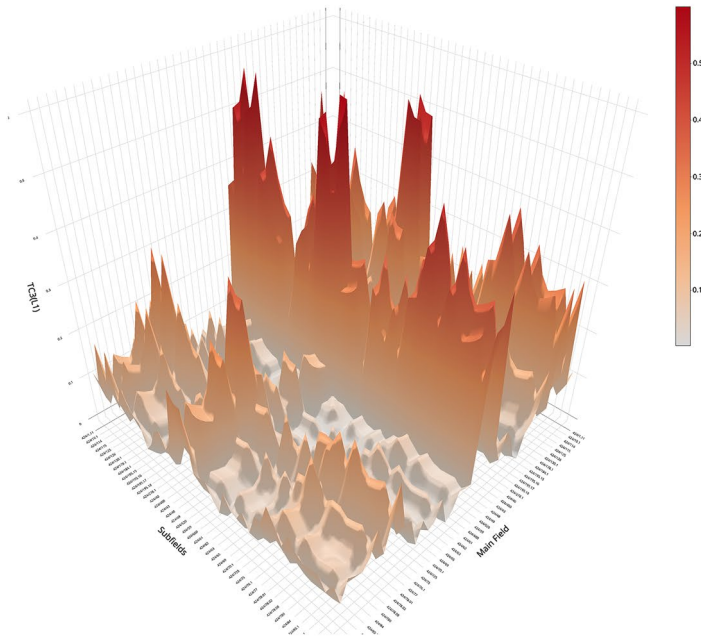
(a) *Count-focused patent landscape*



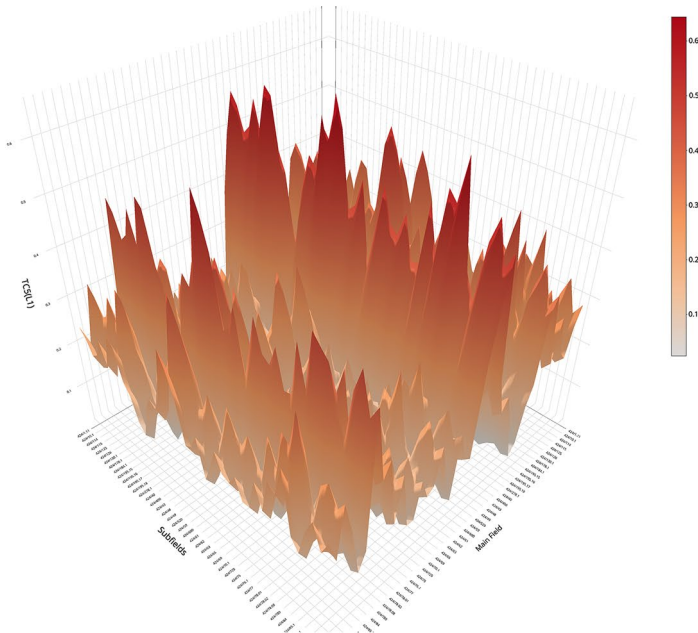
(b) *Novelty-focused patent landscape*



(c) *Value-focused patent landscape for TC3*



(d) *Value-focused patent landscape for TC5*



Appendix 2: Pseudo code of the proposed approach

(a) Measuring the novelty of ideas using the modified LOF

```

LOF (technological configurations_patents, technological configurations_ideas):

For  $p$  in patents:
    neighbours_patent[ $p$ ] = find the  $k$  nearest neighbouring patents to  $p$  via the Manhattan distance
    N_patent[ $p$ ] = calculate the number of  $k$  nearest neighbouring patents to  $p$       /** N_patent[ $p$ ]  $\geq k$  */
    k-distance[ $p$ ] = calculate the Manhattan distance between the  $k$ th nearest neighbouring patent and  $p$ 
End for

For  $p$  in patents:
    For  $n$  in neighbours_patent[ $p$ ]:
        distances_patent[ $p$ ][ $n$ ] = calculate the Manhattan distance between  $p$  and  $n$ 
        reachdist_patent[ $p$ ][ $n$ ] = max(k-distance[ $p$ ]+ $c$ , distances_patent[ $p$ ][ $n$ ]+ $c$ )
        /**  $c$  is a constant to be added to the original distance */
        sum_reachdist_patent[ $p$ ] = sum_reachdist_patent[ $p$ ] + reachdist_patent[ $p$ ][ $n$ ]
    End for
    lrd_patent[ $p$ ] = N_patent[ $p$ ] / sum_reachdist_patent[ $p$ ][ $n$ ]
End for

For  $i$  in ideas:
    neighbours_idea[ $i$ ] = find the  $k$  nearest neighbouring patents to  $i$  via the Manhattan distance
    N_idea[ $i$ ] = calculate the number of  $k$  nearest neighbouring patents to  $i$       /** N_idea[ $i$ ]  $\geq k$  */
    k-distance[ $i$ ] = calculate the Manhattan distance between the  $k$ th nearest neighbouring patent and  $i$ 
End for

For  $i$  in ideas:
    For  $n$  in neighbours_idea[ $i$ ]:
        distances_idea[ $i$ ][ $n$ ] = calculate the Manhattan distance between  $i$  and  $n$ 
        reachdist_idea[ $i$ ][ $n$ ] = max(k-distance[ $i$ ]+ $c$ , distances_idea[ $i$ ][ $n$ ]+ $c$ )
        /**  $c$  is a constant to be added to the original distance */
        sum_reachdist_idea[ $i$ ] = sum_reachdist_patent[ $i$ ] + reachdist_patent[ $i$ ][ $n$ ]
    End for
    lrd_idea[ $i$ ] = N_idea[ $i$ ] / sum_reachdist_idea[ $i$ ]
    For  $n$  in neighbours_idea[ $i$ ]:
        LOF_value[ $i$ ] = (1 / N_idea[ $i$ ] ) * sum(lrd_patent[ $n$ ] / lrd_idea[ $i$ ]))
    End for
End for
End function

```

(b) Estimating the value of ideas using naïve Bayes models

```

NaiveBayes(technological configurations_patents, values_patents, technological configurations_ideas):

For  $i$  in ideas:
    calculate the posterior probability of  $i$  for each value class      /** e.g., L1 and L2 */
    assign  $i$  to the corresponding value class
End for

compute performance metrics      /** e.g., confusion matrix, accuracy, precision, recall, and F1 score */
End function

```

References

- Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy*, 45(1), 81–96.
- Alcacer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flow: The influence of examiner citations. *Review of Economics and Statistics*, 88, 774–779.
- Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62–84.
- Arts, S., & Fleming, L. (2018). Paradise of novelty: Or loss of human capital? Exploring new fields and inventive output. *Organization Science*. <https://doi.org/10.1287/orsc.2018.1216>.
- Arts, S., & Veugelers, R. (2015). Technology familiarity, recombinant novelty, and breakthrough invention. *Industrial and Corporate Change*, 24(6), 1215–1246.
- Basalla, G. (1998). *The evolution of technology*. Cambridge: Cambridge University Press.
- Benner, M., & Waldfogel, J. (2008). Close to you? Bias and precision in patent-based measures of technological proximity. *Research Policy*, 37, 1556–1567.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.
- Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy*, 44(1), 195–205.
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2), 93–104.
- Chaudhuri, S. (2005). *The WTO and India's Pharmaceuticals Industry: Patent protection, TRIPS, and developing countries*. Oxford: Oxford University Press.
- Chen, Y. S., & Chang, K. C. (2010). The relationship between a firm's patent quality and its market value: The case of US pharmaceutical industry. *Technological Forecasting and Social Change*, 77(1), 20–33.
- Dahlin, K. B., & Behrens, D. M. (2005). When is an invention really radical? Defining and measuring technological radicalness. *Research Policy*, 34(5), 717–737.
- Daim, T. U., Rueda, G., Martin, H., & Gerdri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8), 981–1012.
- Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, 25(3), 233–242.
- Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1), 117–132.
- Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy*, 30(7), 1019–1039.
- Gerken, J. M., & Moehle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics*, 91, 645–670.
- Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics*, 36(1), 16–38.
- Harhoff, D., & Wagner, S. (2009). The duration of patent examination at the European Patent Office. *Management Science*, 55(12), 1969–1984.
- Hoberg, G., & Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10), 3773–3811.
- Kauffman, S. (1993). *The origins of order*. Oxford: Oxford University Press.
- 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.
- Kim, J., Kim, S., & Lee, C. (2019). Anticipating technological convergence: Link prediction using Wikipedia hyperlink. *Technovation*, 79, 25–34.
- Kim, J., & Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting and Social Change*, 92, 332–345.
- Kim, J., & Lee, C. (2017). Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, 120, 59–76.
- Lazarevic, A., Ertöz, L., Kumar, V., Ozgur, A., Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM international conference on data mining* (pp. 25–36).
- Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change*, 79(1), 16–29.
- Lee, C., Jeon, J., & Park, Y. (2011a). Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach. *Technological Forecasting and Social Change*, 78(4), 690–702.

- Lee, C., Kang, B., & Shin, J. (2015). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change*, 90, 355–365.
- Lee, C., Kim, J., Noh, M., Woo, H. G., & Gang, K. (2017a). Patterns of technology life cycles: Stochastic analysis based on patent citations. *Technology Analysis and Strategic Management*, 29(1), 53–67.
- Lee, Y., Kim, S. Y., Song, I., Park, Y., & Shin, J. (2014). Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. *Scientometrics*, 100(1), 227–244.
- Lee, C., Kwon, O., Kim, M., & Kwon, D. (2017b). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, 127, 291–303.
- Lee, C., & Lee, H. (2015). Novelty-focussed document mapping to identify new service opportunities. *Service Industries Journal*, 35(6), 345–361.
- 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, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, 127, 291–303.
- Lee, H. J., Lee, S., & Yoon, B. (2011b). Technology clustering based on evolutionary patterns: The case of information and communications technologies. *Technological Forecasting and Social Change*, 78(6), 953–967.
- Lee, C., Park, H., & Park, Y. (2013a). Keeping abreast of technology-driven business model evolution: A dynamic patent analysis approach. *Technology Analysis and Strategic Management*, 25(5), 487–505.
- Lee, C., Song, B., & Park, Y. (2009a). Generation of new service concepts: A morphology analysis and genetic algorithm approach. *Expert Systems with Applications*, 36(10), 12454–12460.
- Lee, C., Song, B., & Park, Y. (2013b). How to assess patent infringement risks: A semantic patent claim analysis using dependency relationships. *Technology Analysis and Strategic Management*, 25(1), 23–38.
- Lee, S., Yoon, B., & Park, Y. (2009b). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29(6), 481–497.
- Lerner, J. (1994). The importance of patent scope: An empirical analysis. *RAND Journal of Economics*, 25, 319–333.
- Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4–15).
- Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. *Research Policy*, 16(2–4), 143–155.
- Nelson, R. R., & Winter, S. (1982). *An evolutionary theory of economic change*. Cambridge: Harvard University Press.
- Park, Y., & Yoon, J. (2017). Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering. *Technological Forecasting and Social Change*, 118, 170–183.
- Rosenberg, N. (1979). Technological interdependence in the American economy. *Technology and Culture*, 20(1), 25–50.
- Schilling, M. A., & Green, E. (2011). Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy*, 40(10), 1321–1331.
- Schumpeter, J. (1939). *Business cycles*. New York: McGraw-Hill.
- Son, C., Suh, Y., Jeon, J., & Park, Y. (2012). Development of a GTM-based patent map for identifying patent vacuums. *Expert Systems with Applications*, 39(3), 2489–2500.
- Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8), 1445–1461.
- Tan, P. N. (2006). *Introduction to data mining*. Delhi: Pearson Education India.
- Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The RAND Journal of Economics*, 21(1), 172–187.
- Usher, A. (1954). *A history of mechanical invention*. Cambridge, MA: Dover.
- USPTO. (2006). Overview of the US patent classification system (USPC), electronic document at <http://www.uspto.gov>.
- Wang, X., Ma, P., Huang, Y., Guo, J., Zhu, D., Porter, A. L., et al. (2017). Combining SAO semantic analysis and morphology analysis to identify technology opportunities. *Scientometrics*, 111(1), 3–24.

- Wang, H. G., Wu, Z., Meng, F. L., Ma, D. L., Huang, X. L., Wang, L. M., et al. (2013). Nitrogen-doped porous carbon nanosheets as low-cost, high-performance anode material for sodium-ion batteries. *Chemsuschem*, 6(1), 56–60.
- Woo, H. G., Yeom, J., & Lee, C. (2018). Screening early stage ideas in technology development processes: A text mining and k-nearest neighbours approach using patent information. *Technology Analysis and Strategic Management*. <https://doi.org/10.1080/09537325.2018.1523386>.
- Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, 90(2), 445–461.
- Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change*, 72(2), 145–160.
- Yoon, B., & Park, Y. (2007). Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information. *IEEE Transactions on Engineering Management*, 54(3), 588–599.
- Yoon, J., Park, H., & Kim, K. (2013). Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. *Scientometrics*, 94(1), 313–331.
- Yoon, B., Phaal, R., & Probert, D. (2008). Morphology analysis for technology roadmapping: Application of text mining. *R&D Management*, 38(1), 51–68.
- Yoon, B., Yoon, C., & Park, Y. (2002). On the development and application of a self-organizing feature map-based patent map. *R&D Management*, 32(4), 291–300.