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

Supplementary materials for

Wenxuan WANG, Yongqin LIU, Xudong CHAI, Lin ZHANG, 2024. Digital twin system framework and information model for industry chain based on industrial Internet. *Front Inform Technol Electron Eng*, 25(7):951-967. https://doi.org/10.1631/FITEE.2300123

1 Three data sources of the industry chain DT system

1. Internet data collection: By mining the Internet information, the system can accurately identify industries and enterprise information to ensure the real-time collection and update of information. For the public data that does not have commercial application programming interface (API) service or services that do not meet the requirements, crawler tools are used to obtain data. The system regularly and continuously crawls data in the cloud, and then pushes it back to the business system, thereby improving the efficiency of data crawling and ensuring the quality and quantity of data. Moreover, the distributed crawler design architecture is adopted to avoid the risk of a potential single point of failure.

2. Enterprise internal data access: Enterprise internal data includes demand formulation, design, production research and development, logistics, operation, etc. The system achieves vertical integration among manufacturing resources, on-site control, workshop execution, enterprise management, industrial cloud platform, and various systems (such as ERP, supervisory control and data acquisition (SCADA), CRM systems, etc.) through digitization, networking, intelligence, and visualization. In this way, the supply, research and development, production, and service processes of products are taken into account to realize data-driven intelligent production and efficient production collaboration control.

3. Industrial system and equipment data access: Based on 5G, software-defined network (SDN), Internet protocol version 6 (IPV6) technologies, and a new generation of smart gateways and smart sensors, the access of heterogeneous devices across industries and domains is realized. The industrial data acquisition interface (the downlink API interface of the industrial cloud platform) is suitable for integrated access of industrial equipment and the industrial Internet, including data interfaces and identification interfaces. The data interface includes file transfer service, time-series data service, and data collection. The identification interface includes gateway identification, equipment asset identification, and identity identification.

2 Preparations for KG relation completion

Specifically, the model takes (A, X, B) as the triple of the relationship to be predicted. First, the model starts to perform calculations from the head entity A and the tail entity B, and finds all entities within the respective step size K to obtain sets $\phi_K(A)$ and $\phi_K(B)$. Second, take the intersection of $\phi_K(A)$ and $\phi_K(B)$, and remove the entities that are only connected to the head entity or tail entity. Then, obtain the set of all entities included in the subgraph as $\{A, B, \phi_K(A) \cap \phi_K(B)\}$. Finally, keep the interconnected edges between the entities in this set, and obtain the final subgraph $\phi(Z)$.

After obtaining the graph structure, to facilitate the calculation of the graph neural network, the entities in the extracted structure are marked using method of double labeling, Any node in the subgraph around the target nodes u and v is represented as $(d(i,u), d(i,v))$, where $d(i,u)$ is the shortest distance from node i to target node u. In this way, the relative positions of other nodes and the target node in the extracted graph structure are obtained.

3 Calculation method for α_i in Eq. (19)

The weights of semantic information and structural information of KG are

$$\begin{cases} a_1 = \text{soft}\max\left(\dfrac{X_a X_a^T}{\sqrt{d}}\right), \\ a_2 = \text{soft}\max\left(\dfrac{\xi\xi^T}{\sqrt{d}}\right), \end{cases} \tag{S1}$$

where X_a is the matrix composed of entity vectors in the KG, ξ is the degree matrix, and d is the distance between nodes.

The weights of the semantic information score and structural information score can be expressed as

$$\alpha_i = \frac{\exp(a_i)}{\sum_{i=1}^{c}\exp(a_i)}, \tag{S2}$$

where c is the number of entities in the KG.

4 Detailed descriptions of datasets in Section 5.1

The two datasets consist of some unstructured data such as product reports, research reports, financial reports, and Internet news from 2010 to 2020, involving 40 industrial chains. The corpus of data set A contains 36 733 sentences and 75 215 triples, and the corresponding KG contains 18 761 entities and 28 304 relationships. The corpus of dataset B contains 13 019 sentences and 28 130 triples, and the corresponding KG contains 5937 entities and 9135 relationships.

5 Data preprocessing for knowledge extraction

Before the experiment of knowledge extraction, the data in the corpus are preprocessed, including character replacement and noise processing. Character replacement is designed to standardize the sentences, that is, to ensure the unity of uppercase and lowercase characters, traditional and simplified characters, and digital Chinese characters in the sentences and triplets. Noise processing mainly includes the unification of the entity format; for example, the date format is unified as "year, month, and day," and the company name needs to be complete. Then the standard data partition method is used to divide the corpus in the data sets A and B into training datasets and test datasets.

6 Data preprocessing for relation completion

Before the relation completion experiment, the link relation in KG of the vector format in datasets A and B is hidden by 15%. The graph embedding vector consisting of all entities and 85% of the link relation is used as the training set. The graph embedding vector consisting of 15% of the link relation and the corresponding entities is used as the test set. Mean reciprocal rank (MRR), Hit@1, Hit@3, and Hit@10 are taken as the evaluation metrics in this paper. MRR represents the average reciprocal rank of all real candidates. Hit@K records the proportion of the valid test triplet ranking in top K predictions.

7 Detailed descriptions of data sources in Section 5.2

This research relies on the "Edge Gateway" project of the Ministry of Industry and Information Technology, and conducts research based on the industry Internet platform of China Aerospace Science & Industry Cloud Co., Ltd. The platform is directly connected to more than 70 enterprises in the field of basic machinery. At the same time, the platform additionally collects relevant data of more than 200 companies based on

the Internet and other means. The data sources include two parts. One part is the structured data collected and cleaned extensively by the industrial Internet platform, and the other part is the data directly obtained by the data layer from the Internet, enterprises, and industrial systems. The second part of the data includes structured data (such as data in enterprise databases and data provided by governments) and unstructured data (such as product reports and Internet news). The structured data can be converted into triples by setting the mapping relationship between the data source and the ontology model. For unstructured data, the BERT-based multi-head selection model proposed in this paper is used to extract triples from the data.