The Sixth International Conference on Design Creativity (ICDC2020)
Oulu, Finland, August 26th – 28th 2020
https://doi.org/10.35199/ICDC.2020.26

Similarity Computation Supporting Creative Activities

Xin Ni[1], Ahmed Samet[2] and Denis Cavallucci[1]

[1]ICUBE/CSIP, INSA de Strasbourg, Strasbourg, France
[2]ICUBE/SDC, INSA de Strasbourg, Illkirch, France

Abstract: Creativity sessions in industry, when they are based solely on people's knowledge, produce less and less value. This is mainly due to the need to further expand the spectrum of knowledge needed to solve problems. We are therefore increasingly witnessing the limits of our knowledge capabilities to meet the demands of today's inventive problem-solving in industry. The research presented in this paper proposes a method of semantic association between problems extracted from an unstructured textual corpus of a patent using Google's Word2vec algorithm followed by cosine similarity to create original pairings between problems from different but semantically close domains. We postulate that such a method is a preamble to the automation of TRIZ and thus avoids the difficulties of not having been updated for a few decades.

Keywords: *TRIZ, Inventive Design Method, Artificial Intelligence, Machine Learning*

1. Introduction

Inventive Design Method (IDM) is from TRIZ theory (Altschuller, 1998) and was created to assist engineers in their invention process (Souili et al., 2015). Besides, the formal knowledge description components using ontologies, such as problems, partial solutions, and parameters (Zanni-Merk et al., 2010) IDM approach contains is mainly hidden in the patent documents. In patents, problems describe unsatisfactory features of existing methods or situations. Partial solutions provide improvements or changes to the defined problems. Each problem may cause one or more contradictions to the patent solves (Ni et al., 2019). This IDM-related knowledge is an important referring resource for engineers to perform R&D activity. Engineers, however, still mainly search existing solutions from different patent documents by manual work. It cannot fit the current rise of infinite and permanent renewal of information and data throughout all domains (Ni et al., 2019). Especially for latent solutions from different domains, it is an obstacle for engineers who do not have abroad experience in the specific domain. Nevertheless, this type of solution might be a creative solution for the given problem. In this work, we assume that, if two problems are similar enough, the corresponding solution from the different domain's problem could be an inventive solution for the given problem. Thus, how to efficiently find out similar problems that are hidden in different domains' patent documents become an important task and it has practical value for the R&D activity.

In order to address this issue, in this paper, we proposed a similarity computation model, called IDM-Similar. It mixes the sentence vector relying on Word2vec neural networks (Mikolov et al., 2013) and cosine metric. We firstly use Patent Extractor (Souili & Cavallucci, 2017) to extract IDM-related knowledge including problems, partial solutions, and parameters from patent documents. Then

Word2vec neural networks are used to achieve each word's vector for the given problem sentence. After that, we average all word vectors that are used to consist of the target sentence to obtain the sentence vector. Cosine similarity between pairwise problems is finally computed in order to retrieve similar problems from different domains' patents.
The final experimental results on U.S. patent datasets show that our model has a promising perspective on the application of similar problems matching from different domains' patents. It is an effective way for engineers to find out similar problems and corresponding solutions from a large amount of patent documents. These different domains' solutions could be creative solutions for the target problem. It will also greatly speed up the whole process of R&D activity. Particularly, the two case studies from different domains show our model's performance.
The paper consists of the following sections. Section 2 introduces the related works of similarity computation in the patent field. Section 3 details the methodology of the Wrod2vec neural networks, Sentence2vec, and cosine similarity metric. The experimental results and case studies on real patent datasets have been shown in Section 4. We finally conclude our work and show perspectives for future work.

2. Related Work

Similarity computation is an important task of Natural Language Processing (NLP). Besides, patent documents are an important carrier of the latest innovative knowledge. It already became a type of valuable resource for the product innovation. Many research works have been involved in this field.
In the beginning, Bibliographic Coupling and Co-citation Analysis methods are proposed by Kessler & Maxwell (1962) and Small & Henry (1973) separately to analyze the similarity among different patents. (Lai et al., 2005) proposed a patent classification system using co-citation analysis to compute the patent similarity. Further, McGill & Joseph (2007) and Mowery (1998) computed the similarity of the firm patents via cross-citation rate when analysing patent citation data. Moehrle (2005) and Bergmann (2008) also used natural language processing methods to extract a subject–action–object–format (SAO) structure in patents first and then built similarity matrices for patents to evaluate the similarity. In addition, some indexes as centrality index, technology cycle index, and technology keyword clusters in patents are also used for in-depth quantitative analysis in order to compute the patent similarity (Yoon et al., 2004).
The above similarity computation approaches on patent documents mainly are applied on wide fields like evaluating the risk of patent infringement (Bergmann et al., 2008), discovering competitive intelligence (Shih et al., 2010), identifying technology opportunities (Yoon et al., 2005), measuring the novelty of patents (Gerken et al., 2012), making the technological roadmap (Lee et al., 2009), detecting the similarity between patent documents, and scientific publications (Magerman et al., 2009), etc..
These similarity computation methods also inspired us to explore a new usage of similarity computation on patent documents. At the same time, we found that few existing similarity computation methods or models which have been used on IDM-related knowledge, especially computing the similarity among different problems in patents. In this paper, we compute the similarity in the field of IDM-related knowledge implementation. The sentence vector relying on Word2vec neural networks and cosine similarity metric are used to improve the efficiency of extracting similar problems and corresponding solutions from a large amount of patent documents. This work will further expand the border of exploring creative solutions.

3. Methodology

In this section, we introduce our similarity computation model, called IDM-Similar. Our work aims to find out similar problems from a large amount of different domains' patents in order to pick up creative solutions for the target problem. A patent extractor tool (Souili & Cavallucci, 2017) is used to extract problems, corresponding partial solutions, and parameters from patent corpus. Then, we compute the similarity between problems. $K_i = \{P_i, PS_i, PA_i\}$ is from i-th patent document where P_i, PS_i, and PA_i are problems, partial solutions, and parameters respectively in the i-th patent document. Given the j-th

problem $P_i^j = (w_i^{j_1}, w_i^{j_2}, \dots, w_i^{j_{|w_i|}})$ in the i-th patent document where $w_i^{j_{|w_i|}}$ is the $|w_i|$-th word in the j-th problem sentence, and we compute its similarity with other considered problems P.

3.1. Word Vector

Word2vec, a two-layer neural networks, is used to generate each word's vector. It can be trained by a large-scale corpus to achieve the vector in the space for each unique word in the corpus. Word vector is positioned in the vector space such that words sharing common contexts in the corpus are located in close proximity to one another in the space (Mikolov et al., 2013). The trained Word2vec model can simplify the processing of the considered text into n-dimensional space vector operation. Thereby, the similarity in vector space can represent the semantic similarity of text. In our work, we trained it using open-source Wikipedia Corpus.

Besides, the skip-gram structure is chosen in our Word2vec to produce a distributed representation of words due to it works well on the large-sized dataset and infrequent words. Skip-gram can predict each context word via the target work. For instance, as shown in Figure. 1, skip-gram predicts the context of the central word "*plurality*" is "*The signal line includes a*" and "*of openings arranged along the*" when it reaches "_ _ _ _ _ *plurality*_ _ _ _ _signal line."

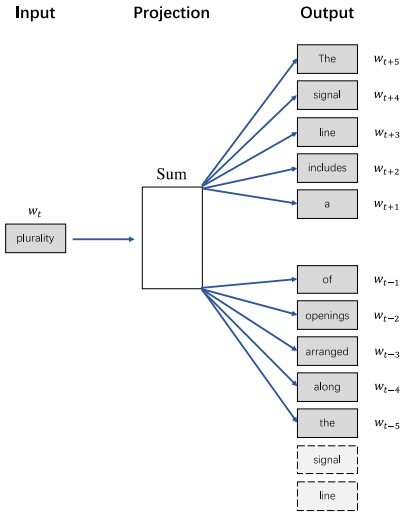

Figure 1. Skip-gram structure

3.2. Sentence Vector

As illustrated in Figure. 2, the sentence vector is combined by corresponding word vectors. We achieve the problem sentence vector $\overrightarrow{\boldsymbol{P}}$ by calculating the average vector of all words in each sentence. The calculation is defined as:

$$\overrightarrow{\boldsymbol{P}} = \frac{\sum_{i=0}^{j} \overrightarrow{\boldsymbol{w_i}}}{j} \tag{1}$$

3.3. Cosine Similarity

We first compute the cosine distance between the given problem's sentence vector $\overrightarrow{\boldsymbol{P_i}}$ and another sentence vector $\overrightarrow{\boldsymbol{P_J}}$:

$$CosineDistance = \frac{\overrightarrow{\boldsymbol{P_i}} \cdot \overrightarrow{\boldsymbol{P_J}}}{|\overrightarrow{\boldsymbol{P_i}}||\overrightarrow{\boldsymbol{P_J}}|} = \frac{\sum_{i,j=1}^{n} \boldsymbol{P}_i \times \boldsymbol{P}_j}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{P}_i)^2} \times \sqrt{\sum_{j=1}^{n} (\boldsymbol{P}_j)^2}} \tag{2}$$

Next, the cosine similarity is defined as:

$$Cosine\ Similarity = 1 - Cosine\ Distance = 1 - \frac{\overrightarrow{\boldsymbol{P_i}} \cdot \overrightarrow{\boldsymbol{P_J}}}{|\overrightarrow{\boldsymbol{P_i}}||\overrightarrow{\boldsymbol{P_J}}|} \tag{3}$$

Overall, if the cosine similarity value is becoming closer to 1, the possibility of similarity between pairs of sentences increases.

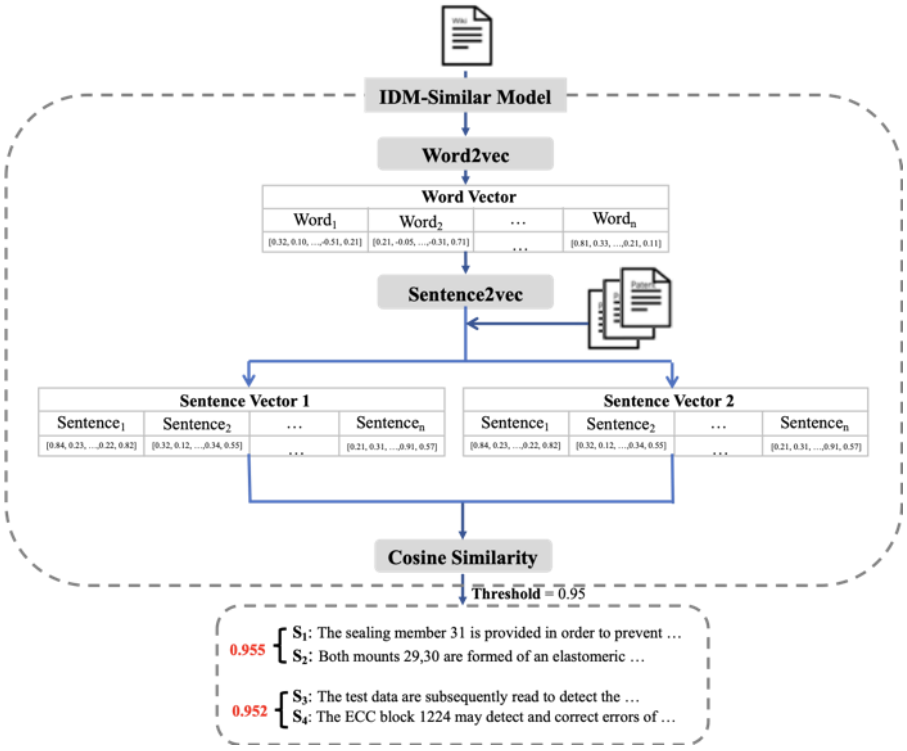

Figure 2. An overview of IDM-Similar model

As illustrated in Figure 2, we first employ the Wikipedia corpus to train Word2vec neural networks in order to achieve each word's vector. After that, we adopt the trained Word2vec neural networks to obtain the sentence vector for each inputting problem sentence. Finally, we apply the cosine similarity metric to achieve the similarity value of each pairwise problems, and then pick up those similar problems whose value are greater than the threshold as well as their corresponding solutions.

4. Experiments

4.1. Datasets and Evaluation Metrics

In the paper, English Wikipedia dataset is used to train the Word2vct neural networks. It only contains regular article text but removes tables and links to foreign language versions. In the end, we evaluate our IDM-Similar model using U.S. Patent Grant dataset (U.S. Patent Dataset, 2017). We leverage utility patent datasets as the testing dataset to evaluate the performance of IDM-Similar model. It contains a total of 6,161 documents.

Finding the "gold-standard" ground truth of evaluating the similarity among different sentences always is an open problem. In this paper, we referred to 3 experts who are respectively from mechanics, chemistry, and architecture to evaluate the experimental results manually. A cross-checking among them is made to ensure the authenticity of final results.

Besides, we set the window size as 5 for training Word2vec neural networks and fix the similarity threshold to 0.95 for performance reasons upon carrying out several tests and optimizing the size of the output.

4.2. Experimental Results

As illustrated in Table 1, Patent Extractor extracted three types of IDM-related knowledge from 6,161 U.S. patent documents. 4,574 problems are used as input dataset to Sentence2vec model. We compute the similarity between any a pair of problem matches via IDM-Similar model. The performance of our

model on U.S. patent dataset is shown in Table 2. From results, IDM-Similar model finally retrieve 1,121 pairs of similar problems when the similarity threshold is set as 0.95. Through three experts' cross-checking, the number of true positive (TP) and false positive (FP) of final results are 1,000 and 121 respectively so that the precision of similarity is 89.21%. It demonstrates that our model can effectively find out similar problems from a large amount of different domains' patent documents.

Table 1. Performance of Patent Extractor on U.S. patent dataset

Model	**Patent Extractor**		
IDM-related Knowledge	Problem	Partial Solution	Parameter
Number	4,574	17,971	29,264

Table 2. Experimental results on U.S. patent dataset

Model	**IDM-Similar**			
Metric	TP	FP	Total	Precision
Number	1,000	121	1,121	89.21%

4.3. Case Study

The objective of this part is to demonstrate the practical value of our model supporting inventive solutions from different domains' patents. Two case studies on chemistry/mechanics and computer/physics domains respectively assess the performance of our model.

1. Chemistry/Mechanics: "Collector for bipolar lithium ion secondary batteries (US9537152B2)" and "Vacuum cleaner with motor between separation stages (US9532691B2)" are two US patents that are from chemistry and mechanics respectively. As shown in Figure. 3, our model finds out a pair of similar problems: "*The sealing member 31 is provided in order to **prevent contact** between the current collectors 11 adjacent to each other inside the battery and prevent a short circuit caused by slight unevenness at edge portions of the single cell layers 19 in the power generation element 21*." and "*Both mounts 29,30 are formed of an elastomeric material and act to **isolate** the second dirt-separation stage 7 and thus the remainder of the dirt separator 3 from the vibration generated by the vacuum motor 6.*"

At the US9537152B2 patent, the inventor proposed a method to provide a resin current collector containing imide group-containing resin for use in a bipolar lithium ion secondary battery, and capable of reducing the absorption of lithium ions inside the current collector. Furthermore, the inventor found out that the permeation and absorption of lithium ions can be significantly reduced by providing, in the resin current collector containing imide group-containing resin, an isolation resin layer containing resin not containing imide group and a metal layer. In the collector, as illustrated in Figure. 3 (left) the sealing member 31 is provided to ensure reliability and safety for a long period of time, which provides the bipolar secondary battery 10 with high quality.

For the normal vacuum cleaner, the suction power generated at the inlet of the vacuum cleaner is significantly less than the suction power generated by the unrestricted vacuum motor due to the pressure drop across each dirt-separation stage as the air is drawn through the vacuum cleaner. In order to solve this problem, as shown in Figure. 4, a new type of vacuum cleaner is invented in the US9532691B2 patent. This vacuum cleaner includes a first dirt-separation stage, a second dirt-separation stage, and a vacuum motor for moving air through the first dirt-separation stage and the second dirt-separation stage. Furthermore, as shown in Figure. 5, in the patent, the axial mount 29 is attached to the front end of the housing 20 and abuts a wall of the second dirt-separation stage 7 so as to form a seal. During use, the axial mount 29 deforms to absorb vibration of the vacuum motor 6 in an axial direction. Besides, the radial mount 30 is attached to the side of the housing 20 and comprises a sleeve 31 that surrounds the housing 20, a lip seal 32 located at one end of the sleeve 31, and a plurality of ribs 33 that extend axially along the sleeve 31. The radial mount 30 abuts a wall of the second dirt-separation stage 7 such that the lip seal 32 forms a seals against the wall.

After analyzing these two patents, we think that the elastomeric material forming a seal by deforming in the US9532691B2 patent might be used in the US9537152B2 patent for preventing the contact and a short circuit between collectors in the battery. Besides, the sealing member in the US9537152B2

patent could possibly be a creative solution for isolating the dirt-separation stage and avoid the vibration of the vacuum motor.

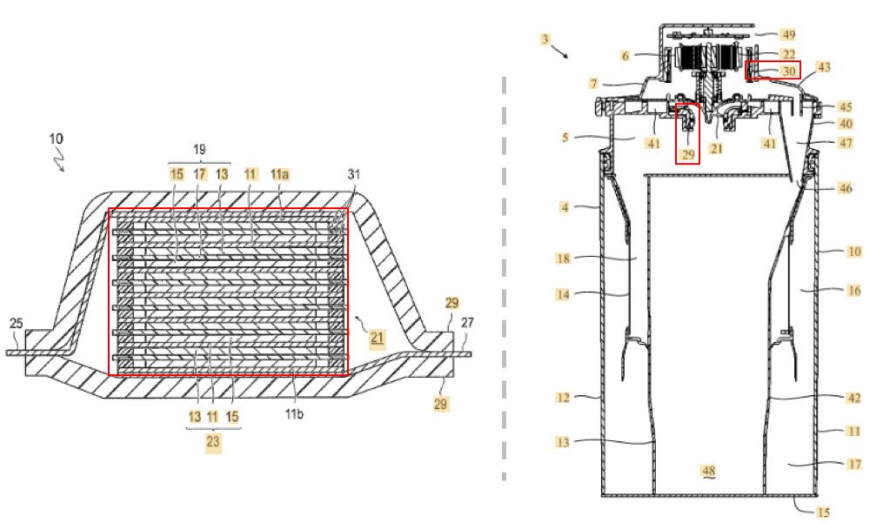

Figure 3. Diagrams of the sealing member (left) and the elastomeric material (right)

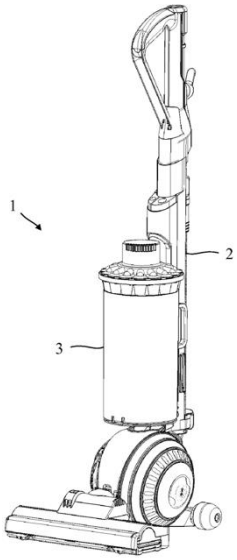

Figure 4. The vacuum cleaner in the US9532691B2 patent

2. Computer/Physics: "Hybrid-HDD with improved data retention (US9536619B2)" and "Semiconductor device and method of fabricating the same (US9536897B2)" are two US patents that are from computer and physics respectively. Two similar problems our model found are illustrated in Figure. 6: "*The test data are subsequently read to detect the possibility of data retention* ***errors*** *that may occur when reading the associated* ***user data***." and "*The ECC block 1224 may* ***detect and correct errors of data*** *which are read out from the memory device 1210.*" We think there is a kind of possibility to add ECC block in US9536897B2 patent into the left device to solve the data retention errors that mentioned in US9536619B2 patent.

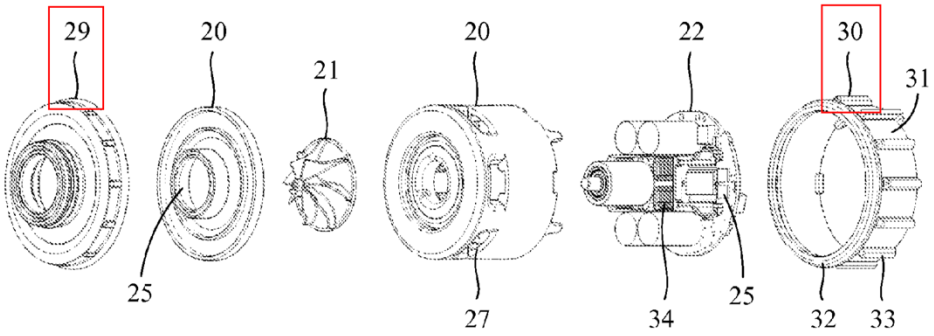

Figure 5. The positions of the axial mount 29 and radial mount 30

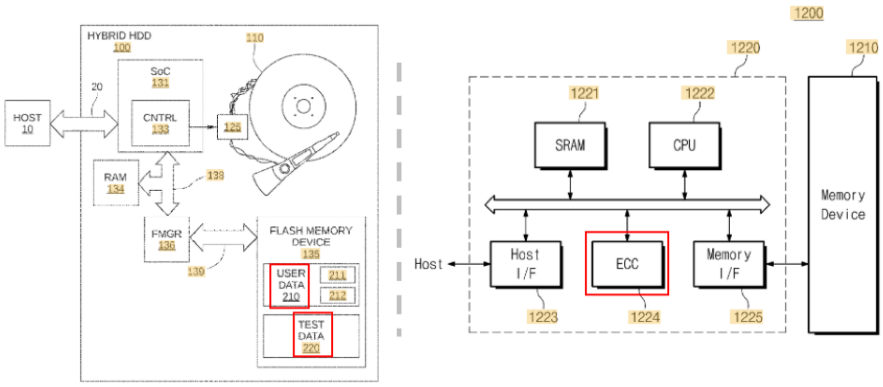

Figure 6. Diagrams of the hybrid HDD (left) and the memory systems (right)

In conclusion of these two cases, we note that the final similar problems from different domains' patents our IDM-Similar model found have a significant practical value for supporting creative solutions.

5. Conclusion and Future Work

The IDM-Similar model is proposed in this paper to effectively extract similar problems from different domains' patent documents. It is helpful for engineers to find out creative solutions from different domains for the target problem. Word2vec neural networks and cosine metric are used in our model. As shown in Section 4, final experimental results show a good performance our model has on real-world U.S. patent datasets. In particular, we demonstrate two cases that the problems could be solved via creative solutions from different domains. This work provides a new way for engineers to effectively access creative solutions from different domain's patents in order to further facilitate R&D activities.
In the future, we will explore to combine other latest neural networks into the IDM-Similar model to further improve the final performance of the model. Besides, how to link problems and inventive solutions from different domains automatically is also an interesting research direction for us.

Acknowledgement

This work is supported by China Scholarship Council (CSC).

References

Altschuller, G. (1998). S.: 40 Principles: TR1Z Keys to Technical Innovation. *Technical InnovationCenter Inc. Worcester MA*.

Souili, A., Cavallucci, D., & Rousselot, F. (2015). A lexico-syntactic pattern matching method to extract IDM-TRIZ knowledge from on-line patent databases. *Procedia engineering*, *131*, 418-425.

Zanni-Merk, C., Cavallucci, D., & Rousselot, F. (2010). Using patents to populate inventive design ontology. In *Proceedings of the TRIZ Future conference* (pp. 52-62).

Ni, X., Samet, A., & Cavallucci, D. (2019, October). An Approach Merging the IDM-Related Knowledge. In *International TRIZ Future Conference* (pp. 147-158). Springer, Cham.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Souili, A., & Cavallucci, D. (2017). Automated Extraction of Knowledge Useful to Populate Inventive Design Ontology from Patents. In *TRIZ–The Theory of Inventive Problem Solving* (pp. 43-62). Springer, Cham.

Kessler, M. M. (1962). *An experimental study of bibliographic coupling between technical papers* (No. 62 673TN1). MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, *24*(4), 265-269.

Lai, K. K., & Wu, S. J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information processing & management*, *41*(2), 313-330.

McGill, J. P. (2007). Technological knowledge and governance in alliances among competitors. *International Journal of Technology Management*, *38*(1), 69.

Mowery, D. C., Oxley, J. E., & Silverman, B. S. (1998). Technological overlap and interfirm cooperation: implications for the resource-based view of the firm. *Research policy*, *27*(5), 507-523.

Moehrle, M. G., Walter, L., Geritz, A., & Müller, S. (2005). Patent-based inventor profiles as a basis for human resource decisions in research and development. *R&D Management*, *35*(5), 513-524.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips. *R&d Management*, *38*(5), 550-562.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37-50.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips. *R&d Management*, *38*(5), 550-562.

Shih, M. J., Liu, D. R., & Hsu, M. L. (2010). Discovering competitive intelligence by mining changes in patent trends. *Expert Systems with Applications*, *37*(4), 2882-2890.

Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change*, *72*(2), 145-160.

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645-670.

Lee, S., Yoon, B., Lee, C., & Park, J. (2009). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change*, *76*(6), 769-786.

Magerman, T., Van Looy, B., & Song, X. (2009). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289-306.

US Patent Dataset, https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/2017/, last accessed 2020/1/10.