Knowledge Graph Construction for "Red Memory" Based on Multi-source Heterogeneous Data Mining (Postprint)

Authors: Guo Jiaxin

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] The process through which the Chinese nation and the Communist Party of China have pursued truth has given rise to red cultural resources. The knowledge organization and mining of these resources to construct "Red Memory" not only enhances national confidence and cohesion, but also constitutes an important pathway to consolidating cultural confidence. To address the issues inherent in red cultural resources—such as wide distribution, diverse sources, heterogeneous types, limited content, and low degree of organization—this study constructs a "Red Memory" knowledge graph based on multi-source heterogeneous data mining, thereby enabling full utilization of red cultural resources. [Method/Process] First, an ontology library for red cultural resources is constructed by designing concepts, relationships, and attributes, thereby completing the knowledge modeling of "Red Memory". Second, red cultural resources are collected through multiple channels, with detailed analysis of their composition and characteristics; subsequently, entity, attribute, and relationship identification and extraction is performed on these multi-source heterogeneous data. Finally, the "Red Memory" knowledge graph is constructed and stored using a graph database. [Results/Conclusion] Through the construction of the "Red Memory" knowledge graph, deep relationship mining of multi-source heterogeneous red cultural resource data can be achieved, the organizational level of red cultural resources can be enhanced, and a foundation can be established for realizing intelligent services for red culture.

Full Text

ChinaXiv Collaborative Journal

Construction of "Red Memory" Knowledge Graph Based on Multi-source Heterogeneous Data Mining

Guo Jiaxin
School of Information Management, Central China Normal University, Wuhan 430079

Abstract:
Purpose/Significance: Red cultural resources emerged from the Chinese nation's and the Chinese Communist Party's pursuit of truth. Organizing and mining these resources to construct "Red Memory" not only enhances national confidence and cohesion but also represents a crucial pathway to strengthening cultural confidence. To address challenges such as wide distribution, diverse sources and types, limited content, and low organization levels in red cultural resources, this study constructs a "Red Memory" knowledge graph based on multi-source heterogeneous data mining to fully utilize these valuable resources.

Method/Process: The approach involves three main steps: first, designing concepts, relationships, and attributes to build an ontology library for red cultural resources, thereby completing the knowledge modeling of "Red Memory"; second, collecting red cultural resources through multiple channels, analyzing their composition and characteristics, and extracting entities, attributes, and relationships from these multi-source heterogeneous data; and finally, storing and constructing the "Red Memory" knowledge graph using a graph database.

Result/Conclusion: The constructed "Red Memory" knowledge graph enables deep relationship mining in multi-source heterogeneous red cultural resource data, improves the organization level of these resources, and establishes a foundation for intelligent red cultural services.

Keywords: Red cultural resources; Knowledge graph construction; Knowledge modeling

Classification Number: G250

---

1. Introduction

Red cultural resources were formed through the Chinese nation's and the Chinese Communist Party's pursuit of truth, resulting in a long historical development cycle that has led to challenges in their development and utilization, including wide distribution, multiple sources, diverse types, limited content, and low organization levels, which hinder users' deep utilization of these resources. In 2012, Google first proposed the concept of the knowledge graph [1], aiming to organize web data from a semantic perspective, build large-scale knowledge bases, and provide intelligent search services. Since then, various companies and research institutions have begun constructing knowledge graphs, such as YAGO from the Max Planck Institute in Germany [2], Google's Knowledge Vault [3], Fudan University's CN-DBpedia [4], and Tsinghua University's XLore [5]. As an important form of knowledge representation, knowledge graphs have gradually

become a crucial component in the transformation and upgrading of various industries from networking to intelligence, offering broad development prospects [3].

As an essential component of China's excellent traditional culture, red cultural resources contain rich revolutionary and historical values and serve as a fundamental support for strengthening cultural confidence [6]. Influenced by the rapid development of electronic technology, many regions have proposed establishing red cultural resource databases, such as the Sichuan Characteristic Cultural Resources Database [7] and the Xibaipo Red Education Resources Basic Database [8]. While these efforts have improved the organization level of red cultural resources to some extent, they remain at the data storage stage with insufficient organization. The knowledge graph, as a new resource organization method, has not been widely applied in the research and utilization of red cultural resources. Therefore, this study collects red cultural resource data with diverse structures and sources, organizes and mines this knowledge, and constructs a "Red Memory" knowledge graph to enhance the organization level of red cultural resources and present them to users in a more intuitive, dynamic, and interconnected manner.

---

2. "Red Memory" Knowledge Graph Construction Process

Red cultural resources represent the noble spirit and material carriers formed during the revolution and construction led by the Chinese Communist Party [9]. They exist not only in the past but also develop in the present, with their connotations continuously deepening along with historical processes and practical needs. Organizing and mining red cultural resources can revive the "Red Memory" embedded within them. A knowledge graph is essentially a structured, semantic knowledge base that uses a graph structure to represent entities, attributes, and their associations in the real world, where nodes represent entities and edges describe semantic relationships between them [10]. There are two primary approaches to constructing knowledge graphs: top-down and bottom-up [11]. The top-down approach involves pre-defining concepts and their relationships to design an ontology library, forming the Schema layer of the knowledge graph, and then matching and populating entities into the predefined ontology Schema layer. The bottom-up approach begins by extracting entities, attributes, and relationships from corpora or datasets, reorganizing similar types of entities, abstracting them into concepts, and finally constructing the Schema layer.

This study combines both top-down and bottom-up approaches to construct the "Red Memory" knowledge graph. First, by observing and comparing various data sources of red cultural resources, we identify the specific data required for the "Red Memory" knowledge graph and collect it through web scraping, manual collection, and other methods from diverse sources including red literature, websites, open datasets, and encyclopedias. Open datasets serve as the

primary source of structured data, encyclopedias provide semi-structured data, while unstructured text is obtained from red literature and vertical red culture websites. Second, we design concepts, relationships, and attributes by analyzing the composition and characteristics of red cultural resource data, using the Protégé tool to construct a red cultural resource ontology library, thereby completing the knowledge modeling of "Red Memory." Third, based on the designed ontology library, we employ different methods for entity, relationship, and attribute extraction according to the different forms of acquired data. Finally, we integrate and process the identified red cultural resource knowledge and store it in the Neo4j graph database, which enables visualization of the knowledge and completes the construction of the "Red Memory" knowledge graph. The overall process is shown in Figure 1 [Figure 1: see original paper].

3. Knowledge Modeling Based on "Red Memory" Ontology Construction

Knowledge modeling is a critical task in knowledge graph construction that involves the logical and systematic organization of knowledge. Building an ontology for knowledge modeling can comprehensively describe the attributes and relationships of entities within the knowledge graph. As an abstract representation model, an ontology can clearly define and describe concepts and their interrelationships, determining the data schema of the knowledge graph and specifying what data exists, such as entity categories, attributes of different entities, and associations between entities [12]. The ontology construction process is relatively complex, requiring adherence to established principles to ensure standardization. The widely recognized ontology modeling specifications are the five principles proposed by T. R. Gruber: clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment [13]. Regarding ontology construction methods, several mature approaches exist, including IDEF-5, Methontology, the seven-step method, and thesaurus-based ontology construction. Among these, the seven-step method offers greater generality [14], and thus this study adopts it, considering the unique characteristics of red cultural resources to construct the "Red Memory" ontology library.

As a special type of cultural resource, red cultural resources possess not only resource attributes but also cultural attributes, along with unique characteristics derived from their deep integration [15], resulting in diverse classification standards. According to Qu Changgen et al. [16], the most fundamental classification method adopted in academia divides red cultural resources into material and spiritual categories. Additionally, some scholars categorize them as dynamic or static types, or use a general/special dichotomy. In practical research, beyond these simple binary classifications, further adjustments are often made according to disciplinary needs. Zhang Taicheng [17] divides red cultural resources into ten major categories based on the principle of "theme-based classification with disciplinary considerations" and following Chinese language conventions:

red sites, artifacts, documents, figures, events, literature and art, architecture, spirit, research, and creation. Zhang Kewei [18], following the national tourism resource classification method, first subdivides red cultural resources into three main categories: sites and relics, buildings and facilities, and cultural activities, which are further divided into ten basic types. Sites and relics include historical event locations and military sites/ancient battlefields; buildings and facilities are divided into five types: cultural activity venues, exhibition halls, stele forests, celebrity former residences and historical memorial buildings, and cemetery parks; cultural activities include three types: figures, events, and literary and artistic works.

Constructing the "Red Memory" ontology library requires careful design of concepts, attributes, and relationships. For "Red Memory," the core is people, so the concept of "Person" is first established as a key concept. Closely related to this is the events that persons experience or participate in, leading to the inclusion of the "Event" concept. Browsing information related to "Red Memory" based on these two thematic concepts reveals that organizations joined by persons are also closely connected to both persons and events, prompting the addition of "Organization" to the ontology list. Additionally, information about persons' former residences, memorial halls, and cemeteries constitutes important concepts, all of which can be viewed as buildings, thus adding the concept of "Architecture." Furthermore, given the cultural attributes of "Red Memory," red literary and artistic works must be included, leading to the addition of the "Resource" concept. After establishing these five concepts, we refer to the aforementioned classification standards and actual collected data to assist in subdividing sub-concepts. The "Person" concept remains independent without subdivision. Since collected event-related data primarily falls into two categories—meetings and wars—events are divided into three types: meetings, wars, and others. Similarly, organizations are divided into schools, legions, political parties, and others; architecture is divided into celebrity former residences, memorial halls, monuments, memorial towers, sites (former sites), cemetery parks, and tombs. Resources are categorized based on carrier form into books, films, paintings, poetry, and songs. Considering these concepts comprehensively, we find that the subdivisions for events, architecture, and organizations have ambiguous boundaries that are difficult to define, and using subclass concepts directly would reduce ontology extensibility. Therefore, we eliminate subclass concepts for events, architecture, and organizations, instead adding a new concept of "Genre," which is divided into event genre, architecture genre, and organization genre, with an "other" option added to each type to ensure the comprehensiveness, accuracy, and extensibility of the constructed ontology.

In summary, the concepts in the "Red Memory" ontology are mainly divided into six categories: Architecture, Event, Genre, Organization, Person, and Resource, with multiple sub-concepts under Genre and Resource. Genre is subdivided into architecture genre, event genre, and organization genre. Analyzing the data for each category reveals distinct characteristics, prompting attribute definitions

based on these features. Table 1 presents selected concepts and attributes from the "Red Memory" ontology model.

Table 1. Partial Concepts and Attributes of the "Red Memory" Ontology Model

Concept	Attributes
Architecture	ArchitectureID, Name, Location, Image, Genre, Description
Event	EventID, Name, StartTime, EndTime, Location, Participants, Description
Genre	OrganizationGenreID (OrganizationType), ArchitectureGenreID (ArchitectureType), EventGenreID (EventType)
Organization	OrganizationID, OrganizationName
Person	PersonID, Name, BirthTime, DeathTime, Gender, Position, Alias, Nationality, Ethnicity, Birthplace, Works, Description
Resource	BookID, Title, PublicationTime, Author, Publisher, ISBN, Description / PoetryID, Title, Author, Content, Description / FilmID, Name, Genre, ReleaseTime, Actors, Director, Screenwriter, Description

In the designed ontology library, concepts and sub-concepts have hierarchical relationships, with sub-concepts possessing different attributes. Entities within sub-concepts have various semantic associations; for example, multiple relationships exist between persons such as "spouse" and "child," while buildings have "commemorate" relationships with persons/events. Based on the aforementioned concept design, we finalize the relationships involved in "Red Memory." Using the ontology construction tool Protégé, we add the defined concepts and relationships to complete the knowledge modeling of "Red Memory." Figure 2 [Figure 2: see original paper] illustrates the partial ontology concepts.

Table 2 . Partial Relationships in "Red Memory"

Relationship	Domain	Range
hasMate	Person	Person
hasChild/isChildOf	Person	Person
commemorateFor	Person/Event	Architecture
hasMember	Organization	Person
hasEventGenre	Event	EventGenre
participateIn	Person	Event

4. "Red Memory" Data Sources and Knowledge Acquisition

Red cultural resources bear witness to the entire process of our Party's development from its founding to its gradual growth [19]. Their long historical development cycle has resulted in varied resource collection, processing, and storage methods, leading to pronounced multi-source heterogeneity in red cultural resource data. Sources for acquiring red cultural resources include libraries, archives, museums, various memorial halls, exhibition halls, and red tourism sites across the country. Additionally, the era of big data has made various web resources important sources for obtaining red cultural resources. Therefore, structured, semi-structured, and unstructured data collected from these sources form the data foundation for constructing the "Red Memory" knowledge graph.

Structured data can be described using numbers or text, has uniform hierarchical or network structures, and is typically stored in relational databases. The structured data for "Red Memory" mainly comes from open datasets, obtained by downloading data to local storage via API interfaces as relational data. The construction of the "Red Memory" knowledge graph is based on this structured data, supplemented by collecting additional data from different sources and structures.

Unstructured data typically comprises text resources preserved in natural language form [20] and represents the richest source of knowledge. Large amounts of text exist in unstructured data sources such as red culture web pages, red tourism web pages, and books. Entity recognition, as the foundation of natural language text processing [21], is a crucial step in knowledge graph construction. Also known as named entity recognition, it involves extracting named referents with specific meanings from corpora, such as person names, place names, and organization names [22]. For the "Red Memory" knowledge graph, the entities to be recognized are those defined in the "Red Memory" ontology model at the schema layer. The most common method for entity recognition is machine learning. We can use web scraping tools to obtain "Red Memory" corpora from web pages, then use segmentation tools for preprocessing tasks such as word segmentation and annotation, followed by word vector conversion of the annotated corpora. Finally, we select training set corpora and train an extraction model through machine learning [23], using the entity recognition model to extract "Red Memory" entities from text. After entity recognition is completed, attribute acquisition can proceed.

The source for acquiring entity attributes in the "Red Memory" knowledge graph is the infobox of entries from various encyclopedia websites. Information in infoboxes is typically semi-structured data with high consistency and completeness, enabling the acquisition of basic person information by simply crawling the corresponding infobox tags from encyclopedia entries. For example, Figure

3 [Figure 3: see original paper] shows the infobox information for the person entry "Yang Zhicheng" from 360 Encyclopedia. Selecting four attributes—Chinese name, foreign name, alias, and nationality—their information can be obtained by browsing the web page source code (see Figure 4 [Figure 4: see original paper]). By parsing the source code, we find that attributes corresponding to persons can be located using "class" tags, allowing us to use Python's BeautifulSoup library to manipulate HTML elements and obtain attribute information for "Yang Zhicheng," resulting in <entity, attribute, attribute_{value}> triples.

Entity relationship identification and extraction follows principles similar to entity recognition. After obtaining "Red Memory" entities, we select statements containing multiple entity objects and perform entity relationship extraction on them. Through the identification and extraction of entities, attributes, and relationships, we ultimately acquire the entities, attributes, and relationships needed to construct the "Red Memory" knowledge graph. Finally, we organize and classify data obtained from different sources and store it in a relational database, with partial data examples shown in Figure 5 [Figure 5: see original paper].

---

5. "Red Memory" Knowledge Storage

Currently, knowledge graph storage is primarily accomplished through graph databases. Storing knowledge graphs in graph databases enables graph data visualization and integrated management through various tools provided by the database, efficiently meeting diverse user needs. Among graph databases, Neo4j is the most widely used due to its excellent performance and simple operation. This study stores the "Red Memory" knowledge graph in Neo4j, where labels represent concepts in "Red Memory," nodes represent entities, and edges describe relationships. Neo4j manages and manipulates knowledge graph data through Cypher commands. Since Cypher provides Load statements for batch importing CSV-format data, we convert "Red Memory" knowledge from the relational database into CSV files for storage and batch import using the following statements.

Batch import of concepts/entities (using "Architecture" as an example):

```
LOAD CSV WITH HEADERS FROM "file:///Architecture.csv" AS line
CREATE (:Architecture{ArchitectureID:line.ArchitectureID, name:line.name, location:line.loca
```

Batch import of relationships (using the "participateIn" relationship between Person and Event as an example):

```
LOAD CSV WITH HEADERS FROM "file:///PersonToEvent.csv" AS line
MATCH (from:Person{PersonID:line.PersonID}), (to:Event{EventID:line.EventID})
MERGE (from)-[r:participateIn{PersonID:line.PersonID, EventID:line.EventID}]->(to)
```

After batch importing the "Red Memory" knowledge from the relational

database into Neo4j, the "Red Memory" knowledge graph is formed, with results shown in Figure 6 [Figure 6: see original paper]. Blue dots represent persons, green dots represent organizations, red dots represent architecture, brown dots represent red resources, and orange dots represent events, with arrows indicating their relationships. Due to the open and interconnected nature of knowledge graphs, new data can be added later using Cypher commands [24], forming a large-scale "Red Memory" knowledge graph that enables intelligent red culture search, knowledge Q&A, knowledge reasoning, and other applications, laying the foundation for intelligent services of red cultural resources.

6. Conclusion

Applying knowledge graphs—a new organization technology—to the development and research of red cultural resources represents an inevitable choice for the advancement of red cultural resource disciplines and a requirement of the digital and intelligent era. This study designs a "Red Memory" ontology library by defining concepts, attributes, and relationships, completing the knowledge modeling of "Red Memory." We collect data from red cultural sources with different structures and origins, perform named entity recognition, relationship extraction, and attribute extraction based on this data to acquire knowledge, obtain "Red Memory" triples, and store them in Neo4j to construct the "Red Memory" knowledge graph. This further improves the organization level of red cultural resources, presents them in a more intuitive and modern manner, and enables the reorganization of fragmented red cultural resources distributed across various locations [25], reviving the "Red Memory" embedded in books, songs, and sites. In future work, we will further research intelligent Q&A, knowledge reasoning, and other applications of the "Red Memory" knowledge graph to meet users' needs for intelligent red cultural services and maximize the value inherent in red cultural resources.

Author Bio: Guo Jiaxin (ORCID: 0000-0002-7243-0291), Master's student, E-mail: jxguo718@163.com

Received: 2019-12-10
Published: 2020-02-25
Responsible Editor: Liu Yuanying

References

[1] SINGHA A. Introducing the knowledge graph: things, not strings[EB/OL]. [2019-04-10]. http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html.

[2] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM, 2007: 697-706.

[3] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//International conference on knowledge discovery and data mining. New York: ACM, 2014: 601-610.

[4] XU B, XU Y, LIANG J, et al. CN-DBpedia: a never-ending Chinese knowledge extraction system[C]//International conference on industrial, engineering and other applications of applied intelligent systems. Berlin: Springer, 2017: 428-438.

[5] WANG Z, LI J, WANG Z, et al. XLore: a large-scale English-Chinese bilingual knowledge graph[C]//International semantic Web conference. New York: ACM, 2013: 121-124.

[6] HUO G, ZHANG RONGXIU. The main characteristics and contemporary value of Chinese red culture[J]. Journal of Shanxi Radio & TV University, 2017(1): 103-105.

[7] WANG MAOCHUN. Exploration on the path of integrating characteristic cultural resources with high-tech[J]. Forum on Chinese Culture, 2015(6): 128-133.

[8] WANG YUPING, ZHANG TONGLE, ZHANG ZHIYONG. Discussion on the construction of Xibaipo red culture resource database[J]. Journal of Hebei Normal University (Philosophy and Social Sciences Edition), 2014, 37(1): 140-145.

[9] LI SHI. Accurately understanding the rich connotation of "red resources"[J]. Political Work Journal, 2005(12): 23.

[10] QI GUILIN, GAO HUAN, WU TIANXING. Research progress on knowledge graphs[J]. Technology Intelligence Engineering, 2017, 3(1): 4-25.

[11] LIU QIAO, LI YANG, DUAN HONG, et al. Survey on knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.

[12] MA CAN. Research and implementation of knowledge graph construction for smart courts[D]. Guizhou: Guizhou University, 2019.

[13] GRUBER T R. Toward principles for the design of ontologies used for knowledge sharing?[J]. International journal of human-computer studies, 1995, 43(5/6): 907-928.

[14] YUE LIXIN, LIU WENYUN. Comparative study of domain ontology construction methods at home and abroad[J]. Information Studies: Theory & Application, 2016, 39(8): 119-125.

[15] ZHANG TAICHENG. On red cultural resources[J]. Red Cultural Resources Research, 2015, 1(1): 1-11.

[16] QU CHANGGEN, WEN JIELU. Review of red cultural resources research[J]. Journal of Zhejiang Sci-Tech University (Social Sciences Edition), 2019, 42(2): 179-187.

[17] ZHANG TAICHENG. On the classification of red cultural resources[J]. Journal of China Executive Leadership Academy of Jinggangshan, 2017, 10(4): 137-144.

[18] ZHANG KEWEI. Research on the industrialization of Yimeng red cultural resources[D]. Jinan: Shandong University, 2010.

[19] XU QINGLING. Research on key technologies of human geographic information integration and visualization[D]. Fuxin: Liaoning Technical University, 2012.

[20] GUO WENLONG. Research and implementation of knowledge graph construction for traditional Chinese medicine prescriptions[D]. Lanzhou: Lanzhou University, 2019.

[21] ZHANG XIAOYAN, WANG TING, CHEN HUOWANG. Research on named entity recognition[J]. Computer Science, 2005(4): 44-48.

[22] WANG LIANGYI. Research on knowledge graph construction for carbon trading domain based on web data[D]. Ma'anshan: Anhui University of Technology, 2018.

[23] JIANG BINGCHUAN, WAN GANG, XU JIAN, et al. Large-scale geographic knowledge graph construction from multi-source heterogeneous data[J]. Acta Geodaetica et Cartographica Sinica, 2018, 47(8): 1051-1061.

[24] WU XUEFENG, ZHAO ZHIKAI, WANG LI, et al. Construction of knowledge graph for coal mine roadway support domain[J]. Industry and Mine Automation, 2019, 45(6): 42-46.

[25] [Reference 25 is missing from the original text]

Note: Figure translations are in progress. See original paper for figures.