Characterizing patent big data upon IPC: a survey of triadic patent families and PCT applications

Jewel X. Zhu[1,2], Minghan Sun[3], Shelia X. Wei[1,2*] and Fred Y. Ye[1,2*]

*Correspondence:
151070078@smail.nju.edu.cn;
yye@nju.edu.cn

[1] School of Information
Management, Nanjing University,
Nanjing 210023, China
[2] Jiangsu Key Laboratory of Data
Engineering and Knowledge
Service and International Joint
Informatics Laboratory, Nanjing
University–University of Illinois,
Nanjing 210023, China
[3] School of Intellectual Property,
Nanjing University of Science
and Technology, Nanjing 210094,
China

Abstract

Research objective: Triadic patent (TP) families and Patent Cooperation Treaty (PCT) applications are often used as datasets to measure innovation capability or R&D internationalization, but their concordance is unclear, which is the main issue in this study.

Methods: We collect the global TP and PCT data from the Derwent Innovations Index (DII), and a total of 1,589,172 TP families and 4,067,389 PCT applications are retrieved. Based on International Patent Classification (IPC) codes, we compare these two big datasets in three parts: IPC distribution, IPC co-occurrence network, and nation-IPC co-occurrence network. In order to understand the overall similarities and differences between TP and PCT, we make the basic statistics of the global data and w-core defined based on the w-index. Furthermore, the w-cores are visualized and the global similarities are calculated for the detailed concordance and differences.

Findings: The result shows that the w-core is suitable to select the core part of big data and TP and PCT get high concordance. Meanwhile, in technological convergence, some specific technical fields (e.g. chemistry, medicine, electronic communication, and lighting technology) and countries/regions (e.g. Germany, Japan, China, and Korea), there are a few differences.

Practical implications: TP families are very similar to PCT applications in terms of reflecting innovation capability or R&D internationalization at a macro level, but when it comes to technological convergence, specific research topics, and countries/regions, the choice may depend on the purpose of the research.

Keywords: Triadic patent families, PCT applications, IPC, Patent statistics, Patentometrics

JEL Classification: O32, O33, O34

Introduction

Patents, which contain 90–95% of the global technical information, represent valuable technical inventions and provide academia and industry with a reliable basis. Compared with other technical documents, patents are more authoritative and up-to-date. A large number of researchers have already used patent data to analyze current and future technological trends. However, with the explosive growth of patents and the massive

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 2 of 17

influx of low-quality patents, the number of patents is no longer an effective measure to investigate the state of innovation and trends in technologies or industries, so researchers have begun to look for some appropriate indicators that represent high-quality patents, where the number of triadic patent (TP) families or the Patent Cooperation Treaty (PCT) applications is frequently used.

The triadic patent (TP) families refer to a set of patents filed at three major patent offices, namely the European Patent Office (EPO), the Japan Patent Office (JPO), and the United States Patent and Trademark Office (USPTO) [1].

Meanwhile, the Patent Cooperation Treaty (PCT) is an international treaty with more than 150 Contracting States. It is possible for an invention to seek patent protection in plenty of countries at the same time by submitting a single "international" patent application via the PCT rather than several separate national or regional patent applications. The granting of patents remains under the control of the national or regional patent offices in what is called the "national phase" [2].

As cross-border patent applications, TP families and PCT applications are important datasets to investigate national or regional innovation capabilities, evaluate industrial development status, and measure cross-border knowledge flow, whether in working papers and reports [3–7] or journal papers [8–10]. On the one hand, although there are some studies to choose TP families or PCT applications as datasets, these studies only focused on a part of PCT and TP applications, such as some patents related to a specific topic or applied for at a certain period. Therefore, in this study, we intend to collect and investigate the global TP families and PCT applications with a million-level volume. On the other hand, there does not exist paper to compare TP families and PCT applications, so it is worth knowing if the TP families and PCT applications get concordance. In a word, we propose to quantitatively explore the TP families and PCT applications based on the global data and understand their concordance from a global perspective in this study.

Literature review

In this section, we review some studies about three aspects, namely TP families and PCT applications, IPC co-occurrence network and nation-IPC co-occurrence network, where the nation refers to the earliest priority country or region, and the h-index and w-index, to understand the current research situation and research gap.

TP families and PCT applications

Patent applications were considered to have the inclination that applicants tend to file patents in their home country's patent office, which is called "home advantage bias" [11]. As multinational applications, TP families were able to balance the home advantage of domestic applicants/inventors in the 1990s [12], so as to more objectively show the innovation strength of a country or a region. After examining the extent of the 'home advantage' effect in the USPTO and the EPO patent data and the TP families, there was a conclusion that TP families could be used as a satisfactory alternative to the USPTO and the EPO for measuring R&D internationalization [13]. On this basis, many papers have conducted empirical studies on TP as an innovation dataset [14–20]. Tahmooresnejad and Beaudry studied the relationship between the structure and characteristics of TP families and patent value, and

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 3 of 17

believed that the structure and characteristics of the patent families played an important role in explaining the high value of patents [21].

As is a key indicator of technological and innovative strength, the number of TP families per country was a function of technological specialization and (national) patenting strategies [22]. Based on TP families, the potential future convergences among technologies can be predicted by using Adamic/Adar similarity between IPC codes [23]. It was also proved that international filings, especially TP, were important to capture variations in research productivity [24]. Recently, the number of TP has continued to be an important indicator for measuring innovation. The registration of TP families was used as an innovation output variable along with the number of research article citations and patent citations to measure knowledge spillover efficiency [25]. Sun et al. used the TP database for 24 innovating countries between the years 1994 and 2013 to investigate the effects of technological innovation within certain countries on the energy efficiency performance of neighboring countries [26]. The number of TP families was selected as the output variable to analyze the relationship between regulation and R&D efficiency [8]. Higham et al. linked citation network layers through TP families and observed that these layers contain complementary, rather than redundant, information about technological relationships [27]. Wei et al. combined TP families and technology life cycle theory to define the grey-rhino model [10].

Similar to TP families, PCT applications were often used to measure innovation output [28–30], innovation capability [31, 32] and international knowledge diffusion [33]. As early as 2008, based on the 138,751 patents filed in 2006 under the PCT, Leydesdorff used IPC codes to analyze the relations among technologies at different levels of aggregation [34]. As a representative of patent activities, PCT applications were also used to study the technological growth of countries [35] or the development of the industry [36, 37], etc. By combining patent data from PCT and EPO, Kers studied trends in genetic patent applications in order to identify the trends in the commercialization of research findings in genetics [38]. The participation of PCT applications in patent portfolios and a country's degree of concentration of PCT application filings were used to evaluate the commercial potential of university patenting [39]. Schmoch analyzed China's technological performance based on the transfer of China's PCT applications [9]. Roszko-Wojtowicz et al. adopted PCT applications per billion GDP as one of the variables to describe the effects of innovative activity [40]. Based on the case of Siemens' PCT applications, Ervits utilized the revealed technological advantage (RTA) index to measure the extent of the technological diversification of patent output [41].

In general, there have been many studies based on TP families or PCT applications in recent years, but there is no paper to compare these two datasets from the global perspective. Hence, we focus on the issue of shaping the relations between the global TP families and PCT applications to know how to profile the TP families and PCT applications and whether they get concordance or non-concordance.

IPC co-occurrence network and nation-IPC co-occurrence network

Compared with simple quantitative statistical analysis, patent network analysis can provide more comprehensive, objective and accurate technical intelligence for the management of research and development activities [42].

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 4 of 17

Patent network analysis can not only show the technical relationship between research subjects such as patents, enterprises, technical fields, countries or regions [43, 44], but also present the knowledge exchange [45], technical cooperation [46, 47], the knowledge maps [48] and technology development trends [49, 50]. In addition, the patent network provided clear data insights for comparative studies of different patent databases [51].

Furthermore, patent networks can be shown as one-mode, two-mode or even higher-mode. One-mode patent networks only include similar entities, such as IPC co-occurrence networks. When applying for a patent, the IPC codes [2] of the technical field corresponding to the patent are given. The structure of the IPC is divided into eight sections, and each section is subdivided into class, subclass, group, and subgroup [52]. A single patent can be granted multiple IPC codes. IPC co-occurrences network analysis was used to identify the convergence of technologies [53, 54], or to predict the pattern of technological convergence [23]. Two- and higher-mode patent networks include different sets of entities, and due to such unique feature, the two-mode network was essential to analyze the links among two disjoint node sets [45, 55, 56, 57]. The nation-IPC two-mode network that combines IPC information with the source country/region information of the patent was effective to identify the technological advantages of different countries/regions [58, 59].

In addition to visualization, network analysis provides rich quantitative indicators for patent comparative analysis, including measures of nodes and links within a network and inter-network similarity such as cosine similarity [60].

The h-index and w-index

The h-index is an index proposed by Hirsch [61] to evaluate the academic influence of scholars [61], which is defined as: A scientist has index h of his or her N_p papers have at least h citation each and the other $\left(N_p - h\right)$ papers have $\leq h$ citations each. The core part intercepted according to the h-index is called h-core [62], and each paper in h-core has at least h citations [63]. There are two main reasons why the h-index is popular. On the one hand, the h-index has the advantages of simplicity and stability. On the other hand, it can accurately grasp the common power-law phenomenon in informatics [64], naturally intercept the top data, and comprehensively balance quantity and influence [65, 66]. Now, the h-index has fully entered the research and application of academic evaluation, information measurement and other fields [14, 15, 66, 68, 69, 70]. The h-index was also introduced into the network node measure [71], and soon gained wide application [72, 73]. As links began to be recognized as playing a key role in the network [74], researchers found that the h-index, as the most characteristic method for extracting top information, was very suitable for measuring high-strength important links in the network, and h-strength (h_s) came into being. Its definition is as follows: the h-strength of a network is equal to h_s, if h_s is the largest natural number such that there are h_s links each with strength at least equal to h_s in the network [75]. The h-strength can significantly simplify complex networks and effectively select the main link structures. However, the h-index and h_s are powerless when extracting core information within very large-scale data and networks, and then the w-index and the generalized w-index were proposed.

The w-index is an improvement on the h-index [76], which focuses more on the evaluation of researchers' high-impact papers than the h-index. It can be defined as follows: If

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 5 of 17

w of a research's papers have at least $10w$ citations each and the other papers have fewer than $10(w + 1)$ citations, his/her w-index is w. On this basis, Egghe expanded 10 in the w-index to any natural number greater than or equal to 1 and proposed the generalized w-index (w_a) in 2011 [77]. When $a = 1, w_a = h$. For the same data set, the larger a is, the smaller w_a is, and the corresponding value of the w_a th source is larger. That is to say, the generalized w-index pays more attention to the top data than the h-index, and it can extract an appropriate level of core especially when faced with huge data. Then, if we combine the generalized w-index with h-strength, we can select a suitable core network from the network of large-scale data.

Methodology

Methods and data applied in this paper are displayed as follows.

Method

We compare TP and PCT in the following three parts: IPC distribution, IPC co-occurrence network and nation-IPC co-occurrence network, where the nation refers to the earliest priority country or region. We propose to use the generalized w-index to extract the core part of datasets. There are three main reasons why we choose the generalized w-index. Firstly, given that the TP and PCT datasets are very large, we deem that it is necessary to focus on the core part. Secondly, although the h-index is very famous and popular, the w-index is more suitable for big datasets because the constant a can be adjusted. Finally, the generalized w-index considers two important aspects of datasets, namely the number of sources (including IPC categories, IPC-IPC links, and Nation-IPC links) and the number of items for each source (see below for detailed representations).

Specifically, we define the w-core based on the generalized w-index.

The generalized w-index, denoted w_a, for $a \geq 1$ is the largest rank r $= w_a$, such that all sources on rank 1, …, r all have at least aw_a items. Following the concept of the generalized w-index, we introduce a new definition of w-core.

Definition   (*w-core*) A set of sources is divided into two groups by the generalized w-index. The first group with w sources each having at least aw_a items is w-core, and the rest of the sources, each having less than aw_a items, is w-tail. If there exists w-core as a subnetwork, we directly call it a w-core network. When the networks change among citation network, co-citation network, co-occurrence network and so on, the w-core can be extended to various w-cores.

In this paper, the w-index is applied to IPC distribution and co-occurrence networks to extract the w-cores. In the part of IPC distribution, an IPC category is a source and patents corresponding to this IPC category are items of this IPC category. In the part of IPC co-occurrence network, an IPC-IPC link is a source, and patents in which these two IPC categories co-occur are items of this IPC-IPC link. The sources and items of nation-IPC co-occurrence network are similar to IPC co-occurrence network. The detailed operation is as follows: first, for the IPC distribution, all IPC categories are sorted in descending order by the number of items in each IPC category. Similarly, for the IPC co-occurrence network and nation-IPC co-occurrence network, all links are sorted in

descending order by the number of items in each link which is called the strength of links. Second, the maximum rank r is decided based on r $= w_a$, where the top r IPC categories or links have at least aw$_a$ items. The w-core consists of the top r IPC categories or links. The constant a depends on the volume of the dataset, and we can adjust the value of a to extract the w-core of IPC distribution or co-occurrence networks effectively.

Cosine similarity, which is a measure of similarity between two individuals using the cosine value of the angle between two vectors in vector space, is adopted to investigate the global situation. The value range of cosine similarity is $[-1, 1]$. The higher the cosine similarity, the more similar the two vectors become. When the value is 1, the angle between these two vectors is 0, which means these two vectors exactly coincide. The value of cosine similarity is independent of the length of the vector, and only related to the direction of the vector, so the disparity in the amount of TP families and PCT applications can be ignored.

Thus, for two n-dimensional vectors A and B, the cosine similarity between them is:

$$s(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{1}$$

In this study, we use cosine similarity to measure the global similarity of TP families and PCT applications in IPC distribution, IPC co-occurrence network and nation-IPC co-occurrence network. The TP and PCT are two vectors with the same dimensions. For three different parts, the dimensions of vectors are IPC categories, IPC-IPC links or nation-IPC links, and the values of dimensions are the number of patents in each IPC category or the strength of links. Then, the cosine similarity of TP and PCT can be calculated based on Eq. (1).

Data

All patent data in this study are retrieved from the Derwent Innovations Index (DII). This database is currently one of the most comprehensive databases of international patent information in the world, published by Thomson Derwent Publishing Company. Every week, 25,000 patent documents published by more than 40 countries, regions and patent organizations and 45,000 patent citations are included in the database. Derwent, a world-class large patent database, provides a standardized and reliable data source for large-scale patentometric research.

The search strategy of TP families is "PN=(US*) AND PN=(JP*) AND PN=(EP*)" and the search strategy of PCT applications is "PN=(WO*)". It should be noted that the PCT came into effect in 1978, so the earliest PCT application appeared in 1978, and there were not many TP families before 1978. Therefore, we limit the search time range to after 1978, and the retrieval date is October 1, 2021. A total of 1,589,172 TP families and 4,067,389 PCT families are retrieved, and the data volume of PCT applications is as high as 2.56 times that of TP families. Figure 1 shows the basic situation of the data.

In Fig. 1, the left part is the number of families of TP and PCT in every priority year. We can see that the number of PCT rises rapidly, while the number of TP rises relatively slowly and even shows a downward trend in recent years, which may be because the application process for TP is more complicated than that for PCT. The right part is the
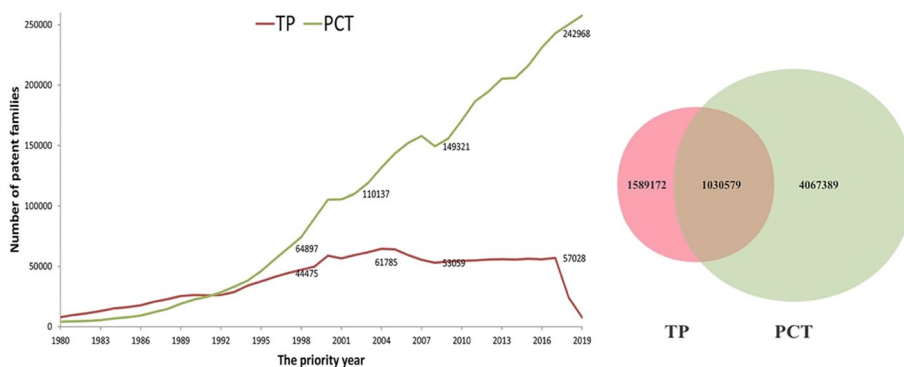
Zhu *et al. Journal of Big Data* (2023) 10:85

Page 7 of 17

Fig. 1 The basic situation of data

Venn diagram of TP and PCT, and they share 1,030,579 patent families which account for 64.85% of TP, 25.34% of PCT, and 22.28% of their union. It can be seen that the degree of overlap between TP and PCT is relatively high.

Furthermore, the broad flowchart of research is shown in Fig. 2. In the next section, we present the basic statistics of the global data and w-core, visualize the w-core and calculate the global similarity.

Results and discussion

The results are also divided into three parts, namely the IPC distribution, IPC co-occurrence networks, and nation-IPC co-occurrence networks. In the three parts, we will discuss the w-cores and global situations respectively.

As the quantities of both TP and PCT exceed one million, after repeated testing, it is found that the appropriate w-cores can be selected when $a = 100$. In order to understand overall similarities and differences between PCT and TP, the basic statistics of global data and w-cores are shown in Table 1, which includes the average, standard deviation, minimum, median, maximum, quartile and the Spearman Correlation between PCT and TP. In Table 1, IPC means IPC distribution, Co-IPC is IPC co-occurrence network, and Nation-IPC is nation-IPC co-occurrence network. In addition, N indicates the sample size, and the value of N in w-cores also means the value of w_{100}.

As shown in Table 1, firstly, the values of these statistics indicators of PCT are all higher than those of TP, excluding the minimum and Q1 in global data, because the data volume of PCT is bigger than that of TP and PCT is more discrete than TP. Secondly, the values of minimum, Q1, median, and Q3 of three parts in global data are very small, which indicates that most IPC categories have a few patents and most links have weak strength. However, the values of those indicators in w-cores are much higher than those in the global data, which to some extent means the w-index and w-core can extract the core part of the global data. Thirdly, the three values of w_{100} of PCT are greater than that of TP, because PCT applications are much more than TP families. Finally, according to the Spearman Correlation, we find that PCT and TP have a strong positive correlation for either global data or w-cores.
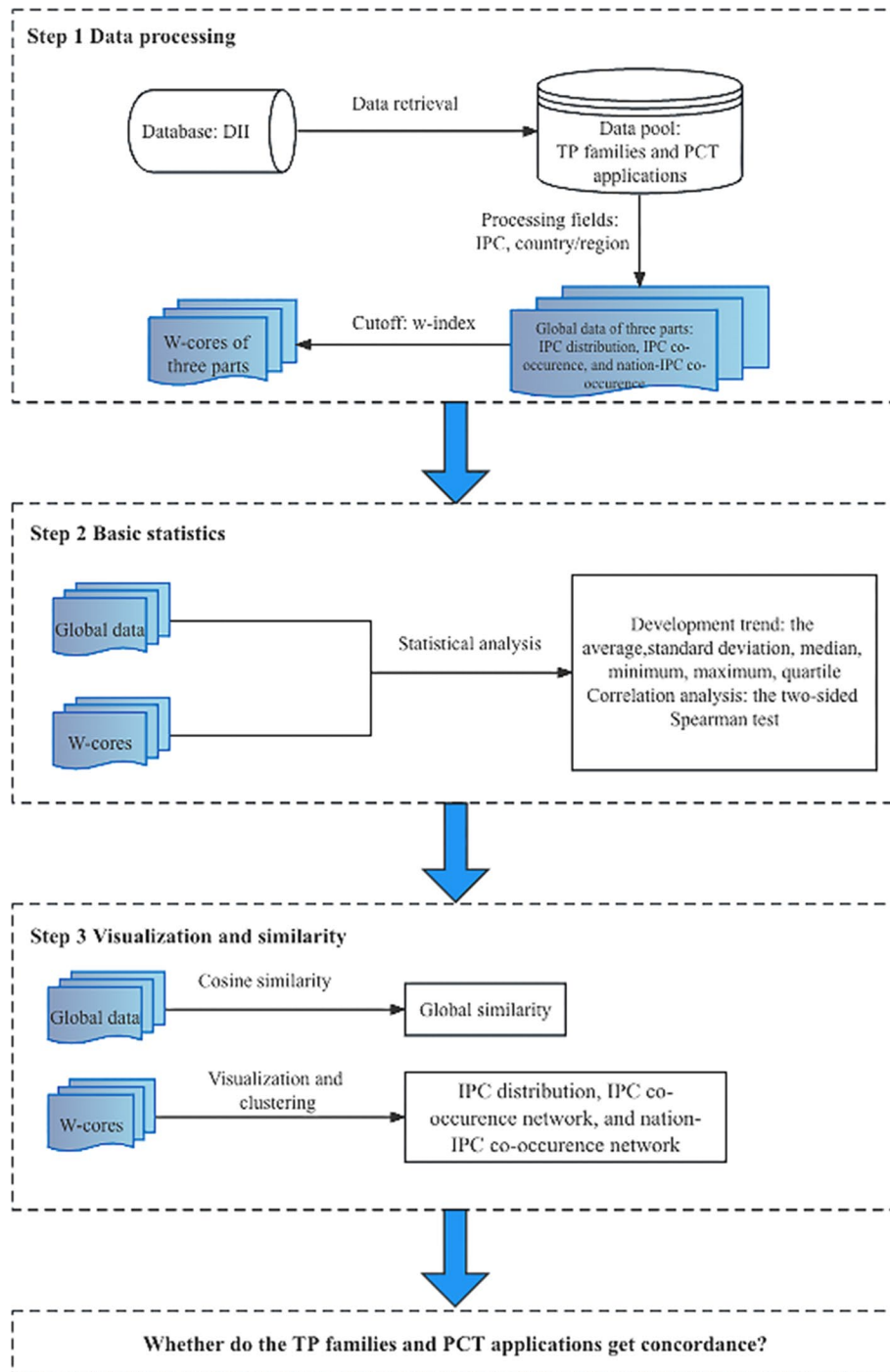
Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 8 of 17

Fig. 2 The flowchart of the research

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 9 of 17

Table 1 The basic statistics of global data and w-cores

Type	N	Min.	Q1	Med.	Q3	Max.	Avg.	Std.	Correl.
Global									
IPC									
PCT	2374	0	1	2	341.5	420,059	4382.39	20,831.09	0.838**
TP	2374	0	1	2	158	208,380	2172.52	10,131.21	
Co-IPC									
PCT	137,286	0	1	4	19	247,641	96.82	1298.37	0.860**
TP	137,286	0	1	2	11	135,449	61.57	769.09	
Nation-IPC									
PCT	36,610	0	1	6	46	203,726	284.08	2697.79	0.791**
TP	36,610	0	0	1	12	98,021	140.87	1342.38	
W-core									
IPC									
PCT	155	15,508	20,999	30,136	59,299	420,059	53,570.61	63,128.09	0.891**
TP	111	11,317	14,096	21,421	41,334	208,380	34,522.39	32,397.93	
Co-IPC									
PCT	125	12,533	15,886	20,315	32,214	247,641	30,234.93	27,866.84	0.852**
TP	101	10,182	12,547.5	15,603	23,955	135,449	21,012.50	15,939.64	
Nation-IPC									
PCT	123	12,415	14,912	20,212	35,365	203,726	32,188.28	31,223.67	0.813**
TP	91	9321	12,189	15,525	24,403	98,021	20,717.88	14,454.40	

The Correl. is the correlation coefficient between PCT and TP, derived from a two-sided Spearman test

*$p < 0.05$; **$p < 0.01$. The correlation between the w-cores of PCT and TP is calculated based on the overlap of two w-cores

The basic statistics present the overall situation, while detailed information of PCT and TP needs to be further shown. Hence, in the following sections, we visualize the w-cores of PCT and TP and calculate the global similarity of the three parts to make sense of the specific similarities and differences.

IPC distribution

The w-cores of TP and PCT have 111 and 155 IPC categories respectively, and 107 IPC categories in the w-core of TP are included in the w-core of PCT. The 107 IPC categories shared by the w-cores of TP and PCT mainly distribute in the front of the w-core of PCT. 48 IPC categories only appear in the w-core of PCT because the data volume of PCT is larger and there are more patents belonging to each IPC category. Meanwhile, 4 IPC categories only appear in the w-core of TP. Actually, they also distribute in PCT, but they have not entered the w-core because of their relatively small numbers.

The overlap of w-cores of IPC distribution of TP and PCT is shown in Fig. 3. The vertical axis is the number of patents in each IPC category and the horizontal axis is the descending order of IPC categories of PCT. The green column is the IPC distribution of PCT, the red column is the IPC distribution of TP and the green line is the distribution of PCT* (see below).

According to Fig. 3, we know that the w-cores of IPC distribution of TP and PCT get high concordance. First, TP and PCT keep similar w-cores as shown in Fig. 3. Second, several IPC categories have a wealth of patents, such as G06F and A61K, while
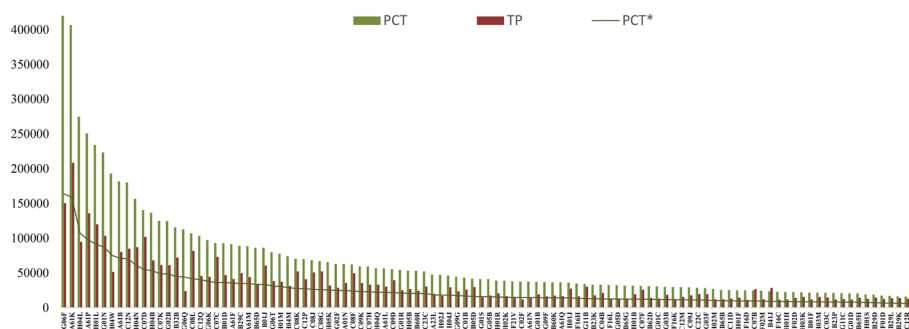
Fig. 3 The overlap of w-cores of IPC distribution of TP and PCT

the number of patents in most IPC categories is low relatively. Third, TP and PCT maintain similar distribution trends. In a lot of IPC categories, if the percentage of TP is high, that of PCT tends to be high. In addition, based on Eq. (1), we calculate the cosine similarity of the global IPC distribution of TP and PCT and the similarity is 0.968, which further indicates TP and PCT are alike.

However, a few differences exist. In all IPC categories in Fig. 3, PCT is higher than TP, because the data volume of PCT is much higher than that of TP, which is about 2.56 times the number of TP. Therefore, in order to make the comparison more intuitive, we divide the number of PCT applications in each IPC category by 2.56 to obtain PCT*, which can ignore the disparity in the number of TP and PCT. However, from Fig. 3 we can see that TP is always slightly higher than PCT*. The reason is the broader technical convergence of TP: each TP family has 3.24 IPC categories on average, while the average number of IPC categories in PCT is only 2.56, which is 0.79 times that of the former. When focusing on specific IPC categories, we find that there are still some differences between TP and PCT*. On the one hand, some categories of TP are much higher than PCT*, such as A61K (preparations for medical, dental, or toilet purposes), A61P (specific therapeutic activity of chemical compounds or medicinal preparations), C07D (heterocyclic compounds), C08L (compositions of macromolecular compounds), and C07C (acyclic or carbocyclic compounds), C07B (general methods of organic chemistry; apparatus therefor), B01J (chemical or physical processes, e.g. catalysis or colloid chemistry), C08F (macromolecular compounds obtained by reactions only involving carbon-to-carbon unsaturated bonds), which are related to chemistry and medicine. On the other hand, four categories of TP, which belong to electronic communication, are lower than PCT*. They are G06F (electric digital data processing), H04L (transmission of digital information), H04W (wireless communication networks) and G06K (recognition of data; presentation of data; record carriers; handling record carriers) respectively. In recent years, with the rapid development of electronic communication [77, 79, 80], the patents corresponding to these IPC categories seem to be more inclined to PCT, perhaps because PCT makes international patent applications faster and more convenient. All these differences are at the micro level, while the IPC distributions of TP and PCT are similar on the whole.

Zhu *et al. Journal of Big Data* (2023) 10:85

Page 11 of 17

Table 2 The basic data of the IPC co-occurrence network

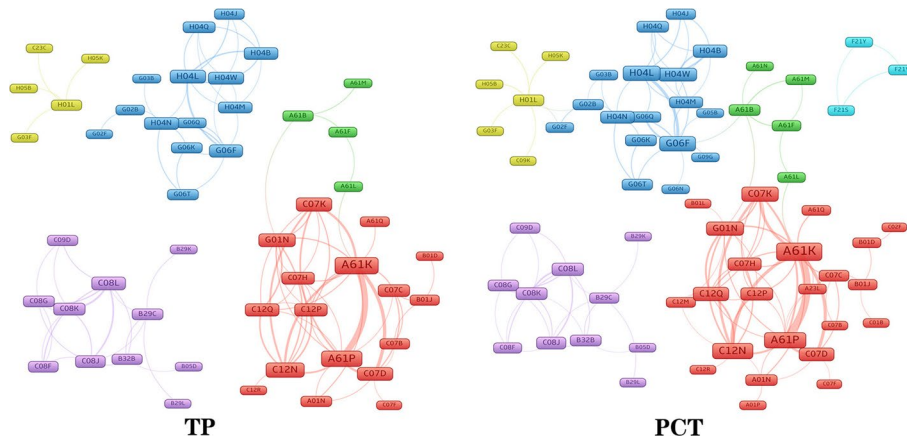| Co-IPC	Global			W-core		
	Nodes	Links	Frequency	Nodes	Links	Frequency
TP	2004	115,037	8,453,047	51 (2.54%)	101 (0.09%)	2,122,263 (25.11%)
PCT	2085	127,535	13,291,821	65 (3.12%)	125 (0.10%)	3,779,366 (28.43%)

Fig. 4 The w-cores of IPC co-occurrence networks of TP and PCT

IPC co-occurrence network

The basic data of the global network and the w-core of the IPC co-occurrence network are shown in Table 2.

In order to focus on the most important part of networks, Fig. 4 shows the w-cores of the IPC co-occurrence network of TP and PCT, where the rectangular box is the IPC category and different colors represent different clusters. The larger the rectangle box, the more times it co-occurs with other boxes. Similarly, if the link between two IPC categories is thick, they co-occur many times.

In Fig. 4, we can see that TP has five clusters and PCT has six clusters, but their clusters are very similar. For TP and PCT, the largest cluster is the red group represented by A61K, which is the field of medicine. The second largest cluster, colored blue, mainly includes H04W and H04L, which is communication technology. In addition, the purple group is chemical technology, electrical technology is represented by yellow and medical treatment and diagnosis technology is the green cluster which is closely linked to the red cluster. Furthermore, the cosine similarity of the global IPC co-occurrence networks of TP and PCT is 0.975, so they are highly similar in terms of IPC co-occurrence.

Nevertheless, there are also some differences. PCT has more nodes and its w-core network is more intensive than TP, which may be related to numerous PCT applications. The light blue cluster only appears on the right side of the PCT w-core network, including three IPC categories, namely F21Y (relating to the form or the kind of the light sources or the color of the light emitted), F21S (non-portable lighting devices; systems thereof; vehicle lighting devices specially adapted for vehicle exteriors) and F21V (functional features or details of lighting devices or systems thereof; structural combinations

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 12 of 17

Table 3 The basic data Nation-IPC co-occurrence network

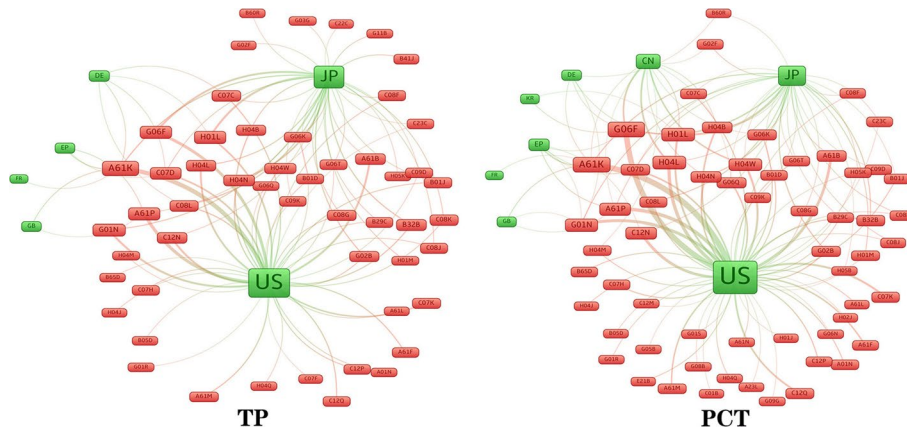| IPC-Nation	Global			W-core		
	Nodes	Links	Frequency	Nodes	Links	Frequency
TP	2110	23,837	5,157,334	58 (2.75%)	91 (0.38%)	1,885,327 (36.56%)
PCT	2228	54,550	10,400,329	53 (2.38%)	109 (0.20%)	2,941,705 (28.28%)

Fig. 5 The w-cores of nation-IPC co-occurrence networks of TP and PCT

of lighting devices with other articles). These IPC categories point to lighting technology, indicating that this technology is more inclined to PCT.

Nation-IPC co-occurrence network

The basic data of the global network and the w-core of the Nation-IPC co-occurrence network are shown in Table 3.

In the same way, Fig. 5 also displays the w-cores of the nation-IPC co-occurrence network of TP and PCT. The green boxes are countries or regions and the red boxes are IPC categories.

We find that the w-core of the nation-IPC co-occurrence network of TP is similar to that of PCT. In two subgraphs of Fig. 5, the applications of PCT and TP in the United States include the most IPC categories, which means patents from the United States involve wide fields at present. The second country is Japan, so its technical fields are broad too. In addition, two w-cores have some same countries or regions, namely Germany, Europe, France and Great Britain.

To compare the similarity of global nation-IPC co-occurrence networks of TP and PCT, we count the number of dimensions in the vector of some representative countries/regions in global networks, and calculate their cosine similarity. The results are presented in Table 4.

Generally speaking, whether these countries/regions or the whole network, their similarities in the TP and PCT are very high. Combined with Fig. 5 and Table 4, Japan and Germany deserve attention. Although Japan has high similarity (0.970) in the global networks of TP and PCT, Japan in the two w-core networks has some

Table 4 The similarity of five representative countries/regions in TP and PCT

Indicators	Global	US	JP	DE	EP	CN
The number of dimensions	36,610	1823	1235	1089	765	657
Similarity	0.935	0.972	0.970	0.892	0.987	0.978

differences. Japan has more IPC categories in the w-core of TP than that in the w-core of PCT. Contrarily, Germany has similar structures in two w-core networks, but its similarity of the global network is lower than that of other countries/regions.

However, like Fig. 4, the nodes of PCT are more and the w-core network is denser than that of TP. The reason should also be related to the large number of PCT applications. China and Korea only appear in the core network of PCT, so they tend to submit PCT applications.

In this section, we present the similarities and differences between TP families and PCT applications in terms of IPC distribution, IPC co-occurrence networks, and nation-IPC networks, based on three methods: statistical analysis, network visualization, and cosine similarity. We find that the w-core is suitable to select the core part of big data. The datasets of TP families and PCT applications are very similar in these three parts for either global data or w-cores, but there are some micro differences as said before. Thus, at a macro level, TP families and PCT applications get high concordance concerning their ability to reflect innovation capability or R&D internationalization, but when it comes to technological convergence, specific research topics and countries/regions, the choice may depend on the purpose of the research.

Conclusion and limitation

According to the above analysis, we have three main contributions. First, the w-core is a useful concept to characterize the core of important patents and patent networks. Second, we profile the w-cores and global situations of the TP families and PCT applications, and characterize their concordance from three parts, IPC distribution, IPC co-occurrence network and nation-IPC co-occurrence network respectively. Although the data volume of TP and PCT varies greatly, the results show that TP and PCT are very similar as a whole. Hence, if we want to observe the innovation capability, R&D internationalization, technical structure or development trend of a country/region or an industry, the analysis result based on TP is similar to PCT, which means TP and PCT can replace each other to a certain extent. Third, the TP and PCT are different in technological convergence, some specific fields (e.g. chemical, medicine, electronic communication and lighting technology) or countries/regions (e.g. Germany, Japan, China, and Korea), so that it is necessary to choose TP or PCT based on different research purposes.

The comparison between TP and PCT is still a relatively primary study, and there are certainly some limitations. Firstly, we simply use basic statistics and network visualization, but there are many different statistical methods and network indicators, such as regression, clustering and centrality, which can be used to further portray the TP families and PCT applications. Secondly, we characterize PCT and TP from three parts, the IPC distribution, IPC co-occurrence networks, and nation-IPC co-occurrence networks,

which only involve IPC and countries/regions of TP families and PCT applications. However, citations and contents of patents both play important roles in patent analysis, so we need to focus on diverse information about patents to answer if they are similar. Finally, because of delays in patent applications and publications [81], it is difficult to cover all TP families and PCT applications, especially in recent years. Generally speaking, we hope to be able to extend our study to patent citations and contents based on various statistical methods and network indicators to explore whether TP and PCT get concordance from different perspectives.

Abbreviations

TP	Triadic patent
PCT	Patent Cooperation Treaty
DII	Derwent Innovations Index
IPC	International Patent Classification
EPO	European Patent Office
JPO	Japan Patent Office
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization
OECD	Organization for Economic Cooperation and Development

Author contributions
JXZ collected and processed data and wrote the paper, MS assisted data processing, SXW wrote the paper, and FYY initiated the idea, designed the research and wrote the paper. All authors read and approved the final manuscript.

Availability of data and materials
The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

References
1. OECD. Triadic patent families (indicator); 2022b. Retrieved 28 March from https://data.oecd.org/rd/triadic-patent-families.htm.
2. WIPO. Protecting your inventions abroad: frequently asked questions about the patent cooperation treaty (PCT); 2020. Retrieved 28 March from https://www.wipo.int/pct/en/faqs/faqs.html.
3. OECD. Patents in environment-related technologies: technology diffusion and patent protection (Edition 2019); 2019. Retrieved 28 March from https://www.oecd-ilibrary.org/environment/data/oecd-environment-statistics/patents-in-environment-related-technologies-technology-diffusion-and-patent-protection-edition-2019_493d1053-en.
4. OECD. Main science and technology indicators; 2022a. Retrieved 28 March from https://www.oecd-ilibrary.org/science-and-technology/main-science-and-technology-indicators_2304277x.
5. WIPO. Global innovation index 2021, 14th edition tracking innovation through the COVID-19 crisis; 2021a. Retrieved 28 March from https://www.wipo.int/publications/en/details.jsp?id=4560.
6. WIPO. WIPO technology trends 2021 assistive technology; 2021b. Retrieved 28 March from https://www.wipo.int/publications/en/details.jsp?id=4541&plang=EN.

Zhu *et al. Journal of Big Data*   (2023) 10:85

Page 15 of 17

7.   WIPO. World intellectual property indicators 2021; 2021c. Retrieved 28 March from https://www.wipo.int/publicatio ns/en/details.jsp?id=4571.

8.   Nam M, Ko J, Lee J. Analysis of the relationship between regulation and R&D efficiency using quantile regression. In: International conference on big data and smart computing (BigComp); 2022, January 17–20, Daegu, South Korea.

9.   Schmoch U, Gehrke B. China's technological performance as reflected in patents. Scientometrics. 2022;127(1):299– 317. https://doi.org/10.1007/s11192-021-04193-6.

10.   Wei SX, Zhang HH, Wang HY, Ye FY. Identifying grey-rhino in eminent technologies via patent analysis. J Data Inf Sci. 2023. https://doi.org/10.2478/jdis-2023-0002.

11.   Dernis H, Khan M. Triadic patent families methodology; 2004. https://doi.org/10.1787/443844125004.

12.   Frietsch R, Schmoch U. Transnational patents and international markets. Scientometrics. 2010;82(1):185–200. https:// doi.org/10.1007/s11192-009-0082-2.

13.   Criscuolo P. The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. Scientometrics. 2006;66(1):23–41. https://doi.org/10.1007/s11192-006-0003-6.

14.   Chen DZ, Huang WT, Huang MH. Analyzing Taiwan's patenting performance: comparing US patents and triadic pat- ent families. Malays J Lib Inf Sci. 2014;19(1):51–70 (**<Go to ISI>://WOS:000331270100005**).

15.   Chen M, Mao SW, Liu YH. Big data: a survey. Mobile Netw Appl. 2014;19(2):171–209. https://doi.org/10.1007/ s11036-013-0489-0.

16.   Clark J, Huang HI, Walsh JP. A typology of 'innovation districts': what it means for regional resilience. Camb J Reg Econ Soc. 2010;3(1):121–37. https://doi.org/10.1093/cjres/rsp034.

17.   Ganda F. The impact of innovation and technology investments on carbon emissions in selected organisation for economic co-operation and development countries. J Clean Prod. 2019;217:469–83. https://doi.org/10.1016/j.jclep ro.2019.01.235.

18.   Kumazawa R, Gomis-Porqueras P. An empirical analysis of patents flows and R&D flows around the world. Appl Econ. 2012;44(36):4755–63. https://doi.org/10.1080/00036846.2010.528375.

19.   Luintel KB, Khan M. Heterogeneous ideas production and endogenous growth: an empirical investigation. Can J Econ Revue Can D Econ. 2009;42(3):1176–205. https://doi.org/10.1111/j.1540-5982.2009.01543.x.

20.   Wada T. Cognitive distances in prior art search by the triadic patent offices: empirical evidence from international search reports.proceedings of the international conference on scientometrics and informetrics. 15th International Conference of the International-Society-for-Scientometrics-and-Informetrics (ISSI) on Scientometrics and Informet- rics, Bogazici Univ, Istanbul, Turkey; 2015.

21.   Tahmooresnejad L, Beaudry C. Capturing the economic value of triadic patents. Scientometrics. 2019;118(1):127–57. https://doi.org/10.1007/s11192-018-2959-4.

22.   Sternitzke C. Defining triadic patent families as a measure of technological strength. Scientometrics. 2009;81(1):91– 109. https://doi.org/10.1007/s11192-009-1836-6.

23.   Lee WS, Han EJ, Sohn SY. Predicting the pattern of technology convergence using big-data technology on large- scale triadic patents. Technol Forec Soc Change. 2015;100:317–29. https://doi.org/10.1016/j.techfore.2015.07.022.

24.   de Rassenfosse G, de la Potterie BVP. A policy insight into the R&D-patent relationship. Res Policy. 2009;38(5):779–92. https://doi.org/10.1016/j.respol.2008.12.013.

25.   Bae J, Chung Y, Lee J, Seo H. Knowledge spillover efficiency of carbon capture, utilization, and storage technology: a comparison among countries. J Clean Prod. 2020;246:119003. https://doi.org/10.1016/j.jclepro.2019.119003.

26.   Sun HP, Edziah BK, Kporsu AK, Sarkodie SA, Taghizadeh-Hesary F. Energy efficiency: the role of technological innova- tion and knowledge spillover. Technol Forec Soc Change. 2021;167:120659. https://doi.org/10.1016/j.techfore.2021. 120659.

27.   Higham K, Contisciani M, De Bacco C. Multilayer patent citation networks: a comprehensive analytical framework for studying explicit technological relationships. Technol Forec Soc Change. 2022;179:121628. https://doi.org/10.1016/j. techfore.2022.121628.

28.   Barragan-Ocana A, Gomez-Viquez H, Merritt H, Oliver-Espinoza R. Promotion of technological development and determination of or biotechnology trends in five selected Latin American countries: an analysis based on PCT pat- ent applications. Electron J Biotechnol. 2019;37:41–6. https://doi.org/10.1016/j.ejbt.2018.10.004.

29.   Furkova A. Implementation of MGWR-SAR models for investigating a local particularity of European regional innova- tion processes. Central Eur J Oper Res. 2021. https://doi.org/10.1007/s10100-021-00764-3.

30.   Liu JP, Lu K, Cheng SX. International R&D spillovers and innovation efficiency. Sustainability. 2018;10(11):23. https:// doi.org/10.3390/su10113974. (**Article 3974**).

31.   Ervits I. Geography of corporate innovation: Internationalization of innovative activities by MNEs from developed and emerging markets. Multinatl Bus Rev. 2018;26(1):25–49. https://doi.org/10.1108/mbr-07-2017-0052.

32.   Murphy KJ, Elias G, Jaffer H, Mandani R. A study of inventiveness among society of interventional radiology mem- bers and the impact of their social networks. J Vasc Interv Radiol. 2013;24(7):931–7. https://doi.org/10.1016/j.jvir. 2013.03.033.

33.   Miguelez E, Temgoua CN. Inventor migration and knowledge flows: a two-way communication channel? Res Policy. 2020;49(9):13. https://doi.org/10.1016/j.respol.2019.103914. (**Article 103914**).

34.   Leydesdorff L. Patent classifications as indicators of intellectual organization. J Am Soc Inform Sci Technol. 2008;59(10):1582–97. https://doi.org/10.1002/asi.20814.

35.   Kumar R, Tripathi RC, Tiwari MD. A case study of impact of patenting in the current developing economies in Asia. Scientometrics. 2011;88(2):575–87. https://doi.org/10.1007/s11192-011-0405-y.

36.   Ardito L, D'Adda D, Petruzzelli AM. Mapping innovation dynamics in the Internet of Things domain: evidence from patent analysis. Technol Forecast Soc Chang. 2018;136:317–30. https://doi.org/10.1016/j.techfore.2017.04.022.

37.   Zhang F, Zhang X. Patent activity analysis of vibration-reduction control technology in high-speed railway vehicle systems in China. Scientometrics. 2014;100(3):723–40. https://doi.org/10.1007/s11192-014-1318-3.

38.   Kers JG, Van Burg E, Stoop T, Cornel MC. Trends in genetic patent applications: the commercialization of academic intellectual property. Eur J Hum Genet. 2014;22(10):1155–9. https://doi.org/10.1038/ejhg.2013.305.

39. Zdralek P, Stemberkova R, Matulova P, Maresova P, Kuca K. Commercial potential of university patents through patent cooperation treaty application. In: International conference on social sciences and humanities (SOSHUM), Kota Kinabalu, Malaysia; 2016, Apr 19–21.

40. Roszko-Wojtowicz E, Danska-Borsiak B, Grzelak MM, Plesniarska A. In search of key determinants of innovativeness in the regions of the Visegrad group countries. Oecon Copern. 2022;13(4):1015–5. https://doi.org/10.24136/oc.2022.029.

41. Ervits I. The effect of co-patenting as a form of knowledge meta-integration on technological differentiation at Siemens. Eur J Innov Manag. 2023. https://doi.org/10.1108/ejim-11-2022-0605.

42. Albino V, Ardito L, Dangelico RM, Messeni Petruzzelli A. Understanding the development trends of low-carbon energy technologies: a patent analysis. Appl Energy. 2014;135:836–54. https://doi.org/10.1016/j.apenergy.2014.08.012.

43. Sternitzke C, Bartkowski A, Schramm R. Visualizing patent statistics by means of social network analysis tools. World Patent Inf. 2008;30(2):115–31. https://doi.org/10.1016/j.wpi.2007.08.003.

44. Van Der Valk T, Gijsbers G. The use of social network analysis in innovation studies: mapping actors and technologies. Innovation. 2010;12(1):5–17. https://doi.org/10.5172/impp.12.1.5.

45. Chang S-B, Lai K-K, Chang S-M. Exploring technology diffusion and classification of business methods: using the patent citation network. Technol Forec Soc Change. 2009;76(1):107–17. https://doi.org/10.1016/j.techfore.2008.03.014.

46. Chen JH, Jang SL, Chang CH. The patterns and propensity for international co-invention: the case of China. Scientometrics. 2013;94(2):481–95.

47. Sun Y. The structure and dynamics of intra- and inter-regional research collaborative networks: the case of China (1985–2008). Technol Forec Soc Change. 2016;108:70–82. https://doi.org/10.1016/j.techfore.2016.04.017.

48. Lee S, Kim MS. Inter-technology networks to support innovation strategy: an analysis of Korea's new growth engines. Innovation. 2010;12(1):88–104.

49. Kumari R, Jeong JY, Lee BH, Choi KN, Choi K. Topic modelling and social network analysis of publications and patents in humanoid robot technology. J Inf Sci. 2019;47(5):658–76.

50. Liu W, Li F, Bi K. Exploring and visualizing co-patent networks in bioenergy field: a perspective from inventor, transnational inventor, and country. Int J Green Energy. 2022;19(5):562–75. https://doi.org/10.1080/15435075.2021.1948418.

51. Baumann M, Domnik T, Haase M, Wulf C, Emmerich P, Rösch C, Zapp P, Naegler T, Weil M. Comparative patent analysis for the identification of global research trends for the case of battery storage, hydrogen and bioenergy. Technol Forec Soc Change. 2021;165:120505. https://doi.org/10.1016/j.techfore.2020.120505.

52. Leydesdorff L, Kushnir D, Rafols I. Interactive overlay maps for US patent (USPTO) data based on international patent classification (IPC). Scientometrics. 2012;98(3):1583–99.

53. Curran C-S, Leker J. Patent indicators for monitoring convergence—examples from NFF and ICT. Technol Forec Soc Change. 2011;78(2):256–73. https://doi.org/10.1016/j.techfore.2010.06.021.

54. Kim MS, Kim C. On a patent analysis method for technological convergence. Proc Soc Behav Sci. 2012;40(40):657–63.

55. Borgatti SP, Everett MG. Network analysis of 2-mode data. Soc Netw. 1997;19(3):243–69.

56. Kim DH, Lee BK, Sohn SY. Quantifying technology–industry spillover effects based on patent citation network analysis of unmanned aerial vehicle (UAV). Technol Forec Soc Change. 2016;105:140–57. https://doi.org/10.1016/j.techfore.2016.01.025.

57. Zhang G, Tang C. How could firm's internal R&D collaboration bring more innovation? Technol Forec Soc Change. 2017;125:299–308. https://doi.org/10.1016/j.techfore.2017.07.007.

58. Rassenfosse GD, Dernis H, Guellec D, Picci L, Potterie BVPDL. The worldwide count of priority patents: a new indicator of inventive activity. Melbourne Inst Work Pap Ser. 2012;42(3):720–37.

59. Shubbak MH. Advances in solar photovoltaics: technology review and patent trends. Renew Sustain Energy Rev. 2019;115:109383. https://doi.org/10.1016/j.rser.2019.109383.

60. Zhang RJ, Ye FY. Measuring similarity for clarifying layer difference in multiplex ad hoc duplex information networks. J Inform. 2020;14(1):10. https://doi.org/10.1016/j.joi.2019.100987. (**Article 100987**).

61. Hirsch JE. An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA. 2005;102(46):16569–72. https://doi.org/10.1073/pnas.0507655102.

62. Rousseau R. New developments related to the Hirsch index. Science Focus. 2006;1:23–5 (**in Chinese**). An English translation is available online at http://eprints.rclis.org/6376/.

63. Ye FY, Rousseau R. Probing the h-core: an investigation of the tail-core ratio for rank distributions. Scientometrics. 2010;84(2):431–9. https://doi.org/10.1007/s11192-009-0099-6.

64. Egghe L. (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Oxford (UK): Elsevier.

65. Egghe L. The Hirsch index and related impact measures. Ann Rev Inf Sci Technol. 2010;44:65–114. https://doi.org/10.1002/aris.2010.1440440109.

66. Norris M, Oppenheim C. The h-index: a broad review of a new bibliometric indicator. J Doc. 2010;66(5):681–705. https://doi.org/10.1108/00220411011066790.

67. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. J Informet. 2017;11(4):959–75. https://doi.org/10.1016/j.joi.2017.08.007.

68. Chen HC, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. Mis Quart. 2012;36(4):1165–88 (**Go to ISI>://WOS:000311525500010**).

69. Egghe L. Theory and practise of the g-index. Scientometrics. 2006;69(1):131–52. https://doi.org/10.1007/s11192-006-0144-7.

70. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. The Leiden Manifesto for research metrics. Nature. 2015;520(7548):429–31. https://doi.org/10.1038/520429a.

71. Zhao SX, Rousseau R, Ye FY. h-Degree as a basic measure in weighted networks. J Informet. 2011;5(4):668–77. https://doi.org/10.1016/j.joi.2011.06.005.

72. Schubert A. A Hirsch-type index of co-author partnership ability. Scientometrics. 2012;91(1):303–8. https://doi.org/10.1007/s11192-011-0559-7.
73. Zhao SX, Ye FY. Exploring the directed h-degree in directed weighted networks. J Informet. 2012;6(4):619–30. https://doi.org/10.1016/j.joi.2012.06.007.
74. Jasny BR, Zahn LM, Marshall E. Connections INTRODUCTION. Science. 2009;325(5939):405–405. https://doi.org/10.1126/science.325_405.
75. Zhao SX, Zhang PL, Li J, Tan AM, Ye FY. Abstracting the core subnet of weighted networks based on link strengths. J Am Soc Inf Sci. 2014;65(5):984–94. https://doi.org/10.1002/asi.23030.
76. Wu Q. The w-index: a measure to assess scientific impact by focusing on widely cited papers. J Am Soc Inform Sci Technol. 2010;61(3):609–14. https://doi.org/10.1002/asi.21276.
77. Egghe L. Characterizations of the generalized Wu- and Kosmulski-indices in Lotkaian systems. J Informet. 2011;5(3):439–45. https://doi.org/10.1016/j.joi.2011.03.006.
78. Sarkar JLVR, Majumder A, Pati B, Panigrahi CR, Wang W, Qureshi NMF, Su C, Dev K. I-Health: SDN-based fog architecture for IIoT applications in healthcare. IEEE/ACM Trans Comput Biol Bioinform. 2022. https://doi.org/10.1109/tcbb.2022.3193918.
79. Wang W, Chen Q, Yin Z, Srivastava G, Gadekallu TR, Alsolami F, Su C. Blockchain and PUF-based lightweight authentication protocol for wireless medical sensor networks. IEEE Internet Things J. 2022;9(11):8883–91. https://doi.org/10.1109/jiot.2021.3117762.
80. Yang Y, Wang W, Yin Z, Xu R, Zhou X, Kumar N, Alazab M, Gadekallu TR. Mixed game-based AoI optimization for combating COVID-19 with AI bots. IEEE J Sel Areas Commun. 2022;40(11):3122–38. https://doi.org/10.1109/jsac.2022.3215508.
81. Milanez DH, Lopes de Faria LI, do Amaral RM, Leiva DR, Rodrigues Gregolin JA. Patents in nanotechnology: an analysis using macro-indicators and forecasting curves. Scientometrics. 2014;101(2):1097–112. https://doi.org/10.1007/s11192-014-1244-4.

Publisher's Note