RIETI Discussion Paper Series 21-E-025

New Indicator of Science and Technology Inter-Relationship by Using Text Information of Research Articles and Patents in Japan (Revised)

MOTOHASHI, Kazuyuki
RIETI

KOSHIBA, Hitoshi
NISTEP / AIST

IKEUCHI, Kenta
RIETI

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 21-E-025
March 2021
Revised version August 2021

New indicator of science and technology inter-relationship by using text information of research articles and patents in Japan[1]

Kazuyuki Motohashi (RIETI)
Hitoshi Koshiba (NISTEP, AIST)
Kenta Ikeuchi (RIETI)

Abstract

In this study, the text information of academic papers (about 2.3 million) published by Japanese authors and patents filed with the Japan Patent Office (about 12 million) since 1990) 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 extracted using cosine similarity. A time trend and sector specific linkage of science and technology are identified by using the count of neighbor patents (papers) for each paper (patent). It is found that the number of papers with similar contents of patents decreased over time while the trend of patent counts with similar contents of paper is relatively stable. It is also found that the scope of scientific discipline by papers is relatively stable, while the technology fields by patents shows more dynamic patterns. This paper proposes a new methodology of measuring science and technology interlinkage by using textual information as a complement to traditional indicators based on non-patent literature citations of patents.

Keyword: Text analysis; patent information; research paper; science and technology linkage

JEL Classification：O31, O34

---

[1] This study was conducted as part of the “Digitalization and Innovation Ecosystem: Holistic Approach” project 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. An early version of this paper was published as NISTEP Discussion Paper No. 175 (in Japanese). The authors also acknowledge financial support from MEXT/JSPS KAKENHI (Grant Numbers: 18H03631, 18K12787).

1. Introduction

The growing importance of scientific knowledge in innovation can be seen across many industries. In the pharmaceutical industry, an 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 progress of genomic science. In the electronic device industry, as the LSI production process is refined, understanding the characteristics of nanoscale materials has become indispensable. Additionally, recent advances in machine learning (AI) using big data have caused processes of innovation to evolve not only in the manufacturing industry but also in the financial and service industries. In these fields as well, scientists at universities and public research institutes play an important role (Motohashi, 2019).

Hitherto, non-patent literature citations have been used to measure the proximity between innovation and science (Narin and Norma, 1985; Schmoch 1997). These non-patent literature citations were considered to be an index showing the extent to which an invention, filed as a patent, was created based on discoveries in the scientific papers cited in the patents. This index is known as the scientific linkage of the patent. However, while this index presents the relationship of science toward invention, it is not able to show the two-way relationship between science and innovation. As an analogy of science linkage using information from citations, one conceivable idea is to use the information of patents cited in research papers. However, the nature of citations that scientific papers require differs from that of patents, where the novelty factor of the invention is the focus, that is, scientific papers, which fulfill the requirements for scientific knowledge such as objectivity and replicability, tend to be used as the citations that form the basis of scientific developments, and it is unusual for patents, which may overlook those principles as long as the ideas are novel and have industrial applicability, to be cited. Therefore, it is not possible to use citation information to uncover the relationship inventions have toward science.

Another approach to show proximity between patents and scientific papers is to search for pairs of research paper and patents that express the same findings or inventions. There are methods such as extracting patents and research papers published simultaneously (Lissoni et al., 2013) or using text mining to identify patents and research papers with similar content (Magerman et al., 2015), and these have been used for analyzing academic inventors (e.g., the impact of patents on research paper productivity). Also, there are research results from a study where a large-scale database was created, connecting research paper authors with patent applicants, and science–technology linkage was measured using information on pairs of patents and research papers from the same researcher (Ikeuchi et al., 2015).

In this study, we chose the latter approach of using text mining to extract pairs of research papers and patents similar to one another and performed a comprehensive analysis on how they relate to the advancement of science and technology in Japan. Specifically, we used titles and

abstracts from research papers and patents published by Japanese authors and inventors in the years 1990–2018. We grouped the documents with high-content similarity and clarified the mutual relationship between research papers and patents in the context of the advancement of science and technology. In this paper, Chapter 2 describes our method of analysis and Chapter 3 presents an outline of our obtained data, as well as the results from a cluster analysis. In Chapter 4, we use citation information from research papers and patents to perform an evaluation of a similarity index via text mining—the method we use in this present paper. In Chapter 5, we present a relation index of science and technology and show the trends in the relationship between the two in Japan. Finally, we present our conclusion and describe future issues to be examined.

2. Data sets and text mining techniques

2.1. Data sets

In this paper, to comprehensively observe the interlinking relationship between Japanese science and technology, we used the following data sets:

- Research paper information: Papers included in Science Citation Index expanded from Clarivate's Web of Science, published between 1990 and 2017, and containing at least one Japan-based author.
- Patent information: Patents filed with the Japan Patent Office and included in the PATSTAT2020 Spring Version (those for which English-translated title and abstract information are available).

Regarding the number of documents, we used 2,342,987 research papers and 12,037,068 patents, for a total of14,380,055 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.

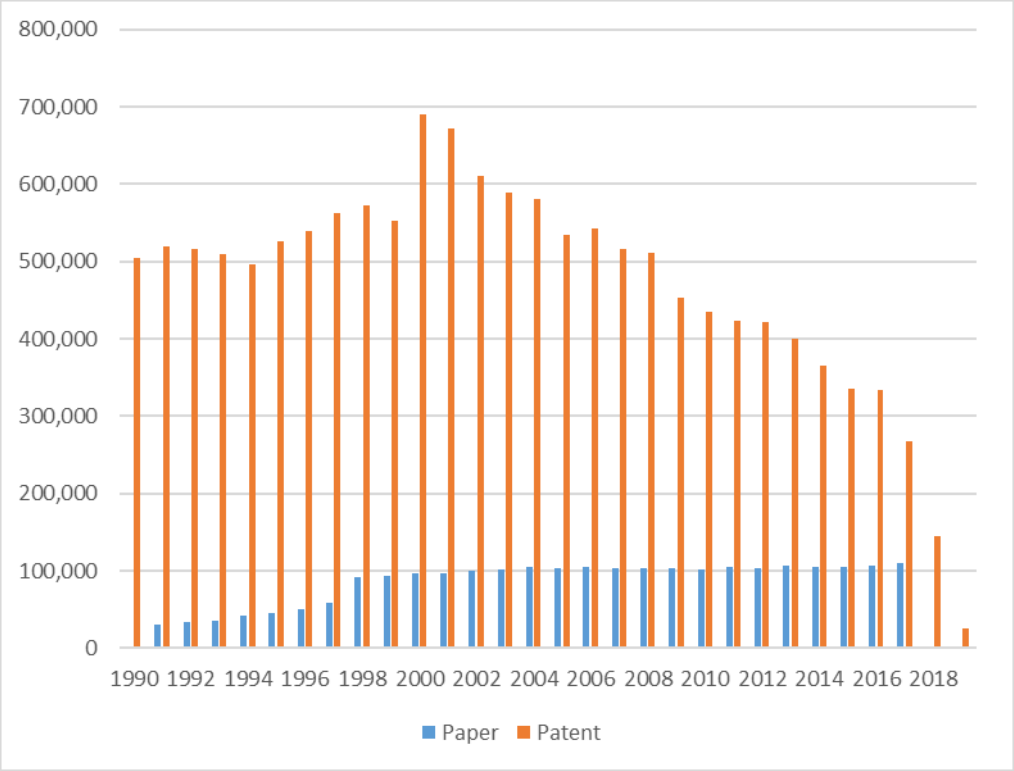

Figure1 : Numbers of papers and patents by application/publication year

2.2. Text mining method and clustering results

To create document embedding vectors that represent the content of each document, we used two steps. We first created embeddings for each word and then aggregated them by document. First, we extracted only the nouns that appear in a total of ~14.3 million titles and abstracts and used FastText (Joulin, 2016; Bojanowki, 2017) to create embedding vectors for words other than common words and rare words. Next, we took the average of these word vectors to obtain a document embedding vector for each document. Regarding embedding results for the words, we conducted cluster analysis using the K-means method and confirmed, by visual checking, that semantically similar words belonged in the same cluster (for details, see Motohashi, Koshiba, and Ikeuchi, 2021).

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 UMAP (McInnes et al., 2018). The result is shown in Figure 2.

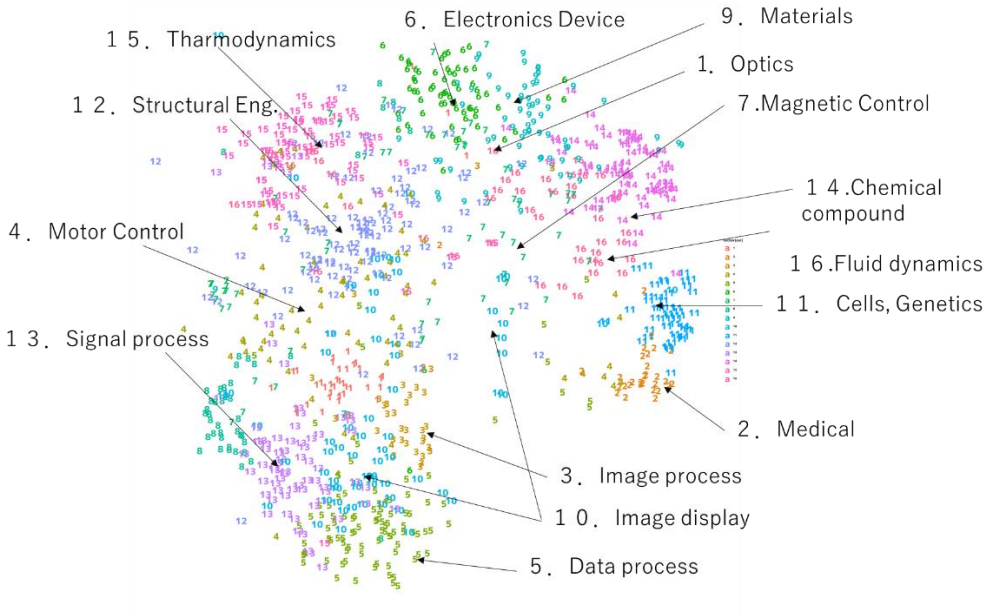

Figure 2: Visualization of document distribution with clustering results

2.3. Visualizing the relationship between research papers and patents

Figure 2 shows the results of a technical mapping that combines both research papers and patents. Figures 3-1, 3-2, and 3-3 distinguish between research papers and patents and show the change over time. In these figures, red indicates the location of patents, while 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 are also seen distributed across the fields of optics, fluid processing, and video display. However, fields relating to mechanics (e.g., motion control, structural mechanics, and thermodynamics), electronic devices, and image processing are mostly covered by patents.

With regard to changes over time, differences can be seen particularly between the 1990s and 2000. As the number of research papers is increasing with respect to the total number of documents, expansion of research papers in technical fields can be observed. This trend is particularly notable in the fields of chemistry/materials (compounds, metallic materials). It is also observed that research papers have been published in fields such as thermodynamics, which was previously only covered by patents. The relationship between research papers and patents per field will be analyzed in detail in Chapter 5.

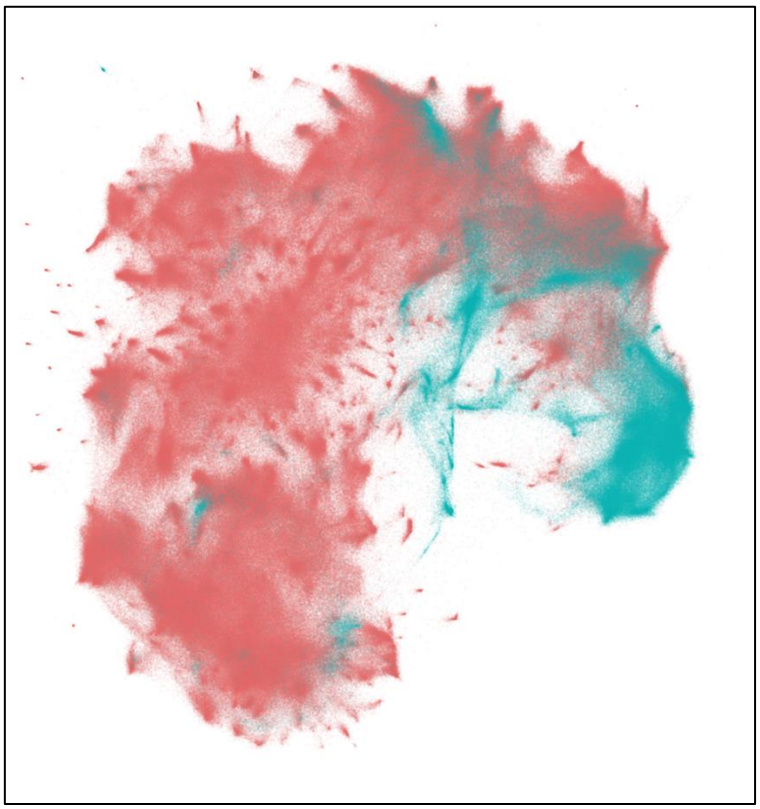

Figure 3-1: Paper and Patent Mapping (1990's)

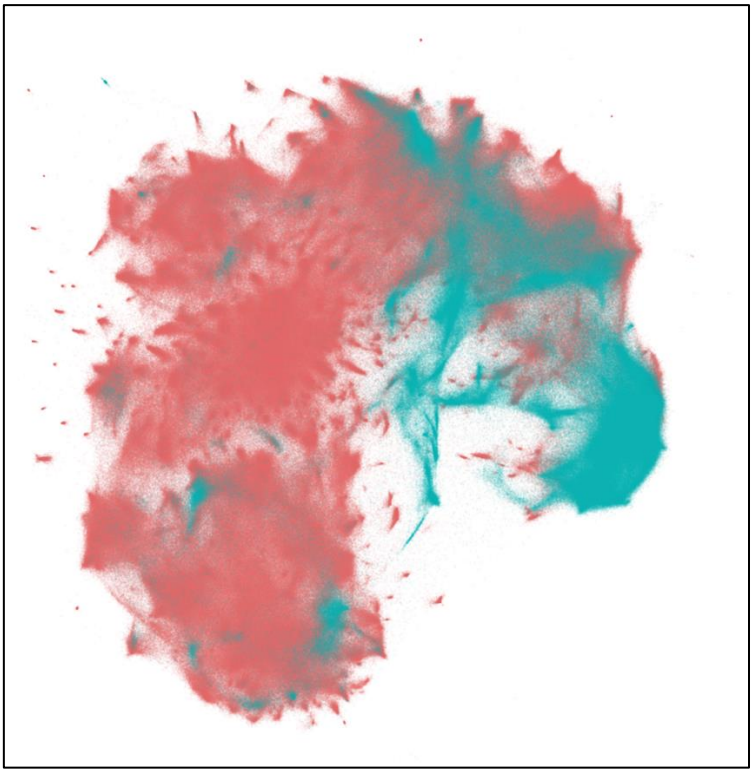

Figure 3-2: Paper and Patent Mapping (2000's)

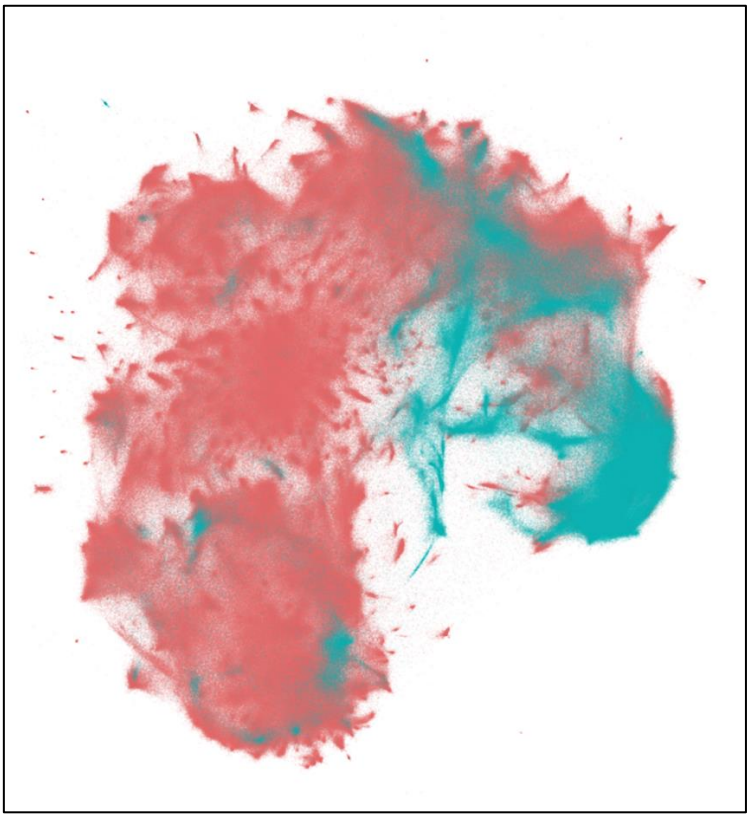

Figure 3-3: Paper and Patent Mapping (2010's)

3. Characteristics of document embedding data

Here, we analyze the characteristics of the document embedding data created from the text information of research papers and patents. As presented in Chapter 3, using the results from the cluster analysis, the embedding information for the words were evaluated visually (we confirmed that similar words were gathered in the same cluster). Here, we attempt a more formal quantitative evaluation. Specifically, using citation information from papers and patents, we confirm that the similarity between citation pairings (pairings of cited documents with the documents that include the citation) is significantly increased. We also confirm, using information from the JSPS Kakenhi "Report on the Research Results," that there is a high degree of similarity between research papers and patents generated from the same-Kakenhi project. Finally, using information on neighboring literature by NGT, we also examine whether cosine similarity between documents is affected by distribution in the technical space.

First, as contrasting samples to show that, for citation pairs, cosine similarity is significantly high for results from the same research project, we randomly selected 10,000 pairs for the three patterns of "paper–paper," "paper–patent," and "patent–patent," and looked at the distribution of cosine similarity. In Figure 4, the decile values for 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). Also, looking at 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). The reason is that the word vectors are not used directly, but the embedding data is used, so there is a certain degree of correlation between the words in the first place. However, compared to the results (the median of randomly extracted cosine similarities was ~0.5) of conducting the same processes using patent documents written in Japanese (Motohashi, Koshiba, and Ikeuchi, 2019), cosine similarity is higher, showing that there is room for improvement in the pretreatment (removal of stop words, n-gram conversion of technical terms, etc.) of the literature written in English used in this present paper.

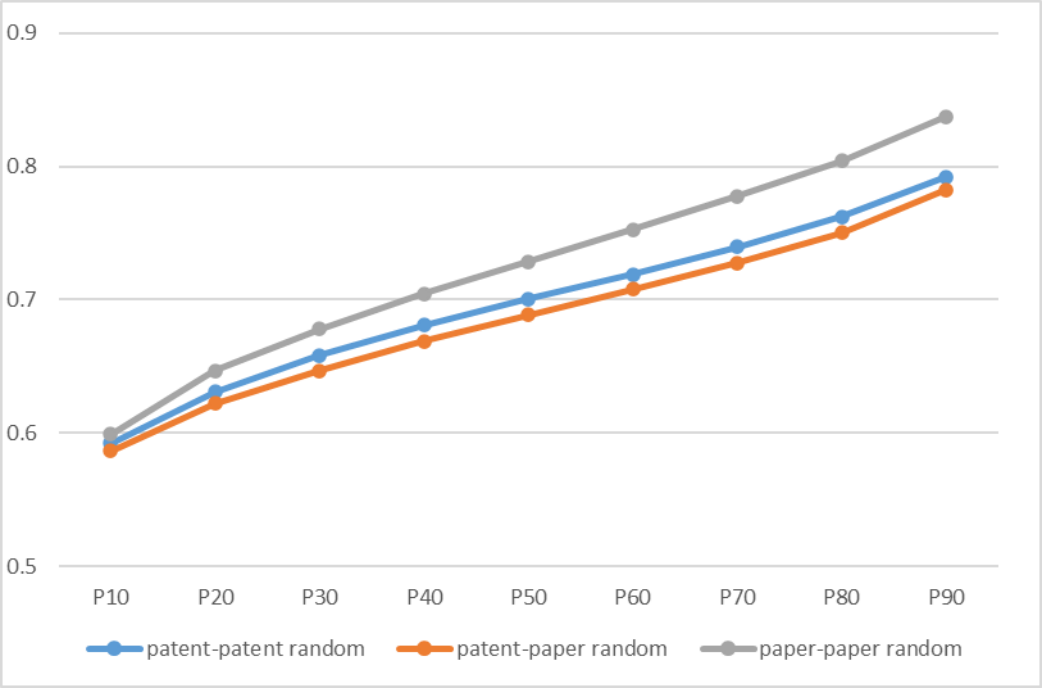

Figure 4: Distribution of cosine similarity of random sampled pairs

Next, we compare the distribution in Figure 4 with the cosine similarity between documents for citation pairs and same-Kakenhi-project outcomes. Figures 5-1, 5-2, and 5-3 show the circumstances for the “paper–paper,” “paper–patent,” and “patent–patent” pairs, respectively. In all pairs, the cosine similarity between citation pairs and same-project outcomes is higher compared to random pairings, confirming the validity of the document embedding.

Also, looking at the difference between document types, 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, 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.

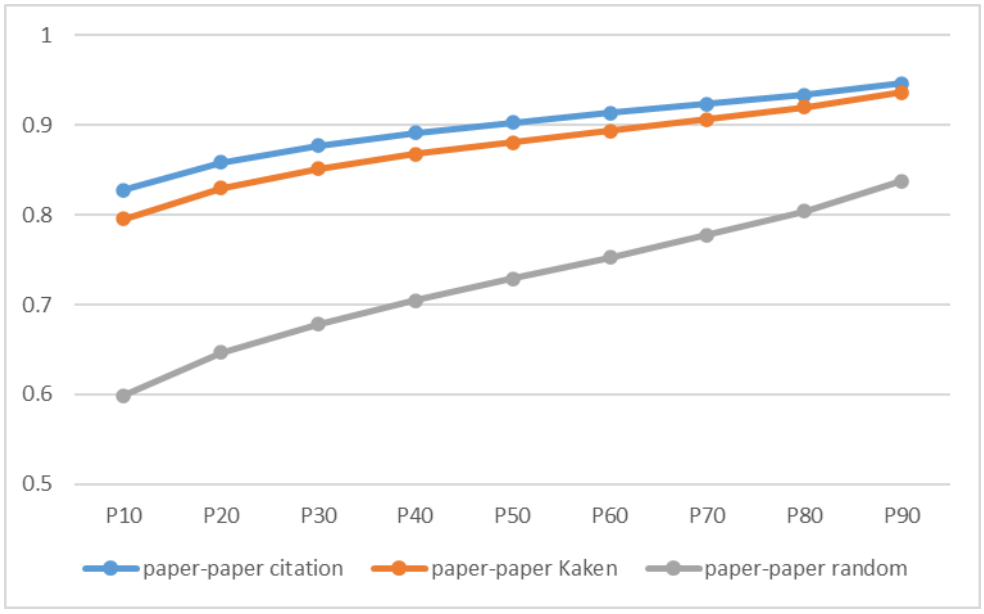

Figure 5-1 : Comparison of cosine similarity distribution (paper-paper)

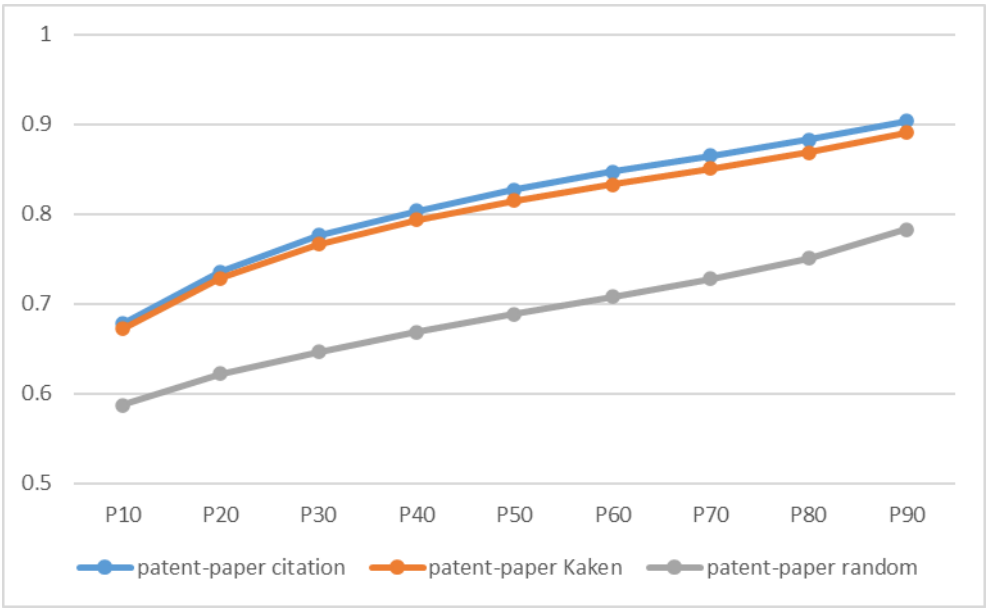

Figure 5-2 : Comparison of cosine similarity distribution (paper-patent)

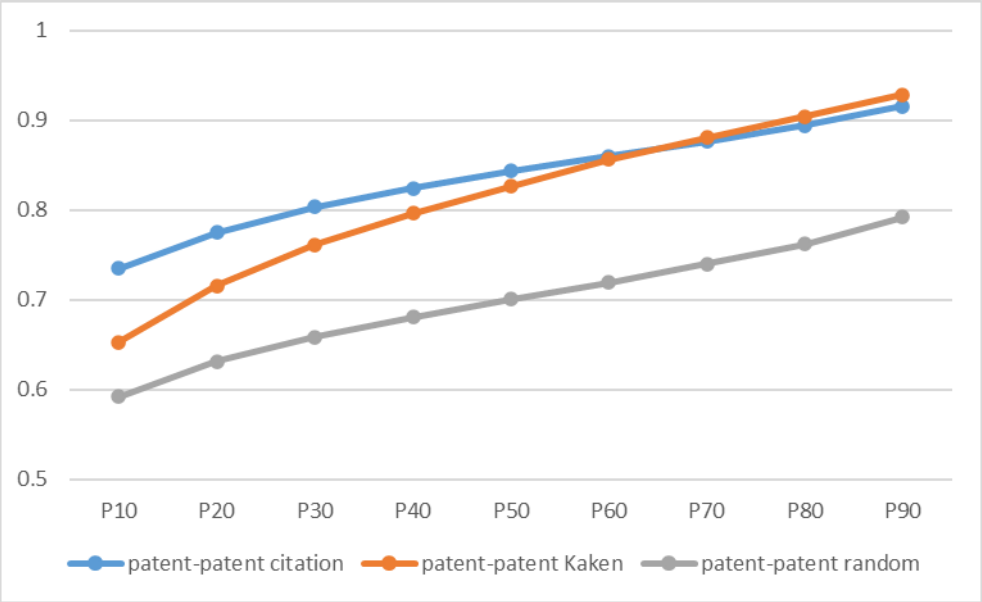

Figure 5-3 : Comparison of cosine similarity distribution (patent-patent)

Finally, we will look at the characteristics of cosine similarity of neighboring documents, using NGT. The NGT is an algorithm where a specific quantity of neighboring documents is searched for out of a large number of documents. In this present study, we extracted 200 neighboring documents based on cosine similarity. Table 1 shows the distribution of cosine similarity for the 100th and 200th documents. Firstly, one will notice that the cosine similarities between the 100th and 200th document are almost the same (for example, median value is 0.899 for the 100th and 0.893 for the 200th). This is because the document embedding vector has a high number of dimensions: 300 dimensions (the relationship between the radius and volume of a 300-dimension supersphere). Additionally, 0.9, which is the median value for cosine similarity of 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, indicating similarity regarding content.

	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

Table 1: Distribution of the cosine similarities of neighboring documents for the 100th and 200th documents.

The difference in cosine similarities with the neighboring documents of the 200th document is attributable to the difference in distribution density in the technical space where the research papers and patents are distributed. Documents with high-cosine similarity with the 200th document indicate that research papers and patents are more densely distributed around the 200th document. 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 Figure 6, neighboring documents are divided into four groups based on their cosine similarities with the 200th document (groups Q1–Q4, starting with documents with low cosine similarity, i.e., sparse technical spatial density), 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. Also, the effect of spatial density is greater in the groups of sparse density (e.g., Q1).

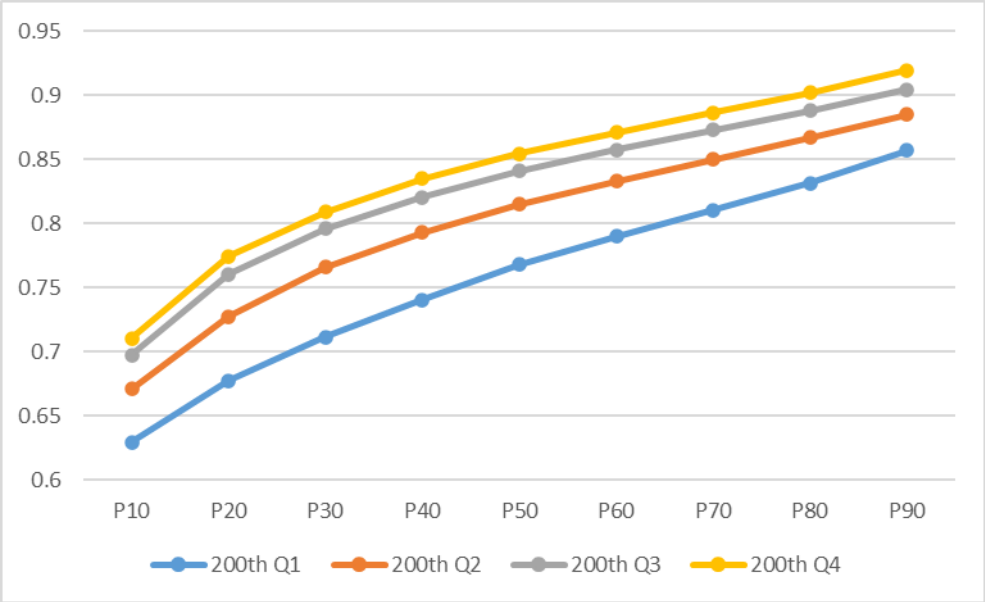

Figure 6: Cosine similarity distribution of citation paper by content density

4. Correlation analysis of science and technology according to neighboring document information

For each paper or patent, it is possible to analyze the correlation between science (papers) and technology (patents) by counting the number of neighboring papers/patents near the technical space. Here, the average number of neighboring patents is calculated by the academic field of the research papers (22 categories based on the classification of academic fields by WoS) and by the technical field of the patents (35 categories in accordance with WIPO). As for the neighboring documents, documents with a cosine similarity of 0.9 or higher were selected from the first 200 documents. With regard to the cosine similarity (based on the document embedding of papers and patents) being 0.9 or higher, this number is the value of the highest 10th percentile of the cosine similarities of "paper–patent" citation pairs or same-Kaken-project pairs (P90 in Figure 5-2). This means documents with a high degree of content similarity are being extracted. Additionally, as Table 1 shows, half or more of the total documents neighboring the 200th document have a cosine similarity of 0.9 or higher (the median cosine similarity for the documents neighboring the 200th document is 0.911). These patents can be considered to be located in a place of relatively high-document density in the technical space. The number of neighboring patents will be determined in advance when performing NGT, and in this study, only neighboring patents within the first 200 documents were extracted. Thus, the 200th document was set to be the limit for the practical reason of data constraint. However, if we handled all neighboring documents with a cosine similarity of 0.9 or higher, there would potentially be an enormous quantity of documents. To eliminate outliers that would have a significant effect on the whole, our view is that, either way, it is necessary to set

a threshold with respect to the number of documents that neighbor one document.

Firstly, we present the overall trend regarding the correlation between science and technology in Figure 7. In this graph, the samples are divided into three categories (1990s, 2000s, and 2010s), according to the year of publication (year of application) of the paper (patent), the mean number of patents (papers) for each paper (patent) is aggregated for each technical field (academic field), and the average value between fields is taken. In order to mitigate the bias associated with the time trend of paper and patent counts, 20,000 samples are selected randomly for each of publication year of paper and patent from 1991 to 2017 (20,000*2*27=540,000 in total) for subsequent analysis.

First, there are higher numbers of neighboring papers (15 to 20 out of 200 papers/patents) around a patent as compared to that of neighboring patent (around 5 out of 200 papers/patents). It is natural to see this since the paper is distributed in limited field in technology space (such as cells/genetics and medical) as compared to the patent (Figure 2). In addition, it is found that the number of neighboring papers per patent is decreasing (the red line), while the number of neighboring patents per paper shows a bit of up and down pattern (the blue line). That is, the overlapping field of science (paper) and technology (patent) is shrinking from the viewpoint of technology side, but it is no so that extent from the viewpoint of science side.

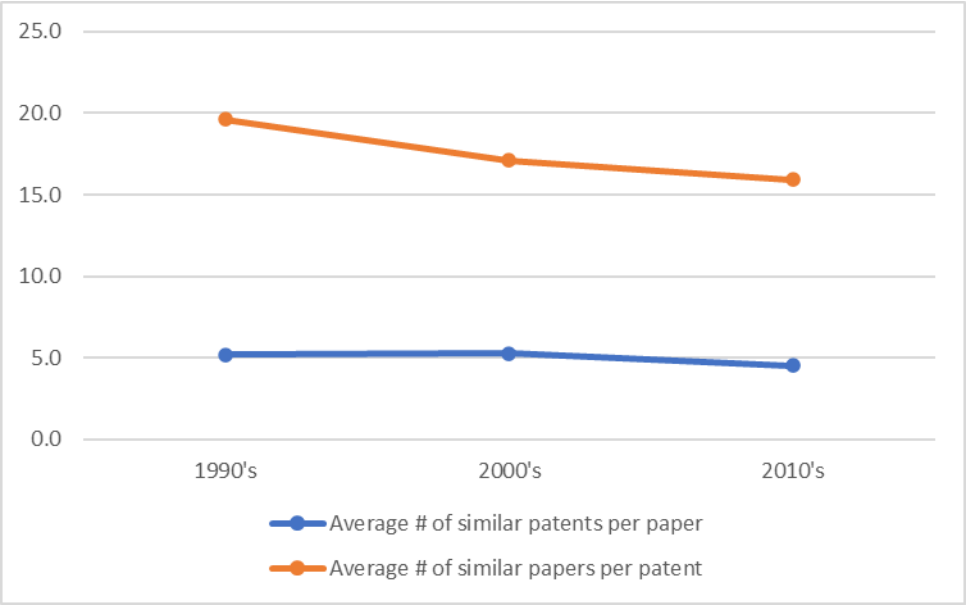

Figure 7: Trend of neighbor patent(paper) counts

In order to look at this point in more detail, we separate each patent/paper intensity by the timing of publication of neighboring papers (Figure 8-1) and patents (Figure 8-2). It is natural to presume that the neighboring patents (papers) at the same timing of compared papers (patents) would

contribute more to the aggregated each intensity. The fact that the share of contribution of 1990's neighbor papers and patents (blue portion in each figure) decreases is consistent with such presumption. However, it should be noted that the shares of newer neighbor papers are relatively large in Figure 8-1. For example, the contribution of 1990's papers, 2000's papers and 2010's papers in 1990's patents are 6.5, 7.5 and 5.6, respectively. In addition the contribution of 1990's patents is the largest among that of other decades in Figure 8-2, suggesting the larger portion of neighbor patents around a paper is relatively old ones.

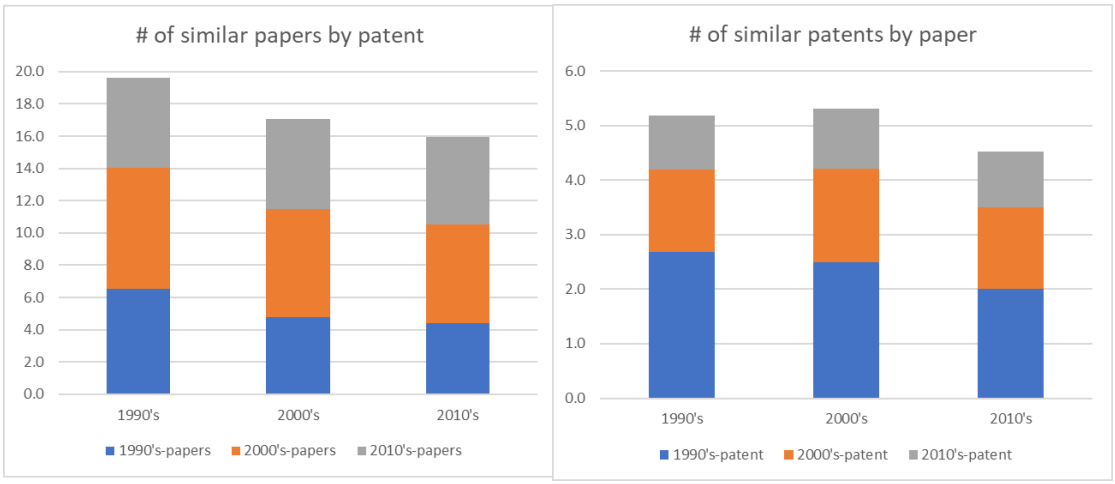

Figure 8 -1 and Figure 8-2: Trend of neighbor paper (patent) counts by cohort

The presumption that overlapping field of science and technology comes from the same timing of publication is based on the idea of both science and technology progressing and evolving parallelly. However, what is found in the above two graphs is that scientific fields covered by papers are relatively stable, while technology fields by patents are more dynamic in the technology space defined by our document embedding results.

Finally, we look at the science and technology correlation intensity by scientific and technology fields. As well as showing relative difference of intensity across the fields, we look into the timing of publication between neighboring papers (patents) and the patent (paper). If the neighboring document is published before the relevant document, it can be interpreted that the relevant document could be created based on this neighboring document. Conversely, if the neighboring document is published later, it can be interpreted that the relevant document influenced the neighboring document.

It is found that "analysis of biological materials", "biotechnology" and "pharmaceutical" are top scientific fields where many neighboring patents can be found (Figure 9-1), while in "energy", "engineering", "chemical engineering", "computer science" and "material", relatively larger share of neighboring papers are found (Figure 9-2). In addition, there is an academic field (computer

science) with its “AFTER” region larger than its “BEFORE” region and some academic fields with the opposite tendency (materials engineering, chemistry) in Figure 9-1. Regarding the former, it can be interpreted that it is a field with a strong tendency for scientific progress to influence technology (patents). Regarding the latter, it can be interpreted that these are fields where more scientific development is observed in areas where technological progress is made. Meanwhile, when looking at the number of neighboring research papers by technical field, the ratio of “BEFORE” and “AFTER” is well balanced in most fields.

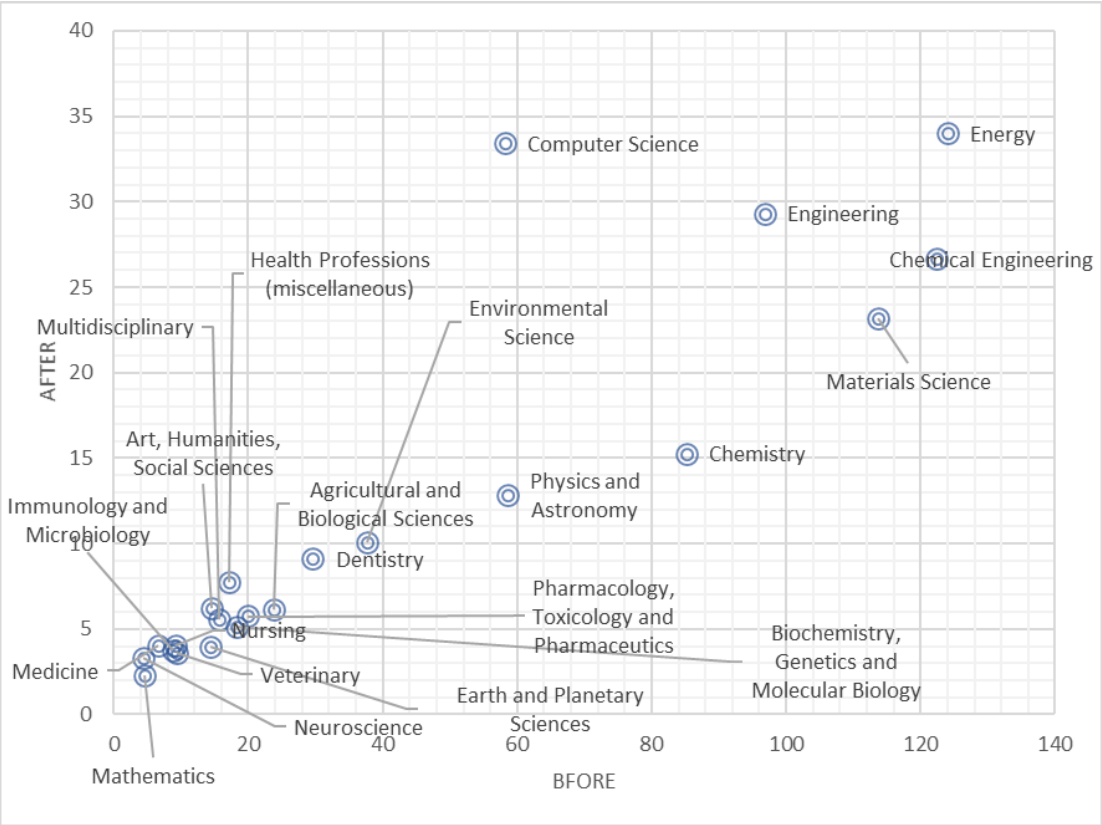

Figure 9-1 : Number of neighbor patents by timing of application (before and after)

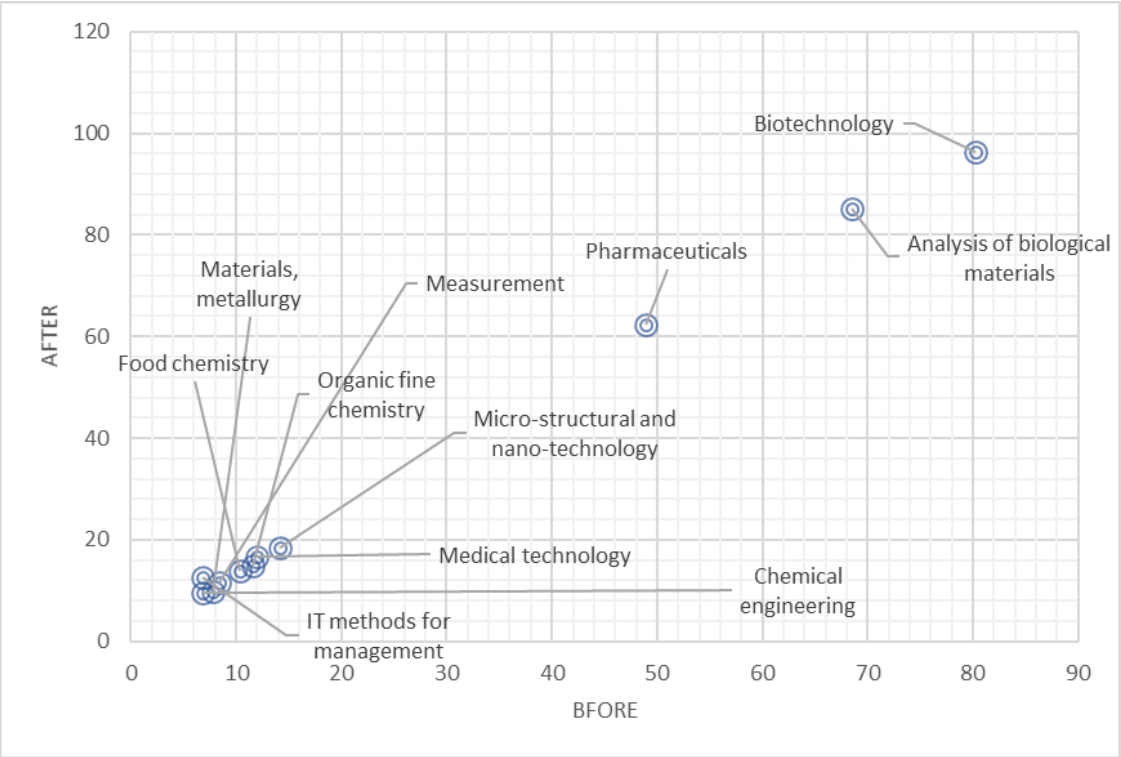

Figure 9-2 : Number of neighbor papers by timing of publication (before and after)

5. Conclusion

In this paper, we analyzed the two-way relationship between science (research papers) and technology (patents) using text data from 2.3 million published papers and 12 million filed patents, since the year 1990. Specifically, we created document embedding vectors using the titles and abstracts for each document and used cosine similarity to extract neighboring documents (documents with a cosine similarity of 0.9 or higher and also within the top 200 similar documents). The relationship between research papers and patents was quantified using the number of neighboring patents (research papers) for each research paper (patent).

As a result, we observed a trend where the number of research papers in the vicinity of patents decreased, and the number of research patents in the vicinity of research papers is relatively stable, from the 1990s to the 2000s and 2010s. We interpret this trend to represent that, overall, scientific fields covered by papers are relatively stable, while technology fields by patents are more dynamic in the technology space defined by our document embedding results. We also found that the scientific fields closely relating to technology were chemical engineering, materials engineering, energy, engineering, chemistry, and computer science, and the technical fields closely related to science were biomaterials, biotechnology, and pharmaceutics. The latter finding has been shown in previous literature using non-patent document (research paper) citations in patents. However, the former finding can be considered to be a new finding that is not found in previous research.

The purpose of this present paper is to show a comprehensive trend regarding the progress of science and technology in Japan. The embedding information gained from a large quantity of research papers and patents can be used for a variety of applicative studies in the future. In Japan, major institutional reforms were conducted in the 2000s, such as national testing laboratories becoming independent administrative agencies in 2001, and national universities becoming incorporated national universities in 2004. We surmise that these institutional reforms impacted the trends of science–technology interrelation shown in this paper. An analysis on the institutions to which research paper authors and patent applicants belong, and on the relationship between the two, can be expected to lead to useful insights on the impacts of these institutional reforms.

One promising future topic might be research, where the most recent natural language processing methods are used to acquire more accurate expressions of content and these are then applied to examine the trajectories of science and technology. In our present study, we chose a bag of words approach, where we obtained embeddings for single words and aggregated them by document. However, using a method such as Bi-directional Encoder Representation with Transformation (BERT), which has been used with increasing frequency in recent years, will make it possible to also include context information of the words into their embedding information. Implementing BERT, which is a deep learning model with a large number of parameters, will require a huge amount of computer resources; however, in recent years the Google team has made public a BERT learning model that is based on patent literature from around the world (Srebrovic and Yonamine, 2020). Based on the estimation results here, in the future, we would like to consider research that delves into the details of content-proximity.

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017), Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146, arXiv:1607.04606

Goto, A. and K. Motohashi (2007) “Construction of a Japanese Patent Database and a first look at Japanese patenting activities,” Research Policy, 36(9), 1431-1442.

Ikeuchi, K. Motohashi, R. Tamura and N. Tsukada (2017), Measuring Science Intensity of Industry using Linked Dataset of Science, Technology and Industry, RIETI Discussion Paper, 17-E-056

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T (2016).: FastText.zip: Compressing text classification models, arXiv preprint, arXiv:1612.03651

Lissoni, F, F. Montabio and L. Zirulia (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., B.V. Looy and K. Debackere (2015) “Does involvement in patenting jeopardize one’s academic footprint? An analysis of patent-paper pairs in biotechnology,” Research Policy, 44(9), 1702-1713.

McInnes, L., Healy, J., and Melville, J (2018).: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv preprint (2018), arXiv:1802.03426

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 and 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)

Narin, F. and E. Noma (1985) “Is technology becoming science?” Scientometrics, 7, 368-381.

Schmoch, U. (1997) “Indicators and relations between science and technology,” Scientometrics, 38(1), 103-116.

Srebrovic, R. and Yonamine J. (2020), Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery, November 2020, Google Cloud Blog, How AI improves patent analysis | Google Cloud Blog