Journal of Visual Languages and Computing 48 (2018) 19–28

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage

Jinhua Dou[a,b,c], Jingyan Qin[a,b,*], Zanxia Jin[a], Zhuang Li[a]

[a] School of Computer and Communication Engineering, University of Science and Technology Beijing, No. 30, Xueyuan Road, Haidian District, Beijing, China
[b] School of Mechanical Engineering, University of Science and Technology Beijing, No. 30, Xueyuan Road, Haidian District, Beijing, China
[c] School of Art & Design, Tianjin University of Technology, No. 391, Binshui Xidao, Xiqing District, Tianjin, China

ARTICLE INFO

Keywords:
Intangible cultural heritage
The 24 solar terms
Domain ontology
Knowledge graph
Natural language processing
Deep learning

ABSTRACT

Intangible cultural heritage (ICH) is a precious historical and cultural resource of a country. Protection and inheritance of ICH is important to the sustainable development of national culture. There are many different intangible cultural heritage items in China. With the development of information technology, ICH database resources were built by government departments or public cultural services institutions, but most databases were widely dispersed. Certain traditional database systems are disadvantageous to storage, management and analysis of massive data. At the same time, a large quantity of data has been produced, accompanied by digital intangible cultural heritage development. The public is unable to grasp key knowledge quickly because of the massive and fragmented nature of the data. To solve these problems, we proposed the intangible cultural heritage knowledge graph to assist knowledge management and provide a service to the public. ICH domain ontology was defined with the help of intangible cultural heritage experts and knowledge engineers to regulate the concept, attribute and relationship of ICH knowledge. In this study, massive ICH data were obtained, and domain knowledge was extracted from ICH text data using the Natural Language Processing (NLP) technology. A knowledge base based on domain ontology and instances for Chinese intangible cultural heritage was constructed, and the knowledge graph was developed. The pattern and characteristics behind the intangible cultural heritage were presented based on the ICH knowledge graph. The knowledge graph for ICH could foster support for organization, management and protection of the intangible cultural heritage knowledge. The public can also obtain the ICH knowledge quickly and discover the linked knowledge. The knowledge graph is helpful for the protection and inheritance of intangible cultural heritage.

1. Introduction

Intangible cultural heritage (ICH) is a precious historical and cultural resource of one's country. "*Intangible Cultural Heritage means the practices, representations, expressions, knowledge, skills – as well as the instruments, objects, artifacts and cultural spaces associated therewith – that communities, groups and, in some cases, individuals recognize as part of their cultural heritage* [1]." Protection and inheritance of ICH are important to sustainable development of the national culture. There are many existing problems in the process of ICH protection and inheritance. For instance, with the development of social economy, certain ICH items have lost their survival environment- modern technology replaces traditional handicrafts, and it has a considerable impact on intangible cultural heritage. Additionally, the public also lacks interest in traditional intangible culture. These factors have led to the extinction of certain intangible cultural heritage items.

With the development of information technology, e.g., database technology, internet of things (IOT), virtual reality (VR) and 3D scanning technology, the digital intangible cultural heritage protection methods have quickly grown. A public cultural service platform, database and websites of ICH were built by such organizations as culture management departments and relevant institutions. Because of this platform's flexible operation, convenient storage and management of information resources, the ICH digital service platform based on database technology has become an important way to protect and inherit intangible cultural heritage.

There is rich and diverse intangible cultural heritage in China. National level intangible cultural heritage consists of a total of 1836 items [2–5]. In addition, there are also many regional intangible cultural heritage items, such as provincial level intangible cultural heritage

* Corresponding author at: School of Computer and Communication Engineering, School of Mechanical Engineering, University of Science and Technology Beijing, No. 30, Xueyuan Road, Haidian District, Beijing, China.
E-mail addresses: doujinhua6971@163.com (J. Dou), Qinjingyanking@foxmail.com (J. Qin), Jinzanxia_go@163.com (Z. Jin), lz_ustb@sina.com (Z. Li).

https://doi.org/10.1016/j.jvlc.2018.06.005
Received 19 January 2018; Received in revised form 10 June 2018; Accepted 26 June 2018

items. Many databases are dispersed in various regions, and database resources are not well-integrated. Furthermore, there are still ICH data items being added every year. These factors present difficulties to the management of intangible cultural heritage information. Most of the existing cultural service platform based on database technology is presented using traditional network architecture and an interface layout that makes it difficult for the public to acquire knowledge quickly.

At the same time, massive data are generated from new media terminals every day, e.g., web platforms and mobile terminals. A large amount of ICH information appears ubiquitous; the public wants to master cultural knowledge but is not able to grasp the key knowledge quickly because of the massive and fragmented nature of the data. However, the ICH agencies and the government struggle to obtain the public's requirements dynamically because of the diversity of the population. This difficulty is disadvantageous to managing the ICH information and providing the appropriate cultural services content to public.

The ICH Knowledge Graph was designed for extraction, management, analysis and visualization of the key knowledge from a large amount of ICH data. This graph could help to organize, manage, and make massive information more understandable for related management departments and the public. We would like to construct a knowledge graph for ICH knowledge protection and dissemination in the research. The ICH domain ontology was defined with the help of intangible cultural heritage experts and knowledge engineers to regulate the concept, attribute and relationship of ICH knowledge. In this study, intangible cultural heritage data was obtained, and domain knowledge was extracted from ICH text data using the Natural Language Processing (NLP) technology. The ICH Knowledge Graph based on ICH knowledge base was explored, which may foster support for intangible cultural heritage management, protection and dissemination. The pattern and characteristics behind the intangible cultural heritage were presented by the ICH knowledge graph. The high quality information was provided to the public, and the people were able to discover the ICH knowledge quickly because of semantic association of knowledge. The knowledge graph is helpful to protect and inherit the intangible cultural heritage.

2. Research aim

Aimed at assisting the intangible cultural heritage knowledge management, protection and dissemination for culture management department and relevant institutions (while helping the public obtain the intangible cultural heritage knowledge quickly), the ICH domain ontology/schema was constructed, and a large amount of intangible cultural heritage data was obtained. The ICH knowledge was extracted based on ICH domain ontology model and Natural Language Processing (NLP) technology. The ICH knowledge base was constructed, and the ICH knowledge graph was developed to provide the appropriate service content to meet user needs.

3. Related works

Several studies and culture service institutions developed big data technology for cultural heritage related fields. Castiglione et al. [6,7] described CHIS (Cultural Heritage Information System), which was used to query, browse and analyze cultural digital content from a set of distributed and heterogeneous repositories. A big data infrastructure was proposed to manage digital cultural items. Colace et al. [8] described a PATCH system, which was applied to cultural heritage smart scenarios using pervasive technologies. Zhang et al. [9,10] proposed a big data analysis platform for facilitation of public digital culture services. The big data collection and analysis framework for public digital culture sharing service platforms was proposed. A public digital cultural service system in China was proposed by the government to provide the public with equal services [11]. The public culture data collection, storage and analysis were emphasized in a public digital cultural service system. Several public culture service platforms were constructed, such as the intangible cultural heritage service platform sponsored by the Chinese National Academy of Arts [12]. Most of the existing research was focused on the traditional network architecture and seldom considered the humanized interface of human-computer interaction. Simultaneously, the culture service system or platform was rarely built based on domain features. With the support of the knowledge graph, we can provide more direct answers to users and present it in a more user-friendly manner.

The knowledge graph is designed to describe the entities and relationships of the objective world. Dömel [13] introduced the idea of a web map to provide navigation support for hypertext browsers. The knowledge graph was first proposed by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources [14]. The Google knowledge graph data was based on wikidata and freebase databases [15], as well as public databases. Microsoft developed Bing search engine based on the Satori knowledge base, which could provide a variety of search services, e.g., search products of web, video, image and map. Microsoft Person cubic meter is a new type of social search engine which can automatically extract such information as names, location names, and organization names. The tool has a developed algorithm that automatically calculates the possibility of a relationship existing between inputs. Facebook Graph Search is a semantic search engine which is designed to give answers to users by natural language queries rather than a list of links. Sogou Search Engine named “knowledge cube”, integrates massive internet fragmentation information, mines the most core information and displays them to the users. Zhixin-schemas from Baidu support both entity query and entity recommendation. Pujara et al. [16] proposed knowledge graph identification based on ontology-aware partitioning to obtain better results. Arnaout and Elbassuoni [17] proposed a general framework that extended both the searched knowledge graph and triple-pattern queries for effective searching of RDF knowledge graphs. Fionda et al. [18] introduced the formalism regarding the web of linked data graph, presenting the MaGe tool, map framework and relevant examples.

Domain ontology is important to help domain experts regulate and annotate knowledge in their fields. Ontology is a philosophical theory and it defines a set of representational primitives to model domain knowledge or discourse in the context of computer and information sciences [19]. Ontology is a specification for modeling concepts, an abstract model describing the objective world, and a formal definition of the concepts and their linkages. Ontology includes class (concepts), slots (roles or properties), and facets (role restrictions). Ontology and a set of individual instances of classes constitute a knowledge base [20].

The CIDOC conceptual reference model (CIDOC CRM) [21, 22] provides definitions and a formal structure for describing the implicit and explicit concepts, and relationships used in cultural heritage documentation. Messaoudi et al. [23] developed an ontological model for the reality-based 3D semantic annotations of building conservation states. The Saint Maurice church of Caromb in the south of France was tested using this model to integrate unique spatial representation information about material and alteration phenomena. Aimed at making rock art data more accessible and more visible, the rock-art database project [24] explored new ways to perceive rock art through a collaborative, ontological and information visualization approach. Yang et al. [25] proposed the public cultural knowledge platform frame with the knowledge graph. With the help of public culture experts and knowledge engineers, they defined the ontology model of public cultural knowledge and the concept of public cultural knowledge, including person, object, location, time, event and organization. These studies provided references for the development of intangible cultural heritage knowledge graph in China. The ICH knowledge graph in our study is primarily intended to provide ICH knowledge networks to the