Industrial Data Fusion Method based on Semantic Data Dictionary for Digital Twin

Yang Liu*[a,b,c], Tianshi Zhang[a,b,c]
[a]Shenyang Institute of Automation, Chinese Academy of Sciences; [b]Key Laboratory of Networked Control Systems, Chinese Academy of Sciences; [c]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, China
* Corresponding author: liuy@sia.cn

ABSTRACT

Industrial digital twin is the key support for the transformation of industry to intelligence. The realization of digital twin relies on the integration of whole-process data, which can provide real-time and intelligent decision-making optimization for production management. This paper proposes a digital twin data fusion method based on semantic data dictionary. Firstly, a modeling method of domain data dictionary is designed, and a semantic method is introduced to realize the construction of semantic data dictionary. Then, based on the semantic data dictionary, a semantic similarity calculation method based on multivariate distance weighting is proposed. Finally, the algorithm is tested and verified. The experimental results show that the method in this paper has a significant effect on automatic data fusion in the construction of industrial digital twins.

Keywords: data dictionary, semantic modeling, digital twin, data fusion, industrial data

1. INTRODUCTION

With the rapid development of a new generation of information technology, the deep integration of IT technology and OT technology is an important trend in industrial development. In order to meet the individualized and intelligent development needs of the industry in the process of development, countries around the world have successively launched their own industrial development strategies[1]. The bottleneck problem that is generally faced at present is how to realize the integration and interaction of physical space and information space in the process of industrial production[2]. Therefore, digital twin is considered as an important means to solve this problem.

The core of digital twin technology is model and data[3]. It is mainly based on model, driven by data. Through the fusion analysis of data, it simulates the behavior of physical entities in the real environment, so that industrial production can truly realize visualization, Intelligent. For example, applying the fused data to the R&D process will improve the accuracy and reliability of product design; in the manufacturing process, it will provide optimized production decisions for the production process; in the maintenance process, it will provide favorable conditions for the maintenance and failure prediction of finished products. Decision basis, effectively improve the reliability and availability of products[4].

The data dictionary is a method for defining and describing the meaning of data, and its purpose is to make a detailed description of various object classes and their properties in the field. Simply put, a data dictionary is a collection of information describing data, a collection of definitions for all data elements used in a system. The semantic data dictionary can not only define the meaning of data, but also express the relationship between data, which is an important basis for realizing industrial data fusion. Data dictionaries are usually classified and formulated according to sub-fields. The fusion of heterogeneous data based on classified data dictionaries in different fields is one of the means for digital twins to achieve unified understanding at the data level.

The formulation of the current data dictionary presents different characteristics on the OT side and the IT side. On the OT side, it often appears in the form of standards, such as IEC 61360, IEC 61987, IEC 61850, etc., which define related objects and attribute definitions in the subdivision field of industrial equipment. On the IT side, the definition of the data dictionary is often defined in the form of XML standards, usually only the framework is defined, but the entity content is not defined much. ID parsing is the application form of IT data dictionary, and the description objects are often entities.

Third International Symposium on Computer Engineering and Intelligent Communications (ISCEIC 2022), edited by Xianye Ben, Proc. of SPIE Vol. 12462, 124620P · doi: 10.1117/12.2661161

Reference[5] proposes a data dictionary learning method constrained by ordinal locality for mixed frequency data classification. Reference [6] focuses on business/data dictionary to improve data governance for industries and propose a method of applying these natural language processing approaches to the business/data dictionary. Reference [7] proposes an optimized data dictionary updating learning-based compressed data collection algorithm to improve data classification. With the development of information technology, the deep integration of IT-side technology and OT-side data dictionary technology and application to the entire manufacturing life cycle will become the key to supporting the realization of digital twins.

2. RELATED WORK

Data fusion is called sensor fusion in industry. It mainly fuses the data collected from sensors, and comprehensively analyzes and intelligently fuses sensor data or knowledge with different sources and semantic ambiguity under certain rules [8].Reference[9] proposes multi-source data fusion technology is an important cornerstone of research on big data and artificial intelligence. Reference[10] uses data fusion method which is easy to realize is used to fuse and mine the massive big data. Reference[11] proposes a BP neural network method to improve fast data fusion and accurate data acquisition operation. This obtains a consistent interpretation and accurate description of the target object, and then provides better decision-making information for the industrial digital twin. However, with the expansion of the coverage of the industrial digital twin, the current industrial data fusion refers to the people involved in the production process. The integration of all relevant data of machine, material, method and environment also brings new challenges to the realization of industrial digital twin.

Data fusion technology is first generated and developed to serve military needs. The US Department of Defense believes that data fusion involves detecting, estimating and combining data from multiple sources at multiple levels in order to obtain real-time, accurate and complete battlefield situation and threat estimates. It emphasizes three characteristics of data fusion: First, data fusion is a hierarchical fusion process, and the acquired multi-source data is classified and re-fused according to specific categories. Second, data fusion is a whole process, including data detection, The whole process of correlation, estimation and processing analysis; The third is the result of data fusion is to realize the whole battlefield situation estimation. Based on this conceptual framework, the famous JDL (Joint Directors of Laboratories) data fusion model is proposed. The JDL model achieves the goal of data fusion through five different levels of processing modules including information preprocessing, object estimation, situation estimation, threat estimation and processing correction. Based on the JDL model, researchers proposed the DFIG (Data Fusion Information Group) model. Compared with the JDL model, the DFIG model increases the influence of human decision-making factors in the fusion process, and configures the fusion database resource management and target task management in a unified manner, extending the high-level fusion target from situation estimation to a set of situation estimation, impact assessment, process Optimization, user optimization and goal management all in one. The OODA loop model proposed by John Boyd includes four parts: observation-orientation-decision-execution loop. Each stage is executed sequentially, which cannot well reflect the parallelism of data fusion processing. Bedworth et al proposed the waterfall model, which sequentially performs the processes of data acquisition, signal processing, feature extraction, pattern processing, situational assessment, and decision-making. It is like a waterfall. The model emphasizes too much on low-level information processing, and does not Clear feedback mechanism.

Reference[12] proposes digital twin as a new potential technology has received much attention from both academia and industry increasingly. Digital twin technology[13] paves a way for achieving cost-efficient trial and performance-optimal management. The data fusion applied to the industrial digital twin is a typical multi-source heterogeneous data fusion, which emphasizes eliminating the differences between different data sources and providing an undifferentiated global unified view for multi-source heterogeneous data. According to the types of models used in data fusion, multi-source data fusion methods are divided into structured methods [14] and semantic methods [15]. Structural methods are mainly based on information structure and do not consider the semantic connection between information, and focus on solving the problem of structural heterogeneity of data to be fused. This method is simple to implement and suitable for scenarios with fixed information sources, but has the disadvantage of poor scalability. The starting point of the semantic method to solve the problem is the semantic association between information. By uniformly describing the semantics of the information and its association, the effect of global representation can be achieved, and services such as data sharing and information query can be realized. This method has good scalability and is suitable for dynamic data sources that are constantly changing. It can perform data query work at the semantic level, but its disadvantage is that the process is more complicated and the implementation is more difficult. OLA [16-17] is a data fusion system that maps from a grammatical point of view. The system mainly analyzes the constituent elements in the ontology, and uses the comprehensive value of the Hamming Distance and the tagging distance between elements to represent the semantic correlation between different elements. However, the calculation of similarity can only be carried out in the same feature space, and elements of different types cannot be compared, and the generality is poor.

The remainder of this paper is organized as follows. We first introduce related works and methods of semantic data fusion in industrial digital twins in Section 2. Then we design the domain data dictionary construction and semantic modeling methods in Section 3. Then we propose a semantic similarity calculation method based on multivariate distance weighting in Section 4. Then we conduct experimental verification in Section 5 using industrial data. Finally, we conclude this article in Section 6.

3. DOMAIN DATA DICTIONARY CONSTRUCTION AND SEMANTIC MODELING

3.1 Building the Domain Data Dictionary

For industrial digital twin data fusion, the multi-source heterogeneous data mainly comes from the production site, so the construction of the domain data dictionary is carried out according to the characteristics of the field application. According to the analysis of the human-machine-material-loop object on the scene, in order to support the description of the object itself and the description of the object relationship, the following 7 types of data dictionaries are defined:

- *Class dictionary* $\mathrm{Dic}_{\mathrm{class}}$
- *Attribute dictionary* $\mathrm{Dic}_{\mathrm{properity}}$
- *Unit dictionary* $\mathrm{Dic}_{\mathrm{unit}}$
- *Data type dictionary* $\mathrm{Dic}_{\mathrm{type}}$
- *Relationship dictionary* $\mathrm{Dic}_{\mathrm{relation}}$
- *Class-attribute dictionary* $\mathrm{Dic}_{\mathrm{c\_p}}$
- *Class-Relation-Class Dictionary* $\mathrm{Dic}_{\mathrm{c\_r\_c}}$

Each data dictionary must contain ID, Description, original, and version information. ID refers to the unique identifier of the data, description is used to describe the data, original is used to describe the source of the data, and version is used to describe the version of the data dictionary.

In particular, the attribute dictionary also needs to distinguish attribute classification. According to the industrial heterogeneous data source mode, it is divided into structural attribute classification Cons, functional attribute classification *Func*, performance attribute classification *Perf*, location attribute *Loc*, and business attribute *Buss*.

Among the seven types of dictionaries defined, they can be divided into basic type data dictionary and mapping type data dictionary. The former includes class dictionary, attribute dictionary, unit dictionary, data type dictionary and relation dictionary, and the latter includes class-attribute dictionary and class-relationship dictionary. The relational structure of the data dictionary is shown in Figure 1.

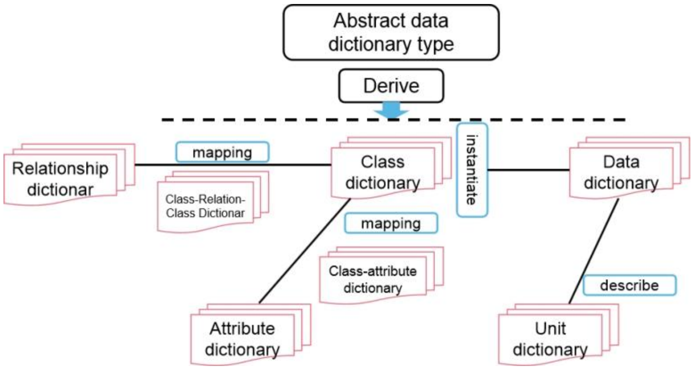

Figure 1. Data dictionary relation diagram

The abstract data dictionary type is a basic type, from which five basic type data dictionaries and two mapping data dictionaries are derived. Among them, through the mapping connection operation between the class dictionary and the attribute dictionary, a class-attribute dictionary will be generated, and the effective mapping relationship will be recorded. The class-attribute mapping dictionary can add entries through operations, or manually insert entries.

Let $C = \{c_j | j \in [1..n]\}$ represent the class dictionary and $A = \{a_i | i \in [1..n]\}$ represent the attribute dictionary. When performing automatic mapping, the Cartesian product must be performed first in (1):

$$CA' = C \times A = \{(a_i, c_j) | a_i \in A \wedge c_j \in C\} \qquad (1)$$

Given a constraint $\mathrm{f(k)} = \begin{cases} 0, & \text{k is not eligible} \\ 1, & \text{k is eligible} \end{cases}$

Get: $\mathrm{CA} = \{\mathrm{ca} | \mathrm{ca} \in \mathrm{CA'} \wedge \mathrm{f(ca)} = 1\}$

The above is the principle of automatically deriving the class-attribute dictionary from the class dictionary to the attribute dictionary under the given conditions.

Similarly, although there is no data-unit dictionary, in practical applications, the relationship between data and units is usually "described". Therefore let $D = \{d_i | i \in [1..n]\}$ represent the data dictionary and $U = \{u_j | j \in [1..n]\}$ represent the unit dictionary. When performing automatic mapping, a Cartesian product must be performed first in (2):

$$DU' = D \times U = \{(d_i, u_j) | d_i \in D \wedge u_j \in U\} \qquad (2)$$

Given a constraint $f(k) = \begin{cases} 0, & k\ is\ not\ eligible \\ 1, & k\ is\ eligible \end{cases}$

Get: $DU = \{du | du \in DU' \wedge f(du) = 1\}$

The above is the principle of automatically deriving the description relationship from the data dictionary to the unit dictionary under the given conditions. Relationships between other data dictionaries can be deduced in a similar way.

3.2 Modeling of Semantic Description Based on Data Dictionary

At present, the semantic models for intelligent service fusion mainly include SensorML proposed by OGC and SWE system proposed by W3C, and SWE is the most widely used. SWE's semantic description system for Sensing Web only includes three types of elements: time, space and subject, among which the subject coverage is relatively vague, and for the whole process of interconnected production with clear processes, a clearer division of the subject will be conducive to the analysis of data. Precise extraction and use. Therefore, this project analyzes the usage patterns and habits of the interconnected production data stream, and combines the hotspot mining technology to analyze and extract the key factors of semantic usage and hotspot attributes, so as to realize the extraction of the key factors of interconnected production metadata, including time, location, Users, operation objects, functions, operation types and operation descriptions (mapped to *when, where, who, which, what, how, do what*) are seven categories of factors. By extracting the above key factors, the interconnected production of heterogeneous elements is realized. Semantic encapsulation of data.

Build a hierarchical semantic usage architecture based on a general semantic model, the formal description is in (3)

$$\begin{gathered} SD_{Frame} = \{D, S, C, Rep, Res | D \subset Dic, S \subset Dic, \\ C \subset Dic, Rep \subset Dic, Rep \subset Dic\} \end{gathered} \qquad (3)$$

- *Dic:* dictionary;
- *D:* data meaning, data function, data value and timestamp, source, device and group, operation type;
- *S:* device status meaning, data source device status
- *C:* device association information, the data source device is associated with other devices;
- *Rep:* report information, data active production report record;
- *Res:* reserved item

Taking the device semantic description as an example, its formal description is in (4)

$$SD\_DeL(X_{Dev}) = \{x | x \in SD\_Frame\} \qquad (4)$$

4. DATA FUSION METHOD BASED ON MULTIVARIATE DISTANCE WEIGHTED SEMANTIC SIMILARITY CALCULATION

In industrial digital twin data fusion, there are a lot of problems that the same data has different expressions in different systems. This is caused by the inconsistent mapping of data after the system has been integrated for many times. The association will result in a large number of invalid, inaccurate or inconsistent associations. For example, the AGV running trajectory data is usually managed in the RCS system, and its running status is often synchronized in the PLC and MES systems. The AGV running in these three places Trajectory and status data both refer to the same AGV, but the application timeliness and requirements are different, resulting in different data storage and expression methods. The same AGV data obtained from different systems is completely inconsistent in form and description. Therefore, the definition of semantic similarity is It is considered to be an effective method to solve the merging of heterogeneous descriptions of homologous data.

There are various types of equipment in the industrial digital twin production site, and there are many devices with more than one device. The same type of data cannot be guaranteed to have a unique source. If all the data is Cartesian product and then saved during semantic calculation, it is not necessary. The waste of resources is also serious, and it will also lead to low runtime efficiency when querying. Taking the production process of a petroleum refining plant as an example, the temperature transmitter is used in multiple production links on site, and the data it generates is updated regularly, and the source is not single. In order to improve efficiency, improve matching, and reduce error rates , it is necessary to distinguish the source, purpose, quality, etc. of the data. In order to better calculate the semantic similarity of data and reduce the repetition and error rate, this paper proposes a weighted semantic similarity calculation method based on multivariate distance. The typical semantic distance such as distance is calculated, and then weighted to output a more reasonable and optimized semantic distance, as shown in Figure 2.

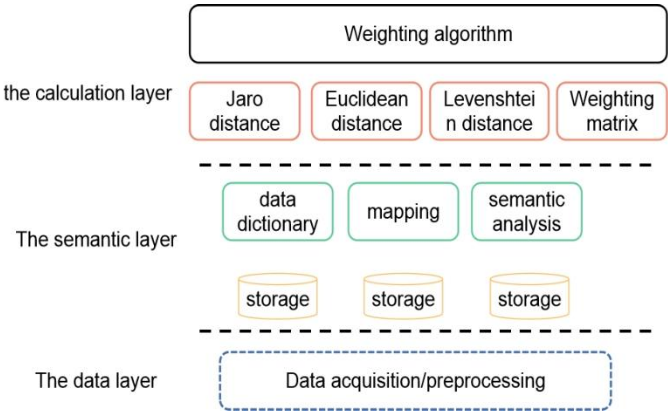

Figure 2. A framework of semantic similarity calculation method based on multivariate distance measurement weighting

From the perspective of the algorithm, it can be divided into three layers, including the data layer, the semantic layer and the computing layer.

The data layer mainly completes data acquisition and data-related preprocessing. Data preprocessing may include but is not limited to data cleaning, data dimensionality reduction, data filling, etc., to prepare for future association, calculation, and expansion.

The bottom layer of the semantic layer includes redundant storage devices, which can be implemented by real-time database or relational database; this layer also includes data dictionary, mapping and semantic analysis. The data dictionary is defined in the previous section and includes seven types of industrial production factor data dictionaries: class dictionary, attribute dictionary, unit dictionary, data type dictionary, relation dictionary, class-attribute dictionary, class-relation-class, and mapping function completion data dictionary automatic build.

The top layer is the calculation layer, which is divided into Jaro distance algorithm, Euclidean distance algorithm, and Levenshtein distance algorithm. After each distance is calculated by the calculation layer, it is multiplied by the weighting matrix, and the final suggested distance and the original three distances are output through the weighting algorithm.

The weighting operation is as follows in (5)

$$D = \frac{W \times (O + J + L)}{3} \tag{5}$$

Among them, D is the output comprehensive distance value, O is the Euclidean space distance matrix, J is the Jaro distance matrix, L is the Levenshtein matrix, W is the variable weighting matrix, and finally the quadruple (D, J, O, L) is output, which is the actual weighted distance output.

A neural network method is used to train the distance threshold of the dataset, and the threshold quadruple F_{th} is obtained. Set the fusion rule R in (6)

$$R = \begin{cases} 1, & if(D, J, O, L) < F_{th} \\ 0, & otherwise \end{cases} \tag{6}$$

When the value of the fusion rule is 1, the node fusion of multi-source data can be realized, otherwise it is considered to be two data of different types or meanings and cannot be fused.

5. EXPERIMENT AND RESULTS

The essence of industrial digital twin data fusion is to realize the unambiguous, barrier-free interconnection and intercommunication of the whole process data in the production process, including the full coverage of human-machine-material-law data. However, the business process of the industrial production process is not organized in the order of the human-machine-material law circle. Therefore, the formulation of the classification data dictionary according to the relevant systems of the industrial production process will help the unified definition and description of the actual production data. , that is, the factors such as human, machine, material, law and other factors are dispersed into each system according to the application scope and characteristics of the system, which can achieve full coverage of the production data dictionary in the industrial digital twin. Generally speaking, the industrial digital twin will include on-site acquisition, WMS related to raw material logistics, MES related to production task allocation, report data related to task flow and statistics, and QMS system related to production quality. Based on the above classification, this paper constructs a production process data dictionary including seven categories, as shown in the Figure 3 below, which is the definition of the data dictionary related to inventory management in the WMS system, including the inventory management entry itself and its associated relationship. Class and related description properties.

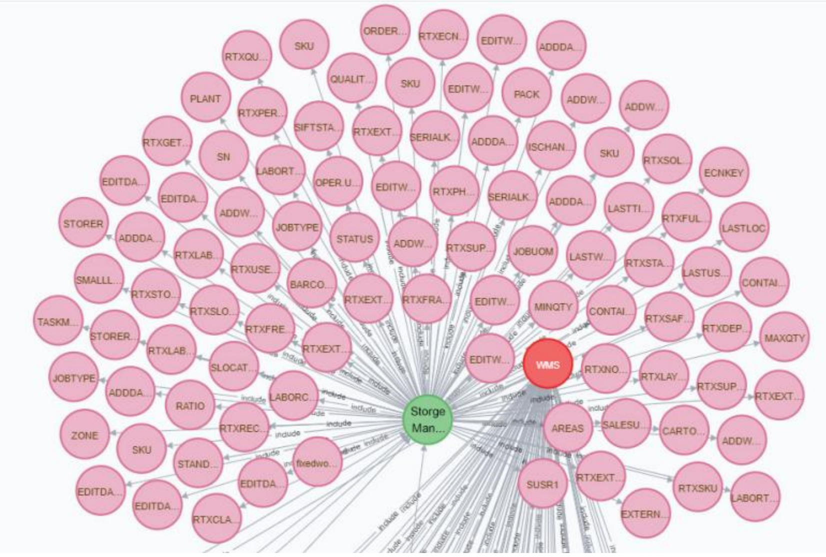

Figure 3. Data dictionary example

As can be seen from Figure 3, the semantic data dictionary can not only express the meaning of the entry itself, but also express the relationship between attributes and other classes, and will uniformly express and explain the multi-source and heterogeneous actual production data. Association building provides strong support.

The data set tested in this paper includes a total of 12,123 entry nodes and 22,812 association relationships. The data comes from the enterprise production system WMS, MES and QMS. When the data comes from one or more production systems and the amount of data is different, the accuracy of the actual data matching is different. When the amount of data is larger and the data source system is single, the accuracy rate is higher, as shown in Table 1.

TABLE I. THE ACCURACY OF DATA MATCHING WITH DIFFERENT DATA VOLUME AND SOURCE

Data volumes	**Data Source**				
	MES	***MES+QMS***	***MES+WMS***	***QMS+WMS***	***MES+WMS+QMS***
1000	75.2%	68.1%	69.5%	69.6%	58.3%
3000	97.5%	78.5%	80.3%	79.6%	66.9%
5000	98.6%	85.6%	88.6%	89.3%	70.5%
10000	99.8%	89.9%	91.2%	90.7%	72.6%

It can be seen from Table 1 that when the data source is the mixed data of MES+WMS+QMS and the data volume is 10,000, the data matching accuracy rate is 72.6%. This low accuracy rate cannot meet the application needs of enterprises. The production process data dictionary, under the guidance of the dictionary, uses the actual database table of the enterprise production line to perform automatic semantic matching and association relationship construction. The data is randomly selected from the three systems, and the data matching accuracy results are as follows in Figure 4.

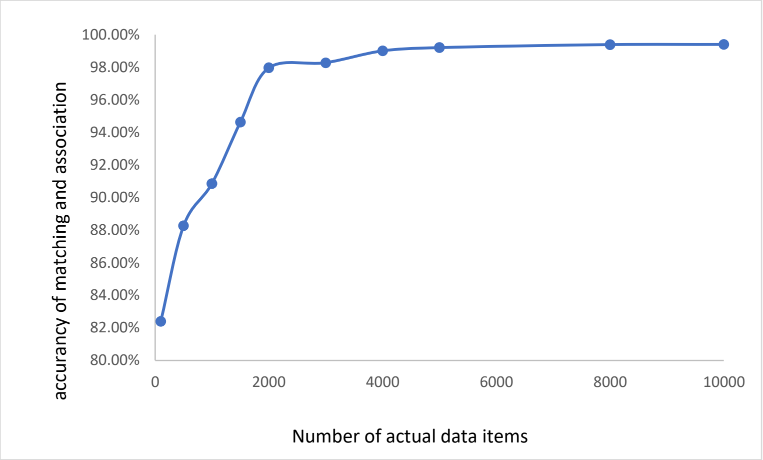

Figure 4. Actual data matching and association accuracy based on data dictionary

As can be seen from the Figure 4, based on the domain data dictionary, the actual production data can be described in a unified way and the association relationship can be constructed. The accuracy of matching and association construction increases gradually with the increase of the number of actual data items, indicating that when there is a more comprehensive business After participating in the production data of the data dictionary, the matching and fusion performance of the data dictionary will be improved. When the amount of data reaches more than 1000, the accuracy rate can reach more than 99%, which is much higher than the data matching accuracy rate 72.6% without the data dictionary. It proves the effectiveness of this method.

6. SUMMARY

This paper proposes a digital twin data fusion method based on domain data dictionary. Firstly, aiming at the fusion goal of industrial digital twin data, according to the characteristics of industrial digital production process, a data dictionary model in the field of production process data is designed, a semantic method is introduced, and a semantic description modeling method based on data dictionary is proposed. Then, based on the semantic data dictionary, a data fusion method based on multivariate distance weighted semantic similarity calculation is proposed. Finally, according to the data fusion requirements of industrial digital twins, the fusion method proposed in this paper is tested by using the actual production system data. In future work, we will continue to carry out research work on automatic semantic construction of data dictionary in a wider range of fields.

ACKNOWLEDGMENT

This research was funded by the National Key R&D Program of China under Grant 2018YFB2003200, and National Natural Science Foundation of China 92067110.

REFERENCES

[1] Yan Jianlin, Kong Dejing. Study on “Industrial Internet” and “Industrie 4.0”. Strategic Study of CAE,2015,17(07):141-144.

[2] Tao Fei, Cheng Ying, Cheng Jiangfeng. Theories and technologies for cyber-physical fusion in digital twin shop-floor. Computer Integrated Manufacturing System, 2017, 23(08): 1603-1611.

[3] Henry Canaday, Li Yun. The key to digital twin technology is data. Aviation Maintenance & Engineering, 2019(10): 15-16.

[4] Zhuang Cunbo, Liu Jianhua, Xiong Hui. Connotation, architecture and trends of product digital twin. Computer Integrated Manufacturing Systems, 2017, 23(04): 753-768.

[5] Yu, Q. Yang, G. Wang and Y. Xie, "A Novel Discriminative Dictionary Pair Learning Constrained by Ordinal Locality for Mixed Frequency Data Classification," in IEEE Transactions on Knowledge and Data Engineering.

[6] P. W. Shin, J. Lee and S. H. Hwang, "Data Governance on Business/Data Dictionary using Machine Learning and Statistics," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 547-552

[7] J. Chen, F. Zhou, Z. Guo and J. Wan, "Compressed Data Collection Method for Wireless Sensor Networks Based on Optimized Dictionary Updating Learning," in IEEE Access, vol. 8, pp. 205124-205135, 2020

[8] Zhou Zhenjuan, Diao Lianwang. An Online iterative Clustering Method for Multi Sensor Consistent Data Fusion. Computer Measurement & Control, 2021,29(02): 251-255.

[9] J. Wang, Y. Lan, S. Zhang, Y. Xia, C. Wu and L. Chen, "Knowledge Graph for Multi-source Data Fusion Topics Research," 2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), 2020, pp. 1-5

[10] P. Chu, Z. Dong, Y. Chen, C. Yu and Y. Huang, "Research on Multi-source Data Fusion and Mining Based on Big Data," 2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), 2020, pp. 606-609

[11] W. Wang and H. Gong, "Fast fusion method of marine environment vector data based on BP neural network," 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing), 2021, pp. 1-7

[12] Y. Zhu, D. Chen, C. Zhou, L. Lu and X. Duan, "A knowledge graph based construction method for Digital Twin Network," 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI), 2021, pp. 362-365

[13] O. G. Brylina, N. N. Kuzmina and K. V. Osintsev, "Modeling as the Foundation of Digital Twins," 2020 Global Smart Industry Conference (GloSIC), 2020, pp. 276-280

[14] Chen J, Song S, Yu H. An indoor multi-source fusion positioning approach based on PDR/MM/WiFi. Aeu-International Journal of Electronics and Communications, 2021,135:153733

[15] Liu X, Jiao L, Li L, et al. Deep multi-level fusion network for multi-source image pixel-wise classification. Knowledge-Based Systems, 2021,221:106921.

[16] Euzenat J, Guégan P, Valtchev P. OLA in the OAEI 2005 alignment contest. Workshop on Integrating Ontologies, 2005:1789-1808.

[17] M. T. Khadir A D A W. XMap++: A novel semantic approach for alignment of OWL-Full ontologies based on semantic relationship using WordNet. International Symposium on Innovations in Information and Communications Technology, 2011,12(9):13-18.