RIETI Discussion Paper Series 23-E-024

Identifying Technology Opportunity Using a Dual-attention Model and a Technology-market Concordance Matrix

MOTOHASHI, Kazuyuki
RIETI

ZHU, Chen
University of Tokyo

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

RIETI Discussion Paper Series 23-E-024
March 2023

Identifying Technology Opportunity Using a Dual-attention Model and a Technology-market Concordance Matrix*

Kazuyuki Motohashi (RIETI, The University of Tokyo)

Chen Zhu (The University of Tokyo)

Abstract

To understand the role of new technologies in innovation, it is crucial to develop a methodology that links technology and market information. Conventionally, the relationship between technology and the market has been analyzed using a technology-industry concordance matrix, but the granularity of market information is confined by industrial classification systems. In this study, we propose a new methodology for extracting keyword-level market information related to firms' technology. Specifically, we developed a dual-attention model to identify technical keywords from firms' websites. We then vectorized the market information (extracted keywords) and technology information (patents) using word embedding to construct technology-market concordance matrices. Matrices were generated based on a group of high-growth companies that suggest new technologies and market opportunities in the automotive, electronics, and pharmaceutical industries.

Keywords: technology opportunity discovery; dual attention model; technology market concordance

JEL codes: C45, O32

* This study was conducted as part of the project "Research on innovation ecosystem formation process" 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 financial support from MEXT/JSPS KAKENHI (Grant Number: 18H03631)

1. Introduction

Technology opportunity discovery (TOD) has become an essential process for most enterprises by which they can extend their technology and product portfolio and understand the potential threats in the surrounding environment. The existing studies on TOD analysis can be divided into two mainstreams: a) identifying and exploring emerging technologies and business fields (Kwon et al., 2017; Lee et al., 2022) and b) diversifying into new areas by exploiting existing technologies and products (ETPs), which is akin to the concept of concentric diversification (Lee et al., 2008; Yoon et al., 2015; Kim et al., 2017). As opposed to the former, which stresses radical innovation, opportunity development based on ETPs is more like an incremental innovation process with a lower uncertainty risk and a higher success rate, which forms a more sustainable growth path for companies (Cantwell and Piscitello, 2000; Zook and Allen, 2001). Unlike small businesses whose flexibility breeds their innovation competitiveness, radical innovation ideas can hardly survive the bureaucratic layers of large firms, where corporate inertia results in a strong resistance against undertaking novel and risky projects (Audretsch and Vivarelli, 1996). In this regard, while exploring emerging fields provides firms a way to comprehend the novel trends in the fields, TOD approaches that exploit ETPs would be a more pragmatic and predictable strategy that large firms can implement straightaway to capitalize opportunities and pursue stable growth.

Chronologically, TOD analysis has been initialized at Georgia Tech since 1990 to gain insight into emerging technological fields using bibliometric analysis of patent data, augmented with experts' judgment (Porter and Detampel, 1995). Later on, as the volume of data increased tremendously, methodological studies on TOD shifted towards developing automated analytic tools, particularly by using text mining techniques (Yoon et al., 2014; Lee et al., 2020; Lee et al., 2022). Among various TOD methods, patent data, which are well structured to store technological knowledge across a broad spectrum of fields, has become the primary source for studying technological opportunities. Admitting patent data can be a good proxy for monitoring technological advancement and innovation activities, but it has limited ability to portray the market-side information.

To overcome this, Arora et al. (2013) conducted a web content analysis to study the commercialization process of emerging technologies by crawling information from companies' websites. Park and Geum (2022) collected technical news articles as complementary materials to reflect market-side opportunities. Indeed, with the broad expansion of the Internet, it has become normal for firms to exhibit their commercial activities (i.e., introducing products and services, key technologies and personnel, etc.) through publicly accessible websites, which can thus be a valuable supplement to exploring market-side opportunities (Gök et al., 2015). Furthermore, Kinne and Axenbeck (2020) use hyperlink information from web pages to analyze the network structure of AI start-up networks in Germany.

This paper combines companies' patent information with their publicly available websites to investigate technology opportunities linked with their product/service activities. Web data can create substantial chances for researchers to study firm-level innovation activities since (1) the massive amount of business activity information disclosed through

company websites in recent decades; and (2) it is an unobtrusive data source that can be accessed without contacting research subjects directly (Gök et al., 2015). However, unlike well-structured patent data, company websites often include miscellaneous information, ranging from primary business information (i.e., products and services) to more general information (i.e., a factory site and recruitment). The noise information embedded in the raw web data may blur the results of the TOD analysis. To address this issue, we propose a dual-attention model that can automatically identify technical information (i.e., products or firms' technological capabilities) from the raw web data. The attention mechanism has widely been used in computer vision (CV) and natural language processing (NLP). The mechanism is self-explanatory, miming cognitive attention that concentrates more on important parts of the data. Existing studies have demonstrated that the attention mechanism is capable of extracting related keywords for a given training objective (Tang et al., 2019; Sun and Lu, 2020). Building on this, we introduce a dual-attention structure where two attention layers are applied at the webpage-level and word-level, respectively, to extract desired webpages and technical keywords. We then show the extracted web information can outperform the raw data through a classification task.

Finally, we create a novel indicator by using extracted web data associated with the patent information to analyze technological and business opportunities for technical fields. Historically, the interactions between patent technology and industrial outputs are analyzed by using the concordance matrix of technology classification (e.g., International Patent Classification [IPC]) and industry classification (e.g., International Standard Industrial Classification [ISIC]) (Johnson, 2002; Schmoch, 2003). In the existing methodology, based on the linked dataset of patent and industrial output at the patent applicant level, the granularity of matrix information is constrained by the technology/industry classification of the original dataset. Instead, our approach, based on textual information from both patents and web pages, allows us to investigate the relationship between technology and market at a finely grained product/service level.

The remainder of this study is organized as follows: Section 2 presents the research background, reviewing existing studies on technology opportunities as well as the empirical methodologies used in this paper. Section 3 explains the data and methodology. Section 4 applies this methodology to web pages and patent information of publicly traded firms in Japan. In this section, as an application of our methodology to link product/service with related technologies, new technology and market opportunities for automotive, electronics, and pharmaceutical industries are analyzed. Finally, Section 5 concludes this study with management implications and our research limitations.

2. Literature review

2.1. Technology opportunity discovery

Technological opportunity is defined as the chance to advance a particular technological field (Klevorick et al., 1995). Olsson (2005) discusses technological opportunities in terms of three types of technological advancement: incremental innovation, radical innovation, and discoveries. For incremental innovation, enterprises leverage technological or commercial opportunities in close proximity to their current business

activities. Studies in this field typically address TOD's high success rate and low risk, making it highly valued by profit-driven investors. Lee et al. (2008) proposed an effective technology road mapping (TRM) approach to support technology and product planning decision-making using text mining techniques on patent data. TRM-related approaches are inherently designed for incremental innovation because they assume that improvements should be made to existing products. Yoon et al. (2015) outlined four TOD paths derived from an organization's ETPs, including new product launches (existing technology to applicable products) and new generation technology development (existing products to adoptable technology). The two listed paths are indeed analogous to the concept of concentric diversification, a detailed discussion by Kim et al. (2017), highlighting that diversification strategies built on companies' core competencies can result in higher commercial success. Lee et al. (2020) introduced a product landscape analysis using word2vec to identify technological opportunities embedded in firms' existing products. Another stream of research focuses on radical innovation and discoveries, emphasizing the exploration of emerging technologies. Radical innovation and discoveries are characterized by risk and uncertainty; however, they also present opportunities to create a new technological paradigm and, therefore, yield substantial economic benefits (Dosi, 1988). Porter and Detampel (1995) pioneered the use of a bibliometric analysis of patent data to identify emerging technologies. To demonstrate that the scientometric approach can supplement conventional expert-centric approaches (e.g., Delphi), Bengisu and Nekhili (2006) conducted a keyword-based search of publications and patents to validate emerging technologies forecasted by experts. Text mining techniques were employed to automate this process. Kwon et al. (2017) applied a latent semantic analysis to provide an overview of emerging technologies. Lee et al. (2022) exploited a deep learning-based approach associated with network indicators to discover new technological opportunities.

Notably, patent data dominate TOD analysis because they store a broad spectrum of technological knowledge. Yoon et al. (2014) explored technological opportunities by constructing technology and product morphologies from collected patent data. Wang and Chen (2019) detected potential opportunities based on the notion that emerging technologies are outliers in a patent corpus. Undeniably, patent data are useful for TOD analysis; however, a single data source cannot present the most up-to-date and non-technological information (e.g., market-side information). To address this issue, Kwon et al. (2018) suggested a morphological analysis using Wikipedia data to study the generation of new ideas. Other studies have shown that technical news and company websites can be informative supplements for patent data (Arora et al., 2013; Park and Geum, 2022).

2.2. Web mining

Researchers rely on web data because of (1) the tremendous amount of publicly available information disclosed through websites and (2) the unobtrusive data collection method that avoids potential bias induced by direct contact (Webb et al., 1966). Web mining techniques can be divided into three broad categories: web content mining, web structure mining, and web usage mining (Miner et al., 2012). This study focused on web content mining, which analyzes textual data scraped from websites. Recent studies have shown

that web content data reflects public and market perceptions. Veltri (2013) conducted a semantic analysis of Twitter comments to study public understanding of nanotechnology. Park and Geum (2022) investigate the market-side driving forces of technology convergence using technical news articles. Since it is common for companies to use websites to exhibit their products and technical competence to their customers or potential investors, corporate websites have become a good alternative for studying industrial innovation activities (i.e., the commercialization of technology). Libaers et al. (2010) analyzed the keyword occurrence on company websites to frame a taxonomy of business models for small and median-sized enterprises (SMEs). Youtie et al. (2012) presented a keyword-based analysis of SMEs' web content to understand their commercialization strategy for emerging nanotechnologies. Arora et al. (2013) used a similar keyword search technique to examine the commercialization process of emerging graphene technologies. It should be noted that web content mining studies rely heavily on keyword-based approaches. This is because company websites usually embody miscellaneous information, ranging from primary business information to general announcements. Although keyword-based methods can circumvent noisy information embedded in a web corpus, they may also induce potential information loss. To address this issue, we demonstrate how the attention mechanism can be applied to extract the desired content.

2.3. Attention mechanism

The attention mechanism has been applied to various CV and NLP tasks to improve the model performance (Niu et al., 2021). It was designed to imitate complex human biological attention mechanisms to identify distinctive features of the input data. Human attentional mechanisms can be divided into two categories: stimulus-driven and goal-directed attention (Corbetta et al., 2002). Stimulus-driven attention refers to people who are more likely to be attracted to louder voices that stand out from the background. By contrast, goal-directed attention enables humans to select or extract relevant objects based on a predetermined target. Applications of the attention mechanism in deep learning belong to the latter category and are more focused on certain parts of the data according to specific training tasks. Tang et al. (2019) demonstrated that attention-based deep learning models could extract relevant keywords using clinical progress notes. Ding and Luo (2021) proposed a hybrid attention model to identify keyphrases that represent the main topics of a given document. Studies using attention-based models to extract keywords generally rely on the attention weights computed for the tokens of a given input sequence. The method of computing attention weights can also be extended to different abstract levels, known as a hierarchical attention network (HAN). Yang et al. (2016) proposed a HAN to generate document vectors by applying attention mechanisms at both the word and sentence levels. In this regard, a document is composed of important sentences aggregated by identified keywords. Wu et al. (2019) constructed a three-tier attention network for recommending items in which attention layers were employed at the word, sentence, and review levels. In this manner, the proposed dual-attention model is also a HAN that focuses on the website structure of a company and is trained with a novel objective to extract technical-related keywords from the company. Details of the model structure are described in the next section.

2.4. Technology Industry Concordance Matrix

Between 1972 and 1995, Canadian patent examiners assigned patent filings to industries of origin and used codes. This information examines the linkage of a particular type of technology with product/service information as the input and/or output. Accordingly, the Yale Technology Concordance (YTC) links eight IPC sections to 25 industries (Everson and Putnam 1988). Kortum and Putman (1997) used the YTC to predict patent counts by industry from 1983 to 1993. In 2002, Johnson provided an additional concordance between the IPC and ISIC codes, referred to as the OECD Technology Concordance (OTC). Johnson (2002) used the industry of origin (IOO) and industry of use (IOU) codes as the basis for the YTC and translates them into the Canadian SIC system. To make the results compatible with international data, the Canadian SIC system was translated into ISIC codes in the second step.

Another approach proposed by Schmoch et al. (2003) uses the industrial classification of patent applicants. Specifically, these authors assigned IPC codes to 44 industrial fields based on NACE industries and the patent activities of approximately 3,000 firms active in these industry sectors. This methodology for the technology industry concordance table has been applied by several scholars in subsequent studies such as Dorner and Harhoff (2017) in Germany and Ikeuchi et al. (2017) in Japan. Neuhäusler et al. (2019) compared existing studies to evaluate performance differences between datasets and methodologies.

3. Data and methodology

3.1. Data

This study uses company websites and patent portfolios to discover technological opportunities. We collected financial statements of 4, 532 listed companies released by the Japanese Financial Service Agency. We then used a publicly available website[1] to search for the company's homepage using its unique corporate identifier (*houjinbango*). From these datasets, we retrieved 3,829 listed companies for which web information was available.

Kinne and Axenbeck (2020) designed a web mining framework to collect web content from a given company. They claimed that a website is a company's presence on the Internet and consists of a collection of webpages. The highest level of a company's website is on its homepage. In this study, a company's web content includes (1) content presented on its homepage (e.g., www.company-name.com) and (2) content presented on subpages that are directly connected to its homepage. Note that crawled subpages would have the same domain as the homepage (e.g., www.company-name.com/products), and other subpages, such as www.twitter.com/company-name, would be excluded. As a result, the collected webpages of a company are referred to as its web portfolio. In addition to web information, we gathered companies' patent portfolios from the database maintained by the Ministry of Economy, Trade, and Industry (METI)[2]. However, because the METI

[1] https://houjin.jp
[2] https://info.gbiz.go.jp/hojin/DownloadTop#

only discloses the patent application numbers for each company, we further linked it with the Japan Patent Office (JPO) data[3] to obtain textual information such as patent titles and abstracts. In summary, there were 1,789 companies with at least one patent and 2,040 companies without.

3.2. The dual-attention model

Wu et al. (2019) demonstrated that the attention mechanism can be customized in terms of different data structures to identify the desired features of the input data. The proposed dual-attention model is trained with the binary objective of high-tech versus non-high-tech companies. In this study, high-tech companies are defined as those with at least one patent. The observations show that the terminologies presented on the websites of high-tech companies are distributed heterogeneously among their non-high-tech counterparts (i.e., food companies and wholesalers). Based on this, the model is expected to identify technically related keywords on high-tech company websites. Figure 1 depicts the structure of the model, in which two attention layers are applied at the page and word levels. At the bottom, the word-level attention layer assigns weights to each webpage token. Specifically, the ith webpage of a company is calculated using the weighted average of the word vectors:

$$p_i = \sum_{j}^{n} \alpha_{ij}^{w} w_{ij}$$

where w_{ij} is the word embedding for the jth word in the ith webpage, and α_{ij}^{w} is the attention weight assigned to w_{ij}. The attention weight for each token is given by

$$\alpha_{ij}^{w} = softmax_j(\boldsymbol{a}_i^{w}) = \frac{\exp(a_{ij}^{w})}{\sum_j \exp(a_{ij}^{w})}$$

$$a_{ij}^{w} = \frac{w_{ij}^{T} V}{\lambda_w}$$

where a_{ij}^{w} is the corresponding attention score for w_{ij}, V is a weight vector that needs to be trained, and hyperparameter λ_w is a scaling factor. Note that attention weights were generated by converting attention scores to a scale of 0–1 using the softmax function. Subsequently, a company vector is derived similarly by aggregating the webpage vectors. Formally, the company vector is given by

$$c = \sum_{i}^{m} \alpha_i^{p} p_i$$

where α_i^{p} is the attention weights for the ith webpage. The attention weight for each webpage is defined as follows:

[3] https://www.gazette.jpo.go.jp/scciidl010

$$\alpha_i^p = sparsemax_i(\boldsymbol{a}^p)$$
$$a_i^p = \frac{p_i^T W}{\lambda_p}$$

where a_i^p is the corresponding attention score for p_i, W is the trainable weight vector, and λ_p is a scaling factor that must be determined in advance. Instead of using a softmax function to normalize the attention scores for page-level attention, we chose a sparsemax function. In contrast to the conventional softmax function, the sparsemax function converts real values into sparse probabilities. Because a company's website usually covers miscellaneous webpages, using sparsemax enables the algorithm to be more concentrated on crucial webpages, which also helps save one parameter for the following keyword extraction process. One may also notice that the above setting implicitly requires that a company can maximally have m webpages, and value n limits the maximal number of words on each webpage. Hence, padding and cutting strategies were required before training the model.

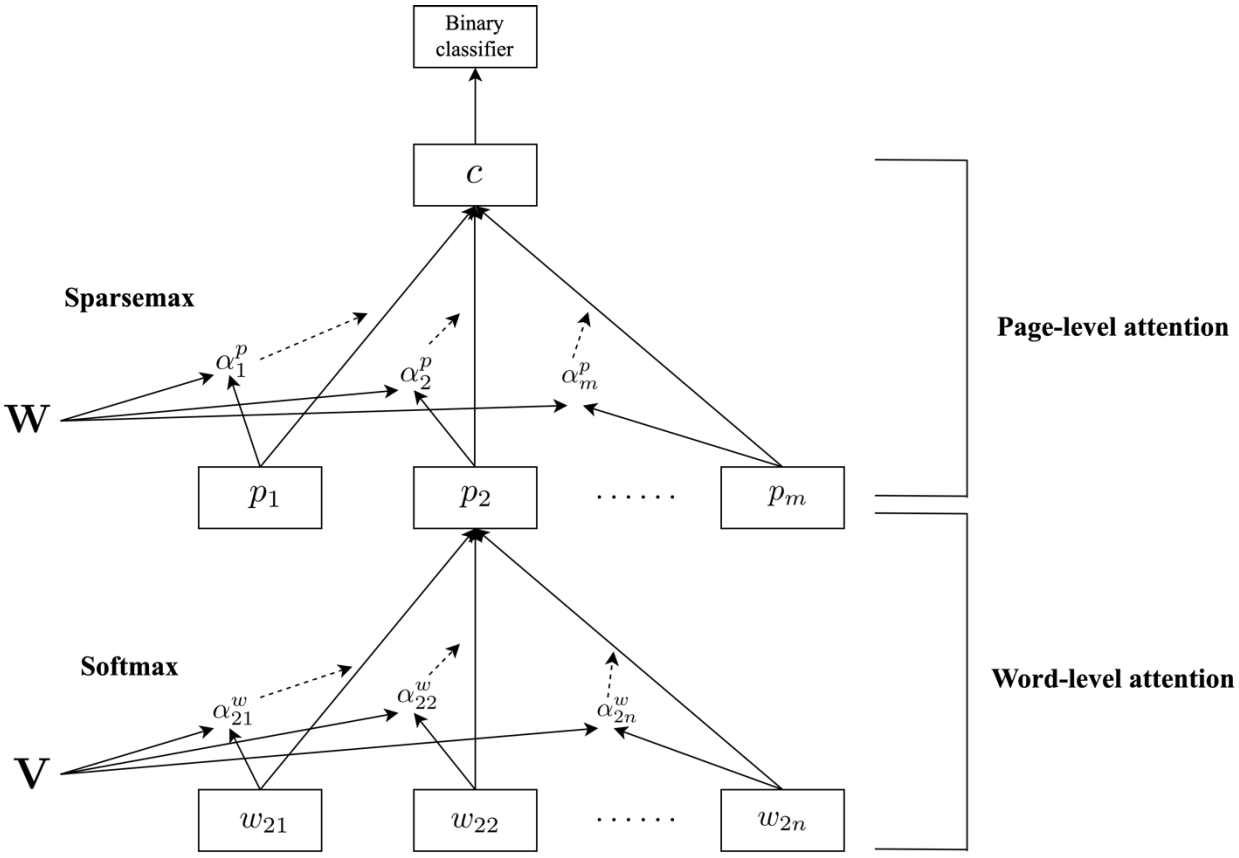

Figure 1. The structure of the dual-attention model

3.3. Keywords extraction

In this section, we extract keywords based on the attention weights obtained from the page and word attention layers. By using the sparsemax function, the page selection process keeps webpages with nonzero attention weights. The keyword-extraction process was implemented for the selected webpages. The main reason for using softmax for word-level attention is that it provides a relatively smoother probability distribution. In contrast, sparsemax eliminates all elements with small weights. Based on this, we introduce a threshold value that can be determined by the users. The threshold θ is determined by

$$\theta = Percentile(\max_{j}(\alpha_{ij}^{w}) - \alpha_{ij}^{w}, k)$$

where k is the percentile. The words with the attention weights less than θ would be extracted. The inclusion of a user-defined parameter k adds flexibility to the proposed method (Eilers et al., 2019; Tang et al., 2019). In this case, fewer keywords would be identified if they were set too high; if they were set too low, more information would be retained.

3.4. Evaluation of extracted web information

By using the proposed method, a company's web portfolio would consist of extracted keywords instead of raw or mixed web content. In this section, we demonstrate that the extracted web information improves the identification of high-tech companies through classification tasks. The setup is as follows: First, the inputs for the classification models were extracted from web information and raw web data. We used two latent Dirichlet allocation (LDA) topic models to separately convert these two data sources into a vector format. The LDA model discovers hidden topics based on the underlying word distribution embedded in the given document (Blei et al., 2003). A company vector is equivalent to a topic distribution derived from the LDA model. Subsequently, classification models such as logistic regression (LR), support vector machine (SVM), and random forest (RF) are used to compare the discernibility of high-tech companies using different approaches.

4. Results

4.1. Training process and web content extraction

In this study, attention weights derived from the dual-attention model were used to extract technical-related web contents. First, because company webpages may be written in multiple languages, we applied a pre-trained language identification model[4] that returns a discrete probability distribution of languages used in the given document. We retained webpages written mainly in Japanese. The open-source package MeCab[5] was used to clean the collected web documents, including removing stop words and punctuation. Subsequently, we apply cutting and padding strategies to calibrate the formats of the

[4] https://fasttext.cc/docs/en/language-identification.html
[5] https://pypi.org/project/mecab-python3/

companies' webpages. The size of the company's web portfolio was set to 32, and each webpage contained 768 tokens. The selection of the hyperparameters and source codes can be found in GitHub[6]. Finally, keywords were extracted based on the aforementioned procedures by setting the threshold value of k to 30. Appendices A and B provide examples of the page attention weights and keywords, respectively.

4.2. Evaluation of extracted web content

In this subsection, we evaluate the Web keywords extracted through the classification task of discerning high-tech companies. We constructed two LDA topic models to generate company vector representations based on extracted web keywords and raw web content. In this case, the number of topics was set to 16. Indeed, our additional experiments showed that the following results were stable in terms of different choices for the number of topics. Three commonly used classification models–(LRs, SVMs, and RFs)–were employed with default settings, and the labeled datasets were split into training and test sets at proportions of 0.7 and 0.3. The results were compared based on four metrics: precision, recall rate, F1 score, and accuracy. Table 1 lists the training and test results for the extracted web keywords and raw web content. The results of the dual-attention model consistently outperformed those of the raw data by a portion around ten percent. The RF model achieved the best performance with a 0.91 training accuracy of 0.85, and a test accuracy.

Metrics	Train		Metrics	Test	
	Dual-attn	Raw		Dual-attn	Raw
(1) LR			(1) LR		
Precision	0.83	0.76	Precision	0.84	0.75
Recall	0.83	0.75	Recall	0.84	0.75
F1 score	0.83	0.75	F1 score	0.84	0.75
Accuracy	0.83	0.75	Accuracy	0.84	0.75
(2) SVM			(2) SVM		
Precision	0.85	0.77	Precision	0.85	0.77
Recall	0.85	0.77	Recall	0.85	0.76
F1 score	0.85	0.76	F1 score	0.85	0.76
Accuracy	0.85	0.77	Accuracy	0.85	0.76
(3) RF			(3) RF		
Precision	0.91	0.81	Precision	0.85	0.77
Recall	0.91	0.81	Recall	0.85	0.77
F1 score	0.91	0.81	F1 score	0.85	0.77
Accuracy	0.91	0.81	Accuracy	0.85	0.77

Table 1. Comparison of the classification results

[6] https://github.com/zhujohn9604/The-dual-attention-model

4.3. Opportunity discovery exercise

In this section, we analyze potential technology opportunities using the extracted web information associated with patent data. First, we introduce two concordance matrices: technology-market matrix A and market-technology matrix B such that

$$m = At$$
$$t = Bm$$

where m and t represent the company's market and technology vectors, respectively. In this study, the market and technology vectors of a given company are generated by vectorizing its web portfolio (extracted web keywords) and patent portfolio using a pre-trained Word2Vec model [7]. The Word2Vec model can generate distributed word representations in which similar words are geometrically close to each other (Mikolov et al., 2013). The technology vector of a company is calculated by aggregating all its patent vectors, which are generated by aggregating the corresponding word vectors and the market vector. Then, the two aforementioned concordance matrices can be derived as

$$\min_A \sum_{i \in S} (m_i - At_i)^2$$
$$\min_B \sum_{i \in S} (t_i - Bm_i)^2$$

where m_i and t_i are the true market and technology vectors of company i, At_i and Bm_i are the estimated market and technology vectors, respectively, and S represents a set of companies with at least one patent. A and B are concordance matrices representing the conversion of information from technology to product/service or vice versa. The difference from existing concordance matrices (Johnson, 2002; Schmoch 2003) is that our matrix is based on textual information for both sides, so that the granularity of technology/industry classification can be flexibly determined by the dimension of document embedding.

In our technology opportunity discovery exercise, we constructed A and B matrices for new firms in a stock market and then apply these matrices to estimate the technology/market vectors of established firms ($\widehat{t_{est}}$, $\widehat{m_{est}}$) based on their actual market/technology vectors (m_{est}, t_{est}). We use the logic of the industry life cycle theory, in which new firms play a dominant role in the emerging embryotic stage of a new industry, whereas large established firms become the main players in the process of industry maturity (Klepper, 1997). In this regard, matrix A reflects how new firms convert their technology into a product/service, and matrix B does so in the opposite direction. Therefore, the estimated vector ($\widehat{t_{est}}$, $\widehat{m_{est}}$) represents the technology (market) pursued by new firms based on existing firms' market (technology). Finally, technology/market opportunities for existing firms can be captured as the difference between the estimated ($\widehat{t_{est}}$, $\widehat{m_{est}}$) and actual vectors (t_{est}, m_{est}). (See Figure 2).

[7] https://fasttext.cc/docs/en/crawl-vectors.html

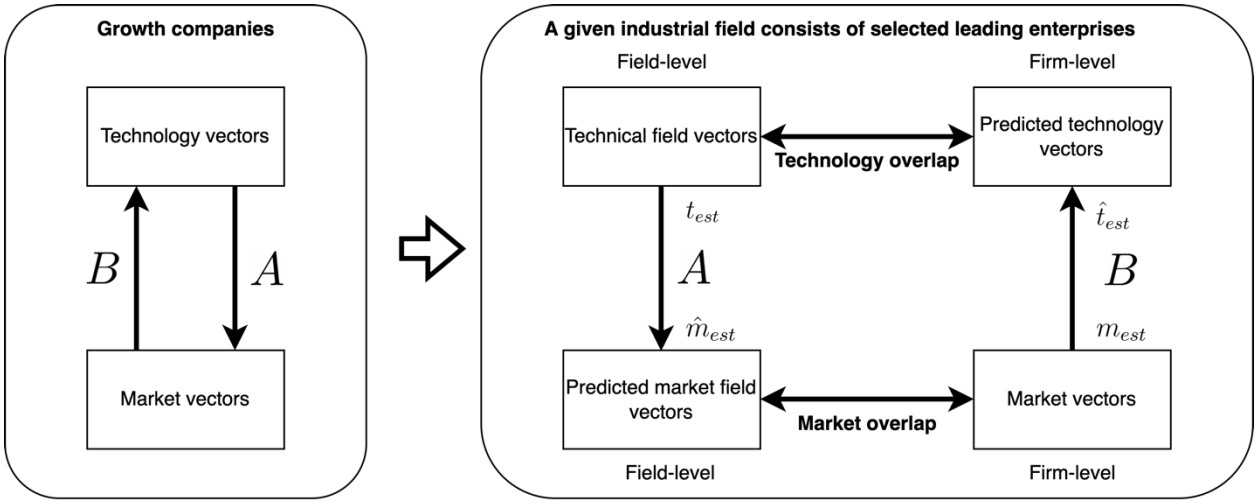

Figure 2. A framework for discovering opportunities exercise

In the actual implementation of this methodology, we use the information of a group of 94 high-growth companies to estimate matrices A and B. On April 4, 2022, Japan Stock Exchange Group (JPE) has restructured its existing market divisions into three new market segments: Prime Market, Standard Market, and Growth Market. The growth Market comprises companies with high growth potential and relatively high investment risk. Based on this, we assume that growth companies explore the emerging technological and business opportunities captured by the two concordance matrices estimated using their technology and market vectors.

Subsequently, the two mapping matrices are applied to analyze the potential opportunities for big enterprises in the prime market by JPE in three industrial fields: automobile, electronics, and pharmaceutics, as established firms. For each industrial field, several leading enterprises were manually selected based on the Toyo Keizai industry classification. We selected 58 automobile, 15 electronic, and 15 pharmaceutical products. Finally, we constructed opportunity measures on both the technological and market sides. Here, we use 1 minus the cosine similarity of $\widehat{t_{est}}$ and t_{est} and $\widehat{m_{est}}$ and m_{est}, respectively, as the distance between the two vectors. Figures 3-1, 3-2 and 3-3 show the results for these three industries. In these figures, we present the opportunity measures by the technology classification of patents (applied by established firms in each industry) based on World Intellectual Property Organization (WIPO) 35 technology classes (Schomoch, 2008).

In the automotive industry, “computer technology” and “control” are identified as market opportunities, while “surface technology, coating” is identified as technology opportunities. Market opportunities can be implemented as new services for autonomous driving. On the technology opportunity side, new technologies can be identified by new firms in the field of surface coating. In the electronics industry, “IT management methods are identified as market opportunities that might be induced by the upward trend of machine learning/AI applications. In contrast, “semiconductors” is identified as a technology opportunity where new technologies are evolving, and R&D managers in big electronics firms should pay particular attention. In the pharmaceutical industry,

technology and market opportunity measures are almost positively correlated, and it is relatively difficult to identify some technology/market opportunities compared to the other two industries. But some products/services using the “IT method” and “Computer Technology” can be identified for large firms to investigate as new business development.

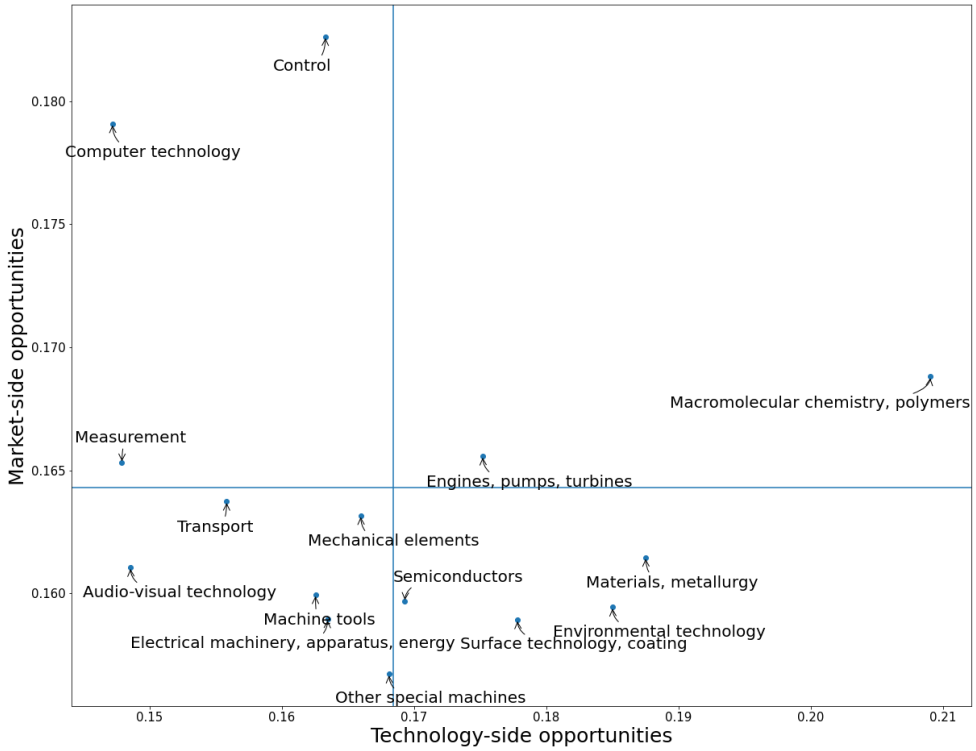

Figure 3-1. Automobile Industry

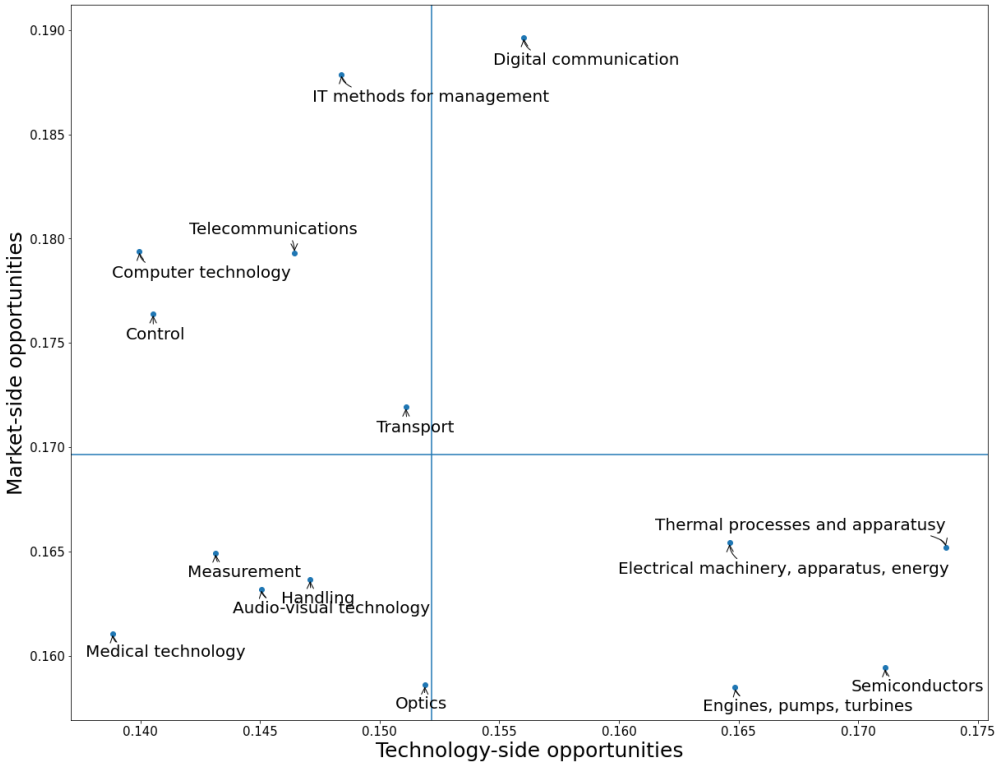

Figure 3-2. Electronics Industry

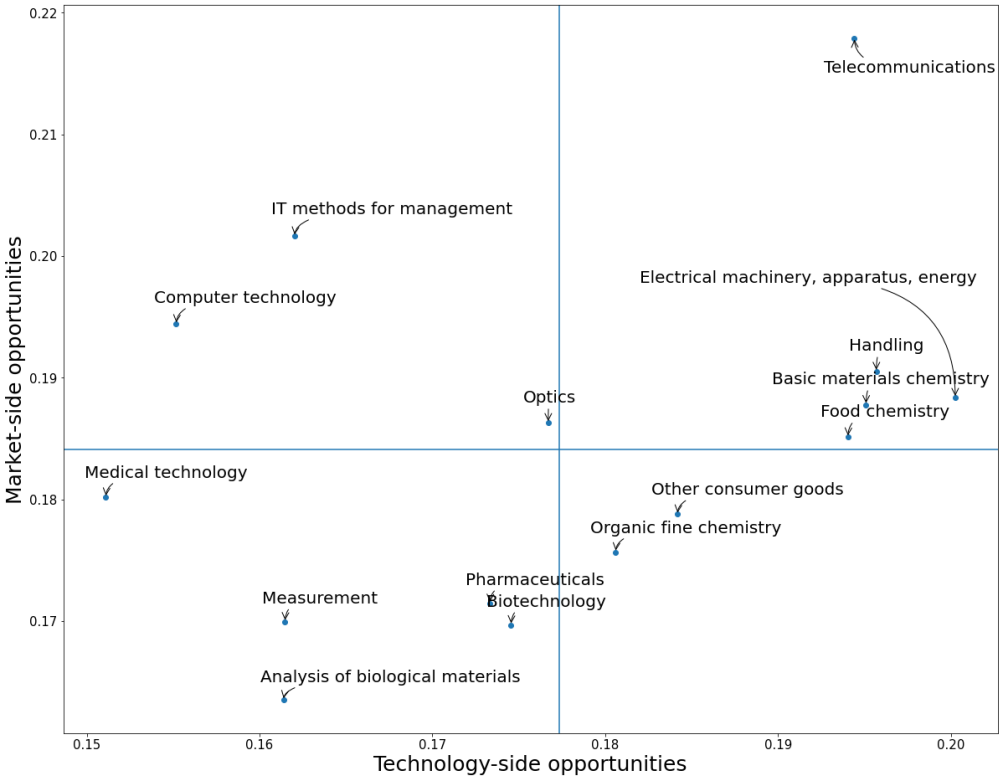

Figure 3-3. Pharmaceutical industry

5. Conclusion

This study presents a new approach for identifying technology and market opportunities using companies' patents and web information. Existing studies have demonstrated that company websites are valuable sources for monitoring commercial activities and for allowing researchers to explore market-side opportunities. However, because companies often disclose miscellaneous information on their websites, the noisy information embedded in the raw web data may obscure the results of the TOD analysis. To address this issue, we proposed a dual-attention model that automatically selects technical webpages and keywords from raw web portfolios. We then validated the quality of the extracted content using a classification task, and the results suggested that the extracted web information outperformed raw web information by a large margin. Subsequently, the extracted web keywords were combined with patent data to form a technology-market concordance matrix. We applied this matrix to identify new technologies and market opportunities for large companies in the automotive, electronics, and pharmaceutical industries. The information derived from the exercise itself is important for technology/market strategies of high-tech firms in these industries. However, our methodology can be used for more narrowly defined specific industries or even for some individual companies to construct their technology management strategies. During this process, the input to the dual-attention model must be modified. This study uses information on all listed companies in Japan, with label data with/without patents in general. However, it would be better to use the webpage information of firms in a particular industry and/or use specific types of patents (by technology class) as labels, depending on the user's interests. Additionally, the technology market index can be manipulated. In this study, we use growth market-listed firm information as a basis, but it could be replaced by other types of new firms, such as high-tech startups in Silicon Valley. Alternatively, it may be useful to use competitor firms' technology market concordance to conduct technology competitive analysis. The managerial contribution of this study is the proposal of a methodology for a technology management analysis with various applications.

Despite its contributions, this study had some limitations. First, this study used company websites as the main proxy for monitoring market-side opportunities; however, a time-trend analysis was missing because all websites were collected at the current time. It may be necessary to collect companies' historical webpages to pursue a more dynamic opportunity analysis, which is challenging. In this regard, we suggest that future work consider the Wayback Machine, a digital archive that stores snapshots of historical webpages worldwide. Second, the proposed technology-market concordance matrix is estimated based on a list of growth companies with patents and websites. Indeed, many technology-driven start-ups do not have patents. In this case, the proposed indicators may not be able to embed the emerging technology-market linkages held by these companies.

Appendixes

Appendix A. The webpage-level attention weights of a company's web portfolio

The Web portfolio of a randomly selected company	Attn-weights
http://www.hakuto.co.jp/irinfo/	0.0
http://www.hakuto.co.jp/irinfo/announce/	0.0
http://www.g5-hakuto.jp/	0.0
http://www.hakuto.co.jp/profile/ethic/	0.0
http://www.hakuto.co.jp/news/2022/20221024.html	0.0
http://www.hakuto.co.jp/news/2022/20221111.html	0.0
http://www.hakuto.co.jp/products/equipment/	0.41
http://www.hakuto.co.jp/profile/outline/strategic/index.html	0.0
http://www.hakuto.co.jp/profile/	0.0
http://www.hakuto.co.jp/products/components/	0.1
http://www.hakuto.co.jp/profile/outline/embedded/index.html	0.0
http://www.hakuto.co.jp/news/2022/20221102.html	0.0
http://www.hakuto.co.jp/news/2022/20221013.html	0.0
http://www.hakuto.co.jp/eco/	0.0
http://www.hakuto.co.jp/products/	0.0
http://www.hakuto.co.jp/site_map/	0.0
http://www.hakuto.co.jp/news/2022/20221020.html	0.0
http://www.hakuto.co.jp/contact/	0.03
http://www.hakuto.co.jp/profile/outline/advanced/index.html	0.0
https://www.process.hakuto.co.jp/	0.18
http://www.hakuto.co.jp/	0.0
http://www.hakuto.co.jp/privacy/	0.0
http://www.hakuto.co.jp/products/chemicals/	0.27
http://www.hakuto.co.jp/products/devices/	0.0
http://www.hakuto.co.jp/sitepolicy/	0.0
http://www.hakuto.co.jp/news/	0.0
http://www.hakuto-vacuum.jp/	0.0
http://www.hakuto.co.jp/news/	0.0
http://www.hakuto.co.jp/	0.0
http://www.hakuto.co.jp/news/2022/20221115.html	0.0
http://www.hakuto.co.jp/irinfo/	0.0

Appendix B. Examples of extracted keywords

URL	Extracted keywords (translated words)
http://www.hakuto.co.jp/products/equipment/	太陽 (solar) 機器 (machine) デバイス (device) テクノロジー (technology) イオン (Ion) 電気 (electronic) スタンダード (standard) 薄膜 (film) コンポーネント (component) 光通信 (Optical communication) ケミカル (chemical) 材料 (material) 製品 (product) 装置 (device) 成形 (mold) 探す (search) 試験 (experiment) メーカー (manufacturer)
http://www.hakuto.co.jp/products/chemicals/	電気 (electronic) ケミカル (chemical) 粘着 (adhesion) 製品 (product) ブランド (brand) 機器 (machine) デバイス (device) ボイラ (boiler) 精製 (purification) フェルト (felt conditioner) 薬品 (drugs) ODM コンポーネント (component)
https://www.process.hakuto.co.jp/	研究 (research) PVD 機器 (machine) 問合せ (inquiry) Compact CL モジュール (module) 起こす (induce) MS situ 生産 (production) エッチング (etching) ガス (gas) FA 量産 (mass production)制御 (control) 材料 (material) イオンビームミリング (Ion beam milling) 製品 (product) シリーズ 装置 (device) 半導体 (semiconductor) セル 定性 (qualitative) 用途 (use) 完結 (completion) 有機 (organic) MBE CVD リフロー (reflow) LTI PVA ブレイクスルー (breakthrough)

References

Audretsch, D., Vivarelli, M., 1996. Firm size and R&D spillovers: evidence from Italy. Small Business Economics 8, 249–258.

Arora, S., Youtie, J., Shapira, P., Gao, L., Ma, T., 2013. Entry strategies in an emerging technology: A pilot web-based study on graphene firms. Scientometrics, 95(3), 1189–1207.

Bengisu, M., Nekhili, R., 2006. Forecasting emerging technologies with the aid of science and technology databases. Technol. Forecast. Soc. Change 73 (7), 835–844.

Blei, D., A. Ng, A., Jordan, M., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.

Cantwell, J., Piscitello, L., 2000. Accumulating technological competence: its changing impact on corporate diversification and internationalization. Ind. Corp. Chang. 9, 21–51.

Corbetta, M., Shulman, G.L., 2002. Control of goal-directed and stimulus-driven attention in the brain, Nat. Rev. Neurosci. 3, 201–215.

Dosi, G., 1988. Sources, Procedures, and Microeconomic Effects of Innovation. Journal of Economic Literature 26, 1120–1171.

Dorner, M., Harhoff, D., 2017. A Novel Technology-Industry Concordance Table Based on Linked Inventor-Establishment Data. SSRN Electronic Journal.

Ding, H., Luo, X., 2021. Attentionrank: Unsupervised keyphrase extraction using self and cross attentions. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1919–1928.

Everson, R., Putnam, J., 1988. The Yale-Canada patent flow concordance. Yale University Economic Growth Centre Working Paper.

Eaton, J., Kortum, S., 2002. Technology, Geography, and Trade. Econometrica, 70 (5), 1741-1779.

Eilers, K., Frischkorn, J., Eppinger, E., Waltera, L., Moehrlea, M.G. (2019). Patent based semantic measurement of one-way and two-way technology convergence: The case of ultraviolet light emitting diodes (UV-LEDs). Technol. Forecast. Soc. 140, 341-353.

Gök, A., Waterworth, A., Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics 102 (1), 653–671.

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.

Johnson, D., 2002. The OECD Technology Concordance (OTC). Patents by Industry of Manufacture and Sector of Use. OECD STI Working Papers, 2002/5, Paris

Kortum, S. and Putnam, J., 1997. Assigning Patents to Industries: Tests of the Yale Technology Concordance. Economic Systems Research, 9 (2), 161-176.

Klepper, S., 1997. Industry life cycles, Industrial and Corporate Change, 6 (1), 145-181.

Kwon, H., Kim, J., Park, Y., 2017. Applying LSA text mining technique in envisioning social impacts of emerging technologies: the case of drone technology. Technovation 60, 15–28.

Kim, H., Hong, S., Kwon, O., Lee, C., 2017. Concentric diversification based on technological capabilities: link analysis of products and technologies. Technol. Forecast. Soc. 118, 246–257.

Kwon, H., Park, Y., Geum, Y., 2018. Toward data-driven idea generation: application of Wikipedia to morphological analysis. Technol. Forecast. Soc. 132, 56–80.
Klevorick, A.K., Levin, R.C., Nelson, R.R., Winter, S.G., 1995. On the sources and significance of interindustry differences in technological opportunities. Res. Policy 24, 185–205.
Kinne, J., Axenbeck, J., 2020. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125 (3), 2011–2041.
Lee, C., Jeon, D., Ahn, J.M., Kwon, O., 2020. Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent product database. Technovation 96-97.
Lee, M., Kim, S., Kim, H., Lee, J., 2022. Technology opportunity discovery using deep learning-based text mining and a knowledge graph. Technol. Forecast. Soc. Chang. 180, 121718.
Lee, Sungjoo, Lee, Seonghoon, Seol, H., Park, Y., 2008. Using patent information for designing new product and technology: keyword based technology roadmapping. R&D Manag. 38, 169–188.Libaers, D., Hicks, D., Porter, A. L., 2010. A taxonomy of small firm technology commercialization. Industrial and Corporate Change.
Miner, G., Elder, J., I. V., Hill, T., Nisbet, R., Delen, D., Fast, A., 2012. Practical text mining and statistical analysis for non-structured text data applications. New York: Academic Press.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint.
Neuhäusler, P., Frietsch, R., Kroll, H., 2019. Probabilistic concordance schemes for the re-assignment of patents to economic sectors and scientific publications to technology fields. Discussion Papers Innovation Systems and Policy Analysis Nr. 60. Karlsruhe: Fraunhofer ISI.
Niu, Z., Zhong, G., Yu, H., 2021. A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62.
Olsson, O., 2005. Technological opportunity and growth. J. Econ. Growth 10, 35–57.
Porter, A.L., Detampel, M.J., 1995. Technology opportunities analysis. Technol. Forecast. Soc. 49 (3), 237–255.
Park, M., Geum, Y., 2022. Two-stage technology opportunity discovery for firm-level decision making: GCN-based link-prediction approach. Technol. Forecast. Soc. Chang. 183, 121934.
Sun, X., Lu, W., 2020. Understanding attention for text classification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3418–3428.
Schmoch, U., 2008. Concept of a Technology Classification for Country Comparisons – Final Report to the World Intellectual Property Organization (WIPO). URL: http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf, accessed: 2016-02-04.
Schmoch, U., Laville, F., Patent, P., Frietsch, R., 2003. Linking Technology Areas to Industrial Sectors – Final Report to the European Commission, November 2003.
Tang, M., Gandhi, P., Kabir, M., Zou, C., Blakey, J., Luo, X., 2019. Progress notes classification and keyword extraction using attention-based deep learning models with BERT. arXiv preprint arXiv:1910.05786.

Veltri, G., 2013. Microblogging and nanotweets: Nanotechnology on twitter. Public Understanding of Science, 22(7), 832–849.
Webb, E., Campbell, D., Schwartz, R., 1966. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally.
Wang, J., Chen, Y.J., 2019. A novelty detection patent mining approach for analyzing technological opportunities. Adv. Eng. Inform.
Wu, C., Wu, F., Liu, J., Huang, Y., 2019. Hierarchical user and item representation with three-tier attention for recommendation, in: NAACL-HLT (1), Association for Computational Linguistics, 2019, pp. 1818–1826.
Yoon, J., Park, H., Seo, W., Lee, J.-M., Coh, B.-Y., Kim, J., 2015. Technology opportunity discovery (TOD) from existing technologies and products: a function-based TOD framework. Technol. Forecast. Soc. Chang. 100, 153–167.
Yoon, B., Park, I., Coh, B.-y., 2014. Exploring technological opportunities by linking technology and products: application of morphology analysis and text mining. Technol. Forecast. Soc. Chang. 86, 287–303.
Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E., 2016. Hierarchical attention networks for document classification. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pp. 1480–1489.
Youtie, J., Hicks, D., Shapira, P., Horsley, T., 2012. Pathways from discovery to commercialization: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. Technol. Anal. Strateg. Manag. 24 (10), 981-995.
Zook, C., Allen, J., 2001. Profit From the Core: Growth Strategy in an Era of Turbulence. Harvard Business School Publishing, Boston, MA.