Journal of the Maharaja Sayajirao University of Baroda
ISSN:0025-0422

Acquaintance Graph-Based aware System for Data Link mistakes.

Dr.D.M.M.Vianny[1], T.Ramya Priya[2], P.Sowjanya[3], D Uma [4]
[1] Professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[2-4] Assistant Professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

Abstract: ***In view of the large scale of the data system of State Grid Corporation of China, multiple links and long cycles of data flow between business systems, lack of fault warning mechanism of links, unable to take measures to adjust links in advance, affecting the transmission efficiency of data links, etc., this paper constructs a fault domain ontology of data links, and combines the Jena reasoning mechanism to define inference rules that are in line with the field of data links. The link fault knowledge graph is constructed, and an improved similarity calculation method is designed to realize the fault warning of the data link, and improve the operation and maintenance efficiency of the data link.***

Keywords: Data Link, Knowledge Graph, Knowledge Reasoning, Fault Warning

1. Introduction

With the rapid development of the ubiquitous power Internet of Things construction and the continuous advancement of data platform construction, the State Grid of China has abundant data resources, more and more data providers and data users, but there is a lack of link fault early warning mechanism in data link monitoring and data management, which cannot take advance measures to adjust the link to improve the efficiency.

At present, most of the research on fault prediction is mainly reflected in the view of hardware entity, and the research on software data level fault analysis is relatively less. In literature [1], the predictive analysis model and processing process of electric vehicle power battery faults are established to realize the diagnosis and predictive analysis of battery faults. Literature [2] analyzed the research status of PHM technology for Marine diesel engines at home and abroad based on different fault prediction methods. At the software level, literature [3] built the equipment failure analysis and prediction system based on big data technology and improved FP-growth algorithm. Based on some data of SG-UEP longitudinal link in the data full link monitoring system, this paper takes the associated fault warning in the intelligent analysis field of data link as the target, uses Protege tool to build the fault domain ontology of data link, and uses Jena knowledge inference machine to carry out knowledge reasoning on the link fault knowledge graph. The OWL file obtained after reasoning is imported into the Neo4j graph database. Jaccard coefficient, cosine similarity coefficient and Pearson correlation coefficient are used to calculate the similarity between ontologies, and an improved similarity calculation method is designed through analysis and comparison. The correlation relationship between various faults is analyzed by the similarity of fault causes, and the possible correlation faults of data link are predicted. The above research results can effectively guarantee the high-quality and efficient operation of the data full-link monitoring system, and improve the monitoring efficiency and maintenance efficiency of the data full-link monitoring tools.

2. Knowledge Graph

2.1 Concept and Current Situation of Knowledge Graph

With the vigorous development of artificial intelligence technology, as an important part of artificial intelligence, the knowledge representation is also widely developed and applied, among which the knowledge graph is particularly prominent. Knowledge graph mainly describes and excavates potential correlation relationships through its visualization function among different entities. Originated in the early semantic network proposed in the 1960s, knowledge graph as a graph model of all kinds of correlation relations in the world is essentially a knowledge network of the entity and attributes through the relationship connection and organization, the basic unit is "entity-relationship-entity-entity" or "entity-relationship- attributes" triples [4]. Because of the unique advantages of knowledge graph in relationship representation, it is able to link large amounts of different kinds of information together and form relational networks so that users can analyze problems through the relationship perspective. In recent years, with the support of relevant technologies, knowledge graph in biomedical, management, economics, scientific metrology, military and many other fields have been popularized and applied, including common sense reasoning, entity extraction, prediction and analysis.

Developed on the basis of traditional knowledge base, knowledge graph is divided into open domain knowledge graph and vertical domain knowledge graph [5]. Open domain universal knowledge graph is constructed from a large amount of encyclopedic knowledge, so open domain universal

knowledge graph pays more attention to the breadth of knowledge. Vertical domain knowledge graph is oriented to specific fields, such as e-commerce domain knowledge graph applied to commodity shopping guide, financial domain knowledge graph applied to investment consulting investment research decision analysis application, etc. Vertical domain knowledge graph pays more attention to the depth of knowledge, requires higher knowledge quality, and knowledge structure is more complex.

2.2 Data Link Fault Warning Framework Based on Knowledge Graph

The establishment of knowledge graph of data link fault warning can give full play to the advantages of knowledge graph in visualization analysis and reasoning prediction function according to the internal correlation relationship built by knowledge graph. At first, this paper constructed the link fault domain ontology model, according to the actual situation of fault ontology may correspond to the cause of the problem, using the Jena framework set custom reasoning rules, based on the constructed and reasoned ontology model, the data files generated by Jena inference machine were imported into the Neo4j graph database, and then realize the construction of knowledge graph. After comparing several traditional similarity calculation methods, an improved similarity calculation method is designed to optimize the original reasoning model and finally realize the application of fault warning. The general flow chart of data link fault field is shown in Figure 1.

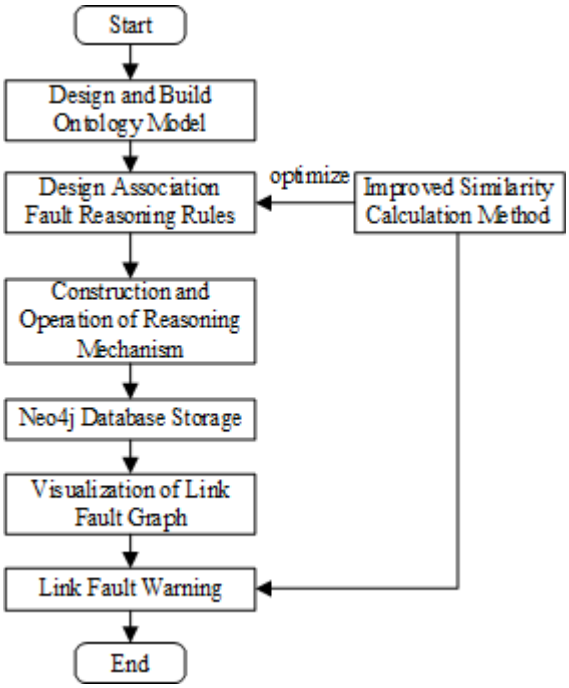

Figure 1: Flowchart for constructing domain knowledge graph of data link fault warning

3. Construction of Fault Domain Knowledge Graph of Data Link

3.1 Ontology Construction

In this paper, the seven-step decision-making method is adopted to construct the ontology [6], which constructs the concepts in the ontology model and the structural relations among the concepts in the top-down way. Protege5.5.0 was selected as the modeling tool for this paper. Protege is researched and constructed by Stanford University. It mainly uses graphical buttons to realize ontology modeling [7], with simple interface and convenient operation. Ontology representation language can provide modeling source language for ontology construction. It is a language that marks the basic concepts and relations between concepts in the objective world into forms that can be understood by computers [8]. It also has the function of intellectual reasoning. By comparing and analyzing several ontology representation languages, this paper chooses OWL language to describe ontology. The OWL language, a W3C recommendation, treats Web resources as presentation objects. Ontology model based on OWL language can not only formally express classes and structural relations among classes, but also realize knowledge reasoning function. The following will describe the specific steps of constructing the data link fault warning ontology model.

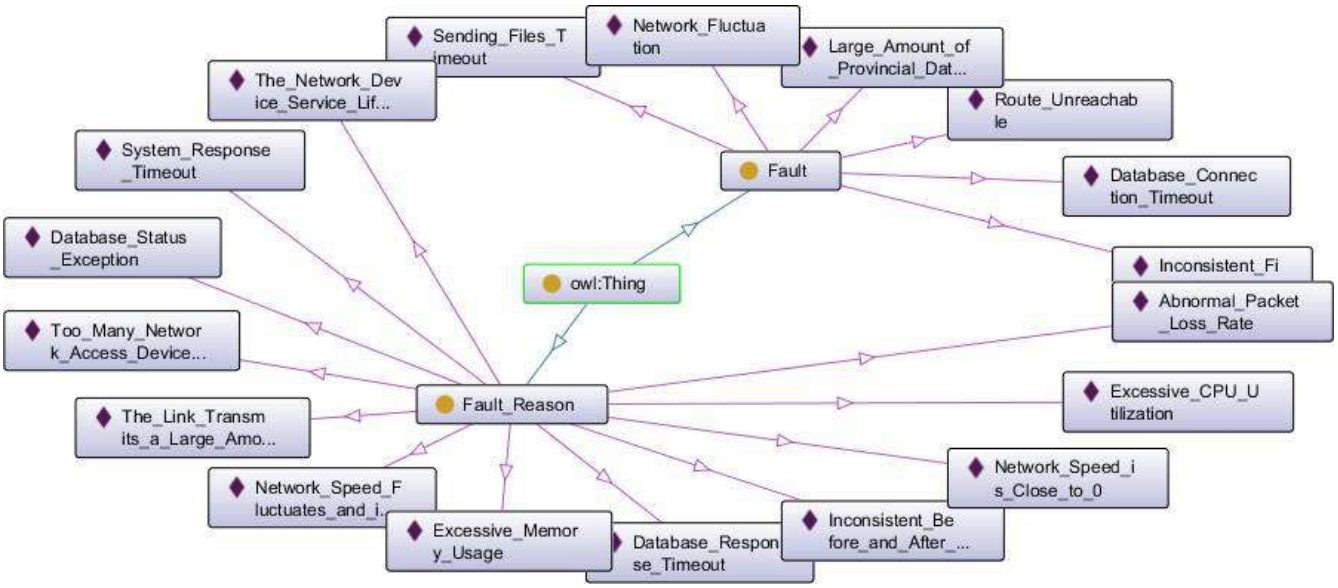

Figure 2: Ontology hierarchy

(1) Define the field and scope.
In the ontology model of data link fault warning to be constructed in this paper, the concepts, the hierarchical relationship between concepts, attributes, constraints, instances and so on are all obtained from the analysis and design of SG-UEP vertical link, a data exchange platform under State Grid Information and Communication Corporation.

(2) Consider reuse of existing ontologies.
Before the construction, we read a large number of literature to find out whether there is a reusable ontology model for data link fault warning. Meanwhile, this paper also finds a large number of useful ontology libraries, such as WordNet, DAML and Ontolingua. The results show that there is no ontology model suitable for this study, and none of them can be reused directly. Therefore, this paper needs to rebuild the ontology model based on data link fault warning.

(3) Get the terminology.
According to SG-UEP longitudinal link failure condition, analyzing common faults with available in the following six: database connection timeout, sending files timeout, route unreachable, network fluctuation, large amount of provincial data and inconsistent field types, its corresponding fault reasons system response timeout, abnormal state of database, network equipment, use fixed number of year is too long and abnormal packet loss rate, etc. 12. In this way, the basic hierarchy of data link fault warning ontology is obtained.

(4) Define classes and their hierarchy.
Faults and fault causes are conceptualized into corresponding classes. All the topmost classes are defined as Thing, and all classes edited at ontology construction can only be subclasses of Thing. The hierarchical structure of data link fault warning ontology is shown in Figure 2.

(5) Define attributes and attribute constraints.
Build properties and constraints, including Object properties (object property) and their constraints. The ontology built in this article contains the object attribute relationship "Cause", setting the constraint: the body (definition domain) of the relationship is the fault cause class, and the object (value domain) is the fault class.

(6) Create an instance.
By using ontology editing tool Protege5.5.0 to edit an instance of ontology concept, the ontology of data link fault warning can be constructed.

(7) Ontology verification.
Inference machines such as FaCT++, Pellet and Hermit are provided in Protege construction tool [9]. The initial ontology model constructed in the previous step may have semantic contradictions, so it cannot be directly applied. In this paper, FaCT++ inference machine is used to verify the initial ontology to eliminate logical errors in the body, ensure the consistency of semantic logic of the ontology, and ensure the normal operation of subsequent Jena reasoning. Then, the graphical plug-in OntoGraf can be used to display the ontology model and determine the formal structure of data link fault warning ontology.

3.2 Knowledge Reasoning

Developed by the HP Laboratory Semantics Network Research Project, Jena is a major open source project in the field of Semantics Network, and is an Java application framework. Storage triples based on RDF graph model are commonly used to realize RDF data management and rule-based ontology reasoning. Specific structure [10] in the Jena framework is mainly divided into six parts:

1) RDF API: builds the RDF model to realize the processing of RDF files and models;

2) Ontology reader and writer / parser: parse XML syntax-based files such as RDF, RDFS, OWL;
3) Persistent storage scheme of the RDF model;
4) Reasoning API:: Rule-based reasoning, and the reasoning subsystem is used for reasoning in the retrieval process;

5) (5)Storage API: supports storage in local or tridatabases in the form of OWL files or relational databases;
6) Query API: supports SPARQL query language used for query search of information.

The Jena query structure is shown in the figure below.

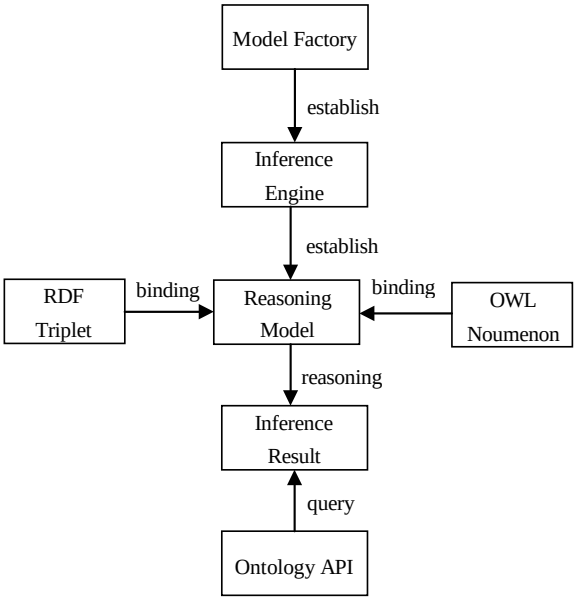

Figure 3: The Jena query structure

The Jena reasoner itself has two internal rules engines: the forward chain inference RETE engine and a tabled datalog engine [11]. They can either run independently or use the forward chain as a forerunner to the back chain engine. Both inference engines need to define their behavior with a rulesset. Jena itself contains some general rules that can be used to check the correctness of relationships between different classes, attributes and properties; users can also make their own rules to create reasoning machines to supplement the general rules to meet the personalized needs of the users. The data link fault ontology semantic custom query rules are as follows:

Rule1: (? p: causes? c) (? p: causes? c2) -> (? c: -related? c2)
Rule1 indicates that if a cause of fault causes faults 1 and fault 2, fault 1 is related to fault 2.

3.3 Knowledge Graph Construction

The ontology model of the link fault domain should be constructed, and the defined inference rules should be formulated according to the possible fault causes corresponding to the fault ontology in the actual situation, combined with the Jena inference mechanism. Next, we need to import the OWL file generated by Jena inference machine into the Neo4j graph database, so as to realize the construction and visualization of the knowledge graph.

Neo4j graph database is a non-relational database based on Java language. It stores the relationships between different entities through graph theory. It is compatible with ACID properties [12] and also supports other programming languages. Neo4j stores data in diagrams rather than tables and can work with diagrams containing billions of nodes and relationships simultaneously, making it suitable for enterprise-level production environments. Currently, Neo4j graph database has been used in many fields, such as project management, software analysis and so on. Compared with relational databases, graph databases are better at dealing with highly correlated, repetitive and complex data with low structure. Cypher is a graph database query language [13], which is equivalent to SQL language in relational database. It has high query efficiency and can efficiently query and update graph database. The features supported by Neo4j are as follows:

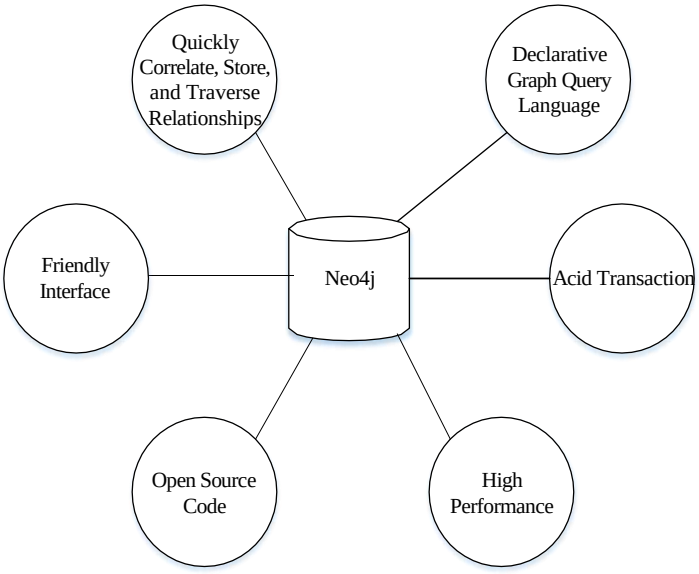

Figure 4: Features of Neo4j

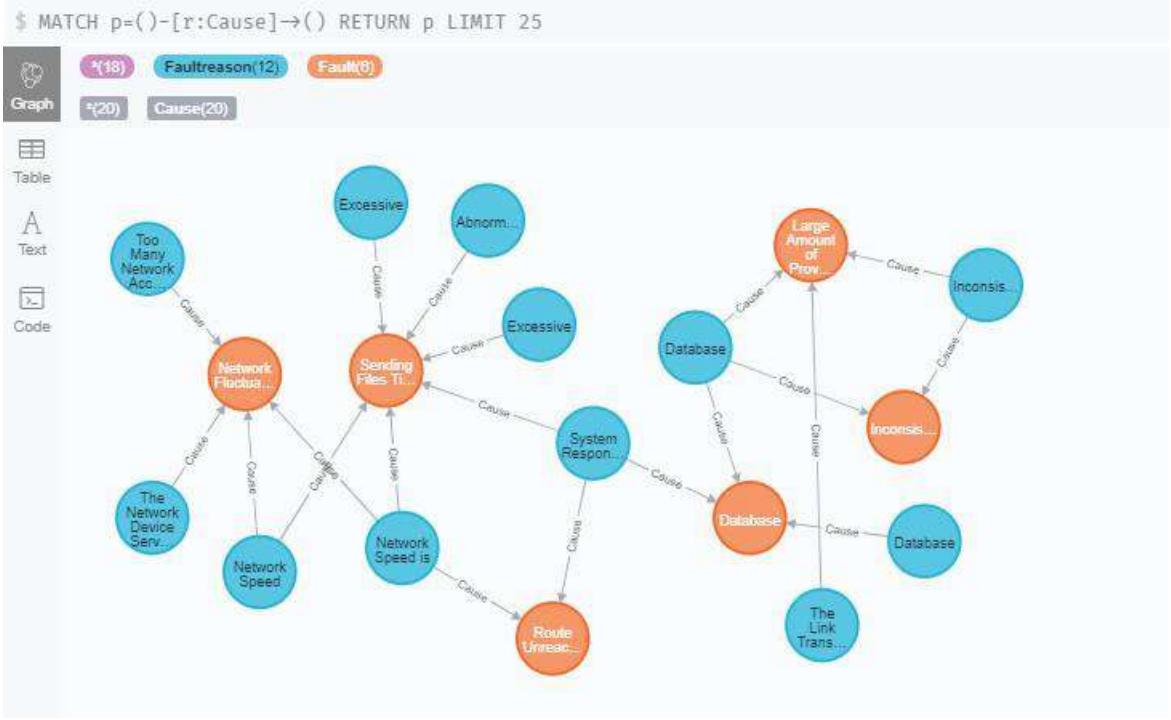

Figure 5: Overall structure of data link fault knowledge graph

Figure 5 shows the overall structure of the final data link fault knowledge graph. Can be seen from the graph data link fault fault and the fault reasons of knowledge map contains nodes, node fault and the fault reasons and the cause of relationship between the relationship between different failure, can be seen from the map of in addition to the field type inconsistent fault, the fault existing in the cause of the problem are different degree of correlation between.

4. Research on Data Link Fault Warning

To realize the warning of data link faults, it is necessary to formulate the corresponding knowledge inference rules and deduce the possible association relations between ontologies. The key of the reasoning model is to calculate the similarity between faults. In order to better realize the data link fault warning, the fault ontology is fully considered, several similarity methods are compared and analyzed, and an improved similarity calculation method is proposed.

4.1 Traditional Similarity Calculation Method

Similarity is comparing the similarities between two things. Usually by calculating the distance between features of things. Similarity calculation methods mainly include:

Jaccard coefficient: the proportion of the intersection elements of two sets in the union [14]. Its calculation formula is as follows:

$$\mathrm{sim}_{\mathrm{Jac}}(i,j)=\frac{\left|S_i \cap S_j\right|}{\left|S_i \cup S_j\right|} \qquad (1)$$

$$\mathrm{sim}_{\mathrm{Cos}}(i,j)=\cos(A,B)=\frac{A\cdot B}{\|A\|\cdot\|B\|} \qquad (2)$$

Pearson correlation coefficient: The correlation coefficient R in the correlation analysis calculates the cosine Angle of the space vector after X and Y are normalized based on their own population respectively. The calculation formula is as follows:

$$\mathrm{sim}_{\mathrm{Pea}}(i,j)=r(X,Y)=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x)^2}\cdot\sqrt{n\sum y^2-(\sum y)^2}} \qquad (3)$$

4.2 Improved Similarity Calculation Method

Through the comparison of the above calculation results, it is found that the Jaccard coefficient only cares about whether the common features between ontologies are consistent, and cannot measure the specific value of the difference. The cosine similarity is more to distinguish the difference from the direction, but is not sensitive to the absolute value. The cosine value of the Angle between two vectors is only used as a measure of the difference between two ontologies. Pearson correlation coefficient focuses more on measuring the linear correlation between two variables X and Y [10].Therefore, this paper comprehensively considers the fault attributes in the link fault knowledge graph and these similarity calculation methods. By analyzing the proportion of each influencing factor and setting its corresponding weight value, a new method is proposed to calculate the similarity between ontology concepts. The formula is as follows:

$$\mathrm{sim}(i,j)=\alpha\mathrm{sim}_{\mathrm{Jac}}(i,j)+\beta\mathrm{sim}_{\mathrm{Cos}}(i,j)+\gamma\mathrm{sim}_{\mathrm{Pea}}(i,j) \qquad (4)$$

Cosine similarity: the cosine value of two vectors, which is calculated as follows:

Among them, α, β and γ represent the adjustment coefficient of Jaccard coefficient, cosine similarity and Pearson correlation coefficient respectively, where α+β+γ=1.

5. Experimental Results and Analysis

According to the above calculation method and the actual attributes of link faults, the similarity between the fault ontology constructed in this paper is calculated. Jaccard coefficient, cosine similarity and Pearson correlation coefficient between faults are as follows (Fault1-6 stands for database connection timeout: 、sending files timeout、route unreachable、network fluctuation、large amount of provincial data and inconsistent field types):

Table 1: Jaccard coefficient between faults

	Fault1	Fault2	Fault3	Fault4	Fault5	Fault6
Fault1	1	0.14	0.25	0	0.2	0
Fault2	0.14	1	0.33	0.25	0	0
Fault3	0.25	0.33	1	0.2	0	0
Fault4	0	0.25	0.2	1	0	0
Fault5	0.2	0	0	0	1	0.67
Fault6	0	0	0	0	0.67	1

Table 2: Cosine similarity between faults

	Fault1	Fault2	Fault3	Fault4	Fault5	Fault6
Fault1	1	0.26	0	0	0.2	0
Fault2	0.26	1	0	0	0	0
Fault3	0	0	1	0	0	0
Fault4	0	0	0	1	0	0
Fault5	0.2	0	0	0	1	0
Fault6	0	0	0	0	0	1

Table 3: Pearson correlation coefficient between faults

	Fault1	Fault2	Fault3	Fault4	Fault5	Fault6
Fault1	1	-0.19	0.25	-0.41	0.11	0.26
Fault2	-0.19	1	0.45	0	-0.58	-0.45
Fault3	0.25	0.45	1	0.16	-0.26	-0.20
Fault4	-0.41	0	0.16	1	-0.41	-0.32
Fault5	0.11	-0.58	-0.26	-0.41	1	0.77
Fault6	0.26	-0.45	-0.20	-0.32	0.77	1

Can be seen through the comparison and analysis, the proposed similarity calculation method, this paper has a detailed analysis of the relationship between the concept of ontology structure, based on the vector space, linear relationship with the traditional and the characteristics of the individual attributes of the approach to consider various factors affecting the contrast experiment to verify the hybrid semantic calculation method can more accurately to determine the degree of similarity between concepts. The effectiveness of the proposed method is verified by comparing with traditional methods. The experimental results are shown in Table4.

Table 4: Comparison of similarity values between some concepts

Similarity	Jaccard Coefficient	Cosine Similarity	Pearson Correlation Coefficient	Method of This Paper
Sim(Fault1, Fault2)	0.14	0.26	-0.19	0.12
Sim(Fault1, Fault5)	0.2	0.2	0.11	0.19
Sim(Fault5，Fault3)	0.33	0	0.45	0.31
Sim(Fault3, Fault4)	0.2	0	0.16	0.18
SimFault5, Fault6)	0.67	0	0.77	0.61

It can be seen from the experimental results that the more identical attributes, the closer the word frequency vectors and the more similar the linear correlation degree are among the ontology instances, the more similar the ontology instances are. The results obtained by calculating semantic similarity based on a single method are sometimes inaccurate, and cannot distinguish all kinds of attributes of instances more comprehensively. For example, from the calculation results of sim (database connection timeout, large amount of provincial data) and sim (route unreachable, network fluctuation), it can be seen that the results are the same under the calculation method based on Jaccard coefficient. This is because the ontology of "database connection timeout" and "large amount of provincial data" and "route unreachable" and "network fluctuation" share the same feature number, but this calculation method does not consider the semantic relationship of the concept itself, so the results obtained are inaccurate. Secondly, from the calculation results of the sim(database connection timeout and large amount of provincial data), the Jaccard coefficient and cosine similarity of "database connection timeout" and "large amount of provincial data" are the same. However, through the calculation results of Pearson coefficient and combined with the linear correlation degree of the two ontology vectors, we will find that the correlation degree of the two will be relatively lower. The semantic similarity calculation method proposed in this paper is based on the traditional calculation method, and comprehensively considers the influence factors such as the relationship between attributes, vector space, linear relationship and the relative weight of the three ontologies. The experimental results are relatively reasonable compared with other methods.

6. Conclusion

Aiming at the problems of large data scale, multiple links and lack of link fault warning mechanism in the data link under the data center of State Grid Corporation of China, this paper analyzes and designs a knowledge graph based on the fault domain of data link and realizes the warning application of data link fault. The main research work of this paper is as follows:

1) According to the partial data of SG-UEP longitudinal link on the data exchange platform of State Grid, Protege ontology modeling tool was used to build the fault domain ontology of data link; The custom inference rules were built and loaded into Jena inference machine. The generated OWL files were imported into the Neo4j graph database, and the visualization function of Neo4j was used to build the knowledge graph of data fault domain.

2) According to the logic relationship between the concepts of the fault body model, analyze the similarity of the fault cause through the improved similarity calculation method, and then obtain the degree of correlation between the faults, and combine the results with the rules formulated by the Jena reasoner to analyze the associated fault nodes.

Through comparison, the improved similarity calculation method meets the actual situation and provides higher reference value for failure prediction, realizing fault early warning, realizing efficient display of data link graph and improving link operation and maintenance efficiency.

References

[1] Hua Zhan. “Fault diagnosis and prediction analysis of electric vehicle power battery” [J]. Electromechanical technology, 2021 (02): 60-61 + 74.

[2] Yun Ke, Enzhe Song, Chong Yao, Quan Dong. “Overview of marine diesel engine fault prediction and health management technology” [J]. Journal of Harbin Engineering University, 2020, 41 (01): 125-131.

[3] Haiyang Wang, Zhaojiang Chi, Pengfei Cai. “Design and practice of distribution network fault emergency repair analysis and prediction system based on big data technology” [J]. Power big data, 2020, 23 (06): 63-68.

[4] Li Wang. “Research on intelligent knowledge support of urban rail transit construction safety management based on knowledge map” [D]. China University of mining and technology, 2019.

[5] Kim Eun Hee, Yoon Hee Chul, Lee Ji Hyun, Kim Hee Soo, Jang Young Eun, Ji Sang Hwan, Cho Sung Ae, Kim Jin Tae. “Prediction of gastric fluid volume by ultrasonography in infants undergoing general anaesthesia”[J]. British Journal of Anaesthesia, 2021, 127(2).

[6] Jia Zhu. “Design and research of coal mine monitoring and early warning model based on ontology and association rules” [D]. Anhui University of technology, 2019.

[7] Minhui Peng, Li Si. “Investigation and Analysis on the application of prot é g é ontology construction tool ”[J]. Library and information work, 2008 (01): 28-30.

[8] Yihua Ni. “Research on ontology based knowledge integration technology for manufacturing enterprises” [D]. Zhejiang University, 2005.

[9] Ke Du, Weimin Lin, Jinghong Guo, Feng Huang, Li Huang. “Research on power grid reasoning expert system based on ontology”[A]. American Applied Sciences Research Institute, AASRI. Proceedings of 2012 Conference on Modelling, Identification and Control (MIC 2012 V3)[C].American Applied Sciences Research Institute, AASRI: Society for the application of intelligent information technology, 2012:6.

[10] Meir M Lehman，Juan F Ramil. “Rules and Tools for Software Evolution Planning and Management” ［J］. Annals of Software Engineering，2001， 11(1): 15 44.

[11] Xufei Nie. “Research on the application of ontology and rule-based reasoning in logistics” [D]. Tianjin University, 2007.

[12] Yisong Ma, Zhigang Wu. “Power big data modeling and analysis based on neo4j [J]. New technology of electrical energy”, 2016, 35 (02): 24-30.

[13] Li Feng. “Constructing middle school Chinese poetry knowledge map based on neo4j map database” [D]. Shaanxi Normal University, 2019.

[14] Ziqi Yan, Qiong Wu, Meng Ren, Jiqiang Liu, Shaowu Liu, Shuo Qiu. “Locally private Jaccard similarity estimation” [J]. Concurrency and Computation: Practice and Experience, 2019, 31(24).

[15] Kim Eun Hee, Yoon Hee Chul, Lee Ji Hyun, Kim Hee Soo, Jang Young Eun, Ji Sang Hwan, Cho Sung Ae, Kim Jin Tae. “Prediction of gastric fluid volume by ultrasonography in infants undergoing general anaesthesia”[J]. British Journal of Anaesthesia, 2021, 127(2).