A Patent Semantic Representation Using Technical Compound Sentences

Shuxuan Xiang[1], Jin Mao[2,3,*] and Gang Li[2,3]

[1]Laboratory of Data Intelligence and Interdisciplinary Innovation, Nanjing University, Nanjing 210000, China
[2]School of Information Management, Wuhan University, Wuhan 430072, China
[3]Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China

Abstract

The claims of a patent define the scope of exclusive rights to an invention, containing all essential technical features reflecting the novelty and non-obviousness. Current patent text mining methods have not fully leveraged patent claims by considering the expression of technical features in patent claims. In this study, we clarify the textual structure of patent claims and model the claims in a patent as a tree by capturing the denpendency relationships among the patent claims. We derive patent technology compound sentences (TCS), then propose a novel patent semantic representation based on TCS. To evaluate the proposed patent representation, we apply relational and direct strategies of empirical evaluation to a dataset of USPTO. The results show that our TCS-based and quantity-quality-weighted representation for patents outperforms other methods on task of P2P similarity and automated IPC symbol classification, which suggest that TCS enables more efficient use of technical information of the patent claim. The potential application of the novel representation in novelty analysis is discussed as well. The foundamental patent representation method using TCS could unleash the value of patent claims as technical information resource, and have many potentials in improving many subsequent tasks of patent mining.

Keywords

Claim tree, patent semantic representation, technical compound sentence

1. Introduction

Patent documents are valuable resources for technology text mining. As a combination of legal and technical terms, patent text differs significantly from other types of documents as scientific articles [1, 2]. The characteristics of patent text should be considered and utilized in patent text mining. To this end, many recent techniques of patent mining have increasingly employed a few methods like information fusion and text reorganization [3, 4]. As an important element in patent document, patent claim outlines the scope of an invention's exclusive rights and include all essential technical elements that demonstrate its novelty and non-obviousness. Patent claim has been exploited by many applications of patent mining, including patent infringement detection [5, 6, 7], patent evaluation [8, 9, 10], patent classification and clustering [11, 12, 13], patent information representation, etc. Therefore, it is an essential task to design text processing methods of patent claims by fully leveraging their features. However, current studies have not yet clarified the textual structure of patent claims, nor designed improved methods to deal with patent claims. In this study, we propose a method of patent technology compound sentences (TCS) to structure patent claims, then apply it to design a novel patent semantic representation. We evaluate the proposed patent semantic representation on a patent dataset. The fundamental patent representation method based on TCS could unleash the resource value of patent claims, and have many potentials in improving many subsequent tasks of patent mining.

2. Related work

For patent semantic representations, terms and phrases [14, 15] or original text [16, 17, 18, 19, 20] are used as the input. Keywords extraction and subject-action-object (SAO) analysis are leveraged to describe the technologies embedded in the patent texts. These methods, however, could be unable to capture the relationships within the technical concepts and might overlook some of the technical specifics. The original text may be a superior option in terms of information integrity with the advancement of NLP techniques. Title and abstract of patent are desirable sources of technical information, yet the claim of patent alone is able to achieve state-of-the-art results [12]. Recently, a growing body of research has concentrated on applying patent claim in patent semantic representation for its delicate writing [3, 14, 18, 19, 20, 21]. Yet the virtue of patent claims' characteristics on NLP tasks are not always valued, and the particularities of patent claim are not dealt with properly. There have been some

***PatentSemTech'23: 4th Workshop on Patent Text Mining and Semantic Technologies**, July 27, 2023, Taipei, Taiwan.*

*Corresponding author.
xsx@smail.nju.edu.cn (S. Xiang); danveno@163.com (J. Mao); imiswhu@aliyun.com (G. Li)
0000-0002-3259-7169 (S. Xiang); 0000-0001-9572-6709 (J. Mao); 0000-0002-8336-4891 (G. Li)

further studies which optimize the input by attending to characteristics that distinguish patent text from other text types, such as information enhancement with patent citation [3], or input transformation according to claim structure [20]. These methods leverage idiosyncrasies of claim text to some extent. To our knowledge, little research on patent semantic representation utilizes the specific structure and internal logic of technical information within patent claim. Therefore, we contribute to the research on patent semantic representation by providing an embedding method that can capture the nuance internal logic of patent claims.

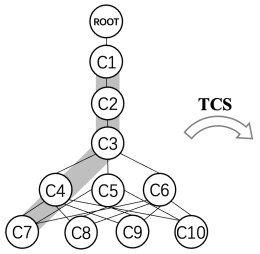
3. A representation using technical compound sentence

3.1. The tree structure of patent claims

The claims of patent can be classified into independent claims and dependent claims. Independent claims describe different embodiments or aspects, uses, or methods of producing the invention. Dependent claims refer back to and further limit another claim or the claims in the same application, to further limit the scope and complete the description with more details. The technological embodiments of dependent claims are embedded in the independent claims. With such structure, the patent claims can be model as a tree. Typically, each patent claim is provided as a separate numbered sentence, and the referenced claim is easily identified in the sentence. Theoretically, it is easy to identify the dependencies of patent claims and construct the tree structure of claims [22, 23, 24]. We refer such tree structure of claims as claim tree. In a claim tree, a claim follows serial dependency refers to the previous claim, and a claim follows parallel dependency refers to claim or claims before the previous one. Serial dependency between claims adds into the depth, and parallel dependency adds into the breadth, resulting in varying structures.

3.2. Construction of Technical Compound Sentence

The logical connections between technicalities embodied in the claims are reflected by the dependencies of claims. Therefore, a path from the root to the leaf nodes in claim tree denotes a chain of claims that together provide a full statement of an aspect, use, or method of fabricating the invention. A technical compound sentence (TCS) is constructed by combining the claims of the path in sequence. It is capable of grasping the progressive and explanatory relationships of claims, as well as the superior and subordinate relationships between technical concepts and technicalities. Furthermore, TCS enables the disambiguation of claims following the serial dependency and claims following the parallel dependency. The claims following the serial dependency add into the length of TCS, i.e., the technicalities volume of a full description. The claims following the parallel dependency add into the count of TCS, i.e., generalize and thus expand the scope of a patent. As shown in fig.1, the example patent claim can be break down into 12 TCSs, and each of them consists of 5 claims.



No	Technical Compound Sentence (TCS)
1	C1-C2-C3-C4-C7
2	C1-C2-C3-C4-C8
3	C1-C2-C3-C4-C9
4	C1-C2-C3-C4-C10
5	C1-C2-C3-C5-C7
6	C1-C2-C3-C5-C8
7	C1-C2-C3-C5-C9
8	C1-C2-C3-C5-C10
9	C1-C2-C3-C6-C7
10	C1-C2-C3-C6-C8
11	C1-C2-C3-C6-C9
12	C1-C2-C3-C6-C10

Figure 1: Claim Tree and TCS

3.3. Patent Representation Learning using TCS

We develop a method for semantic representations of patent based on technical compound sentence (TCS). The embedding vector of a patent is the weighted average of the embedding vector of its TCSs, where the weights are based on the quantity *Q(s)* and quality *F(s)* of the information the TCS contains. The representation is obtained through

$$\vec{P} = \frac{1}{\sum_{s \in S} W(s)} \sum_{s \in S} \vec{s} \times Q(s) \times F(s) \qquad (1)$$

A patent claim can be represented as a graph where nodes are terms of the claim. The graph-of-words of patent claim *C* is defined as*G = (V,E)* where V is the set of nodes that represents the nouns and verbs of *C* and *E* is the set of edges which represents the co-occurrence of the words in a 1-size window. Information quantity *Q(s)* of a TCS is determined by its cover of level *H(s)* and cover of breadth *R(s)* of the claims it includes. Cover of level *H(s)* is the maximum depth of a claim that form the TCS in the claim tree, which is positively related with more technical details. And cover of breadth *R(s)* is measured by radius of subgraph of the TCS G_s , which can describe the scope of technical information the TCS contains. Information quantity *Q(s)* is calculated with

$$Q(s) = H(s) \times R_s \qquad (2)$$

As for the information quality F(s) of a TCS, the k-core approach is employed [25], which focus on cohesiveness and connections of nodes (terms). The c_i-core of G is a subgraph G_{c_i}, in which the degree of nodes is greater than or equal to c_i. In the G_{c_i}, for the edge $D(v_m,v_n)$ linking the term v_m and v_n of G, its weight equals to the number of co-occurrences of two terms, and its core degree is c_i. Weight of the edge linking two terms and the core where those two terms appear are combined to calculate the information quality $F(s)$ as

$$F(s) = \sum_{i=1}^{k} \sum_{\substack{(\mathrm{v_m},\mathrm{v_n})\in \mathrm{s} \\ (\mathrm{v_m},\mathrm{v_n})\in G_{c_i}}} D\left(v_m, v_n\right) \times c_i \qquad (3)$$

The TCSs are then embedded using a custom Bert+SimCSE-unsup model, and the claim representation is obtained by taking weighted average of the TCSs embeddings. The whole process is illustrated in Figure 2.

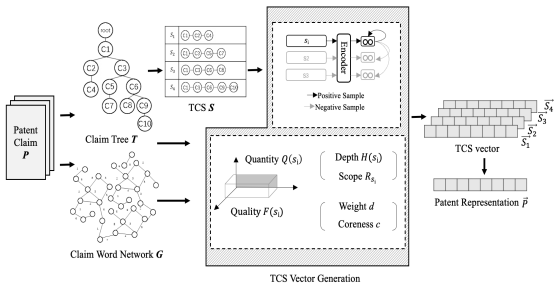


Figure 2: Patent Semantic Representation Using TCS

4. Experiments

4.1. Datasets

With the help of the Patent Public Search tool provided by the United States Patent and Trademark Office (USPTO), we gather claims, descriptions, and IPC assignments of 2114 patents that were submitted between 2016 and 2017 and contained the terms "quantum computing", "quantum computer" and "quantum computation" in their abstracts.

4.2. Evaluation

We apply "relational" and "direct" methodologies to evaluate the TCS-based and quantity-quality-weighted representation for patents [18]. The former method assesses the similarity of two items from the semantic representation and regularly used observable metrics such as IPC assignments [26, 27]. The correlation between the two similarities is investigated. The latter method analyzes the representation's performance in the prediction of the associated IPC classes [7]. Firstly, we demonstrate the benefit of TCS and the weighting strategy, by comparing with: (i.) full text of claim; (ii.) the first claim; (iii.) TCS + unweighted average; (iv.) TCS + quantity weighted average; (v.) TCS + quality weighted average. One should notice the above methods share the Bert+SimCSE-unsup model for embeddings. For good measure, other baseline models include: (vi.) PatentSBERTa [20]; (vii.) Technological Signature [18]; (viii.) Doc2vec [28]; (ix.) tfidf-Mittens [29]; (x.) Mittens+WR [30]. Each IPC of a patent can be represented by a tree for it comprises a hierarchically organized taxonomy, and the IPC tree of a patent is structured by additionally inserting a root node to unify the trees of all assigned IPC codes. The dissimilarity space embedding (DSE) is adapted for IPC representation [26, 31], which transform the IPC tree into a vector space. Given a distance function d, the dissimilarity space embedding of IPC is defined as

$$\varphi_n(c) : G \to \Re^n \varphi_n(c) = (d(c_1, c), d(c_2, c), \ldots, d(c_n, c)) \qquad (4)$$

Tree edit distance (TED) is employed as distance function. It is given by the minimal cost sequence of all operations including insertion, deletion, and relabeling transforming one tree to another. Then we calculate similarity by dot product of two representation vector. Besides, the absolute value of difference between 1 and the ratio of two similarities (i.e., the similarity derived from the representation and IPC assignment), which takes the form of

$$\mu = \left| \frac{\vec{p_1} \cdot \vec{p_2}}{\varphi_n(c_1) \cdot \varphi_n(c_2)} - 1 \right| \qquad (5)$$

is adopted in the variance analysis. Using TCS as the input format considerably improves the overall performance, as illustrated by Table 1. Additionally, the performance of the model is further enhanced by the weighting of quantity and quality developed on TCS. As a result, TCS with weight increases the model's efficiency for task of p2p similarity, and the use of TCS alone is able to boost the performance of patent representation in an observable way. We apply Z-test on μ to compare the average score of two patent semantic representations, and thus to testify the outperformance of embedding using TCS and the weight strategy based on TCS. As Table 2-4 depicts, the p-values are all less than 0.001, indicating that the null hypotheses are rejected and the differences across the models are not chance variations. We could come to the conclusion that TCS facilitates more effective use of technical information in the patent claim and could be effective in organizing technical information of patents. In addition, based on TCS, the weight of quantity and quality can result in superior patent semantic representation, allowing the representation to maintain a balance

Table 1
Performance of Patent Semantic Representations (i.)

Method	Relevance(%)	p-value
First claim	24.55	0.0043
Full claim	22.43	0.0032
TCS + unweighted average	26.37	0.0033
TCS + quantity weighted	27.67	0.0032
TCS + quality weighted	26.41	0.0031
TCS + quantity and quality weighted	27.72	0.0031
PatentSBERTa	13.63	0.0035
Technological Signature	17.90	0.0021
Doc2vec	21.72	0.0036
tfidf-Mittens	19.25	0.0035
Mittens+WR	22.16	0.0032

Table 2
Result of Z-Test (i.)

	TCS + unweighted average	Full-claim
Avg.	0.5797	0.6678
Std.	1.1598	1.5396
Z value	-18.4620	
P value (one-sided)	0.0000	

Table 3
Result of Z-Test (ii.)

	TCS + unweighted average	First claim
Avg.	0.5797	0.5929
Std.	1.1598	1.3834
Z value	-2.9395	
P value (one-sided)	0.0016	

Table 4
Result of Z-Test (iii.)

	TCS + weighted	TCS + unweighted
Avg.	0.5409	0.5797
Std.	0.9152	1.1598
Z value	-10.6267	
P value (one-sided)	0.0000	

between highlighting the key details and elaborating the full scope.

Furthermore, we examine whether the generated vectors can function as inputs for automated IPC symbol classification for the main section (In this case, binary classification of section G and section H). An artificial neural network (ANN) is deployed [18], which takes the representations as input and predicts the main section of the patent. Table 5 demonstrates that our method outperforms the baseline methods on this task, which indicates the capability of the presented method in semantic representation and proves the TCS as well as the weighting strategy effective.

Table 5
Performance of Patent Semantic Representations (ii.)

Method	Loss	Acc(%)	Pre(%)
TCS + quantity and quality weighted	0.489	74.65	66.67
PatentSBERTa	0.611	65.45	64.60
Technological Signature	0.605	69.18	62.50
Doc2vec	0.598	69.44	53.80
tfidf-Mittens	0.597	72.05	66.29
Mittens+WR	0.638	64.93	52.20

4.3. Application

We apply technical compound sentence (TCS) on novelty analysis. Innovation consists in carrying out new combinations. Actually, innovation is fundamentally the combination of facts, concepts, techniques, theories, goals, etc. [32]. Thus, for novelty analysis, the combinations held by the patent are vital and the combinations should be considered when conducting patent semantic search in novelty analysis. Patent claims define the boundary for an exclusive right granted by the patent office, and we may express the same thing by saying that each patent occupies a certain inventive space of the protecting parts of technologies that exclude other inventions. A TCS derived from a patent claim tree, naturally, describes a relatively separate segment of the entire space the claim defines, which means it contains the implicit combinations of an aspect or method the patent right intends to protect. Therefore, the relevant patents can be located and identified by matching similar TCS. By applying TCS embedding as the query, we are able to retrieve more of relevant items which might be novelty-prejudicial to the target patent for novelty assessment. Thus, TCS could improve the recall of patent retrieval in patent semantic search in novelty analysis.

5. Conclusion

A technical compound sentence (TCS) is composed of a set of claims that on the path from the root to the leaf nodes in a claim tree. The experiment's findings demonstrate that the employment of TCS enhances the performance of patent semantic representation. This indicates the capability of the TCS in technical information organization of patents. Additionally, the balance of emphasizing the key information and elaborating the full scope is achieved by the weight of quantity and quality built on TCS, which further improves the semantic representation. For future work, we will further explore the uses of TCS in the field of patent text mining, attempting to achieve efficient processing, interpretation, and utilization of patent texts.

References

[1] S. Casola, A. Lavelli, Summarization, simplification, and generation: The case of patents, Expert Systems with Applications 205 (2022).

[2] J. Wang, W. Lu, L. HanTong, A two-level parser for patent claim parsing, Advanced Engineering Informatics 29 (2015) 431–439.

[3] J. Qi, L. Lei, K. Zheng, X. Wang, Patent analytic citation-based vsm: Challenges and applications, IEEE Access 8 (2020) 17464–17476. doi:10.1109/ACCESS.2020.2967817.

[4] Y. Chi, H. Wang, Establish a patent risk prediction model for emerging technologies using deep learning and data augmentation, Advanced Engineering Informatics 52 (2022).

[5] C. Lee, B. Song, Y. Park, How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships, Technology analysis and strategic management 25 (2013) 23–28.

[6] H. Jang, S. Kim, B. Yoon, An explainable ai (xai) model for text-based patent novelty analysis, Available at SSRN (2023). doi:http://dx.doi.org/10.2139/ssrn.4341594.

[7] L. Du, W. Liu, K. Xiao, S. Gao, Y. Han, Technical function-effect based patent multi-to-one negation game model, in: 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2022, pp. 1443–1448. doi:10.1109/CSCWD54268.2022.9776122.

[8] S. Wittfoth, Measuring technological patent scope by semantic analysis of patent claims–an indicator for valuating patents, World Patent Information 58 (2019).

[9] A. J. H, Modeling patent clarity, Research Policy 51 (2022).

[10] E. Novelli, An examination of the antecedents and implications of patent scope, Research Policy 44 (2015) 493–507.

[11] D. H. Milanez, L. I. L. de Faria, R. M. do Amaral, J. A. R. Gregolin, Claim-based patent indicators: A novel approach to analyze patent content and monitor technological advances, World Patent Information 50 (2017) 64–72.

[12] J. Lee, J. Hsiang, Patent classification by fine-tuning bert language model, World Patent Information 61 (2020).

[13] S. Huang, H. Ke, W. Yang, Structure clustering for chinese patent documents, Expert Systems with Applications 34 (2008) 2290–2297.

[14] Z. Qiu, Z. Wang, Construction and application of patent technical element dependency network, IEEE Transactions on Engineering Management (2022) 1–15. doi:10.1109/TEM.2022.3227175.

[15] S. Yun, W. Cho, C. Kim, S. Lee, Technological trend mining: identifying new technology opportunities using patent semantic analysis, Information Processing and Management 59 (2022).

[16] Z. Qiu, Z. Wang, What is your next invention?—a framework of mining technological development rules and assisting in designing new technologies based on bert as well as patent citations, Computers in Industry 145 (2023).

[17] deGrazia Charles AW, J. P. Frumkin, N. A. Pairolero, Embracing invention similarity for the measurement of vertically overlapping claims, Economics of Innovation and New Technology 29 (2020) 113–146.

[18] D. S. Hain, R. Jurowetzki, T. Buchmann, P. Wolf, A text-embedding-based approach to measuring patent-to-patent technological similarity, Technological Forecasting and Social Change 177 (2022).

[19] L. Lei, J. Qi, K. Zheng, Patent analytics based on feature vector space model: A case of iot, 2019. arXiv:1904.08100.

[20] H. Bekamiri, D. S. Hain, R. Jurowetzki, Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert, 2021. arXiv:2103.11933.

[21] T. Roh, Y. Jeong, B. Yoon, Developing a methodology of structuring and layering technological information in patent documents through natural language processing, Sustainability 11 (2017).

[22] Y. T. Demey, D. Golzio, Search strategies at the european patent office, World Patent Information 63 (2020).

[23] L. FuRen, C. K. Chen, L. SzuYin, A hybrid patent prior art retrieval approach using claim structure and description, 2014, pp. 231–248.

[24] J. Rossi, M. Wirth, E. Kanoulas, Query generation for patent retrieval with keyword extraction based on syntactic features, 2019. arXiv:1906.07591.

[25] H. Mirisaee, E. Gaussier, C. Lagnier, A. Guerraz, Terminology-based text embedding for computing document similarities on technical content, 2019. arXiv:1906.01874.

[26] K. Riesen, H. Bunke, Graph classification based on vector space embedding, International Journal of Pattern Recognition and Artificial Intelligence 23 (2009) 1053–1081.

[27] Y. M. G. Costa, D. Bertolini, A. S. B. Jr., G. D. C. Cavalcanti, L. E. S. Oliveira, The dissimilarity approach: a review, Artificial Intelligence Review 53 (2020) 2783–2808.

[28] J. H. Lau, T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation, 2016. arXiv:1607.05368.

[29] N. Dingwall, C. Potts, Mittens: An extension of glove for learning domain-specialized representations, 2018. arXiv:1803.09901.

[30] E. Kawin, Unsupervised random walk sentence embeddings: A strong but simple baseline, in: Proceedings of the Third Workshop on Representation Learning for NLP, 2018, pp. 91–100. URL: https://aclanthology.org/W18-3012. doi:10.18653/v1/W18-3012.

[31] K. Frerich, M. Bukowski, S. Geisler, R. Farkas, On the potential of taxonomic graphs to improve applicability and performance for the classification of biomedical patents, Applied Sciences 11 (2023).

[32] H. Michael, T. Kiessling, M. Moeller, A view of entrepreneurship and innovation from the economist "for all seasons": Joseph s. schumpeter, Journal of Management History 16 (2010) 527–531.