Research on Fusion of Electric Power Industry and External Data Based on Knowledge Map

Libo Fan, Yi Wang*, Xiang Fang, Jian Jiang
Yi Wang: ruth_wang@hotmail.com

Libo Fan: huoqiupsx8@163.com

Xiang Fang: kuih25630@163.com

Jian Jiang: lvu6830@163.com
State Grid Hangzhou Power Supply Company, Hangzhou, Zhejiang 310000, China,

Research on Key Technologies of Data Governance for enterprise-level Power Brain Center System Based on Data Center (5211HZ2001WF)

Abstract—Aiming at the problem of multi-source heterogeneous data fusion and utilization between the power industry and the outside, an application idea of internal and external data fusion based on knowledge map is proposed. Through the analysis of the existing mainstream technical routes of text information extraction and knowledge map construction, this paper puts forward an optimized text information extraction scheme and knowledge map construction scheme, and realizes the support of multi-source heterogeneous data fusion modeling. The research results have been applied in a provincial electric power company for risk identification in the field of engineering project process management, and the feasibility of the proposed technical scheme has been verified.

Keywords-knowledge map; Electric power industry; External data fusion

1 INTRODUCTION

Although the company is actively developing the application of big data research, and has started a number of applied research projects, the overall data application is limited, and most of the majors are confined to the professional category. The integration of all kinds of data across majors and businesses is obviously insufficient, and the potential value of a large amount of data has not been fully tapped. Therefore, it is urgent to plan and control the enterprise data in a unified way from the overall situation of the enterprise business, so as to promote the integration of enterprise data in data, system and business, thus providing a logical and unified data support for all business applications of the company, providing complete data resources, efficient analysis and calculation ability and a unified running environment for all kinds of analysis and decision-making applications of the company, and supporting the comprehensive development of enterprise-level data analysis and application[1].

The company will carry out the research on big data fusion technology of operation monitoring business based on business system. By analyzing the current situation and existing problems of data fusion of transportation supervision business, it will build a data demand framework and

ICICA 2022, December 02-04, Chongqing, People's Republic of China

DOI 10.4108/eai.2-12-2022.2327981

establish a data fusion technology system. Based on the business system, it will solve the problems of cross-professional data fusion and application in combination with the actual needs of transportation supervision business management and application, so as to promote the further mining of data value and lay a technical foundation for the company's big data application [2].

2 KNOWLEDGE MAP TECHNOLOGY

Knowledge map is a graph data structure composed of nodes and edges, which is essentially a structured semantic knowledge base. By connecting the multi-source heterogeneous information in the professional business field, the entity relationship network is obtained, and the knowledge map can provide a powerful ability to analyze problems from the perspective of "relationship". At first, the knowledge map was mainly used to solve intelligent question-answering problems related to entities, and now it has been widely used in the fields of information retrieval and data mining. In the electric power industry, the ability of relational calculation and graph mining based on knowledge map, combined with data mining methods such as anomaly analysis, static analysis and dynamic analysis, can be used in the fields of anti-fraud, inconsistency verification, power grid equipment failure analysis, disaster prevention and early warning, master data quality optimization, etc[3].

The construction of knowledge map can be roughly divided into two main steps. Firstly, combined with the business knowledge in related fields, the key conceptual entities in business fields and their relationships are described as knowledge map ontology. Then, using the technologies of entity extraction, entity disambiguation and linking, entity relationship extraction, knowledge reasoning, etc., the information fields related to entities and relationships are extracted from the actual business data, disambiguated and fused, and then "filled" according to the ontology of knowledge map, so as to obtain knowledge map data examples and store the knowledge map. The main challenges of the application of knowledge map technology include the low level of automation in the process of knowledge map construction and the data noise caused by errors and redundancies in the data itself [4].

3 DATA FUSION BASED ON KNOWLEDGE MAP

3.1 Text preprocessing

In order to carry out the analysis and knowledge map construction based on unstructured text, some data preprocessing steps are needed, including:

1)Chinese word segmentation. The Chinese text to be analyzed is usually a continuous text sequence, which cannot be analyzed directly. It needs to be cut into meaningful words through Chinese word segmentation. Chinese word segmentation technology is widely used in the fields of text analysis and mining such as information retrieval. The proposed method mainly adopts the sequential tagging method based on statistical language model. Its basic principle is: firstly, prepare tagging data, and then model and train parameters of the text to be segmented based on the part-of-speech and statistical features of manual tagging, so that the model can describe the distribution of word elements relative to the context to a certain extent; Then, the model is used to predict the probability of word segmentation in the text to be extracted, and the word with

the highest probability is taken as the result of word segmentation analysis. The representative of this kind of sequence labeling algorithm is CRF algorithm [5].

2)Construct a word vector model. Computers can't directly process text, images, sounds and other contents, but they need to be converted into digital features before they can be processed. Word vectors are digital features that the computer can process after the words in the text are converted. Word embedding technology is adopted. The basic process is to embed the text into a mathematical space, so that the semantically similar words in the text are represented by similar vectors. The specific model adopted is word2vec.

3.2 Text information extraction

Aiming at the shortcomings of key information extraction in existing texts, this paper puts forward an innovative method, which can make full use of the advantages of existing methods, while avoiding their disadvantages to a great extent, so it has wide applicability. Its main principles and steps are as follows:

1) Using the characteristics of high accuracy and small matching range of "rule extraction", a small number of rules are written to match a small number of but accurate extraction objects from a large number of corpora, which can be imported as a follow-up process.

2) Cut out a certain proportion (such as 80%) of the extraction result obtained in step 1, and use it as the training corpus to be imported into the "automatic sequence labeling" method instead of the "manual labeling" process.

3) Using the training corpus obtained in step 2, and combining with the open source "automatic sequence labeling" algorithm, build an "extraction model".

4) Using the result of step 3, automatically extract the remaining corpus (such as 20%) cut out in step 2, and automatically judge the extraction result; If the accuracy of the automatic judgment of the model has not reached the business requirements, go to step 1, and write more regular expressions to form more "annotated corpora" to be imported as model training; If the accuracy of the automatic judgment of the model has reached the business requirements, the process is stopped, and the model is deployed and applied as the final model of text extraction [6].

3.3 Knowledge Map Construction

As shown in Figure 1, Neo4J provides the whole process support for the creation and application of knowledge map. In the data warehousing stage, the extracted named entities and entity relationships can be imported into the graphic database by batch import; In the stage of knowledge exploration, using Cypher language to query all nodes and relationships can not only obtain the whole knowledge map, but also search the required information of nodes and relationships, which can provide personalized knowledge services for users. In the application integration stage, you can call Neo4J's RREST API interface by programming to further develop the knowledge map interface. The basic framework of data fusion is based on the carding results of the company's integrated business system, and constructs a data fusion model from the perspective of business processes and business entities. It is the basis of data fusion application and the premise of supporting the five research contents of the application guide. Therefore, the basic framework of data fusion will also be included in the scope of this study.

The basic framework of data fusion includes two models: business process based data fusion model and business agent based data fusion model. These two models are based on graphic database technology and realize data fusion on different carriers through the concept of "Vertex and Edge".

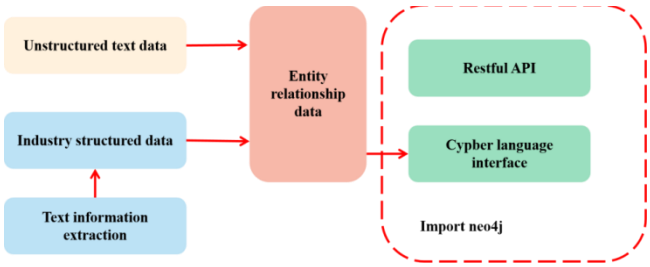

Figure 1. Storage and Utilization of Knowledge Map Based on Neo4J.

3.4 Knowledge Extraction from Reference Documents of Application Programming Interface

API reference documents belong to semi-structured data, and different HTML tags represent different types of API entities. The API tags contain attribute information of API entities, such as function description, parameters, return value and return value type. Therefore, this paper develops an AH entity and attribute analysis module to analyze and identify the relationship between application programming interface entities [7]. Specifically, through static analysis and identification of application programming interface classes or methods/functions and their inclusion relationships; According to the declaration rules of classes, the inheritance relationship between classes and base classes is extracted by regular expressions, and the entities and relationships of application programming interfaces are identified and extracted, as shown in Tables 1 and 2.

Table 1 The entity information table of the program programming interface reference document

API entity and atribute information
API entity set={Modules, classes, functions/methods, domain concepts}
Attribute set of module={Module name, description}
Property set of class={Module name, description, parameters, variables}
Property set of the function={Function name, description, parameter, return value, return value type}
The property set of the method={Method, description, parameter, return value, return value type, yield}

Table 2 Relationships information table of the program programming interface reference document

Relational information
Relation set={Include, inherit mention, overload}
Include={(Module, module)(Module, class).(Module, function),(Class, method),(Class, domain)}
Inherit={(Class, class)}
Mention={(Module, domain concept),(Class, domain concept),(Method, domain concept),(Function,domain concept).(Domain, domain concept)}
Overload=((Method, method)}

The domain concept implied in the functional description of application programming interface can be identified by the existing domain concept dictionary. Because this article focuses on the PyTorch framework. And identify the domain concept in the functional description of the application programming interface. That is, the domain concept dictionary is added to the user-defined named entity identification list of the tool to match and identify the domain concepts in the functional description of the application programming interface. Each API corresponds to a functional description, which implies domain concepts. Therefore, the "reference" relationship between API and domain concepts can be extracted from this corresponding structure [8].

4 APPLICATION EXAMPLES

The proposed technical scheme has been put into practice in the field of power grid project management risk early warning. Based on the power industry and external heterogeneous data involved in the process of power grid project management, the risk knowledge map is constructed, and the project management risk prediction based on the knowledge map is carried out.

4.1 Risk Analysis of Related Party Transaction Compliance

Get the equity relationship between power companies and related suppliers from the industrial and commercial data, maintain it in the knowledge map, and form "company-company equity relationship". Combine it with "company-project winning relationship" and "company-project purchasing relationship", and you can screen out the number, amount and proportion of the projects purchased by the same Party A's unit, which are won by Party B's unit with equity relationship. Compared with the set threshold, you can identify the scale of related party transactions.

1)For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

4.2 Risk control during project implementation

The key attributes of the project (including project number, project name, project contract signing time, project starting time, project agreed time limit, project actual completion time, etc.) are extracted from different electronic documents generated during the project execution, and integrated into the knowledge map. By querying the graph rules, you can screen out which projects are signed later than the start time, that is, the risk of "backsigning verification"; There is also the risk of "project delay" when the project has been carried out for longer than the contract period. It is also possible to further compare the credit information data corresponding to the risk items. If the above risks exist, and the suppliers also have known risks in credit information, the risk level will be increased accordingly. This scenario reflects the value of knowledge map to multi-source data fusion [9].

4.3 Risk Analysis of Project Unit's Performance and Insufficient Bearing Capacity

Extract the key attributes in the historical public announcement and contract data, such as Party A, Party B, project start time, project duration, project amount, etc., and integrate them into the

knowledge map, so that the number and amount of projects undertaken by the project unit in a certain period can be summarized through the map rules. By comparing the current data with the historical data of the same period, the situation that the number and amount of projects undertaken in the same period exceed a certain threshold can be screened out, and the possible risk of "insufficient bearing capacity of Party B" can be identified[10].

This project mainly uses the graphic database technology to establish the mapping relationship with the unified data center of the whole business, the source business system, and even the document management system and other multiple and heterogeneous data storage locations. Through the traversal and index algorithms, it builds a data fusion model based on the business value chain and relationship chain. The data fusion model can realize the calculation, retrieval and other applications of the relationship between data without data fusion and modification on the physical level. By using deep learning technology and combining the application practice of the business department on the fusion model, the model is trained and upgraded by using machine training algorithms, so that it can quickly adapt to the application needs of different businesses for data fusion, thereby improving the cross professional The cross system data application has laid a solid foundation for the company to carry out data mining and application more deeply and widely.

5 CONCLUSIONS

The fusion analysis of industry data and external data is one of the main trends of the development of big data technology in depth. Against the background of the actual demand of power industry, aiming at the fusion analysis of external unstructured data and internal data in the industry, this paper provides an extraction method of unstructured text relationship information based on natural language processing and a construction method of fused knowledge map, which realizes the fusion of multi-source heterogeneous data required to support business and provides a solid knowledge data foundation for upper-level advanced business applications. The pilot application of risk identification in the field of engineering project process management in a provincial electric power company can better integrate internal and external heterogeneous data into the knowledge map, and realize the verification of typical risks by mining the map rules, which verifies the effectiveness and feasibility of the proposed scheme. Only the process of building a knowledge map based on industry data and external data is emphasized in the front, but the graph-based mining analysis such as graph neural network and graph embedding is not discussed in depth, which is also the direction of further research.

REFERENCES

[1] Li, L. , Gan, J. , & Guo, X. . (2019). Research on the development roadmap of wind power industry under high renewable energy penetration. IOP Conference Series: Earth and Environmental Science, 384(1), 012018 (8pp).

[2] Gao, M. , Wang, J. , Zhang, W. , Wang, D. , Peng, Z. , & Lu, H. . (2021). Research on the importance of data enhancement technology in power document understanding. Journal of Physics: Conference Series, 1827(1), 012041-012047.

[3] Pang, R. , Yang, Y. , Zhang, P. , Yuan, Y. , & Liu, D. . (2021). Dynamic evaluation technology of bridge based on multi-source heterogeneous data fusion. IOP Conference Series Earth and Environmental Science, 638(1), 012064.
[4] Yu, J. , Zhang, Y. , Wu, Y. , & Mao, L. . (2021). Research on the practical application of visual knowledge graph in technology service model and intelligent supervision. Journal of Physics: Conference Series, 1982(1), 012040-012045.
[5] Xiong, J. , Liu, G. , Liu, Y. , & Liu, M. . (2021). Oracle bone inscriptions information processing based on multi-modal knowledge graph. Computers & Electrical Engineering, 92(6), 107173.
[6] Zhang, X. , & Guo, P. . (2021). Research on e-commerce logistics and traditional industry integration mode based on big data. Journal of Physics Conference Series, 1744(4), 042052.
[7] Hu, J. , Zhang, N. , Shang, Y. , Yu, J. , & Zhang, W. . (2021). Research on power equipment system of knowledge graph under electric energy in smart grid. IOP Conference Series: Earth and Environmental Science, 714(4), 042034 (7pp).
[8] Gao, L. C. , Yao, L. M. , Yang, Z. W. , & Zheng, F. . (2021). Research on hybrid index method of double-layer b+ tree for power big data considering knowledge graph. Journal of Physics: Conference Series, 1771(1), 012004 (8pp).
[9] Wang, K. , Zhang, R. , Song, L. , Lan, H. , Wu, Y. , & Pan, J. , et al. (2021). Research on intelligent technology of dispatching and control to ensure power supply based on multivariate information. Journal of Physics: Conference Series, 1846(1), 012023 (9pp).
[10] Qiu, S. , & Yang, L. . (2021). Research on the mode of short video project teaching and media talent training in colleges and universities under 5g industry chain based on multimedia technology. Journal of Physics: Conference Series, 1992(2), 022057-022059.