Expert Systems With Applications 210 (2022) 118564

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Measuring latent combinational novelty of technology

Xiaoling Sun *, Na Chen, Kun Ding

Institute of Science of Science and S.&T. Management, Dalian University of Technology, Dalian 116024, China

ARTICLE INFO

Keywords:
Technological novelty
Knowledge combination
Link prediction
Hierarchical similarity

ABSTRACT

Novelty is considered an important driving force of scientific and technological innovation. How to measure novelty has drawn much attention in recent years. The comprehensive measurement of the novelty could help to identify novel patents as soon as possible and reduce the risk of delayed identification of important technologies. This research introduces a comprehensive measure that could identify novel technology using IPC in patents from a knowledge combinational perspective. The existing methods for measuring novelty commonly use the co-occurrence of knowledge pairwise combinations and identify novelty by assessing the new pairings that did not exist. Besides considering the number of direct co-occurrence of knowledge combinations to evaluate novelty, the proposed method integrates indirect link probability and hierarchical similarity in the IPC tree structure. The feasibility of the measure is demonstrated by applying it to the patent data in the field of Artificial Intelligence (AI). Compared with previous measures, the proposed measure could capture the latent distance between knowledge pairings and identify more novel combinations. The relationship between novelty and citations in the AI field shows that: High-novelty/high-conventional patents have a higher average number of citations and a higher probability of being "hit" patents, indicating that novel patents build on prior knowledge have a relatively higher future impact.

1. Introduction

Scientific and technological innovation plays an important role in promoting the development of society. In radical technological innovation, novelty is an important driver. It is of great practical significance to have a set of methods to identify novelty and avoid delayed recognition of important patents (Stephan, Veugelers, & Wang, 2017; Van Raan, 2004; Wang, Veugelers, & Stephan, 2017).

Novelty is an essential feature of creative ideas, the building blocks of which are often embodied in existing knowledge (Uzzi, Mukherjee, Stringer, & Jones, 2013). Novelty and conventionality are sometimes displayed together, as novel ideas can be difficult to absorb and communicate (Uzzi et al., 2013). For example, Newton presented his laws of gravitation using well-accepted geometry rather than his newly developed calculus. The existing methods usually measure novelty from a combinational perspective of original knowledge (Boyack & Klavans, 2014; Carayol, Llopis, & Lahatte, 2016; Foster, Rzhetsky, & Evans, 2015; Veugelers & Wang, 2019; Wang et al., 2017). The knowledge components include cited journals, technology codes, keywords, and so on. The novelty is identified by searching for new or unexpected combinations. The prior research usually only considers the co-occurrence of pairwise combinations, which is likely to underestimate the combination probability. For example, two knowledge components may not co-occur before, but they could be highly related and have a high probability to co-occur, making the combination not remarkably novel. Instead of just considering the frequency of co-occurrence, methods need to be able to capture the latent distance between knowledge components.

In this paper, we propose a new comprehensive method to measure the novelty of the technology, which complements the previous studies. Following the existing research, we consider novelty as the unusual combination of pre-existing knowledge components. IPC codes are used to represent knowledge components of patents. To measure the combination probability, link prediction methods in complex network studies inspire us to quantify the indirect relationships besides the direct co-occurrence relationship. The methods could measure how likely-two components are likely to co-occur based on the IPC co-occurrence network structure. In addition to co-occurrence relationships, we also consider the hierarchical-level relationship between knowledge components. The hierarchical-level distance (McNamee, 2013) is also an important factor that influences the combination probability, as it is relatively easier to search in familiar domains than across unfamiliar domains.

* Corresponding author.
E-mail addresses: xlsun@dlut.edu.cn (X. Sun), nchen@mail.dlut.edu.cn (N. Chen), dingk@dlut.edu.cn (K. Ding).

https://doi.org/10.1016/j.eswa.2022.118564
Received 7 May 2021; Received in revised form 8 August 2022; Accepted 12 August 2022
Available online 18 August 2022

The remainder of the paper is organized as follows: First, we begin by reviewing the literature on scientific novelty and technological novelty from the perspective of knowledge combination. Next, we introduce the proposed indicator to measure technological novelty. Then the experimental data and the results of our analysis are presented. Finally, we conclude the findings and their implications for future research.

2. Related work

Many scholars study the novelty of science and technology from the perspective of knowledge combination and believe that the combination of new knowledge components and existing knowledge components is the main source of novelty. Papers and patents are regarded as the carriers of scientific knowledge and technological knowledge respectively, which have been commonly used to identify the novelty. In the field of document-level or sentence-level novelty mining, novelty detection aims at identifying whether a test sample is novel or unusual compared to a previously observed sample (Silva, Vieira, Martínez, & Paiva, 2021; Tsai & Kwee, 2011). Here we focus on the combinational novelty of the technology. As the research about the combinational novelty of science is also very related, we summarize the related work from the knowledge combinational perspective for both science and technology.

2.1. Combinational novelty of science

In the measurement of scientific novelty, papers are used as carriers of scientific knowledge. Previous studies mainly used information such as knowledge sources (e.g. cited references) and research fields (e.g. classification numbers) of the paper as proxy indicators of combinational novelty and discussed its relationship with the citations of the paper (Azoulay, Graff Zivin, & Manso, 2011; Boudreau, Guinan, Lakhani, & Riedl, 2016; Hofstra et al., 2020; Shibayama & Wang, 2020; Uzzi, Mukherjee, Stringer, & Jones, 2013; Wagner, Whetsell, & Mukherjee, 2019).

The cited references of papers contain diverse knowledge. The combinations of cited journals are commonly used to study the novelty of scientific papers. Uzzi et al. (2013) hypothesized that high-impact papers were likely to cite novel combinations of existing knowledge, which were embedded in and supported by conventional knowledge. They defined novelty by identifying atypical combinations of cited journals, that is, novel or unusual combinations of co-cited articles at the journal level. Boyack and Klavans (2014) replicated Uzzi's finding using slightly different methods and found that atypical co-cited journal combinations, and thus citation rates, were highly dependent on discipline and journal effects. Lee, Walsh, and Wang (2015) adapted Uzzi's method for the study of creativity in scientific teams and found that team characteristics had different effects on novelty and the impact of papers. Stephan et al. (2017) and Wang et al. (2017) also tested atypicality using a "commonality score" to measure the expected number of co-citations. Wang et al. (2017) measured novelty by making unprecedented combinations in the referenced journals and considered the difficulty of making such combinations. Schilling and Green (2011) analyzed the Dewey decimal code information of the references and found that search scope, depth, and atypical connections between different research domains significantly increased the impact of a paper. These studies preliminarily showed that the papers with appropriate novelty and conventionality had better academic influence.

Recent studies quantitatively examined the convergent validity of several novelty scores, wondering whether these novelty scores could measure what they were proposed to measure (Bornmann, Tekles, Zhang, & Ye, 2019; Tahamtan & Bornmann, 2018).

2.2. Combinational novelty of technology

The combinational novelty of technology refers to the degree of difference compared the combination of new knowledge components and existing knowledge components in the process of technological invention, with the previous combination of knowledge components in this field. In the technological innovation literature, measuring technological novelty from a knowledge combination perspective has been extensively studied and validated using patent information (Arthur, 2007; Fleming, 2001; Verhoeven, Bakker, & Veugelers, 2016; Youn, Strumsky, Bettencourt, & Lobo, 2015; Zhang et al., 2017; Kim, Cerigo, Jeong, & Youn, 2016; Harrigan, Di Guardo, Marku, & Velez, 2017; He & Luo, 2017; Chai & Menon, 2019).

The technology classification codes in the patents are commonly used to measure novelty. For patent examination, technology classification codes are used by patent examiners when searching for relevant prior art during the patent application examination process. Technology classification codes are underutilized data resources used to identify unique technical capabilities, define technical space, mark the arrival of technical novelty and measure technical complexity. Using patent classification codes to describe the combination process of inventions makes it possible to assess the novelty of technology. Fleming (2001) believed that technological novelty came from the combination of new classification codes and existing classification codes, and the negative binomial regression model using patent citation data showed that the knowledge local search process reduced the success rate of inventions, but increased the probability of making breakthroughs. Strumsky and Lobo (2015) followed Fleming's approach and introduced another type of novel patent labeled as 'technological originations'. These patents were classified in at least one technology class that was newly introduced by the US patent office. Valverde, Solé, Bedau, and Packard (2007) utilized technology codes to characterize the combinational process of invention and assessed the novelty of inventions. Verhoeven et al. (2016) introduced a measure of novelty in knowledge origins, based on if the patent cited patents or papers from research areas that were never cited before in its patent class. They built on the work of Arthur (2007) and found that combining the combinational novelty and the novelty in knowledge origins was powerful in identifying breakthrough inventions. Youn et al. (2015) found that the combinational inventive process showed two phenomena: "exploitation" (improvement of existing technology combination) and "exploration" (introduction of new technology code combination). Kim et al. (2016) quantitatively studied the novelty distribution of technology pairs and the connection between the novelty profile of an invention and its future impact. He and Luo (2017) identified a novelty 'sweet spot' in which the mix of novel combinations of prior technologies favored an invention's eventual success.

To the best of our knowledge, although there were many studies measuring novelty from a knowledge combinational perspective, most of which focused on counting the frequency of co-occurrence pairs. Research has started to consider the knowledge distance between class pairs (Luo, Sarica, & Wood, 2021). The previous studies inspired us to propose a more comprehensive method to measure combinational novelty, making the identification of novelty more reliable.

3. Methodology

3.1. Defining technological novelty

Technology is identified as having novelty in combination if the combinations of components are unlikely and different from those embodied in previous technologies.

We use IPC codes that a patent is assigned to as a proxy for the knowledge components of the invention and the pairwise combinations of IPC codes as a proxy for the combinational process. IPC is suitable for assessing the combinational process of technology because it classifies according to the complete technical information contained in the patent application (Gruber, Harhoff, & Hoisl, 2013). IPC has a hierarchical structure as illustrated in Fig. 1. We have choices to rely on different

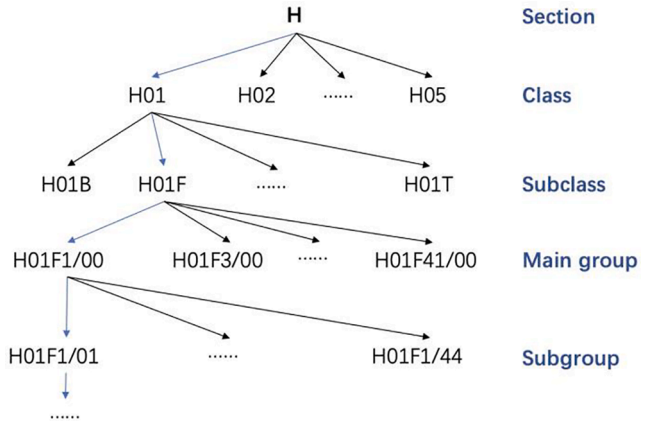

Fig. 1. Example of IPC hierarchical structure.

levels of IPC codes. Main group-level IPC codes are often used (Verhoeven et al., 2016) in the research, while in this paper we want to explore deeper. Subgroup-level IPC codes, compared to main group-level codes, define more detailed technical subject areas within the main groups. We would like to reveal the knowledge combination reflected by subgroup-level IPC codes.

For each IPC pair of a patent, we firstly count the frequency of the pair in previous patents before the application year. A patent has "Novelty in Co-occur (*NC*)" when it contains at least one pair of IPC codes that did not previously co-occur or rarely co-occur. Then, the connection between IPC pairs is extended to long-distance relationships using link prediction methods. A patent has "Novelty in Link (*NL*)" when it contains at least one pair of IPC codes that is unlikely to connect based on previous relationships. Besides the co-occurrence relationship, the innate hierarchical-level similarity in the IPC tree (Fig. 1) between IPC codes is also considered. A patent has "Novelty in Similarity (*NS*)" when it contains at least one pair of IPC codes that has a long distance in the IPC hierarchical tree structure.

For pairwise combination (c_1, c_2), the above three types of novelty are measured by three probabilities and combined into a pairwise combination probability $p_{combine_t}(c_1, c_2)$. The lower the probability is, the higher novelty of the combination.

3.2. Pairwise combination probability

The probability of pairwise combination (c_1, c_2) is defined as Eq.1 using a linear weighted model, combining three aspects including co-occurrence probability p_{occur}, link probability p_{link} and hierarchical-level similarity $p_{similarity}$. The three probabilities are normalized using the max–min normalization method and α and β are the adjustment factors to control the proportion of three probabilities. To reduce subjectivity, the adjustment factors are set objectively by the method detailed in the following subsection Entropy-based weighting model. The probability is computed based on the data before t, which is the application year of the patent.

$$p_{combine_t}(c_1, c_2) = \alpha * p_{occur_t}(c_1, c_2) + \beta * p_{link_t}(c_1, c_2) + (1 - \alpha - \beta) * p_{similarity}(c_1, c_2) \quad (1)$$

In co-occurrence probability, several statistical measures could be used, including mutual information, z-score, and frequency. Here we use the normalized frequency to compute the co-occurrence probability of the given IPC pair based on the observation in previous data, as shown in Eq.2, where $f_{C_1C_2}$ is the observed frequency of the pair in the data before the application year t.

$$p_{occur_t}(c_1, c_2) = (f_{C_1C_2} - f_{min})/(f_{max} - f_{min}) \quad (2)$$

To measure the indirect connections of the IPC pair, the link prediction method is applied here to measure how likely it will occur based on the current knowledge network structure. In the knowledge network, nodes represent IPC codes and two IPC codes are connected if they co-occurred in one or more patents. We study the proposed measure in the context of link prediction in an unsupervised manner. The problem definition is: Given a snapshot of a network at time t, predicting links that will be added to the network during the interval from time t to a given future time t' (Liben-Nowell & Kleinberg, 2007). There are some representative methods in network analysis, for example, common neighbors, Adamic/Adar indicator (Adamic & Adar, 2003), and Katz indicator (Katz, 1953). According to previous studies (Liben-Nowell & Kleinberg, 2007; Sun, Lin, Xu, & Ding, 2015), Adamic/Adar has been proved to be a relatively simple and effective method, which refines the simple counting of common neighbors by weighting rarer neighbors more heavily. Therefore, Adamic/Adar is adopted to measure the link probability of two IPC codes, computed by Eq.3. $\Gamma(k)$ denotes the set of neighbors of node k in the knowledge network.

$$p_{link_t}(c_1, c_2) = \sum_{k \in \Gamma(c_1) \cap \Gamma(c_2)} 1/(log|\Gamma(k)|) \quad (3)$$

In the measurement of hierarchical-level similarity, we consider both the length of the shortest path and depth in the IPC hierarchical tree of two IPC codes. For example, the length between two IPC codes "H01B1/01" and "H01F1/01" in the hierarchical tree is 6. In general, the longer the distance between two concepts is, the lower the similarity is. Depth refers to the depth of the LCA (Lowest Common Ancestor) of two concepts. For "H01B1/01" and "H01F1/01", the LCA is "*H*01", which is the second level of the tree, therefore the depth of these two codes is 2. The similarity is not only related to the length of the shortest path but also related to the conceptual level. In the case of the same length, the similarity increases with the decrease of the conceptual level in the hierarchical structure. Inspired by the previous study (Liu, Zhou, & Zheng, 2007), the hierarchical-level similarity is measured by Eq. (4).

$$p_{similarity}(c_1, c_2) = F(l, d) = f(d)/(f(d) + f(l)) = d/(d + l) \quad (4)$$

Where l is the shortest path length and d is the depth of the LCA of two IPC codes. f is the transfer function for d and l. Similarity is assumed to be a process of comparing common features to all the features, that is, the sum of common and different features. Transfer functions of the depth and the shortest path length between two IPC codes can be used to substitute their common and different features respectively. When $d = 0$, the two IPCs have no common features. When $l = 0$, the two IPCs are the same. The interval of similarity is [0, 1]. The simplest linear f function: $f(x) = x$ is used here, which could satisfy the requirements.

3.3. Entropy-based weighting model

The contributions of the three probabilities to the comprehensive indicator are different. We use the entropy method (Shannon, 1950) to determine the weights of adjustment factors. Entropy is a measure of uncertainty in information theory. The larger the amount of information is, the smaller the uncertainty is and the smaller the entropy is. According to the characteristics of entropy, the randomness and disorder of an event could be determined by calculating the entropy value. The entropy value is used to determine the degree of discreteness of an indicator here. The larger the degree of discreteness of the indicator is, the smaller the entropy value is and the larger the impact (weight) of the indicator on the comprehensive evaluation is. The steps of the entropy-based weighting model are shown below:

1. If all patents in year t have a total of n IPC pairs and each IPC pair has three indicators (p_{occur}, p_{link} and $p_{similarity}$), and x_{ij} is the value of the j-th index of the i-th IPC pair ($i = 1,2,3 \ldots, n$, $j = p_{occur}, p_{link}, p_{similarity}$) ;

2. Normalize indicators: The measurement units of indicators are not uniform, so they must be normalized before calculating the comprehensive indicator. The absolute values of the indicators need to be

converted into relative values, to solve the homogeneity of heterogeneous indicators. In addition, as the values of positive and negative indicators represent opposite meanings, different algorithms are used for data normalization.

If j is a positive indicator:

$$x_{ij}^{'} = \frac{x_{ij} - min\{x_{1j}, \cdots, x_{nj}\}}{max\{x_{1j}, \cdots, x_{nj}\} - min\{x_{1j}, \cdots, x_{nj}\}} \quad (5)$$

If j is a negative indicator:

$$x_{ij}^{'} = \frac{max\{x_{1j}, \cdots, x_{nj}\} - x_{ij}}{max\{x_{1j}, \cdots, x_{nj}\} - min\{x_{1j}, \cdots, x_{nj}\}} \quad (6)$$

The three indicators here are all positive indicators and Eq.5 is used to normalize the indicators.

4. Calculate the proportion of the i-th IPC pair in the indicator under the j-th index:

$$p_{ij} = x_{ij}^{'} / \sum_{i=1}^{n} x_{ij}^{'} \quad (7)$$

5. Calculate the entropy of the j-th index, where $k = \frac{1}{\ln(n)}, e_j \geq 0$:

$$e_j = -k \sum_{i=1}^{n} p_{ij} log(p_{ij}) \quad (8)$$

6. Calculate information entropy redundancy:

$$d_j = 1 - e_j \quad (9)$$

7. Calculate the weight of each indicator:

$$w_j = d_j / \sum_{j=1}^{3} d_j \quad (10)$$

8. Calculate the comprehensive score of each IPC pair i:

$$p_{combine_t}(i) = \sum_{j=1}^{m} w_j * p_{ij} \quad (11)$$

The weights are used to set the adjustment factors α and β in Eq.1, where $w_1 = \alpha$, $w_2 = \beta$, $w_3 = 1\text{-}\ \alpha\text{-}\ \beta$.

3.4. Technological novelty

The steps to calculate the novelty of a patent are illustrated in Fig. 2 and detailed as follows:

1. List all IPC codes of the patent and all pairs of IPCs;
2. Calculate $p_{occur_t}(c_1, c_2)$, $p_{link_t}(c_1, c_2)$, and $p_{similarity}(c_1, c_2)$ of each pair (c_1, c_2), according to Eq.2–4. The measures are all normalized using Eq.5;
3. The entropy method is used to weight the three normalized indicators and obtain the probability of each IPC combination $p_{combine_t}(c_1, c_2)$.
4. Normalize $p_{combine_t}(c_1, c_2)$ using Eq.12, and the value is scaled (multiplied by 10) to fall within the same range as Uzzi's z-score (Uzzi et al., 2013). z_{score} is the final index to evaluate the novelty of the IPC pair.

$$z_{score} = (p_{combine} - \xi_{exp}) / \sigma_{Var} \quad (12)$$

Where ξ_{exp} is the expected value of $p_{combine}$, and σ_{Var} is the standard deviation.

The above method gives a distribution of pairwise combination probabilities based on the patent's IPC codes. The technological novelty can be measured by this distribution. Following the work by Uzzi et al. (2013), to characterize a patent's tendency of conventional and novel combinations of prior work, the median z_{score} and the 10th percentile z_{score} of the distribution are utilized to characterize the central tendency (conventionality) of a patent's combinations and the more unusual combinations (novelty). "Median conventionality" (whether the patent's median z_{score} is in the upper or lower half of all median z_{score}) and "tail novelty" (whether the patent's 10th percentile z_{score} is above or below zero) are used to classify patents into four types: high median conventionality/high tail novelty ($C + N +$); high median conventionality/low tail novelty ($C + N$-); low median conventionality/high tail novelty (C-$N +$); low median conventionality/low tail novelty (C-N-).

4. Experiments and analysis

4.1. Empirical data

To test the feasibility of the measure, we retrieved and used the patent data in the AI field. The data was obtained from the incoPat[1] scientific and technological innovation information platform, which is a retrieval system covering a large amount of patent information worldwide. According to the classification of artificial intelligence by Venture Scanner company (focusing on the emerging technology research) and China's "White Paper on Artificial Intelligence Standardization 2018", combined with the relevant literature research and search results test of keywords in the database, the search strategies were constructed as follows:

((TIAB = ("Artificial Intelligence*" OR "Machine Learning*" OR "Deep Learning*" OR "Knowledge Map*" OR" Natural language processing*" OR "Human-computer interaction*" OR "Computer vision*" OR" Biometrics Fingerprint Recognition*" OR" Fingerprint Recognition*" OR "Face Recognition*" OR "Speech Recognition*" OR "Image Recognition*" OR "Virtual reality*" OR "augmented reality*" OR "chip technology*" OR "Internet of Things*" OR "information retrieval and recommendation*" OR "Data Mining*" OR "Gesture control* " OR "smart robot*"))) NOT ((IPC-SUBCLASS=("H04M")) OR (IPC-SUBCLASS=("A61B")) OR (IPC-SUBCLASS=("G08G")) OR (IPC-SUBCLASS=("G05B")) OR (IPC-SUBCLASS=("G09B")) OR (IPC-SUBCLASS=("B60R"))).

The dataset contains 292,275 patents in AI, which was retrieved on Jan 14th, 2020. The trend of the number of patents is shown in Fig. 3. The patent data applied in 2018 and 2019 is incomplete due to the 18-month delay in publishing patent applications.

4.2. Illustrative study of technological novelty

A case study is used to illustrate the measurement and effectiveness of the measure. The selected patent ("US15038030") was applied by Facebook Corporation in November 2014, the title of which is "Control device for controlling a camera arrangement and a method for controlling an augmented reality application program of the camera arrangement". The exemplary patent has 12 IPC codes, therefore, there are 66 pairs of IPC combinations. Fig. 4 shows the distribution of the values of all IPC combinations. The median value of the distribution is −7.02, representing the conventional tendency of the IPC pairings, while the 10th percentile value is −8.83, which represents the novel tendency of the IPC pairings.

Table 1 shows the scores of some IPC pairings using the three indicators and the final z_{score}. The smaller the z_{score} is, the more novel of the IPC pairing is. For example, *A63F13/219* and *H04N5/232* pair, did not appear before 2014 (application year), therefore, $p_{occur_{2014}}(A63F13/219, H04N5/232)$ is zero. The classification code *A63F13/219* (used to target a specific area on the display, such as a light gun) belongs to Part A (required for human society), while *H04N5/232* (the device that controls the camera) belongs to Part H (electricity), making $p_{similarity}(A63F13/219, H04N5/232) = 0$. According to the

[1] https://www.incopat.com.

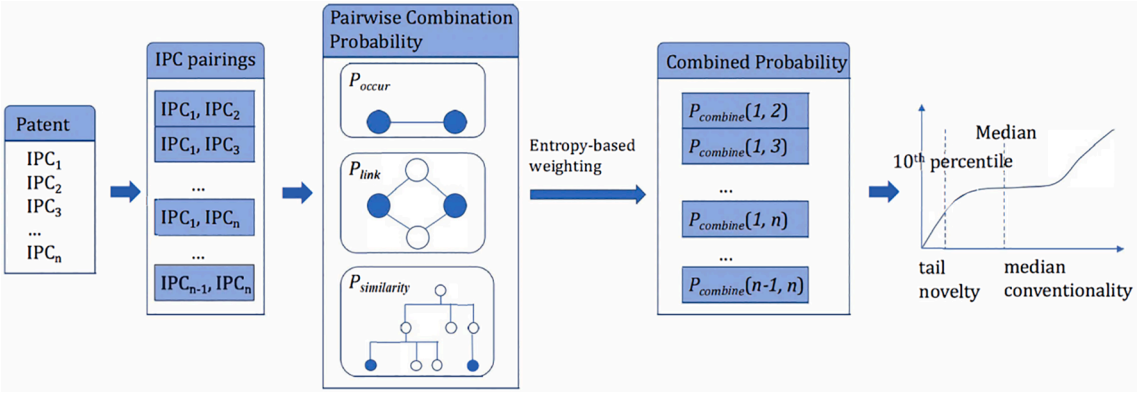

Fig. 2. The flow diagram.

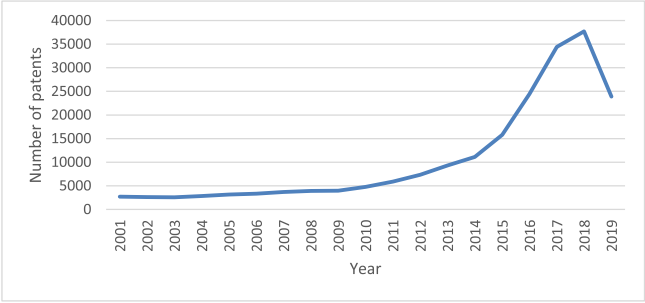

Fig. 3. The trend of the number of patents in the AI field.

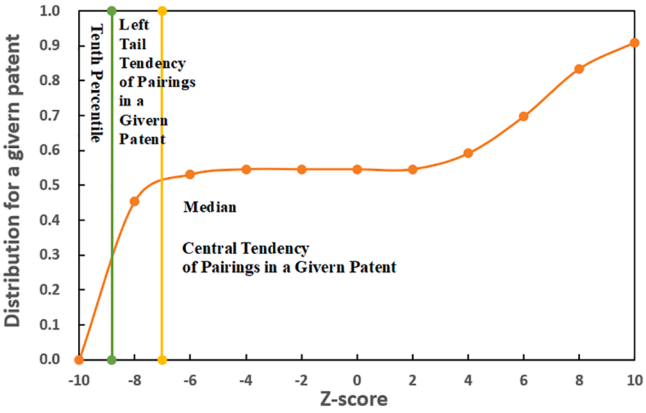

Fig. 4. Novelty and conventionality of a patent. The median value captures the central tendency in combining prior work, while the 10th percentile value captures the relatively unusual pairings.

previous IPC pairings, the link probability is calculated using link prediction algorithm $p_{link2014}(A63F13/219, H04N5/232) = 7.695$. In contrast, the larger the z_{score} is, the more conventional the combination of the IPC pairing is. For example, the *H04N5/225* and *H04N5/232* pair co-occurred very frequently, and the link possibility and hierarchical-level similarity are also very high, as *H04N5/225* (television camera) and *H04N5/232* (the device that controls the camera) are the children and grandchildren of *H04N5/222* (television system) in the IPC tree.

The advantage of this measure is the consideration of latent similarity between pairings. Let us compare the pairings *A63F13/219 - H04N5/232* and *G03B17/02 - G03B17/56*. If we only consider if the pairings appear or not before the application year, they all seem to be novel combinations. However, *G03B17/02* and *G03B17/56* are so close in the IPC tree and also have common neighbors they both have co-occurred with. The proposed measure could capture this latent distance between pairings, identifying more novel combinations.

In Fig. 5, we studied all the patents from the year 2001 to 2019. The scores were computed and plotted every 5 years in terms of median z_{score} and 10th percentile z_{score}. The distribution patterns changed little over the four periods. Compared with the study of the previous paper (Uzzi et al., 2013), in which fewer than 5 % of papers had median z_{score} below 0, the result with more than 33 % of patents had median z_{score} below zero, showed a relatively high proportion of novelty. If we used the 10th percentile z_{score}, 45 % of the patents in 2016–2019 and 55 % of the patents in 2006–2010 had scores below zero. Compared with previous results, technology in AI showed a relatively high proportion of novel combinations.

4.3. Evaluation of the measure

The proposed measure produced a ranking list of patents by tail novelty probability. As there was no ground-truth data about the novelty of patents, we applied an evaluation indicator AUC (area under the receiver operating characteristic curve) (Hanley & Mcneil, 1982) to evaluate the performance of the proposed measure. AUC does not focus on specific scores and only focuses on the ranking results, which makes it particularly suitable for the effect evaluation of ranking problems. It could measure how likely the predicted value of a randomly selected positive case is higher than that of a randomly selected negative case. The AUC was computed as $(n'+0.5n'')/n$, where n was the sample size, n' was the number of times the proposed measure was correct, and n'' was the number of times the proposed measure thought the compared two patents were equal in novelty but the truth was not. The higher AUC is over 0.5, the better the measure is.

Four researchers in the related field of artificial intelligence were invited to label the relative novelty of two patents by giving detailed information. For example, if two patents P_A and P_B published in the same year were randomly selected for comparison, the researchers need to judge the relative novelty of the two patents based on their own experience and relevant materials. If P_A was more novel than P_B, then (P_A, P_B) was labeled as 1; otherwise, it was labeled as -1; if P_A and P_B were equally novel, it was labeled as 0. The four researchers labeled the data separately. From 2001 to 2019, 50 pairs of patents were randomly selected each year, and a total of 950 samples were generated in the evaluation set. Four researchers labeled the relative novelty of each pair as $-1,0,1$. The agreement between the two researchers was computed by the proportion of the same labels in the results. If n is the number of times two researchers have the same labels for the pairs, the agreement between these two researchers should be $n/950$. All the agreements between pairs of researchers were over 85 %. If there was any inconsistency, the result was determined by the majority of the labeled results.

If the relative novelty of (P_A, P_B) ranked by the proposed method is the same as the judgment of manual labeling, the method is correct. To

Table 1
Examples of IPC pairing scores for the illustrative patent.

IPC Pairings	p_{occur}	p_{link}	$p_{semantic}$	z_{score}	Meaning
A63F13/245 - G03B17/56	0	7.235	0	-8.852	
A63F13/2145 - G03B17/56	0	7.269	0	-8.842	
H04N5/225 - A63F13/245	0	7.607	0	-8.737	More Novel Combinations
A63F13/65 - G03B17/56	0	7.623	0	-8.732	
A63F13/219 - H04N5/232	0	7.695	0	-8.709	
G03B17/02 - G03B17/56	0	8.029	0.667	12.298	
A63F13/213 - A63F13/2145	7	7.569	0.667	13.207	
A63F13/213 - A63F13/219	8	7.598	0.750	15.980	More Conventional Combinations
A63F13/90 - A63F13/98	0	7.072	0.833	17.226	
H04N5/225 - H04N5/232	186	48.831	0.857	58.920	

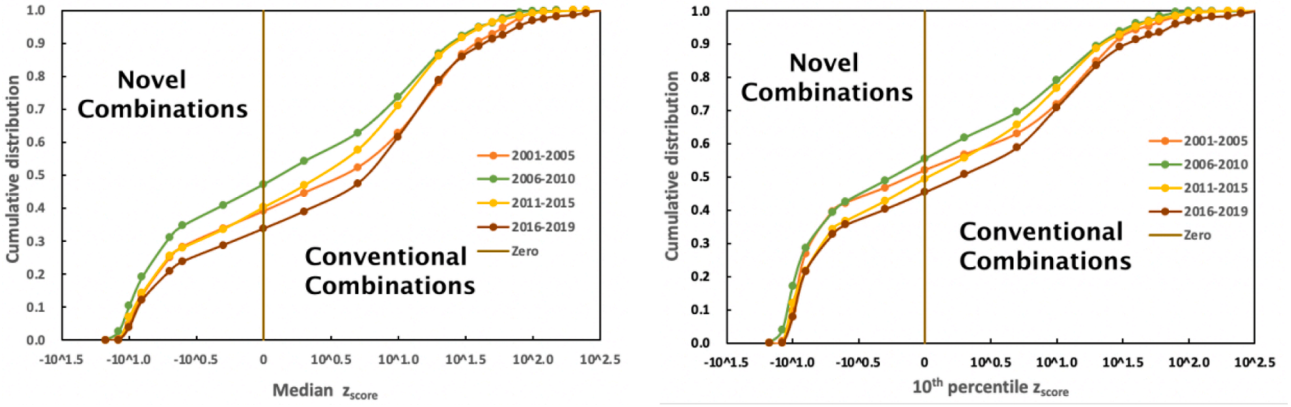

Fig. 5. The patents are studied every 5 years from 2001 to 2019. The median z_{score} and 10th percentile z_{score} of the patents are plotted respectively.

demonstrate the efficiency of the measure, the accuracies of three single indicators and the proposed measure are shown in Table 2. The Spearman pairwise correlations between them are also shown. The comprehensive measure is positively correlated with three single indicators, and the significance level is 0.01. The accuracy of the proposed measure is the highest, indicating that the integration of latent distance between knowledge components indeed has benefits in identifying novel patents.

4.4. Technological novelty and citations

Here we studied the correlation between novelty and citations which were usually used to represent impact.

To have an intuitive understanding, the top patents ranking by citations and novelty in 2001–2005 were listed in Tables 3 and 4. Table 3 showed the top ten patents ranking by citations. 8/10 of top patents were high novelty type. Table 4 showed the top ten patents ranking by novelty value. The citations of patents with top novelty value were not always ranking high, indicating that the novel patents had certain risks and uncertainty. The scatter plot of novelty and citations (Fig. 6), showed there was a positive relationship between the number of

Table 2
Spearman correlations and accuracies of the indicators.

Indicator	p_{occur}	p_{link}	$p_{semantic}$	$p_{combine}$	AUC
p_{occur}	1				0.852
p_{link}	0.595	1			0.877
$p_{similairy}$	0.353	0.245	1		0.868
$p_{combine}$	0.622	0.540	0.703	1	0.934

Table 3
Top ten patents ranking by citations from 2001 to 2005.

Ranking	Application number	IPC	Novelty Type	Citations
1	US09766560	*G06F15/173; G06F11/30; H04L12/24; H04L29/06*	*C-N+*	826
2	US09825152	*G06F15/18; G06F1/00; G06F1/16; G06F3/00; G06F3/01; G06F3/042; G06F17/30; G06F21/00*	*C + N-*	716
3	US10358759	*G06F7/00; G06F17/30; G06Q10/00*	*C + N-*	554
4	US10740242	*H04N5/232; H04N5/76; G06F17/30; H04N1/32; H04N5/765; H04N5/77; H04N5/781; H04N9/804; H04N9/82*	*C-N+*	530
5	US10029225	*A63F13/00; G07F17/32*	*C-N+*	530
6	US10106992	*A63F9/24; G06F19/00; G07F17/32*	*C-N+*	524
7	US09957673	*G06F9/445; G06F15/177*	*C-N+*	522
8	US10519818	*H04N5/225; G06K9/00; G06T7/20; H04N5/262; H04N7/00*	*C-N+*	514
9	US10269050	*G06F17/60; G06Q30/00; G06T11/20*	*C-N+*	513
10	US10783378	*H04N5/232; G06F17/30; H04N1/00; H04N1/32*	*C-N+*	489

Table 4
Top ten novel patents from 2001 to 2005.

Ranking	Application number	IPC	Novelty Value	Citations
1	CN03808908.4	*G10L19/02; H04S3/00*	−11.526	41
2	JP2003407410	*G10L15/28; B66C13/00; G10L15/00*	−11.438	4
3	JP2003333628	*G06T17/40; A63G31/00; H04N13/00*	−11.353	12
4	CN03140255.0	*H05B37/02; G06F15/00*	−11.330	13
5	CN03808938.6	*G10L21/00; G10L17/00; H04K1/00*	−11.275	37
6	US10612781	*G06K9/62; H04N7/18; H04Q1/00; G06K9/68*	−11.258	103
7	US10644270	*G06F15/16; G06K9/00; G07C11/00; H04L12/58; H04L29/00*	−11.070	102
8	US10334989	*G10L21/00; H04R1/10; G10L15/00; G10L15/26; G10L15/28*	−11.034	64
9	CN03807057.X	*G10L11/02; G10L21/02; H04R1/40; H04R1/46*	−10.977	20
10	CN03100436.9	*G10L19/00; H03M7/00*	−10.957	15

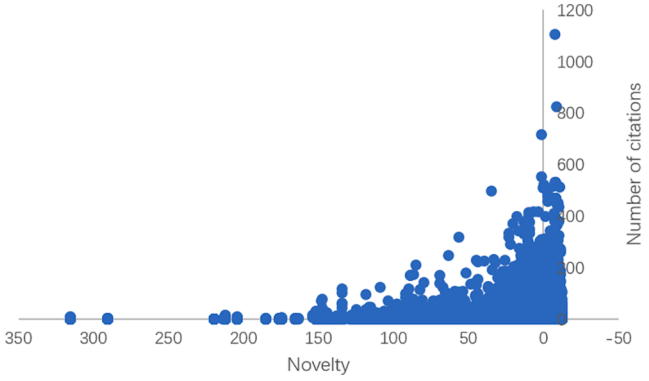

Fig. 6. The relationship between novelty and the number of citations.

citations and novelty.

4.4.1. Novelty vs Average citations

Fig. 7 showed the average number of citations of four novelty types in different periods. In the later three periods, no significant difference was found by comparing the average citations of different types of patents. From 2001 to 2005, the average number of citations of high conventionality/high novelty ($C+N+$) patents was the highest, and the average number of citations of low conventionality/low novelty (*C-N-*) patents was the lowest. Averaging the citations of all patents during 2001–2019 showed the same trend. The result here is consistent with prior studies, indicating high novelty combined with high conventionality could have a larger impact measured by average citations.

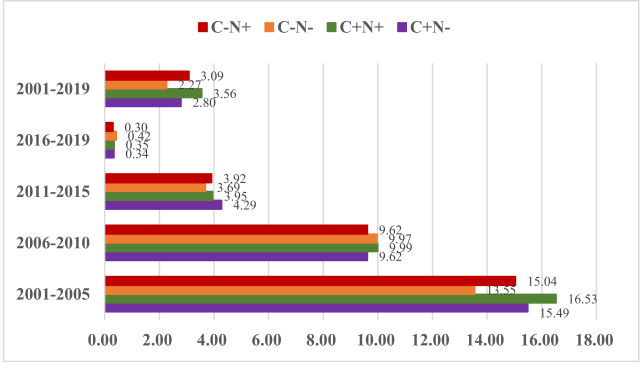

Fig. 7. Novelty type and the average number of citations.

4.4.2. Novelty vs Top citations

In addition to average citations, we also explored the top citations. As the citations of a patent in the database are the cumulative citations after the patent was published, the number of citations of a newly published patent is relatively lower. As shown in Fig. 8, the critical values of the top citations in each period were quite different. For example, during the period 2001 to 2005, patents with citations no less than 100 could be ranked in the top 1 %, while patents from 2016 to 2019 could be ranked in the top 1 % with citations no less than 3. Here, we studied the relationship between patent novelty and top citations in the two periods 2001–2005 and 2006–2010 when the citations of patents had already accumulated for a long time.

Fig. 9 showed the number of patents of four types in the two periods. The top 1 %, 5 %, and 10 % highly cited patents were selected in the two periods from 2001 to 2005 and 2006 to 2010 and the relationship between citations and novelty/conventionality of the patents was analyzed, as shown in Fig. 10. The "hit" probability was defined by the probability of a patent being in the top x% citations conditional on four types. For example, the "hit" probability of type C + N-, was computed by the number of patents of type C + N- in top x% divided by the total number of patents of type C + N-. This figure showed broadly consistent patterns both over time and "hit" patents. Specifically, the figure showed that high conventionality combined with high novelty ($C+N+$) outperformed the other categories, regardless of whether "hit" patents were defined as top 1 %, 5 %, or 10 % by citations. One exception was *C-N-* type in the 2006–2010 period had a slightly higher probability than others.

Overall, the finding is consistent with the prior studies, showing that high tail novelty combined with high median conventionality ($C+N+$) is required to be a "hit" patent.

5. Conclusion and future work

In this paper, we proposed a new method to measure the novelty of technology from the knowledge combination perspective, in which the co-occurrence relationship, the link probability, and the hierarchical-level relationship between knowledge components were considered. The novelty and conventionality of a patent were measured by the 10th percentile z_{score} and the median z_{score} of the distribution of IPC combination values. The method was validated using the patents in the AI field. We randomly selected 950 pairs of patents and labeled the relative novelty. According to the labeled results, the predicted accuracy using the AUC measure was 93.4 %, showing that the integration of different aspects of knowledge combination indeed had benefits in identifying novel patents compared with three single indicators.

The research about the relationship between patent novelty and citations showed that novelty and citations had a positive relationship. The results about the relationship between the average number of

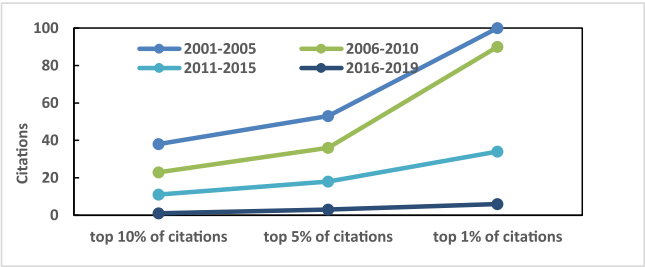

Fig. 8. The critical values of the top *x*% citations in each period. The patents that have a higher number of citations than the critical values of the top *x*% citations belong to the top citations.

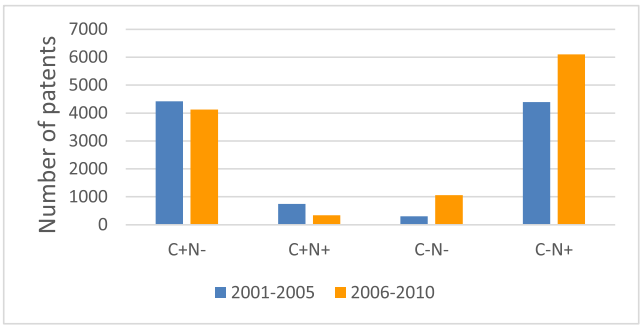

Fig. 9. The number of patents of four types in the two periods.

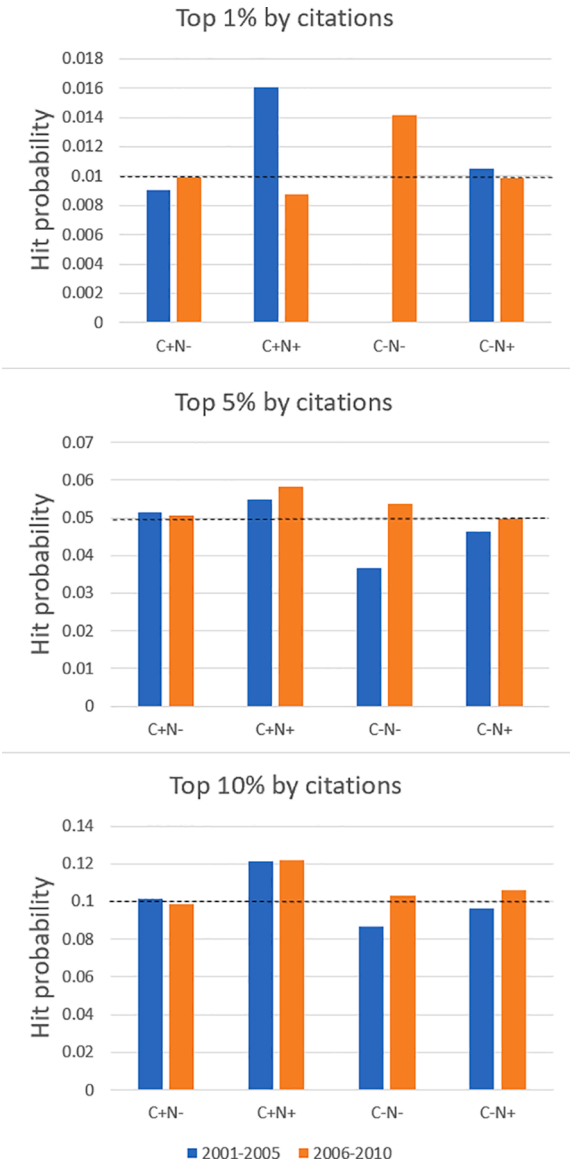

Fig. 10. The probability of being a "hit" patent of different types in each period. The dashed line represents the expected hit probability.

citations and novelty type, and the probability of being a "hit" patent of different types, were consistent with prior studies. The patents of high novelty combined with high conventionality ($C+N+$) had a larger impact measured by citations. As the impact reflected by citations needs a long time to accumulate, only the ex-post indicators (such as the number of citations) are not appropriate for identifying novel patents. The ex-ante indicators such as the proposed measure in this paper could discover novel patents in advance to monitor potential novelty in time.

The research on the identification of technological novelty has important practical significance: Firstly, for science and technology novelty searchers, the measure could help them to avoid the subjectivity and non-comprehension of manual retrieval to judge the novelty of patents; Secondly, it has a certain guiding role for the research strategy of the R&D team. When conducting the research of cutting-edge technologies, it is necessary to consciously evaluate the novelty and conventionality of the knowledge combination and increase the possibility of breakthrough technologies; Thirdly, patent inventors, when researching and writing patent documents, should focus on combining conventionality and novelty of knowledge to help generate high-impact technologies.

There are also some limitations and future directions of this study. Firstly, the patents of the AI field were used in this paper, and whether the findings still hold for other fields is not clear. Future studies will apply the indicator to other fields to validate the results and demonstrate the feasibility and universality of the method. Secondly, the evaluation method was a compromise choice due to the lack of ground-truth data. Better-designed validation approach will be explored in future work. Thirdly, the proposed method still has several directions for further improvement. For example, instead of using IPC as a proxy for the knowledge components of technology, the content analysis of technology is also very worthy of study. Technological novelty may also come from scientific knowledge, which could be an influencing factor when identifying novelty. As novelty is one of the necessary conditions for breakthrough innovation, based on the measurement of technological novelty in this paper, new relevant indicators could be built to detect emerging technological breakthroughs that have a significant impact on scientific and technological development.

CRediT authorship contribution statement

Xiaoling Sun: Conceptualization, Methodology, Writing – original draft, Supervision. **Na Chen:** Methodology, Data curation, Investigation. **Kun Ding:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is partially supported by grant from National Natural Science Foundation of China (No. 71704019) and the Planning Fund for Liaoning Social Science (No. L17CGL009).

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks, 25* (3), 211–230.
Arthur, W. B. (2007). The structure of invention. *Research Policy, 36*(2), 274–287.
Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics, 42*, 527–554.
Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics, 13* (4), Article 100979.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science, 62*(10), 2765–2783.
Boyack, K. W., & Klavans, R. (2014). Atypical combinations are confounded by disciplinary effects. *Proceedings of the Science and Technology Indicators Conference 2014 Leiden "Context Counts: Pathways to Master Big and Little Data"*.
Carayol, N., Llopis, O., & Lahatte, A. (2016). Capturing scientific novelty through paper keyword combinations. *Proceedings of the 21ST International Conference on Science and Technology Indicator (STI)*.
Chai, S., & Menon, A. (2019). Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy, 48*, 733–747.
Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), Article 117132.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 0003122415601618.
Gruber, M., Harhoff, D., & Hoisl, K. (2013). Knowledge recombination across technological boundaries: Scientists vs Engineers. *Management Science, 59*(4), 837–851.
Hanley, J. A., & Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.
Harrigan, K. R., Di Guardo, M. C., Marku, E., & Velez, B. N. (2017). Using a distance measure to operationalise patent originality. *Technology Analysis & Strategic Management, 29*, 988–1001.
He, Y., & Luo, J. (2017). The novelty 'sweet spot' of invention. *Design Science, 3*, E21.
Hofstra, B., Kulkarni, V. V., Galvez, M. N., He, B., Jurafsky, D., & Mcfarland, D. A. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences, 117*(17), 9284–9291.
Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.
Kim, D., Cerigo, D. B., Jeong, H., & Youn, H. (2016). Technological novelty profile and invention's future impact. *EPJ Data Science, 5*, 8.
Lee, Y. N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(3), 684–697.
Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.
Liu, X., Zhou, Y., & Zheng, R. (2007). Measuring Semantic Similarity in Wordnet. *International Conference on Machine Learning and Cybernetics*, 3431–3435.
Luo, J., Sarica, S., & Wood, K. L. (2021). Guiding data-driven design ideation by knowledge distance. *Knowledge-Based Systems, 218*, Article 106873.
McNamee, R. C. (2013). Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy, 42*(4), 855–873.
Schilling, M. A., & Green, E. (2011). Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy, 40*(10), 1321–1331.
Shannon, C. E. (1950). The mathematical theory of communication. 1963. *Bell Labs Technical Journal, 3*(9), 31–32.
Shibayama, S., & Wang, J. (2020). Measuring originality in science. *Scientometrics, 122*, 409–427.
Silva, S. R., Vieira, T., Martínez, D., & Paiva, A. (2021). On novelty detection for multi-class classification using non-linear metric learning. *Expert Systems with Applications, 167*, Article 114193.
Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature, 544*(7651), 411–412.
Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy, 44*(8), 1445–1461.
Sun, X., Lin, H., Xu, K., & Ding, K. (2015). How we collaborate: Characterizing, modeling and predicting scientific collaborations. *Scientometrics, 104*(1), 43–60.
Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics, 12*(3), 906–930.
Tsai, F. S., & Kwee, A. T. (2011). Experiments in term weighting for novelty mining. *Expert Systems with Applications, 38*(11), 14094–14101.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472.
Valverde, S., Solé, R. V., Bedau, M. A., & Packard, N. (2007). Topology and evolution of technology innovation networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 76*, Article 056118.
Van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics, 59*(3), 467–472.
Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy, 45*(3), 707–723.
Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy, 48*(6), 1362–1372.
Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy, 48*, 1260–1270.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436.
Youn, H., Strumsky, D., Bettencourt, L. M. A., & Lobo, J. (2015). Invention as a combinatorial process: Evidence from US patents. *Journal of the Royal Society Interface, 12*(106), 20150272.
Zhang, Y., Qian, Y., Huang, Y., Guo, Y., Zhang, G., & Lu, J. (2017). An entropy-based indicator system for measuring the potential of patents in technological innovation: Rejecting moderation. *Scientometrics, 111*(3), 1925–1946.