Identification of core technological topics in the new energy vehicle industry: The SAO-BERTopic topic modeling approach based on patent text mining

Jianxin Zhu[1,2], Yutong Chuang[1], Zhinan Wang[1,2,*], Yunke Li[1]

[1] *Harbin Engineering University, School of Economics and Management 150001, China*
[2] *Key Laboratory of Big Data and Business Intelligence Technology, Ministry of Industry and Information Technology*

Abstract

In the new energy vehicle industry, precise identification of core technologies is the key to promoting innovation and maintaining market competitiveness. In this article, a comprehensive approach combining information weight method and SAO-BERTopic topic model is proposed to extract and analyze core technologies from large-scale patent data. Through in-depth analysis of Incopat patent database, we use the information weight method to select high-quality core patents from four dimensions: technological, strategic, law and market value. These selected patents form the basis of the research data, which is then applied to the SAO-BERTopic model, which combines the advanced semantic understanding capabilities of BERTopic with the fine-structured characteristics of SAO analysis, greatly improving the efficiency and accuracy of identifying complex technical topics. This innovation of this research is not only applicable to the analysis of the technological development of the new energy vehicle industry, but also can provide valuable reference for other high-tech industries such as biomedicine, renewable energy and information technology. These fields also need to identify core technologies from a large number of patents. SAO-BERTopic's structured analysis framework can help these industries to insight into technology development trends, identify innovation opportunities, and provide data-driven decision support for enterprises, research institutions and governments, thus playing an important role in technology planning and market commercialization.

Keywords

technical identification, patent analysis, new energy vehicle, SAO-BERTopic

1. Introduction

In a globalized economic environment, the rapid development of the new energy automobile industry has become an important symbol of technological innovation and industrial transformation[1]. China has made remarkable progress in this field[2,3]. However, in the face of the strategic layout and potential containment of traditional automobile powers such as the United States, Europe, Japan and South Korea in terms of technology and industrial chain, the sustainable growth and competitiveness of China's new energy automobile industry presents new challenges [4,5,6,7].

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), April 23~24, 2024, Changchun, China and Online

zhjx@vip.163.com (Jianxin Zhu); chuangyutong@hrbeu.edu.cn (Yutong Chuang); wzn6768@163.com (Zhinan Wang); liyunke1209@163.com (Yunke Li)

In the research of technological progress and innovation, patents and scientific papers are indispensable resources. However, due to the large scale of these literatures, traditional manual analysis methods are difficult to cope with, and expert analysis is subjective. Therefore, based on the large amount of technical information contained in patents [8], many scholars have devoted themselves to using data mining methods in recent years [9]. Among them, the topic model is an approach that can automatically extract key topics from documents, revealing technology trends and areas of innovation. Latent Dirichlet Allocation (LDA) is a word frequency-based models that may struggle to capture semantic complexity [10,11,12,13]. Latent Semantic analysis (LSA) has limitations in handling word sense diversity and polysemy, which can lead to information loss when analyzing specialized technical documents [14,15]. Correlated Topic models (CTM) high computational complexity limits its rapid application on large document collections [11,16] .

In view of the limitations of existing topic models in the identification of core technologies in the new energy vehicle industry, we propose an innovative topic model— SAO-BERTopic. The model combines SAO analysis and BERTopic, an advanced natural language processing technology, to improve the accuracy and efficiency of technology recognition.

The main innovations of this article are:

1. By combining SAO analysis and BERTopic topic model, we construct SAO-BERTopic method, which effectively solves the problem of identifying deep-level technology trends and key innovation points in the new energy vehicle industry.

2. By using the combination of the information weight method and the SAO-BERTopic model, we realize the selection of highly innovative and market-valued core patents from large-scale data in the patent analysis of the new energy vehicle industry for the first time, and improve the efficiency and accuracy of technology identification.

2. Research framework

We adopt an innovative hybrid approach, utilizing a phased scientific process to identify the core technological topics within the new energy vehicle industry. This process is divided into six detailed steps (as shown in **Figure 1**), designed to ensure the efficient and systematic extraction, analysis, and determination of core technologies from a broad range of patent data. After selecting patents related to core technologies using the information weight method from a large pool of patents, the SAO-BERTopic model is applied for an in-depth analysis of the selected patents. This model identifies the core technologies' SAO triplets and defines these triplets as specific topics.

Step One: Patent Selection — The first step is to use Incopat patent database to download patents, and use the information weight method to establish indicators in four dimensions of patent technical, strategic, legal and market value; Step Two: Textual Vector Representation — This step is text preprocessing and then vectorization via BERT(Bidirectional encoder representation from the transformer) model; Step Three: Dimensionality Reduction and Clustering — In this step, UMAP(Uniform manifold Approximation and projection) algorithm is used to reduce the dimensionality of the text vector, and then HDBSCAN(hierarchical densi-based applied spatial clustering with noise) algorithm is used to cluster the topic; Step Four: SAO Triplet Extraction — Using the C-TF-IDF(class-based word frequency-inverse document frequency) algorithm, SAO triples are extracted from the clusters and the topic representation is refined using the MMR(Maximum marginal correlation) algorithm to balance correlation and diversity; Step Five: Result Validation — The Calinski-Harabasz index and Davies-Bouldin index were used to compare the LDA, BERTopic and SAO-LDA topic models to evaluate the clustering effectiveness. In addition, dimensionality reduction visualization via UMAP provides an intuitive comparison of cluster distributions; Step Six: Strategic Recommendations — Based on the results, recommendations are made from the corporate, industry and governance perspectives.

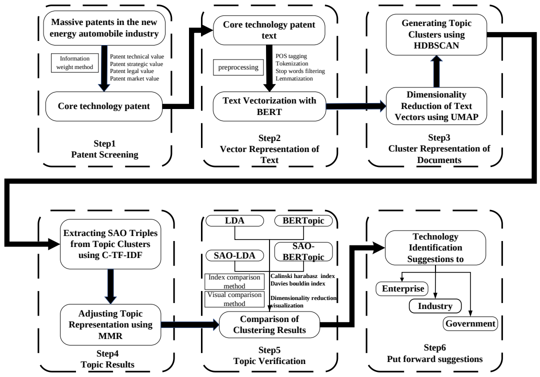
Massive patents in the new energy automobile industry
Information weight method
Patent technical value
Patent strategic value
Patent legal value
Patent market value
Core technology patent
Step1
Patent Screening
Core technology patent text
preprocessing
POS tagging
Tokenization
Stop words filtering
Lemmatization
Text Vectorization with BERT
Step2
Vector Representation of Text
Generating Topic Clusters using HDBSCAN
Dimensionality Reduction of Text Vectors using UMAP
Step3
Cluster Representation of Documents
Extracting SAO Triples from Topic Clusters using C-TF-IDF
Adjusting Topic Representation using MMR
Step4
Topic Results
LDA
BERTopic
SAO-LDA
SAO-BERTopic
Index comparison method
Visual comparison method
Calinski harabasz index
Davies bouldin index
Dimensionality reduction visualization
Comparison of Clustering Results
Step5
Topic Verification
Technology Identification Suggestions to
Enterprise
Industry
Government
Step6
Put forward suggestions

Figure 1: Overall research process

3. Analysis results

This article uses the Incopat patent database to collect data pertinent to the new energy vehicle industry, the topics are finally refined into five core topics.

Topic 0 is "Energy Management and Power Transmission Technology in Hybrid Electric Vehicles". Topic 1 is "Electric Vehicle Charging Systems and Energy Management", where the SAO triples contained in topic 1 focus on describing the functionalities of receiving, managing, and providing energy in charging panels and systems. Topic 2 is "Battery System Integration and Energy Efficiency Management", where the SAO triples collectively depict various aspects of energy storage system design, management, and application, including modular composition, energy control and management, system power supply, and integration with electric vehicles. Topic 3 is "Electric Vehicle Battery Pack Configuration and Structural Design", covering various aspects of the physical configuration, structural design of battery packs, and how these designs impact battery pack performance. Topic 4 is "Cathode Materials and Composition for Secondary Batteries", focusing on the composition and design of cathode materials for batteries, especially secondary batteries, which directly impact battery performance aspects like capacity, energy density, charge/ discharge speed, and lifespan.

4. Conclusions

This research successfully applied the information weight method to screen high-quality core patents from a large-scale patent dataset, and employed the SAO-BERTopic model to identify core technological topics from the filtered data. Our empirical analysis results highlight the superior clustering performance of SAO-BERTopic in identifying technological topics in the new energy vehicle domain compared to traditional models. The approach presented in this research, through its efficient processing of technical terms and complex concepts and its precise screening of patent data, demonstrates the ability to identify core technologies and drive innovation across a wide range of fields. Therefore, this research not only has a direct contribution to the technological progress of the new energy vehicle industry, but also provides a new perspective and tool for technology identification and innovation management in various fields, which is helpful to promote the technological development and innovation strategy formulation in a wider range of fields.

References

[1] Y. Li, The impact of economic systems and financial systems on new energy vehicle industry, Adv. Econ. Manag. Polit. Sci. 33 (2023) 162-170. doi:10.54254/2754-1169/33/20231622.

[2] Z. Liu, H. Hao, X. Cheng, F. Zhao, Critical issues of energy efficient and new energy vehicles development in china, Energy Policy 115 (2018) 92-97. doi:10.1016/J.ENPOL.2018.01.006.

[3] P. Yu, J. Zhang, D. Yang, X. Lin, T. Xu, The evolution of China's new energy vehicle industry from the perspective of a technology-market-policy framework, Sustainability 11 (2019) 1711. doi:10.3390/SU11061711.

[4] M. Kendall, Fuel cell development for new energy vehicles (NEVs) and clean air in china, Prog. Nat. Sci. Mater. Int. 28 (2018) 113-120. doi:10.1016/J.PNSC.2018.03.001.

[5] T. Yang, C. Xing, X. Li, Evaluation and analysis of new-energy vehicle industry policies in the context of technical innovation in china, J. Clean. Prod. 281 (2021) 125126. doi:10.1016/J.JCLEPRO.2020.125126.

[6] X. L. Xu, H. H. Chen, Exploring the innovation efficiency of new energy vehicle enterprises in china, Clean Technol. Environ. Policy 22 (2020) 1671-1685. doi:10.1007/S10098-020-01908-W.

[7] S. Chen, Y. Feng, C. Lin, Z. Liao, X. Mei, Research on the technology innovation efficiency of China's listed new energy vehicle enterprises, Math. Probl. Eng. 2021 (2021) 6613602. doi:10.1155/2021/6613602.

[8] C. Lee, A review of data analytics in technological forecasting, Technol. Forecast. Soc. Change 166 (2021) 120646. doi:10.1016/J.TECHFORE.2021.120646.

[9] N. Su, Z. Tan, Review and vision for the future of the research on technology opportunity analysis methods, Inf. Stud. Theory Appl. 43 (2020) 179–186.

[10] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993-1022.

[11] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 5228-5235. doi:10.1073/PNAS.0307752101.

[12] J. Lafferty, D. Blei, Topic models, in: A. N. Srivastava and M. Sahami (Eds.) Text mining, Chapman and Hall/CRC, New York, NY, 2009, pp. 71-93. doi:10.1201/9781420059458.CH4.

[13] M. Hoffman, F. Bach, D. Blei, Online learning for latent dirichlet allocation, in: 24th Annual Conference on Neural Information Processing Systems 2010, NeurIPS, New York, NY, 2010, pp. 856–864.

[14] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391-407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[15] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Process 25 (1998) 259-284. doi:10.1080/01638539809545028.

[16] J. Lafferty, D. Blei, Correlated topic models, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, 2005, pp. 147–154. doi:10.5555/2976248.2976267.