Research on the method of data architecture construction for joint intelligent simulation

Suike Li, Yongchao Zhao and Haijun Wang

National Defense University, Beijing, 100091, China
*Corresponding author’s e-mail: lisk@mail.nwpu.edu.cn

Abstract. In view of the data problems restricting the joint intelligent simulation, this paper constructs a data classification system. By mining and sorting the data, rules and models in cases and experiences, and combining with the basic data and ability data of the relevant system simulation model, it carries on the digital and systematic annotation, constructs the conceptual model, establishes the segmentation and part of speech standards, and adopts the methods of concept mining, attribute mining and instance mining. The data system of joint intelligent simulation is established by mining and instance mapping, and the multi-layer logical template and intelligent matching rules are used to solve the data problems of joint intelligent simulation.

1. Introduction
In the complex system, the traditional equipment single performance simulation has been unable to truly reflect the practical efficiency of contemporary equipment, and the equipment with superior single performance index is not necessarily powerful in practical application. The key factor that affects the success or failure is the equipment system with the best overall efficiency, which is related by the information system. Model and simulation method has become the mainstream method to study system of systems countermeasure. At the same time, due to the complexity, uncertainty, chaos, emergence, integrity and other characteristics of the system, the traditional model and simulation methods cannot meet the requirements of complex system model. Intelligent simulation has the ability of self-learning, self-adaptation and self-organization. With the improvement of the theoretical system of intelligent simulation science and technology, the deepening of the technical system and the popularization of application, it provides a new method to solve the problem of complex large-scale system model[1]. The core of intelligent simulation is learning algorithm, and the basis is data. Intelligent simulation system starts with data, trains with data, and expresses with data. At present, the system integration of various data sources has become a key problem restricting the joint intelligent simulation, which has not been well solved for a long time.

This paper aims at building a data system for joint intelligent simulation, analyses the data representation of each model data source, automatic construction and management of data system, intelligent retrieval based on data system, self-evolution of data system for cognition, etc., and tries to break through various technical barriers in the construction of existing data system, and form an unstructured and related joint intelligent simulation model according to the methodology of system.
2. Joint intelligent simulation and data architecture requirement analysis

2.1. The purpose of joint intelligent simulation
At present, there are many simulations for single equipment or multiple equipment of the same type, such as the evaluation of aircraft performance and vehicle mobility [2]. A single simulation system or simulator only has a single simulation function. A large number of simulation systems supporting different applications have different systems, different standards and poor interoperability, which cannot realize the integrated modelling and simulation of joint environment and joint action. Therefore, joint intelligent simulation is urgently needed to meet the requirements of future system of systems confrontation.

2.1.1. The purpose of the joint simulation. The joint simulation is to break through the barriers between simulation systems in different fields, at different levels, by different means and for different purposes, comprehensively integrate infrastructure resources such as training ground, laboratory and equipment test range, continuously strengthen the interconnection among simulation systems, information systems and data centers, and strive to build a comprehensive simulation system that can meet the multi-purpose requirements of training, application and test integrated joint simulation environment [3].

2.1.2. The purpose of the intelligence. Artificial intelligence aided simulation is used to introduce artificial intelligence into simulation, assist simulation modelling, simulation interaction and simulation analysis, and improve the level of intelligence and integration of system modelling and simulation [4]. It is a kind of modeling and simulation system which adopts the deep integration of advanced artificial intelligence technology and modeling and simulation technology as the technical means and aims at optimizing the overall performance of complex system modeling, simulation operation and result analysis. Its development is promoted by hardware, software, algorithm and application of high performance modeling and simulation system. It has many characteristics, such as multi sample analysis, super real-time simulation, causal sequence coordination, complex model solution, and it realizes system demonstration, system efficiency evaluation and application research in complex electromagnetic environment.

2.2. Data architecture requirement analysis of joint intelligent simulation
Due to the high complexity of its mechanism, intelligent system and artificial intelligence system are often difficult to establish its system principle model through mechanism (analytical method), but need to simulate and simulate its internal mechanism through a large number of experiments and application data. The modelling method based on big data intelligence is a kind of method that uses massive observation and application data to realize effective simulation modelling of intelligent system with unclear mechanism. The data involves the daily construction, training and test accumulated platform efficiency, confrontation efficiency, personnel ability, model, rule and other ability data and relevant auxiliary decision-making calculation model. The simulation data comes from the quantitative expression of the relevant information of each simulated objective object, which refers to the recording of the nature, state and mutual relationship of all kinds of objective things involved in the whole process of simulation a physical symbol or combination of physical symbols.

The data system of joint intelligent simulation provides support for joint simulation by mining and sorting the data, models and components of each simulation system, and combing the rules, aiming at solving the problems of knowledge memory and logical reasoning. At the same time, combined with data engineering, it uniformly represents, manages and queries the elements of power, ability and resources to form the data elements of joint simulation To build a general data generation mechanism and build a general data generation tool.
3. Framework of data architecture for joint intelligent simulation

In order to meet the needs of various business processing and management, and improve the level of automatic processing and management, many application systems have been built in the field of simulation, and many of the key data that need to be shared and interacted are enclosed in independent systems, forming what is called "information islands". In order to break the "information island" phenomenon, realize the integration and sharing of data for a wide range of data sources, and build a data system for joint intelligent simulation, systematic technical methods are needed to carry out systematic data analysis and processing. This paper focuses on the following three aspects, as shown in Figure 1.

![Figure 1. Framework of data architecture construction technology for joint intelligent simulation.](image_url)

3.1. Construction of digital and systematic annotation classification system

The main task is to digitize and standardize the rules and other unstructured resources and the simulation model data. In the digital part, it is necessary to label all potential data items of events, cases and experiences according to the relevant formats, and label, extract and structure the content with special meaning such as basic data, capability data, model data and rule data. In the part of constructing classification system, according to the essential attributes of joint simulation, a generic concept is divided into several kinds of concepts, which are divided into different types of data systems according to the same, different, related and other attributes, so as to clarify its due position and mutual relationship in the data system, and establish a sequential, standardized and systematic data classification and annotation system [5].

3.2. Construction of conceptual model of data system

Based on the data classification system of joint simulation, a big data system oriented to joint simulation is constructed. The logical relationship of data in the process of joint simulation is constrained and organized by establishing a conceptual model, and the relationship between concepts is organized by attributes with practical meaning, not just the simple relationship between upper and lower levels. One is the establishment of data classification architecture, the other is the generation of ontology conceptual model. For the first part, we need to analyse and mine the differences and relations between data items, design the attribute set of concepts, and reasonably arrange the distribution of attributes among concepts. For the second part, the extracted concepts are sorted out, the attribute structure between concepts is
designed, the compatibility of instance extraction methods between cross domain data sources is considered, and the general matching and linking of top-down and bottom-up instances are achieved.

3.3. Construction of entity library for joint intelligent simulation
According to the conceptual model of data system for joint intelligent simulation, a data system database covering all elements of joint intelligent simulation is constructed through concept mining, attribute mining and instance mining. The specific work of concept mining includes the analysis and mining of concept instances such as data, rules and models according to the concept model. The specific work of attribute mining includes extracting attributes from the above concepts and analysing the properties of these attributes. The concrete work of case mining is to mine the instances of the above concepts and the relationships among them. Through the above steps, the core part of the joint intelligent simulation data system is constructed, and then the content of the data system is expanded according to the multi-source data associated with it, and in turn, the massive multi-source data is linked.

4. Key technologies of data architecture for joint intelligent simulation
According to the conceptual model and classification architecture, there are many technologies involved in the construction of joint intelligent simulation data system for data engineering, joint data resources, business data resources, command information system and all kinds of simulation systems. The key technologies include domain keyword table construction technology, concept mining technology based on rule mapping and statistical learning, case mining technology based on rule induction and sequence annotation, automatic case mapping technology and information retrieval technology.

4.1. Construction technology of domain keyword list
Domain keyword extraction will be constructed by the method of word net and statistical learning model. The word web contains many semantic relations between words, Semantic relations include: synonym, similar, antonym, opposite, hypernym, hyponym, meronym, attribute, material, implication and derived from, cause, also see, user, possessor, goal, location, etc.

Word net can achieve fast segmentation of common words, while statistical learning model can achieve ambiguity elimination through the optimal choice of word segmentation. In the unsupervised keyword extraction method, by using the univariate, binary and ternary grammar rules, based on the linguistic characteristics of the extended TF-IDF (term frequency inverse document frequency) keyword weighting calculation method, the extraction of unknown words is implemented and the optimization strategy of keyword extraction is implemented.

4.2. Automatic Concept Mining Technology of Data system
The named instance recognition technology based on Hidden Markov Model (HMM) is adopted to label the concept of unstructured text. As shown in figure 2, implied variable $S_t$ represents named instance representation, including named instance start identifier, named instance topic identifier, named instance end identifier, and unnamed instance identifier. There is a linear dependency relationship between implied variables, observation variable $O_t$ represents the observable words in the current text, and there is a generation relationship between implied variables and observation variables.

![Figure 2. Schematic diagram of HMM](image)
The distributed model is used to identify the explicit and implicit relationship between upper and lower concepts. Binary classification distribution methods are used, such as neural network (NN), logistic regression (LR) and support vector machine (SVM), to train the data model of implicit upper and lower concept relationship.

4.3. Automatic Case Mining Technology of Data system

Semantic annotation based on similar rule induction mainly includes rule learning module, annotation module and interpretation module. The rule learning module uses SRL (Similarity based Rule Learner) algorithm, uses cross validation technology to select the size of the rule context window, and establishes the corresponding initial annotation rule set. Then, in the process of rule learning, SRL always searches for the most similar rules for induction, so as to avoid the problems existing in the traditional method of random induction. Finally, SRL prunes the redundant rules and outputs the learned rule set. The annotation module uses the learned annotation rules to annotate. The most important SRL algorithm includes four sub processes: pre-processing, rule set initialization, rule induction and rule pruning. The input is the labelled training document, and the reprocessing module uses natural language processing (NLP) technology to pre-process the training document. The rule set initialization module takes the pre-processed result as the input, uses the dynamic window technology to select the window size, and generates the initial rule set. In each iteration, the algorithm tries to find the most recent rules in the process of rule induction. Finally, the rules are pruned to evaluate the learned rules, delete the low credibility rules, and output the annotation rule set.

4.4. Automatic Case Mapping Technology of Data system

Automatic instance mapping is to automatically associate the instances of two datasets, and use multi-level instance features to automatically identify the consistency of instances. The main features of automatic instance mapping are semantic relevance and semantic similarity, so as to associate simulation capability elements with massive multi-source data.

4.4.1. Semantic relevance calculation.

The correlation degree between two instances is calculated by the instance correlation degree. Given two instances $e_1$ and $e_2$, the calculation formula of semantic correlation degree is as follows:

$$\text{SmtAss}(e_1, e_2) = 1 - \frac{\log(\max(\|E_1\|, \|E_2\|)) - \log(\|E_1 \cap E_2\|)}{\log(\|W\|) - \log(\min(\|E_1\|, \|E_2\|))}$$

In formula (1), $E_1$ and $E_2$ are the set of instances directly connected with $e_1$ and $e_2$, and $W$ is the complete set of instances. From the formula, we can see that the numerator is similar to the Jaccard correlation. We can conclude that when two instances are connected with more common instances, the semantic correlation between instances is higher. The semantic relevance of an instance is the average distance between the instance and the ontology knowledge base instance set $\Gamma_d$, the formula is as follows:

$$\text{SA}(e) = \frac{\sum_{cc \in \Phi_d} \text{SmtAss}(cc, e)}{|\Phi_d|}$$

4.4.2. Semantic similarity calculation.

The significance of semantic similarity calculation is to analyze the similarity of instances in the concept set. In this way, calculating the semantic similarity of instances becomes calculating the semantic similarity of two super class sets. The steps of semantic similarity calculation are as follows:

Step.1: Calculate the super class set $\Phi_{e_1}$, $\Phi_{e_2}$ of two instances $e_1$ and $e_2$.

Step.2: Calculate a superclass set to another superclass set target class as follows:

$$e(C_1) = \arg \max_{C_2 \in \Phi_{e_2}} \text{sim}(C_1, C_2)$$

(3)
In formula (3), $sim(C_1, C_2)$ is the semantic similarity between two classes, and $\varepsilon(C_1)$ is the class in $\Phi_{e_2}$ that can maximize the similarity between $\Phi_{e_1}$ and $\Phi_{e_2}$. The semantic similarity of two classes is calculated as follows:

$$sim(C_1, C_2) = \frac{2 \times \log(p(C_1))}{log(p(C_1)) + log(p(C_2))} \quad (4)$$

In formula (4), $C_0$ is the root of the minimum concept tree of $C_1$ and $C_2$, and $P(C)$ is the probability that a class belongs to the minimum concept tree with $C$ as the root.

Step.3: Calculate the semantic similarity from one class to another as follows:

$$sim(\Phi_{e_1} \rightarrow \Phi_{e_2}) = \frac{\sum_{C_i \in \Phi_{e_1}} sim(C_i, \varepsilon(C_1))}{|\Phi_{e_1}|} \quad (5)$$

Step.4: Calculate the semantic similarity between two instances, the formula is as follows:

$$SmtSim(e_1, e_2) = \frac{sim(\Phi_{e_1} \rightarrow \Phi_{e_2}) + sim(\Phi_{e_2} \rightarrow \Phi_{e_1})}{2} \quad (6)$$

Step.5: At the last step, the average semantic similarity between an instance and all other instances of ontology a knowledge base is calculated, the formula is as follows:

$$SS(e) = \frac{\sum_{\varepsilon_{e_k}} SmtSim(cc,e)}{k} \quad (7)$$

4.5. High performance information retrieval technology for massive data

The strategy of keyword retrieval based on ontology information is adopted. Firstly, the semantic information of RDF(Resource Description Framework) data is fully described by constructing an index structure suitable for retrieving massive RDF knowledge data. Then, semantic recognition is carried out according to the keywords entered by users, and the user's query intention is further mined. Finally, by improving the existing sorting algorithm and increasing the weight of ontology semantics in sorting, the relevance between RDF retrieval results and user queries is judged to further optimize the sorting results.

5. Conclusion

Data is the key content of simulation system construction, and data system is the basic content of joint intelligent simulation, which has important practical significance. In order to solve the data problem of joint intelligent simulation, the purpose of joint intelligent simulation and the requirements of joint intelligent simulation data system are first defined. And then, the basic data and capability data related to intelligent simulation model are considered, the digital and systematic annotation is carried out, the concept model is constructed, the segmentation and part of speech standards are established, the data system is established by using concept mining, attribute mining, instance mining and instance mapping. At last, the multi-layer logic template and intelligent matching rules are elaborated, and the basic content and requirements of the data problem of joint intelligent simulation are answered. The theory and method provide a way of thinking for the development of data architecture of joint intelligent simulation, so as to ensure the efficiency of joint intelligent simulation.

Acknowledgments

This paper is one of the phased achievements of the national social science fund military science project "Research on intelligent evolution and application of joint operation command information system" (2019-SKJJ-C-063).

References

[1] Li B.H., Chai X.D., Zhang L., et al. (2018) Preliminary study on Modelling and simulation technology for new artificial intelligence system. Journal of system simulation, 02: 349-362.

[2] Fu Y.H., Xing J.J. (2015) Research on joint test of weapon equipment system. Military operations
research and systems engineering, 29: 65-70.

[3] Chen S.Q., Fan R.X., Huang B., et al. (2016) Implementation method and engineering application of AMESim and fluent joint simulation technology. Missile and space delivery technology, 01: 92-97.

[4] Hu X.F..(2013) Challenge and thinking of modelling and Simulation in the era of big data. Military operations research and systems engineering, 27: 5-12.

[5] Shi Y.ZH., Wang SH.T., Wang J., et al. (2013) Fast classification algorithm for non-static large data sets based on minimum inclusion sphere. Control and decision, 28: 1065-1072.