Automatic Construction of Technology Function Matrix

Xiang Shi, Zikun Feng, Jiawei Liu, Qikai Cheng* and Wei Lu

Wuhan University, No. 299, Bayi Road, Wuhan, Hubei, 430072, China

Abstract

The Technology Function Matrix (TFM) is a typical patent analysis method that is widely used to detect high-value technology and to locate technical gaps in a specific field. Early TFM construction methods were either based on manual work or machine learning (ML) models. However ML-based models often require large-scale annotated datasets, which are labor-intensive and time-consuming. Therefore, there is a great practical need for low-cost and efficient construction of the TFM. In this paper, we propose a framework for automatically constructing a TFM that requires only a small amount of labeled data. First, we adopt a semi-supervised strategy that comprehensively uses the semantic dependency parser and the pre-trained language model to extract function and technology phrases. Second, a large-scale dictionary of upper and lower categories and synonyms is adopted to merge the related function and technology phrases. Finally, we build an interactive system to visualize the TFM construction process. Compared with traditional methods, our method can significantly improve the performance of technology and function phrase extraction. Furthermore, our system can help experts correct TFM construction results and analyze the current state of technology development in a certain field.

Keywords

technology function matrix, entity recognition, semantic dependency parsing, pre-trianing language model

1. Introduction

A patent is an important carrier of technology information. Through an analysis of the technical means, technical problems, and technical functions in patent documents, we can uncover existing high-value technology and the potential development needs of future technologies and functions. However, with the rapid development of science & technology, technical fields continuously subdivide and cross, and a large number of patent documents have accumulated, which significantly increases the difficulty of patent analysis. This is especially true for those patent analysis methods that must access the content level of patent documents, such as the Technology Function Matrix (TFM). It is difficult to construct the TFM by merely relying on domain experts and information analysts[1]. Therefore, there is an urgent need for an automatic patent analysis tool to enable the accurate mining and correlation of high-value technology and function information.

The TFM[2], also known as the technology function map, is a widely used patent analysis method. As its name suggests, the TFM consists of two dimensions: technology and function. The technical dimension reflects the existing technical means in a given field, whereas the functional dimension reflects the functions that can be improved by the existing technical means. The intersection between the technical dimension and the functional dimension reflects the number of relevant patents or patent applicants, i.e., a technical means is adopted to improve a certain function. The greater the number of intersections, the more popular this kind of technology research in the current field. If the number is small or even zero, it would indicate that this kind of technology is a research gap. It can be seen that the TFM is an important basis for patent analysis, such as high-value technology discovery and potential technology function prediction.

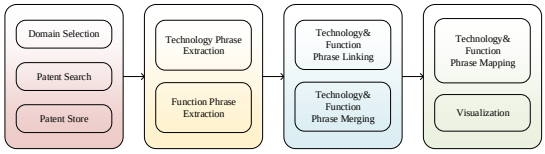

Figure 1: Construction process of TFM

In recent years, researchers have shown an increased interest in the automatic construction of TFM. The construction process of TFM generally comprises four steps[3]: patent document retrieval, technology and function phrase extraction, technology and function phrase merging, and TFM visualization, as shown in Figure 1. Among them, the extraction of technology and function phrases is the core research question for the automatic construction of TFM. Recently, a variety of technology and function phrase extraction approaches have been proposed, which can be divided into four main categories: rule-based, statistics-based, grammar structure-based, and machine learning-based. For instance, Hui and Yu

3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 20-24, 2022, Cologne, Germany and Online

*Corresponding author.
EMAIL: coding@whu.edu.cn (X. Shi); zikunfeng@whu.edu.cn (Z. Feng); laujames2017@whu.edu.cn (J. Liu); chengqikai0806@163.com(Q. Cheng); weilu@whu.edu.cn (W. Lu)

[4] first explored how to extract technology and function phrases by constructing a conceptual model and semantic templates. Chao et al. [5] and Trappey et al. [6] regarded technology and function phrases as keywords in the patent text and used the TF-IDF feature to extract them. Moreover, Choi et al. [7] focused on the analysis of the semantic structure of patent documents. They utilized the Subject-Action-Object (SAO) structure, which describes the semantic relationship between technology and function, to simultaneously extract the technology and function phrases. In a similar work, He et al. [8] regarded technology and function phrases as a Semantic Role Labeling (SRL) task and then extracted the corresponding technology and function phrases through an analysis of the predicate verbs and lexical parts of speech in a text. Teodoro et al. [9] considered the extraction of technology and function phrases as a sequence annotation task and introduced machine learning algorithms, e.g., Conditional Random Field (CRF), to extract phrases. In addition to the extraction-based methods introduced above, Cheng and Wang [10] adopted a classification-based method to map patent documents to patent classification systems, such as International Patent Classification (IPC) and Cooperative Patent Classification (CPC), and they were used to represent technology and function phrases. Although these methods indeed help to automatically construct the TFM to a certain extent, there are still some deficiencies, such as the undesirable recognition accuracy and the requirement for a large amount of annotation data. At the same time, some studies constructed the TFM without post-processing for the extracted technology and function phrases, which may introduce noise into the TFM.

To address these issues, we propose a semi-supervised learning method to extract technology and function phrases. Experiments show that our method has significant improvement in technology and function phrases recognition. In addition, we build a prototype system that supports human-computer interaction to assist experts in analyzing the current state of technology development in a certain field. In summary, our contributions are:

- We present a practical technology framework to automatically construct the technology function matrix.
- We propose a semi-supervised method to integrate the semantic dependency parser and the pre-trained language model to extract technology and function phrases. This not only reduces the labor cost, but also ensures the accuracy of technology and function phrase extraction.
- We build an interactive and visualized TFM construction system that automatically extracts technology and function information from patent documents in certain fields.

2. Methods

As the construction process of the TFM shown in Figure 1, our goal is to generate a TFM given a set of patent documents $\mathcal{D} = \{d_1, d_2, \cdots, d_n\}$ in a certain field. For patent retrieval, the patent documents set $\mathcal{D}$ is obtained from websites by using a web crawler. For the extraction of technology and function phrases, a semi-supervised strategy is introduced by combining the semantic dependency parser and the pre-trained language model (PLM). For the merging of technology and function phrases, we build a dictionary of upper and lower categories and synonyms by using classification systems such as IPC or CPC. Finally, we visualize the TFM with the bubble chart. We will depict each step of our proposed method in detail.

2.1. Patent Retrieval

We choose the Espacenet[1] patent retrieval system, which is open to the public and published by the European Patent Office (EPO), as the main data source to acquire the set $\mathcal{D}$ of patent documents. Then, by means of HTML analysis, 88,576 patent documents in the field of "New Energy Vehicles" are crawled. These patent documents mainly included vital information such as IPC, CPC, and Chinese abstracts. IPC and CPC are used as key references for assigning labels to technology and function phrases. Chinese abstracts are used to extract technology and function phrases, which are then used to construct the TFM.

2.2. Function Phrase Extraction

For the extraction of function phrases, we introduce a two-step pipeline method. First, we adopt a PLM-based method for recognizing function sentences. Then we use a semantic dependency parser and a template to recognize function phrases based on function sentences.

2.2.1. Function sentence recognition

We formulate function sentence recognition as a text classification task that contains only two classes $\mathcal{Y} = \{Yes, No\}$. Given a sentence $\boldsymbol{x} = \{w_1, w_2, \cdots, w_m\}$, our first step is to convert the sentence to the input sequence $\{[\text{CLS}], w_1, w_2, \cdots, w_m, [\text{SEP}]\}$, and then we use BERT[11] model to encode the input sequence to obtain the contextual representation of tokens $\{h_{[\text{CLS}]}, h_{w_1}, h_{w_2}, \cdots, h_{w_m}, h_{[\text{SEP}]}\}$. Finally, the hidden vector of classifier token $h_{[\text{CLS}]}$ is fed to a MLP to compute the probability distribution over the class set $\mathcal{Y}$.

$$\hat{y^c} = \text{softmax}(W_2^c(\text{relu}(W_1^c h_{[\text{CLS}]} + b_1^c)) + b_2^c) \quad (1)$$

[1] https://worldwide.espacenet.com/

2.2.2. Function phrase recognition

Based on the semantic dependency parser, we use a bootstrapping strategy to extract a function-phrase-matching template. The strategy consists of four main steps:

1. We manually select a few seed words and function sentences. Then we use the pre-trained Word2Vec[12] model to recall similar words to expand the seed vocabulary.
2. We analyze function sentences through the semantic dependency parser. Matching templates are created by combing seed words with the parts of speech and dependency between words.
3. We use the matching template to extract the function phrases and expand the seed vocabulary by evaluating the results.
4. Repeating step (2) and step (3). After each round of template updating, calculate the F1 score of the function phrase extraction. If the F1 score is increased, the new matching template will be retained.

2.3. Technology Phrase Extraction

Different from a function phrase, a technology phrase may appear in various positions of a document. Instead of directly utilizing the semantic dependency parser to construct matching templates, we choose to combine some characteristic words to automatically generate annotated samples. Specifically, we extract technology phrases by using a span-based named entity recognition model (Span-BERT[13]).

2.3.1. Training set generation

As shown in Figure 2, we use semantic dependency parser to generate the training set for technology phrase extraction. First, we retrieve some technology words related to the field of "New Energy Vehicles" from the Internet, and we then use them as the core words to construct the training set. Second, we extract the content containing these words from the patent documents. Third, we take the left-most five words and the right-most five words in the sentence as the context of the word. Moreover, we use ancestor words and sub-words as lexical features. Ancestor word specifically refers to the syntactic parent of the core word, whereas sub-word refers to the word that removes modifier. As seen in Figure 2, we select "improved genetic algorithm" as the core word. The ancestor word and the sub-word are "operating" and "algorithm", respectively. Finally, we concatenate these characteristic words to form the training data.

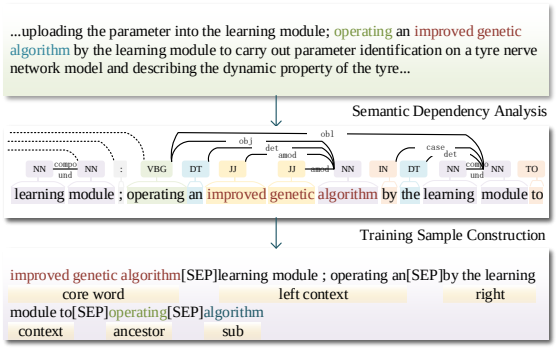

Figure 2: Training set generation.

2.3.2. Technology phrase recognition.

Based on the construction of the training set, we consider technology phrase recognition as a span classification task (the span here represents the core word). Firstly, we fed the input into the BERT encoder to obtain the representation of the span. Then a series of representations are aggregated using a fusion function $f(e_i, e_{i+1}, \cdots, e_{i+k})$. Note that in this paper, we adopt the max-pooling as the fusion function. Finally, we concatenate the aggregated representation with the span width embedding to form the span representation, where $\oplus$ denotes concatenation and w_k denotes width embedding:

$$e(s) := f(e_i, e_{i+1}, \cdots, e_{i+k}) \oplus w_k \tag{2}$$

The span representation is fed into a softmax classifier to predict whether the span is a technology phrase or not:

$$\hat{y^s} = \text{softmax}(W^s e(s) + b^s) \tag{3}$$

Similar to the function word extraction, we can continuously optimize the training sets until the model is sufficiently effective.

2.4. Technology and Function Phrase Merging

Technology and function phrases merging or linking is an important post-processing step that aims to unify words having similar semantics and make the TFM more accurate. Therefore, we construct a large-scale dictionary of upper and lower categories and synonyms. The following methods are adopted to construct the dictionary:

- *Directory tree crawling.* It refers to crawling existing classification systems, such as ACM and IPC, by web crawlers as the basis for constructing upper-class and lower-class dictionaries.

- *Abbreviation recognition.* It refers to using the maximum entropy model to identify various abbreviations for words and establish a synonymous relationship between words, such as "Support Vector Machine" and "SVM".
- *Domain triplet recognition.* It refers to using knowledge extraction techniques to extract the hyponymy and synonymy relations between phrases, such as (A, ISA, B), etc.
- *Suffix tree pattern recognition.* It refers to using a suffix tree string matching algorithm to find words having the same sub-words, such as "LDA" and "author LDA".

2.5. Visualization

The last step is to construct and visualize the TFM. We use the co-occurrence relationship between technology and function phrases in patent documents to map them and calculate the co-occurrence frequencies. Then, we develop a prototype system to represent the construction process of the TFM and use a bubble chart to demonstrate the results.

3. Experiment

3.1. Dataset and Implementation Details

To evaluate our proposed framework, we manually annotated 1,000 function sentences, 532 function phrases, and 907 technology phrases from patent data as a test set. For evaluating the function sentence recognition performance, we compare the BERT model with the traditional sentence classification algorithm, including the Naive Bayes and Word2Vec-based Multilayer Perceptron (Word2Vec+MLP). For evaluating the performance of the function phrase extraction, we compare our methods with the variant only using the SAO structure to extract function phrases. When selecting the best models for technology phrase extraction, we compare our method with a method without using characteristic words.

For BERT encoder version, we adopt the bert-base-chinese[14], which is pre-trained on a Chinese corpus. The batch size is set as 4. We adopt the Adam optimizer[15] and set the learning rate as 5e-5. In the Span-Bert model, we set the dimension of the width embeddings w_k as 25.

3.2. Results and Analysis

3.2.1. Function Sentence Recognition Results.

We use the standard classification evaluation metric, accuracy, to report the results of function sentence recognition, as shown in Table 1. The results show that Bert outperforms the traditional classification models in terms of function sentence identification. By using a PLM, we can improve the overall performance of the function phrase extraction.

Table 1
Comparative experimental results of function sentence recognition.

Algorithm	Accuracy
Naive Bayes	65.86
Word2Vec+MLP	65.67
Bert	89.13

3.2.2. Function Phrase Extraction Results.

Because the extraction of function phrases is essentially a NER task, *Precision, Recall*, and *F1 score* are adopted to evaluate the performance. As shown in Table 2, we can observe a significant improvement compared with the baseline (SAO). Previous work[16] usually perceives the function phrase as a "Verb + noun (or noun phrase)" pair. However, this assumption brings considerable noise to the function phrase recognition. For example, in Figure 2, "operating algorithm" is a "Verb + noun" pair, but it is not a function phrase. To alleviate this problem, we screen some trigger words to constrain the process of extracting function phrases, which reduces the noise.

Table 2
Experimental results of function phrase extraction.

Algorithm	Precision	Recall	F1 score
SAO	20.14	26.16	22.76
SDP + Template	56.83	48.59	52.39

3.2.3. Technology Phrase Extraction Results.

Table 3 shows that models trained on ancestor words and sub-words achieve better extraction performance. Compared with the model without using the features of the ancestor words and sub-words, the full version of our method has about 4% improvement in terms of F1 score. This means that the result of semantic dependency parsing can be regarded as a priori knowledge that provides some semantic information for the recognition of technology phrases.

3.3. System Overview

According to the technology framework for the TFM construction, we built a prototype system. The workflow of our system include the five steps: 1) *project creation.*

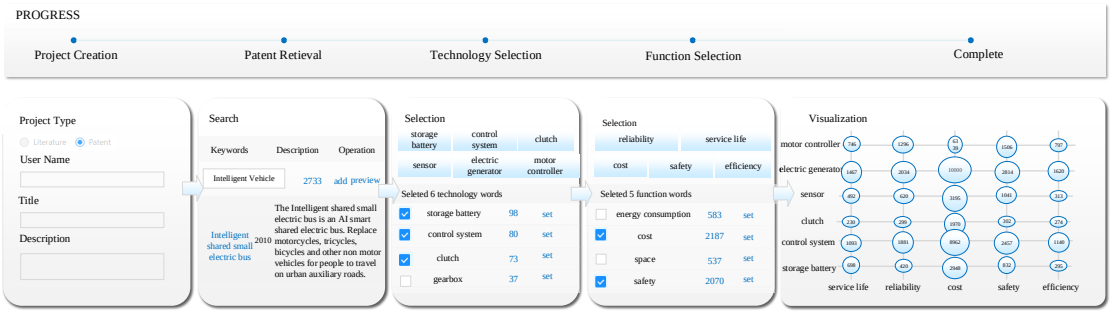

Figure 3: Workflow of TFM construction system.

Table 3
Technology Phrase Extraction Measurement

Algorithm	Precision	Recall	F1 score
Span-BERT	35.55	81.25	49.46
+ Ancestor	53.81	52.14	51.61
+ Sub	46.94	63.50	52.96
+ Ancestor + Sub	47.90	60.52	53.48

Figure 4: System Interface.

inputting the description of the project; 2) *patent retrieval.* inputting keywords to select the patent documents in a specific field; 3)*technology selection.* selecting the technical means that apply in specific fields; 4) *function selection.* selecting the function that corresponds to these technical means; 5) *TFM generation.* visualizing the TFM in the form of a bubble chart. The overall workflow is shown in Figure 3.

An anonymous online platform[2] is established to demonstrate the process and the results of our system. Figure 4 shows the system interface. Users can follow the operation process mentioned above to construct a TFM and review the results provided by the system. Moreover, users can test the effectiveness of the constructed TFM by using the tool set interface.

[2]http://124.70.200.79:8088/index.html

4. Conclusion

In this paper, we propose a framework for automatically constructing the TFM. We show that by combining semantic dependency parser with a template and a PLM, we can effectively extract the technology and function phases from patent documents with only a small number of annotated data. Based on the proposed framework, we also develop a prototype system that supports human-computer interaction. It can help experts and information analysts to grasp the development status of a certain technical field quickly and accurately, and provide support for the scientific and technological strategies of enterprises and countries.

In the future, we will further explore the solution in low-resource settings and improve the performance of technology and function phrase extraction and resolution. Moreover, we will extend the visual interaction to high-dimensional data visualization.

References

[1] T.-Y. Cheng, A new method of creating technology/-function matrix for systematic innovation without expert, Journal of technology management & innovation 7 (2012) 118–127.

[2] S.-I. Suzuki, Introduction to patent map analysis, Japan Patent Office,Asia-Pacific Industrial Property Center (2011). URL: https://www.jpo.go.jp/e/news/kokusai/developing/training/textbook/document/index/Introduction_to_Patent_Map_Analysis2011.pdf.

[3] Y. Yang, G. Ren, Hanlp-based technology function matrix construction on chinese process patents, International Journal of Mobile Computing and Multimedia Communications (IJMCMC) 11 (2020) 48–64.

[4] B. Hui, E. Yu, Extracting conceptual relationships from specialized documents, Data & Knowledge Engineering 54 (2005) 29–55.

[5] M.-H. Chao, A. J. Trappey, C.-T. Wu, Emerging technologies of natural language-enabled chatbots: A review and trend forecast using intelligent ontology extraction and patent analytics, Complexity 2021 (2021).

[6] A. J. Trappey, C. V. Trappey, U. H. Govindarajan, A. C. Jhuang, Construction and validation of an ontology-based technology function matrix: technology mining of cyber physical system patent portfolios, World Patent Information 55 (2018) 19–24.

[7] S. Choi, J. Yoon, K. Kim, J. Y. Lee, C.-H. Kim, Sao network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells, Scientometrics 88 (2011) 863–883.

[8] Y. He, Y. Li, L. Meng, A new method of creating patent technology-effect matrix based on semantic role labeling, in: 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), IEEE, 2015, pp. 58–61.

[9] D. Teodoro, J. Gobeill, E. Pasche, P. Ruch, D. Vishnyakova, C. Lovis, Automatic ipc encoding and novelty tracking for effective patent mining, in: The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, 2010.

[10] T.-Y. Cheng, M.-T. Wang, The patent-classification technology/function matrix-a systematic method for design around (2013).

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] R. Rehurek, P. Sojka, Software framework for topic modelling with large corpora, in: In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, Citeseer, 2010.

[13] M. Eberts, A. Ulges, Span-based joint entity and relation extraction with transformer pre-training, arXiv preprint arXiv:1909.07755 (2019).

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[15] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[16] W. Ki, K. Kim, Generating information relation matrix using semantic patent mining for technology planning: a case of nano-sensor, IEEE Access 5 (2017) 26783–26797.