Research on Intelligent Construction Methods for TRIZ Analysis Matrices

Authors: Xing Sisi, Qian Li, Qian Li

Date: 2022-06-09T00:00:00+00:00

Abstract

[Objective] Based on deep learning and domain knowledge graph technologies, this study investigates an automatic construction method for TRIZ analysis matrices to provide knowledge support for technological innovation. [Method] First, oriented towards domain requirements, the generic "technical elements"are extended and refined into entity and relation categories to complete the schema layer design of the domain knowledge graph; then, based on the BERT pre-trained language model, an intelligent tool for automatically identifying knowledge entities and relations from patent literature is researched and designed to achieve automatic entity and relation extraction; finally, a graph database is utilized to complete the construction of the domain knowledge graph, and based on knowledge queries, the required knowledge entities and relations are retrieved to realize the automatic construction of TRIZ analysis matrices. [Results] Through empirical validation, the BERT-MH+CRF entity recognition model constructed in this paper for patents in the thin-film magnetic head technology domain achieves an F1-score of 84.93%, and the BERT-MH+softmax relation extraction model achieves an F1-score of 63.7%. [Limitations] The application of knowledge reasoning techniques is lacking, unable to reveal potential knowledge associations between entities, and the quality of the constructed TRIZ analysis matrices remains insufficient. [Conclusion] The proposed method can achieve the goal of automatic construction of TRIZ analysis matrices and can provide effective knowledge support for technological innovation.

Full Text

Research on Intelligent Construction Method of the TRIZ Analysis Matrix

Xing Sisi[1],[2], **Qian Li**[1],[2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract

[**Objective**] Based on deep learning and domain knowledge graph technology, this study investigates an automatic construction method for the TRIZ analysis matrix to provide knowledge support for technological innovation. [**Methods**] First, addressing domain requirements, we extend and refine the generic "technical elements" into entity and relationship categories, completing the schema layer design of the domain knowledge graph. Then, based on the BERT pre-trained language model, we design an intelligent tool to automatically identify knowledge entities and relationships from patent documents, achieving automatic entity-relationship extraction. Finally, we construct the domain knowledge graph using a graph database, and automatically query and retrieve required knowledge entities and relationships based on knowledge queries to realize the automatic construction of the TRIZ analysis matrix. [**Results**] Empirical validation shows that the BERT-MH+CRF entity recognition model constructed for thin-film magnetic head technology patents achieves an F1 score of 84.93%, while the BERT-MH+softmax relationship extraction model achieves an F1 score of 63.7%. [**Limitations**] The lack of knowledge reasoning technology application prevents the revelation of potential knowledge associations between entities, and the quality of the constructed TRIZ analysis matrix remains insufficient. [**Conclusions**] The proposed method can achieve the goal of automatic TRIZ analysis matrix construction and provide effective knowledge support for technological innovation.

Keywords: TRIZ Analysis Matrix, Patent Literature, Pre-training Techniques, Entity Recognition, Relation Extraction

1 Introduction

China is undergoing a strategic transformation from a "manufacturing giant" to an "intelligent manufacturing powerhouse," with an innovation-driven development strategy centered on scientific and technological innovation elevated to a national strategy. The innovation demands from both the nation and enterprises continue to grow, primarily manifested in the need for rapid product iteration, intensifying technological convergence, and centralized aggregation of innovation knowledge. Against this backdrop, innovation is no longer merely an idea arising from individual inspiration but increasingly requires scientific methods and evidence for breakthroughs [1]. The TRIZ innovation methodology, through mining and analyzing large-scale patent data, has formed a systematic approach to guide invention and innovation, enabling accurate analysis and identification of core problems while improving innovation efficiency and quality, making it

one of the most effective innovation methods available.

The system interaction analysis matrix represents the core technology for functional analysis within the TRIZ innovation methodology framework. This matrix traditionally relies on manual extraction of component and function knowledge entities and their corresponding relationships from patent literature, followed by matrix construction for systematic functional analysis to guide technological innovation. However, facing rapid global technological progress and the explosive growth of patent applications, the conventional system interaction analysis matrix suffers from insufficient knowledge elements, slow manual construction speed, and high labor costs, resulting in low effectiveness in promoting technological innovation through TRIZ methods and difficulty in meeting the growing innovation demands of the nation and enterprises.

Addressing this context and development needs, this study proposes an intelligent construction method for the TRIZ analysis matrix under the guidance of TRIZ innovation methodology. We extend the knowledge elements covered by the system interaction analysis matrix for specific domains, defining this domain-oriented matrix containing richer knowledge entities and relationships as the TRIZ analysis matrix. Specifically, we employ BERT pre-trained language model technology to automatically identify knowledge entities and relationships from massive patent data, completing domain knowledge graph construction. For specific requirements, we automatically retrieve required knowledge entities and relationships from the domain knowledge graph based on knowledge queries and reasoning, achieving automatic construction of the TRIZ analysis matrix. This approach addresses both the knowledge supply issue of TRIZ innovation methods and provides support for the continuous development and improvement of TRIZ methodology.

2 Research Status

This study aims to achieve intelligent construction of the TRIZ analysis matrix, focusing on technical solutions for automatic extraction of knowledge entities and relationships from patent literature. Therefore, we systematically investigate relevant research on knowledge extraction for TRIZ analysis matrices, including general scientific knowledge extraction methods and patent-oriented knowledge extraction approaches.

2.1 Scientific Knowledge Extraction Research

Knowledge extraction methods for scientific literature have evolved through dictionary and rule-based approaches [2-4], statistical machine learning [5-6], deep learning [7-10], pre-trained language model-based methods, and comprehensive extraction techniques. In 2018, Google researchers Devlin et al. [11] proposed the BERT model, which employs a bidirectional Transformer Encoder structure pre-trained on large-scale public corpora to obtain more robust pre-trained character vectors, significantly improving model performance. Reference [12] ap-

plied BioBERT to extract biological entities from 29 million PubMed abstracts, achieving state-of-the-art performance. Zhang et al. [13] pre-trained BERT on Chinese clinical corpora and used the resulting word embeddings as input features for a BiLSTM-CRF model to address breast cancer named entity recognition. Reference [14] proposed a model that leverages both pre-trained BERT language models and target entity information for relation classification tasks, achieving significant improvements over contemporary methods. Tang et al. [15] utilized BERT pre-trained language models to build a BERT-BiGRU-CRF sequence labeling model for joint entity-relation extraction from financial text corpora. In summary, "pre-training + fine-tuning" technology, by incorporating general and effective linguistic knowledge encoding, substantially enhances the performance of entity and relation extraction.

2.2 Patent-Oriented Knowledge Extraction Research

Since the pioneering work of Tsourikov et al. [16], various methods have been proposed for patent knowledge extraction, including SAO (Subject-Action-Object) methods, ontology-based approaches, and statistical machine learning methods. Hu et al. [17] identified generic TRIZ technical information (technical problems, solutions, functions, and effects) from patent literature based on SAO structural semantic analysis and LDA clustering methods. H.B. Kim et al. [18] extracted technical problems and solution entities from patents using the SAO method. Li et al. [19] identified five entity types—materials, products, methods, efficacy, and applications—from nano-fertilizer domain patents using the SAO method. Machine learning models for extracting knowledge entities and relationships from patent literature primarily include maximum entropy models, SVM models, and CRF models. Li et al. [20] proposed a patent function information extraction method based on lexical analysis, syntactic analysis, and maximum entropy classification models. Nanba et al. [21] identified technology and effect information in academic literature and patents using SVM methods. Lai et al. [22] designed an ontology structure for rice breeding methods based on TRIZ theory, applying SVM models to identify breeding methods in patents and CRF models to identify rice varieties.

The above review reveals that knowledge extraction technologies for TRIZ analysis matrices, particularly entity recognition and relation extraction for patents, remain far from mature. Existing approaches suffer from issues such as lack of annotated datasets, insufficient types of extracted patent knowledge, low automation levels, and performance requiring improvement. Therefore, this study aims to improve entity recognition and relation extraction for patents using pre-trained language model technology.

3 Design of Intelligent Construction Method for TRIZ Analysis Matrix

The framework of the intelligent construction method for the TRIZ analysis matrix proposed in this study is illustrated in Figure 1 [Figure 1: see original paper], consisting of three main modules: domain knowledge graph schema design for TRIZ analysis matrices, patent-oriented knowledge entity and relationship extraction methods, and automatic TRIZ analysis matrix construction based on domain knowledge graphs.

The workflow proceeds as follows: (1) Under the guidance of TRIZ innovation theory and facing the knowledge requirements of specific domains, we extend and refine the generic "technical elements" (technical problems, solutions, functions, and effects) in semantic TRIZ into domain-specific knowledge entity and relationship categories, completing the domain knowledge graph schema design. (2) Based on the BERT pre-trained language model, we design intelligent tools to automatically identify knowledge entities and relationships required for TRIZ analysis matrix construction from patent documents, achieving automatic patent entity and relation extraction. (3) After simple fusion of knowledge entities and relationships, we construct the domain knowledge graph using a graph database. For specific requirements, we automatically query and retrieve required knowledge entities and relationships from the domain knowledge graph based on knowledge queries and reasoning, thereby realizing automatic TRIZ analysis matrix construction.

This framework provides a universal method workflow and model architecture across domains, clarifying the domain knowledge graph schema design process, establishing a generic continual pre-training BERT model, entity recognition fine-tuning model, and relation extraction fine-tuning model architecture, and offering standardized domain knowledge graph construction methods. When applied to specific domains, only domain requirements and patent corpora are needed to construct TRIZ analysis matrices following this framework.

3.1 Domain Knowledge Graph Schema Design for TRIZ Analysis Matrix

Semantic TRIZ utilizes semantic technologies to automatically or semi-automatically model technical information implicit in patents, effectively representing patent-specific technical knowledge such as "technical problems, technical solutions, technical functions, and technical effects" [23]. Referencing the semantic TRIZ model, this study clarifies knowledge elements in the TRIZ analysis matrix knowledge model as four major functional semantic types: technical problem, technical solution, technical function, and technical effect, and conducts domain knowledge graph schema design. The specific process is shown in Figure 2 [Figure 2: see original paper]. First, we conduct requirement analysis for specific domains, extending and refining the generic knowledge elements (technical problem, technical solution, technical function, technical

effect) in the TRIZ analysis matrix knowledge model into domain-specific knowledge entity and relationship categories. On this basis, we establish mappings between knowledge entities and their relationships with knowledge nodes and relations in the graph, formulate unified semantic relation classification specifications, and complete the domain knowledge graph schema design to guide subsequent domain knowledge graph and TRIZ analysis matrix construction.

When designing knowledge graph schemas for different domains, the diversity and variability of domain requirements lead to different innovation elements corresponding to technical problems, technical solutions, technical functions, and technical effects across domains, thereby forming domain-specific TRIZ analysis matrix knowledge models. The subsequent sections will demonstrate domain knowledge graph schema design for the thin-film magnetic head technology domain, detailed in "4.2.1 Schema Design for Thin-Film Magnetic Head Technology Domain."

3.2.1 Technical Path for Knowledge Entity and Relationship Extraction

Based on BERT pre-trained language model technology, we design intelligent tools to automatically identify knowledge entities and relationships required for TRIZ analysis matrix construction from patent documents. The technical path for knowledge entity and relationship extraction is shown in Figure 3 [Figure 3: see original paper], comprising three stages: pre-training, continual pre-training, and fine-tuning. The pre-training stage directly introduces the original BERT model trained by Google using a 12-layer Transformer Encoder on Wikipedia and Book Corpus. The continual pre-training stage further trains the original BERT model on large-scale unlabeled domain patent corpora to learn more patent syntactic structure knowledge and domain-specific knowledge, obtaining domain-specific BERT word embeddings. The fine-tuning stage trains and optimizes the entity recognition model and relation extraction model separately using small amounts of task-specific annotated data. This study aims to improve entity recognition and relation extraction model performance while reducing the model' s dependency on annotated data during the fine-tuning stage.

3.2.2 BERT Model Continual Pre-training

BERT models have three primary training and usage modes: "pre-training + fine-tuning," "pre-training + continual pre-training + fine-tuning," and "re-pre-training + fine-tuning" (Figure 4 [Figure 4: see original paper]). The ACL 2020 Best Paper nomination "Don' t Stop Pretraining: Adapt Language Models to Domains and Tasks" [24] conducted extensive language model pre-training experiments, demonstrating that continual pre-training on target domain datasets can improve pre-trained language model performance on domain tasks. Chen et al. [25] found significant differences between patent datasets from different technical domains through comparative analysis, and observed that domain-

specific word embeddings perform better on domain tasks compared to general patent-domain embeddings.

Therefore, this study adopts the "pre-training + continual pre-training + fine-tuning" path for BERT models. For specific domain patent literature entity-relationship characteristics, we conduct continual pre-training of the original BERT model on unlabeled patent corpora from target domains, enabling the model to learn more patent syntactic structure knowledge and domain-specific knowledge, thereby obtaining domain-specific BERT word embeddings to achieve optimal entity recognition and relation extraction performance.

3.2.3 BERT-CRF Entity Recognition Model Design

Guided by sequence labeling principles, this study employs the BERT-CRF entity recognition model with architecture shown in Figure 5 [Figure 5: see original paper]. The model consists of three components: feature representation, feature encoding, and label decoding. The feature representation step uses BERT to generate distributed vector representations of input text. The feature encoding step transforms BERT' s output vectors through a linear layer to extract sentence semantic features, converting hidden state sequence vector dimensions to annotation label quantity dimensions to obtain prediction scores for each label. The label decoding step uses CRF for decoding, taking features extracted by BERT as input and considering contextual label dependencies to obtain the entity label sequence with maximum conditional probability.

3.2.4 BERT-softmax Relation Extraction Model Design

Relation extraction adopts a multi-label classification approach, constructing a BERT-softmax relation extraction model. Specifically, BERT provides sentence word embedding representations and learns semantic features, which are then fed into a softmax classifier to output relation classification results. The relation classification model is trained using annotated corpora, taking sentences containing entity pairs as input and outputting relation categories.

Additionally, since this study employs a pipeline approach for entity recognition and relation extraction, the entity pair generation stage iteratively produces numerous entity pairs without actual relationships, which can interfere with relation extraction model training. To enable the relation extraction model to better learn features of unrelated entity pairs, we assign a special "no_{relation}" type to entity pairs without relationships as negative samples added to the training set, treating them equally with other relation types during training to improve relation extraction model performance.

3.3 Automatic TRIZ Analysis Matrix Construction Based on Domain Knowledge Graph

The automatic TRIZ analysis matrix construction process based on domain knowledge graphs is shown in Figure 6 [Figure 6: see original paper], comprising two stages: domain knowledge graph construction and TRIZ analysis matrix construction.

First, we perform simple fusion of extracted knowledge entities and relationships, ensuring each entity has a unique ID through an entity ID_{Name} mapping table. We merge identical knowledge entities using entity IDs and remove duplicate triples. On this basis, we store entity and relationship data through a graph database, converting them into nodes and relationships in the database to complete domain knowledge graph construction. Using the Neo4j graph database, we conduct interactive queries and associative reasoning for specific requirements, automatically retrieving required knowledge entities and relationships from the domain knowledge graph to achieve automatic TRIZ analysis matrix construction, providing effective support for subsequent system problem analysis, technical efficacy analysis, technology evolution analysis, and technology path analysis to accelerate technological innovation.

4 Empirical Study

Thin-film magnetic heads in the computer hard disk domain can significantly reduce the distance between the head and disk, increase data density, and improve accuracy, holding important significance for China' s high-performance magnetic recording industry development. Therefore, this study conducts an empirical investigation in the thin-film magnetic head technology domain under the aforementioned domain-general TRIZ analysis matrix intelligent construction methodology, specifically including: 1) constructing BERT models for patent knowledge extraction in the thin-film magnetic head technology domain; and 2) conducting empirical research on TRIZ analysis matrix construction for this domain.

4.1.1 Experimental Data

This experiment employs the TFH-2020 dataset constructed by Chen et al. [25] for patent entity recognition and relation extraction in the thin-film magnetic head technology domain. TFH-2020 is an annotated patent dataset for hard disk thin-film magnetic head technology. The dataset comprises 1,010 patent abstracts from the thin-film magnetic head technology domain retrieved from the United States Patent and Trademark Office (USPTO), containing 22,742 entities and 17,421 semantic relationships. The dataset defines 17 entity type specifications and 15 semantic relation type specifications. Entity types include: material flow, information flow, energy flow, system, component, attribute, shape, material, state, position, measurement object, value, scientific concept, function, effect, result, and other. Relation types include: instance,

alias, positional relation, part relation, causal relation, construction, attribute, in···manner, formation, comparison, measurement, operation, generation, and purpose.

4.1.2 BERT Model Continual Pre-training Experiments

"Magnetic head" is a hypernym of "thin-film magnetic head." Considering patent literature volume, this study conducts BERT model continual pre-training experiments for the magnetic head domain in computer hard disks. Specifically, we obtain large-scale unlabeled patent corpora in the magnetic head domain and perform continual pre-training on the original BERT model parameters, enabling the model to learn more patent syntactic structure knowledge and magnetic head domain knowledge, resulting in magnetic head domain BERT word embeddings—BERT-MH.

(1) Patent Corpora for Continual Pre-training
This experiment selects the Derwent Innovations Index (DII) as the patent corpus for BERT model continual pre-training. Using "magnetic head" as the search term in the title field, we retrieved 44,705 patent records (including titles and abstracts). After regex matching, filtering, and cleaning, we extracted title and abstract texts for each patent record to form an unlabeled patent corpus in the magnetic head domain. The final corpus size is 48.7MB, containing 431,210 sentences.

(2) Continual Pre-training for Magnetic Head Domain BERT Model
This study employs the Transformers library in PyTorch for BERT model continual pre-training. The constructed magnetic head domain unlabeled patent corpus is randomly divided into training and validation sets at a 9:1 ratio. Following techniques suggested in [11], we run additional pre-training steps on the corpus starting from existing BERT checkpoints. Key parameter settings for the continual pre-training experiments are shown in Table 1 .

The entire training process was completed on a Linux server with Tesla V100 GPU. After training, we obtained the PyTorch version of the magnetic head domain BERT model with .bin extension. This study refers to the BERT word embeddings obtained through continual pre-training on magnetic head domain unlabeled patent corpora as BERT-MH, used for subsequent knowledge entity recognition and relation extraction experiments.

4.1.3 Knowledge Entity Recognition Experiments

In the knowledge entity recognition experiments, the TFH-2020 dataset is randomly divided at the document level into training, validation, and test sets at a 6:2:2 ratio. The distribution of patent documents and sentences in each set is shown in Table 2 .

The knowledge entity recognition experiments design two tasks: model performance comparison and annotation data dependency experiments, to validate

the effectiveness of the technical path in improving model performance and reducing dependency on annotated data.

(1) Model Performance Comparison Experiment
We construct entity recognition models using BERT-softmax and BERT-CRF architectures for comparative experiments to achieve optimal performance. Experimental element settings are shown in Table 3 . To obtain the best-performing entity recognition model, we combine different experimental elements, yielding four deep learning models for evaluation. Model performance on entity recognition tasks is shown in Table 4 , where the BiLSTM-CRF+MH-46K model is the entity recognition model adopted in the TFH-2020 dataset paper.

The results demonstrate that the BERT-MH+CRF model achieves optimal performance on patent entity recognition in the thin-film magnetic head technology domain, proving that the proposed "pre-training + continual pre-training + fine-tuning" technical path can effectively improve patent entity recognition model performance. Further analysis yields three conclusions: 1) Compared with the BiLSTM-CRF+MH-46K architecture used in the dataset paper, our model architecture shows significant performance improvement, indicating that dynamic word embeddings can learn more feature knowledge than static embeddings; 2) Under the same downstream neural network architecture, BERT-MH outperforms BERT, demonstrating that continual pre-training on domain-specific unlabeled corpora is an effective approach to improve downstream model performance for domain tasks; 3) Under the same BERT pre-trained language model, CRF outperforms softmax, showing that model performance can still be improved by integrating other networks for specific tasks.

(2) Annotation Data Dependency Experiment
Using the best-performing BERT-MH+CRF model, we repeat entity recognition experiments with 50%, 40%, 30%, 20%, and 10% of annotated data. During training, all model parameters remain identical except for the epochs parameter. Evaluation employs both micro-average and macro-average approaches, with results shown in Table 5 and Figure 7 [Figure 7: see original paper].

The results show that as annotated data volume decreases, micro-average F1 scores decline slightly, while macro-average F1 scores drop significantly from 68.42% (full annotated data) to 44.40% (10% annotated data). This suggests that when datasets suffer from severe imbalance, reducing annotated data volume disproportionately affects entity categories with fewer samples. However, overall results indicate that continual pre-training on large-scale unlabeled domain patent corpora can reduce model dependency on annotated data during fine-tuning.

4.1.4 Knowledge Relation Extraction Experiments

(1) Entity Pair Generation
Entity pairs are generated traversing at the sentence level (avoiding self-

combination) and added to a candidate set. Entity pair generation rules filter the candidate set, removing obviously impossible semantic relation pairs. For remaining pairs, we match them with standard relation information in sentences to generate correct relation types, assigning a special "no_{relation}" type to pairs without any relation annotation. This process generated 205,119 entity pairs for relation extraction, with distribution shown in Table 6 . Notably, "no_{relation}" type entity pairs total 183,140, accounting for 89.3% of all pairs, creating an extremely imbalanced distribution that poses significant challenges for relation extraction model training.

(2) Relation Extraction Model Performance
Experiments adopt the BERT-softmax architecture for relation extraction models. BERT pre-trained language models include both the original BERT and the continually pre-trained magnetic head domain BERT-MH model. We combine BERT and BERT-MH models with softmax classifiers to identify the optimal relation extraction model. After training, model performance on the test dataset is shown in Table 7 , where the BiGRU-HAN model is the relation recognition model adopted in the TFH-2020 dataset paper.

The "with no_{relation}" rows in Table 7 show evaluation data considering the "no_{relation}" category, with the BERT-MH+softmax model achieving 89.7% F1. However, the "without no_{relation}" rows more accurately reflect true performance on patent relation extraction tasks. The BERT-MH+softmax model achieves optimal performance with 63.7% F1, slightly higher than BERT+softmax (63.56%) and substantially outperforming BiGRU-HAN (41.5%). This demonstrates that the "pre-training + continual pre-training + fine-tuning" technical path effectively improves relation extraction model performance, and domain-specific BERT embeddings better enhance downstream model performance on domain tasks compared to general-domain embeddings.

Notably, the BERT-MH+softmax relation extraction model' s F1 performance still falls short of state-of-the-art levels, likely due to: 1) Patent texts contain far more entities than general texts, resulting in a much higher proportion of unrelated entity pairs and increased training difficulty; 2) Error propagation issues in pipeline methods, where incorrect entity recognition results inevitably lead to semantic relation extraction errors and reduced model performance.

4.2 Empirical Study on Thin-Film Magnetic Head Technology Domain TRIZ Analysis Matrix Construction

Applying the aforementioned thin-film magnetic head technology domain patent knowledge extraction models—BERT-MH+CRF for entity recognition and BERT-MH+softmax for relation extraction—we conduct an empirical study on TRIZ analysis matrix construction for this domain.

4.2.1 Schema Design for Thin-Film Magnetic Head Technology Domain

Thin-film magnetic head technology patents primarily concern magnetic head system structures, working mechanisms, and the position, attributes, and material composition of components. Therefore, for the thin-film magnetic head technology domain, technical problems typically refer to patent research objects including material flow, information flow, and energy flow; technical solutions typically refer to thin-film magnetic head system structures including component attributes, shapes, materials, and positions; technical functions refer to functions implemented by the magnetic head system; and technical effects refer to system impacts, effects, and generated results.

Based on this analysis and referencing entity and relation types defined in the TFH-2020 dataset, we specify entity categories for the thin-film magnetic head technology domain knowledge graph as: material flow, information flow, energy flow, system, component, attribute, shape, material, position, function, effect, and result—12 knowledge entity types in total. The correspondence between knowledge entities and the four major knowledge elements is shown in Table 8 . Relation categories are specified as: alias, positional relation, part relation, causal relation, construction, attribute, in···manner, formation, operation, and purpose—10 knowledge relation types. The resulting conceptual model of the thin-film magnetic head technology domain knowledge graph is shown in Figure 8 [Figure 8: see original paper].

4.2.2 Patent Data Acquisition for Thin-Film Magnetic Head Technology Domain

Using "thin film magnetic head" as the search term in the title field of the Derwent Innovations Index database, we retrieved 1,732 patent records related to thin-film magnetic head technology (including titles and abstracts). After data cleaning and statistics, we obtained a patent data document of 2.88MB as the patent corpus for TRIZ analysis matrix construction in this domain. The corpus contains 15,666 valid sentences with 372,650 words, averaging 23.8 words per sentence.

4.2.3 Knowledge Entity Recognition and Relation Extraction Results

After data preprocessing, the BERT-MH+CRF model identified 70,844 entities from the thin-film magnetic head technology domain patent corpus, with entity category distribution shown in Table 9 . Following entity recognition, 335,596 entity pairs were generated. Relation extraction experiments using the BERT-MH+softmax model yielded relation category distribution shown in Table 10 .

4.2.4 Thin-Film Magnetic Head Technology Domain Knowledge Graph Construction

We perform simple fusion of extracted knowledge entities and relationships, ensuring each entity has a unique ID through an entity ID_{Name} mapping table. After simple deduplication, we obtain 14,014 entities, and after removing "no_{relation}" relationships, we retain 20,863 valid relationships.

The deduplicated entity set and relationship set are structurally stored with corresponding labels in CSV format. Using Neo4j graph database' s neo4j-admin import tool, we batch import CSV-format entity and relationship files, successfully importing 14,014 entity nodes and 20,863 relationships to construct the thin-film magnetic head technology domain knowledge graph shown in Figure 9 [Figure 9: see original paper] (left). The knowledge graph exhibits an overall clustered structure with only a few entity relationships scattered peripherally. Figure 9 (right) shows detailed entity and relationship structures.

4.2.5 Thin-Film Magnetic Head Technology Domain TRIZ Analysis Matrix Construction

Recording density and frequency are critical indicators for evaluating thin-film magnetic head performance. Based on this, we demonstrate TRIZ analysis matrix construction around the innovation requirement of "how to improve thin-film magnetic head recording density and frequency." First, we identify the desired technical effect as high recording density and frequency in thin-film magnetic heads. Using "high recording density and frequency" as the keyword, we query the constructed knowledge graph with the statement: "match(x) where x.name =' high recording density and frequency' return x."

The query identifies the "high recording density and frequency" entity, which is directly associated with the "invented thin-film magnetic head" entity. Further expansion reveals direct associations with "width," "pole part," "track narrowing," and "schematic structure" entities, suggesting that the pole part is a key component for achieving high recording density and frequency. We continue expanding entities associated with "pole part." After manual assessment of relevance and removal of less relevant entities, the final query results are shown in Figure 10 [Figure 10: see original paper].

The results show that "high recording density and frequency" directly relates to "invented thin-film magnetic head," which in turn directly relates to "schematic structure," "pole part," "width," and "track narrowing." Based on these key knowledge entities and relationships, we construct the TRIZ analysis matrix shown in Table 11 .

Analysis of this TRIZ analysis matrix reveals that the effect of "high recording density and frequency" co-occurs with the result of "track narrowing" and the attribute of "width,"while the experimental realization of these objectives relates to the "pole part" component of the "thin-film magnetic head." This suggests

that high recording density and frequency may be related to the width of the pole part track, and reducing the track width of the thin-film magnetic head pole part could serve as an exploration direction for improving magnetic head recording density and frequency.

5 Conclusion

This study addresses the issues of slow speed, high cost, and insufficient knowledge elements in traditional manual construction of system interaction analysis matrices by proposing an intelligent construction method for TRIZ analysis matrices. Under the guidance of TRIZ innovation theory, we extend the knowledge content of system interaction analysis matrices for specific domains, define knowledge entity and relationship categories, and complete domain knowledge graph schema design for TRIZ analysis matrices. Based on the "pre-training + continual pre-training + fine-tuning" technical path, we employ BERT pre-trained language model technology to automatically identify knowledge entities and relationships from patent literature, constructing domain knowledge graphs. For specific requirements, we automatically query and retrieve required knowledge entities and relationships from domain knowledge graphs, ultimately achieving automatic TRIZ analysis matrix construction. Empirical validation demonstrates that the proposed method can achieve the goal of automatic TRIZ analysis matrix construction and provide effective knowledge support for technological innovation.

However, this study has certain limitations: (1) The pipeline approach for entity recognition and relation extraction suffers from error propagation issues; (2) The TRIZ analysis matrix construction process primarily relies on interactive queries to obtain required knowledge entities and relationships, lacking knowledge reasoning technology application that prevents revealing potential knowledge associations between entities; (3) The quality of constructed TRIZ analysis matrices remains insufficient. Future work will introduce knowledge reasoning technology to deeply mine potential knowledge associations in knowledge graphs, improve TRIZ analysis matrix quality, and provide enhanced support for technological innovation.

References

[1] Zeng Bei, Lü Jianqiu. Analysis of Research Hotspots and Frontiers of TRIZ in China Based on CiteSpace[J]. Science and Technology Management Research, 2019, 39(18): 260-
[2] Kiela D, Guo Y, Stenius U, et al. Unsupervised discovery of information structure in biomedical documents[J], 2015, 31(7): 1084-1092.
[3] Guo Y, Silins I, Stenius U, et al. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review[J], 2013, 29(11): 1440-1447.
[4] Guo Y, Reichart R, Korhonen A. Improved information structure analysis of

scientific documents through discourse and lexical constraints[C]. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 928-937.
[5] Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[J], 2003.
[6] Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition[C]. COLING 2002: The 19th International Conference on Computational Linguistics, 2002.
[7] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J], 2016.
[8] Yoon W, So C H, Lee J, et al. Collabonet: collaboration of deep neural networks for biomedical named entity recognition[J], 2019, 20(10): 55-65.
[9] Yu Li, Qian Li, Fu Changlei, et al. Research on Fine-grained Knowledge Element Extraction from Text Based on Deep Learning[J]. Data Analysis and Knowledge Discovery, 2019, 3(01): 38-45.
[10] Gao H, Gui L, Luo W. Scientific literature based big data analysis for technology insight[C]. Journal of Physics: Conference Series, 2019: 032007.
[11] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J], 2018.
[12] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J], 2020, 36(4): 1234-1240.
[13] Zhang X, Zhang Y, Zhang Q, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods[J], 2019, 132: 103985.
[14] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]. Proceedings of the 28th ACM international conference on information and knowledge management, 2019: 2361-
[15] Tang Xiaobo, Liu Zhiyuan. Research on Joint Extraction of Financial Text Sequence Labeling and Entity Relations[J]. Information Science, 2021, 39(05): 3-11.
[16] Tsourikov V M, Batchilo L S, Sovpel I V. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures: Google Patents, 2000.
[17] Hu Zhengyin, Liu Chunjiang, Wei Ling, et al. Design and Practice of Domain Patent Technology Mining System for TRIZ[J]. Library and Information Service, 2017, 61(01): 117-124.
[18] Kim H, Hyeok Y, Kim K. Semantic SAO network of patents for reusability of inventive knowledge[C]. 2012 IEEE International Conference on Management of Innovation & Technology (ICMIT), 2012: 510-515.
[19] Li Xiaoman, Zhang Xuefu, Song Hongyan, et al. Research on Patent Literature Technical Element Recognition Method—Taking the Nano-fertilizer Domain as an Example[J]. Library and Information Service, 2020, 64(06): 59-68.
[20] Li Weichao. Research on Patent Function Information Extraction Method[D]. Hebei University of Technology, 2013.
[21] Nanba H, Kondo T, Takezawa T. Automatic creation of a technical trend

map from research papers and patents[C]. Proceedings of the 3rd international workshop on Patent information retrieval, 2010: 11-16.
[22] Lai Yingxu, Li Yajuan, Liu Jing. Construction of Application Knowledge Base for Rice Breeding Methods Based on Ontology[J]. Journal of Beijing University of Technology, 2019, 45(12): 1181-1191.
[23] Zhang Xian, Hu Zhengyin, Ru Lijie, et al. Research on Patent Technology Supply-Demand Information Association Knowledge Organization Mode[J]. Library and Information Service, 2016, 60(08):
[24] Gururangan S, Marasović A, Swayamdipta S, et al. Don' t stop pretraining: adapt language models to domains and tasks[J], 2020.
[25] Chen L, Xu S, Zhu L, et al. A deep learning based method for extracting semantic information from patent documents[J], 2020, 125(1): 289-312.

Note: Figure translations are in progress. See original paper for figures.