Scientometrics (2024) 129:2159–2179
https://doi.org/10.1007/s11192-024-04949-w

Measuring science and innovation linkage using text mining of research papers and patent information

Kazuyuki Motohashi[1,2] · **Hitoshi Koshiba**[3] · **Kenta Ikeuchi**[2]

Received: 17 April 2023 / Accepted: 18 January 2024 / Published online: 29 February 2024

Abstract
In this study, the text information of academic papers published by Japanese authors (about 1.7 million papers) and patents filed with the Japan Patent Office (about 12.3 million patents) since 1991 are used for analyzing the inter-relationship between science and technology. Specifically, a distributed representation vector using the title and abstract of each document is created, then neighboring documents to each are identified using the cosine similarity. A time trend and sector specific linkages within science and technology are identified by using the count of neighbor patents (papers) for each paper (patent). It is found that the science intensity of inventions (the number of neighbor papers for patents) increases over time, particularly for university/PRI patents and university–industry collaboration patents over the 30 years studied. As for university/PRI patents, the institutional reforms for the science sector (government laboratory incorporation in 2001 and national university incorporation in 2004) contributed to the interactions between science and technology. In contrast, the technology intensity of science (the number of neighbor patents by paper) decreases over time. It is also found that the technology intensity of life science papers is rather low, although they have a significant impact on subsequent patents. However, there are some scientific fields which are affected by technological developments, so that the state of science and innovation interactions is heterogeneous across the fields.

Keywords Text analysis · Patent information · Research paper · Science and technology linkage

Mathematics Subject Classification 68T30 · 68T50

JEL Classification O31 · O34

✉ Kazuyuki Motohashi
kazuyukimot@gmail.com

[1] The University of Tokyo, Tokyo, Japan

[2] RIETI, Tokyo, Japan

[3] NISTEP, Tokyo, Japan

Introduction

The increasing importance of scientific knowledge in innovation can be observed across industries. In the pharmaceutical industry, well known for having a high degree of scientific linkage, the importance of science in new drug development processes is increasing due to the advancement in genomic science (Pisano, 2006). In the electronic device industry, as the large-scale integration production process is miniaturized, understanding the characteristics of nanoscale materials has become indispensable. Furthermore, a recent advancement in machine learning (AI) is achieved through corporate scientists, who make scientific publications and patent inventions, simultaneously (Motohashi, 2019; Hartmann & Henkel, 2020).

The linkage of science and technology has been attracted many scholars in the past, mainly to see whether science-technology linkage leads to novel innovation. For example, Veugelers and Wang (2019) show a positive correlation between scientific novelty and the technological impact of science-based innovation. Cassiman et al. (2018) analyze factors that bridge the gap between scientific knowledge and innovation, focusing on the relationship between the flow of human resources between industry and academia. Consequently, science-based information obtained by a company is more likely to lead to new market innovations (Mention, 2011; Kobarg et al., 2018). However, incorporating scientific knowledge into the innovation process involves a significant technology and market risk for companies (Arora et al., 2016). Therefore, a proper measurement of science and technology interactions is critical to understanding subsequent process of innovation, materializing such science based technology into new products and services.

A typical approach to measure science and technology interaction is based on the information of patent citing to non-patent literature (NPL), mainly research papers, and the majority of studies cited above use this NPL citation based index (Narin and Noma, 1985; Schmoch, 1997; Marx and Fuegi, 2020). However, there are some shortcomings regarding the measure of the linkage between science and innovation. First, it is well known that papers with a substantial volume of citations get more mentions in research papers (Mathew effect; Merton, 1973). Such bias with patent citations should be smaller, given that the patent citations are regulated by the patent system, such as disclosure obligations of relevant prior arts in US patent law. However, it has been found that disproportionally large number of citations are made to some specific patents (Kuhn et al., 2020) and the citation patterns could be geographically biased toward more concentration as compared to the patterns by textual similarity measures (Feng, 2020). These findings suggest that the NPL citations in patent documents suffer from some bias associated with their nature by citing practices by patent applicants and examiners.

Second, NPL citation represents the science used in a patented technology, but it lacks the information of how technology impacts science. In other words, it is unable to show the two-way relationship between science and innovation. While there exists some information of patents cited in research papers, such a scenario is rather rare. Additionally, the nature of citations in scientific papers differ from those of patents, where the novelty factor of the invention is the focus. Put differently, scientific papers, which fulfill the requirements for scientific knowledge, including objectivity and replicability, tend to be used as the citations that form the basis of scientific developments. Therefore, the two-way information between papers/patents to the other patents/papers provides inconsistent information of the interlinkage of science and innovation, even though such information is available.

Therefore, this study relies on the textual information of research papers and patents to delineate the relationship between science and innovation. Specifically, we used the titles and abstracts from research papers and patents published by Japanese authors and inventors between 1991 and 2017 to determine the content similarity across scientific papers and patents. We grouped the documents with high-content similarity and clarified the mutual relationship between research papers and patents.

The content similarity between research papers and patents have been investigated in some specific technology fields, such as Magerman et al. (2015), or by using academic inventors, involving both research papers and patents (Ikeuchi et al., 2017; Lissoni et al., 2013). However, to the best of our knowledge, there is no such work that combines research papers and patents to establish the interlinkage of science and innovation across science and technology fields, while a large-scale content examination of patent text information has been conducted and evaluated in the past (Arts et al., 2017; Younge & Kuhn, 2016). This study fills the aforementioned research gap by providing systematic information in connection to science and innovation interlinkage across technology and scientific fields based on the document contents in two forms.

The remainder of this study is organized as follows: "Data sets and text mining techniques" section presents an outline of our obtained data as well as a visualization of the overlapping of science and technology over time. In "Checking contents similarity by document embedding" section, we use the citation information from research papers and patents to perform an evaluation of a similarity index via text mining. In "Neighbor documents-based indicator of science and technology linkage" section, we present a relation index of science and technology linkage based on the neighbor patent or paper information. "Dynamic analysis of science and technology coevolution" section extends this analysis to observe the changes in the mutual interactions in relation to the science and technology field. Finally, we conclude our study with some research limitations in "Conclusion" section .

Data sets and text mining techniques

Data sets

In this study, to comprehensively observe the interlinkage between Japanese science and technology, we used the following data sets:

Research paper information: Papers included in the Science Citation Index expanded from Clarivate's Web of Science, published between 1991 and 2017, containing at least one Japan-based author.
Patent information: All patents filed to the Japan Patent Office (Goto and Motohashi, 2017) at PATSTAT2020 Spring Version (those for which English-translated title and abstract information are available).

Regarding the number of documents, we used 1,696,338 research papers and 12,330,725 patents, forming a total of 14,027,063 documents.

Figure 1 shows the changes in the number of documents by publication year (for patents, the application year). The number of patents shows a declining trend since 2000, while the number of research papers remains stable, with ~ 100,000 publications per year.

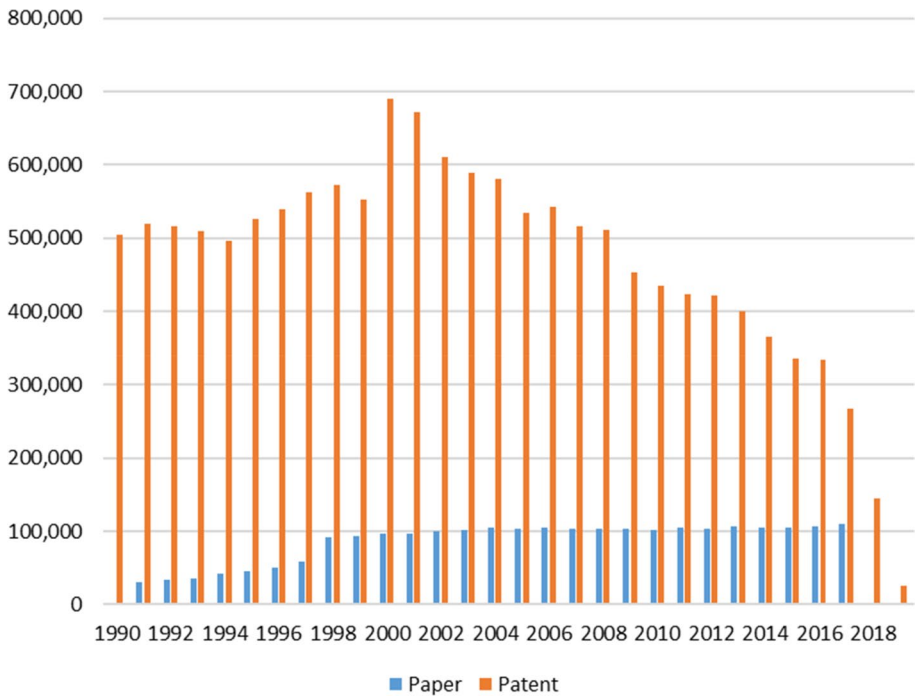


Fig. 1 Numbers of papers and patents by application/publication year

Text mining method and clustering results

We followed two steps in creating document embedding vectors that represent the content of each document. First, we extracted only the nouns that appear in a total of approximately 14 million titles and abstracts (all paper and patents described above) and used FastText (Bojanowski et al., 2017; Joulin et al., 2016) to create embedding vectors for words other than common and rare words with a dimension of 300. Here the common words are about 170 stop words recorded in Python NLTK library, and the rare words are those occurs only one time in our whole corpus. Second, this embedding vectors for each of words are aggregated into document level, by summing up of all word vectors within each document, and normalizing the document vectors. Regarding the embedding results for the words, we conducted cluster analysis using the K-means method and confirmed, by visual checking, that semantically similar words belong to the same cluster (for details, see Motohashi et al., 2019).

The embedding results of the words were aggregated for each document. We clustered these using the K-means method (classification with 16 clusters) and compressed the results into two dimensions using the uniform manifold approximation and projection (UMAP) technique (McInnes et al., 2018) (Fig. 2).

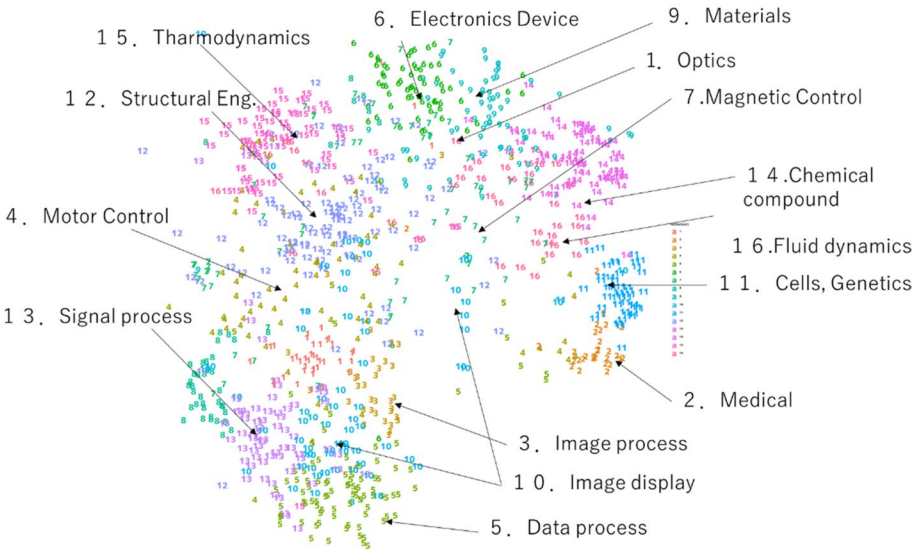


Fig. 2 Visualization of document distribution with clustering results

Visualizing the relationship between research papers and patents

Figures 3 compares the distribution of papers and patents in the same science-technology space as one of Fig. 2, between papers/patents published/applied in 1990s and those in the 2010s, where the red indicates the location of patents, while the blue indicates the location of research papers.

Overall, a large percentage of the research papers were related to life science (cells/genes, medicine) and chemistry/materials (chemical compounds, metal ingredients). They were evidently distributed across the fields of optics, fluid processing, and video display. However, fields relating to mechanics (motion control, structural mechanics, and thermodynamics), electronic devices, and image processing are mostly covered by patents.

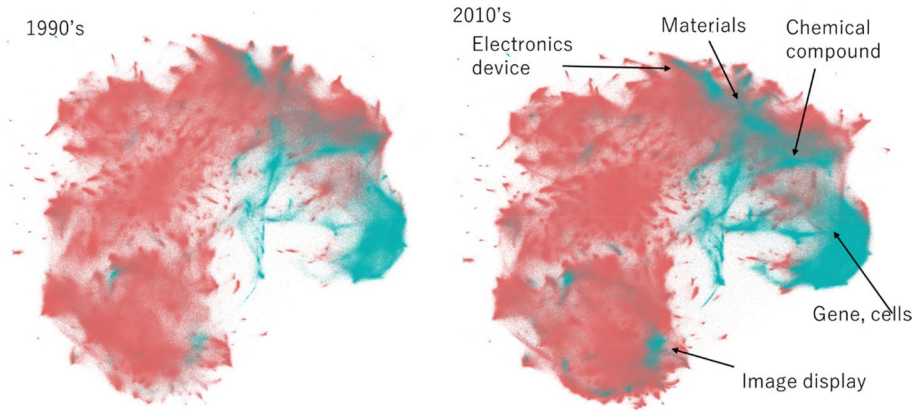


Fig. 3 Paper and patent mapping (1990s and 2010s)

Regarding changes over time, as the number of research papers increases with respect to the total number of documents, an expansion of research papers in technical fields can be observed (overlapping area of red and blue). This trend is particularly notable in the fields of chemistry/materials (compounds, metallic materials). It was also evident that research papers had been published in fields such as electronics devices and image displays, which were previously only covered mainly by patents.

Checking contents similarity by document embedding

We evaluated the document embedding results by using citation pairs (paper–paper, paper–patent, and patent–patent) and document pairs in the same research project. Specifically, the cosine similarity of randomly selected pairs are compared with that of those pairs whose contents are similar each other.

First, the distribution of the cosine similarity of randomly selected 10,000 pairs for the three patterns of "paper–paper," "paper–patent," and "patent–patent" is presented in Fig. 4, where the decile values for cosine similarity of each pair are plotted. Looking at the median values (median, P50), "paper–paper" has the highest value at 0.73, followed by "patent–patent" (0.70), and finally, "paper–patent" (0.69). Additionally, considering the 10th percentile (P10), the respective values are ~ 0.6, and it can be seen that cosine similarity between randomly extracted samples is distributed in a relatively narrow region (the width between the 10th and 90th percentiles being ~ 0.2). In a similar exercise using patent abstract information in Japanese, the median of randomly extracted cosine similarities was ~ 0.5 (Motohashi et al., 2019). Therefore, the cosine similarity in this study (using an English abstract) is relatively higher than that of Japanese. This may be due to the difference in the vocabulary of both languages.

Further, we compared these distributions (in Fig. 4) with those by the document pairs of citing and cited relationship, as well as those of the outputs from the same research project. The citation pair information is taken from the patent literature and NPL citations of patents (patent–patent, and patent–paper pairs) and paper citations in each research paper

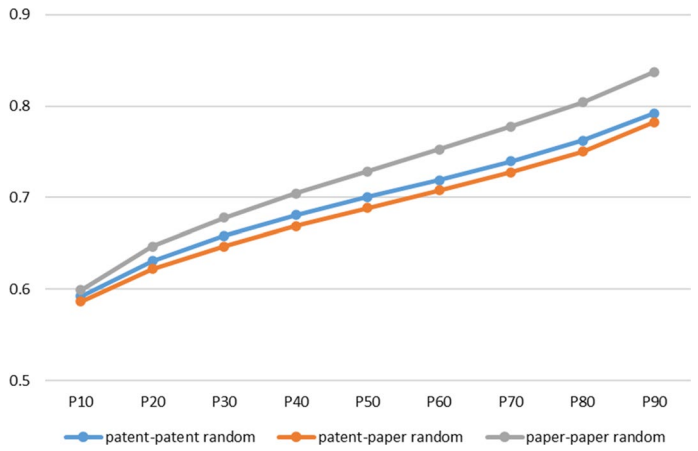
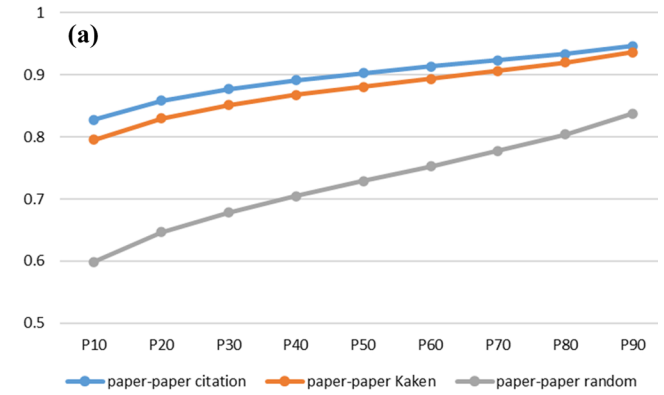
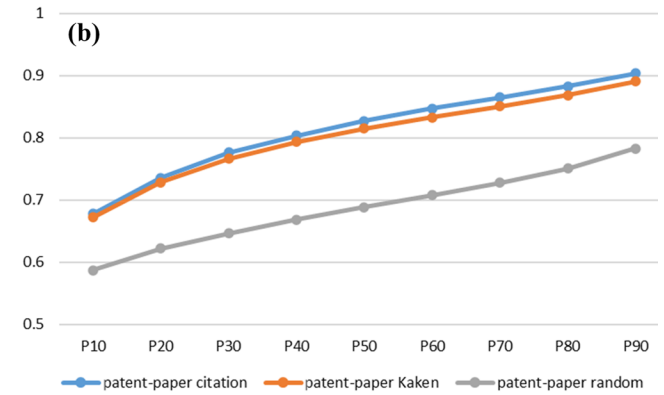


Fig. 4 Distribution of cosine similarity of random sampled pairs





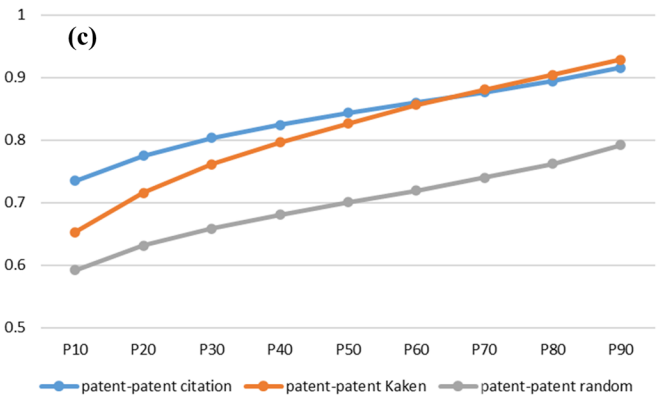


Fig. 5 **a** Comparison of cosine similarity distribution (paper–paper). **b** Comparison of cosine similarity distribution (paper–patent). **c** Comparison of cosine similarity distribution (patent–patent)

(paper–paper citation). In addition, we take the information of papers and patents with similar contents from the JSPS Kakenhi "Report on the Research Results," where documents (papers and patents) are declared as the outputs within an identical funded research project. The cosine similarity between those pairs (citing pairs and pairs in an identical project) is supposed to be greater than that of randomly selected pairs due to the similarity of document contents. Figure 5a–c compare the results for the "paper–paper," "paper–patent," and "patent–patent" pairs, respectively.

It is confirmed that the cosine similarity between citation pairs and same-project outcomes is higher in all figures. In addition, it is found that the citation pairs and same-project pairs between research papers provide information with high homogeneity (0.8 or greater even at the 10th percentile). However, for some other pairs, 10th percentile values are under 0.7, which is a value lower than the median value for random pairs. Additionally, the distributions for citation pairs and same-project pairs are almost the same, except for those between patents. With regard to "patent–patent" pairs, the variation in same-Kakenhi-project pairs is greater than for citation pairs. Arts et al. (2017) has validated the accuracy of content similarities derived from using patent abstract information. Therefore, the variation of cosine similarity infers the substantial variance of content similarities among citation pairs of documents (together with those with research outcomes within the same project).

Neighbor documents-based indicator of science and technology linkage

Identification of neighbor documents in terms of the document contents

The science and innovation linkage can be detected by the coexistence of research papers and patents with similar embedding vectors. To identify such fields, we extracted the neighbor documents for each of the 14 million documents (papers or patents). We applied the neighborhood graph and tree (NGT) for Indexing High-dimensional Data algorithm, where a certain number of neighboring documents can be efficiently searched, out of several documents (Iwasaki, 2011). In this study, we extracted 200 neighboring documents based on cosine similarity with each of the 14 million focal documents.

The number of the neighbor documents, a global parameter of the NGT algorithm, reflects the area of document search in the document content space. The greater number of neighbor documents allows for a more comprehensive search of the documents, but it requires more computer costs in terms of the process speed and the storage area.

The distribution of the cosine similarity for the 100th and 200th documents with each of the focal documents is presented in the Table 1. It is found that the values of the cosine similarities between the 100th and 200th document are almost the same (for example, the median value is 0.899 for the 100th and 0.893 for the 200th). Our embedding vector has a dimension of 300, and the cosine similarity is proportional to the Euclidian distance between a focal document and its neighboring one. The number of neighbor documents is proportional to the volume of a hypersphere with a dimension of 300, which is the 300th power of the radius from the position of a focal document.

This raises a pertinent question of the result of enlarging the number of neighbor documents to 1000 for example. For this we assumed that the cosine similarity of the 200th neighbor is 0.9 and the documents are evenly distributed in a 300 dimensional space. Then,

Table 1 Distribution of the cosine similarities of neighboring documents

	100th	200th
1%	0.843	0.834
5%	0.870	0.863
10%	0.881	0.875
25%	0.899	0.893
50%	0.916	0.911
75%	0.932	0.928
90%	0.944	0.941
95%	0.951	0.948
99%	0.961	0.958

the question entails determining the extent to which the cosine similarity decreases. First, the cosine similarity is converted to the radius (Euclidian distance by $2(1 - \cos)$), to get 0.2.[1] Then, the radius of the hypersphere with 1000 documents inside (x) can be derived from the following equation:

$$\frac{(x)^{300}}{(0.2)^{300}} = \frac{1000}{200}$$

From the estimated x, the radius of a 1000 document search can be obtained to be 0.201076, which is only 0.53% larger than that of 200 documents (0.2). However, the 200 neighbor results lead to 2.8 billion (20*14 million) observations with a tsv file size of about 70 GB. In the case of 1000 neighbor extractions, the output size becomes five times as above, so that the cost performance to increase the size of neighbors is very poor. Therefore, we stick to the size of 200 neighbors for subsequent analysis.

The number of neighbor documents also depends on the search objective, that is, the extent of the similarity to be required for document search. In this regard, the cosine similarity of 0.9 for the 200th document corresponds to the proximity in the 60th percentile for "paper–paper," the 90th percentile for "paper–patent," and the 80th percentile for "patent–patent" based on citation pairs (Fig. 5a–c in the previous section). Therefore, by extracting 200 neighbor documents, it provides an opportunity to investigate the overlapping paper and patent documents within the content similarity comparable to the citing and cited pair ones.

Another potential concern would be that up to 200 similar documents is too many. As is analyzed in Table 1, the distribution of cosine similarity with 100th documents and 200th ones is not so different. But how about looking at only smaller documents such as 10 or 20? In order to address the issue of false positive, we have conducted a sensitivity analysis of the numbers neighbors to the degree of similarity. Since the patent documents contain detail information about technological classification, we constructed the similarity measure by looking at the share of same IPC subclass pairs in all neighbor patents. Table 2 shows the results, indicating the share by the number of neighbors (column) and the quadrant group by cosine similarity of the 200th document (row). It should be noted that the quadrant is created by the threshold cosine similarity measures of 0.892, 0.90 and 0.927, indicating that the density of document distribution becomes greater for higher quadrant.

[1] Our embedding vector is normalized, as its norm = 1.

Table 2 Share of same IPC subclass pairs by # of neighbors

	10	20	50	100	200
Q1	30.6%	28.1%	25.0%	22.8%	20.7%
Q2	33.1%	31.0%	28.4%	26.5%	24.7%
Q3	36.5%	34.6%	32.2%	30.5%	28.8%
Q4	42.9%	41.3%	39.4%	37.9%	36.6%

First, it is natural to see that the same IPC share decreases by the number of neighbor increasing, showing that the larger the size of neighbor is, the more likely it suffers by false positive errors. Second, the same IPC share increases by the document density around the focal document. It should be noted that the same IPC share of 200 documents (36.6%) of Q4 (the most dense area) is greater than that of 10 documents (30.6%) of Q1 (the least dense area). Therefore, the likelihood of false positive is not so sensitive to the size of neighbor documents, as compared to the variation of document density across science-technology space. Given that the false negative error is higher for smaller numbers of neighbors for use, we have decided 200 neighbors for subsequent analysis of science technology interactions.

It is conceivable that the cosine similarity of citation pairs would also be affected by this state of technical spatial density. This is because it is highly likely that a document with higher cosine similarity is cited among documents that are located in a place with high-technical spatial density. In Fig. 6, neighboring documents are divided into four groups based on their cosine similarities with the 200th document (Groups Q1–Q4, same as in the Table 2), and the distribution of cosine similarities with the citation pairs of documents in each group (decile values) are observed. As hypothesized, documents located in a dense technical space (e.g., a document in Q4) have a high-cosine similarity with their cited documents. Additionally, the effect of spatial density is greater in the groups with sparse density (e.g., Q1).

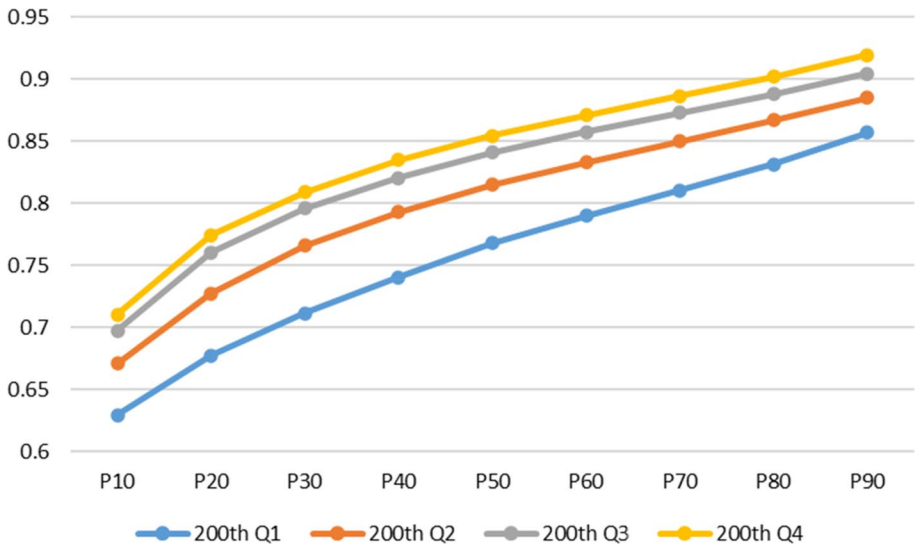


Fig. 6 Cosine similarity distribution of citation paper by content density

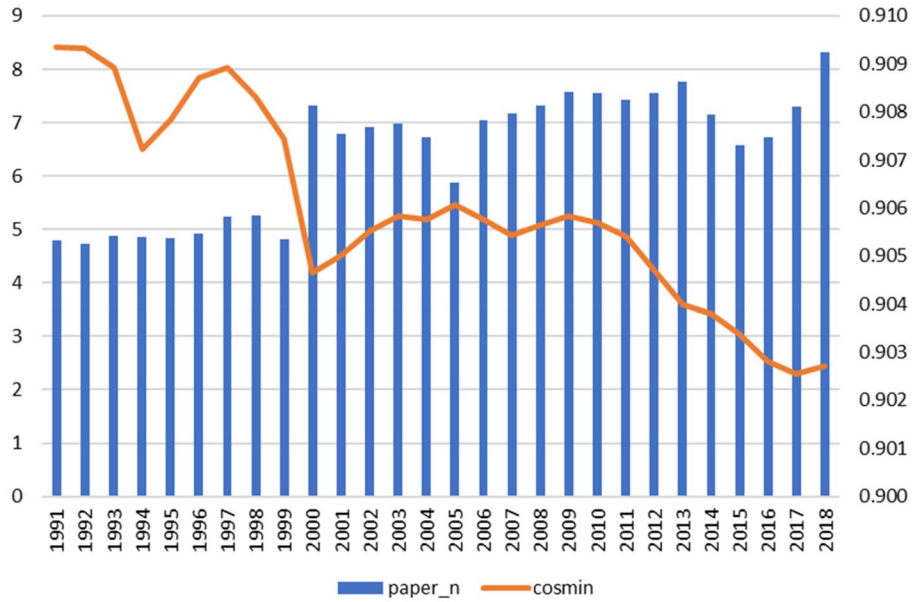


Fig. 7 Number of neighbor papers and cosine similarity in 200th for patents

Counting neighbor papers of patent documents

Counting the number of papers in 200 neighbor documents by each patent allows us to investigate the degree of scientification of inventions. The mean of the number of neighbor papers is 5.88 (out of 200 neighbor documents). Its median value is 0, and 7,171,041 patents (58.2% of the total number of patents) do not have any paper (paper count = 0) in 200 neighbor documents. As indicated in Figs. 1, 2, and 3, there are substantial technological fields with no overlap with the distribution of scientific papers.

Figure 7 shows the mean value of the number of neighbor papers by patent application year, together with the cosine similarity in the 200th document (minimum cosine similarity in 200 neighbor documents). In general, an increasing pattern of the mean paper counts is found, implying that the science intensity of invention increases over time. It should be noted that there are more neighbor papers published before the patent application as the application year become later. Therefore, the increasing pattern of the neighbor paper could be interpreted by the fact that more scientific papers existed before patents to be applied in later years than the scientific papers published after the patent applications in earlier years. In other words, science influences technology more than technology affects science. In addition, the minimum cosine similarity (200th document) decreases over time, implying that newer patents are applied in relatively sparse areas in the technology space.

Figure 8 shows the mean neighbor paper counts by patent applicant type. Here, we distinguished the patents by firm only, university or public research institute (PRI) only, and joint application of firm and university/PRI. We eliminated all patents by other patterns of applicant compositions, such as those involving individual inventors. The increasing pattern of overall paper counts is driven by the patents involving university or PRI, as is the share of patents by university or PRI increases from 0.7% in the 1990s and 1.6% in the 2000s to 1.7% in the 2010s. It should be noted that the substantial institutional reforms

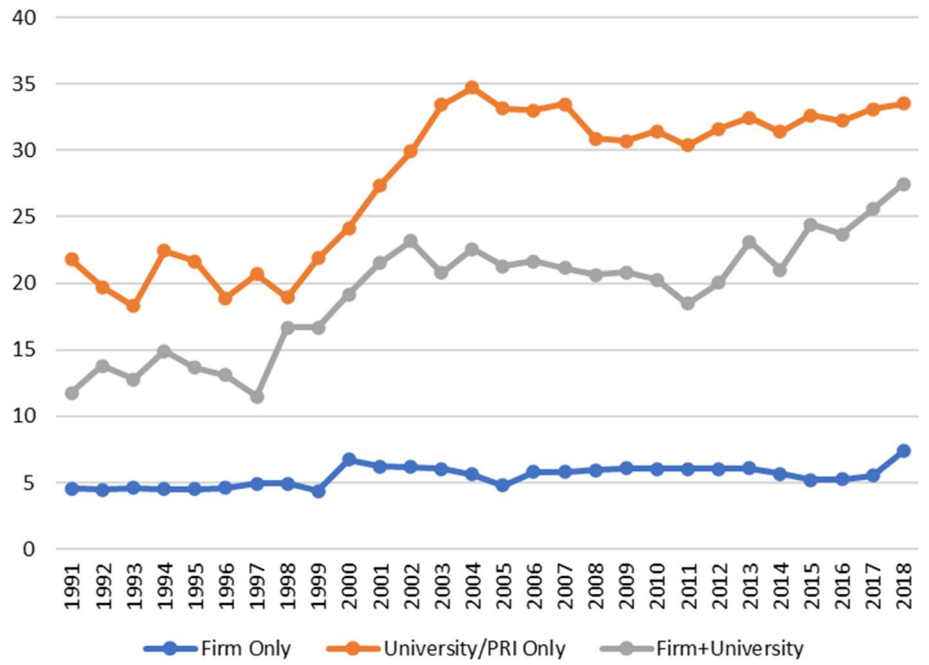


Fig. 8 Neighbor paper counts by patent applicant type

in public science sector was introduced in 2001 for the central government laboratories and in 2004 for national universities. Both types of entities, which used to belong to the government, became an independent agency, and they started applying the patents after these years (Motohashi & Muramatsu, 2012). As the central government laboratories and national university focus more on basic science in their research activities, compared to local government laboratories and private universities, the mean science index by paper counts increased after years of these new players started patenting (2001 for government laboratories and 2004 for national universities).

Furthermore, it should be noted that the science index by firm's patents is relatively stable, while that of university–industry joint application patents (UI patents) increases over time. The increase in the science index of UI patents is found not only in the 2000s but also after the 2010s. The pattern of private firms' retreat from basic science in the US (Arora et al., 2016) is also found in Japan. However, private firms substitute their in-house scientific activities by joint research with universities and PRIs.

Finally, changes in the science index are analyzed by technology field. Figure 9 shows the neighbor paper counts by technology field of focal patent in the 1990s and 2010s (classified into the 35 categories used by WIPO for its annual report; Schomoch, 2008).

First, the scientification of inventions has a highly skewed distribution with very high intensity in life science fields, such as biotechnology and pharmaceuticals. In addition, the intensity is relatively high for micro-structural and nano, fine chemistry, and IT methods/management. In contrast, the mean value of neighbor science paper counts is only less than 10 (5% of the total of 200 documents) in most of the technology fields.

Second, the science index in the 2010s is higher than that of the 1990s for most of the technology fields. It is important to note that the majority of the neighbor papers for the

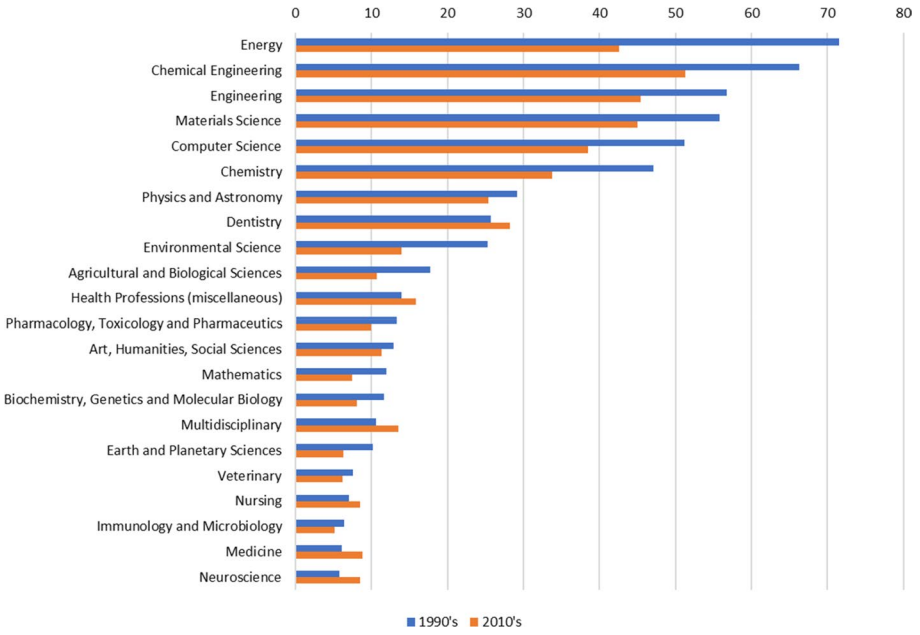


Fig. 9 Mean paper counts by technology field

patents in the 2010s are published before their applications and vice versa for the patents in the 1990s. Therefore, a higher science index in the 2010s suggests that new inventions are born in the technology fields where some scientific understandings have been achieved beforehand. Alternatively, technology generally relies on science. This is typically the case for technology fields, where substantial differences in the level of science indices in two periods exist, such as pharmaceuticals and nano technology.

Counting neighbor patents of paper documents

The interlinkage between science and technology can also be observed by counting neighbor patents around each research paper. The mean neighbor patent counts is 23.05 (out of 200 documents). The median value is 2 and the number of papers without any neighbor patents is 618,238 (36.4% of all research papers). Again, there are substantial areas for pure science without any patent applications nearby.

Figure 10 shows the trend of mean neighbor patent counts, together with the cosine similarity with the 200th neighbor document (minimum cosine similarity for all neighbor documents). In contrast to the science index of patent, the technology index of paper decreases over time. Such an overall trend is the reverse side of the coin of the increasing trend of the science index of patent, that is, technology relies on science, but not so much for vice versa.

Figure 11 shows the technology index of paper by author affiliation type, firm only, university/PRI only, or joint publication of firm and university/PRI. The technology index decreases for paper by all sectors, but such a trend is clearer for papers by private

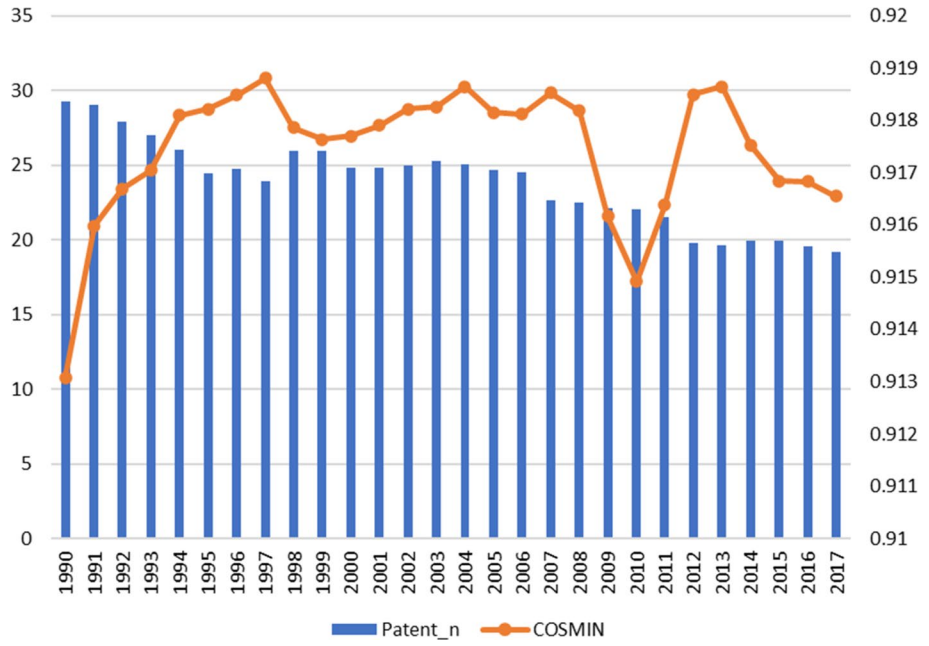


Fig. 10 Number of neighbor patents and cosine similarity in 200th for papers

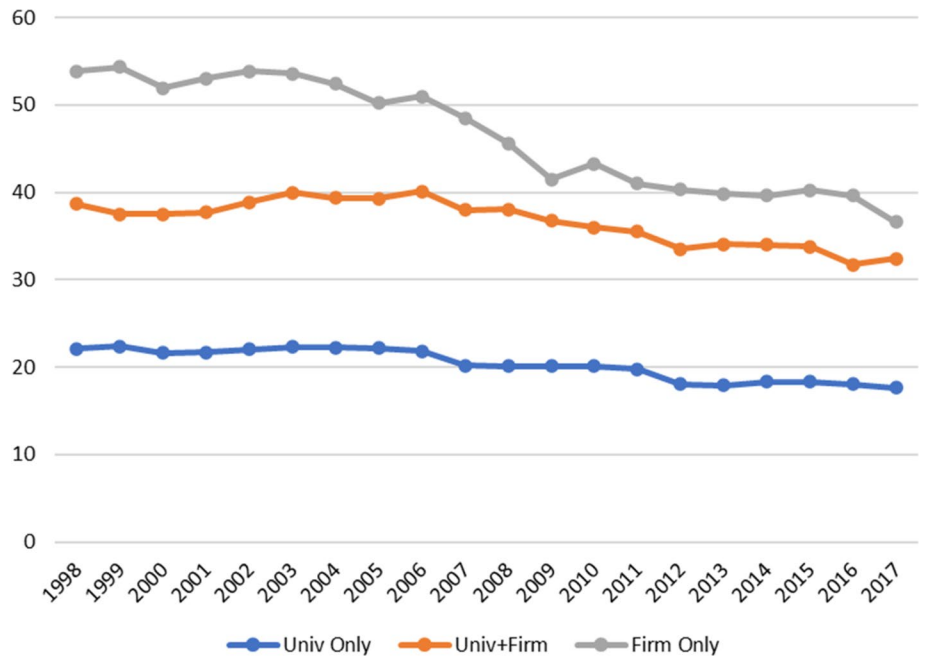


Fig. 11 Neighbor patent counts by paper author affiliation type

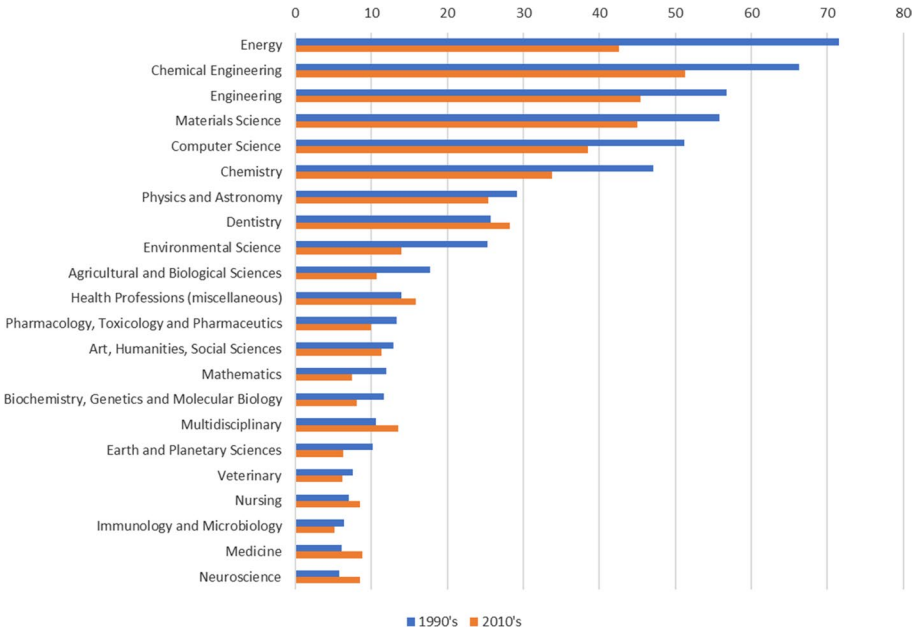


Fig. 12 Mean patent counts by scientific discipline

firm authorship. It should be noted that the number of papers by private firms decreases over time (4408 in 2000 v.s. 2226 in 2017), while the number of joint paper with university/PRI increases (5386 in 2000 v.s. 5980 in 2017). In the process of such structural change, the technology indices of firm only and joint paper with university/PRI are converged to a similar level. This can be interpreted as that a firm substitutes in-house research activities to joint activities with public research organizations.

For the papers involving university/PRI authors, the technology index is stable in the 2000s due to the institutional reforms in national laboratory and university in the early 2000s. It starts declining around 2007 after a temporary shock.

Finally, Fig. 12 shows the technology intensity in the 1990s and 2010s by the scientific field of research paper (based on Web of Science subject code). The scientific fields with high technology index in the 1990s, such as energy, chemical engineering, and material science, have relatively greater impacts on subsequent inventions, while there are many fields with smaller numbers of technology applications. It should be noted that the technology intensities in life sciences, such as biochemistry and pharmacology, are relatively small, even though the science index of such applications is high. Put differently, the scientific frontier in these fields expands very rapidly, where only small part of them can be a basis of subsequent innovations.

Dynamic analysis of science and technology coevolution

In this section, we discuss the dynamics of science and technology evolution over 30 years and how the results are interpreted in the foregoing sections.

Table 3 shows the changes in the distribution of neighbor documents for patents and research papers. COSMIN is the cosine similarity of the 200th neighbor document, and Radius is the Euclidian distance converted by 2(1 − COSMIN) covering the neighbor search in a 300 dimensional technology space. This radius should be adjusted by the number of documents, as the time trend of patent applications and research paper publications exists, if such a trend comes from changes in application/publication propensity, given the same technological or scientific findings. Radius-adj shows the adjusted values in the 2010s by using the following equation.

$$\text{Radius}_{\text{adj}}^{2010} = \text{Radius}^{1990} * \sqrt[300]{\frac{\#\text{of doc}^{2010}}{\#\text{of doc}^{1990}}}$$

Regarding the patent, the mean adjusted radius to the 200th neighbor document increases from 0.1793 to 0.1911, representing an increase of 6.5%. In the case of the original value (0.1918 in the 2010s), it increases by 7.0%. In contrast, the same measure is relatively stable for research papers (0.1645 to 0.1661, 1.0% growth). These findings imply that the technological frontier measured by patent expands to more sparse places in the technology space. In other words, the area covered by patent applications expands its space. In contrast, the size of the scientific frontier covered by research papers is relatively stable.

Next, its dynamics in the science-technology space is analyzed. The science intensity of patents increases on average (from 4.71 neighbor papers in the 1990s to 7.40 neighbor papers in the 2010s), while the share of patents with no research paper as a neighbor document does not change significantly (58.1% in the 1990s and 59.0% in the 2010s). In contrast, the technology intensity of papers decreases on average (from 25.7 neighbor patents in the 1990s to 20.2 neighbor patents in the 2010s), and the share of papers with no patent as a neighbor increases (from 20.0% in the 1990s to 37.0% in the 2010s).

The Fig. 13 graphically explains the changes of the area covered by papers and patents with one dimension science-technology space representation with the document quantity by area. "Paper 1990s" stands for the papers published in 1990s, while "Paper 2010s" does for the papers published in 2010s, and the same for patents. Therefore, for example, the science and technology linkage for "Paper 1990s" can be shown in the overlapping areas with both "Patent 1990s" and "Patent 2010s".

Frist, the share of patents with no research paper does not change over time, by looking at $(a–b)/a$ for the 1990s and $(d–c)/d$ for the 2010s. Second, the science intensity, represented by the share of A to the whole area of triangle for the 1990s and the share of A + B + C to the whole area of rectangle for the 2010s, increases over time. It should be noted that the quantity of scientific papers as a neighbor is represented by the

Table 3 Changes in the distribution of neighbor documents

	COSMIN (a)	Radius (b)	# of docs (c)	Radius-adj (d)
Patent in 1990s	0.9103	0.1793	6,472,191	0.1793
Patent in 2010s	0.9041	0.1918	2,207,567	0.1911
Paper in 1990s	0.9177	0.1645	420,412	0.1645
Paper in 2010s	0.9170	0.1659	612,618	0.1661

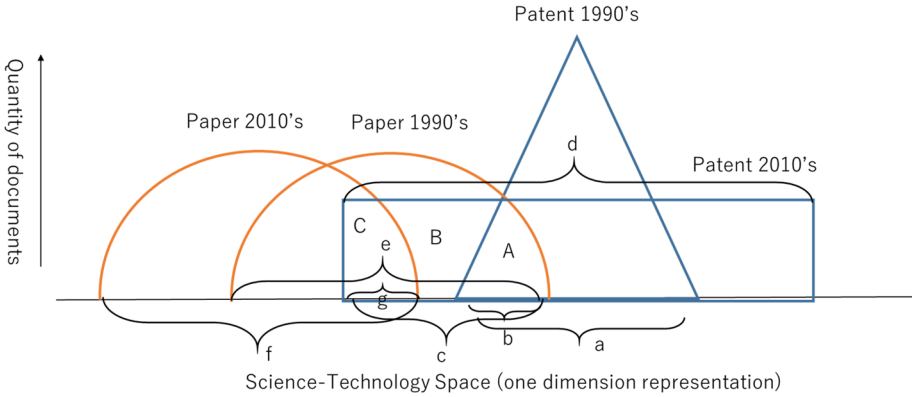


Fig. 13 Changes in the relationship between science and technology

overlapped area of each of patent area (1990s and 2010s), with a whole area covered by both 1990s and 2010s papers. Hence the area A for the 1990s and the area A + B + C for the 2010s.

Similarly, the technology linkage from the viewpoint of papers can be looked as follows. First, the increase of the share of papers with no patent as a neighbor is described as $(c–e)/e$ (1990s) to $(f–g)/f$ (2010s). Second the decrease in the technology intensity can be described as the share of the A + B + C to the whole area of half circle for the 1990s and the C to the whole area of half circle for the 2010s.

In summary, technology evolutions reflected in patents occurs in the direction of scientific fields, as well as an opposite direction (non scientific fields). Therefore, science based technological development has been progressed, but there are some other areas of technological developments, regardless of scientific findings. In contrast, scientific frontier expansion occurs toward to the left direction in Fig. 13 only, indicating the technology has little influence over scientific progress, at least at the macro level.

Finally, we examined the dynamics of technology changes in relationship with science and technology interactions by technology or science field. First, we measured the technological (scientific) changes in patents (papers) by considering how an individual technology (science) field moves from the 1990s to the 2010s. More specifically, we calculated the centroid vector for each technology (science) field in the 1990s and 2010s and used the 1 minus cosine similarity of these vectors as the degree of technological (scientific) changes of each field. Figure 14 shows the scatter graph of this measure and the changes in neighbor papers from the 1990s to the 2010s of 35 technology fields.

Except for "furniture and games" and "digital communication," the technology fields with growing science intensity, such as "pharmaceuticals," "organic fine chemicals," "biotechnology," and "nano-technology," show a relatively greater level of dynamics in the contents of patents. Two exceptions may be explained by the market change, such as video games and mobile telecommunications services, together with technological advancements related to new products. But the other technology fields rely heavily on scientific findings. In addition, we cannot find any technology field with high science intensity and relatively less dynamics in technological change. Therefore, it leads us to say that the science is one of important factors behind technological dynamics.

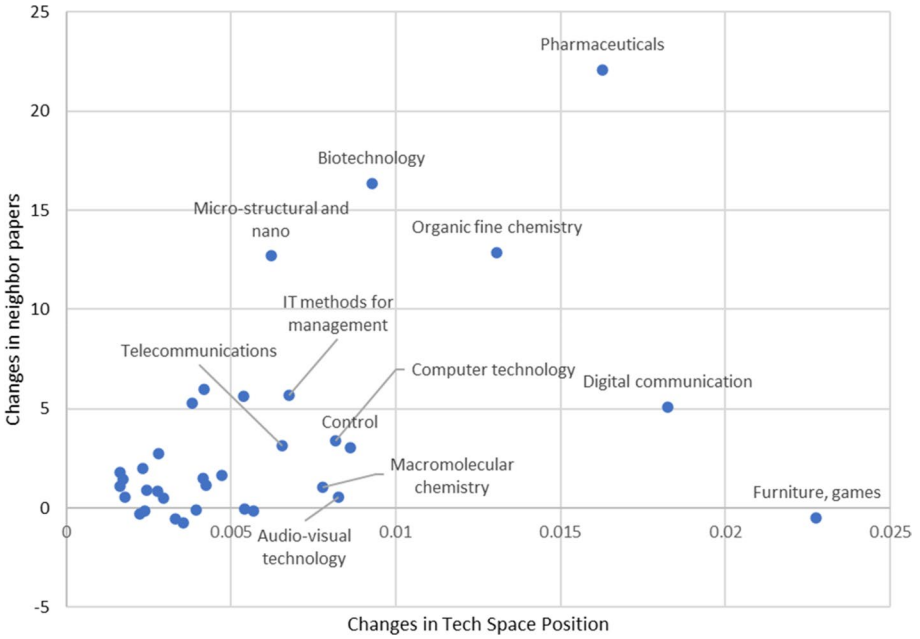


Fig. 14 Change of science intensity and technology space position for patents

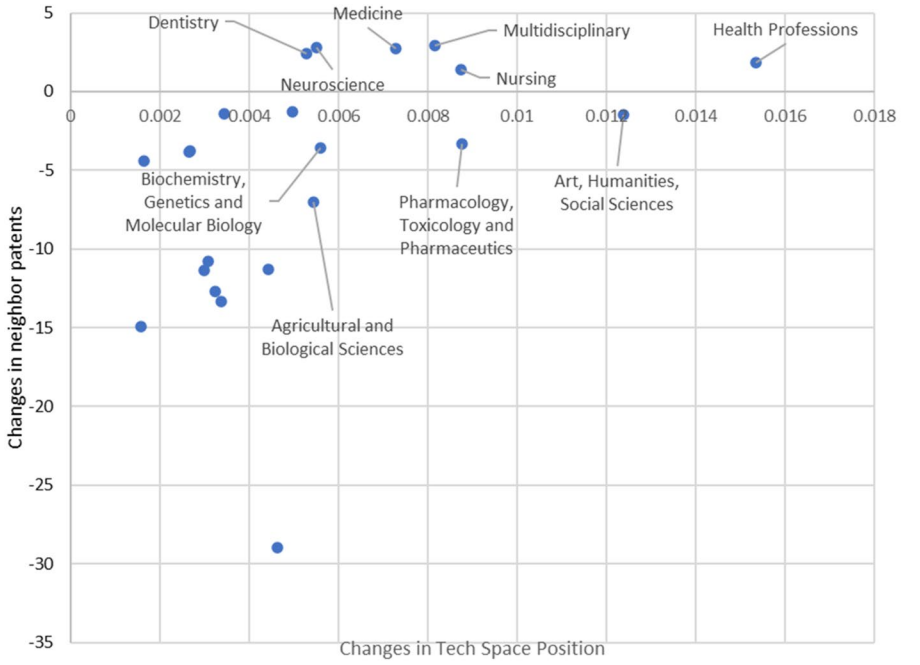


Fig. 15 Change of technology intensity and technology space position for papers

Figure 15 shows the dynamics of science and technology intensity by scientific field. As the proportion of neighbor patents to research paper decreases over time in general, the changes in neighbor patents become negative from the 1990s to the 2010s for most of the scientific field. However, there are some fields without substantial decrease or even some increase in that count, such as "health professions," "nursing," and "multidisciplinary." It should be noted that the degree of technology position change is relatively higher for these fields. As for the scientific fields contributing to science-based innovation (mainly life science fields), the reverse impact (from innovation to science) is unclear. However, it should be noted that there are some scientific fields, such as "health professions" and "nursing," where technological change and industry applications lead to their dynamics.

Conclusion

In this study, we analyzed the two-way relationship between science (research papers) and technology (patents) using text data from 1.7 million published papers and 12.3 million filed patents since 1990. Specifically, we created document embedding vectors using the titles and abstracts for each document and used cosine similarity to extract 200 neighboring documents by using the NGT algorithm. The relationship between research papers and patents was quantified using the number of neighboring patents (research papers) for each research paper (patent).

It was found that the scientification of inventions (the number of neighbor papers for patent) increased over time, particularly for university/PRI patents and university industry collaboration patents over these 30 years. As for university/PRI patents, the institutional reforms for the science sector (government laboratory incorporation in 2001 and national university incorporation in 2004) affect the interactions between science and technology.

In contrast, the technology intensity of science (the number of neighbor patents by paper) decreased over time. It was also found that the technology intensity of science, having significant impact on subsequent innovations, such as life science, was rather low. However, there are some scientific fields where related technological developments affect their scientific progress. Therefore, while there is substantial heterogeneity by the technology and science field, there is a presence of some two-way interaction between science and technology (innovation).

In Japan, major institutional reforms were conducted in the 2000s, such as national laboratories becoming independent administrative agencies in 2001 and national universities becoming national university incorporations in 2004. It was evident that such institutional reforms increase the science intensity of inventions. However, there is little sign of science being influenced by such institutional reforms. After the reforms, central government laboratories and national universities got involved in substantial technology commercialization activities, but there is little evidence of their research activities being biased toward application orientation instead of basic science.

This study proposes a new methodology and science/innovation two-way interaction by using research paper and patent text information. However, there are some limitations in our research. First, we measure scientific findings by research papers, instead of controlling for the heterogeneity of their contents. For example, we found that the technology intensity of scientific papers in "energy" or "chemical engineering" is high, while that of "mathematics" or "genetics" is low. This observation could be interpreted by the type of research papers, an application-oriented paper or a basic science one. Here, further investigations

are needed based on some conceptual works related to the taxonomy of science/technology (e.g., Stokes, 1997).

Another limitation of our work is the methodology of document embedding. We chose a bag-of-words approach, where we obtained embeddings for single words and aggregated them by document. The most serious problem with this methodology is that the embedding vector for each word is consistent over time. In this regard, we need to consider the context of the word used in each document. There has been tremendous progress in the methodology of contextual word embedding, such as bidirectional encoder representation with transformation (BERT). Recently, BERT has been used for patent text analysis, and it has been found to work efficiently in distinguishing the difference between similar patents (Lee & Hsiang, 2020; Li et al., 2017). In addition, all of such embedding techniques assume that the document content is represented as a point in certain vector space. However, science and technology progress occurs in technology conversion of multiple technical components. Taking into account mixed components for document content representation is necessary with this regard. It is another venue of potential future research.

Acknowledgements Open access funding provided by The University of Tokyo. This study is conducted as a part of the Project "Digitalization and Innovation Ecosystem: Holistic Approach" undertaken at the Research Institute of Economy, Trade, and Industry (RIETI). The authors would like to thank Professor Nagaoka and RIETI discussion paper seminar participants for their helpful comments. The authors also acknowledge support from JSPS KAKENHI (Grant Number JP18H03631).

Funding Open Access funding provided by The University of Tokyo. This study is funded by the Research Institute of Economy, Trade, and Industry (RIETI), under the project "Digitalization and Innovation Ecosystem: Holistic Approach", together with JSPS KAKENHI (Grant Number JP18H03631).

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Arora, A., Cohen, W. M., & Walsh, J. P. (2016). The acquisition and commercialization of invention in American manufacturing: Incidence and impact. *Research Policy, 45*(6), 1113–1128.

Arts, S., Cassiman, B., & Gomez, J. C. (2017). Text matching to measure patent similarity. *Strategic Management Journal, 39*(1), 62–84.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Cassiman, B., Veugelers, R., & Arts, S. (2018). Mind the gap: Capturing value from basic research through combining mobile inventors and partnerships. *Research Policy, 47*(9), 1811–1824.

Feng, S. (2020). The proximity of ideas: An analysis of patent text using machine learning. *PLoS ONE, 15*(7), e0234880.

Goto, A., & Motohashi, K. (2007). Construction of a Japanese Patent Database and a first look at Japanese patenting activities. *Research Policy, 36*(9), 1431–1442.
Hartmann, P., & Henkel, J. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries, 6*(3), 359–381.
Ikeuchi, K., Motohashi, K., Tamura, R., & Tsukada, N. (2017). Measuring science intensity of industry using linked dataset of science, technology and industry. RIETI Discussion Paper, 17-E-056.
Iwasaki, M. (2011). Proximity search using approximate K nearest neighbor graph with a tree structured index. *Journal of Information Processing Society of Japan, 52*(2), 817–828. in Japanese.
Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models, arXiv preprint, arXiv:1612.03651
Kobarg, S., Stumpf-Wollersheim, J., & Welpe, I. M. (2018). Universityindustry collaborations and product innovation performance: The moderating effects of absorptive capacity and innovation competencies. *The Journal of Technology Transfer, 43*(6), 1696–1724.
Kuhn, J., Younge, K., & Marco, A. (2020). Patent citations reexamined. *Rand Journal of Economics, 51*(1), 109–132.
Lee, J., & Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. *World Patent Information, 61*, 101965.
Li, S., Hu, J., Cui, Y., & Hu, J. (2017). DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics, 117*(2018), 721–744.
Lissoni, F., Montabio, F., & Zirulia, L. (2013). Inventorship and authorship as attribution rights: An enquiry into the economics of scientific credit. *Journal of Economic Behavior and Organization, 95*, 49–69.
Magerman, T., Looy, B. V., & Debackere, K. (2015). Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology. *Research Policy, 44*(9), 1702–1713.
Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal, 41*(9), 1572–1594.
McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426
Mention, A. (2011). Co-operation and co-opetition as open innovation practices in the service sector: Which influence on innovation novelty? *Technovation, 31*(1), 44–53.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.
Motohashi, K. (2019). Science and technology co-evolution in AI: Empirical understanding through a linked dataset of scientific articles and patents. RIETI Discussion Paper 20-E010.
Motohashi, K., Koshiba, H., & Ikeuchi, K. (2019). A method of extracting content information from patent documents and comparison of their characteristics by applicant type by using the vector space model of distributed expressions, NISTEP Discussion Paper 175, December 2019, NISTEP, Japan (**in Japanese**).
Motohashi, K., & Muramatsu, S. (2012). Examining the university industry collaboration policy in Japan: Patent analysis. *Technology in Society, 34*(2), 149–162.
Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics, 7*, 368–381.
Pisano, G. (2006). *Science business: The promise, the reality, and the future of biotech*. Harvard Business School Press.
Schmoch, U. (1997). Indicators and relations between science and technology. *Scientometrics, 38*(1), 103–116.
Schomoch, U. (2008). Concept of a technology classification for country comparisons: Final report to the World Intellectual Property Organization (WIPO), Fraunhofer Institute for Systems and Innovation Research, Karlsruhe, Germany, June 2008.
Stokes, D. E. (1997). *Pasteur's quadrant—Basic science and technological innovation*. Brookings Institution Press.
Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy, 48*(6), 1362–1372.
Younge, K., & Kuhn, J. (2016). Patent-to-patent similarity: Vector space model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2709238