Scientometrics (2020) 125:1877–1897
https://doi.org/10.1007/s11192-020-03710-3

Identification of promising inventions considering the quality of knowledge accumulation: a machine learning approach

Uijun Kwon[1] · **Youngjung Geum**[1]

Received: 15 November 2018 / Accepted: 3 September 2020 / Published online: 21 September 2020

Abstract
The identification of promising inventions is an important task in technology planning practice. Although several studies have been carried out using patent-based machine learning techniques, none of these have used the quality of knowledge accumulation as an input for identifying promising inventions, and have simply considered the number of backward citations as the link with previous knowledge. The current study therefore aims to fill this research gap by predicting promising inventions with patent-based machine learning, using the quality of knowledge accumulation as an important input variable. Eight criteria and 17 patent indicators are used as input variables, and patent forward citations are employed as the output variable. Six machine learning techniques are tested on 363,620 G06F patents filed between January 1990 and December 2009, and the results show that the quality of knowledge accumulation is the most important variable in predicting emerging inventions.

Keywords Promising technology · Technology forecasting · Patent analysis · Machine learning · Patent indicator

Introduction

Technology has formed a central part of modern innovation, and this is especially true in the current commercial arena, where many technology-intensive products based on artificial intelligence are designed, developed and launched onto the market. It is evident that new innovative ideas are primarily recognised based on technological advances and improvements; in the literature, these are referred to as technology-push innovations (Chau and Tam 2000; Uriona-Maldonado et al. 2010). For this reason, the existing literature has been almost unanimous in support of the value of technology planning. This is critically important, since planning innovation generally begins under conditions of extreme uncertainty; this is known as a fuzzy front end (Brem and Voigt 2009), and needs to be dealt with using a well-planned and integrative planning method.

✉ Youngjung Geum
yjgeum@seoultech.ac.kr

[1] Department of Industrial and Information Systems Engineering, Seoul National University of Science and Technology, Seoul, South Korea

When planning technological innovation, the most important step involves an understanding of technology trends and identification of promising technological inventions. This is especially important in view of the rapid changes in technological development trends and the limited resources of each firm. Since the time and resources available for new product development are not endless, it is critically important to understand promising technologies within the industry. This task is known as technology forecasting, and has been widely discussed, both in theory and in practice (Daim et al. 2006). As Porter et al. (1991) have noted, there are several technological forecasting methods, and these can be generally classified into three types: direct methods, which directly forecast the parameters used to measure technological developments based on expert opinion or trend extrapolation; correlative methods, which use scenarios or cross-impact analysis to measure the correlative characteristics of technological factors; and structural methods, which focus on the cause and effect relationships that affect growth (Porter et al. 1991; Daim et al. 2006).

Recently, many studies of the identification of promising technologies have examined patents. Patents are outputs of technological innovation in a knowledge-based economy, and are important indicators for industrial R&D activities (Lerner 1994; Daim et al. 2006). Patents also reflect the cumulative processes involved in technological changes and up to 80% of all technological knowledge can be assumed to be described in patent applications (Teichert and Mittermayer 2002; Geum et al. 2017). For this reason, many studies have used patents in the identification of promising technology (Park et al. 2007; Yoon and Kim 2011; Joung and Kim 2017; Suominen et al. 2017; Lee et al. 2018).

There are two main research streams in the patent-based identification of promising technology. The first of these focuses on the preliminary work necessary to identify promising technology, such as patent-based trend analysis (Yoon and Kim 2011; Breitzman and Thomas 2015; Joung and Kim 2017). The second stream is more closely associated with the actual identification of promising technology. Daim et al. (2006) integrated patent analysis with several well-known technology forecasting methods, including scenario planning, growth curves, analogies, and system dynamics, in order to forecast emerging technologies. Geum et al. (2013) employed a content-based novelty detection technique to identify promising technologies. In this study, text mining was carried out to analyse the contents of patents, and several novelty detection techniques were used to identify unexpected and novel patterns in these patents. Ju and Sohn (2015) proposed a hierarchical quality function deployment (QFD) framework for identifying emerging technologies and related business models. Lee et al. (2015) integrated local outlier factor (LOF) and text-mining techniques to identify novel technologies, while You et al. (2017) developed a trend forecasting model using a patent-based time-series analysis.

Within the second research stream, machine learning techniques are recently employed to predict promising technologies meeting industrial needs for data-driven approaches (Trappey et al. 2012; Wu et al. 2016; Kyebambe et al. 2017; Lee et al. 2018). Trappey et al. (2012) extracted relevant patent quality indicators and analysed the quality of patents using a back-propagation neural network model. Wu et al. (2016) developed a framework for automatic patent quality analysis and classification system using several methods such as self-organizing maps, principal component analysis, and support vector machine. Kyebambe et al. (2017) set out to identify emerging technologies using several input variables including the number of claims, patent citations, technology lifecycle, patent class, and the similarity between cited patents. Lee et al. (2018) used machine learning techniques to identify emerging technologies in the early stages of innovation. Using novelty, science intensity, growth speed, scope and coverage, and development effort/capabilities as the main input categories, these authors extracted 18 patent indicators matching these input

categories. These were then used as input variables for a feed-forward multilayer neural network that was employed to forecast the number of citations.

Despite contributions from previous studies of promising technology identification, a specific research gap remains. Most prior studies have used the linkage with current knowledge, measured as the number of backward citations, as their input variable, meaning that if there is a high level of knowledge accumulation for developing a particular technology, it is likely to be a promising technology (Kyebambe et al. 2017; Lee et al. 2018). In work by Schoenmakers and Duysters (2010), certain patents (i.e. the promising technologies examined in this study) have more backward citations than others. Breitzman and Thomas (2015) also showed that a patent is likely to be an emerging technology when considering previous linkage to the hot patents. This means that knowledge accumulation can be considered to be an important characterising variable when identifying promising technology.

However, previous studies using machine learning to predict promising technologies have examined only the extent of knowledge accumulation, without considering its quality. Even if different patents contain the same number of backward citations, the impact of knowledge accumulation will differ significantly according to the patent quality of each backward citation. Here, the 'quality' of backward citation can be measured in many different ways, but the body of literature has been almost unanimous in using patent forward citation is a good proxy for measuring technological impact (Azagra-Caro et al. 2017). Since patent citations can reflect the degree to which papers are part of the technological state-of-the art (Meyer 2000, p. 425), representing the body of understanding in knowledge structure (Nelson 1998; Meyer 2000), knowledge accumulation quality can be measured by the number of forward citations of backward citation patents.

In view of this, the current study aims to predict promising technologies using patent-based machine learning, taking into consideration the quality of knowledge accumulation. This study examines not only the number but also the quality of backward citations in order to measure the quality of knowledge accumulation in technological innovation. This study also employs several machine learning techniques and compares the results of prediction.

The remainder of this study is organised into four parts. Section 2 contains a *literature review* that provides the context for this study in terms of the ways in which identification of promising technology has been achieved in previous studies. Section 3 describes the *proposed approach*, and deals with the general and detailed processes involved, including data collection and machine learning. To illustrate our proposed approach, a *case study* is conducted in Sect. 4. Finally, Sect. 5 presents a *conclusion*, in which the contributions and limitations of this study are summarised.

Literature review

Identification of promising technology

The identification of promising technology has long been considered a critical task. Two questions have been actively discussed in the literature: what are promising technologies, and how can these be identified?

The definition of promising technologies has been addressed in many studies. Rotolo et al. (2015) defined emerging technologies using five characteristics: radical novelty, fast growth, coherence, prominent impact, and uncertainty and ambiguity. Verhoeven et al. (2016) characterise technological inventions using two dimensions of technological

novelty. Based on the assumption that technologies differ from each other in terms of how they combine existing components and principles, these authors defined two types of novelty: novelty in recombination, and novelty in knowledge origins. Using these aspects as a basis, they proposed patent-based operationalisation. From a data-driven perspective, promising technology is defined as a collection of highly cited patents in many studies (Noh et al. 2016; Song et al. 2017). As an extension of a technological perspective that defines promising technology, Song et al. (2017) defined promising technology as "technology that is likely to have a substantial impact on other technologies as well as those that can respond to market needs" (Song et al. 2017, p. 119) by integrating the perspective of market needs.

The second question of how to identify promising technologies has also been addressed by many scholars. Since a patent is a powerful proxy for technological innovation, previous studies on identifying promising technologies mainly rely on patent-based approaches. Daim et al. (2006) integrated a patent bibliometric analysis with various forecasting tools, including growth curve analysis and system dynamics. Yoon (2008) used patent information in a morphological analysis, which is a well-known tool for idea generation, and developed a software prototype. Kim and Seol (2012) used patent data to identify core technologies with an integrated approach involving association rule mining (ARM), an analytic network process (ANP), and data envelopment analysis (DEA). The co-occurrence, relatedness and cross-impact were measured based on a co-classification of patent information.

The co-classification and co-citation of patents are also frequently used to identify promising technologies. Noh et al. (2016) employed the concept of RFM (recency, frequency and monetary) to identifying promising technologies. Song et al. (2017) attempted to identify emerging technology from patent bibliographic coupling using the three criteria of impact, applicability, and sustainability, and to measure technological and market characteristics for each technology. Wang and Duan (2011) also identified core technologies in the electric vehicle industry using patent co-citation information. Kim and Bae (2017) attempted to forecast promising technology using patent analysis; a clustering analysis was conducted using patent classification, and each cluster was then examined using patent indicators. Park et al. (2007) developed technology portfolios using a patent IPC distribution vector, and applied collaborative filtering to the company's portfolios in an analysis of technological opportunity. Joung and Kim (2017) suggested a keyword-based analysis for monitoring emerging technologies in which a technical keyword-context matrix was constructed and keyword pairs were identified to measure the relatedness between keywords. They also suggested a flowchart to check whether or not a keyword was promising. Li et al. (2018) suggested a framework for monitoring the development trend of emerging technologies and for identifying emerging technologies. For this purpose, they used both a patent database and Twitter, with a focus on perovskite solar cell technology.

Identification of promising technology using supervised machine learning

Although numerous patent-based studies of the identification of promising technology have been carried out, machine learning techniques have rarely been used. Most prior studies have employed text-based topic identification, keyword co-occurrence analysis, patent co-classification analysis and patent co-citation analysis.

The identification of promising technology can be considered a typical classification problem, since it is related to the question of whether or not this technology will become popular. For this reason, supervised machine learning has recently been employed in

several studies to identify promising technologies (Trappey et al. 2012; Kyebambe et al. 2017; Lee et al. 2018).

Trappey et al. (2012) conducted a patent quality analysis using several patent-based indicators including application length, IPCs, UPCs, foreign citation, forward citation, backward citation, claims, independent claims, patent families, technology cycle time, science linkage, and the length of specification. A principal component analysis (PCA) was then conducted to identify key impact factors from several patent indicators, and a back-propagation neural network model was implemented for these PCA results. Kyebambe et al. (2017) used supervised machine learning techniques with several input variables, such as the number of claims, number of patent citations, number of citations of non-patent literature, technology lifecycle, patent class, and similarity between cited patents. It is notable that this work uses the number of patent citations and number of citations of non-patent literature as discriminating characteristics to identify promising technologies. Similarly, Lee et al. (2018)'s work also tried to identify emerging technologies using machine learning. Five categories were used as patent input indicators: novelty, science intensity, growth speed, scope and coverage, and development effort and capabilities. This work also used the number of backward citations and the number of non-patent literature references.

Although several previous works have tried to identify emerging technologies, the quality of knowledge accumulation has not been taken into account; these approaches simply use backward citations as discriminating characteristics, without considering the quality of the citations.

Proposed approach

Research framework

This study focuses on identifying promising technologies considering the quality of knowledge accumulation. Previous studies that predict promising technologies have examined only the extent of knowledge accumulation, without considering its quality. However, the consideration of its quality of backward citation is critically important. For example, the technological strength of a patent that cites other very highly cited patents will be different from one that cites other patents with few citations, as shown in Fig. 1. This figure shows why quality of knowledge accumulation should be considered. Therefore, our study aims to introduce the concept of knowledge accumulation quality, i.e. number of citation of backward citation patents. Not only considering the number of backward citation, the quality of backward citation is also considered.

Figure 2 illustrates the overall process used in this paper. First, input and output variables are defined in order to identifying promising technologies. The primary task is to define the variables characterising the concept of a promising technology; this is very important, since the inputs and outputs used in identifying promising technologies are determined based on these variables. These variables can be divided into eight types: newness, technological knowledge base, technological generality, technology lifecycle, technology protection coverage, technological scope, technical strength of inventors, and technological activity.

When the characterising variables have been defined, data are collected and preprocessed in order to reflect these variables in the modelling process. When data preprocessing

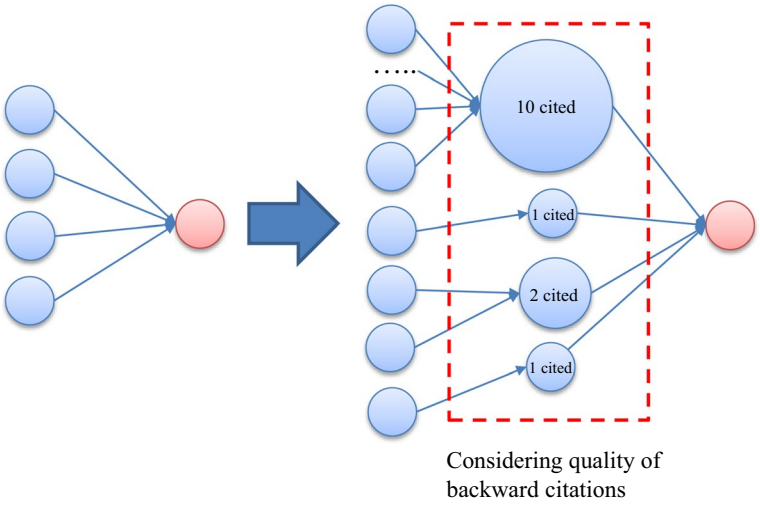

Fig. 1 Core concept of our paper

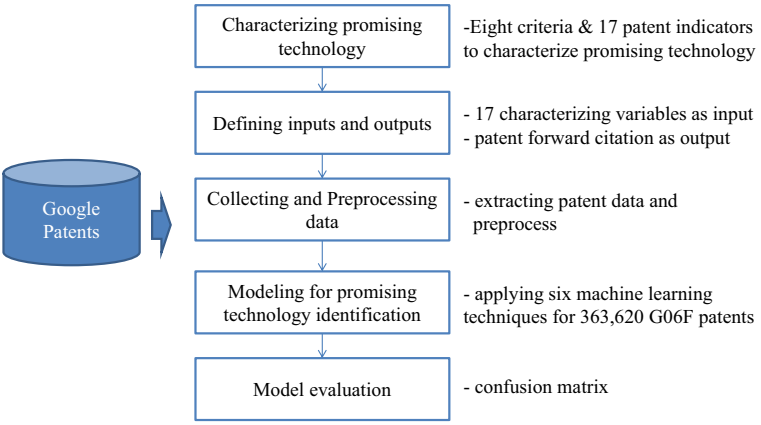

Fig. 2 Overall process

is complete, classification of promising technology is carried out using several machine learning techniques. Each model is evaluated based on a confusion matrix.

Variables

Output variables

First, the output variable is defined for the prediction of emerging technologies. Since our research problem involves identifying promising and important technologies, the number of forward citations of a patent is used as the output variable; this approach is strongly supported by previous studies as an indicator of technological strength. When a patent

is frequently cited by other patents, this means that it makes a significant contribution in terms of technological development (Lanjouw and Schankerman 1999). There is a substantial body of literature to show that highly cited patents are generally technological leaders (Geum et al. 2013; Verhoeven et al. 2016).

For this reason, the number of citations (forward patent citations) is used as a quality indicator in this study. However, certain issues need to be examined when dealing with the number of citations; for instance, the number of forward citations accumulates over time. Therefore, the total number of citations over two different periods (5 and 10 years) is used to address this problem.

Input variables

Next, several input variables are identified in order to characterise the technological aspects of a patent. Eight different factors are used to define input variables: (1) recombinative novelty; (2) knowledge accumulation; (3) technological life cycle; (4) technology protection coverage; (5) technological jurisdiction; (6) technical capability of the assignee; (7) technological generality; and (8) technological activity. Relevant indicators are prepared for each of these eight criteria, and a total of 17 indicators are prepared as input variables for machine learning.

Recombinative novelty Many scholars have proposed that technological novelty arises from the recombination and synthesis of existing technologies (Fleming and Sorenson 2001). With respect to this theory, some researchers (Rosenkopf and Nerkar 2001; Shane 2001) have found that the number of technological classes a patent cites outside of its own technology class can be used as a proxy measure of novelty (Verhoeven et al. 2016). This means that the overlap between classes can be considered as a measure of technological novelty. Similarly, Fleming et al. (2007) suggested a concept of generative creativity, which is defined as the occurrence of new combinations of subclasses of patents.

Following prior works, novelty in recombination can be measured as a new combination of IPC groups that had not previously been connected. However, in order to calculate the novelty in recombination, collecting all patents for a relevant IPC classification is required, which would be very time-consuming. In view of the complexity of calculating all IPC combinations for all patents in the relevant IPC classification, this study instead calculates the recombinative novelty for an IPC combination of reference patents (backward cited patents), as shown in Table 1.

Table 1 Calculation example of novelty recombination

Patent number	IPC	IPC pair	Backward citation patent	IPC pair for backward citation patent	Novelty in recombination
7,117,073	B60Q1	B60Q1-B60R21	6,919,822	B60Q1-B60R21	x
	B60R21	B60Q1-B62D15		B60Q1-B62D15	x
	B62D15	B60Q1-G01S5		B60Q1-G08G1	o
	G01S5	B60Q1-G06F19		B60Q1-B60K35	o
	…	…	…	…	…
					Total = 7

Knowledge accumulation Whether or not a certain technology is promising can be determined based on the extent of knowledge accumulation, as many studies have noted (Harhoff et al. 2003; Verhoeven et al. 2016; Lee et al. 2018). The number of backward citations is generally used to measure the extent to which a certain technological knowledge builds on existing knowledge (Verhoeven et al. 2016). In work by Schoenmakers and Duysters (2010), highly cited patents (i.e. the promising technologies in this study) had more backward citations than other patents. This means that knowledge accumulation can be considered an important characterising variable to identify promising technology. Harhoff et al. (2003) showed that the number of backward citations, including both patent and non-patent references, is related to the value of the patent. Lee et al. (2018) used the number of backward citations, which is a proxy for knowledge accumulation, as the input variable in the prediction of emerging technologies. It is important to note that it is sometimes difficult to use knowledge accumulation as the decision criterion for characterising promising technology, since in some cases, patents with no backward citations represent radical innovations rather than non-promising technologies (Ahuja and Lampert 2001; Banerjee and Cole 2011).

In this study, two types of indicators are employed to measure knowledge accumulation. First, the extent of knowledge accumulation is measured by the number of backward citations. Second, the quality of knowledge accumulation is measured, addressing the limitations of previous studies. To reflect the quality of knowledge accumulation, two indicators are added: the average number of forward citations that the backward cited patents received, and the maximum number of citations that the backward cited patents received. The reason to use the number of forward citations of each backward citation patent as the knowledge accumulation quality is that the body of literature has been almost unanimous in using patent forward citation is a good proxy for measuring technological impact (Azagra-Caro et al. 2017). Whether the two variables to measure knowledge accumulation are helpful or not in order to predict the patent quality will be tested in the experiments.

Technology lifecycle Technology lifecycle is measured by the technology cycle time (TCT), which is the median age of the patent references cited in the patent (Karki 1997; Kayal and Waters 1999; Huang et al. 2015). The more recent the cited patents, the more quickly a new generation of inventions is replacing the current one (Kayal and Waters 1999). In other words, this factor measures both the speed of technological innovation and the immediate impact of research on technological innovation (Huang et al. 2015).

Technology protection coverage This is defined as the number of claims in the patent. Following previous work, each claim represents a distinct contribution of a given patent (Tong and Frame 1994). It also describes the essential features of an innovation and is the subject of legal protection (Sheremetyeva 2003). A higher number of claims means that this innovation is protected in terms of its specified and diversified aspects; thus, the number of claims can provide information about technological activities and is related to the quality and the value of the innovation output (Lanjouw and Schankerman 1999). Both independent and dependent claims may be prepared, representing the core characteristics of innovation and detailed explanations, respectively. Therefore, both types of claim are used as indicators of technology protection coverage.

Technological jurisdiction Technological jurisdiction indicates the number of countries in which a patent is protected for a given invention (Putnam 1996; Harhoff et al. 2003). According to the principle of territorial privilege for jurisdiction, patents are protected

within each individual country, and thus if patent rights are to be protected in several countries, a patent has to be filed in each. The size of the patent family is frequently employed as a quality indicator for measuring patent value, since it represents both technological importance and innovative value in terms of the fees, translation and legal costs involved in applying for and maintaining each patent in multiple countries (Putnam 1996; Lanjouw and Schankerman 1999, 2004; Harhoff et al. 2003).

Technical capability of assignee The technical capability of the assignee is an important criterion in identifying promising technologies, since patents originating in highly innovative firms are likely to have high value (Ernst 2003). In this study, two indicators are used to measure technical capability of assignee: assignee activity and the technical strength of the assignee. Assignee activity is defined as the total number of patents registered by an assignee; a higher number of patents indicates that a certain institution/company is trying to develop and innovate technologies. The technical strength of the assignee is defined as the number of citations of patents for which this assignee has applied; the number of citations has been used for some time as an important indicator in the literature (Lanjouw and Schankerman 1999).

Technological activity Technological activity measures the extent to which technological development and innovation is active. It can be defined twofold: the number of patents in the relevant patent IPC, and its increase compared to that in the previous year. According to Ernst (2003), patent activity is defined as the number of patent applications in a specific technological fields. Considering that a certain IPC itself is a technological field, the number of patents in the relevant IPC can be represented as technological activity. This indicator is also useful in reflecting the different conditions of each patent registered in different years.

Technological generality Technological generality is defined as diversity of technological characteristics, and is measured by the number of IPC classes to which a given patent belongs. In many studies, the number of IPC classes is used to measure the scope of a certain technology, and is related to the technological and economic value of a patent (Lerner 1994; Matutes et al. 1996; Harhoff et al. 2003). This indicator has recently become particularly important since many technological innovations now arise from a recombination of existing technologies and technological convergence.

Using eight criteria, 17 patent indicators are defined as input variables in order to reflect the characteristics for promising technologies. Table 2 summarizes input variables used in this study.

Data collection and preprocessing

Patent data were collected using the Google BigQuery platform, and relevant patent indicators were extracted from the patent database. Each dataset was then preprocessed, including outlier processing and data normalisation.

Table 2 Input variables used in this study

Type	Indicator	Variable	Operational definition	Reference
Novelty	Recombination novelty	Novelty recombination (NR)	Number of new combinations compared to previous IPC pairs in backward-citation patents	Fleming and Sorenson (2001), Fleming et al. (2007), Gruber et al. (2013), Verhoeven et al. (2016)
Previous knowledge accumulation	Extent of knowledge accumulation	Patent reference (PR)	Number of backward citations	Harhoff et al. (2003), Callaert et al. (2006), Lanjouw and Schankerman (2001), Verhoeven et al. (2016), Lee et al. (2018)
		Non-patent reference (NPR)	Number of non-patent backward citations	
		Prior patent reference average (PPRa)	Average number of backward citations of backward cited patents (reference patents)	
		Prior patent reference max (PPRm)	Maximum number of backward citations of backward cited patents (reference patents)	
	Quality of knowledge accumulation	Patent reference forward citation average (PRFa)*	Average number of forward citations of backward cited patents (reference patents)	Ernst (2003)
		Patent reference forward citation max (PRFm)*	Maximum number of forward citation of backward citation patent (reference patents)	
Technology lifecycle	Technology cycle time	Technology cycle time (TCT)	Median age of backward citations	Narin (1999), NSF (1996)
Technology protection coverage	Number of claims	Independent claims (IC)	Number of independent claims	Tong and Frame (1994), Lanjouwand and Schankerman (1999), Trappey et al. (2012)
		Dependent claims (DC)	Number of dependent claims	
Technological jurisdiction	Number of patent family	Patent family (FP)*	Number of patent family	Putnam (1996), Lanjouw and Schankerman (1999), Harhoff et al. (2003), Lanjouw and Schankerman (2004)

Table 2 (continued)

Type	Indicator	Variable	Operational definition	Reference
Technical capability of assignee	Assignee activity	Assignee's patent count (AP)	Number of patents of assignee	Ernst (2003)
	Technical strength of assignee	Assignee patent forward citation average (APFa) *	Average number of citations of assignee's patents	Ernst (2003), Lee et al. (2018)
		Assignee patent forward citation max (APFm) *	Maximum number of citations of assignee's patents	
Technological generality	Technological generality	IPC	Number of IPCs for each patent	Lerner (1994), Matutes et al. (1996)
Technological activity	Extent of technological activity	Technology activity (TA)	Number of patents in the relevant IPC main class	Ernst (2003)
	Increasing rate of technological activity	Technological activity increasing rate (TAr)	Increasing rate of technology activity	

*Ex-post variables

Modelling and evaluation

For the modelling, six classification algorithms were used identify promising technologies, and the algorithm with highest performance are finally employed. The Python library scikit-learn (version 0.19.1) is used to implement this. For the evaluation, a confusion matrix is used to evaluate the model; this is a matrix allowing a comparison of the results of the actual and predicted classes, as shown in Table 3.

Based on the confusion matrix, several performance indicators are calculated. The accuracy is measured based on the ratio of correct classifications in the test dataset. Precision is measured as the ratio of true positives within data that are predicted to be positive, and can therefore be seen as a measure of the exactness of the quality of classification (Patil and Sherekar 2013). Recall is the ratio of true positives within the data that are actually positive, which can be considered as completeness (Patil and Sherekar 2013). Finally, the F-measure is defined as a harmonic mean of the precision and recall.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FN}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$F_1 = 2 \times \frac{\mathrm{Pr}\,ecision \times \mathrm{Re}call}{\mathrm{Pr}\,ecision + \mathrm{Re}call}$$

Results

Data

Patent data were collected from the Google Patent website. Patents whose IPC subclass is G06F (electric digital data processing) are collected since the number of patents in this subclass is the highest. This technological area also forms the basis of the technologies of big data analytics and the Internet of Things, two important technological developments in engineering practice.

Patents from January 1990 to December 2009 were collected. Based on this, two different datasets were prepared: dataset-5 and dataset-10. For example, when there are three

Table 3 Confusion matrix

		Predicted	
		Positive	Negative
Observed	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Table 4 Example of preparing 5-year dataset and 10-year dataset

Patent	Year	5-year dataset (Forward citation for 5 years)	10-year dataset (Forward citation for 10 years)
A	1995	Relevant year: 1995–2000 Use	Relevant year: 1995–2005
B	2000	Relevant year: 2000–2005 Use	Relevant year: 2000–2010 Delete from this database
C	2005	Relevant year: 2005–2010 Delete from this database	Relevant year: 2005–2015 Delete from this database

patents (A registered in 1995, B registered in 2000, and C registered in 2005), the collection of forward citation is explained in Table 4.

Because forward citations during selected (5 or 10) years after each patent' registered year are collected, time lag problem can be resolved. In 5-year dataset, forward citations during 5 years after each patent' registered year are collected. This is same in other dataset that deals with 10-year dataset.

At first, 363,620 target patents were collected from January 1990 to December 2009, and again 677,750 backward citation patents were collected. The same procedures were used for collecting forward citations (5-year and 10-year from the registered year) for each backward cited patents. Two types of datasets are used, according to the forward-citation counting period. When the patent is too recent to meet the requirement, this patent was dropped from the dataset. As the result, two types of datasets were obtained, as shown in Table 5.

Since the quality of knowledge accumulation is used as an important input for modelling, backward cited patents (i.e. those cited by the target patents) were also required, and 677,750 backward cited patents were therefore collected using Google BigQuery. Finally, based on this patent information, 17 input variables and three output variables were prepared for modelling. All variables were normalised using min–max normalisation.

The number of forward citations are defined as an output variable to measure the quality of a patent. To ensure reliability, 5-year and 10-year cumulative citations were used. Each output variable is transformed to a binary value (0, 1), using the top 10% of the citations as a threshold, since the purpose of this study is to predict whether or not a patent involves promising technology, which is a classification problem. In the literature, there are several works in which promising technologies are defined based on the number of citations, for

Table 5 Dataset explanation

Dataset type	Explanation	Patent count
5-year dataset	Including patents whose registration year plus 5 year does not exceed 2009 (Registration year + 5 year < 2009)	209,727
10-year dataset	Including patents whose registration year plus 10 year does not exceed 2009 (Registration year + 10 year < 2009)	91,288

Table 6 Results of classification

Model		Accuracy	AUC
LR	Logistic regression	0.9021	0.5508
SVM	Support vector machine	0.9059	0.5597
DT	Decision tree	0.8636	0.6420
RF	Random forest	0.9059	0.5984
ADA	Adaptive boosting	0.9040	0.6035
XGB	Extreme gradient boosting	0.8704	0.7820

Table 7 Results of classification

Period	Prediction	Accuracy	Precision	Recall	F1
Five years	Not promising	0.8704	0.9597	0.8931	0.9252
	Promising		0.4172	0.6709	0.5144
	Average		0.9041	0.8703	0.8831
Ten years	Not promising	0.8940	0.9444	0.9374	0.9409
	Promising		0.4726	0.5040	0.4878
	Average		0.8972	0.8940	0.8955

which the decision criterion is the top 1%~10% (Tijssen et al. 2002). Thus, if the number of citations exceeds top 10%, the value is set to one; otherwise, it is set to zero.

Results of modelling

Of the entire dataset, 70% were used as a training set, and the remaining 30% were used as the test set. Ten-fold cross validation was used to check the generalisability of the model. Six different models were employed to compare their performance; Table 6 shows that the XGBoost model gave the best results.

As part of the model optimisation, a grid search was applied to identify the optimal parameter. Table 7 presents the final results of modelling. The accuracy of identification of promising technologies was around 0.87 to 0.89 which is quite good enough; this value decreased when the number of five-year citations was used as the output variable, but increased when the number of 10-year citations was used (note that the output variable was then transformed to a binary variable using the top 10% citation threshold).

Interpretation

To quantify the relative importance of each input variable, the Gini importance is calculated to evaluate the importance of each variable. The Gini importance is defined based on the improvement in the Gini index, and is a measurement of the total reduction in the error in a decision tree classifier. Table 8 shows variable importance in the 10-year model.

The input variables with the highest Gini importance are the average number of forward citations of backward cited patents, the family patent, and the assignee patent forward citation average. It is notable that the quality of knowledge accumulation (a variable added in this study) has the highest importance. This means that the quality of

Table 8 Top 10 important variables (for 10-year model)

Criterion	Variable	Name of variable	Importance
Quality of knowledge accumulation	Average number of forward citations of backward cited patents	Patent reference forward citation average (PRFa)	0.2213
Technological jurisdiction	Number of patent family	Family Patent(FP)	0.1331
Technical capability of assignee	Technical strength of assignee	Assignee Patent Forward citation average(APFa)	0.1048
Technology protection coverage	Number of claims	Independent Claims(IC)	0.1007
Technology protection coverage	Number of claims	Dependent Claims(DC)	0.0478
Technical capability of assignee	Technical strength of assignee	Assignee Patent Forward citation max(APFm)	0.0453
Technical capability of assignee	Assignee activity	Assignee's patent count (AP)	0.0425
Technology lifecycle	Technology cycle time	Technology cycle time (TCT)	0.0414
Technological activity	Extent of technological activity	Technology activity (TA)	0.0359
Technological generality	Technological generality	IPC	0.0357

knowledge accumulation is critical in the development of innovative technologies; in other words, patents that cite qualified and highly cited patents appear to be the most promising technologies.

Another important aspect of identifying promising technologies is the technical capability of the assignee. This means that companies with strong prior technological experience appear to develop innovative and promising technologies, since both the number of patents and number of citations of the assignee's patents can explain the technical strength of the assignees. The extent of knowledge accumulation also seems to be important in identifying promising technologies; this is in alignment with the results of previous works, which have found that knowledge accumulation (the number of backward citations) is important in the identification of promising technologies (Schoenmakers and Duysters 2010; Verhoeven et al. 2016; Wu et al. 2016; Kyebambe et al. 2017; Lee et al. 2018). Another important variable is the technological jurisdiction, as discussed in many studies (Putnam 1996; Lanjouw and Schankerman 1999, 2004; Harhoff et al. 2003).

Evaluation

To validate the model, our result is compared with the model not considering knowledge accumulation. The reference model is developed without knowledge accumulation quality. The result is shown in Table 9. Generally, the model performance is better in the model considering knowledge accumulation quality. In most measures including accuracy, precision, recall, and F1, models considering knowledge accumulation quality outperforms the models without knowledge accumulation quality.

Discussion

Important indicators for promising technologies

According to our result, quality of knowledge accumulation and technical capability of assignee are determined to be important variables to decide promising technologies. Therefore, further investigation is required to check why they are important, and whether our results are in line with previous literatures.

Table 9 Result without considering knowledge accumulation quality

Period	Prediction	Accuracy	Precision	Recall	F1
Five years	Not promising	0.8490	0.9601	0.8679	0.9117
	Promising		0.3711	0.6837	0.4811
	Average		0.9000	0.8490	0.8676
Ten years	Not promising	0.8730	0.9403	0.9172	0.9286
	Promising		0.3902	0.4763	0.4290
	Average		0.8852	0.8730	0.8786

Knowledge accumulation quality

There have been several studies to mention the importance of backward citations in predicting promising technologies (Schoenmakers and Duysters 2010; Breitzman and Thomas 2015; Verhoeven et al. 2016), thus backward citation has been considered as an important indicator for representing knowledge accumulation. Cammarano et al. (2017) mentioned that a cited patent can represent a piece of existing knowledge that the citing patents have built.

Our result is in line with previous literatures in terms of considering knowledge accumulation as an important indicator for predicting patent quality. What is notable in our study is to show the importance of not only the extent of knowledge accumulation, but also the quality of knowledge accumulation, i.e. how many citations does a backward citation patent received. This is very important implication because it shows that patents that cite highly cited patents are also likely to be qualified and promising patents. The result can be also employed for identifying important technological evolution path, by analysing the trajectories of highly cited patents or classifying citation patterns of patents.

Technological power of assignee

There have been several works to discuss the technological power of assignees (Putnam 1996; Ernst 2003; Ernst 2003; Lee et al. 2018). Ernst (2003)'s work employed company-related measures for understanding patent quality, such as co-operation intensity, share of granted patents, or technological scope. Lee et al. (2018) employed core area know-how and peripheral area know-how for predicting emerging technologies. In this study, the number of patents in a technology field of interest issued by an assignee is used as a proxy for core area knowhow, and number of patents in other technology fields issued by an assignee is used as a proxy for peripheral area know-how. Our result is in line with previous studies that technological power of assignee is very important to decide whether a patent is promising or not.

Additional considerations

Consideration of technological novelty

Since the novelty is one of the most important criteria for patent application, one has to consider novelty to predict the patent quality. Our study employed recombinative novelty, which is measured by the number of new combinations compared to previous IPC pairs. This is based on the previous literatures that technological novelty can be measured by the number of patent classes outside of its own technological class (Fleming et al. 2007; Rosenkopf and Nerkar 2001; Verhoeven et al. 2016). However, the contents-based novelty is not considered in this study, which is a limitation of this study. Therefore, the use of contents analysis tools such as natural language processing or topic modelling can improve the results.

Use of alternative measures for the output variable

In this study, the forward citation count is used as the output variable, since the patent forward citation has been extensively used to represent the technological value (Karki 1997;

Lanjouw et al. 1998). However, this is simply one form of value, but not necessarily private value. Many scholars have mentioned that technological impact and commercial value are two different and separate issues, which means patents with high forward citation do not necessarily mean that they are enough to bring profits to their owners (Thomas 1990). Therefore, it is required to test the model based on the commercial value, not only based on the technological value. To measure commercial value of patents, several measures can be employed such as patent renewal fee or renewal decisions (Schankerman 1991; Bessen 2008) or patent transactions or reassignment (Choi et al. 2015; de Marco et al. 2017; Graham et al. 2018; Huang et al. 2018). Therefore, comparing the results with technological value derived from patent citation and market value derived from patent renewal decisions or patent transactions can provide more valuable insights and makes the research more concrete.

Conclusion

This study attempts to predict promising technologies using patent-based machine learning, considering the quality of knowledge accumulation. For this purpose, the number of forward citations of backward cited patents (i.e. the number of citations of citing patents) is used as an input variable. As a result, it is proved that the quality of knowledge accumulation plays a key role in predicting promising technologies.

This study contributes to the field in three ways. First, this study enables firms to predict emerging technologies using patent-based machine learning techniques, making long-term systematic technological planning possible. Second, knowledge accumulation quality is used as an important input for predicting emerging technologies, and it is confirmed that this is a very important variable. The number of backward citations has previously been employed to predict patent quality in some studies, but its significance was considered to be low compared to the number of forward citations, which is a direct measure of technological strength. However, the current study proves that the quality of knowledge accumulation, i.e. the types of patents employed and their technological strength, should be considered very important in the prediction of emerging technologies. Finally, the technique described in this study can replace expert-based evaluation, which is typically a time-consuming and expensive task. This study employs a data-driven approach in which the input and output variables are all drawn from patents, meaning that the entire process can be easily automated as a form of computer-based evaluation.

Despite these contributions, however, this study has several limitations. First, this study uses the number of forward citations of a patent as an output variable. Even if the number of patent citations is an important proxy for technological strength, it cannot reflect commercial factors, i.e. the ways in which this technology can be effectively utilised in the marketplace. Therefore, alternative measures such as renewal fee or patent transactions can be excellent alternatives to measure patent value. These limitations can be addressed in future research. Future work may therefore incorporate market perspectives in the prediction of emerging technologies. Second, this study has a data imbalance problem, in that emerging technologies are very rare compared to ordinary technologies in the training dataset. Future work should therefore develop appropriate sampling techniques and use a weighted machine learning approach. Third, information can be lost when considering the quality of accumulated knowledge, since this study uses only the average and maximum numbers of (forward) citations of backward cited patents. Thus, more information should be considered

in future work, such as the distribution of patent citations during certain periods. Fourth, the case study is conducted for only one class, and extension and generalisation to other classes is left for future work.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIP) (NRF-2017R1E1A1A01077324). This work is based on the thesis submitted by Uijun Kwon for a master's degree at Seoul National University of Science and Technology (SeoulTech), Seoul, Korea, 2018

References

Ahuja, G., & Morris Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal, 22*(6–7), 521–543.

Azagra-Caro, J. M., Barberá-Tomás, D., Edwards-Schachter, M., & Tur, E. M. (2017). Dynamic interactions between university-industry knowledge transfer channels: A case study of the most highly cited academic patent. *Research Policy*, *46*(2), 463–474.

Banerjee, P. M., & Cole, B. M. (2011). Globally radical technologies and locally radical technologies: The role of audiences in the construction of innovative impact in biotechnology. *IEEE Transactions on Engineering Management, 58*(2), 262–274.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, *37*(5), 932–945.

Brem, A., & Voigt, K. (2009). Integration of market pull and technology push in the corporate front end and innovation management—Insights from the German software industry. *Technovation, 29*(5), 351–367.

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.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics*, *69*(1), 3–20.

Cammarano, A., Michelino, F., Lamberti, E., & Caputo, M. (2017). Accumulated stock of knowledge and current search practices: The impact on patent quality. *Technological Forecasting and Social Change*, *120*, 204–222.

Chau, P. Y., & Tam, K. Y. (2000). Organizational adoption of open systems: A 'technology-push, need-pull' perspective. *Information & Management, 37*(5), 229–239.

Choi, J., Jang, D., Jun, S., & Park, S. (2015). A predictive model of technology transfer using patent analysis. *Sustainability, 7*(12), 16175–16195.

Chu, Y. T., & Su, H. N. (2015). Understanding inter-assignee dynamics of technological development. In *2015 Portland international conference on management of engineering and technology (PICMET)* (pp. 783–792). IEEE.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, 73*(8), 981–1012.

De Marco, A., Scellato, G., Ughetto, E., & Caviggioli, F. (2017). Global markets for technology: Evidence from patent transactions. *Research Policy, 46*(9), 1644–1654.

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information, 25*(3), 233–242.

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly, 52*(3), 443–475.

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy, 30*(7), 1019–1039.

Geum, Y., Jeon, J., & Seol, H. (2013). Identifying technological opportunities using the novelty detection technique: A case of laser technology in semiconductor manufacturing. *Technology Analysis & Strategic Management*, *25*(1), 1–22.

Geum, Y., Kim, M., & Lee, S. (2017). Service technology: Definition and characteristics based on a patent database. *Service Science, 9*(2), 147–166.

Graham, S. J., Marco, A. C., & Myers, A. F. (2018). Patent transactions in the marketplace: Lessons from the USPTO patent assignment dataset. *Journal of Economics & Management Strategy, 27*(3), 343–371.

Gruber, M., Harhoff, D., & Hoisl, K. (2013). Knowledge recombination across technological boundaries: Scientists vs. engineers. *Management Science*, *59*(4), 837–851.

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363.

Huang, M. H., Yang, H. W., & Chen, D. Z. (2015). Increasing science and technology linkage in fuel cells: A cross citation analysis of papers and patents. *Journal of Informetrics, 9*(2), 237–249.

Huang, H. C., Su, H. N., & Shih, H. Y. (2018). Analyzing patent transactions with patent-based measures. *In 2018 Portland international conference on management of engineering and technology (PICMET)* (pp. 1–12). IEEE.

Joung, J., & Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological Forecasting and Social Change, 114*, 281–292.

Ju, Y., & Sohn, S. Y. (2015). Patent-based QFD framework development for identification of emerging technologies and related business models: A case of robot technology in Korea. *Technological Forecasting and Social Change, 94*, 44–64.

Lanjouw, J. O., Pakes, A., & Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. *The Journal of Industrial Economics, 46*(4), 405–432.

Lanjouw, J. O., & Schankerman, M. (1999). The quality of ideas: Measuring innovation with multiple indicators (No. w7345). *National Bureau of Economic Research*.

Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal, 114*(495), 441–465.

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., 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.

Lerner, J. (1994). The importance of patent scope: An empirical analysis. *RAND Journal of Economics, 25*(2), 319–333.

Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2018). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting & Social Change*. https://doi.org/10.1016/j.techfore.2018.06.004.

Karki, M. M. S. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269–272.

Kayal, A. A., & Waters, R. C. (1999). An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. *IEEE Transactions on Engineering Management, 46*(2), 127–131.

Kim, C., & Seol, H. (2012). On a patent analysis method for identifying core technologies. *Intelligent decision technologies* (pp. 441–448). Berlin, Heidelberg: Springer.

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change, 125*, 236–244.

Meyer, M. (2000). Does science push technology? *Patents Citing Scientific Literature. Research policy*, *29*(3), 409–434.

Narin, F. (1999). *Tech-line background paper*. Haddon Heights, NJ: CHI Research.

Noh, H., Song, Y. K., & Lee, S. (2016). Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. *Telecommunications Policy, 40*(10–11), 956–970.

Park, Y., Yoon, B., & Le, S. (2007). A organizational dynamics-the idiosyncrasy and dynamism of technological innovation across industries: Patent citation analysis Af: 160. *Operations Research Management Science*, *47*(1), 25.

Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications, 6*(2), 256–261.

Porter, A. L., Roper, A. T., Mason, T. W., Rossini, F. A., & Banks, J. (1991). *Forecasting and management of technology* (Vol. 18). Hoboken: Wiley.

Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: Boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal, 22*(4), 287–306.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827–1843.

Sheremetyeva, S. (2003). Natural language analysis of patent claims. In *Proceedings of the ACL-2003 workshop on patent corpus processing* (Vol. 20, pp. 66–73). Association for Computational Linguistics.

Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change, 115*, 131–142.

Shane, S. (2001). Technological opportunities and new firm creation. *Management Science, 47*(2), 205–220.

Song, K., Kim, K., & Lee, S. (2017). Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents. *Technological Forecasting and Social Change.*

Schankerman, M. (1991). How valuable is patent protection? Estimates by technology field using patent renewal data (No. w3780). *National Bureau of Economic Research*.

Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy, 39*(8), 1051–1059.

Teichert, T., & Mittermayer, M. A. (2002). *Text mining for technology monitoring* (pp. 596–601). Cambridge, UK: Proceedings of IEEE international engineering and management conference.

Thomas, P. (1999). The effect of technological impact upon patent renewal decisions. *Technology Analysis & Strategic Management, 11*(2), 181–197.

Tijssen, R., Visser, M., & Van Leeuwen, T. (2002). Benchmarking international scientific excellence: Are highly citedresearch papers an appropriate frame of reference? *Scientometrics*, *54*(3), 381–397.

Tong, X., & Frame, J. D. (1994). Measuring national technological performance with patent claims data. *Research Policy, 23*(2), 133–141.

Trappey, A. J., Trappey, C. V., Wu, C. Y., & Lin, C. W. (2012). A patent quality analysis for innovative technology and product development. *Advanced Engineering Informatics, 26*(1), 26–34.

Uriona-Maldonado, M., de Souza, L. L. C., & Varvakis, G. (2010). Focus on practice service process innovation in the Brazilian electric energy sector. *Service Business, 4*(1), 77–88.

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy, 45*(3), 707–723.

Wu, J. L., Chang, P. C., Tsao, C. C., & Fan, C. Y. (2016). A patent quality analysis and classification system using self-organizing maps with support vector machine. *Applied Soft Computing, 41*, 305–316.

Wang, X., & Duan, Y. (2011). Identifying core technology structure of electric vehicle industry through patent co-citation information. *Energy Procedia, 5*, 2581–2585.

You, H., Li, M., Hipel, K. W., Jiang, J., Ge, B., & Duan, H. (2017). Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics, 111*(1), 297–315.

Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. *Expert Systems with Applications, 35*(1–2), 124–135.