

An entropy-based indicator system for measuring the potential of patents in technological innovation: Rejecting moderation

Yi Zhang¹, Yue Qian², Ying Huang², Ying Guo^{2*}, Guangquan Zhang¹, Jie Lu¹

¹Decision Systems & E-Service Intelligence Research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

²School of Management and Economics, Beijing Institute of Technology, Beijing, P. R. China

Email addresses: yizhang.uts@gmail.com; qianyuedc@163.com; huangying_work@126.com; guoying_bit@163.com (*); guangquan.zhang@uts.edu.au; jie.lu@uts.edu.au.

Abstract: How to evaluate the value of a patent in technological innovation quantitatively and systematically challenges bibliometrics. Traditional indicator systems and weighting approaches mostly lead to “moderation” results; that is, patents ranked to a top list can have only good-looking values on all indicators rather than distinctive performances in certain individual indicators. Orienting patents authorized by the United States Patent and Trademark Office (USPTO), this paper constructs an entropy-based indicator system to measure their potential in technological innovation. Shannon’s entropy is introduced to quantitatively weight indicators and a collaborative filtering technique is used to iteratively remove negative patents. What remains is a small set of positive patents with potential in technological innovation as the output. A case study with 28,509 USPTO-authorized patents with Chinese assignees, covering the period from 1976 to 2014, demonstrates the feasibility and reliability of this method.

Keywords Patent analysis; Indicator system; Bibliometrics; Technological innovation; Entropy.

Introduction

Following Schumpeter’s observations on Business Cycles (Schumpeter 1939), an invention is considered as a process of recombination (Fleming 2001), and theoretical and systematic explanations of technological innovation have become a crucial scholarly topic for innovation management. The suitability of patents for indicating technological innovation has been discussed considerably since the 1990s and even before (Basberg 1987). A number of patentometric indicators have been applied to measure technological innovation from diverse econometric perspectives. Based on statistics and empirical studies, determinants of patent value regarding economic potential were observed, in which both quantitative indicators [e.g., backward patent citations, non-patent citations, the number of inventors, and the number of co-assignees (Sapsalis et al. 2006)] and qualitative ones [e.g., technical importance, inventing difficulty, and learning value for competitors (Reitzig 2003)] are involved.

How to evaluate the value of a patent quantitatively and systematically also challenges bibliometrics. As a pioneering study, Pavitt (1985) pursued the argument of de Solla Price (1983) on the practical needs to explain new empirical data provided by measurement systems, and foresaw positively on using patent statistics in analyzing technological activities for policy making. Patent indicators were then widely introduced to measure patent value, which could constitute technological value, or direct and indirect economic value (Lee 2009). Such indicators involve not only patent statistics but also legal status information sometimes. For a wide range of science, technology, innovation, and policy (STIP) studies, these indicators are selected to evaluate a corpus of patents that represents a given technological area or entities such as country, organization, and individual (Narin and Hamilton 1996; Meyer and Tang 2007; Zhang et al. 2014b).

Indicator systems are not unfamiliar for econometrics, which usually apply regression-based statistical models to measure the relationships between economic outcomes and bibliometric indicators, but a bibliometric indicator system to automatically identify meaningful patents and patent portfolios remains elusive. On one hand, blending patent citation/co-citation analysis and social network analysis to seek patents at traffic hubs is one mainstream approach to identify “key” patents (Choi and Park 2009; Funk and Owen-Smith 2016), despite the fact that certain limitations of citation analysis have already been argued by Rip (1988). On the other hand, the engagement of multiple indicators also introduces issues (e.g., how to weight those indicators). Delphi-based or Analytic Hierarchy Process (AHP)-based qualitative approaches can be helpful in some sense (Bozbura et al. 2007). However, even if we ignore the bias possibly resulting from subjective opinions of experts, these traditional weighting approaches could mostly lead to “moderation” results; that is, patents ranked to a top list can well be neither those with the highest forward citations nor those with the most active cross-national collaborations, but they will have good-looking values on all indicators. In Chinese philosophy, such a phenomenon is summarized as the Doctrine of the Mean, but it is definitely not good for indicating innovation potential.

Aiming to address the above concerns, orienting patents authorized by the United States Patent and Trademark Office (USPTO), this paper constructs an entropy-based indicator system to measure the potential of patents in technological innovation. One basic target is to identify significant patents with high technological innovation rather than those multi-dimensional moderate ones. Our endeavours include (1) proposing an indicator model for USPTO patents, which contains three macro-level perspectives: technological perspective, legal perspective, and economic perspective. Each perspective is constituted by a number of patent indicators. Descriptive statistics and correlations are used to identify dynamic indicators and high-coupled indicators, which help configure the priorities of indicators in a ranking system. (2) Shannon’s entropy (Shannon 1948), well-known as a coefficient for measuring complexity and uncertainty, is introduced to quantitatively weight indicators. Its basic weighting criterion is that the more common an indicator is the less weight it would have. In other words, patents with irregular indicator values would be ranked higher. (3) Based on the ranking performed by entropy-based weights, a group of patents with negative innovation potential is first identified. We apply a collaborative filtering technique to measure similarities between all remaining patents and the set of negative patents, and patents sharing a high similarity value with negative patents are identified as noise and will be set as negative patents for next iteration. The output of our method is a small set of ranked patents, which can indicate the potential of a patent’s technological innovation from diverse dimensions.

This method can be used to seek patents with technological values and innovative potential. We applied our method to 28,509 USPTO patents with Chinese assignees, covering the period from 1976 to 2014. A number of patents with distinctive potential in technological innovation were identified, and the results demonstrate the feasibility and reliability of our method, which holds abilities to explore insights to support Research & Development (R&D) plans and strategic management in a wide range of government and industry sectors.

This paper is organized as follows. We review previous studies in the Related Works section, which include patent analysis and indicators in economics, patent indicators in bibliometrics, and patent ranking systems. The Methodology section follows and presents our method using an entropy-based indicator system for measuring the potential of USPTO patents in technological innovation. The Empirical Study section applies the method to identify distinctive patents with the potential of technological innovation from a corpus of USPTO patents. Finally, we provide an in-depth discussion on the technical implication of the method, limitations, and future directions in the Discussion and Conclusions section.

Related Works

We review related works from the following three aspects: patent analysis and indicators in economics, patent indicators in bibliometrics, and patent ranking system.

Patent analysis and indicators in econometrics

Patent statistics, serving as a crucial indicator of innovation, have been used to measure technological change since the 1980s (Basberg 1987; Archibugi and Planta 1996; Fleming 2001), and regarding to the life cycle of a technology patents focus on the development stage that links academic research with actual applications (Martino 2003). Credit to Hall (2002), the development of the U.S. Patent Citations Data File further pushed these efforts forward. Recently, indicator systems with patent statistics are widely constructed to characterize technological innovation from the perspective of economics, e.g., Grimaldi et al. (2015) focused on the strategic information of patents in analyzing the value of patent portfolios, and Verhoeven et al. (2016) integrated patent citations and classifications to evaluate patents from the perspectives of recombination and knowledge origins. In addition, as a hotspot in economic studies, researchers have deeply conducted the interactions among patents, market value, and R&D via both theoretical and empirical studies (Hall et al. 2005).

Patent indicators in bibliometrics

Using patent indicators for bibliometrics can also date back to the 1980s (de Solla Price 1983; Pavitt 1985). Patent statistics, acting as one of the most significant elements in patent analysis, demonstrate incredible power in a wide range of STIP studies, e.g., analyzing technological landscapes (Chen et al. 2005), identifying the technological relationships between scientific and technological communities (Guan and He 2007), and measuring technological or innovative capabilities (Narin and Hamilton 1996; Meyer and Tang 2007). These studies usually addressed concerns on given entities, such as organization, region, and country. Furthermore, the use of patent indicators in bibliometrics, gaining benefits from text mining and network analysis, exceeds that of econometrics, e.g., text elements (including single words, terms, and subject-action-object structures) are involved in patent analysis (Choi et al. 2011; Yoon et al. 2013; Zhang et al. 2014b), and the development of patent maps, based on citations and International Patent Classification (IPC) codes, provides a novel way to measure technological similarity for multidisciplinary studies (Kay et al. 2014; Leydesdorff et al. 2014). In particular, citation statistics (e.g., the number of citations, co-citations, and citation rate) are highlighted, which have become the most important and the widest-used indicator of evaluating scientists (Hirsch 2005; Cronin and Meho 2006; Egghe 2006) and journals (Braun et al. 2006; Vinkler 2013).

Patent ranking systems

Bibliometrics usually closely relate to actual requirements in STIP studies, and indicator systems have been extended to a broad range, e.g., evaluating scientific and technological activities (Lee 2009), and profiling leading individuals or organizations (Waltman et al. 2012). Ranking has become an increasing application of indicator systems, in which the targets of the systems include not only patents but also a number of scientific publications, journals, websites, and topics, and both quantitative and qualitative methodologies are introduced, e.g., network analysis, time series analysis, fuzzy decision-making approaches, Analytical Hierarchy Process (AHP), and questionnaire survey (Glänzel and Thijs 2012; Iwami et al. 2014; Xu et al. 2014; Wang and Hsieh 2015). Concentrating on ranking needs, previous studies might either heavily depend on individual indicators (that might exaggerate their influence) or easily get into trouble in engaging multiple indicators, but both led to “moderation” results, with good-looking values on all indicators.

Methodology

Orienting USPTO patents, this paper constructs an entropy-based indicator system, including three main models: (1) a patent indicator model, involving a number of patent indicators and containing three macro-level perspectives: technological perspective, economic perspective, and legal perspective; (2) an entropy-based weighting model, to quantitatively define weights for the indicators and initialize a set of negative patents, that is, those ranked at the end of the queue; and (3) a collaborative filtering model, to remove patents sharing a high similarity with the negative patents, in which an iterative collaborative filtering approach is engaged. The framework of the entropy-based indicator system is given in Figure 1.

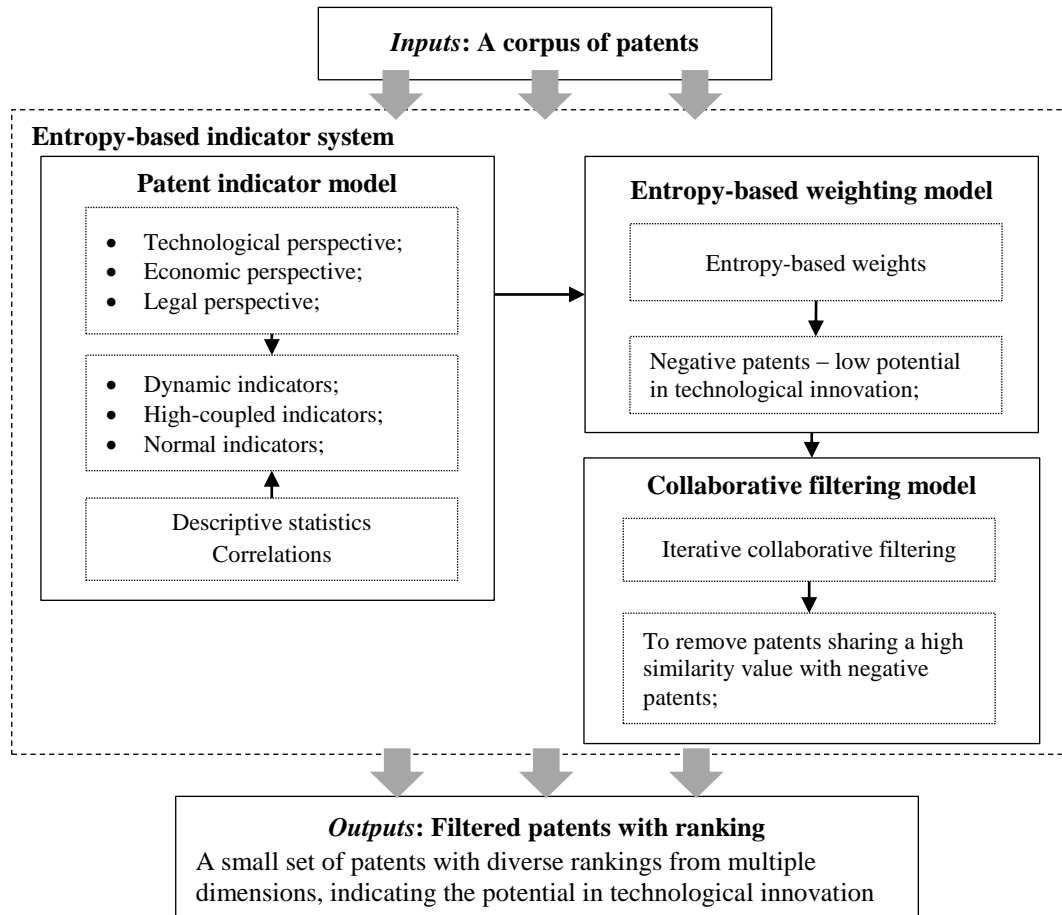


Fig. 1. Framework of the entropy-based indicator system

Patent indicator model

With reference to the given features of USPTO and patent indicators used in econometrics and bibliometrics, we in particular selected eleven indicators for evaluating the potential of a patent's technological innovation (e.g., the number of inventors, the number of patent families, the number of legal transactions, the number of claims, the number of patent references, the number of non-patent references, the number of citations, the number of IPCs, the number of terms, and the time gap between the application year and the issue year, and the number of assignees). Despite the fact that technological perspective is the one we emphasize, the interactions among technological, economic, and legal issues cannot be ignored. At this stage, three macro-level perspectives are highlighted (e.g., technological

perspective, economic perspective, and legal perspective). The classification of the eleven indicators to the three perspectives is given in Table 1.

Table 1. Classification of the eleven patent indicators to technological, economic, and legal perspectives

	<i>Indicator</i>	<i>Tech.</i>	<i>Eco.</i>	<i>Legal</i>
1	# inventors		√	
2	# patent families		√	√
3	# legal transactions		√	√
4	# claims	√		
5	# patent references	√	√	
6	# non-patent references	√		
7	# citations	√		
8	# IPCs	√		
9	# terms	√		
10	Time gap*	√		
11	# assignees		√	

Note. #: the number of; Tech.: technological perspective; Eco.: economic perspective; Legal: legal perspective; Time gap: the time gap between the application year and issue year.

(1) Inventors and assignees

The information of patent inventors and assignees plays a crucial role in exploring knowledge/technology transfer, and their engagement belongs to a company's R&D strategy (Agrawal 2006). Comparably, the information of inventors emphasizes the R&D capability and inner collaboration of a company, while multiple assignees of a patent can indicate related technological collaboration patterns home and abroad (Guellec and de la Potterie 2001; Lei et al. 2013). At this stage, we used the number of inventors and the number of assignees as two economic indicators, following the hypothesis: *the more inventors/assignees a patent has the higher economic value it might have.*

(2) Patent family

Inventors will seek legal protection for their invention, and such protection can be authorized from diverse patenting authorities. The hypothesis that *the more economic and legal value an invention has to the inventor, the more broadly the invention will be fielded* has been approved, i.e., patents with a large number of patent families can be particularly valuable (Harhoff et al. 2003; Wang 2007). It is common to use patent family data to analyze the internationalization of a technology, forecast its application, and estimate patent value (Martínez 2011). Despite the fact that USPTO does not provide the information of patent families, one feasible solution is to collect such information from the Derwent World Patents Index (DWPI). Thus, we still added the number of patent families as one indicator covering both economic and legal perspectives in our indicator system.

(3) Legal transactions

Previous studies attempting to identify the determinants of patent value emphasized the information of renewal, i.e., the applicant of a patent will pay the maintenance fee to renew the protection when the term expires (Guellec and de la Potterie 2000). As an example, patents applied from USPTO after December 11, 1980, will require a maintenance fee after 3.5, 7.5, and 11.5 years to remain in force beyond 4, 8, and 12 years, respectively (Bessen

2008). In addition, it is also fruitful to delve into the transfer of patent rights, since patents have become one of the most crucial resources in technology mergers and acquisitions (Makri et al. 2010). This paper counted all changes of the legal status of a patent and defined it as the most important indicator from the legal perspective. The hypothesis here is *the more legal transactions a patent has the higher economic and legal value it might have*. Note that legal transaction data is not available for all patent data sources, so it is necessary to consider possible influence while applying this system to other patent data sources rather than USPTO.

(4) Patent claims

A patent claim describes the legal protection area of a patent, which usually contains a number of novel technological features (Schmoch 1993), and analyzing patent claims is considered as a good way to explore technological performance (Lee et al. 2007). However, from the professional view of patent examiners, one critical concern here is a patent with too long claims can be unclear and will have to be rewritten, so patent claims might have limited relationships with technological innovation but can indicate legal meaning. At this stage, we introduced the number of claims to be an indicator of legal perspective. Thus, the hypothesis is *the more claims a patent has the higher legal value it might have*.

(5) Patent citations

Patent citations are one of the most significant indicators for measuring the value of a patent, which indicate innovation from two aspects: 1) the interactions among inventions, inventors, and assignees; and 2) the importance of individual patents (Hall et al. 2005). It is common to use backward citation and forward citation to distinguish two types of patent citations. The former one relates to references that a patent cites, while the latter one is used to describe how many times a patent was cited by others (Von Wartburg et al. 2005). This paper used “the number of patent references” and “the number of non-patent references” to represent backward citations, and “the number of citations” to reveal forward citations.

We further distinguish backward citations due to the following reason: citing previous patents can be a way to align with certain technological flows and obtain more opportunities to earn economic value, while citing non-patent references (e.g., scientific publications) is to illustrate its close relation with the frontiers of knowledge and the potential in technological innovation (Funk and Owen-Smith 2016). In addition, non-patent references can also be considered as a breakthrough point for investigating the science-technology linkages (Tijssen 2001). However, it is necessary to consider the role of patent examiners in evaluating backward citations, e.g., Alcacer and Gittelman (2006) and Azagra-Caro et al. (2009) argued that diverse influences exist between applicant-inset and examiner-inset backward citations when tracing knowledge flows and science-technology links. At this stage, despite the fact that the use of the number of backward citations can be counterevidence, we list the two indicators in our indicator system, in which we assume the number of patent references as an indicator emphasizing the economic value of a patent and the number of non-patent references as an indicator to highlight a patent’s technological value. They both follow the hypothesis that *the more patent/non-patent references a patent has the higher economic or technological value it might have respectively*. However, we are fully aware of and highlight the arguments on these two indicators, and it is required to discuss the practicability before applying them to any actual cases.

The importance of patent citations (forward citations) has been discussed from multiple dimensions, and the evidence from previous studies strongly supports the conclusion that patent citations can be a crucial indicator to measure a patent’s technological importance (Gittelman and Kogut 2003; Harhoff et al. 2003; Hall et al. 2005). We in particular selected the number of citations as one of the most important indicators for technological value, and

the hypothesis is that *the more citations a patent receives the higher technological value it might have*.

(6) IPCs and terms

IPC provides a hierarchical taxonomy system reflecting existing technological categories and sub-categories (Zhang et al. 2016), and it has been widely used as an indicator to measure the technological scope of a patent (Reitzig 2004). In addition, IPCs also play an active role in measuring technological distance between patents or patent portfolios (Zhang et al. 2016). We used the number of IPCs as a technological indicator, with the hypothesis that *the more IPCs a patent has the higher technological value it might have*.

It is not common to use terms in patent indicator systems because of semantic complexity and insufficient term cleaning techniques. In this paper, we introduced a term clumping process (Zhang et al. 2014a) to remove noise and consolidate technological synonyms, and identified core technological terms to reflect meaningful technological information contained in patent documents. At this stage, the number of terms is involved in the system to indicate the technological value of a patent, and the hypothesis is that *the more core technological terms a patent has the higher technological value it might have*.

(7) Time gap

One observation from the empirical work of Mowery et al. (2002) is that patent citations peak during the first four years after the issue date of a patent. Therefore, one consideration here is whether a patent application that is issued rapidly can be better than the one issued slowly. At this stage, we used the time gap between the issue year and the application year of a patent to indicate its technological value. Our hypothesis is that *the lower the time gap a patent has the higher technological value it might have*.

Based on the entire dataset, descriptive statistics and correlations are applied, in which we will emphasize the following parts of the analytic results: (1) standard deviation—certain indicators with a high value of standard deviation will be set as “dynamic indicators,” which means the value of this indicator in the dataset is extremely unstable and this indicator might be used to seek special patents; (2) the value of correlation—certain pairs of indicators with a high value of correlation will be set as a “high-coupled pair of indicators,” which means the indicators in the pair are highly correlated and only one indicator of the pair might have a relatively high weight.

Entropy-based weighting model

Expert knowledge is indispensable for weighting indicators in most indicator systems, in which qualitative approaches (e.g., Delphi and AHP) are broadly used to engage experts. Despite the fact that a number of efforts (e.g., fuzzy set and multiple criteria decision making) are involved to reduce subjective bias and improve the performance of related indicator systems (Bozbura et al. 2007; Wang and Hsieh 2015), traditional weighting approaches still face one critical issue: highly ranked patents can only be those have good-looking values on all indicators rather than those with distinctive values in certain individual indicators.

Entropy, well-known as a coefficient for measuring complexity and uncertainty, was first introduced from thermodynamics to information theory by Shannon (1948). Concentrating on studies in bibliometrics, certain interesting applications exist, e.g., Leydesdorff (2002) built up an entropy-based indicator to measure the “heat” in the dynamics of science, and Chen and Chang (2012) applied entropy to investigate the influence of technological diversification on technological competition. The basic weighting criterion of entropy is that *the more common an indicator is the less weight it would have*. Therefore, an entropy-based weighting

model varies with actual datasets and highlights the dynamics of related indicators, which can be a way to explore special patents with potential in technological innovation.

We followed the definition given by Grupp (1990) and denoted a set of patents as M and a set of indicators as N , and thus a matrix $X = M^T N$ exists, where $x_{i,j}$ is used to represent the value of the j -th indicator of the i -th patent ($i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$). The stepwise process of the entropy-based weighting model is described as follows:

Step 1: to normalize $x_{i,j}$ as $a_{i,j}$, in which a max-min normalization approach is used (in particular since the time gap prefers a fewer value rather than a larger value as other indicators, for the value of the time gap, $a'_{i,j} = 1 - a_{i,j}$):

$$a_{i,j} = \frac{x_{i,j} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

where $\max(x_j)$ and $\min(x_j)$ is the maximum and minimum value of the j -th indicator.

Step 2: to calculate an entropy H_j :

$$H_j = -\frac{1}{\ln m} \sum_{i=1}^m a_{i,j} \ln a_{i,j}$$

Step 3: to transfer the entropy H_j to the weight w_j of related indicator:

$$w_j = \frac{1 - H_j}{n - \sum_{j=1}^n H_j} \quad (0 \leq w_j \leq 1, \sum_{j=1}^n w_j = 1)$$

Step 4: to rank patents based on the weights calculated automatically, and set the patents ranked at the end of the queue as negative patent (i.e., those with low potential in technological innovation). A threshold δ with certain strategies based on actual data will be used to decide the selection of negative patents.

Compared to traditional weighting approaches, the entropy-based weighting model fully takes the situation of actual data into consideration and automatically calculates the weights of all indicators. More importantly, the main purpose of this model is not to identify significant patents, as what traditional approaches do, but to identify negative patents. Our concern here is that patents with good-looking values in all indicators could not be those with high potential in technological innovation, but patents with low values in all indicators must be those with limited contribution to technological innovation. Therefore, similar to certain search strategies, our design here is to apply the entropy-based weighting model to help narrow down our targets, and provide related resources (e.g., a set of negative patents) for an iterative process of further filtering in the collaborative filtering model.

Collaborative filtering model

Collaborative filtering is a common technique for recommender systems, which recommends items based on shared interests between users or shared features between items (Lu et al. 2015). Collaborative filtering techniques have been widely used for social network analysis (Mao et al. 2016), e-business applications (Shambour and Lu 2012), and big data analysis (Jiang et al. 2011). We followed the basic rules of item-based collaborative filtering techniques and constructed the collaborative filtering model to filter patents sharing a high similarity with negative patents.

Following the definition given in the entropy-based weighting model, after normalization a patent can be represented as M_i , where $a_{i,j}$ is the value of the j -th indicator of M_i , and the

set of entropy-based weights is W , where w_j is the weight of the j -th indicator. The collaborative filtering model is described as follows:

Step 1: to reconstruct M as M^W , with the engagement of W :

$$M^W = M \circ W$$

where $a_{i,j}^W$ is the value of the j -th indicator of M_i^W and $a_{i,j}^W = a_{i,j} \times w_j$.

Step 2: to divide M^W into two sets—the set of negative patents M^N and the set of remaining patents M^R :

$$\begin{aligned} M^N \cup M^R &= M^W \\ M^N \cap M^R &= \emptyset \end{aligned}$$

Step 3: to measure the similarity $S(M_{iR}, M_{iN})$ between each patent M_{iR} of M^R and each negative patent M_{iN} in M^N , in which the traditional cosine measure (Salton and Buckley 1988) is used:

$$S(M_{iR}, M_{iN}) = \cos(M_{iR}, M_{iN}) = \frac{M_{iR} \cdot M_{iN}}{|M_{iR}| |M_{iN}|}$$

$$|M_{iR}| = \sqrt{\sum_{j=1}^n (a_{iR,j}^W)^2}, |M_{iN}| = \sqrt{\sum_{j=1}^n (a_{iN,j}^W)^2}$$

where $|M_{iR}|$ and $|M_{iN}|$ is the norm of the vector M_{iN} and M_{iR} respectively, $a_{iR,j}^W$ and $a_{iN,j}^W$ is the value of the j -th indicator of M_{iN} and M_{iR} , respectively, and n is the total number of applied indicators.

Step 4: for each remaining patent M_{iR} , to set the maximum value of $S(M_{iR}, M_{iN})$ as the similarity $S(M_{iR}, M^N)$ between M_{iR} and the set of negative patents M^N :

$$S(M_{iR}, M^N) = \max[S(M_{iR}, M_{iN})]$$

Step 5: to compare $S(M_{iR}, M^N)$ with a threshold τ , if $S(M_{iR}, M^N) \geq \tau$, the patent M_{iR} is marked as “negative;”

Step 6: to remove all the patents in M^N , and set the patents with the label of negative as new negative patents and move to M^N .

Step 7: to end the iteration if the total number of remaining patents is less than a threshold θ or there is no new patent marked the label of negative. Or else, return to Step 3:

$$\text{Num}(M^R) \leq \theta \text{ or } \text{Num}(M^N) = 0$$

where $\text{Num}(M^R)$ is the number of the patents in M^R .

The output of the collaborative filtering model is a small set of patents, with potential in technological innovation. However, differing from traditional approaches, we did not rank these patents but provided certain dimensions to select distinctive patents (e.g., good citations & frequent legal transactions, large-scale engagement of inventors & assignees, and a large number of involved IPCs & terms). At this stage, certain benefits can be gained from such design: (1) to emphasize the distinctive outcomes of a patent in certain individual indicators; (2) to explore interactions between/among multiple indicators; and (3) to provide a way to evaluate patents at a macro-level and leave space to engage expert knowledge more effectively than a traditional large-scale manual patent indexing process.

Empirical Study

The design of the entropy-based indicator system orients USPTO patents, however, we downloaded the USPTO patents of our empirical dataset from the DWPI database¹ by the search strategy “Database = US Grant (the database of patents issued by USPTO) AND PAOD (the address of patent assignees) = CN AND PY (publication year) >=1976 AND PY <= 2014.” Our consideration includes: (1) we emphasize the empirical patents are from the same jurisdiction, i.e., the same priority patent authority, the same examination systems, and the same language, so USPTO is our basic focus; (2) we prefer a broad sample of patents with multiple technological domains to show the ability for our system to generate novel insights, so we collect all patents with Chinese assignees; (3) the global view of DWPI provides additional information on patent family and citation, which can be necessary indicators in our design, and the rewrite of DWPI would help reduce the number of technological synonyms and further benefit the identification of core terms; and (4) DWPI is integrated with Web of Science, and our previous work in data pre-processing (including the use of VantagePoint) matches perfectly here. Therefore, we finally decided to collect patents from DWPI rather than the website of USPTO. In addition, we invited one patent examiner from the Intellectual Property Office of China (SIPO) and one researcher from the Beijing Institute of Technology, who has focused on patent analysis for nearly ten years, as our experts to provide professional consultation for our study.

Data and patent indicators

A raw set with 33,585 patents was first retrieved, but considering some patents without a title or an abstract, we only collected 28,509 patents with the both fields to run a term clumping process (Zhang et al. 2014a) to remove noise (e.g., conjunctions, prepositions, and pronouns) and common terms in patents (e.g., “description,” “use,” and “drawings”), and consolidate technological synonyms based on the stem (e.g., singular and plural, and the part of speech). The stepwise results of the term clumping process are given in Table 2.

Table 2. Stepwise results of the term clumping process

<i>Step</i>	<i>Description</i>	<i># Terms</i>	<i>% Reduce</i>
1	Raw terms after natural language processing	493,856	N/A
2	Basic cleaning—to remove noise	458,234	7.2%
3	Basic cleaning—to remove common terms in patents	445,745	2.7%
4	Stem-based consolidation	405,822	9.0%
5	Pruning—to remove terms appearing in only one patent	70,034	82.7%

Note: #Terms: the number of terms; % Reduce: the proportion of reduction.

As shown in Table 2, we reduced the scale of the term amount from 493,856 to 70,034 by using certain thesauri and association rules. Note that the target of the term clumping process is to handle terms rather than individual words, since the limitation of current natural language processing techniques is that some adjectives might be combined with existing terms and produce new synonyms. At this stage, Step 4 in some sense was designed for such an issue. Although there might be a number of low frequency terms containing valuable innovative information, it is also reasonable to imagine that a term will be meaningless if it only appeared once in several decades.

Eleven patent indicators were selected: the time gap between the application year and the issue year of a patent, the number of inventors, patent families, legal transactions, claims,

¹ <http://thomsonreuters.com/en/products-services/intellectual-property/patent-research-and-analysis/derwent-world-patents-index.html>

patent references, non-patent references, citations, IPCs, terms, and assignees. We constructed a 28509×11 matrix linking individual patents and the eleven indicators. Descriptive statistics and entropy-based weights are given in Table 3, and Table 4 reports correlations.

Table 3. Descriptive statistics and entropy-based weights

<i>Indicator</i>	<i>Min.</i>	<i>Max.</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean (*)</i>	<i>Std. Dev. (*)</i>	<i>Weight</i>
1 # inventors	1	31	3.0	2.1	.066	.069	0.03
2 # patent families	1	132	49.5	56.5	.371	.431	0.06
3 # legal transactions	0	8	1.6	1.1	.196	.133	0.23
4 # claims	0	145	13.8	7.7	.095	.053	0.17
5 # patent references	0	792	14.2	24.4	.018	.031	0
6 # non-patent references	0	886	4.2	16.1	.005	.018	0
7 # citations	0	695	5.7	12.1	.008	.017	0.13
8 # IPCs	0	28	1.5	.9	.053	.032	0.12
9 # terms	0	304	13.0	11.2	.043	.037	0.10
10 Time gap	0	11	2.1	1.3	.810	.122	0.13
11 # assignees	1	28	1.4	.9	.015	.033	0.03

Note: Mean (*) and Std. Dev. (*) are based on the results after normalization.

Table 4. Correlations

<i>Indicator</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
2	.088	1								
3	-.040	-.215	1							
4	.062	.015	.020	1						
5	.072	.105	-.049	.127	1					
6	.110	.175	-.039	.098	.472	1				
7	.086	.102	.149	.135	.248	.179	1			
8	.079	.123	.058	.018	.046	.117	.117	1		
9	.205	.013	-.013	.117	.105	.229	.084	.127	1	
10	.057	.291	-.102	.043	.099	.115	.167	.126	-.008	1
11	.101	-.143	.010	-.010	.010	.018	.027	.043	.114	-.106

Note: The indicators follow the numbers given in Table 3.

Certain findings observed from Tables 3 and 4 include: (1) based on the normalized results, the number of patent families peaks the largest standard deviation by a significant margin, and the number of legal transactions and the time gap follow. Thus, we set the three indicators as *dynamic indicators*, which indicate that their related values vary significantly with different patents; (2) the correlation between the number of patent references and the number of non-patent references is very high, and they both correlate with the number of citations in a relatively high level. At this stage, considering the important of forward citations and the argument on backward citations, we set the number of patent references and the number of non-patent references as *high-coupled indicators*, and remove them from the indicator system; and (3) a correlation also exists between the number of patent families and the time gap. We can imagine if an applicant is willing to apply patents from different patent authorities, with diverse rules and regulations, that it is common to get a delay resulting from

some unexpected issues. However, after consulting with our experts, we decided to only pay attention to these two indicators rather than to set them as high-coupled indicators as well.

Entropy-based weighting and collaborative filtering

Following the process of the entropy-based weighting model, we calculated the weights of the nine indicators, given in Table 3. The number of legal transactions (legal and economic value), the number of claims (technological and legal value), the time gap (technological value), and the number of citations (technological value) were weighted as the top 4 indicators, and the numbers of IPCs and terms followed—the two indicators both related to technological value. At this stage, the six indicators can be considered as the main indicators of this system. Especially, although we set the number of patent families as one dynamic indicator, considering a balance between the two indicators patent family and time gap, it is reasonable to only give time gap a high weight.

The collaborative filtering model followed, and the parameters and related strategies are described as follows: (1) *the initial size of negative patents*: we set δ as the 5% of the entire dataset and collected 142 patents with the lowest ranking in the queue. Then, aiming to minimize the initial size of the set of negative patents to avoid exaggerative identification of noise, we used the means of the six main indicators to be six additional thresholds (i.e., once a patent has a value in either of the six main indicators more than a related threshold, we would remove the patent from the set of negative patents); (2) *the overflowing range of the similarity measure*: we set τ as 0.9—a relatively conservative upper line for similarity measure; and (3) *the size of the outputs*: we set θ as the 10% of the entire dataset—it is acceptable if the number of the innovative patents is far less than this threshold. The collaborative filtering model ended after the 3-round iteration, with 751 remaining patents as the output. The stepwise results of the iterative process in the collaborative filtering model are given in Table 5.

Table 5. Stepwise results of the term clumping process

<i>Iteration</i>	<i>#Remaining patents</i>	<i># Negative patents</i>	<i>% Reduce</i>
1	28,504	5	0.02%
2	17,679	10,825	37.98%
3	751	16,928	95.75%

Note: #Terms: the number of terms; % Reduce: the proportion of reduction.

It is interesting that we only obtained 5 patents in the first iteration; one explanation for this phenomenon could be that there were only a few patents with bad-looking values in all these six main indicators, even those ranked at the bottom of the queue. This observation might be able to endorse a finding that if an invention can be patented, it is definitely equipped certain features from technological, economic, or legal perspectives e.g., covering a number of IPCs (i.e., technological classes and sub-classes), containing a number of technological terms, or claiming a number of technological significances for legal protection. Such features might be further approved by the time (e.g., being cited by follow-up inventions, or being maintained or transferred). In addition, it is meaningful to bring down the initial size of negative patents, since the collaborative filtering model is mostly based on a patent's composition with the values in the nine indicators rather than its semantic content and the iterative process would exponentially increase the size of negative patents. Therefore, an accurate initial set of negative patents can be a guarantee, and sometimes, engaging expert knowledge to help identify the initial set of negative patents can be an alternative option.

Aiming to further filter patents from diverse requirements and in a visual way, we generated six three-dimensional maps as examples, given in Fig. 2. It is easy to identify valuable patents (marked as red nodes) with distinctive values in selected indicators from these maps. In addition, based on our case, certain insights on exploring interactions between selected indicators are summarized: (1) there is no significant evidence to support a direct relationship between citation and the other main indicators such as patent family, legal transaction, and time gap. However, relatively weak negative linkages seem to exist between citation and claim, assignee, and core terms, that is, those patents with a high number of citations usually have limited claims, assignees, and core terms; and (2) specifically considering the two indicators we removed from our indicator system (i.e., the numbers of patent references and non-patent references), Fig. 2 indicates that patents with a large number of patent references cannot increase the opportunity to be cited, despite the fact that several exceptions also exist.

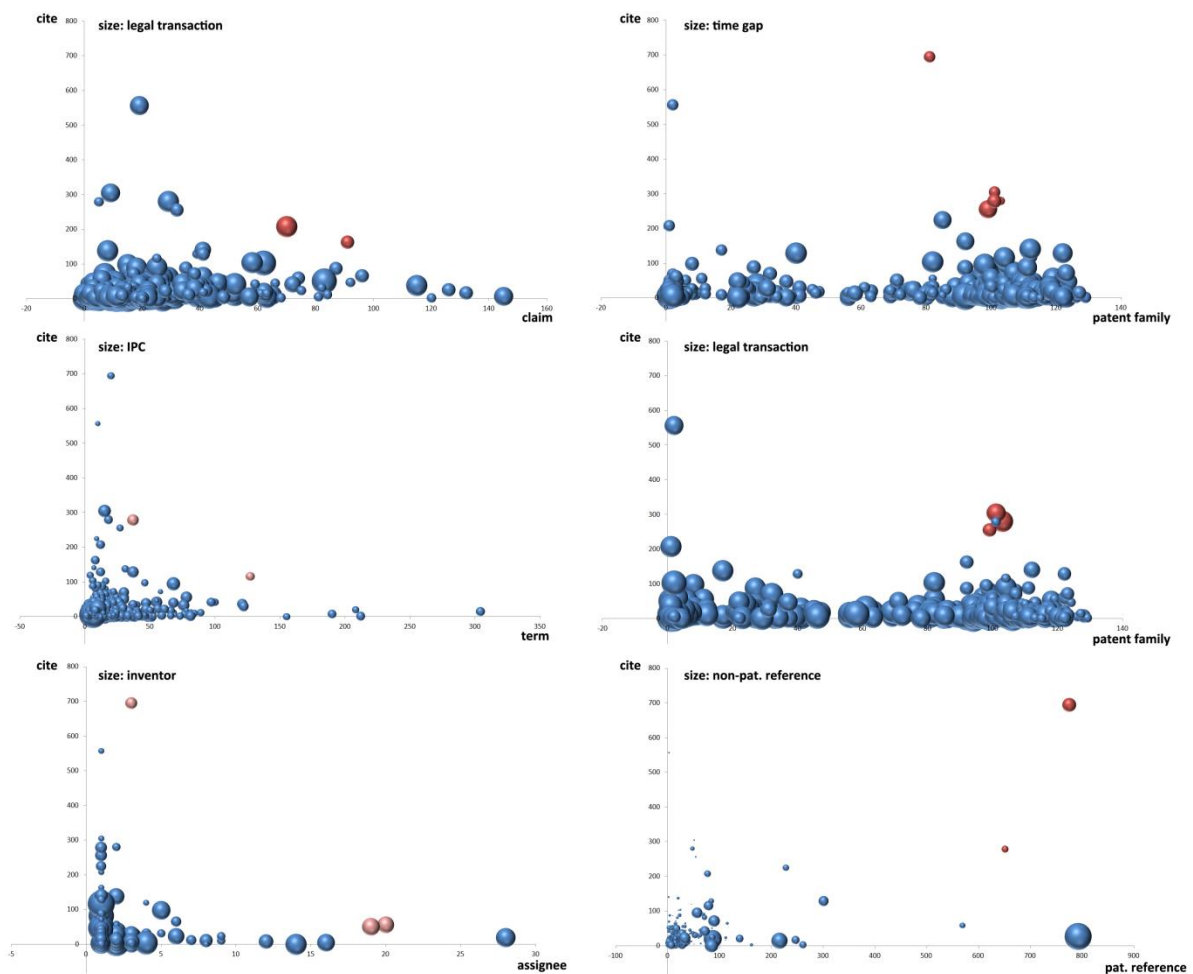


Fig. 2. Three-dimensional maps for filtering patents

Validation measures

Aiming to better demonstrate the effectiveness of our method, we conducted two ways to validate the results of our method: (1) to compare the results derived from our method with the results ranked from a traditional weighting approach (e.g., AHP), and (2) to investigate a case study on certain patents that were identified by our method but were not ranked in the top list by traditional weighting approaches. The comparison can be used to demonstrate the fact that our method can do what traditional approaches do, and can do even better, when the

case study indicates our method holds the ability to discover “underlying” patents with the potential in technological innovation.

(1) Comparison with an AHP-based weighting approach

We followed the basic steps of the AHP fundamental scale proposed by Saaty (1990). The pairwise values and the matrix are given in Table 6. Then, we used the Priority Vector (P.V.) as the vector of weights for the nine indicators (we also removed the number of patent references and the number of non-patent references here), and ranked the raw 28,509 patents with these weights.

Table 6. Pairwise comparison matrix

<i>Indicator</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>P.V.</i>
1	1	1/4	1/2	1/2	1/8	2/5	1/2	2	1/4	0.04
2	4	1	3	6	1/4	2	1/2	4	2	0.17
3	2	1/3	1	2	1	2	2	5	1/2	0.13
4	2	1/6	1/2	1	2/5	1	1/2	4	2	0.08
7	8	4	1	5/2	1	2	1/2	8	3	0.23
8	5/2	1/2	1/2	1	1/2	1	1/2	4	2	0.10
9	2	1/2	1/2	2	1/2	2	1	6	1/4	0.10
10	1/2	1/4	1/5	1/4	1/8	1/4	1/6	1	1/4	0.03
11	4	1/2	2	1/2	1/3	1/2	4	4	1	0.12

$$\lambda_{max} = 0.088$$

Note: The indicators follow the numbers given in Table 3.

It is clear that these expert knowledge-based weights are different from the entropy-based weights. The number of citations was weighted as the most prior indicator, and the number of patent families, the number of legal transactions, the number of terms, and the number of assignees were emphasized. Despite the fact that both weighting approaches highlight the importance of citations and legal transactions in evaluating a patent’s value, the inconsistency exists on the weights to patent family, assignees, and the time gap between the application date and the issue date. Actually, it is interesting that legal transactions and citations are the only two indicators that cannot be directly generated by patent applicants, and at this stage, the two indicators can be more objective than the remaining nine indicators. Since both sets give a priority to the two indicators, we in particular focused on the two indicators and designed a way to measure the accuracy of the two methods.

We first ranked raw patents with only the number of citations and the number of legal transactions respectively, and labeled a small set of patents in a top *n* list. On one hand, we searched these patents in the remaining patents of the 2-round and 3-round iterations² of our method, and recorded the number of patents existing in the two sets, respectively. On the other hand, we simulated the iterative process of our method and, based on the ranking given by the AHP-based weighting approach, labeled the top 1 to 751 patents as the 3-round iteration and the top 752 to 17,679 patents as the 2-round iteration. We defined *Accuracy* as the indicator of validation measures, which can be calculated as follows:

$$Accuracy = \frac{\#3R + \#2R \times 0.5}{\#Raw}$$

² Note that patents in the 2-round iteration will exclude patents in the 3-round iteration, i.e., there are 751 patents in 3R and 16,928 patents in 2R.

where #Raw is the total number of the patents in the top n list, and #3R and #2R is the number of patents that respectively exist in the 2-round and 3-round sets of remaining patents.

Based on the two indicators for the number of citations and the number of legal transactions, the accuracy of the AHP-based weighting approach and the entropy-based weighting approach is given in Table 7.

Table 7. Accuracy of the AHP-based and entropy-based weighting approaches (in indicators: the number of citations and the number of legal transactions)

<i>Method</i>	<i>Indicator</i>	<i># Raw</i>	<i># 3R</i>	<i># 2R</i>	<i>Accuracy</i>
AHP	# citations	10	6	4	0.80
		15	8	7	0.77
	# legal transactions	18	9	9	0.75
		38	10	28	0.63
Entropy	# citations	10	6	4	0.80
		15	9	6	0.80
	# legal transactions	18	18	0	1.00
		38	38	0	1.00

It can be argued that our method weighted the number of legal transactions as a crucial indicator with the highest weight, which can result in the perfect accuracy of our method in this indicator. However, although our expert panel thought that citations were the most important indicator, our method still performed better than the AHP-based weighting approach in citations. This comparison can act as a fair stage to compare the effectiveness of our method with other approaches, and the results indicate the strength of our method in identifying patents with distinct values in certain individual indicators.

(2) Case study on windfall patents

We define windfall patents as those searched by our method but neither having good-looking values in all indicators nor being ranked in a top list by traditional weighting approaches. Comparing the 3-round set of remaining patents identified by our method to the top 751-patent list ranked by the AHP-based weighting approaches, only 169 patents were the same. Therefore, it is interesting and promising to delve into the 582 windfall patents and confirm whether they really were the ones with potential in technological innovation.

We specifically selected five windfall patents and listed their information in Table 8. It is obvious that all of them cannot be traditionally considered to be outstanding candidates since they have shortages in certain indicators, and don't have either sufficient citations or don't cover a broad range of IPCs. However, insights dug out by case studies can be good evidence to endorse the potential of these patents in technological innovation.

Table 8. Samples of windfall patents

	<i>Indicator</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>
	<i>Patent Number</i>									
1	US20040188519A1	1	1	5	70	208	3	12	2	1
2	CN1341604A	1	2	8	15	16	1	19	3	1
3	EP1369820A2	3	40	1	39	129	3	12	7	1
4	WO2002027542A1	2	92	1	15	4	1	6	11	1
5	US20030152651A1	5	103	1	1	37	4	121	10	1

Note: The indicators follow the numbers given in Table 3.

Patent 1 has a good citation count, but its ranking would be heavily influenced by the number of patent families and also the number of IPCs in traditional weighting approaches. One piece of intriguing evidence here is that the only assignee of this patent is Kepler Energy Ltd., which was founded by three senior academics from the University of Oxford³. Definitely, Oxford can theoretically support the technological value of this patent, and acting as the core of technology transfer activities (from leading universities to industry sectors), the value of these patents has been well-approved.

The only distinctive value of Patent 2 is the number of legal transactions, but it is surprising to retrieve this patent from our study. This patent was applied by a researcher from Sichuan University, China, and was mostly based on one of his publications in *Nature Biotechnology* (Qiu et al. 2003). Despite a dispute on academic fraud in 2004, a follow-up publication of this author in *Nature Biotechnology* (Qiu et al. 2007) helped exonerate the author from such accusations and supported the technological innovation of this research. At the same time, the continuous maintenance records until present also demonstrated its economic and legal value⁴.

Patents 3, 4, and 5 have distinctive values in certain individual indicators, but limited legal transaction and a large time gap between the application date, and the issue date weakens their rankings. Generally, Patents 3 and 4 relate to techniques of data processing (i.e., video and image encoding, and information retrieval), and they were applied by the Microsoft Cooperation and Intel Cooperation, respectively. Patent 5, with a large number of core terms and patent families, was applied by the Talsy Group, a listed company in the biomedicine sector and specifically focusing on herbal medicine⁵. Obviously, the two world-leading IT companies and the China-leading herbal medicine company endorse the technological value of their patents.

It is clear that the five windfall patents indicate significant value in technological innovation, and they could be the evidence to demonstrate the reliability of our method. However, we also noticed that in the set of outputs there were still a number of patents that have either expired for years or could not explore any significance in technological innovation. At this stage, we concluded that our method could be used as an effective tool for filtering noise and bringing down the size of target patents from more than ten thousand to several hundred, in which a scale engaging expert knowledge to manually identify innovative patents becomes valuable and efficient.

Discussion and Conclusions

This paper constructed an entropy-based indicator system for evaluating a patent's potential in technological innovation. Based on eleven patent indicators from technological, economic, and legal perspectives, Shannon's entropy was used to quantitatively weight these indicators and rank patents to identify a set of negative patents, with bad-looking values in almost all indicators. A collaborative filtering technique was introduced to measure the similarity between all remaining patents and negative patents in an iterative process, in which the patents sharing a high similarity value with the set of negative patents were identified as new negative patents that would be used in the next iteration. A small set of filtered patents were

³ <http://www.keplerenergy.co.uk/about-us.html>

⁴ http://www.thomsoninnovation.com/tip-innovation/recordView.do?datasource=T3&category=PAT&&idType=pns&databaseIds=PATENT&fromExternalLink=true&recordKeys=US6942993B2&locale=en_US

⁵ http://www.talsy.com/en_web/index.aspx

the outputs of our method, which were considered to be patents with good potential in technological innovation.

Technical implication

Compared to traditional expert knowledge-based weighting approaches, the significant strength of the entropy-based weighting method is the ability to quantitatively calculate the weights of indicators. Such ability, on one hand, is data-oriented and can adapt to changing data situations and actual requirements. On the other hand, the objectivity of quantitative computing effectively omits bias resulted from those expert knowledge-based qualitative approaches. However, it is notable that since the calculation of weights is totally automatic the design of related indicator systems will become crucial.

The collaborative filtering technique is one of the most representative techniques in the field of recommender systems, and it is fruitful to be introduced to our method. First, it provides empirical support for our hypothesis that patents sharing a high similarity value with negative patents will also have limited potential in technological innovation. Second, similarity measures in collaborative filtering are the same as those in bibliometrics, which makes it easy to implement collaborative filtering techniques in patent documents. In addition, the iterative process in the collaborative filtering model uses a relatively appropriate way to narrow down the scale of target patents to a manual-readable level, which neither exaggeratively removes valuable patents nor remain as too noisy patents.

Limitations and future study

Despite the fact that we seriously addressed concerns while constructing the indicator system and discussed the viability of certain indicators (e.g., the numbers of patent references and non-patent references) in the case study, it is still critical to package them in one framework without empirically evaluating the hypotheses of all involved indicators, although most isolated ones have been tested by some econometric models. We also notice that the viewpoints to some indicators (e.g., patent claims and backward citations) seem inconsistent between academic researchers (e.g., economists) and patent experts, both theoretical and empirical argumentations might be required further. In addition, although we emphasize the adaptability of our method (e.g., the entropy-based weighting model and the collaborative filtering model), the selection of patent indicators needs to focus on single patent authority, in which the same jurisdiction (e.g., related laws, examination systems, and languages) is applied. At this stage, engaging new indicators or exploring different use of existing indicators would be interesting.

The technical limitations of our method and related further improvement that we can anticipate include: (1) we need to configure three thresholds in our current method (i.e., the initial size of negative patents, the overflowing range of the similarity measure, and the size of the outputs), and the three thresholds can be considered as the crucial factors to influence the performance of our method. At the moment, we decided them based on experiments and our previous experience, so one future direction is to use training data to learn the thresholds automatically; (2) considering that the data of our case study covered a broad range of research domains, we only measured the similarity between patents by using the vector of indicators, but when the case comes to a specific scientific or technological domain, the significance of IPCs and core terms needs to be highlighted. Using IPCs and terms to measure technological similarity can greatly help increase the performance of our method; and (3) a thorough approach for validation measures could help further improve our method.

Acknowledgement

We acknowledge support from the Australian Research Council (ARC) under Discovery Project DP140101366, and the National Science Foundation of China under Grant 71503020 and Grant 71673024. We are also grateful to the anonymous reviewers for addressing critical but valuable comments on the evaluation of patent indicators.

References

- Agrawal, A. (2006). Engaging the inventor: Exploring licensing strategies for university inventions and the role of latent knowledge. *Strategic Management Journal*, 27(1), 63-79.
- Alcacer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4), 774-779.
- Archibugi, D., & Planta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, 16(9), 451-519.
- Azagra-Caro, J., Fernández-de-Lucio, I., Perruchas, F., & Mattsson, P. (2009). What do patent examiner inserted citations indicate for a region with low absorptive capacity? *Scientometrics*, 80(2), 441-455.
- Basberg, B. L. (1987). Patents and the measurement of technological change: A survey of the literature. *Research Policy*, 16(2), 131-141.
- Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, 37(5), 932-945.
- Bozbura, F. T., Beskese, A., & Kahraman, C. (2007). Prioritization of human capital measurement indicators using fuzzy AHP. *Expert Systems with Applications*, 32(4), 1100-1112.
- Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169-173.
- Chen, D.-Z., Chang, H.-W., Huang, M.-H., & Fu, F.-C. (2005). Core technologies and key industries in Taiwan from 1978 to 2002: A perspective from patent analysis. *Scientometrics*, 64(1), 31-53.
- Chen, Y.-S., & Chang, K.-C. (2012). Using the entropy-based patent measure to explore the influences of related and unrelated technological diversification upon technological competences and firm performance. *Scientometrics*, 90(3), 825-841.
- Choi, C., & Park, Y. (2009). Monitoring the organic structure of technology based on the patent development paths. *Technological Forecasting and Social Change*, 76(6), 754-768.
- Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C.-H. (2011). SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics*, 88(3), 863-883.
- Cronin, B., & Meho, L. (2006). Using the h- index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57(9), 1275-1278.
- de Solla Price, D. (1983). *Sealing Wax and String: A philosophy of the experimenter's craft and its role in the genesis of high technology*. Paper presented at the Sarton Lecture, AAAS Meeting (May 1983).
- Egghe, L. (2006). Theory and practice of the g-Index. *Scientometrics*, 69(1).
- Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1), 117-132.
- Funk, R. J., & Owen-Smith, J. (2016). A dynamic network measure of technological change. *Management Science*. doi: 10.1287/mnsc.2015.2366
- Gittelman, M., & Kogut, B. (2003). Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science*, 49(4), 366-382.
- Glänzel, W., & Thijs, B. (2012). Using "core documents" for detecting and labelling new emerging topics. *Scientometrics*, 91(2), 399-416. doi: 10.1007/s11192-011-0591-7
- Grimaldi, M., Cricelli, L., Di Giovanni, M., & Rogo, F. (2015). The patent portfolio value analysis: A new framework to leverage patent information for strategic technology planning. *Technological Forecasting and Social Change*, 94, 286-302. doi: 10.1016/j.techfore.2014.10.013
- Grupp, H. (1990). The concept of entropy in scientometrics and innovation research: an indicator for institutional involvement in scientific and technological developments. *Scientometrics*, 18(3-4), 219-239.

- Guan, J., & He, Y. (2007). Patent-bibliometric analysis on the Chinese science—technology linkages. *Scientometrics*, 72(3), 403-425.
- Guellec, D., & de la Potterie, B. v. P. (2000). Applications, grants and the value of patent. *Economics Letters*, 69(1), 109-114.
- Guellec, D., & de la Potterie, B. v. P. (2001). The internationalisation of technology analysed with patent data. *Research Policy*, 30(8), 1253-1266.
- Hall, B. H. (2002). The NBER patent citation data file: Lessons, insights and methodological tools. In A. B. Jaffe & M. Trajtenberg (Eds.), *Patents, Citations and Innovations* (pp. 403-460). Cambridge, MA: MIT Press.
- Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *Rand Journal of Economics*, 16-38.
- Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8), 1343-1363.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 16569-16572.
- Iwami, S., Mori, J., Sakata, I., & Kajikawa, Y. (2014). Detection method of emerging leading papers using time transition. *Scientometrics*, 101(2), 1515-1533. doi: 10.1007/s11192-014-1380-x
- Jiang, J., Lu, J., Zhang, G., & Long, G. (2011). *Scaling-up item-based collaborative filtering recommendation algorithm based on hadoop*. Paper presented at the 2011 IEEE World Congress on Services.
- Kay, L., Newman, N., Youtie, J., Porter, A. L., & Rafols, I. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology*, 65(12), 2432-2443.
- Lee, Y.-G., Lee, J.-D., Song, Y.-I., & Lee, S.-J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics*, 70(1), 27-39.
- Lee, Y.-G. (2009). What affects a patent's value? An analysis of variables that affect technological, direct economic, and indirect economic value: An exploratory conceptual approach. *Scientometrics*, 79(3), 623-633.
- Lei, X.-P., Zhao, Z.-Y., Zhang, X., Chen, D.-Z., Huang, M.-H., Zheng, J., . . . Zhao, Y.-H. (2013). Technological collaboration patterns in solar cell industry based on patent inventors and assignees analysis. *Scientometrics*, 96(2), 427-441.
- Leydesdorff, L. (2002). Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports. *Scientometrics*, 53(1), 131-159.
- Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics*, 98(3), 1583-1599.
- Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Systems*, 74, 12-32.
- Makri, M., Hitt, M. A., & Lane, P. J. (2010). Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal*, 31(6), 602-628.
- Mao, M., Lu, J., Zhang, G., & Zhang, J. (2016). Multirelational social recommendations via multigraph ranking. *IEEE Transactions on Cybernetics*. doi: 10.1109/TCYB.2016.2595620
- Martínez, C. (2011). Patent families: when do different definitions really matter? *Scientometrics*, 86(1), 39-63.
- Martino, J. P. (2003). A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change*, 70(8), 719 - 733.
- Meyer, M., & Tang, P. (2007). Exploring the "value" of academic patents: IP management practices in UK universities and their implications for Third-Stream indicators. *Scientometrics*, 70(2), 415-440.
- Mowery, D. C., Sampat, B. N., & Ziedonis, A. A. (2002). Learning to patent: Institutional experience, learning, and the characteristics of US university patents after the Bayh-Dole Act, 1981-1992. *Management Science*, 48(1), 73-89.
- Narin, F., & Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics*, 36(3), 293-310.

- Pavitt, K. (1985). Patent statistics as indicators of innovative activities: possibilities and problems. *Scientometrics*, 7(1-2), 77-99.
- Qiu, X.-Q., Wang, H., Lu, X.-F., Zhang, J., Li, S.-F., Cheng, G., . . . Zhou, Y.-Q. (2003). An engineered multidomain bactericidal peptide as a model for targeted antibiotics against specific bacteria. *Nature biotechnology*, 21(12), 1480-1485.
- Qiu, X.-Q., Wang, H., Cai, B., Wang, L.-L., & Yue, S.-T. (2007). Small antibody mimetics comprising two complementarity-determining regions and a framework region for tumor targeting. *Nature biotechnology*, 25(8), 921-929.
- Reitzig, M. (2003). What determines patent value?: Insights from the semiconductor industry. *Research Policy*, 32(1), 13-26.
- Reitzig, M. (2004). Improving patent valuations for management purposes validating new indicators by analyzing application rationales. *Research Policy*, 33(6-7), 939-957. doi: 10.1016/j.respol.2004.02.004
- Rip, A. (1988). Mapping of science: Possibilities and limitations. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 253-273). North-Holland: Elsevier Science Publishers B.V.
- Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European journal of operational research*, 48(1), 9-26.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.
- Sapsalis, E., de la Potterie, B. v. P., & Navon, R. (2006). Academic versus industry patenting: An in-depth analysis of what determines patent value. *Research Policy*, 35(10), 1631-1645.
- Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics*, 26(1), 193-211.
- Schumpeter, J. A. (1939). *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process* (Vol. 1). New York: McGraw Hill.
- Shambour, Q., & Lu, J. (2012). A trust-semantic fusion-based recommendation approach for e-business applications. *Decision Support Systems*, 54(1), 768-780.
- Shannon, C. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27(3), 379-423.
- Tijssen, R. J. (2001). Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows. *Research Policy*, 30(1), 35-54.
- Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, 45(3), 707-723. doi: 10.1016/j.respol.2015.11.010
- Vinkler, P. (2013). Comparative rank assessment of journal articles. *Journal of Informetrics*, 7(3), 712-717. doi: 10.1016/j.joi.2013.04.006
- Von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 34(10), 1591-1607.
- Waltman, L., Calero- Medina, C., Kosten, J., Noyons, E., Tijssen, R. J., Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432.
- Wang, B., & Hsieh, C.-H. (2015). Measuring the value of patents with fuzzy multiple criteria decision making: insight into the practices of the Industrial Technology Research Institute. *Technological Forecasting and Social Change*, 92, 263-275. doi: 10.1016/j.techfore.2014.09.015
- Wang, S.-J. (2007). Factors to evaluate a patent in addition to citations. *Scientometrics*, 71(3), 509-522.
- Xu, H., Martin, E., & Mahidadia, A. (2014). Contents and time sensitive document ranking of scientific literature. *Journal of Informetrics*, 8(3), 546-561. doi: 10.1016/j.joi.2014.04.006
- 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.
- Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014a). “Term clumping” for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, 85, 26-39.

- Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., & Yan, A. (2014b). Triple Helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with semantic TRIZ and technology roadmapping. *Scientometrics*, 99(1), 55-75.
- Zhang, Y., Shang, L., Huang, L., Porter, A. L., Lu, J., & Zhu, D. (2016). A hybrid similarity measure method for patent portfolio analysis *Journal of Informetrics*. doi: 10.1016/j.joi.2016.09.006