Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Digital twin system framework and information model for industry chain based on industrial Internet*#

Wenxuan WANG[1], Yongqin LIU[1,2], Xudong CHAI[3], Lin ZHANG‡[1]

[1]*School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China*

[2]*Department of Control Science and Engineering, Jilin University, Changchun 130012, China*

[3]*CASICloud, China Aerospace Science and Industry Corporation Limited, Beijing 100080, China*

E-mail: wangwenxuan0516@126.com; liu_yq@buaa.edu.cn; xdchai@263.net; zhanglin@buaa.edu.cn

Received Feb. 28, 2023; Revision accepted Sept. 11, 2023; Crosschecked Apr. 7, 2024

Abstract: The integration of industrial Internet, cloud computing, and big data technology is changing the business and management mode of the industry chain. However, the industry chain is characterized by a wide range of fields, complex environment, and many factors, which creates a challenge for efficient integration and leveraging of industrial big data. Aiming at the integration of physical space and virtual space of the current industry chain, we propose an industry chain digital twin (DT) system framework for the industrial Internet. In addition, an industry chain information model based on a knowledge graph (KG) is proposed to integrate complex and heterogeneous industry chain data and extract industrial knowledge. First, the ontology of the industry chain is established, and an entity alignment method based on scientific and technological achievements is proposed. Second, the bidirectional encoder representations from Transformers (BERT) based multi-head selection model is proposed for joint entity–relation extraction of industry chain information. Third, a relation completion model based on a relational graph convolutional network (R-GCN) and a graph sample and aggregate network (GraphSAGE) is proposed which considers both semantic information and graph structure information of KG. Experimental results show that the performances of the proposed joint entity–relation extraction model and relation completion model are significantly better than those of the baselines. Finally, an industry chain information model is established based on the data of 18 industry chains in the field of basic machinery, which proves the feasibility of the proposed method.

Key words: Industry chain; Digital twin; Industrial Internet; Knowledge graph; Graph neural network
https://doi.org/10.1631/FITEE.2300123 **CLC number:** TP39

1 Introduction

The spread of COVID-19 and the countercurrent of economic globalization have had a significant impact on the long-established industry chain and supply chain (Yin et al., 2020). The traditional industry chain lacks a systematic risk early warning and prevention mechanism. In addition, there is a lack of corresponding data support for supplementing and strengthening the industry chain (Zhang XY et al., 2022). Therefore, it is necessary to design a data-driven industry chain analysis system to realize dynamic perception of the industry chain, which includes the assessment of the production capacity and supply capacity of the industry chain under different conditions, and the timely discovery of potential risks in the industry chain.

The industrial Internet is the deep integration of the new generation of network information technology and manufacturing. It is an important

‡ Corresponding author

* Project supported by the Ministry of Industry and Information Technology of China (Nos. TC200802C and TC190A445)

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2300123) contains supplementary materials, which are available to authorized users

ORCID: Wenxuan WANG, https://orcid.org/0009-0009-8060-3656; Lin ZHANG, https://orcid.org/0000-0003-1989-6102

infrastructure for realizing digital, networked, and intelligent development of the industry chain (Li et al., 2017). The emergence and development of digital twin (DT) have led to the recognition that DT and industrial Internet are complementary (Kamble et al., 2022). Based on the industrial Internet, DT deeply integrates digital space and physical space, thus effectively promoting the intelligent development of the industry chain. At the same time, the networked connection and collaborative capabilities for industrial elements of the industrial Internet provide an ecological mechanism for the construction of the DT system for the industry chain.

At present, the connotation of DT has expanded from the original complex product life-cycle mirroring and high fidelity to the manufacturing field of workshops and enterprises (Al Faruque et al., 2021). However, how to use DT technology to explicitly express complex industry chain data, to analyze the risk of the breakpoints and blocking points in the industry chain, and to optimize the scheduling of industry chain resources is a research gap. Min et al. (2019) proposed a DT framework for the petrochemical industry based on the industrial Internet, which integrates data processing and machine learning methods. However, this framework is applicable only to the petrochemical industry. Chen et al. (2023) proposed a heuristic multi-cooperation scheduling framework based on blockchain and DT to ensure the production efficiency and information security of the entire manufacturing process. However, this model ignores the upstream and downstream supply of the industry chain, and cannot predict the risks in the supply chain. Cheng et al. (2020) proposed a DT-enhanced industrial Internet (DT-II) reference framework, and discussed the DT-II implementation mechanism for a single industry chain. However, this structure does not consider the interconnection among multiple industry chains, and it is difficult to achieve cross-industry resource sharing. Therefore, in this paper we propose a DT system framework that can comprehensively manage multiple industry chains to integrate industry chain information and achieve comprehensive perception from a macro process to a micro process. Based on this framework, the capabilities of collaborative innovation, anti-risk, and comprehensive management of all links in the industry chain can be improved.

Processing complex data and mining potential features based on the industrial Internet are the keys to building an industrial chain DT system. However, the industry chain is characterized by dynamics, complex relationships, and huge data volumes, making it difficult to calculate, organize, and adjust in time. How to achieve cross-region, cross-field, and cross-factor industrial analysis and layout reconstruction is a challenge.

Knowledge graph (KG) is an effective method for mining complex relationships from industrial data (Ren et al., 2023). However, there are three challenges in extracting industry chain knowledge from the data in the industrial Internet. First, the industrial Internet contains massive multi-source heterogeneous industry chain data, which are distributed on the information islands of different systems, making it difficult to effectively connect. However, there are few general methods for mining industry chain knowledge from a large volume of expert knowledge, public databases, and enterprise manufacturing data. Second, knowledge of the industry chain is extracted mainly from Internet data, enterprise internal management systems, and industrial equipment and products. Therefore, knowledge of the industry chain includes structured, semi-structured, and unstructured information resources. This leads to the lack of a unified structured expression of industry chain knowledge and the inefficiency of data processing. Finally, industry chain knowledge is characterized by strong professionalism and complex relationships. For example, there are complex relationships such as industry upstream and downstream activities, enterprise geographic relationships, processing operations, and parameter semantics. In addition, the hidden relationships in the industry chain knowledge need to be further explored. This challenges the effective preservation of structure and content information of industry chain knowledge through representation learning.

To solve these problems, we propose an industry chain DT system framework that provides decision support for maintaining industry chain stability, strengthening chain, and complementing chain. Based on the industrial Internet, the system can schedule production on demand, predict risks in the industry chain, and analyze capacity bottlenecks. Then, to deal with multi-source heterogeneous data and integrate industry chain knowledge, an industry chain information model based on a KG is designed

which aims at processing multi-source heterogeneous data in the industry chain and extracting knowledge from unstructured data.

The main contributions of this article are as follows:

1. A construction framework for the industry chain DT system is proposed, including the industry chain information model, the industry chain risk assessment analysis model, and the industry chain capacity simulation model. They are used for real-time analysis of the state, robustness, and industry chain capacity feasibility decisions, separately.

2. A KG-based industry chain information model is designed, to extract industry chain knowledge from heterogeneous, multi-source industrial Internet data. The model realizes comprehensive real-time monitoring of the industry chain.

3. An industry chain knowledge ontology database is proposed which provides a basis for the construction of KG of the industry chain. A bidirectional encoder representations from Transformers (BERT) based multi-head selection model is proposed for joint entity–relation extraction of the industry chain from the industry chain data to obtain the original KG.

4. A KG relation completion model based on a graph neural network is proposed. In this model, a relational graph convolutional network (R-GCN) is used to obtain semantic information in the KG, a graph sample and aggregate network (GraphSAGE) is used to obtain the structural information of the local subgraph around the target node in the KG, and the multi-head attention mechanism is used to combine the scoring results of these two kinds of information.

2 Literature review

2.1 Industrial chain management technology

The industrial Internet is a technology that integrates physical and network components, enabling all-factor connection of enterprise information systems, machines, and people (Wang JL et al., 2020). It helps companies gain insights into industrial processes from data to improve productivity, efficiency, and reliability (Qin et al., 2020). At present, the industrial Internet platform has become the core of industrial system operation, which extends the field of enterprise management to all stages of the product life cycle (Menon et al., 2019). The industrial Internet based industry chain optimization objectives are divided mainly into the following three aspects: (1) establish a comprehensive and systematic feature representation model for the industry chain based on the chain-like relationship formed by the logical relationship and the spatial–temporal layout relationship of the industry; (2) identify breakpoints and blocking points in the industry chain and form a consortium with upstream and downstream enterprises in the industry chain to achieve collaborative innovation; (3) organize enterprises in the industry chain to share production capacity, strengthen supply chain coordination, and empower resilient scheduling of the industry chain.

With the capability of high-fidelity modeling and simulation computing/analysis, DT meets the high demand for industry chain optimization under the industrial Internet. In recent years, the connotation of DT has expanded from complex product life-cycle mirroring and high fidelity to the manufacturing field of workshops and enterprises (Kiel et al., 2017). Zhou GH et al. (2020) proposed a general framework for a knowledge-driven DT manufacturing cell towards intelligent manufacturing. It was verified by three application examples about intelligent process planning, intelligent production scheduling, and production process analysis and dynamic regulation. Zhuang et al. (2018) proposed a framework of DT-based smart production management and a controlled approach for complex product assembly shopfloors, and the real-time perception of physical assembly workshop data was studied. Kong et al. (2021) proposed a data construction method to provide stable and efficient data support for the application of the DT system in the manufacturing workshop. Bao et al. (2019) proposed an approach of modeling and operations for DT in the context of manufacturing. Moreover, they provided implementation methods for virtual–physical convergence and information integration for a factory. Lu and Xu (2019) proposed a generic system architecture for cloud-based manufacturing equipment based on DT systems and big data analytics. Sun XM et al. (2020) proposed a DT-driven assembly-commissioning method for high-precision products with multi-discipline coupling. However, most current studies focus on DT

in specific products, production lines, or enterprises. There are relatively few applications of DT in the industry chain.

2.2 Industry chain analysis technology based on a knowledge graph

The industry chain is characterized by dynamics, complex relationships, and huge data volumes, making it difficult to calculate, organize, and adjust in time. At the same time, to conduct a comprehensive analysis of the industry chain, it is necessary to carry out data mining and model construction for the association rules of all elements of the industrial Internet in the complex semantic environment. KG is an effective method for mining complex correlations from industry chain data, and can effectively deal with the dynamically growing massive industry chain data in the industrial Internet. Zhou B et al. (2023) proposed an industrial KG embedding time-series data, which integrates dynamic time-series semantic events and workshop knowledge from technical documents. It effectively associates the dynamic data and static data between production resources. Tan et al. (2021) proposed an automatic construction of a KG for the electric power field and applied it to power equipment inspection. Liu YSY et al. (2022) proposed an industrial KG to integrate resources for manufacturing enterprises, which can recommend resources in low-resource conditions for industrial collaboration and thereby promote the production and operation efficiency of manufacturing enterprises. Hedberg et al. (2020) introduced a method for linking and tracing data throughout the product lifecycle using graphs to form digital threads. Liu MF et al. (2022) established a multilayer manufacturing KG based on the digital thread of manufacturing process data, including device sensing data, production processing data, and business processing data, and realized perception analysis and cognition decision-making in the resource allocation of the manufacturing process. Sarazin et al. (2021) combined KG with an aviation system, proposed an advanced expert system for heterogeneous information network, and developed condition maintenance of a complex system. Guo et al. (2022) proposed a framework for automatic construction of a knowledge base in the machining field based on a KG, solving the time-consuming and labor-intensive issues of traditional frameworks.

However, there is little research on systematic construction of KG for the industry chain. Most research focuses on the specific industries, production quality, and production processes in the industry chain. The Industrial chain cooperation, enterprise geographic relationships, processing operations, and parameter semantics in the industrial Internet remain to be researched. The hidden relationship in the industry chain KG remains to be explored.

3 Digital twin technology for the industrial chain based on industrial Internet

Industry chain is a system involving multiple fields. At present, few scholars consider the construction of DT systems involving multiple fields (products, factories, and upstream and downstream supplies). However, it is difficult for a single-domain DT system to meet the cross-domain interaction requirements in the industry chain, and it is difficult to completely simulate, analyze, and optimize the industrial chain. In addition, the industry chain DT system has to dynamically perceive and analyze, in real time, the state of the industry chain, the robustness of the industry chain, and the feasibility decision of the industry chain capacity. At present, there is no industry chain DT system that can simultaneously predict the risks of the industry chain, analyze the capacity bottlenecks of the industry chain, and organize related production resources.

3.1 System framework

DT is a digital model of a physical object that evolves in real time by receiving data from the physical object, thereby remaining consistent with the physical object throughout its lifecycle (Zhang L et al., 2021). At present, DT systems of different levels represented by 3P (product, production, and process) have been formed, as shown in Fig. 1.

1. Product-grade DT system. The product-grade DT system establishes a dynamic mapping between physical and virtual products, thereby collecting and analyzing the state of physical products in real time and realizing remote diagnosis and predictive maintenance of the product state. At the same time, the real-time simulation technology is used to carry out design optimization iterations in the virtual world, forming a closed loop for continuous improvement of product quality.

2. Production-grade DT system. The production-grade DT system digitizes various elements in the manufacturing process. Then, it forms the connection and mapping between the digital factory and physical factory through the industrial Internet platform. In addition, the system optimizes the production process by converging production process data through the industrial Internet. It realizes industrial intelligent applications at the production site based on cloud–edge–device collaboration, thereby continuously improving the efficiency and quality of the production process.

3. Process-grade DT system. The establishment of a process-grade DT system is based on product-grade and production-grade DT systems. The process-grade DT system applies and extends the DT technology to other business processes of enterprises and even the supply networks. The system connects various heterogeneous systems through the industrial Internet, realizes the integration of complex heterogeneous data, and forms a multi-scale DT of "all regions, all industries, and all time." In addition, the system needs to provide a reliable simulation foundation and continuously optimize the business process of the enterprise through data analysis.

The DT system proposed in this paper is a process-grade DT system. It needs to connect all elements of the industry chain to form a comprehensive interconnection across elements, enterprises, industries, and regions, as shown in Fig. 2.

3.2 Establishment flow of a DT system

The industry chain DT system framework proposed in this paper includes four layers: platform layer, data layer, model layer, and application layer, as shown in Fig. 3. The platform layer is constructed based on the industrial Internet, which is embedded with massive-data storage capabilities compatible with time-series data, structured data, and unstructured data. In addition, the industrial Internet platform provides a basic environment for distributed computing and industrial big data analysis. In the data layer, the data of enterprises, products, equipment, production capacities, investment financial reports, patents, and papers are collected using data collection tools. In addition, data transmission, data processing, and data collection technologies are

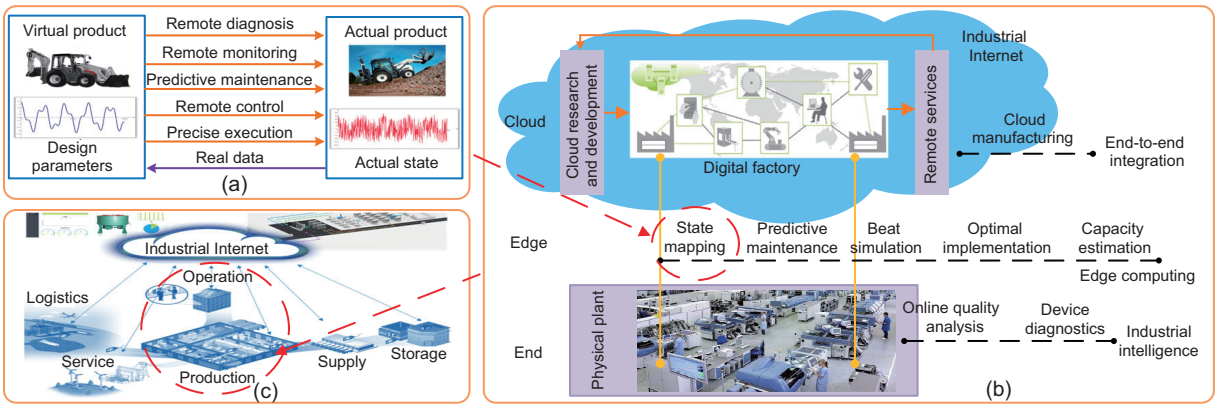

Fig. 1 Digital twin systems: (a) product grade; (b) production grade; (c) process grade

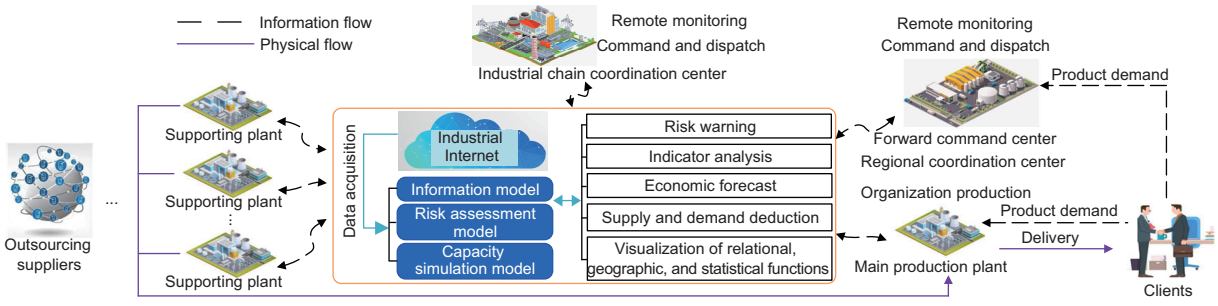

Fig. 2 Technical framework of industry chain analysis based on the industry chain digital twin systems

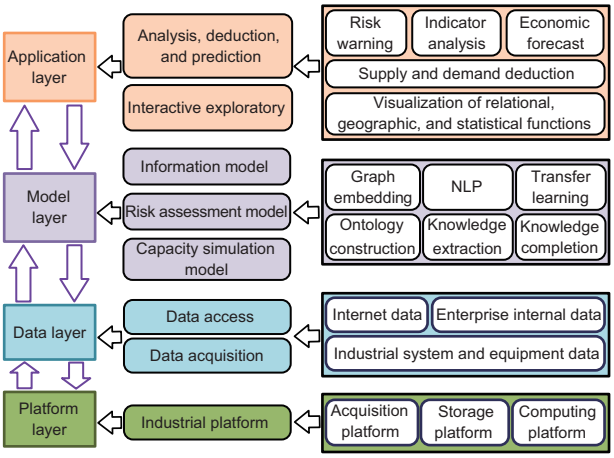

Fig. 3 Digital twin based industry chain analysis technology framework (NLP: natural language processing)

used to form a complete industrial data collection system. In the model layer, through natural language processing technology, entities such as products, industries, upstream and downstream enterprises, and typical companies are identified from the data layer. The identified entities and relationships are stored in the graph database. Then, combined with the industrial economics theories, the industry chain information model, industry chain risk assessment analysis model, and industry chain capacity simulation model are established to form an algorithmic support for the industry chain DT system. In the application layer, powerful visualization plug-ins for the relational, geographic, and statistical functions provide users with an interactive exploration interface. In addition, the functional modules such as risk warning, indicator analysis, economic forecast, and supply and demand deduction are designed to provide applications for various users.

3.2.1 Data collection and access

The industry chain DT system should have information security protection technology and provide a secure hypertext transfer protocol, to realize a safe and reliable two-way connection between devices and platforms. In addition, the system can be equipped with security hardware according to application requirements, providing system-level security monitoring, kernel-level security control, and trusted application program management to realize secure collection, secure transmission, and secure application of device data.

Industrial DT is constructed by simulating a series of entities, relationships, and events surrounding related industries in the objective physical world. It provides an analysis and deduction environment for decision-makers, and then assists them in introducing a series of industrial regulation measures. The objects of industrial DT monitoring are the enterprise development dynamics, product supply and demand, equipment operation, project establishment and execution, talent flow, technological innovation trends, investment events, and financing events from the physical world (capital world, technological world, or industrial world).

There are three data sources for this paper. The first aspect is the Internet public data, including corporate business data, financial report data, papers, and software, forming the macro composition of industry chain data. The second aspect is the enterprise internal management system data, including business flow data in the supply chain management system, enterprise resource planning (ERP) system, and customer relationship management (CRM) system. The third aspect is the data from the industrial systems, equipment, and products (descriptions of the three data sources are provided in the supplementary materials).

3.2.2 Industrial chain DT model construction

The industry chain DT system needs to analyze and simulate multiple industry chain network models, including the industry chain information model, industry chain risk assessment analysis model, and industry chain capacity simulation model. These three models and their analysis tools enable the system's dynamic perception and real-time analysis of the industry chain state, industry chain robustness, and industry chain capacity feasibility decisions.

1. Information model. The industrial chain is a complex system that can be divided into several layers and different scales. The complex DT should be able to display the entities in different scales as needed, and the model, data, and behavior could be adjusted accordingly as the perspective changes (Jia et al., 2022). KG can provide a bridge between DTs of different scales. It is an ideal technology for building an information model of the industrial chain.

2. Risk assessment model. First, obtain data through the information model. In the risk

assessment model, the clustering algorithm is used to analyze the risk factors. Then, organize all the influencing factors into several key indicators, and classify the risk levels. Third, build a comprehensive quantitative index system based on key indicators to predict potential risks in the industrial chain. Finally, explore the relationship between entities and events based on KG and find the breakpoints, to adjust production resources and reduce risks.

3. Capacity simulation model. First, obtain the real-time production status through the information model. Clustering and classification algorithms are used to extract the characteristics of key influencing factors and locate bottleneck production resources. Then, based on the production-grade DT system, the production capacity of each factory is predicted, and the impact of capacity reduction is quantified. Finally, analyze the production capacity of upstream and downstream core enterprises based on KG. Cross-enterprise, cross-industry, and cross-region collaborative analysis can be carried out through the industrial Internet platform to ensure that the output of each link meets the requirements.

3.2.3 Application and closed-loop iteration of the industrial chain DT model

The industrial chain DT model can carry out daily monitoring, risk assessment, and capacity analysis of the industrial chain. In the application layer, the industrial chain information is presented in real time through the visualized KG. The industrial chain risk assessment model evaluates the risk level based on real-time operation status data. When an emergency occurs or the risk level exceeds the set threshold, the system automatically sends a warning message. At the same time, the capacity simulation model analyzes the production capacity data of each enterprise in the industrial chain to identify capacity bottlenecks, and realizes the visual terminal display of capacity data. Then, interact with the industrial chain information model to find enterprises with matching production capacity and alternative products, to realize the rapid reconstruction of the industrial chain under emergency conditions.

After the results are obtained, the model is sent to each production-grade DT system for simulation verification and virtual execution. Compare the simulation results with the preset ones. If the results do not meet the requirements, multiple simulations and iterative analysis can be performed to adjust, optimize, and analyze the DT model. In addition, real-time status information and execution information are collected from the actual scene during execution, and then transmitted to the industry chain DT system for continuous analysis, simulation, evaluation, and optimization, thereby realizing virtual and physical mapping and interaction. Through the interaction and optimization iteration among the industry chain DT model, production-grade DT model, and physical space, the calibration and optimization of the industry chain DT system are realized, and the comprehensive management ability of the industry chain is improved.

4 Construction of the industry chain information model based on a knowledge graph

From Section 3, the industry chain information model is the key to establishing the industry chain DT system, which needs to quickly identify, calculate, organize, and adjust relationships. KG is an ideal technology for building the industry chain information model. It can describe complex relationships between entities through the expression of graph data and can organize and represent data from different dimensions. At the same time, the industrial Internet platform provides storage and expression capabilities for the KG data and the required algorithms and computing power for graph computing. Therefore, based on the industrial Internet platform and KG, the industry chain information model is constructed: (1) collect industry chain data and construct the KG ontology; (2) entity alignment of triples is performed and the structured data are stored in the graph database; (3) according to the characteristics of the collected domain data structure, joint entity–relation extraction of semi-structured and unstructured data is carried out; (4) aiming at the problems of imperfect content and lack of relationship, knowledge completion technology is used to supplement and improve the KG. The construction process of the industry chain information model is shown in Fig. 4.

4.1 Ontology of the industry chain KG

The industry chain KG includes three dimensions: product, enterprise, and supply–demand

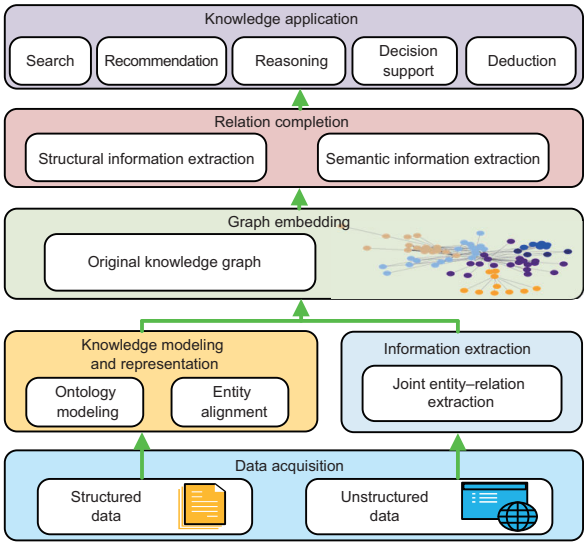

Fig. 4 Framework of industry chain information model construction

relationship. Each dimension contains multiple parameters. The industry chain is described based on these data, as shown in Fig. 5. The concepts of product, enterprise, and supply–demand in the industry chain are defined as $\alpha=\{P_i, B_i, T_i, O_i, S_i\}$, $\beta=\{E_i, N_i, C_i, Q_i, G_i\}$, and $\chi=\{P_i, E_{\mathrm{ui}}, E_{\mathrm{di}}, D_i, A_i\}$ respectively, as follows:

P_i is the unique identification of the product, and the industrial Internet identification analysis system is used for identification. B_i is the bill of material (BOM) factor, that is, the composition of the product. T_i is the technical factor, indicating whether the technology has been mastered. O_i is the production capacity factor, indicating the daily production capacity. S_i is the inventory factor, indicating the current inventory of the product.

E_i is the unique identification of the enterprise, and the industrial Internet identification resolution system is used to identify the enterprise. N_i is the registered information of the enterprise, including its name, registered capital, and legal person. C_i is the credit factor, indicating the credit rating of the enterprise, which depends mainly on the performance evaluation index. Q_i is the quality factor, which indicates the quality management level of the enterprise. G_i is the spatial factor, which represents the enterprise's geographical location information.

E_{ui} and E_{di} are upstream and downstream enterprise identification, respectively. D_i is the distance factor, indicating the supply distance. The upstream and downstream supply distances of the

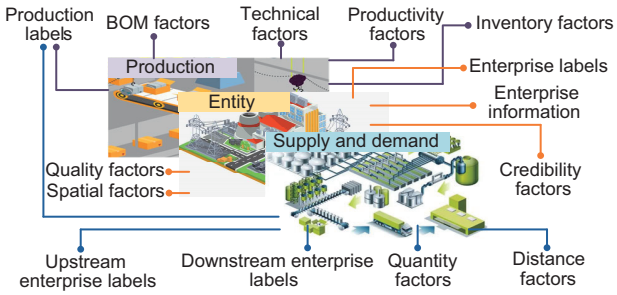

Fig. 5 Ontology framework of the industry chain KG

industry chain affect the scheduling of industrial resources. A_i is the quantity factor, that is, the daily demand quantity of the product.

4.2 Entity alignment

The purpose of entity alignment is to remove the inconsistency of entity names of enterprises, products, and people in heterogeneous data from multiple sources. The current solutions include the text similarity algorithm, thesaurus algorithm, and keyword co-occurrence algorithm. However, there is a large amount of repetitive relational data in KG of the industry chain, and it is difficult for traditional technologies to complete entity alignment efficiently and accurately. In this paper, an entity alignment method based on scientific and technological achievements is proposed. It includes keyword linking and industry entity linking.

1. Keyword linking. The purpose of keyword linking is to form a synonymous thesaurus in the industrial field and provide a basis for industrial knowledge labeling. The principle is to find potential synonyms based on similar phrases that appear in patents and papers, incorporating a small amount of expert evaluation to form the basis for entity alignment. The cosine similarity is used to measure the similarity between entities, in which the threshold is set to judge whether two entities are similar:

$$S(\boldsymbol{X}_i, \boldsymbol{X}_j) = \cos(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j}{\|\boldsymbol{X}_i\| \, \|\boldsymbol{X}_j\|}, \quad (1)$$

where S represents the similarity of entity names, and $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are the entity word vectors. The larger the S value between two entities, the more similar the two entity names.

2. Industry entity linking. On the basis of keyword linking, the related entity information is indexed in the domain. Combined with other

key information, whether the two records point to the same entity in the physical world is jointly determined.

4.3 Knowledge extraction of the industry chain based on the BERT-based multi-head selection model

Considering the professionalism of words in various fields of the industry chain and the advantages of the pre-trained language model, we use a multi-head selection model based on BERT to jointly extract entities and relationships. The model includes a BERT-based encoding layer, a conditional random field (CRF) based sequence labeling layer, and a multi-head selection based relation extraction layer, as shown in Fig. 6. The output of the CRF layer is the result of entity recognition with a BIO labeling strategy, where B represents the element at the beginning of the entity, I represents the element of the rest of the entity, and O represents a nonentity. In the relationship extraction layer, nonentity elements or entities that have no relationship with other entities are marked "N."

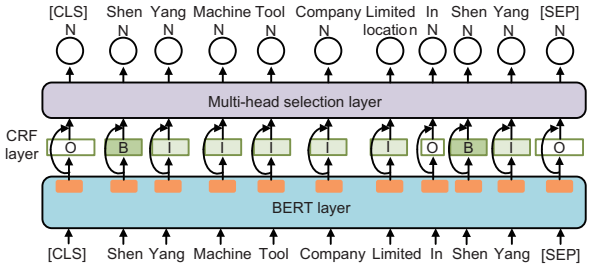

Fig. 6 BERT-based multi-head selection model for joint entity–relation extraction

The BERT model consists of a Transformer encoder (TE) structure stacked with 12 layers (Devlin et al., 2019). The TE includes a self-attention layer and a feed-forward network layer. Through the continuous stacking of multiple TE structures, BERT can output a high-level semantic representation combined with context information.

The self-attention layer is the core of the Transformer. Its main idea is to highlight different parts of the text semantic representation by assigning different weights, which is calculated as

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{soft}\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}, \quad (2)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ represent the input word vectors, and d is the dimension of the input.

To enable the model to learn relevant information in different representational subspaces, BERT uses a multi-head attention mechanism to perform multiple self-attention operations on each word in the input sequence, as follows:

$$\begin{aligned}&\text{Multihead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)\\&=\text{Concat}\left(\text{head}_1,\text{head}_2,...,\text{head}_n\right)\boldsymbol{W}^{\mathrm{O}},\end{aligned} \quad (3)$$

$$\text{head}_i=\text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right), \quad (4)$$

where $\boldsymbol{W}_i^{\boldsymbol{Q}}$, $\boldsymbol{W}_i^{\boldsymbol{K}}$, and $\boldsymbol{W}_i^{\boldsymbol{V}}$ are the weight matrices of head_i, $i=1,2,\cdots,n$, and $\boldsymbol{W}^{\mathrm{O}}$ represents an additional weight matrix.

The CRF layer is added on the BERT output layer to introduce dependencies between the labels and output the most probable label for each split. Assuming that the text word vector sequence is $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n)$ and that the label sequence is $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_n)$ (here, n represents the number of elements in the word vector sequence), the CRF scoring function is

$$S\left(\boldsymbol{X},\boldsymbol{Y}\right)=\sum_{i=1}^{n}\boldsymbol{x}_{i_{\boldsymbol{y}_i}}+\sum_{i=1}^{n-1}\boldsymbol{T}_{\boldsymbol{y}_i,\boldsymbol{y}_{i+1}}, \quad (5)$$

where $\boldsymbol{x}_{i_{\boldsymbol{y}_i}}$ represents the score of word vector $\boldsymbol{x}_i$ on label $\boldsymbol{y}_i$, and $\boldsymbol{T}_{\boldsymbol{y}_i,\boldsymbol{y}_{i+1}}$ is a state transition matrix representing the score from label $\boldsymbol{y}_i$ to $\boldsymbol{y}_{i+1}$.

The probability of the output label sequence can be obtained based on the softmax function:

$$P\left(\boldsymbol{Y}|\boldsymbol{X}\right)=\frac{\mathrm{e}^{S(\boldsymbol{X},\boldsymbol{Y})}}{\sum_{\tilde{\boldsymbol{Y}}\in Y_X}\mathrm{e}^{S(\boldsymbol{X},\tilde{\boldsymbol{Y}})}}, \quad (6)$$

where $\tilde{\boldsymbol{Y}}$ represents the possible tag sequences of sentence $\boldsymbol{X}$, and Y_X represents the set of all possible tag sequences for sentence $\boldsymbol{X}$. During training, the model aims to maximize the log probability function of the correct label sequence:

$$\log\left(P\left(\boldsymbol{Y}|\boldsymbol{X}\right)\right)=S\left(\boldsymbol{X}|\boldsymbol{Y}\right)-\log\left(\sum_{\tilde{\boldsymbol{Y}}\in Y_X}\mathrm{e}^{S(\boldsymbol{X},\tilde{\boldsymbol{Y}})}\right). \quad (7)$$

In the prediction stage, the formula for the output label sequence with the largest total probability is

$$L^*=\arg\max_{\boldsymbol{Y}\in Y_X}S\left(\boldsymbol{X},\tilde{\boldsymbol{Y}}\right). \quad (8)$$

The input of the relation extraction layer based on multi-head selection is $\boldsymbol{z}_i=[\boldsymbol{h}_i;\boldsymbol{x}_i]$, $i=$

$1, 2, ..., n$, where $\boldsymbol{h}_i$ is the output vector of the BERT model. The goal of the relationship extraction layer is to identify the most likely corresponding entity $\boldsymbol{x}_j$ and the most likely corresponding relationship $\boldsymbol{r}_{\mathrm{k}}$ for each word in the input word sequence $\boldsymbol{X}$, as follows:

$$s\left(\boldsymbol{z}_i, \boldsymbol{r}_{\mathrm{k}}, \boldsymbol{z}_j\right) = \boldsymbol{M} \cdot f\left(\boldsymbol{z}_i \boldsymbol{U} + \boldsymbol{z}_j \boldsymbol{G}\right), \tag{9}$$

where $f(\cdot)$ is an element-wise activation function, $\boldsymbol{M}$ is the relation embedding vector, and $\boldsymbol{U}$ and $\boldsymbol{G}$ are the linear transformation matrices of the head and tail entities, respectively.

The probability of $\boldsymbol{x}_j$ being selected as the head of $\boldsymbol{x}_i$ with the relation label $\boldsymbol{r}_{\mathrm{k}}$ between them is

$$P\left(\text{head} = \boldsymbol{x}_j, \text{label} = \boldsymbol{r}_{\mathrm{k}} | \boldsymbol{x}_i\right) = \sigma\left(s\left(\boldsymbol{z}_i, \boldsymbol{r}_{\mathrm{k}}, \boldsymbol{z}_j\right)\right), \tag{10}$$

where $\sigma(\cdot)$ represents the sigmoid function.

We minimize the cross-entropy loss L_{rel} during training as

$$L_{\mathrm{rel}} = \sum_{i=1}^{n} \sum_{j=1}^{m} -\log P\left(\text{head} = \boldsymbol{y}_{i,j}, \text{relation} = \boldsymbol{r}_{i,j} | \boldsymbol{x}_i\right), \tag{11}$$

where m is the number of relations for $\boldsymbol{x}_i$, $\boldsymbol{y}_{i,j}$ is the ground truth vector of the head, and $\boldsymbol{r}_{i,j}$ is the associated relation label of $\boldsymbol{x}_i$.

The construction of the model is based on multi-task training. On one hand, the sequence boundary position of entities is learned through the CRF loss function, and on the other hand, the interaction between entities and relations is enhanced through the multi-head selection loss function.

4.4 Relation completion of the knowledge graph based on a graph neural network

Due to the characteristics of KG, the structural characteristics between various entities imply unknown information. Therefore, it is necessary to identify and infer new relationships from existing data to supplement the knowledge in KG. To fully obtain the information of entities and relationships in KG and improve the accuracy of KG relation reasoning, we consider both the semantic information and graph structure information of KG. Then, the multi-head attention mechanism is used to combine the scoring results of structural and semantic information, as shown in Fig. 7.

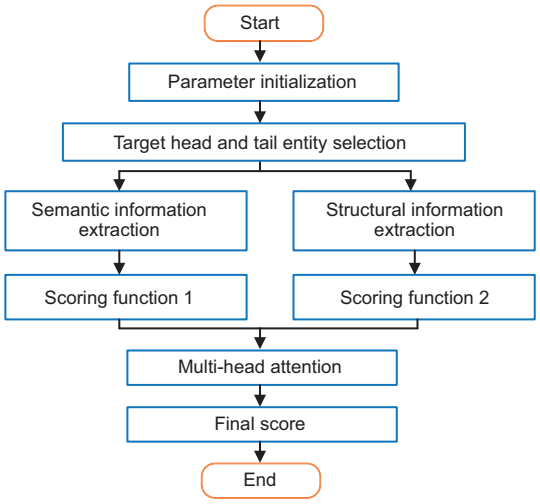

Fig. 7 Knowledge graph completion based on the semantic and structural information of the graph

4.4.1 Relation reasoning based on the semantic information KG

An R-GCN is used to obtain semantic information in KG. The message propagation model in R-GCN is as follows (Schlichtkrull et al., 2018):

$$\boldsymbol{h}_i^{l+1} = \sigma\left(\sum_{r \in \mathbb{R}} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} \boldsymbol{W}_r^l \boldsymbol{h}_j^l + \boldsymbol{W}_o^l \boldsymbol{h}_i^l\right), \tag{12}$$

where N_i^r is the set of subscripts of nodes that have a relationship $r \in \mathbb{R}$ with the i^{th} node in KG, $c_{i,r}$ is the normalized constant value for some particular problems, $\boldsymbol{h}_i^l$ represents a potential representation of node i in the l^{th} layer, and $\boldsymbol{W}_r^l$ and $\boldsymbol{W}_o^l$ are the relation-based weight matrices.

The score function of the predicted triple (s, r, o) is defined as

$$f_1(s, r, o) = \boldsymbol{e}_s^{\mathrm{T}} \boldsymbol{R}_r \boldsymbol{e}_o, \tag{13}$$

where $\boldsymbol{e}_s$ and $\boldsymbol{e}_o$ represent the head entity and tail entity respectively, and $\boldsymbol{R}_r$ is the relation matrix.

The model is trained based on sampled negative examples (Wang Y et al., 2023), considering ω negative samples, and optimized with cross-entropy loss:

$$\begin{aligned} l = -\frac{1}{(1+\omega)\left|\varepsilon'\right|} \sum_{(s,r,o,y) \in T_s} & \left[y_s \log \sigma\left(f_1(s, r, o)\right)\right. \\ & \left. + (1 - y_s) \log\left(1 - \sigma(f_1(s, r, o))\right)\right], \end{aligned} \tag{14}$$

where ε' is the entity sampling range, T_s is the sampling range of triples, and y_s is the label of the triple,

with $y_s = 1$ representing the positive example and $y_s = 0$ representing the negative example.

4.4.2 Relation reasoning based on the structure information KG

When predicting the relationship between two entities, entities that are far apart rarely carry semantic information that is helpful for relationship prediction. Therefore, the model needs to process the dataset in the original KG, and only the local subgraph between two selected entities is considered. As shown in Fig. 8, all relations between two target nodes whose path length is not greater than K form a subgraph. However, many entities in the graph are only unilaterally associated with the head entity or the tail entity, and they need to be eliminated.

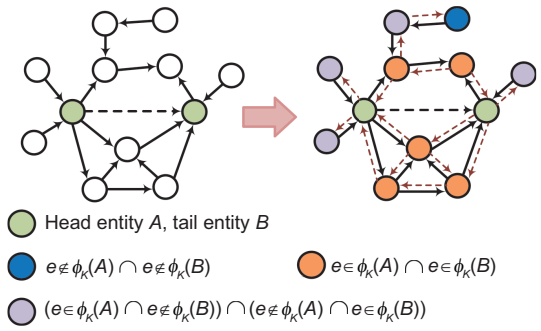

Fig. 8 Subgraph extraction

After finding the target nodes of the KG (the preparations for KG relation completion are provided in the supplementary materials), we convert the entities and relationships into a low-dimensional representation through GraphSAGE, and extract the information of the structure around the target nodes (Hamilton et al., 2017) as follows:

$$\boldsymbol{a}_u^{K'} = \text{Aggregate}^k\left(\left\{\boldsymbol{h}_{u'}^{k-1} : u' \in N(u), \boldsymbol{h}_v^{k-1}\right\}\right), \tag{15}$$

$$\boldsymbol{h}_u^k = \text{Concat}^k\left(\boldsymbol{h}_u^{k-1}, \boldsymbol{a}_u^{K'}\right), \tag{16}$$

where $\boldsymbol{a}_u^{K'}$ represents the aggregated information from neighborhood K' of node u, $N(u)$ represents the set of neighbors of node u, and $\boldsymbol{h}_u^k$ represents a potential representation of node u in the k^{th} layer, which is calculated by combining the information of the hidden layer of the previous layer and the aggregated information of this layer.

The model obtains the structural information of all layers of the target node through GraphSAGE. The final structural information value is the average value of all the node information in the last layer, as follows:

$$\boldsymbol{z}_{\vartheta_{(u,v,r_\text{t})}}^L = \frac{1}{|V|}\sum_{i\in V}\boldsymbol{h}_i^L, \tag{17}$$

where V represents the set of vertices in graph $\vartheta_{(u,v,r_\text{t})}$, r_t represents the relationship to be determined in the triplet, and L represents the number of layers of R-GCN.

Finally, the score function is used to predict the probability of relation r between entity pairs (u, v) as follows:

$$f_2(u,v,r_\text{t}) = \boldsymbol{W}^\text{T}\left[\boldsymbol{z}_{\vartheta_{(u,v,r_\text{t})}}^L \oplus \boldsymbol{h}_u^L \oplus \boldsymbol{h}_v^L \oplus \boldsymbol{e}_{r_\text{t}}\right], \tag{18}$$

where $\boldsymbol{e}_{r_\text{t}}$ represents the KG embedding of the target relationship, and "$\oplus$" is the direct sum symbol.

In this paper, the multi-head attention mechanism is used to combine the results to reasonably reflect the semantic information and structural information in the KG. The calculation of the multi-head attention is shown in Eqs. (2)–(4).

Based on Eqs. (13) and (18), the semantic information score and structural information score of graph relationship reasoning are combined:

$$f = \sum_{i=1}^{c}\alpha_i f_i, \tag{19}$$

where c is the number of entities in KG, and α_i is the weight coefficient (the calculation method for α_i is provided in the supplementary materials).

5 Experiments and analysis

In this section, the knowledge extraction method and relation completion method proposed in this paper are verified. In addition, the industry chain information model in the field of basic machinery is established and analyzed to verify the feasibility of the research method. This research relies on the national-level industrial foundation innovation platform, which has the characteristics of open standards, wide application, high safety, and high reliability. The platform has a complete acquisition, sharing, and application mechanism for big data. In addition, in terms of data privacy protection and data confirmation regulation, it has comprehensive big-data resource management and security service capabilities.

5.1 Analysis of knowledge extraction and relation completion of the knowledge graph

The data are provided by a government department. They are divided into two datasets, namely dataset *A* and dataset *B*. Each dataset contains a corpus and its corresponding KG in vector format (detailed descriptions of datasets are provided in the supplementary materials).

To verify the effect of knowledge extraction, the BERT-based multi-head selection model used in this paper is compared with the LSTM-CRF model, BiLSTM-CRF model, BERT model, and BERT-CRF model. The input of the entity–relation extraction model is the sentences in the corpus, and the output is the triples composed of the head entity, tail entity, and relation (the data preprocessing method for knowledge extraction is provided in the supplementary materials).

The main parameter information of the model in the experiments is determined by training the model and adjusting the parameters constantly. The parameters in the experiments are shown in Table 1. The batch size and learning rate in the experiments are the optimal indicators found on the training set by the grid search. The maximum text length depends on the corpus. The BERT model has 12 hidden layers, and the vector has 769 dimensions. In this paper, five-fold cross-validation is used to evaluate the accuracy of models. Tables 2 and 3 show the average of precision (P), recall (R), and F value (F) of each model on the test set of datasets A and B, respectively.

Table 1 Description of the parameters in the experiments

Parameter	Value
Batch size	32
Maximum sentence length	128
Optimizer algorithm	Adam
Learning rate	2×10^{-5}
Number of layers of Transformer	12

As shown in Tables 2 and 3, BERT is obviously superior to LSTM-CRF and BiLSTM-CRF. Compared with BiLSTM-CRF, under datasets A and B, the F values of BERT are increased by 4.60% and 3.61%, respectively. This is because the BERT model has strong semantic expression ability, and the encoded semantic vector contains more information. Therefore, the BERT model can fully extract the features of the relationship between characters, words, and even sentences, which enhances the generalization ability of the model. This gives the BERT model obvious advantages in knowledge extraction tasks. In addition, after the BERT model is added to the CRF layer, compared with the BERT model, the F values under datasets A and B are increased by 0.30% and 0.16%, respectively. Compared with the BERT-CRF model, the F values of the proposed model are increased by 0.10% and 0.62% respectively, which shows that BERT has better knowledge extraction performance after adding the CRF and multi-head selection mechanism. This is because the model not only fully extracts text features, but also learns the sequence boundary positions of entities through the CRF. In addition, the model enhances the interaction between entities and relationships using the multi-head selection algorithm, improving the performance of entity and relationship recognition. Therefore, the BERT-based multi-head selection algorithm for knowledge extraction can effectively take advantage of the BERT model.

Table 2 Comparison of the entity–relation extraction on dataset *A*

Model	P (%)	R (%)	F (%)
LSTM-CRF	76.33	76.56	75.36
BiLSTM-CRF	77.62	78.87	76.14
BERT	79.02	81.08	79.64
BERT-CRF	78.94	81.57	79.88
BERT-based multi-head selection	**79.13**	**81.67**	**79.96**

Best results are in bold

Table 3 Comparison of the entity–relation extraction on dataset *B*

Model	P (%)	R (%)	F (%)
LSTM-CRF	77.88	78.47	76.66
BiLSTM-CRF	79.39	80.28	77.24
BERT	82.02	82.12	80.03
BERT-CRF	82.23	82.76	80.16
BERT-based multi-head selection	**82.44**	**82.95**	**80.66**

Best results are in bold

In this paper, comparative experiments are also designed to verify the advantages of the adopted relation completion model of the KG. The experiments compare the early classic KG embedding model TransE (Bordes et al., 2013) and the first model DistMult (Yang et al., 2015) using the sum evaluation index scoring function, and also compare

recent innovative models in semantics and structure, such as ComplEx (Trouillon et al., 2016), ConvE (Dettmers et al., 2018), RotatE (Sun Z et al., 2019), R-GCN (Schlichtkrull et al., 2018), A2N (Bansal et al., 2019), and CompGCN (Vashishth et al., 2020).

The relationship completion task is to infer the missing relationship (r) based on the structure of the existing triples (h, r, t) in KG to supplement KG. The input of the relation completion model is the graph embedding vector of KG. The output is the inferred relation (the data preprocessing method for relation completion is provided in the supplementary materials). The experimental results are shown in Tables 4 and 5.

Table 4 Comparison of the relation completion on dataset *A*

Model	MRR	Hit@1	Hit@3	Hit@10
TransE	0.285	0.193	0.283	0.457
DistMult	0.234	0.150	0.255	0.406
ComplEx	0.239	0.153	0.268	0.415
ConvE	0.315	0.230	0.345	0.486
RotatE	0.328	0.233	0.364	0.517
R-GCN	0.241	0.162	0.296	0.404
A2N	0.307	0.225	0.338	0.471
CompGCN	0.334	0.246	**0.378**	0.509
Ours	**0.348**	**0.261**	0.376	**0.525**

Best results are in bold, and the second-best results are with the underline

Table 5 Comparison of the relation completion on dataset *B*

Model	MRR	Hit@1	Hit@3	Hit@10
TransE	0.290	0.128	0.268	0.525
DistMult	0.241	0.184	0.331	0.534
ComplEx	0.294	0.261	0.314	0.526
ConvE	0.325	0.197	0.421	0.520
RotatE	0.314	0.156	0.315	0.522
R-GCN	0.259	0.166	0.271	0.452
A2N	0.308	0.223	0.347	0.511
CompGCN	0.349	0.236	0.440	0.592
Ours	**0.381**	**0.282**	**0.451**	**0.621**

Best results are in bold, and the second-best results are with the underline

The mean reciprocal rank (MRR) and Hit@K (K=1, 3, and 10) are taken as the evaluation metrics in this paper. MRR represents the average reciprocal rank of all the real candidates, and Hit@K records the proportion of the valid test triplet rank in top K predictions. It can be seen from Tables 4 and 5 that the model adopted in this paper achieves the best performance on most of the indicators, which shows that the proposed model has the strongest representation learning ability on KG. Specifically, on dataset A, the proposed model achieves improvements of 4.19% in MRR (Liu YSY et al., 2022), 6.10% in Hit@1 (Dai et al., 2022), and 1.55% in Hit@10 compared with the second-best results. On dataset B, the proposed model achieves improvements of 9.17% in MRR, 8.04% in Hit@1, 2.50% in Hit@3, and 4.90% in Hit@10 compared with the second-best results. This demonstrates the effectiveness of the proposed model in simultaneously learning structural and semantic information in KG. In addition, as shown in the tables, the proposed model is significantly better than the R-GCN model under these two datasets. This is because R-GCN collects only the semantic information of the nodes around the target entity, while the model proposed in this paper adds the improved GraphSAGE to realize the extraction of structural information. The proposed model refines the representation of entities in different semantic spaces through different base graphs and distinguishes the semantic information in different sub-graph structures, thereby improving the relation reasoning performance.

5.2 Case study

In this subsection, 18 industry chains in the field of basic machinery are modeled and analyzed on the developed platform to verify the feasibility of the research method (a detailed description of the data sources is provided in the supplementary materials). The original industry chain KG is constructed by using the above triples, containing 12 211 entities and 18 902 relationships.

Due to omissions in the knowledge extraction process, the relationships between entities in KG are partially missing, and some implicit relationships have yet to be discovered. Therefore, the model proposed in this paper is used to combine semantic information and structural information to complete the KG relationship. The experimental results show that by learning existing link relationships, the model proposed in this paper predicts 2257 new links, and the total number of relationships reaches 21 159, which is 11.94% more than that of the original graph. Finally, Neo4j is used as the graph database and integrated with the front-end visualization framework to complete the development of the platform.

The industry chain information model based on KG can monitor relevant data of the industrial

chain in real time, and update the industrial chain database in real time to provide services in the latest state. In addition, the model can represent information such as the structural form of the industry chain, products in key links, and supplier dependence, thereby assisting decision-makers in adjusting the development direction of the industry in time.

Fig. 9 shows the industry panorama of the 18 industry chains involved in the basic machinery field. The nodes of different colors represent different enterprises, and the larger the node, the larger the scale of the enterprise. The industry panorama visually presents the analysis results of relevant data, and divides enterprises based on regional conditions and advantages. It shows the distribution and relationship of enterprises, the numbers of industry chains, enterprises, accumulated assets, and operating income involved in the basic machinery field. A partial node view of the panorama of industry is shown in Fig. 10. Each node represents an enterprise, and the connection between enterprises represents the upstream or downstream relationship.

The platform built in this paper can display an industry chain portrait based on the industry chain information model, as shown in Fig. 11. The industry chain portrait focuses on each sub-sector, showing its downstream enterprises and related products. In addition, it shows the operation of the industry chain, including assets, income, profit, and growth rate. The industry chain portrait shows the important characteristics of the industry in the form of radar charts, including the number of suppliers, industry scale, industry influence, state-owned asset coverage ratio, and dependence risks. Moreover, it presents key manufacturers and products in the industry in the form of a bubble chart, and the bigger the bubble, the more important it is. The size of the bubble is based on assets, revenue, profit, and total

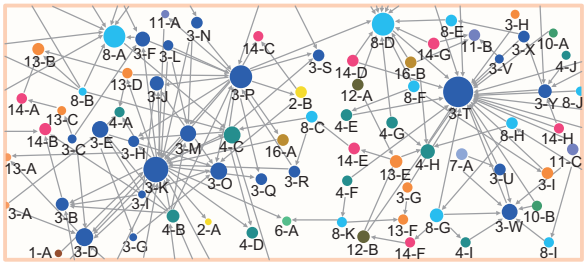

Fig. 10 Partial view of the industry panorama

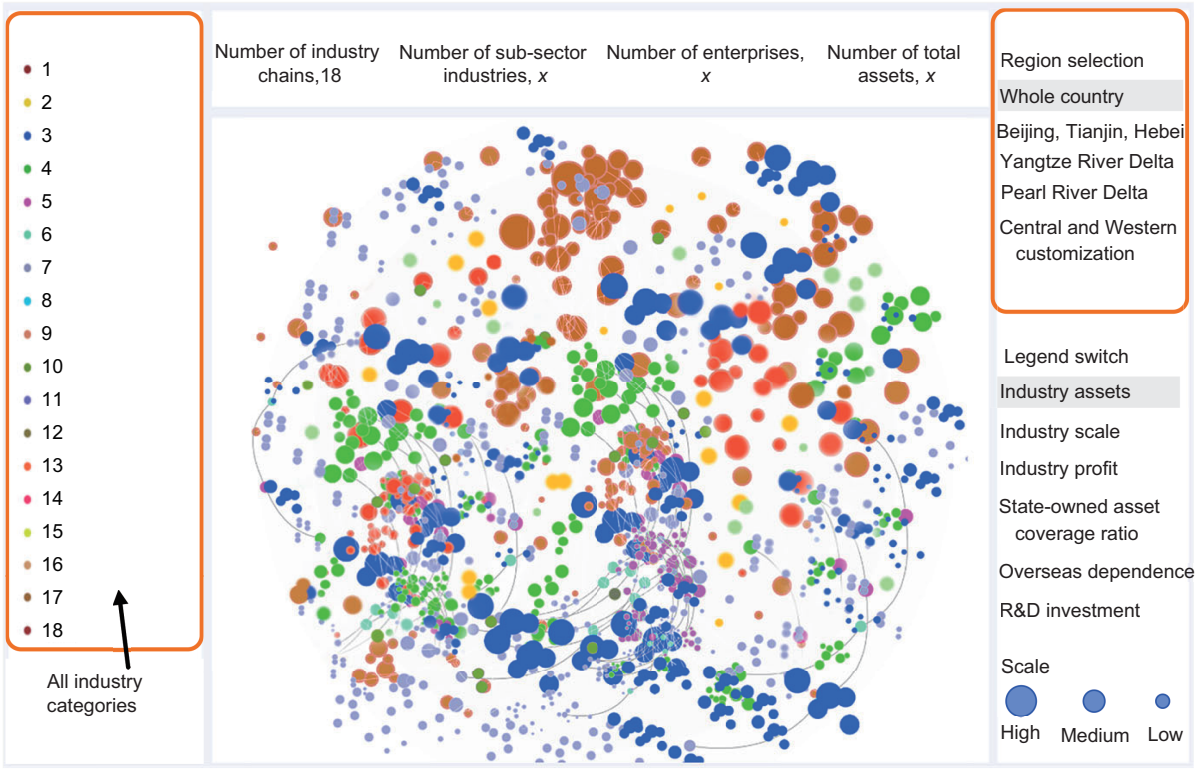

Fig. 9 Industry panorama of the basic machinery field

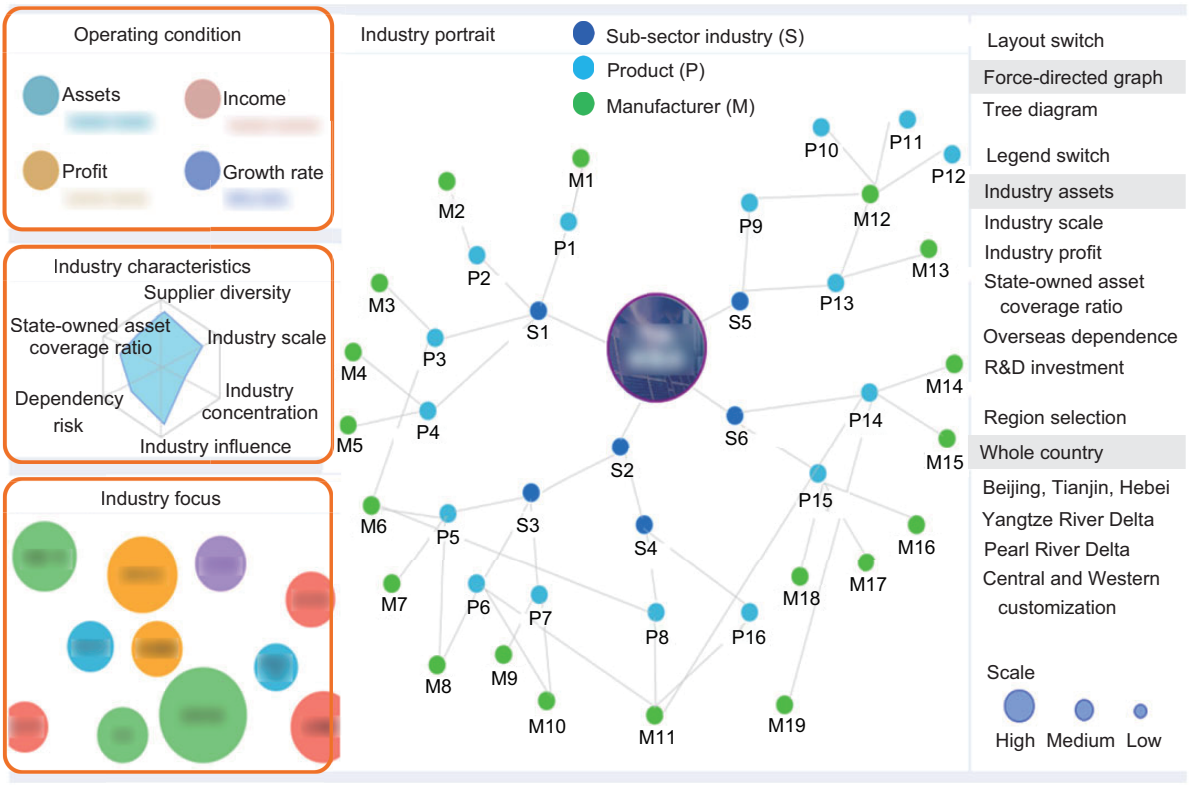

Fig. 11 Industry chain portrait

expenditure of research and development. Therefore, the structure of the industry chain in the industry chain portrait is clear and focused.

Based on the above analysis, the industry chain information model based on KG can combine multi-source data to mine the upstream and downstream correlations of the industry, and display key industrial clusters, industry fields, and macro development trends in multiple dimensions. It is conducive to realizing the integrated development of the industry chain and innovation chain and promoting the development of the industrial digital economy.

6 Conclusions

To improve the ability to collaboratively innovate, avoid risk, and comprehensively govern each link of the industrial Internet, we propose an industrial chain DT system framework and an industrial chain information model based on KG. This paper contributes to industrial chains in three aspects:

1. The construction process of the industrial chain DT system framework supported by the industrial Internet was presented, including data collection, industrial chain DT model construction, and the application of the model. The role of the industrial chain information model and its relationship with the risk assessment model and the capacity simulation model were analyzed. In addition, the iterative optimization method of the industrial chain DT system was discussed.

2. A KG-based industry chain information model construction method was proposed, including the joint entity–relation extraction method based on the BERT-based multi-head selection model and the relationship completion model based on the graph neural network. The models were validated on two industrial chain datasets. Experimental results showed that compared with baseline models, the proposed model achieved better performance on most indicators.

3. Based on the data of 18 industrial chains, an industrial chain information model in the field of basic machinery was established. A visualized industrial panorama was displayed, in which the characteristics of the enterprises were reflected through attributes such as color and size of the nodes. In addition, an industry chain portrait was displayed

which shows the operation of the industry chain through KG. From the above two aspects, the application scenarios and operation mechanism of the industrial chain information model were explained, which proves that the model is an effective tool for industrial chain monitoring and optimization.

The industry chain DT system proposed in this paper includes three models, but only the industry chain information model has been established so far. In future work, we plan to establish a risk assessment model to identify breakpoints and blockages, discover potential risk factors, and analyze weak links in the industry chain. In addition, a production capacity analysis model will be established. It will be used to determine the possibility of production capacity realization after changes in production capacity demand through simulation and deduction analysis, and the model can discover the key path and bottleneck link of production capacity realization. In this way, managers can implement scheduling, resource organization, and capacity planning in advance.

Contributors

Wenxuan WANG and Lin ZHANG designed the research. Wenxuan WANG and Yongqin LIU processed the data. Yongqin LIU and Xudong CHAI drafted the paper. Wenxuan WANG helped organize the paper. Yongqin LIU, Xudong CHAI, and Lin ZHANG revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The publicly available data from companies on the Internet that support the findings of this study can be provided, and are available from the corresponding author upon reasonable request. However, the internal data from companies and the Ministry of Industry and Information Technology of China cannot be shared due to confidentiality.

References

Al Faruque MA, Muthirayan D, Yu SY, et al., 2021. Cognitive digital twin for manufacturing systems. Proc Design, Automation & Test in Europe Conf & Exhibition, p.440-445. https://doi.org/10.23919/DATE51398.2021.9474166

Bansal T, Juan DC, Ravi S, et al., 2019. A2N: attending to neighbors for knowledge graph inference. Proc 57th Annual Meeting of the Association for Computational Linguistics, p.4387-4392. https://doi.org/10.18653/v1/P19-1431

Bao JS, Guo DS, Li J, et al., 2019. The modelling and operations for the digital twin in the context of manufacturing. *Enterp Inform Syst*, 13(4):534-556. https://doi.org/10.1080/17517575.2018.1526324

Bordes A, Usunier N, Garcia-Durán A, et al., 2013. Translating embeddings for modeling multi-relational data. Proc 26th Int Conf on Neural Information Processing Systems, p.2787-2795.

Chen HT, Jeremiah SR, Lee C, et al., 2023. A digital twin-based heuristic multi-cooperation scheduling framework for smart manufacturing in IIoT environment. *Appl Sci*, 13(3):1440. https://doi.org/10.3390/app13031440

Cheng JF, Zhang H, Tao F, et al., 2020. DT-II: digital twin enhanced Industrial Internet reference framework towards smart manufacturing. *Robot Comput Integr Manuf*, 62:101881. https://doi.org/10.1016/j.rcim.2019.101881

Dai GQ, Wang XZ, Zou XY, et al., 2022. MRGAT: multi-relational graph attention network for knowledge graph completion. *Neur Netw*, 154:234-245. https://doi.org/10.1016/j.neunet.2022.07.014

Dettmers T, Minervini P, Stenetorp P, et al., 2018. Convolutional 2D knowledge graph embeddings. https://arxiv.org/abs/1707.01476

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional Transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Guo L, Yan F, Li T, et al., 2022. An automatic method for constructing machining process knowledge base from knowledge graph. *Robot Comput Integr Manuf*, 73:102222. https://doi.org/10.1016/j.rcim.2021.102222

Hamilton WL, Ying R, Leskovec J, 2017. Inductive representation learning on large graphs. Proc 31st Int Conf on Neural Information Processing Systems, p.1025-1035.

Hedberg TDJr, Bajaj M, Camelio JA, 2020. Using graphs to link data across the product lifecycle for enabling smart manufacturing digital threads. *J Comput Inform Sci Eng*, 20(1):011011. https://doi.org/10.1115/1.4044921

Jia WJ, Wang W, Zhang ZZ, 2022. From simple digital twin to complex digital twin part I: a novel modeling method for multi-scale and multi-scenario digital twin. *Adv Eng Inform*, 53:101706. https://doi.org/10.1016/j.aei.2022.101706

Kamble SS, Gunasekaran A, Parekh H, et al., 2022. Digital twin for sustainable manufacturing supply chains: current trends, future perspectives, and an implementation framework. *Technol Forecast Soc Change*, 176:121448. https://doi.org/10.1016/j.techfore.2021.121448

Kiel D, Arnold C, Voigt KI, 2017. The influence of the Industrial Internet of Things on business models of established manufacturing companies—a business level perspective. *Technovation*, 68:4-19. https://doi.org/10.1016/j.technovation.2017.09.003

Kong TX, Hu TL, Zhou TT, et al., 2021. Data construction method for the applications of workshop digital twin system. *J Manuf Syst*, 58:323-328. https://doi.org/10.1016/j.jmsy.2020.02.003

Li BH, Hou BC, Yu WT, et al., 2017. Applications of artificial intelligence in intelligent manufacturing: a review. *Front Inform Technol Electron Eng*, 18(1):86-96. https://doi.org/10.1631/FITEE.1601885

Liu MF, Li XY, Li J, et al., 2022. A knowledge graph-based data representation approach for IIoT-enabled cognitive manufacturing. *Adv Eng Inform*, 51:101515. https://doi.org/10.1016/j.aei.2021.101515

Liu YSY, Gu F, Gu XJ, et al., 2022. Resource recommendation based on industrial knowledge graph in low-resource conditions. *Int J Comput Intell Syst*, 15(1):42. https://doi.org/10.1007/s44196-022-00097-2

Lu YQ, Xu X, 2019. Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services. *Robot Comput Integr Manuf*, 57:92-102. https://doi.org/10.1016/j.rcim.2018.11.006

Menon K, Kärkkäinen H, Wuest T, et al., 2019. Industrial Internet platforms: a conceptual evaluation from a product lifecycle management perspective. *Proc Inst Mech Eng B*, 233(5):1390-1401. https://doi.org/10.1177/0954405418760651

Min QF, Lu YG, Liu ZY, et al., 2019. Machine learning based digital twin framework for production optimization in petrochemical industry. *Int J Inform Manag*, 49:502-519. https://doi.org/10.1016/j.ijinfomgt.2019.05.020

Qin W, Chen SQ, Peng MG, 2020. Recent advances in Industrial Internet: insights and challenges. *Dig Commun Netw*, 6(1):1-13. https://doi.org/10.1016/j.dcan.2019.07.001

Ren L, Li YJ, Wang XK, et al., 2023. An ABGE-aided manufacturing knowledge graph construction approach for heterogeneous IIoT data integration. *Int J Prod Res*, 61(12):4102-4116. https://doi.org/10.1080/00207543.2022.2042416

Sarazin A, Bascans J, Sciau JB, et al., 2021. Expert system dedicated to condition-based maintenance based on a knowledge graph approach: application to an aeronautic system. *Expert Syst Appl*, 186:115767. https://doi.org/10.1016/j.eswa.2021.115767

Schlichtkrull M, Kipf TN, Bloem P, et al., 2018. Modeling relational data with graph convolutional networks. 15th Int Conf on the Semantic Web, p.593-607. https://doi.org/10.1007/978-3-319-93417-4_38

Sun XM, Bao JS, Li J, et al., 2020. A digital twin-driven approach for the assembly-commissioning of high precision products. *Robot Comput Integr Manuf*, 61:101839. https://doi.org/10.1016/j.rcim.2019.101839

Sun Z, Deng ZH, Nie JY, et al., 2019. RotatE: knowledge graph embedding by relational rotation in complex space. https://arxiv.org/abs/1902.10197v1

Tan YP, Xu HF, Yan D, et al., 2021. Automatic construction of knowledge graph and its application in electric power system. 3rd Asia Energy and Electrical Engineering Symp, p.725-729. https://doi.org/10.1109/AEEES51875.2021.9403122

Trouillon T, Welbl J, Riedel S, et al., 2016. Complex embeddings for simple link prediction. Proc 33rd Int Conf on Machine Learning, p.2071-2080.

Vashishth S, Sanyal S, Nitin V, et al., 2020. Composition-based multi-relational graph convolutional networks. Proc 8th Int Conf on Learning Representations.

Wang JL, Xu CQ, Zhang J, et al., 2020. A collaborative architecture of the industrial Internet platform for manufacturing systems. *Robot Comput Integr Manuf*, 61:101854. https://doi.org/10.1016/j.rcim.2019.101854

Wang Y, Hu L, Gao WF, et al., 2023. AdaNS: adaptive negative sampling for unsupervised graph representation learning. *Patt Recogn*, 136:109266. https://doi.org/10.1016/j.patcog.2022.109266

Yang BS, Yih WT, He XD, et al., 2015. Embedding entities and relations for learning and inference in knowledge bases. Proc 3rd Int Conf on Learning Representations, p.141-153.

Yin S, Zhang N, Dong HM, 2020. Preventing COVID-19 from the perspective of industrial information integration: evaluation and continuous improvement of information networks for sustainable epidemic prevention. *J Ind Inform Integr*, 19:100157. https://doi.org/10.1016/j.jii.2020.100157

Zhang L, Zhou LF, Horn BKP, 2021. Building a right digital twin with model engineering. *J Manuf Syst*, 59:151-164. https://doi.org/10.1016/j.jmsy.2021.02.009

Zhang XY, Ming XG, Bao YG, et al., 2022. System construction for comprehensive industrial ecosystem oriented networked collaborative manufacturing platform (NCMP) based on three chains. *Adv Eng Inform*, 52:101538. https://doi.org/10.1016/j.aei.2022.101538

Zhou B, Shen XW, Lu YQ, et al., 2023. Semantic-aware event link reasoning over industrial knowledge graph embedding time series data. *Int J Prod Res*, 61(12):4117-4134. https://doi.org/10.1080/00207543.2021.2022803

Zhou GH, Zhang C, Li Z, et al., 2020. Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing. *Int J Prod Res*, 58(4):1034-1051. https://doi.org/10.1080/00207543.2019.1607978

Zhuang CB, Liu JH, Xiong H, 2018. Digital twin-based smart production management and control framework for the complex product assembly shop-floor. *Int J Adv Manuf Technol*, 96(1-4):1149-1163. https://doi.org/10.1007/s00170-018-1617-6

List of supplementary materials

1 Three data sources of the industry chain DT system

2 Preparations for KG relation completion

3 Calculation method for α_i in Eq. (19)

4 Detailed descriptions of datasets in Section 5.1

5 Data preprocessing for knowledge extraction

6 Data preprocessing for relation completion

7 Detailed descriptions of data sources in Section 5.2