big data and cognitive computing

MDPI

Article

Research on Emerging Technology Identification Methods Based on a Knowledge Graph of High-Value Patents

Chuan Zhan [1], Yang Zhou [2,*] and Yanping Huang [2]

[1] School of International Business, Chongqing Technology and Business University, Chongqing 400067, China; zchuan@ctbu.edu.cn
[2] School of Management Science and Engineering, Chongqing Technology and Business University, Chongqing 400067, China; 2023336007@ctbu.edu.cn
* Correspondence: zhouyang68@ctbu.edu.cn

Abstract

In the context of a new wave of scientific and technological revolution and industrial transformation, this study proposes an emerging technology identification framework that integrates a High-Value Patent Knowledge Graph with Social Network Analysis, aiming to systematically uncover the semantic and structural relationships embedded in patent data and to support national efforts to secure strategic technological advantages. First, patent textual feature scores are extracted using the Doc2Vec model, while indicator feature scores are calculated across the technical, legal, and economic dimensions using the CRITIC weighting method. These two types of scores are then integrated to derive a comprehensive patent value score, and high-value patents are screened according to the Pareto principle. Subsequently, a High-Value Patent Knowledge Graph is constructed based on entity extraction using the BERT-BiLSTM-CRF model and relationship matching techniques. Building upon this graph, centrality analysis is conducted on the nodes, and the results are combined with the rich semantic relationships represented in the knowledge graph to further identify emerging technologies. Taking the New Energy Vehicle domain as an empirical case, a High-Value Patent Knowledge Graph comprising seven types of entities, six types of relationships, and 25,611 triplets is developed, through which six key emerging sub-technology directions are identified. The empirical findings demonstrate the effectiveness and robustness of the proposed approach for emerging technology identification.

Keywords: knowledge graph; high-value patents; emerging technologies; new energy vehicle

1. Introduction

Technological innovation is the fundamental driving force of high-quality economic development, with emerging technologies serving as a critical manifestation now central to national strategic planning. In the context of a new wave of scientific and technological revolution and accelerating industrial transformation, these forward-looking, strategic technologies possess the potential to both guide industrial innovation and shape a nation's future competitiveness. As global technological competition intensifies and monopolization in key domains grows, the accurate identification of emerging technologies is essential for securing a strategic advantage. However, the intrinsic uncertainty and complexity of these technologies pose substantial challenges to the assessment and prediction of their developmental trajectories.

Check for updates

Received: 16 December 2025
Revised: 12 January 2026
Accepted: 26 January 2026
Published: 28 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

The rapid advancement of computer science, electronics, and information technologies has introduced new paradigms for technology identification, particularly those grounded in big data analytics and artificial intelligence. Patents, an extensive and information-intensive corpus, are critical carriers of innovation, yet their sheer volume and informational density make identifying emerging technologies akin to locating a needle in a haystack. Furthermore, existing research often remains focused on macro- and meso-levels, offering limited insight into the intrinsic technological attributes needed for accurate characterization and identification of developmental dynamics.

Against this backdrop, the primary goals of this research are explicitly defined:

1. Construct a comprehensive high-value patent screening system that integrates multi-dimensional text and indicator features to establish a robust foundation for emerging technology identification.
2. Develop an interpretable identification framework that leverages high-value patents to build a patent knowledge graph and, through the incorporation of social network analysis, uncovers concept-level technological relationships beyond conventional lexical representations.
3. Empirically apply and validate the proposed framework in the new energy vehicle sector to reveal potential emerging technologies and provide evidence-based insights to inform technological innovation and policy development.

The subsequent sections of this paper are structured as follows. Section 2 briefly reviews relevant research on high-value patent identification, patent knowledge graphs, and emerging technology identification. Section 3 describes the details of the proposed analytical framework and its validation methods. Section 4 conducts experiments using new energy vehicle as an exemplar field. Section 5 discusses theoretical and practical implications, and suggests future research directions, followed by the conclusion in Section 6.

2. Related Work

2.1. High-Value Patent Identification

Patent literature, serving as a critical repository of technical information, is characterized by inherent legal, technical, and economic attributes, which renders the valuation of patent assets a complex and challenging endeavor. Current research endeavors typically evaluate patent value through the establishment of multi-dimensional indicator systems, utilizing methodologies such as the Analytic Hierarchy Process [1], Factor Analysis [2], and the Entropy Weight Method [3]. These methods construct multi-dimensional indicator frameworks, calculating a comprehensive value score based on various aspects, including technology, law, market, economics, strategy, and management. Nevertheless, these indicator systems suffer from a significant reliance on external metrics, thus exhibiting notable limitations in terms of objectivity, scalability, and their capacity to capture the intrinsic technical innovativeness and complexity embedded within the patent itself.

With the rapid development of natural language processing and machine learning techniques, an increasing number of studies have focused on applying these methods to text-based analysis of patent value. By leveraging text vectorization, topic modeling, and other approaches, latent value information within patent texts can be effectively extracted. For instance, Lee et al. [4] trained a Doc2Vec model to vectorize patent documents and employed methods such as the interquartile range, three-sigma rule, and isolation forest to identify outlier patents. Deng et al. [5] proposed a patent novelty evaluation method that combines Sentence-BERT with the isolation forest algorithm, integrating text vectorization, anomaly detection, and technical metrics for comprehensive assessment of patent novelty. Song and Ran [6] utilized topic models to extract the technological themes

covered by patents for clustering, and identified high-value and outlier patents based on the Catastrophe Progression Method and outlier factor algorithm.

However, the identification of high-value patents essentially involves the evaluation and prediction of a patent's potential value. Assessing this potential requires not only the analysis and judgment of a patent's intrinsic indicator attributes but also the examination of its text features [7]. This is because the patent text itself contains the most critical information regarding its technological value, legal and market information, and commercial worth. Current research into the identification of high-value patents predominantly concentrates on two distinct approaches: data measurement based on indicator attributes or semantic mining directed at the patent text content. Although some existing literature has attempted to integrate both the indicator features and the text features for a comprehensive value assessment, such combined studies remain relatively scarce.

2.2. Patent Knowledge Graphs

The concept of a knowledge graph was proposed by Google in 2012. Essentially, it is a structured semantic network that connects different entities and their relationships in the form of "entity–relation–entity" triples, thereby constructing a large and complex knowledge network. Existing studies have explored the integration of patent data with knowledge graphs. Some patent knowledge graphs focus on external features of patents. Li et al. [8] constructed a patent knowledge graph using patent citations, inventors, applicants, and classification information to measure knowledge proximity. Chen and Deng [9] developed a domain-specific knowledge graph based on patent assignees, inventors, categories, and countries, and proposed an interpretable patent recommendation method. Other knowledge graphs emphasize the deep semantic mining of patent content. Siddharth et al. [10] extracted engineering fact triples from the syntactic characteristics of patent claims and integrated them into a large-scale engineering knowledge graph, which supports retrieval, representation, reasoning, and inference in engineering design. At present, entity recognition and relation extraction from patent texts have been widely studied [11–13], providing a theoretical and technical foundation for the construction of patent knowledge graphs.

2.3. Emerging Technology Identification

A unified consensus on the definition of emerging technologies has yet to be established, and scholars have analyzed and interpreted the concept from different perspectives. Day and Reibstein [14] first proposed the term "emerging technology" emphasizing its innovativeness and disruptive impact on traditional industries, often leading to the creation of new sectors. Cozzens et al. [15] argued that emerging technologies are those that demonstrate great potential but have not yet manifested their value or reached any consensus regarding their impact.

Research on emerging technology identification methods generally falls into three categories: qualitative analysis, quantitative analysis, and a combination of both. In terms of qualitative analysis, commonly used methods include the Delphi method [16], scenario analysis [17], and the Analytic Hierarchy Process [18]. These approaches primarily rely on experts' subjective judgments as criteria for identifying emerging technologies, which may introduce cognitive biases and are less suitable for emerging technology identification in the era of big data.

Quantitative methods, on the other hand, include bibliometrics, patent analysis, and other data-driven approaches. Quantitative approaches can complement the limitations of qualitative methods by providing objective results based on large-scale data. Among them, patent analysis is a representative quantitative method for emerging technology

identification and has become a mainstream approach in the field. Other quantitative techniques, such as text mining and network analysis, are increasingly being integrated with patent-based methods.

The application of text mining in emerging technology identification primarily focuses on extracting semantic information from textual data. The two most commonly used approaches are topic model-based methods and structure-semantic analysis-based methods. Hu et al. [19] developed an LDA2Vec-BERT topic mining model to identify and analyze the evolution of fine-grained emerging and hot technology topics in the blockchain domain. Yun et al. [20] integrated SAOx extraction, clustering for standardization, sequential pattern mining, and comparative analysis to predict new technology opportunities.

Network analysis methods explore the interaction patterns and development trajectories of technological elements by constructing networks either among patents or among technological features. These networks can be classified into citation networks (including citations, co-citations, and bibliographic coupling), co-occurrence networks, and similarity networks, among others. Ning et al. [21] constructed a patent citation network and combined the KeyGraph algorithm with CPM analysis to identify emerging technologies. Li et al. [22] constructed a temporal heterogeneous network, combined with structural and semantic features, to dynamically identify technology opportunities. Zhou et al. [23] constructed a patent citation coupling matrix and employed cosine similarity to measure patent similarity, further identifying outlier patents within the network as candidate emerging technologies.

Given the potential of knowledge graphs in integrating multi-source information, some researchers have begun to adopt knowledge graph-based approaches combined with deep learning for emerging technology identification, aiming to address the limitations arising from neglecting the semantic information in patent texts, which can affect both the accuracy and interpretability of identification. Lee et al. [24] constructed a knowledge graph encompassing technologies, new technology-based enterprises, and their investors to reveal the interrelations among these elements in technology opportunity discovery, and proposed a Technology Opportunity Discovery Index to quantify technological potential. Wei et al. [25] combined knowledge graph techniques with deep neural network algorithms to train an industrial emerging technology identification model, achieving precise identification of emerging technologies in industry. However, the data used in most knowledge graph applications related to emerging technology identification are limited to external patent attributes such as patent titles, types, applicants, inventors, classification codes, or keywords. These studies primarily focus on exploring relationships among patents through knowledge graphs, supplemented by keywords, without deeply mining the fine-grained content of patent documents. From this perspective, there remains substantial research potential for the application of knowledge graphs in this field.

In summary, this study introduces a method that combines indicator-based evaluation with text vectorization for patent value assessment, enabling the selection of high-value patents. A knowledge graph of these high-value patents is then constructed and integrated with social network analysis to identify emerging technologies. By semantically reconstructing patent data, the representation of patent texts shifts from a lexical to a conceptual level, strengthening the semantic associations and logical connections among technological concepts. This method enhances the interpretability of emerging technology identification and provides a solid knowledge foundation for the discovery of prospective technology pathways and the formulation of strategic science and technology decisions.

3. Research Framework

This research framework is divided into four phases—data collection and preprocessing, identification of high-value patents, construction of a high-value patents knowledge graph, and identification of emerging technology—as shown in Figure 1. Initially, patent data for the target domain is collected. For high-value patent identification, both text features (measuring novelty using Doc2Vec) and indicator features (weighted by the CRITIC method across technical, legal, and economic dimensions) are synthesized, and the Pareto principle is used to select high-value patents. Subsequently, a high-value patent knowledge graph is built using the BERT-BiLSTM-CRF model and rule-based methods, and stored in Neo4j. Finally, through centrality analysis of the knowledge graph, nodes exhibiting low degree centrality but high betweenness centrality are regarded as potential technological opportunities, thereby identifying emerging technologies.

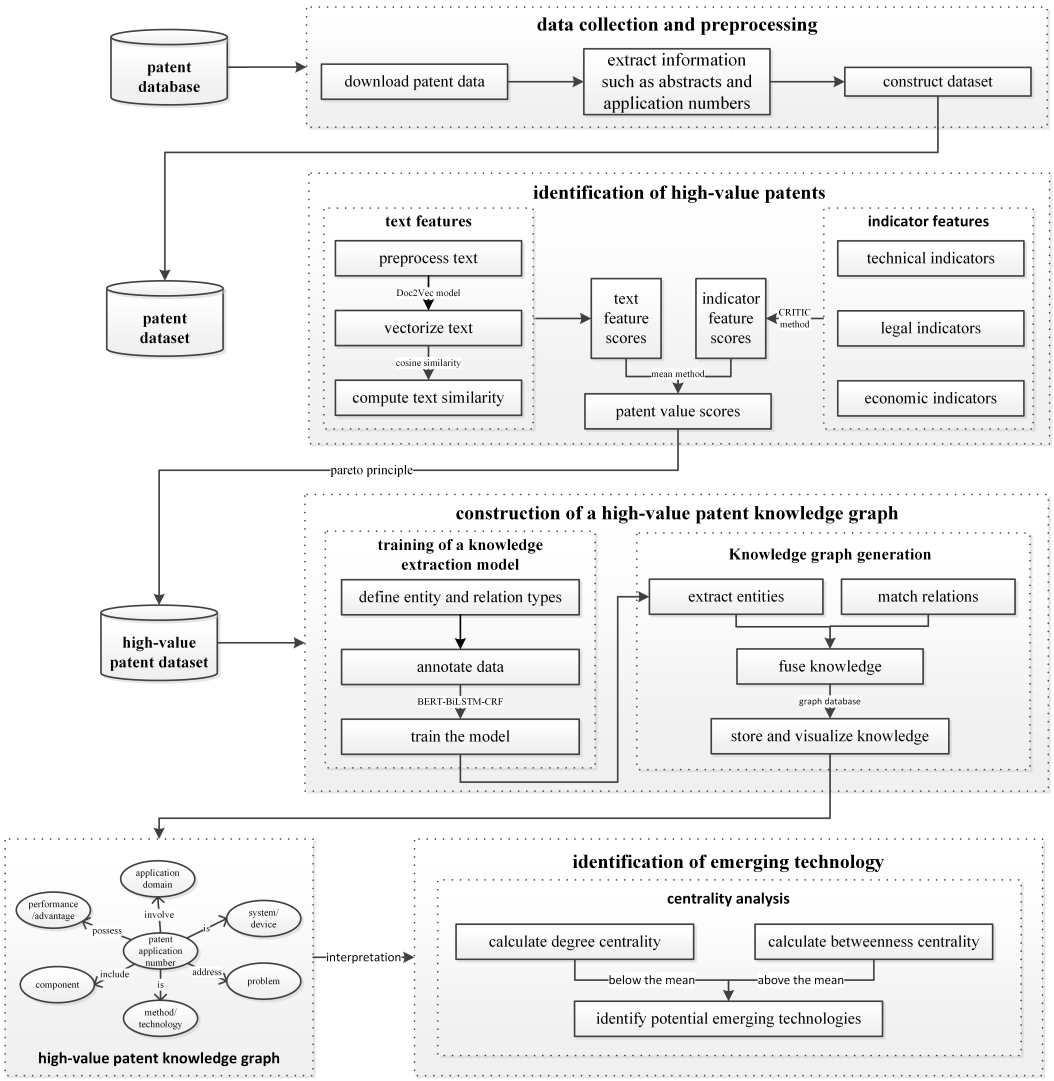


Figure 1. model framework.

3.1. Multi-Feature-Based High-Value Patent Identification

3.1.1. Text-Based Identification

The Doc2Vec model is a distributed document representation method that can directly transform sentences or paragraphs into fixed-dimensional vectors. It captures contextual semantics and situational information at the levels of words, sentences, and even entire paragraphs. In recent years, some scholars have explored the application of the Doc2Vec

model in patent text analysis, such as identifying technological topics [26–28] and analyzing patent similarity [29–31]. In this study, the Doc2Vec model is employed to vectorize patent abstract texts, after which cosine similarity is used to calculate the similarity between patent abstracts:

$$\mathrm{Sim}(A, B_i) = \cos(\theta_i) = \frac{A \cdot B_i}{\|A\|\|B_i\|} \quad (1)$$

where $\mathrm{Sim}(A, B_i)$ represents the semantic similarity between patent text A and the i-th patent text B_i; A and B_i denote the corresponding semantic vectors of the two patent texts being compared; and θ_i is the angle between these two semantic vectors.

The value of a patent can be measured by its novelty [32]. Zhang and Tsai [33] argued that the novelty of a paper is closely related to its most similar paper; the higher the similarity between the two papers, the lower the novelty. This study adopts their proposed method for calculating a novelty index. If a patent is highly similar in textual content to another patent under evaluation, it indicates that the evaluated patent has relatively weak novelty, which may affect its market competitiveness and technological leadership. The maximum similarity can intuitively reflect potential technological overlap or content redundancy and has strong interpretability. Therefore, this study adopts an indicator based on the maximum cosine similarity of patent texts to quantify the value of patent text features:

$$S_{\text{text}} = 1 - \max_{1 \le i \le n} \mathrm{Sim}(A, B_i) \quad (2)$$

where S_{text} denotes the text feature score of the patent, and B_i represents the i-th patent to be compared.

3.1.2. Indicator-Based Identification

The selection of patent value evaluation indicators is a critical step in the assessment model, as their rationality directly affects the accuracy and reliability of the evaluation. Since 1 September 2023, the Recommended National Standard "Guidelines for Patent Evaluation", formulated by the China National Intellectual Property Administration and relevant organizations, has been implemented, providing an extensible and operational set of indicators for patent value analysis. Drawing on the indicators outlined in the Guidelines and the framework proposed in reference [34], which encompasses the dimensions of legal value, technical value, and economic value, this study constructs a patent value evaluation indicator system, as shown in Table 1.

Table 1. Patent value evaluation indicators.

Indicator Dimensions	Database Quantifiable Indicators
Legal value	Number of patent families Number of independent claims Number of claims
Technical value	Number of backward citations Number of IPC classes Number of specification pages Patent lifespan
Economic value	Number of transfers Number of inventors Number of forward citations Number of pledges

The weights of the indicators were determined using the CRITIC method, and the indicator data were subsequently normalized. The normalized values were then multiplied by their corresponding weights, and the sum of these weighted values was calculated to obtain the final indicator feature score:

$$S_{ind} = \sum_{j=1}^{m} w_j \cdot x_{ij} \tag{3}$$

where S_{ind} is the indicator feature score of the patent, x_{ij} denotes the normalized value of the j-th evaluation indicator for the i-th patent, w_j represents the weight of the j-th indicator, and m is the total number of indicators.

3.1.3. Multi-Dimensional Fusion Identification

After calculating scores from both text features and indicator features, this study employs the arithmetic mean to integrate the two types of feature scores, resulting in the final patent value score:

$$S = \frac{S_{text} + S_{ind}}{2} \tag{4}$$

In accordance with the Pareto principle [35], patents ranking within the top 20% are considered to possess high value.

3.2. High-Value Patent Knowledge Graph

The patent abstract serves as the core of a patent document. Reference [36] summarized that patent abstracts primarily provide a concise description of technical features, including the technical structure, application domains, performance characteristics, and constituent components. Accordingly, this study focuses on extracting four types of knowledge from patent abstracts: technical structure knowledge, application domain knowledge, performance knowledge, and component knowledge. Based on these four categories, seven types of patent entities are defined: patent application number, system/device, method/technology, application domain, performance/advantage, problem, and component. The relationships among these patent entities are categorized as is, involves, possesses, solves, and includes. Building on this framework, this study constructs a conceptual patent knowledge graph from the entities and relationships embedded in patent abstracts, as shown in Figure 2.

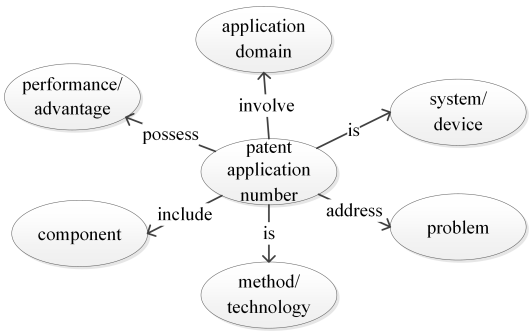


Figure 2. Conceptual diagram of the patent knowledge graph.

Moreover, the most critical step in constructing a knowledge graph is knowledge extraction. Knowledge extraction from unstructured text typically involves entity recognition and relation extraction. Patent abstract texts are highly specialized, and entity names often exhibit semantic similarity. Therefore, this study employs a BERT-BiLSTM-CRF model for the entity recognition task [37]. After selecting the entity recognition model, a subset of the

corpus is manually annotated for training, and the trained model is subsequently applied to extract entities from unannotated texts. For relation extraction, given that the entity types and relation types in this study are predefined and the relation patterns are relatively fixed, a rule-based approach is adopted to classify the relationships between entity pairs.

After the completion of knowledge extraction, knowledge fusion is conducted. Finally, using Python packages such as pandas and py2neo, all entity–relation data are imported into the Neo4j graph database for storage, thereby completing the construction of the high-value patent knowledge graph.

3.3. Network-Based Emerging Technology Identification

Social network analysis can intuitively reveal knowledge units or knowledge clusters, as well as their structural characteristics and evolutionary patterns, and can be employed to track technological development trends and identify frontier hotspots within specific technological domains [38]. When combined with the advantages of knowledge graphs in terms of visual representation and semantic association, this approach enables a more systematic elucidation of the network structure among technological knowledge and its evolutionary features. Within this framework, centrality, as a key metric in network analysis, is widely used to assess the significance and influence of knowledge nodes within the graph. By quantitatively analyzing centrality measures, it becomes possible to identify knowledge nodes that occupy critical hub positions in the technological network, thereby facilitating the discovery of potential emerging technologies [39].

Common centrality measures include degree centrality and betweenness centrality. Degree centrality quantifies the number of immediate neighboring nodes directly connected to a given node. A higher degree centrality indicates that the node is more active within the network, with a greater number of direct connections, and may represent a current research hotspot:

$$D_i = \frac{k_i}{N-1} \tag{5}$$

where k_i denotes the number of edges connecting node i to existing nodes, and $N-1$ denotes the total number of edges connecting node i to all other nodes.

Betweenness centrality measures the position of a node along the shortest paths between other nodes, reflecting its role as a bridge in the flow of information. Nodes with high betweenness centrality may serve as critical intermediaries within the network, controlling the flow of information and knowledge:

$$B_i = \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}} \tag{6}$$

where n_{st}^i denotes the number of shortest paths between nodes s and t that pass through node i, and g_{st} denotes the total number of shortest paths connecting nodes s and t.

In the identification of emerging technologies, these two metrics offer complementary perspectives. Nodes with high degree centrality typically represent current research hotspots that attract widespread attention. In contrast, nodes with high betweenness centrality but low degree centrality may indicate potential emerging technologies [40]. Such technologies have not yet received broad attention but play a critical bridging role within the knowledge network, potentially exerting significant influence in the future. To establish a quantitative criterion, the mean value of each metric is utilized as the threshold. Nodes scoring above the mean are classified as high centrality, while those below are designated as low centrality, thereby enabling the systematic identification of potential emerging technology nodes.

4. Experiment

4.1. Data Collection and Preprocessing

As one of the representative emerging industries, the new energy vehicle (NEV) sector has experienced rapid development under the strong support of relevant policies and the acceleration of technological advancement. Its market scale continues to expand, and technological pathways are constantly evolving, making it a key domain in global technological competition and industrial upgrading. Therefore, conducting research on emerging technology identification within the NEV sector is highly representative and practically significant, as it helps validate the scientific rigor and practical applicability of the identification methodology.

Using the Innojoy patent database as the data source, and following consultations with domain experts, the specific search formula was determined as: (TA = "New Energy Vehicle") OR (TA = "Hybrid Vehicle") OR (TA = "Electric Vehicle"). The retrieval period was set from January 2022 to June 2025, with the search conducted in June 2025. The patent types included Chinese invention patents and utility model patents. After removing duplicate records and patents not relevant to the NEV sector, a total of 17,497 valid patents were obtained.

4.2. High-Value Patent Identification (NEV)

4.2.1. Text Scoring

During the text vector training process, Doc2Vec can be implemented in two modes: DM and DBOW. In the present study, considering factors such as storage efficiency and the relatively short length of the abstract texts after noise reduction, the DBOW mode was selected for training. The experiments were conducted in a Python 3.8 environment using Gensim 4.3.3, with the model parameters summarized in Table 2.

Table 2. Doc2Vec parameter settings.

Parameter	Description	Value
Dm	Model training algorithm: dm = 1 for Distributed Memory (DM) mode; dm = 0 for Distributed Bag-of-Words (DBOW) mode	0
Vector_size	Dimensionality of the feature vectors	300
Window	Maximum distance between the predicted word and context words	5
Min_count	Ignores all words with total frequency lower than this threshold	2
Workers	Number of parallel threads used during training	4
Epochs	Number of training iterations	20

The trained Doc2Vec model was employed to convert textual documents into vector representations, thereby achieving text vectorization, computing similarities between abstracts, and transforming them using the textual feature scoring formula. Due to space limitations, only the top five patent text feature scores are presented, as shown in Table 3.

Table 3. Text feature scores.

Rank	Patent Application Number	Textual Feature Score
1	CN202211544232.4	0.865934565663338
2	CN202211476005.2	0.864687204360962
3	CN202210153575.1	0.864127963781357
4	CN202323535100.8	0.859420135617256
5	CN202211512669.X	0.85793000459671

4.2.2. Indicator Scoring

Value-related indicators across the legal, technological, and economic dimensions can be directly obtained from the Innojoy database. A total of 11 indicators were selected to evaluate the patent indicator features. The CRITIC method was employed to calculate objective weights, yielding the weight coefficients of each indicator within the evaluation system shown in Table 4 and thereby obtaining the patent indicator feature scores.

Table 4. CRITIC method weight calculations.

Dimension	Indicator	Weight Coefficient (%)
Legal	Number of patent families	2.42
	Number of independent claims	1.50
	Number of claims	10.42
Technical	Number of backward citations	34.27
	Number of IPC classes	2.44
	Number of specification pages	27.72
	Patent lifespan	2.05
Economic	Number of transfers	0.34
	Number of inventors	13.70
	Number of forward citations	5.04
	Number of pledges	0.10

The results indicate that the weights of different dimensions in patent indicator value assessment differ significantly. The technical dimension dominates, with number of backward citations and number of specification pages as core indicators, reflecting the patents' innovativeness and technological depth; the legal dimension is represented primarily by the number of claims, reflecting the breadth of protection; and the economic dimension is mainly characterized by the number of inventors, suggesting that team size or collaboration intensity is associated with commercialization potential.

To further verify the robustness of the model, this study conducted a sensitivity analysis of the measurement results from two dimensions: perturbations in indicator data and perturbations in indicator weights. Regarding the perturbation of indicator data, while keeping the indicator weights unchanged, each individual indicator was perturbed by ±5%, ±10%, and ±15% to assess the changes in individual patent values and their rankings. The results show that the Pearson correlation coefficients of patent values and the Spearman correlation coefficients of rankings before and after perturbation were all above 0.99, with the maximum change rate of patent value reaching 14.02% and the maximum change rate in ranking reaching 14.24%, as shown in Figure 3. Regarding the perturbation of indicator weights, while keeping the indicator data unchanged, the weight of each indicator was perturbed by ±5%, ±10%, and ±15%. The analysis indicates that the Pearson and Spearman correlation coefficients remained above 0.99, with the maximum change rate of patent value at 11.49% and the maximum change rate in ranking at 14.24%, as shown in Figure 4. Overall, whether perturbing the indicator data or the indicator weights, the calculated patent values and their rankings exhibit high robustness.

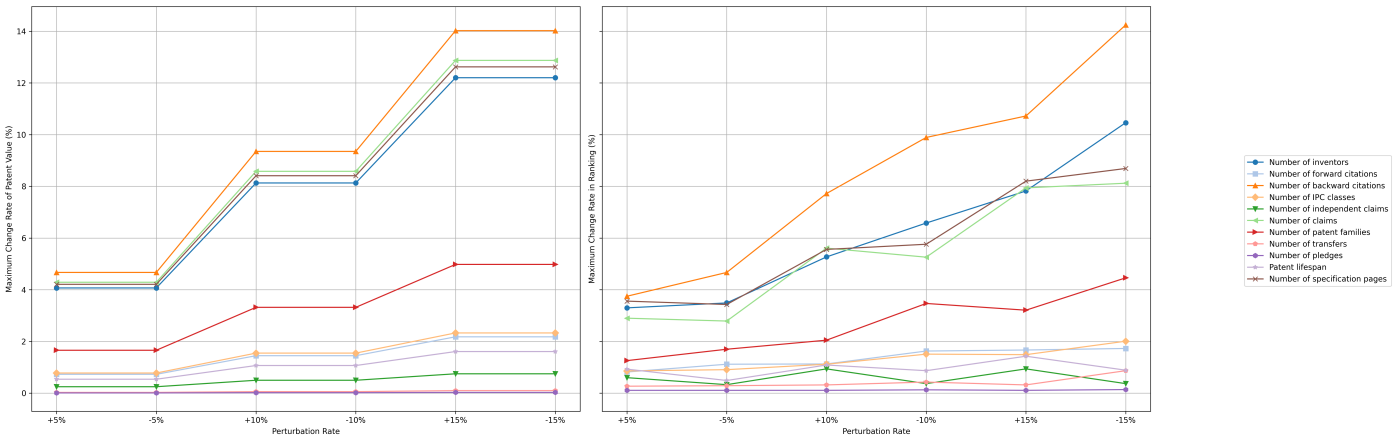


Figure 3. perturbations in indicator data.

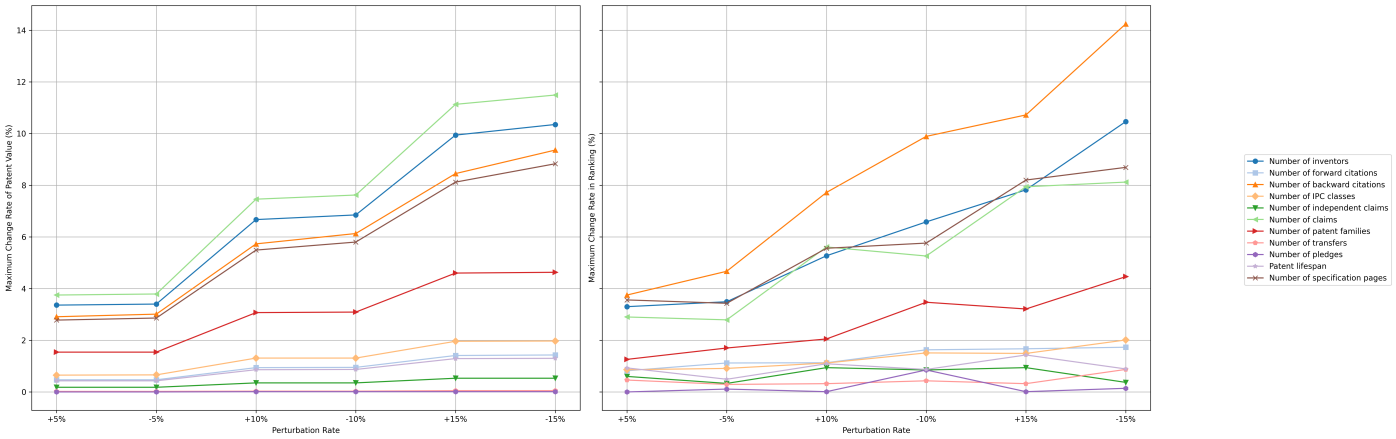


Figure 4. perturbations in indicator weight.

4.2.3. Patent Selection

After calculating the scores based on the text features and the indicator features of the patents separately, the comprehensive patent value score was obtained using the averaging method, with a portion of the results presented in Table 5.

Table 5. Comprehensive score of patent value.

Application Number	Comprehensive Score
CN202310444867.5	0.68
CN202311195503.4	0.66
CN202210717415.5	0.64
CN202311562216.2	0.63
CN202310506870.5	0.62

The patents were ranked in descending order according to their comprehensive scores, and, following the Pareto principle, a total of 3500 high-value patents were selected.

4.3. NEV Patent Knowledge Graph

4.3.1. Entities and Relations

This study defines the types of entities and relations to be extracted by considering the technical terms and semantic relationships present in patent abstracts within the field of new energy vehicles, resulting in a total of seven entity categories and six relation types, as shown in Table 6.

Table 6. The main entities and types of relationships covered by patent abstracts.

Subject	Predicate	Object	Example
patent application number	is	system/device	Discloses a coating device for the manufacturing of complete new energy vehicles.
patent application number	involve	application domain	The invention involves the field of new energy vehicle processing and manufacturing technology.
patent application number	is	method/technology	The invention provides a control method for autonomous vehicles.
patent application number	include	component	The charging device includes a wireless charging module.
patent application number	possess	performance/advantage	Exhibits excellent mechanical strength and impact resistance.
patent application number	address	problem	Addresses the issue of battery performance degradation under low-temperature conditions.

4.3.2. Annotation and Extraction

After identifying 3500 novel high-value patents, 1000 abstracts were selected for manual annotation and verified by domain experts. The annotation corpus employed the "BIO + Named Entity" scheme, where B indicates the beginning of an entity segment, I represents the interior of an entity segment, and O denotes a character that does not belong to any entity. Finally, the manually annotated data were stored in JSON format.

The BERT-BiLSTM-CRF entity extraction model was implemented using the PyTorch deep learning framework and Python 3.8. The parameter settings of the experimental model are presented in Table 7. During training, the annotated data were split into training and validation sets with a ratio of 8:2.

Table 7. BERT-BiLSTM-CRF model parameters.

Parameter	Description	Value
epochs	Number of training epochs	100
max_seq_length	Maximum sequence length	256
batch_size	Number of samples per training batch	64
lr	Learning rate	0.0003
weight_decay	Weight decay rate	0.01
scheduler	Learning rate scheduling strategy	warmup_linear

The model achieved a precision of 0.77, a recall of 0.72, and an F1-score of 0.74 during training. Finally, the trained model was employed to extract entities from unannotated sentences, and a rule-based approach was subsequently used to classify the relationships between entities according to their respective categories, resulting in the generation of high-value patent triplet data.

4.3.3. Integration and Storage

After extracting entities using the BERT-BiLSTM-CRF deep learning model, knowledge integration is required. Knowledge integration refers to the alignment of data to solve the problem where different entity names represent the same entity [41]. For example, "technical field of new energy vehicle charging" and "new energy vehicle charging technical field" both denote the same entity. In this study, the Jaccard similarity coefficient between entities was calculated to determine whether merging should be performed, thereby avoiding the issue of duplicate nodes representing the same attributes in the construction of the knowledge graph.

Currently, knowledge graphs are primarily stored using RDF-based triple stores or graph databases [42]. The former excels at managing and publishing large-scale RDF

data; however, the presence of extensive redundant descriptions among data can lead to inefficiencies. In contrast, Neo4j, as the most widely used graph database, employs an optimized storage mechanism specifically designed for graph-structured data, enabling both efficient storage and significantly enhanced query performance. Moreover, Neo4j provides a more intuitive visualization of entities and their relational structures within a knowledge graph. Therefore, in this study, the high-value patent data were stored using the Neo4j graph database. Based on the entity and relation types defined in Table 6, a high-value patent knowledge graph was ultimately constructed, comprising 23,390 entity nodes and 25,611 relationships.

4.4. Emerging Technology Identification (NEV)

A knowledge graph is represented in the form of knowledge triples consisting of nodes and edges. Nodes, representing entities, include information such as patent application number, system/device, application domain, method/technology, component, performance/advantage, and problem. The edges, as a representation of relational structure, delineates the diverse associations among nodes, including is, involve, include, possess, and address. After constructing a high-value patent knowledge graph, social network analysis can be applied using node-level metrics as the analytical entry point. This approach enables the examination of relationships between nodes and external entities, thereby assessing the influence of nodes within the network.

The Neo4j Community Edition version 5.26.0 was used, with the GDS plugin version 2.13.2. Both betweenness centrality and degree centrality algorithms were executed, and centrality values were computed for all nodes. Here, system/device and method/technology entities with betweenness centrality above the mean and degree centrality below the mean are selected for presentation, as shown in Table 8.

Table 8. Centrality analysis results.

Group	Node Name
charging technologies	charging method, wireless charging method, dynamic wireless charging system, charge–discharge scheduling method
battery management technologies	battery assembly, battery cooling system, battery pack protective structure, battery thermal management method, power battery pack heating method
power systems	range extender, powertrain testing system
vehicle Control	vehicle control method, torque distribution control method, autonomous vehicle control method
manufacturing equipment and inspection tooling	conductive component, filtration device, stamping die, snap-fit structure, rigid component testing fixture
battery swapping facilities	vehicle-mounted platform and associated battery swap station

In subsequent analyses, conditional queries were performed using Neo4j's query language. For example: MATCH (a)-[rel*1..2]-(b) WITH ['wireless charging method', 'charging method', 'dynamic wireless charging system', 'charge–discharge scheduling method'] AS nodeNames, a, rel, b WHERE a.name IN nodeNames RETURN a, rel, b. This approach enables the retrieval of the high-value patent knowledge graph and, by leveraging its rich semantic information, facilitates the analysis of potential technological opportunities, thereby enhancing the interpretability of emerging technology identification.

4.5. Result Evaluation

Taking the nodes "wireless charging methods," "charging methods," "dynamic wireless charging systems," and "charge–discharge scheduling methods" as examples, the

conditional query results are shown in Figure 5. The nodes presented in the figure not only cover these core technologies but also include related auxiliary technologies, such as "neural network model," "joint planning methods for electric vehicle fast-charging stations and dynamic wireless charging systems," and "load curve." This suggests that the future development of charging technologies for new energy vehicles is trending toward higher efficiency and wireless solutions. Specifically, wireless charging technology, combined with neural network models, optimizes charging control to enhance charging efficiency and reduce voltage and current fluctuations, achieving smoother energy transfer. Dynamic wireless charging systems, coordinated with fast-charging station planning, optimize the layout of charging infrastructure, improving charging flexibility and energy utilization. Intelligent charge–discharge scheduling methods, based on grid load data and vehicle status, guide vehicles to charge during off-peak periods, achieving peak shaving and valley filling, and thereby stabilizing grid operation.

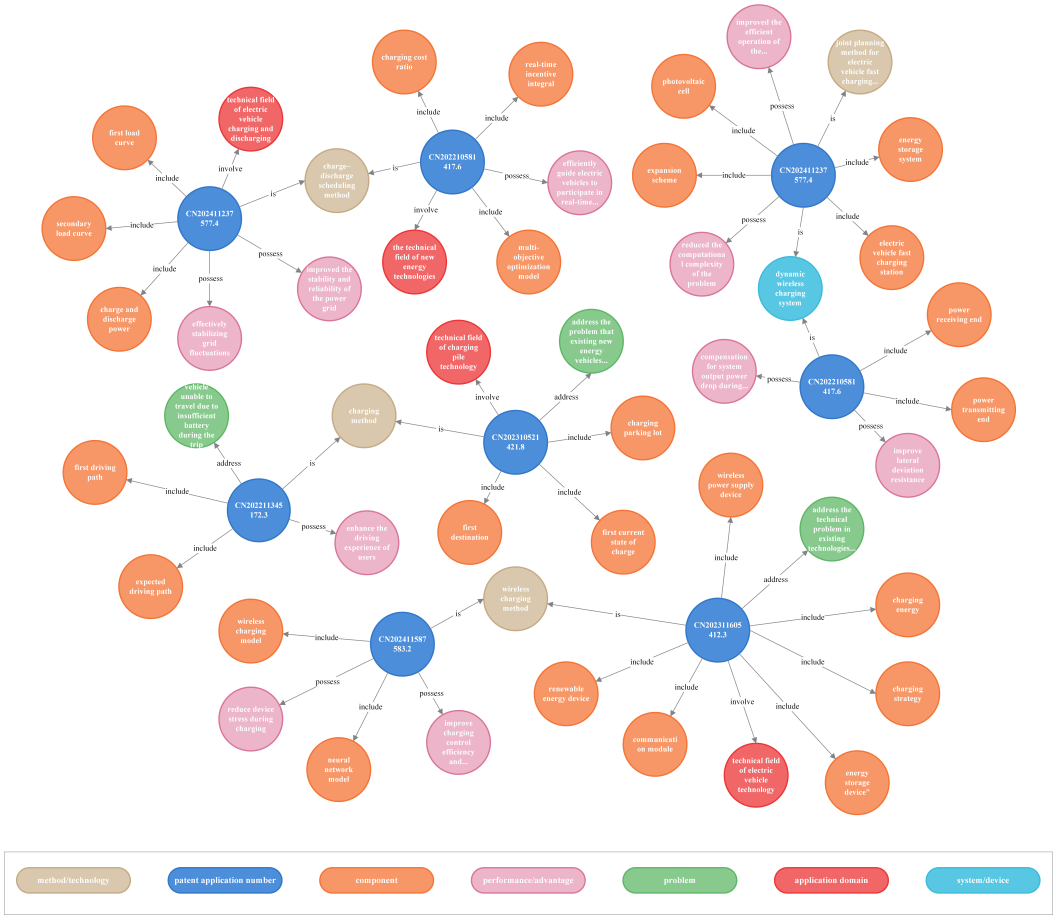


Figure 5. Illustration of search results on charging technology.

To validate the effectiveness of the above results, this study retrieved authoritative national-level policy documents and industry reports to provide objective and reliable references for verifying the identified emerging technologies, as show in Table 9, thereby minimizing subjective bias.

Table 9. Validation of emerging technology identification results.

Emerging Technology	Policy Documents and Industry Reports
Future charging technologies for new energy vehicles are trending toward higher efficiency and wireless operation. Wireless charging with neural networks optimizes control, boosting efficiency and stabilizing energy transfer. Dynamic wireless charging systems and fast-charge station co-planning enhance infrastructure flexibility. Smart charging scheduling uses grid/vehicle data to guide off-peak charging, stabilizing the grid via load shifting.	The "New Energy Vehicle Industry Development Plan (2021–2035)" issued by the State Council emphasizes the advancement of emerging charging technologies, including intelligent and orderly charging, high-power charging, and wireless charging, aiming to enhance charging convenience and product reliability.
Future battery management technologies for new energy vehicles aim to enhance battery safety, stability, and operational efficiency. Power battery heating uses optimized methods for better efficiency and low-temperature performance. Thermal management employs intelligent prediction and dynamic regulation for efficient cooling and temperature control, extending range and lifespan. The battery pack structure integrates optimized supports, cushioning, and enclosures to enhance crash safety.	The "New Energy Vehicle Industry Development Plan (2021–2035)" issued by the State Council emphasizes the need to intensify breakthroughs in critical technologies for power batteries and fuel cell systems, particularly those related to high strength, lightweight design, enhanced safety, reduced cost, and extended service life.
New energy vehicles are expected to evolve toward intelligent mobility in the future. For torque distribution, intelligent optimization algorithms balance economy, power, and maneuverability for precise control, boosting overall vehicle performance and stability. Autonomous driving control utilizes deep learning predictive models integrating perception and data for accurate decision-making and path planning, enhancing safety and intelligence.	The "Notice on Launching Pilot Projects for the Integrated Application of Vehicle–Road–Cloud Collaboration" issued jointly by five ministries proposes promoting multi-scenario applications of autonomous driving technologies through efficient coordination among vehicles, roads, networks, clouds, and maps, thereby accelerating technological breakthroughs and industrialization in the intelligent connected vehicle sector.
The power systems of future new energy vehicles are evolving toward higher levels of stability. Range extender technology focuses on real-time electric balance control with multi-cycle energy management and high integration for low NVH, improving system efficiency, stability, and user experience. Powertrain testing systems prioritize structural cushioning, dust removal, and multi-node sensing to enhance testing accuracy and equipment reliability.	The "Outline of the 14th Five-Year Plan for National Economic and Social Development and the Long-Range Objectives Through the Year 2035" highlights the need to achieve breakthroughs in key technologies such as high-performance power systems for new energy vehicles.
Future manufacturing equipment and inspection tooling are evolving toward higher precision and greater integration. Conductive connectors feature materials/structures for high reliability and heat dissipation. Card-in designs use profiled interlocking for better aging resistance and blind-mating ease. Filters offer high-efficiency, self-cleaning, and multi-scenario adaptability. Stamping dies move toward automation and efficient scrap clearing. Rigidity testing fixtures use integrated structures for high-precision detection.	In the "Guidelines for Equipment Renewal and Technological Transformation in Key Industrial Sectors" issued by the Ministry of Industry and Information Technology, it is stated that priority should be given to equipment used in the production of new energy vehicles and their components, focusing on the four major vehicle manufacturing processes—stamping, welding, painting, and final assembly—as well as component production.
Battery swapping facilities are gradually evolving toward greater intelligence, efficiency, and safety. Optimizing battery swap facilities' design and mechanisms enhances space utilization and convenience. These advancements also boost automation and strengthen safety protection for non-swapping traffic and personnel.	In the "Implementation Opinions on Further Enhancing the Service and Support Capacity of Charging and Battery-Swapping Infrastructure" issued by the National Development and Reform Commission and other departments, it is proposed to deploy battery-swapping stations in accordance with local conditions and to strengthen the service capacity of public charging facilities. The document also requires charging facility owners and residential community management units to strengthen safety management of charging infrastructure and to promptly detect and eliminate potential safety hazards.

5. Discussions

This study develops an emerging technology identification model following the logical process of "data collection and preprocessing— identification of high-value patents—

construction of a high-value patent knowledge graph—identification of emerging technology." During the research, a standardized workflow from high-value patent screening to potential emerging technology recognition was established by leveraging Doc2Vec for text modeling, the CRITIC method, the BERT-BiLSTM-CRF entity extraction model, and centrality calculations. This approach provides both theoretical support and practical guidance for enterprises in formulating technology innovation strategies, as well as for national-level planning of scientific and technological development and the formulation of industrial policies.

Theoretically, this study advances the understanding and methodology of technology identification. A significant achievement is the successful development and application of a comprehensive high-value patent screening system. This system is robust because it integrates both text features and indicator features, moving beyond the reliance on single metrics to provide a more holistic and authoritative measure of patent value, which is crucial for building a foundation for subsequent analysis. The most notable theoretical innovation is the proposed methodology that leverages high-value patents to construct a knowledge graph followed by social network analysis for identifying emerging technologies. This approach achieves semantic reconstruction of patent data, elevating the representation from a mere lexical level to a conceptual level. By utilizing the rich semantic relationships captured within the knowledge graph, the model significantly enhances the semantic correlations and logical connections between technical concepts. This directly mitigates the weak interpretability inherent in many existing black-box or purely statistical methods. Specifically, the mechanism of identifying nodes with low degree centrality but high betweenness centrality as potential technological opportunities is theoretically sound, as it spots niche but highly connective concepts that are often indicative of emergent areas.

In practical terms, the established workflow offers immediate and valuable utility for various stakeholders. For enterprises, the model provides a structured and systematic tool for monitoring their competitive technological landscape. By identifying high-value patents, the system helps R&D departments focus resources on the most promising areas. The final recognition of potential emerging technologies allows companies to proactively adjust their technology roadmap, leading to informed decisions regarding investment and long-term strategic planning. Furthermore, this study took the new energy vehicle sector as its empirical object, and the results show that potential emerging technologies in this field are mainly concentrated in charging technologies, battery management technologies, vehicle control, power systems, manufacturing equipment and inspection tooling, and battery swapping facilities. All the above research findings are supported by existing policies. The findings offer direct, evidence-based guidance for national-level planning of scientific and technological development. Government agencies can utilize the identified clusters to formulate targeted industrial policies, providing significant reference value for technological innovation in this field.

Certainly, the emerging technology identification model proposed in this study has several limitations. Regarding the comprehensiveness of data sources and the control of potential biases, this study relies primarily on patent data and does not incorporate non-patent sources such as academic publications, industry reports, or policy documents. Consequently, the model may lag in identifying frontier technologies that are still at the basic research stage and have not yet been reflected in patent portfolios. Moreover, the distribution of patent data across institutions is uneven, with large enterprises often dominating due to their financial resources and intellectual property strategies, potentially reducing the model's sensitivity to innovations from smaller companies or research institutions. In addition, as the sample data are predominantly derived from Chinese patents, the study may be subject to geographic and language biases, limiting its ability to fully

capture global technological dynamics. With respect to the generalizability of the research framework, this study focuses solely on the field of new energy vehicles, and the selection of the domain is not sufficiently comprehensive. Future work should integrate multi-source data, mitigate biases, and broaden the methodology across technological fields to enhance the accuracy and generalizability of emerging technology identification.

6. Conclusions

Technology identification is crucial for maintaining a leading position in a rapidly evolving technological environment. This study constructs an emerging technology identification framework that identifies potential emerging technologies through the systematic analysis of patents. The framework achieves methodological innovation by integrating textual features and indicator-based features at the patent screening stage, and by constructing a knowledge graph based on high-value patents, further incorporating social network analysis methods to identify emerging technology nodes. To ensure the feasibility and stability of the framework in practical applications, this study adopts a modular, step-by-step analytical strategy, thereby efficiently supporting the in-depth processing of large-scale patent data. At the same time, through a high-value patent screening mechanism, the computational complexity of knowledge graph construction and network analysis is effectively reduced, improving overall computational efficiency. An empirical study in the field of new energy vehicles shows that the emerging technologies identified by this framework are highly consistent with current policy orientations and industrial development practices, verifying the effectiveness and practicality of the proposed method. This research contributes both theoretically, advancing technology identification methodology, and practically, offering valuable guidance for enterprises and national planning.

Author Contributions: C.Z. conceived the research concept. Y.Z. supported model development and manuscript refinement. Y.H. was responsible for writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Authors declare no conflicts of interest.

References

1. Huang, Z.; Li, J.; Yue, H. Study on comprehensive evaluation based on AHP-MADM model for patent value of balanced vehicle. *Axioms* **2022**, *11*, 481. [CrossRef]
2. He, C.; Shi, F.; Tan, R. Evaluation and cultivation method of high-tech value patents for mechanical products. *PLoS ONE* **2024**, *19*, e0298144.
3. Wan, D.; Peng, L.; Zhan, H. Patent Analysis of the Electric Vehicle Battery Management Systems Based on the AHP and Entropy Weight Method. *World Electr. Veh. J.* **2025**, *16*, 218. [CrossRef]
4. Lee, J.; Park, S.; Lee, J. Study on the Technology Trend Screening Framework Using Unsupervised Learning. *Appl. Sci.* **2022**, *12*, 8920. [CrossRef]
5. Deng, N.; Wang, Y.; Yang, Y.; Chen, X. Patent Novelty Evaluation Based on Sentence BERT and Isolation Forest Algorithm. *J. Intell.* **2025**, *44*, 174–182.
6. Song, K.; Ran, C. Research on Technology Opportunity Identification Based on Topic Mining and Patent Evaluation:A Case Study of Smart Agriculture. *Libr. Inf. Serv.* **2023**, *67*, 61–71. [CrossRef]
7. Liu, W.; Li, S.; Cao, Y.; Wang, Y. Multi-task learning based high-value patent and standard-essential patent identification model. *Inf. Process. Manag.* **2023**, *60*, 103327. [CrossRef]
8. Li, G.; Siddharth, L.; Luo, J. Embedding knowledge graph of patent metadata to measure knowledge proximity. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 476–490. [CrossRef]

9. Chen, H.; Deng, W. Interpretable patent recommendation with knowledge graph and deep learning. *Sci. Rep.* **2023**, *13*, 2586. [CrossRef]
10. Siddharth, L.; Blessing, L.T.; Wood, K.L.; Luo, J. Engineering knowledge graph from patent database. *J. Comput. Inf. Sci. Eng.* **2022**, *22*, 021008. [CrossRef]
11. Zuo, H.; Yin, Y.; Childs, P. Patent-kg: Patent knowledge graph extraction for engineering design. *Proc. Des. Soc.* **2022**, *2*, 821–830. [CrossRef]
12. Zhang, L.; Sun, X.; Ma, X.; Hu, K. A New Entity Relationship Extraction Method for Semi-Structured Patent Documents. *Electronics* **2024**, *13*, 3144. [CrossRef]
13. Puccetti, G.; Chiarello, F.; Fantoni, G. A simple and fast method for Named Entity context extraction from patents. *Expert Syst. Appl.* **2021**, *184*, 115570. [CrossRef]
14. Day, G.S.; Reibstein, D.J. *Wharton on Dynamic Competitive Strategy*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
15. Cozzens, S.; Gatchair, S.; Kang, J.; Kim, K.S.; Lee, H.J.; Ordóñez, G.; Porter, A. Emerging technologies: Quantitative identification and measurement. *Technol. Anal. Strateg. Manag.* **2010**, *22*, 361–376. [CrossRef]
16. Shen, Y.C.; Chang, S.H.; Lin, G.T.; Yu, H.C. A hybrid selection model for emerging technology. *Technol. Forecast. Soc. Change* **2010**, *77*, 151–166. [CrossRef]
17. Hussain, M.; Tapinos, E.; Knight, L. Scenario-driven roadmapping for technology foresight. *Technol. Forecast. Soc. Change* **2017**, *124*, 160–177. [CrossRef]
18. Lee, S.; Kim, W.; Kim, Y.M.; Lee, H.Y.; Oh, K.J. The prioritization and verification of IT emerging technologies using an analytic hierarchy process and cluster analysis. *Technol. Forecast. Soc. Change* **2014**, *87*, 292–304. [CrossRef]
19. Hu, Z.; Wang, M.; Han, Y. Multidimensional Indicator Identification and Evolution Analysis of Emerging Technology Topics Based on LDA2Vec-BERT—A Case Study of Blockchain Technology in the Field of Disruptive Technology. *J. Mod. Inf.* **2024**, *44*, 42–58.
20. Yun, S.; Cho, W.; Kim, C.; Lee, S. Technological trend mining: Identifying new technology opportunities using patent semantic analysis. *Inf. Process. Manag.* **2022**, *59*, 102993. [CrossRef]
21. Ning, B.; Zhang, W.; Ma, J.; Yuan, H. Identification and Forsight of Emerging Technologies Based on KeyGraph-CPM: A Case of the Bio-pharmaceutical Industry. *Inf. Stud. Appl.* **2021**, *44*, 146–152. [CrossRef]
22. Li, H.; Liang, H.; Hu, Y.; Liu, X. Technology convergence prediction based on temporal heterogeneous graph neural networks. *Inf. Process. Manag.* **2025**, *62*, 104034. [CrossRef]
23. Zhou, Y.; Dong, F.; Liu, Y.; Ran, L. A deep learning framework to early identify emerging technologies in large-scale outlier patents: An empirical study of CNC machine tool. *Scientometrics* **2021**, *126*, 969–994. [CrossRef]
24. Lee, M.; Kim, S.; Kim, H.; Lee, J. Technology opportunity discovery using deep learning-based text mining and a knowledge graph. *Technol. Forecast. Soc. Change* **2022**, *180*. [CrossRef]
25. Wei, M.; Zheng, R.; Gao, Z.; Wang, X. Research on Industry Emerging Technology Forecast Modeling Based on Knowledge Graph and Deep Neural Networks. *J. China Soc. Sci. Tech. Inf.* **2022**, *41*, 1134–1148.
26. Gao, Q.; Jiang, M. Exploring technology fusion by combining latent Dirichlet allocation with Doc2vec: A case of digital medicine and machine learning. *Scientometrics* **2024**, *129*, 4043–4070. [CrossRef]
27. Jin, Q.; Chen, H.; Zhang, Y.; Wang, X.; Zhu, D. Unraveling scientific evolutionary paths: An embedding-based topic analysis. *IEEE Trans. Eng. Manag.* **2023**, *71*, 8964–8978. [CrossRef]
28. Monnet, N.; Maréchal, L.; Jang-Jaccard, J.; Mermoud, A. Clustering doc2vec output for topic-dimensionality reduction: A MITRE ATT&CK calibration. *arXiv* **2024**, arXiv:2410.11573.
29. Yoon, B.; Kim, S.; Kim, S.; Seol, H. Doc2vec-based link prediction approach using SAO structures: Application to patent network. *Scientometrics* **2022**, *127*, 5385–5414. [CrossRef]
30. Ascione, G.S.; Sterzi, V. A comparative analysis of embedding models for patent similarity. *arXiv* **2024**, arXiv:2403.16630. [CrossRef]
31. Wang, Z.; Guo, W.; Shao, H.; Wang, L.; Chang, Z.; Zhang, Y.; Liu, Z. From technology opportunities to solutions generation via patent analysis: Application of machine learning-based link prediction. *Adv. Eng. Inform.* **2024**, *62*, 102944. [CrossRef]
32. Kalıp, N.G.; Erzurumlu, Y.Ö.; Gün, N.A. Qualitative and quantitative patent valuation methods: A systematic literature review. *World Pat. Inf.* **2022**, *69*, 102111. [CrossRef]
33. Zhang, Y.; Tsai, F.S. Chinese novelty mining. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 1561–1570.
34. Ran, C.; Li, W.; Huang, W. Identifying Technology Opportunities from High-Value Patents in Universities: The Case of Generative Artificial Intelligence. *J. Inf. Resour. Manag.* **2024**, *14*, 103–116. [CrossRef]
35. Pareto, V. *Cours d'Économie Politique*; Librairie Droz: Carouge, Switzerland, 1964; Volume 1.
36. Cao, R.; Cao, S. Research on the Evolution of High-value Patent Technology Innovation Based on Knowledge Graph. *J. Mod. Inf.* **2024**, *44*, 3–17.

37. Chen, W.; Qiu, P.; Cauteruccio, F. MedNER: A service-oriented framework for Chinese medical named-entity recognition with real-world application. *Big Data Cogn. Comput.* **2024**, *8*, 86. [CrossRef]
38. Jun, S. Patent Keyword Analysis Using Bayesian Factor Analysis and Social Network Visualization in Digital Therapy Technology. *Computers* **2025**, *14*, 78. [CrossRef]
39. Yang, Z.; Zhang, W.; Yuan, F.; Islam, N. Measuring topic network centrality for identifying technology and technological development in online communities. *Technol. Forecast. Soc. Change* **2021**, *167*, 120673. [CrossRef]
40. Li, X.; Xie, Q.; Hong, Z.; Huang, L. Study the Development Trends of Emerging Technologies Based on Socially Aware Analysis: A Case of Perovskite Solar Cells Technology. *Sci. Technol. Prog. Policy* **2018**, *35*, 15–24.
41. Yuan, S.; Lu, Z.; Li, Q.; Gu, J. A multi-modal entity alignment method with inter-modal enhancement. *Big Data Cogn. Comput.* **2023**, *7*, 77. [CrossRef]
42. Tamašauskaitė, G.; Groth, P. Defining a knowledge graph development process through a systematic review. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 27. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.