# *Research on Construction Technology of Industry Knowledge Graph*

**Chenguang Liu[1,a,*], Xingxin Li[1,b] and Yongli Yu[1,c]**

[1] *Shijiazhuang Campus of Army Engineering University, Shijiazhuang, 050003, China*
*a. 17370228702@163.com, b. lxx_1226@sina.com, c. yu_yongli@263.net.cn*
**corresponding author: Chenguang Liu*

***Keywords:*** Industry knowledge graph, knowledge representation, knowledge extraction.

***Abstract:*** In recent years, many enterprises pay more attention to industry + knowledge graph. Industry knowledge graph has been well applied in finance, agriculture, medical treatment, e-commerce and other fields. For enterprises, the industry knowledge graph can help industry personnel to answer the task-based needs of the industry, assist various complex analysis applications or decision support, and build industry barriers. An effective construction system of industry knowledge graph can ensure the quality and scale of knowledge base, expansibility and reasoning ability. However, different data modes of different industries and different business needs make it impossible to build a unified system of industry knowledge graph. The purpose of this paper is to propose a general way to construct industry knowledge graph.

## 1. Introduction

It is because of the ability to acquire and form knowledge that human beings can make continuous progress. The important value of knowledge for AI is to equip machines with cognitive ability to understand the world and the industry or field of application. The structure of knowledge graph is similar to the structure of human brain organization knowledge, which is helpful to machine simulation of human thinking mode and knowledge structure for language understanding, visual scene analysis and decision analysis. Therefore, knowledge graph, as the support basis of artificial intelligence, is the only way to realize real human intelligence.

Tim Berners-Lee of the World Wide Web Consortium introduced the concept of Semantic Web in 1998. Adding semantic "metadata" that can be understood by computers to documents on the World Wide Web (e.g. HTML documents, XML documents), making the entire Internet a universal information exchange medium[1]. Knowledge graph, or semantic graphs, are an important realization of semantic web technology for knowledge representation.

The concept of knowledge map was proposed by Google in May 2012 and added to the company's search engine. Google's knowledge base uses semantic retrieval to collect information from multiple sources in order to improve the search quality and user experience of search engines. Then there was a craze for the study of knowledge graph in academia and industry. Knowledge graph are widely used in many industries, and have shown increasing value in the fields of semantic search, intelligent question and answer, and data analysis.

According to the coverage of knowledge, knowledge graph are divided into general knowledge graph and industry knowledge graph (or vertical knowledge graph). Industry knowledge graph is oriented to specific areas, based on industry data construction, emphasizing the depth of knowledge. The industry knowledge graph can be regarded as an industry knowledge base based on semantic technology, and its potential users are professionals in the industry[2]. For companies, industry knowledge graph is a valuable tool and a prerequisite for the application of Semantic Artificial Intelligence (SAI). It can help companies mine the undetected facts behind industry content, data and knowledge, and assist various complex analytical applications or decision support to build industry barriers. Therefore, it is particularly important and challenging to build a high-quality and large-scale industry knowledge graph with professional data models.

## 2. Knowledge Graph Related Technology

### 2.1. Definition and Architecture of Knowledge Graph

#### 2.1.1. Definition of Knowledge Graph

The knowledge graph is a large-scale knowledge base that connects and stores the knowledge of the world with a unified norm. Moreover, knowledge graph has certain reasoning ability, which can help human to discover new knowledge and facts. Compared with general relational databases, the graph structure of knowledge graph is similar to the structure of human brain organization knowledge, which helps machine simulate human thinking to process and understand knowledge.

#### 2.1.2. The Architecture of Knowledge Graph

Logically, we usually divide the knowledge graph into two layers: data layer and schema layer. The data layer is mainly composed of a series of facts, and knowledge is stored in units of facts. For example, the facts are expressed by triples such as (entity 1, relationship, entity 2) and (entity, attribute, attribute value)[3]. The schema layer is built on top of the data layer and regulates the data layer through Ontology. Ontology is a conceptual template of structured knowledge base. The knowledge base constructed by ontology library not only has a strong hierarchical structure, but also has less redundancy.

The knowledge graph construction mode is divided into top-down and bottom-up. Top-down refers to defining ontology database and data schema first, then adding a series of facts to the knowledge base, that is, schema layer before data layer. From the bottom up, the preliminary extracted text analysis data is driven by the data, and the pattern layer of the knowledge base is designed, that is, the data layer and the pattern layer. Most general knowledge graph are built from the bottom up, such as Google's Knowledge Vault. However, for the vertical domain knowledge map, it needs to meet the specific industry expertise and high-quality data, while coping with complex and changeable business needs, so the top-down construction method is mostly used.

### 2.2. Knowledge Representation

Knowledge representation refers to the way in which we code knowledge, beliefs, behaviors, feelings, goals, desires, preferences and other psychological activities in an artificial system. We can evaluate a knowledge representation from three dimensions: clarity, accuracy, and naturalness. A good knowledge representation needs to be unambiguous, with enough detail, and easily understood by humans[4].

### 2.3. Knowledge Extraction

Knowledge extraction involves clear and factual information, which comes from different sources and structures. The methods of knowledge extraction for different data sources are different. The difficulty of using D2R to acquire knowledge from structured data lies in the processing of complex table data. The difficulty of acquiring knowledge mapping from linked data is data alignment. The difficulty of obtaining knowledge wrapper from semi-structured data lies in the automatic generation, update and maintenance of wrapper. Here we mainly talk about the acquisition of knowledge from text, that is, information extraction in our broad sense.

#### 2.3.1. Entity Extraction

Entity extraction or named entity recognition (NER) plays an important role in information extraction[5]. It mainly extracts atomic information elements in text, such as person name, organization/organization name, geographical location, event/date, character value, amount value, etc.

#### 2.3.2. Relation Extraction

Relational Extraction (RE) is one of the tasks of natural language processing. The definition of this task is to return a semantic relationship between two entities given a sentence with two entities. Result from relation extraction task is often used in question answering system and knowledge graph, which is the basic and important task of natural language processing[6].

### 2.4. Knowledge Storage

The task of knowledge storage is to store the matrix at a minimum cost, and to make queries, additions, deletions and other operations as efficient as possible[7]. In the knowledge storage, it is necessary to comprehensively consider the write performance, query performance, and support for reasoning.

The main knowledge storage tools are: Apache Jena TDB、Apache Jena SDB、Eclipse RDF4J(Sesame)、Neo4j、Apache TinkerPop.

## 3. Analysis of System Requirements for Building Industry Knowledge Graph

### 3.1. System Requirements and Feasibility Analysis

The industry knowledge graph has industry characteristics, and the data sources and scene requirements of different industries are inconsistent, so there is no standard evaluation index[8]. However, unlike general knowledge graph, industry knowledge graph have higher requirements in terms of quality, scale and real-time.

(1)Accuracy：A high-quality industry knowledge graph not only has the necessary details, but also these details are clear and specific, without ambiguity.

(2)Completeness：The data of industry knowledge graph need to integrate the industry data and knowledge in an all-round way. The data not only come from the inside of the enterprise, but also need to integrate the potential information outside the enterprise.

(3)Real-time：Industry data iteration and business demand changes require that the industry knowledge graph is dynamically variable, and the knowledge base is regularly updated, modified or improved without manual intervention.

When constructing an industry knowledge graph, there is often no existing industry knowledge graph that can be referenced, only structured and unstructured industry data[9]. Therefore, for the schema layer, it is necessary to design a schema layer for the industry domain according to the existing data schema; for the data layer, knowledge extraction is required to continuously expand and improve the industry knowledge base. At the same time, in the process of expanding the industry knowledge base, it is necessary to select appropriate storage tools to meet different business requirements. Specifically, the requirements for building a system of industry knowledge graph include:

(1)schema Layer Design: The construction of industry knowledge graph needs to be based on the existing data model. Because the schema layer standardizes the data layer by ontology library, the knowledge base constructed by ontology library has not only a strong hierarchical structure, but also a small degree of redundancy. Moreover, the design of a complete and accurate model layer will greatly improve the accuracy and efficiency of subsequent knowledge extraction and knowledge base construction.

(2)Knowledge Extraction：Knowledge extraction is the discovery of knowledge from structured (relational databases, XML) and unstructured (text, document, image) data. The resulting knowledge needs to be machine readable and machine interpretable, and knowledge must be represented in a way that is reasonable. The extraction algorithm needs to recognize the relationship between entities and get a complete triple. At the same time, it is necessary to detect the correctness of the extracted triples and ensure that they are added to the knowledge base. Knowledge extraction requires that the extracted knowledge be machine readable and machine interpretable, and represent knowledge in a way that is reasonable.

(3)Knowledge storage：Knowledge storage tools are used to store industry knowledge, including well-designed model layers with industry characteristics, and large-scale interrelated industry data. It is necessary to select appropriate knowledge storage tools according to the size of industry knowledge graph and different requirements of writing, query and reasoning.

### 3.2. System Structure

In the previous paper, two methods of knowledge graph construction are introduced: top-down and bottom-up. The top-down approach needs to utilize the schema layer of the existing knowledge base and then expand the data layer. The bottom-up approach is to extract the data first, and then build the upper ontology mode through data driving. In the process of building the industry knowledge graph, the data model of most industries is not clear and specific enough, which needs to be sorted out artificially or unstructured and semi-structured data account for a large proportion. Therefore, this paper takes the bottom-up and top-down combination of knowledge base construction methods for analysis. Firstly, driven by data, through the analysis and collation of industry data, the model layer of industry knowledge graph is designed from top to bottom. Then, using the designed schema layer and the table's own semantic structure, we extract triples from the table and import them into the knowledge base to expand the data layer from the bottom to the top. Therefore, our industry knowledge base construction system is mainly divided into three steps：

(1)Data preprocessing: preliminary extraction of data, analysis and induction of data patterns;

(2)Schema Layer Design: The design of knowledge base schema layer is guided by the results of pre-processing analysis.

(3)Data Layer Construction: Knowledge extraction is guided by schema layer, triples are obtained from semi-structured tables and added to the data layer of knowledge base.

Therefore, the general process of building the system of industry knowledge graph is shown in Table 1.

Table 1: System flow of knowledge map construction for Industry.

| |
|---|
| (1) Preliminary Extraction, Analysis of Data Patterns, Definition of Database Storage Tools |
| (2) Based on the existing data model, design the knowledge base schema layer, and add the schema layer to the knowledge repository storage tool to standardize the subsequently expanded data layer. |
| (3) Based on schema layer, namely knowledge base, knowledge extraction algorithm is designed to extract triples from product documents. |
| (4) Determine the accuracy of the triple |
| (5) Add the correct triples to the knowledge base |

The process of building a system of industry knowledge map is shown in Figure 1.

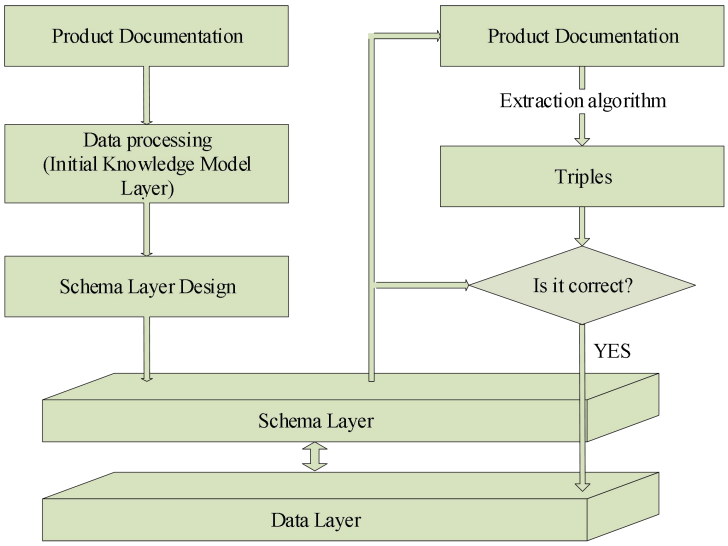


Figure 1: System flow chart of building industry knowledge graph.

From Figure 1, it can be seen that the process of the system is mainly divided into two parts: the conceptual layer design of industry knowledge map and knowledge extraction. Conceptual layer design is completed before knowledge extraction. Knowledge extraction includes table triple extraction, triple filtering and knowledge storage. Therefore, a knowledge extraction system consists of three services:

(1)Table Triple Extraction Service: The table knowledge extraction is designed to identify the entities, relationships, and matching relationships in the table, and extract all possible triples in the table. These triples may be incomplete, incorrect, or even conflicting.

(2)Triple filter service: The triplet filtering service is to match the schema layer ontology library, improve the triples with missing items, delete the incorrect triples, and manually judge the conflicting triples.

(3)Knowledge storage service: The knowledge storage service aims to add the correct and complete triples extracted from the above two steps to the knowledge graph data layer according to the existing schema layer.

## 4. Summary

The industry knowledge map is based on data from within the domain or within the company. It is usually required to rapidly expand the scale, build industry barriers, and the knowledge structure is more complex, usually including ontology engineering and rule-based knowledge. The quality requirement of knowledge extraction is very high. Joint extraction of structured, unstructured and semi-structured data from the enterprise needs manual verification to ensure the quality. Usually, the integration of multi-source areas is an effective means to expand the scale of data. The application form is more comprehensive. In addition to the search question and answer, it also includes decision analysis, business management, etc., and has higher requirements for reasoning and strong interpretability requirements. In view of the high requirements of the industry knowledge graph, its construction process is full of technical difficulties and challenges.

This paper first introduces and analyzes the related technologies of knowledge mapping. Understand the composition and construction of the knowledge map. Then, an ontology-based industry knowledge map construction system is designed and implemented. Aiming at the characteristics of existing industry data, a knowledge mapping framework method combining top-down and bottom-up is adopted. For the construction system of industry knowledge map, we need to further expand the extraction content and scale to further improve the knowledge base. We need to pay more attention to unstructured pure text data, extract entities and relationships among them, expand the knowledge base, and provide sufficient data for knowledge reasoning and question answering system.

## References


*[1] Liu Qiao, Li Yang, Duan Hong, Liu Yao, Qin ZhiGuang. Knowledge Graph Construction Techniques[J]. Computer research and development,2016,53(03):582-600.*

*[2] RuanTong,WangMengJie,WangHaoFen,HuFangHuai. Research on Construction and Application of Vertical Knowledge Graph [J]. Knowledge Management Forum,2016,1(03):226-234.*

*[3] Bengio Y , Courville A , Vincent P . Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798-1828.*

*[4] Foley J , Bendersky M , Josifovski V . Learning to Extract Local Events from the Web[C]// the 38th International ACM SIGIR Conference. ACM, 2015.*

*[5] Singh S, Riedel S, Martin B, et al. Joint inference of entities, relations, and coreference[J]. 2013.*
*Kadry A , Dietz L . Open Relation Extraction for Support Passage Retrieval: Merit and Open Issues[C]// the 40th International ACM SIGIR Conference. ACM, 2017. 6.*

*[6] Pobiedina N . Benchmarking database systems for graph pattern matching[M]// Database and Expert Systems Applications. Springer International Publishing, 2014.*

*[7] Pobiedina N . Benchmarking database systems for graph pattern matching[M]// Database and Expert Systems Applications. Springer International Publishing, 2014.*

*[8] Pobiedina N . Benchmarking database systems for graph pattern matching[M]// Database and Expert Systems Applications. Springer International Publishing, 2014.*

*[9] Xiong W, Hoang T, Wang W Y. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning[J]. 2018.*