Construction and Application Research of Knowledge Graph in Spacecraft Launch

Kou Chao, Liu Tao, Ma Li, Jiang Guoyu, Wang yuchao, Zhao Yu
Xichang Satellite Launch Center, Xichang, 615000

Abstract: In this paper, through analyzing the necessity of constructing a knowledge graph in spacecraft launch, we construct and use a knowledge graph in spacecraft launch, which serves as not only an effective way of solving the problem of "sparse knowledge" and "incomplete knowledge" in a sea of data but also a basis for achieving Semantic Artificial Intelligence (Semantic AI) and complex data analytics and applications. On the basis of understanding the basic technical architecture of knowledge graph, this paper analyzes the construction approaches of the knowledge graph in spacecraft launch in terms of knowledge source, modeling, extraction, fusion, inference, and storage. By means of the knowledge graph, automatic question-answering based on semantic search, fault detection based on machine learning, and portraits of equipment or information system can be achieved at the launch site. In doing so, the intelligent data engineering and construction at the launch site can be improved.

1. INTRODUCTION
The continuous construction of the data center at the spacecraft launch site makes it possible to manage the data resources at the launch site through a dynamic data catalog, thus strengthening data protection in terms of security and confidentiality. As the themes of the data warehouse of the testing mission are continuously expanded, the comprehensive collection and consolidation of historical data and information can be achieved. As the construction of the data center progressed, decision-makers and commanders were drowned in a sea of data. Owing to it, the problem of "sparse knowledge" appeared so that it became increasingly difficult to obtain valuable information. On the other hand, due to the lack of support for data correlation, knowledge discovery, and knowledge reasoning in the process of intelligent analysis and decision-making, the problem of "incomplete knowledge" came into being. In accordance with the engineering and construction direction of intelligent testing mission data and the demand for "expertise and intelligence" in the construction of the back-end of the information system, we carried out in-depth research on the comprehensive utilization of test data and information, realized the expansion from data storage to knowledge storage, and addressed the difficulties in obtaining data and information through traditional channels. As a result, a knowledge graph in spacecraft launch was constructed and utilized, which serves as a basis for the implementation of Semantic AI and various complex analytics and applications.

A knowledge graph is a structured semantic knowledge base, usually in the basic ternary representation form of "entity-relationship-entity" or "entity-attribute-value". The nodes in the knowledge graph represent real-world concepts or entities, while the edges in the network represent semantic relationships between these concepts or entities, which are essentially collections of real-world concepts and relationships. According to the different contents and coverage of knowledge, knowledge graphs can be divided into two types, which are universal knowledge graphs and industry knowledge graphs. Regarding different construction methods such as manual construction, construction based on...
encyclopedic knowledge, and construction based on open knowledge extraction, universal knowledge graphs emphasize the breadth of knowledge, typically including Google Knowledge Graph, YAGO, NELL. In the universal Chinese knowledge graph, CN-DBpedia, which is developed and maintained by the Knowledge Factory Laboratory of Fudan University, Sogou Knowledge Cube, Baidu Zhixin are representative examples. The industry knowledge graph is built based on the data of a specific industry and emphasizes the depth of knowledge, which has been widely applied in e-commerce, medical care, finance, agriculture, and other fields.

2. Technical Architecture of the Knowledge Graph

Logically, the technical architecture of the knowledge graph can be divided into four layers: data layer, knowledge layer, service layer, and application layer, as shown in Figure 1. The data layer mainly solves the problem of data sources, including how to introduce structured, semi-structured, textual data, and multimodal data into the knowledge graph. Through mining a series of data such as historical test data, test documents, teaching materials, design plans, national military standards of various space launch systems in the spacecraft launch field, information from different sources and of different types can be linked to form a relation network and interact with the data stored in the data center at the launch site. The knowledge layer mainly studies and solves the key technologies in the construction and application of knowledge graph, which is the research focus in the whole technical architecture of knowledge graph, including knowledge representation, knowledge modeling, knowledge extraction, knowledge fusion, and knowledge processing. By taking full advantage of relevant technologies in the knowledge layer, the service layer mainly provides service technologies such as knowledge storage, search, computation, and inference for the application layer. The application layer makes use of the knowledge graph system and carries out fault detection, intelligent question-answering, user portrait, and other applications in accordance with users’ needs.

3. Construction of Spacecraft Launch Knowledge Graph

The knowledge graph can be constructed in two approaches: top-down and bottom-up. For the top-down approach, the ontology library and database schema are defined first, and then a series of facts are added to the knowledge base. For the bottom-up approach, high-quality data sources are first extracted by technical means, and then the new schema with a higher degree of confidence is selected, which will be added to the knowledge base after manual review. Universal knowledge graphs are mostly constructed in a bottom-up approach, while industry knowledge graphs are mostly constructed in a top-down approach because of high requirements for domain expertise and complex and changing business needs. The spacecraft launch knowledge graph can be constructed in a top-down manner at the initial stage of construction and can be continuously enriched and improved in a bottom-up manner subsequently.

3.1 Data Sources

Given the special nature of knowledge in the spacecraft launch field, the spacecraft launch knowledge
graph is composed of two parts: domain-specific knowledge and domain-general knowledge.

Domain-specific knowledge mainly refers to the knowledge related to rocket and special equipment models, including not only a series of terms for rockets, charging pipes, special equipment or systems but also the codes representing system units or operating status measurement points (such as telemetry parameters and Filling valves). For these terms or codes have no corresponding knowledge in external corpus and lack representations in the Semantic Web, the semantic computation is hard to perform. However, these codes are usually backed by a series of flight, test, or operating status data, which is unique to the domain-specific knowledge in spacecraft launch. For example, most of the rocket test data and flight telemetry data are characterized as time series data, which are comparable and calculable and have obvious correlations that can form a relationship network of rocket system knowledge.

Domain-general knowledge refers to the knowledge that is closely related to spacecraft launch but whose corresponding data can be found in the external corpus, such as spacecraft orbital knowledge, geophysical parameters, radar performance parameters, and basic communication terms. The extraction of entities, attributes, and relationships of these types of knowledge can be achieved by virtue of the construction method of a universal knowledge graph. Represented by word vectors, these entities or relationships can be mapped into a unified space and dense low-dimensional vectors can be utilized to represent semantic information. By means of natural language processing algorithms, the semantic relationships between entities, between relationships, and between entities and relationships can be efficiently computed so that the organic integration of dispersed knowledge can be achieved.

3.2 Knowledge Modeling

The knowledge graph can be logically divided into two layers: the schema layer and the data layer. The schema layer on top of the data layer stores the condensed knowledge, and the data is usually stored and managed in the form of an ontology library. The schema layer defines the structure of the entire knowledge graph so that its reliability needs to be ensured. The data layer contains a set of relationships between entities, between relationships, and between entities and relationships.

(1) Ontology Construction at the Schema Layer

Ontology was originally a concept in the field of philosophy that referred to the conceptualization of entities and their relationships that exist in the real world. In knowledge graphs, ontology is a form of presenting knowledge in a domain model. The conceptual diagram of ontology in knowledge graphs is shown in Figure 2.

![Figure 2 Conceptual Diagram of Ontology](image)

There are many methods for developing ontology models, including the assessment method, the Skeleton method, METHONTOLOGY, the Seven-step method, and the Berneras method. The top-down Seven-step ontology construction process of the spacecraft launch knowledge graph is shown in figure 3.
There are various tools for ontology construction. The most commonly used is protégé, an open source editor for ontology construction and storage of knowledge graphs. It can assist users not only in the operation and construction of ontologies but also in the visualization of the ontologies generated, including a hierarchical concept tree and a graphical interface of entities, entity attributes, and entity relationships of ontology concepts. Based on the ontology hierarchy, the query model can be used for ontological reasoning. Also, RDF and OWL are supported for reading and writing ontologies.

(2) Abstract Modeling at the Data layer

Abstract modeling at the data layer mainly includes three aspects: First, entity extraction and merging. Taking entities or concepts as the main target, map and merge data from multiple sources. Second, attribute mapping and merging. Use attributes to represent the descriptions of entities in different data sources with the aim of forming an all-round description of entities. Third, entity-relationship extraction. Use relationships to describe the relationships between various types of data abstractly modeled as entities so as to support correlation analysis.

3.3 Knowledge Extraction

Knowledge extraction refers to extracting knowledge from data from different sources and different structures to form knowledge to be deposited in the knowledge graph. In general, the channels and means of knowledge extraction are shown in Figure 4.
Figure 4 Channels and Means of Knowledge Extraction

(1) Structured Data Extraction

Structured data extraction includes linked data extraction and database data extraction. Linked data mainly includes web links and task document links in the data center, which can be directly mapped into the Resource Description Framework (RDF) ternary groups through graph mapping. For database data extraction, D2R tools can be adopted to automatically convert the relational database into the virtual RDF database. The difficulty of the problem lies in automatic corresponding with the knowledge modeling results and merging with other knowledge components.

(2) Semi-Structured Data Extraction

Semi-structured data is mostly presented in the form of tables, lists, web pages generated from templates. Knowledge in the domain of spacecraft launch is more often found in tables or lists of task information. This type of data is usually configured with corresponding wrappers for different structures to complete the data parsing. The wrappers can learn automatically. To ensure reliability, a combination of humans and machines is generally used. Some researchers designed a wrapper for the automated extraction of knowledge from a table. The test result showed that for simple tables, the extraction accuracy of entities, relationships, and ternary groups was over 90%, while for complex tables, the extraction accuracy was over 80%.

(3) Textual Data Extraction

Knowledge extraction tools for textual data include ReVerb, TextRunner, and DeepDive. Knowledge extraction of textual data mainly includes entity extraction and relationship and attribute extraction.

1) Entity Extraction

An entity, also known as a named entity, refers to a noun with a defined meaning. In the spacecraft launch field, an entity can also be a special code. Entity extraction is the automatic identification of named entities from the original corpus. The major methods are shown as follows.

Entity extraction method based on rules and dictionaries. For example, when it comes to organization and personnel, set "**Department", "**Station", "**Office", "**System", "**Position", "**Engineer", etc. The advantages of this entity extraction method are high accuracy and fast search speed, while the disadvantages are limited rule coverage, poor portability, and complex rule construction.

Entity extraction method based on statistics. The basic idea of this method is to use an annotated corpus to train the probability of a certain word as a component of a named entity, and then calculate the probability of the candidate word for being named entities. If the probability is greater than a certain threshold, the word is identified as a named entity. This method has good robustness, flexibility, and portability, while the disadvantage is that it is difficult to annotate the training corpus.

Entity extraction method combining statistics and rules. By combining the above two methods, the complexity and blindness of the rule method can be reduced by probability calculations in the statistical method. On the other hand, the requirements of the statistical method on large-scale corpus can be
reduced by the multiplexing of rules.

Entity extraction method based on deep learning. BIO sequence annotation of text characters in the domain can be conducted first. The long short-term memory (LSTM) in the recurrent neural network (RNN) can be used for entity extraction. The disadvantage of this method is the same as that of the statistical method, which is that a large amount of corpus needs to be annotated for training.

2) Relationship and Attribute Extraction

Extraction method based on rules and pattern-matching. For example, YAGO uses regular expressions to extract relationships or attributes from annotated information and uses pattern matching to identify relationships between entities by manually constructing syntax and semantic rules. The shortcoming is the high workload and the need for a deep understanding and knowledge of a particular domain.

Extraction method based on statistical machine learning. By modeling the pattern of relationships between entities, relationships can be identified with the help of machine learning techniques. Statistical machine learning can be divided into three parts, which are supervised learning, semi-supervised learning, and unsupervised learning in terms of the amount of corpus to be trained and labeled. Unsupervised learning can be achieved by using a clustering-based method to compute the similarity of entity contexts.

(4) Domain-Specific Knowledge Extraction

Domain-specific knowledge in spacecraft launch includes rocket test data, rocket flight data, ground fueling data, equipment health data, and communication monitoring data, which are mostly stored in a time-value manner. Data correlation mining can be conducted by virtue of correlation analysis. For example, for rocket flight telemetry data, in addition to using a rocket data dictionary to extract explicit correlations to form a ternary group, the correlation analysis method can also be used to verify explicit correlations and mine invisible correlations. Commonly used data correlation analysis methods include covariance, Pearson, Spearman, Kendall, and other correlation coefficient matrices, where Pearson's correlation coefficient can be calculated as follows.

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

Pearson’s correlation coefficient ranges from -1 to +1. After taking absolute values, the correlations can be classified into five levels, which are very strong correlation (0.8-1.0), strong correlation (0.6-0.8), moderate correlation (0.4-0.6), weak correlation (0.2-0.4), very weak correlation, or no correlation (0.0-0.2). The correlation analysis was conducted for the outlet pressure of the attitude control reducer and the pressure of the attitude control storage tank in four combat flights of the CZ-3B rocket, both of which were the power system parameters. The two yielded correlation coefficients were above 0.9, which was a very strong correlation and consistent with the analysis results in the rocket parameter dictionary. Based on the results of the correlation analysis to assist domain experts in describing the parameter relationships, a ternary group of telemetry parameters as entities can be formed to be stored in the knowledge graph.

3.4 Knowledge Fusion

Knowledge fusion is a relatively complex task, including the fusion at the schema layer (the fusion of concepts, their contextual relationships, and their attributes) and the data layer. Since the fusion at the schema layer shall be reviewed by domain experts during the ontology construction process, it has a high degree of reliability. Since knowledge sources are scattered and of varying quality, the key to knowledge fusion lies in the fusion at the data layer, including entity disambiguation and entity alignment, which are generally integrated combined with the entity extraction process.

(1) Entity Disambiguation

Entity disambiguation is to address the situation where a certain entity corresponds to multiple named entities in a real language environment. For example, there are multiple "technical room" and "system group" in an organization and the key flight time sequence "TK" value of multiple rocket types. The solutions can be divided into five categories based on data characteristics: disambiguation based on
entity salience, contextual similarity, entity relevance, deep learning algorithms, and special annotated resource [3].

(2) Entity alignment

Entity alignment is to address the situation where an entity corresponds to multiple different names in a real language environment. For example, the ballistic missile shutdown point is also called the ballistic missile feature point in some rocket models, and some parameter codes are written in different letter cases. In the entity alignment process, data can be standardized through data preprocessing, and then the computational complexity can be reduced through partitioning indexes. By means of textual similarity computation, entity pair alignment can be achieved. By taking advantage of structural similarity computation, collective entity alignment can be achieved.

3.5 Knowledge Reasoning

(1) Methods of Knowledge Reasoning

Based on the existing entity and relationship data, knowledge reasoning refers to the process of establishing new relationships between entities through automated reasoning with the aim of expanding and enriching the knowledge network. There are three main approaches. First, reasoning based on ontology. Ontology reasoning is used for new knowledge discovery or conflict detection. Second, reasoning based on rules. Rules engines are used to compile corresponding rules and assist in decision-making through reasoning. Third, reasoning based on representation learning and neural network. Machine learning algorithms are used to embed knowledge in a low-dimensional continuous space so that knowledge reasoning can be achieved through vector computation.

(2) Text Similarity Computation

Reasoning based on representation learning cannot be separated from knowledge computation. Knowledge in the field of spacecraft launch is mainly Chinese document. The text processing falls under Natural Language Processing (NTP), while machine learning is the mathematical manipulation of variables of data. Using a series of variables to represent documents or texts is called text feature extraction in academia, such as the Term Frequency-Inverse Document Frequency (TF-IDF) index, which is intended to reflect how important a word is to a document. Besides, the contextual relationship of words can be used to form word vectors, thus constructing low-dimensional and dense Continuous Bag of Words (CBOW) and Skip-Gram (SG) models. "Similar" words are represented with "similar" vectors, and methods such as Euclidean distance, the inner product of vectors, cosine distance are used to calculate the similarity of the word vectors and the texts. In the field of machine learning, word segmentation plays another important part in the processing of Chinese texts. Python third-party library “Jieba” (Chinese for “to stutter”) is a mature Chinese word segmentation tool. After the segmentation, the application convenience and accuracy of the machine learning model shall be greatly improved [4].

3.6 Knowledge Storage

The knowledge graph can be stored in either a relational database or a graph database. The relational database uses SQL language for queries, but there are problems such as high difficulty in storing and updating ternary groups, data redundancy, and high overhead of complex queries. The graph database is a collection of nodes and relationships, which can effectively manage, store, and update data and their intrinsic relationships, and execute complex queries [5] by means of SPARQL. Commonly used graph databases can be queried at DB-Engines. Currently, the most commonly used is Neo4j.

4. Application of the Spacecraft Launch Knowledge Graph

Both knowledge graph and machine learning are artificial intelligence technology. Although a better performance can be achieved with the two combined in the application process, each has its own focus. Machine learning focuses on solving learning capabilities, which plays an important role in perception, recognition, judgment, thus making artificial intelligence smarter. Knowledge graph focuses on solving reasoning capabilities, which plays an important role in thinking, language, and reasoning, thus making artificial intelligence more knowledgeable. In the field of spacecraft launch, knowledge graphs have
multiple applications.

4.1 Automatic Question-Answering Based on Semantic Search

Search engines are the most typical application of knowledge graphs in practice. Automated question-answering systems that directly and accurately answer questions posed by users in a natural language will gradually become a basic component of search engines. Compared to keyword search, semantic search based on knowledge graphs can provide users with the most desired information and the most comprehensive summary, thus expanding the breadth and depth of the query. The knowledge graph can solve simple factual questions, such as "What are the turbine parameters of the CZ-3A rocket engine?" "What is the sensitivity of the XX radar receiver?" The knowledge graph system can directly return the most desired results for users, and provide the knowledge source and relevant recommendations. The main challenge is to solve complex problems, such as "Which external speed-measuring device in the launch range has the highest accuracy?". The knowledge graph system should have the reasoning ability to understand the area included in the launch range, find the accuracy of all the external speed-measuring devices in the launch range, and then compare them one by one to reach a conclusion. The difficulty of automatic question-answering lies in the semantic parsing and knowledge reasoning of natural language, which is the unique advantage of knowledge graphs. The flow of the automatic question-answering based on the semantic search is shown in Figure 5.

![Figure 5 Flow of Automatic Question-Answering Based on Semantic Search](image)

4.2 Fault Detection Based on Machine Learning

Some researchers [6] proposed a way of building a knowledge graph for unmanned aerial vehicles (UAV) fault detection based on reverse-sequence fault trees, which can model no less than ten failure modes. The failure codes are then used as universally unique codes as inputs to the neural network to construct a convolutional neural network, which can give an intelligent assessment of the healthy states (healthy, sub-healthy, and unhealthy) of UAV flight control with an accuracy of more than 85%. For the sub-healthy and unhealthy states, the reverse-sequence fault tree can be used to find the cause of failure. In rocket fault detection, it is difficult to quickly and intelligently locate the faults before launch and during flight. Therefore, detecting the faults by using the knowledge graph constructed based on telemetry parameters and their correlations combined with machine learning algorithms is a feasible approach. In the ground fueling system, the reasoning ability of the knowledge graph can be utilized to individually code the fueling valve and historical fueling data can be used to train the fault detection neural network model so as to assess the state of the fueling process and detect valve failure in a timely manner.

4.3 System Portrait Based on Knowledge Classification

The system portrait is relative to the user portrait in the social network knowledge graph. A user portrait is a virtual representation of a real user, which is a user model supported by a series of real data tags. Data tags are highly refined characteristics from the analysis of user information, which are used by enterprises for precise product marketing, intelligent recommendations, evaluation of marketing effectiveness, and improvement of user experience. In the spacecraft launch field, on the basis of knowledge classification, the characteristics, index parameters, and attribute classification of rocket internal systems, measuring equipment, and information systems can be abstracted and presented in a
visual way, which contributes to improving the data analysis ability and the overall knowledge level of the system.

5. Difficulties and Outlook
The knowledge graph in the spacecraft launch will reshape the knowledge management within the launch site, simplifying and linking the core knowledge of the launch site. In the process of building and maintaining the knowledge graph, there are difficulties such as high workload, immature algorithms, and problems such as the inability of the internal space network to access the external knowledge graph. With the development of knowledge bases in a big data environment, there are also difficulties and challenges, including computational complexity and data quality challenges, and the difficulty of updating and aligning knowledge. In the early stage of knowledge graph construction, we can first define the application scenarios, focus on solving specific business problems, and organize the data after clarifying the reasoning rules. After the smallest domain knowledge graph is built and successfully applied, the knowledge graph can be gradually improved.

About the Author
Kou Chao, Engineer at Xichang Satellite Launch Center
Mailing Address: Group 8, Box 16, Xichang City, Sichuan Province, China
Postal Code: 615000
Telephone: 18280650092

References
[1] Li Sizhen. Research and implementation of the construction technique of ontology-based knowledge graphs [D]. Beijing University of Posts and Telecommunications. Master's Thesis. June 2019, 25-29.
[2] Chen Cheng, Chen Yueguo, Liu Chen, Lv Xiaotong, Du Xiaoyong. Construction and application of intentional knowledge graphs [J]. Big Data. March 2020.
[3] Wen Pingmei, Ye Zhiwei, Ding Wenjian, Liu Ying, and Xu Jian. A review of research progress on the disambiguation of named entities [J]. Data Analysis and Knowledge Discovery. 2020(9).
[4] Tang Gen. Data science: from linear regression to deep learning [M]. Posts and Telecom Press. June 2018.
[5] Kang Jiehua, Luo Zhangxuan. Research on RDF data storage based on Neo4j [J]. Information Technology. 2015(6).
[6] Liu Hanze. Research on intelligent fault detection flight control system for UAVs based on machine learning [D]. University of Electronic Science and Technology of China. Master's Thesis. May 2019, 20-44.