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
Available online 01 August 2018
1045-926X/ © 2018 Elsevier Ltd. All rights reserved.

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

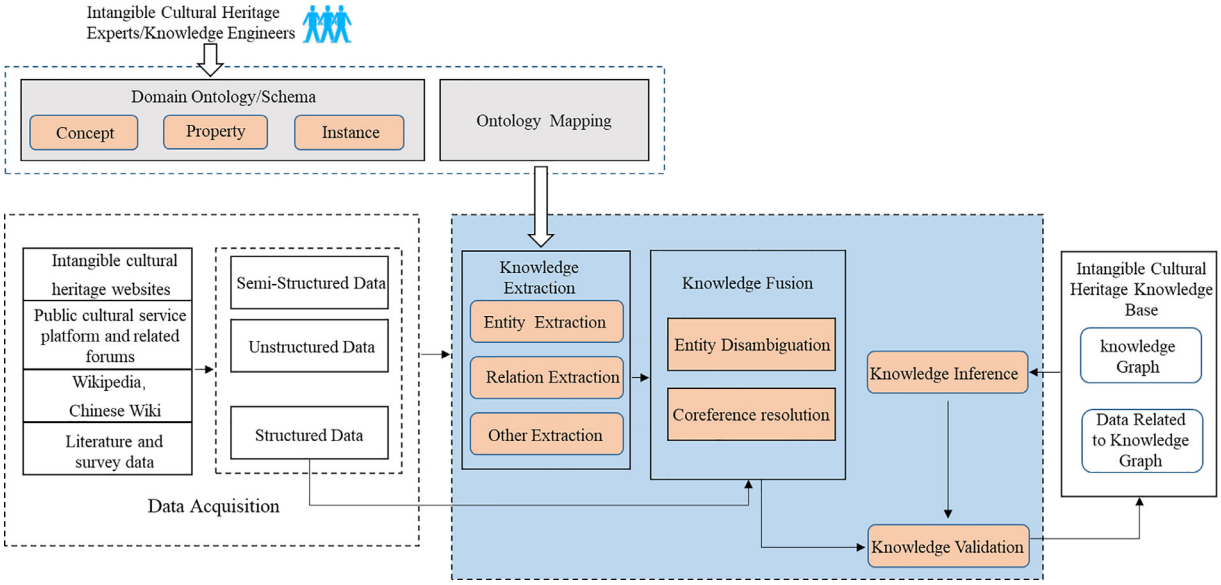


**Fig. 1.** Technology architecture of knowledge graph for intangible cultural heritage.

users from the massive network data, in a more intuitive way. The ICH knowledge can be discovered, fused, and analyzed from the explored knowledge graph.

## 4. Research methodologies

The primary research process includes data acquisition and preprocessing, domain ontology/schema construction, knowledge extraction, knowledge fusion, knowledge inference and knowledge validation. Fig. 1 shows the technology architecture of the knowledge graph for intangible cultural heritage.

### 4.1. Data acquisition and preprocessing

The intangible cultural heritage data was obtained from various data sources. The main data sources were Chinese intangible cultural heritage websites, such as national intangible cultural heritage websites, and every provincial intangible cultural heritage website. The articles, images, audios, and videos of these websites were obtained. A large amount of ICH text data was pulled from the public cultural service platform and related internet forums. The other data sources came from Wikipedia or Chinese Wikis, e.g., Hudong Baike, Baidu Baike and Soso Baike. The published academic articles and documents from digital library were also provided part of the ICH data. The structured data, semi-structured data and unstructured data were obtained from these data sources. Next, data preprocessing, including data cleaning, data integration and transformation, was executed, and the valid intangible cultural heritage data was obtained.

### 4.2. Intangible cultural heritage domain ontology /schema building

Domain ontology describes the concepts and relationships between concepts in particular fields (such as medicine and geography). The concept/class, relation and instance are included in the ontology knowledge base. Intangible cultural heritage, as one of the contents of public culture, concepts, relation and instances of it were defined with the help of intangible cultural heritage experts and knowledge engineers. The CIDOC CRM (Conceptual Reference Model) provides theory references for the construction of intangible cultural heritage ontology /schema and communication of cultural content [21,22,26].

### 4.3. Knowledge extraction

The basic unit for the knowledge graph is composed of "Entity – Relation – Entity" triples.

$$\mathbf{G} = (\mathbf{E}, \mathbf{R}, \mathbf{S})$$

E is the set of entity in a knowledge base, different entity types represented by |E|

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{|E|}\}$$

R is the set of relation in a knowledge base, different relationship types represented by |R|

$$\mathbf{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{|R|}\}$$

S represents the set of triples in a knowledge base.

$$\mathbf{S} \subseteq \mathbf{E}^*\mathbf{R}^*\mathbf{E}$$

Knowledge extraction is a technology that automatically creates knowledge, such as entity and entity relation from structured data sources (relational databases, XML), semi-structured data sources and unstructured data sources (text, documents, images) [27,28]. Knowledge extraction is the foundation of constructing the knowledge graph. Knowledge extraction includes entity extraction, relation extraction and other extractions such as attribute extraction. In this study, we extracted the ICH knowledge mainly from the ICH text data.

Several methods of entity extraction, such as heuristic algorithm, k-nearest neighbor (KNN) , conditional random field (CRF) and clustering algorithm, are all commonly used in related studies [29,30,31]. These methods of entity extraction do not perform as well in terms of accuracy and recall – it is difficult to obtain higher learning accuracy with these methods.

In recent years, convolutional neural networks (CNNs) were used in Natural Language Processing to realize semantic analysis, and several studies have achieved excellent results [32–35]. ICH text entity recognition can be simply understood as a sequence annotation problem: a sentence is given and each character in the sequence of the sentence is marked. Entity extraction is first executed from a massive text data set. During the process of text sequence annotation, traditional CNN models could not overwrite the entirely input data. To cover all of the input information, it is necessary to add more convolution layers, which increase the difficulty of model training. Fisher Yu and Vladlen Koltun [36] proposed the dilated convolutions model, which could quickly

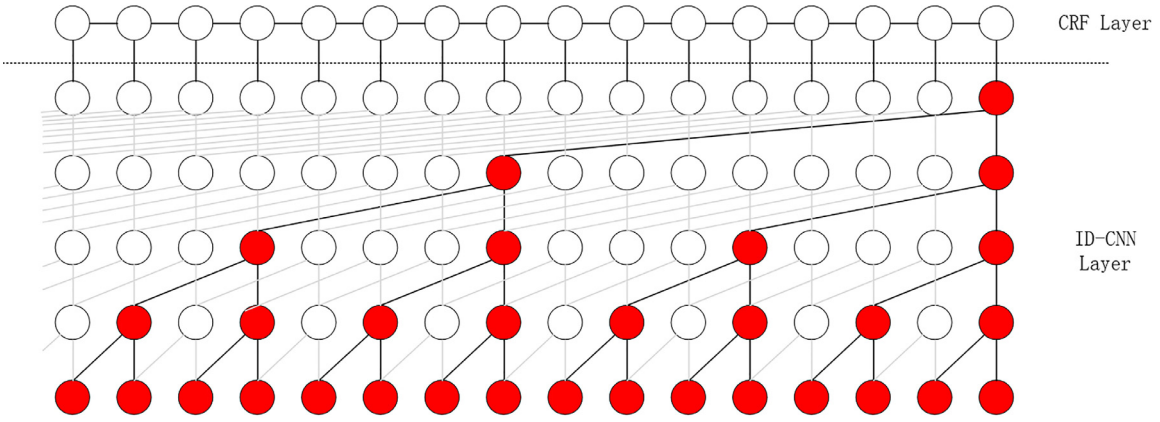


**Fig. 2.** Model of ID-CNNs combined CRF algorithm. A dilated CNN block with maximum dilation width 8 and filter width 2.

cover all of the input data. These researchers also presented an application of dilated convolutions for sequence labeling. Emma Strubell et al. [37] proposed the Iterated Dilated Convolutional Neural Networks (ID-CNNs), based on the dilated convolutions model, to realize fast and accurate entity recognition. The ID-CNNs architecture repeatedly applied the same block of dilated convolutions to token-wise representations. ID-CNNs were trained to aggregate context from the entire document, which realized even more accurate results. In this study, the ID-CNNs model was combined with CRF algorithm to recognize Chinese language entities of intangible cultural texts. Fig. 2 shows the model of ID-CNNs combined with CRF algorithm; a dilated CNN block with maximum dilation width 8 and filter width 2 was adopted to extract the ICH entity.

The deep learning methods using CNN or Bi-directional LSTM were considered solutions to the present relationship extraction. Nguyen and Grishman [38] introduced a convolutional neural network for relation extraction, which automatically learned features from sentences and minimized the dependence on external toolkits and resources. These researchers emphasized the relation extraction problem with an unbalanced corpus. Zhou et al. [39] proposed Attention-Based Bidirectional Long Short-Term Memory Networks (Att-BLSTM) to capture the most important semantic information in a sentence. Yankai Lin et al. [40] explored a sentence-level attention-based CNN model for relation extraction. Lee [41] proposed LSTM conditional random fields (LSTM-CRF) models to extract the named entity and achieve better results. However, these methods were applied only to English language corpus.

Most of the existing literature and code are trained for the English corpus, but a great majority of Chinese intangible cultural texts are the Chinese Corpus. Cho et al. [42] proposed Gated Recurrent Unit (GRU) model, which was a variation of LSTM. GRU maintained the effect of LSTM and the structure was simpler. The Att-BLSTM model was improved by replacing the LSTM with GRU in the model structure and each Chinese character in the sentence was input as character embedding. In this study, the bidirectional GRU model was used for character-level attention. Fig. 3 shows the bidirectional GRU model with attention. The sentence-level attention-based CNN model was improved by replacing the CNN with the Bi-directional GRU model in the model structure [43]. Fig. 4 shows the architecture of the sentence-level attention-based Bi-GRU model. For this study, the character-level attention-based Bi-GRU model and the sentence-level attention-based Bi-GRU model were all adopted to extract the Chinese language entity relation of ICH.

### 4.4. Knowledge fusion

Knowledge fusion consists of entity disambiguation and coreference resolution. Although entity and relation information was extracted, there was considerable redundancy or error information. **Entity**

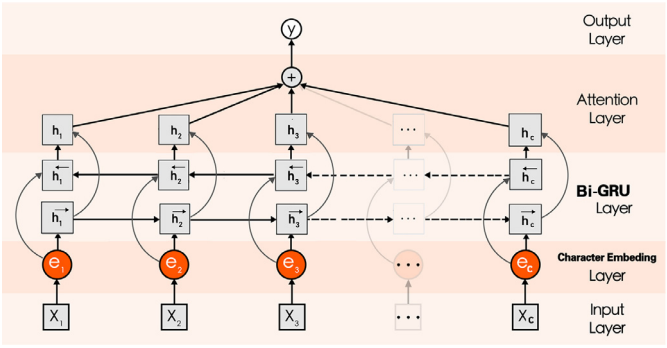


**Fig. 3.** Bidirectional GRU model with attention based on Att-BLSTM model [39,43].

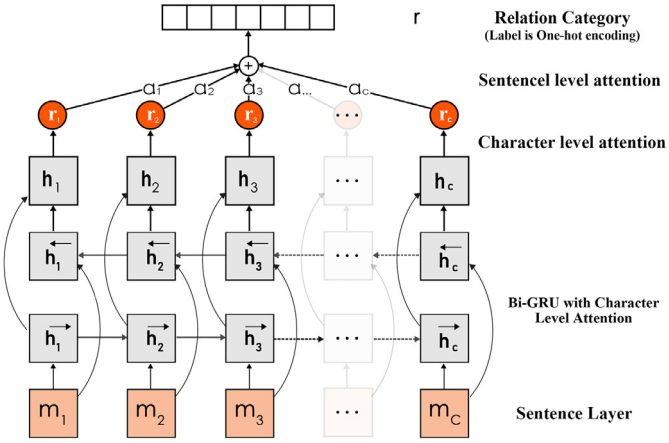


**Fig. 4.** Architecture of sentence-level attention-based Bi-GRU, where $m_c$ and $r_c$ indicate the original sentence for an entity pair and its corresponding sentence representation, $a_c$ is the weight given by sentence-level attention, and r indicates the representation of relation category [40,43].

**Disambiguation** based on entity linking methods was selected to solve ambiguous problems of entities, e.g., one named entity could correspond to multiple entity concepts. There were also existing problems in which there were multiple terms corresponding to the same entity. **Coreference resolution** based on cluster/ classification algorithm methods was adapted to solve this problem.

### 4.5. Knowledge inference

The knowledge inference infers new and unknown knowledge according to the existing entity-relationship triples in the knowledge base.

Available methods include deep learning, probability model and logic methods. The new relation is learned using the methods of analyzing the relations with rule-based reasoning. The knowledge inference is finished and the new links between ICH entities are established, thereby expanding and enriching the knowledge networks.

### 4.6. Knowledge validation

Structured knowledge is produced from various types of ICH data. However, invalid knowledge or false entity relationship may also be produced from the ICH data. Knowledge validation is important to ensure the credibility of knowledge. In this study, the knowledge was verified with the help of intangible cultural heritage experts, knowledge engineers and related domain experts.

Finally, the ICH knowledge base, including the knowledge graph and related data, was constructed based on the above methods. With the emergence of new data resources, the content of the knowledge base was constantly updated. The knowledge graph and various data sources were managed by the knowledge base. The entity-relationship triples and attribute value pair (AVP) were included in ICH knowledge base. Service applications associated with ICH knowledge was provided based on the knowledge graph.

## 5. Experiment and results

The 24 Solar Terms are a unique cultural heritage item for the working people of our country, China. This heritage item reflects the changes of the season and guides the agricultural activities. The 24 Solar Terms refers to 24 specific seasonal changes in the Chinese lunar calendar, which is based on changes in the earth's position on the ecliptic. The ecliptic is measured in 360 longitudinal degrees, and it is divided into 24 equal segments; a solar term corresponds to the earth moving forward 15° along the ecliptic.

The 24 Solar Terms include " Spring begins "," The rains "," Insects awaken "," Vernal Equinox "," Clear and bright "," Grain rain "," Summer begins "," Grain buds "," Grain in ear "," Summer solstice "," Slight heat "," Great heat "," Autumn begins "," Stopping the heat "," White dews"," Autumn equinox"," Cold dews"," Hoar-frost falls"," Winter begins"," Light snow"," Heavy snow"," Winter solstice"," Slight cold"," Great cold". On December 1, 2016, The 24 Solar Terms of China was incorporated in UNESCO's intangible cultural heritage list [44].

There is a large amount of knowledge and social practice concerning the 24 Solar Terms. The 24 Solar Terms are a combination of astronomy, agriculture, phenology and folklore, and they derive a large number of related age-old seasonal culture, including the relevant proverbs, songs, legends, traditional production tools, living appliances, handicrafts, calligraphy, painting and other works of art, as well as festival culture, production rituals and folk customs. It is necessary to mine the knowledge of the 24 Solar Terms and construct knowledge graph for protection and dissemination of 24 Solar Terms culture.

### 5.1. Experimental environment settings

**Hardware Environment:**
Processor: Xeon E5 CPU;
Memory: 126 G;
Disk: 2.0 TB.
**Software Environment:**
Operating System: Ubuntu 14.04;
Deep learning framework: TensorFlow;
Development language: Python 2.7.

### 5.2. Domain ontology building for the 24 solar terms

CIDOC CRM described the concept/class of culture heritage with entity and described the relation of culture heritage concept using property. First, we regulate 24 Solar Terms knowledge based on CIDOC CRM ontology model with the help of domain experts. The concept/class of 24 Solar Terms primarily includes **actor, thing, event, time and place. Actor** mainly refers to person and organization, e.g., management department at all levels, committees, training organization, intangible cultural heritage inheritor, related scholars and expert, cultural institution managers. For example, Zhengming Bao is the inheritor of “Shiqian Spring”. “Shiqian Spring” is included on the extension list for the 24 solar terms. **Thing** is mainly composed of man-made things and legal objects, e.g., a **physical object** or **conceptual object**, including relevant proverbs, songs, handicrafts, books, literature associated with the 24 Solar Terms and various objects used in the 24 Solar Terms activities. **Event** mainly includes various folk activities, related meetings, seminars and forum activities, e.g., sacrificial ceremony and collective ceremonies occurrence in some solar terms. For example, ancestor worship occurs at “Clear and bright”. **Time** is used to define the temporal extent of instances, e.g., February 3–5 of every year is the time span of when “Spring begins”**. Place** may be the location of the declaration area, related exhibition locations, related organization locations, or place of event occurrence. For example, Huayuan is the place where the Catch the Autumn Festival of the Miao nationality occurred.

Next, we extracted the property from the CIDOC CRM model to define the relation among concepts of the 24 Solar Terms. The main relation is shown in Table 1. A part of the 24 Solar Terms entity and the relation diagram was also constructed, as shown in Fig. 5.

### 5.3. Character embedding training

At first, the character embedding was trained using the Skip-Gram

**Table 1**
Mainly relation of the 24 Solar Terms.

| Property ID | Property name | Property ID | Property name |
|---|---|---|---|
| P1 | is identified by (identifies) | P67 | refers to (is referred to by) |
| P2 | has type (is type of) | P69 | has association with (is associated with) |
| P4 | has time-span (is time-span of) | P70 | documents (is documented in) |
| P5 | consists of (forms part of) | P82 | at some time within |
| P7 | took place at (witnessed) | P89 | falls within (contains) |
| P11 | had participant (participated in) | P94 | has created (was created by) |
| P12 | occurred in the presence of (was present at) | P108 | has produced (was produced by) |
| P14 | carried out by (performed) | P116 | starts (is started by) |
| P16 | used specific object (was used for) | P128 | carries (is carried by) |
| P17 | was motivated by (motivated) | P165 | incorporates (is incorporated in) |
| P35 | has identified (was identified by) | P166 | was a presence of (had presence) |
| P37 | assigned (was assigned by) | P167 | at (was place of) |
| P45 | consists of (is incorporated in) | P171 | at some place within |
| P55 | has current location (currently holds) | P192 | initiated by (initiates) |

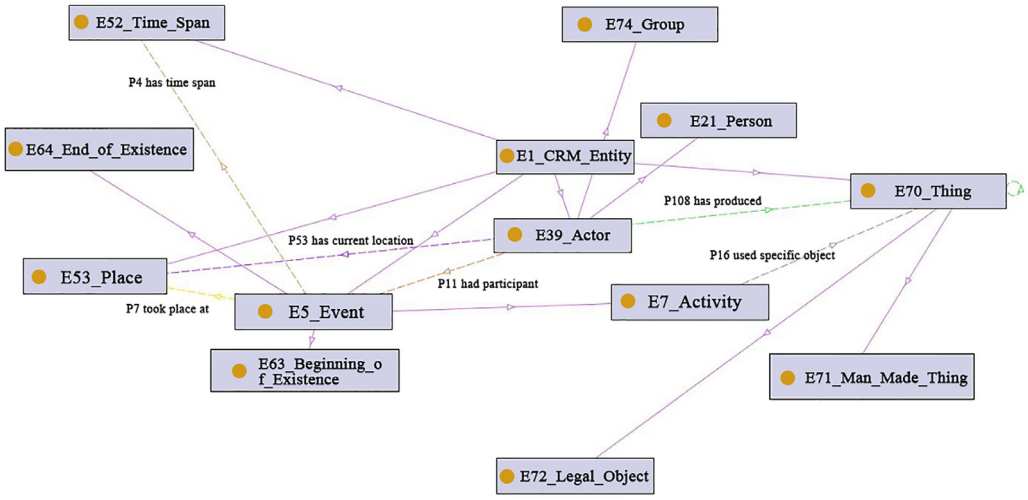


**Fig. 5.** Part of the relation diagram of the 24 Solar Terms.

model of word2vec [45,46]. In other words, the input is the character embedding of a particular character, and the output is a context character embedding corresponding to the particular character.

ICH corpus database includes data from 1836 National Intangible Cultural Heritage items. We organized the ICH corpus database as the training text, and then separated the text by inserting a space into a character-by-character form. The training text was trained with the word2vec tool to obtain the character embedding. The result of the character embedding for ICH is shown in Fig 6. Each character occupies one line and is represented by one hundred-dimensional embedding.

### 5.4. The 24 solar terms corpus processing

The 24 Solar Terms text was gathered from various data sources, e.g., Chinese intangible cultural heritage websites, Wikipedia, Baidu Baike, articles and documents, and internet forum. A large number of texts were tagged according to the previously identified domain ontology of the solar terms. The manual annotation method was adopted, and a total of twenty-one people, including intangible cultural heritage experts, knowledge engineer, relevant researchers and students, took part in this work.

Next, we converted the tagged data formats using the Perl programming language. The data format of the training data, verified data and test data is shown in Fig. 7. The text is marked by character, e.g., Letter B represents the first character of the entity name and Letter C represents the entity category in B-C. Letter I represents the non-first entity character and Letter C represents the entity category in I-C. The letter O represents a non-entity.

二 B-C
十 I-C
四 I-C
节 I-C
气 I-C
申 O
报 O
单 O
位 O
中 B-ORG
国 I-ORG
农 I-ORG
业 I-ORG
博 I-ORG
物 I-ORG
馆 I-ORG

**Fig. 7.** Format of training data, validation data and test data.

### 5.5. Model training

The ID-CNNs model combined with CRF algorithm was adopted to carry out sequence annotation. The training data, validation data and test data were input, and the well-trained entity recognition model was obtained through this training process. The subsequent process

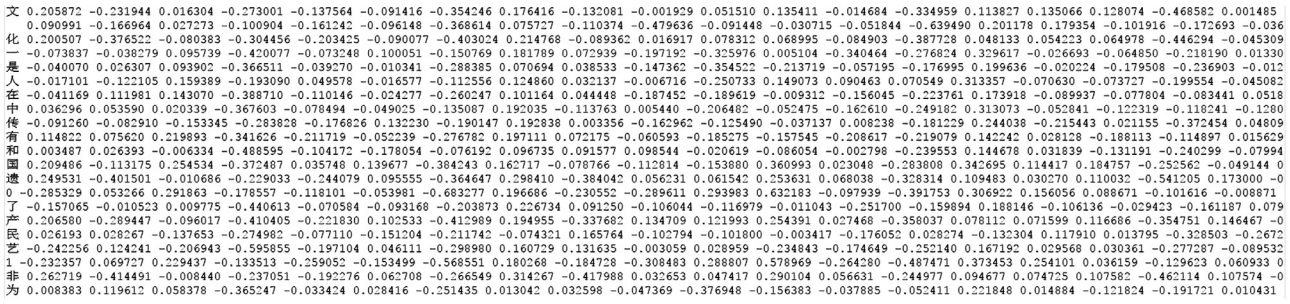


**Fig. 6.** Character vector of intangible cultural heritage.

**Table 2**
Part of knowledge of the 24 solar terms.

| Relation | Entity range | Entity name | Entity range | Entity name |
|---|---|---|---|---|
| was a presence of | conceptual object | 二十四节气 (the 24 Solar Terms) | physical thing | 淮南子·天文训 (Huainanzi·Astronomical theory) |
| was created by | physical thing | 淮南子·天文训 (Huainanzi·astronomical theory) | person | 刘安 (liu an) |
| is incorporated in | conceptual object | 二十四节气 (the 24 Solar Terms) | conceptual object | 人类非物质文化遗产代表作名录 (UNESCO Intangible Cultural Heritage Lists) |
| has time span | conceptual object | 人类非物质文化遗产代表作名录 (UNESCO Intangible Cultural Heritage Lists) | time | 2016年11月30日 (November 30, 2016) |
| carried out by | conceptual object | 二十四节气 (the 24 Solar Terms) | organization | 中国农业博物馆 (China Agricultural Museum) |
| is type of | conceptual object | 二十四节气 (the 24 Solar Terms) | conceptual object | 民俗 (folk custom) |
| consists of | conceptual object | 二十四节气 (the 24 Solar Terms) | conceptual object | 立春 (Spring begins) |
| is associated with | conceptual object | 立春 (Spring begins) | conceptual object | 九华立春祭 (the sacrifice ceremony at spring begins in Jiuhua) |
| took place at | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) | place | 浙江衢州 (Quzhou, Zhejiang) |
| consists of | place | 浙江衢州 (Quzhou, Zhejiang) | place | 梧桐祖殿 (The ancestral temple of Wu Tung) |
| performed | person | 汪筱联 (Wang xiaolian) | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) |
| is associated with | person | 刘魁立 (Liu kuili) | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) |
| is associated with | person | 王霄冰 (Wang xiaobing) | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) |
| has produced | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) | physical thing | 春牛图(Spring Cow Drawing) |
| motivated | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) | event | 迎春 (welcome spring) |
| motivated | conceptual object | 九华立春祭(the sacrifice ceremony at spring begins in Jiuhua) | event | 咬春 (biting the spring) |

included calling the model, entering a new text of the 24 Solar Terms, and obtaining the entity of the new text. Next, we extracted the relation among ICH entity based on the Bi-GRU model. Next, more entities, relations and instances were extracted, based on deep learning model of NLP methods. A number of the final results are outlined as follows:

{'string': '在国际气象界, 二十四节气被誉为中国的第五大发明. 2017年5月5日, 二十四节气保护联盟在浙江杭州拱墅区成立.',
'entities': [{'end': 12, 'start': 7, 'word': '二十四节气', 'type': ' C '}, {'end': 33, 'start': 24, 'word': '2017年5月5日', 'type': ' TM '}, {'end': 43, 'start': 34, 'word': '二十四节气保护联盟', 'type': 'F'}, {'end': 51, 'start': 44, 'word': '浙江杭州拱墅区', 'type': ' L '}]}

实体1:二十四节气
实体2: 二十四节气保护联盟
关系:加入
实体1: 二十四节气保护联盟
实体2: 浙江杭州拱墅区
关系:发生于某地
实体1: 二十四节气保护联盟
实体2: 2017年5月5日
关系:时间

{'string': '2006年5月20日, “二十四节气”作为民俗项目经国务院批准列入第一批国家级非物质文化遗产名录.',
'entities': [{'end': 10, 'start': 0, 'word': '2006年5月20日', 'type': ' TM '}, {'end': 17, 'start': 12, 'word': '二十四节气', 'type': ' C '}, {'end': 22, 'start': 20, 'word': '民俗', 'type': ' C '}, {'end': 47, 'start': 35, 'word': '国家级非物质文化遗产名录', 'type': ' D '}]}

实体1: 二十四节气
实体2: 2006年5月20日
关系:时间
实体1: 二十四节气
实体2: 国家级非物质文化遗产名录
关系:加入
实体1: 二十四节气
实体2: 民俗
关系:类型

Notes:

在国际气象界, 二十四节气被誉为中国的第五大发明. 2017年5月5日, 二十四节气保护联盟在浙江杭州拱墅区成立: In the international meteorological field, the 24 Solar Terms are known as the fifth invention of China. May 5, 2017, The 24 solar terms protection league was established in Gongshu District, Hangzhou, Zhejiang.

2006年5月20日, “二十四节气”作为民俗项目经国务院批准列入第一批国家级非物质文化遗产名录: May 20, 2006, “The 24 solar terms” as a folk-custom project approved by the State Council included in the first batch of national intangible cultural heritage list.

实体: Entity
关系: Relation
二十四节气: The 24 solar terms;
二十四节气保护联盟: The 24 solar terms protection league
浙江杭州拱墅区: Gongshu District, Hangzhou, Zhejiang
民俗: Folk-custom
国家级非物质文化遗产名录: National intangible cultural heritage list

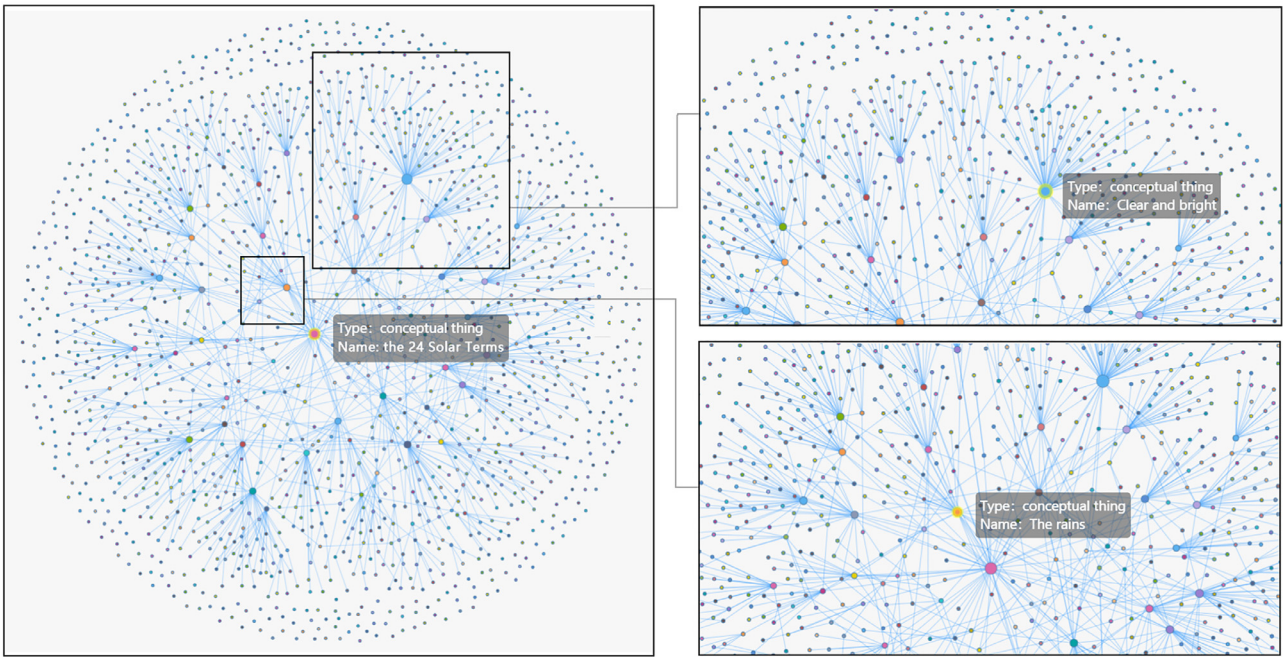


**Fig. 8.** Knowledge graph visualization of the 24 Solar Terms.

加入: Incorporates
发生于某地: Took place at
时间: Has time-span
类型:Is type of

### 5.6. Construction of the 24 solar terms knowledge graph

With deep learning methods used in the Natural Language Processing technology, we obtained approximately 1500 entities and relationships from large amounts of the 24 solar terms text data. Part of the knowledge of the 24 solar terms is shown in Table 2. After entity disambiguation and coreference resolution, the valid knowledge of the 24 solar terms was obtained. There were a number of errors among the obtained knowledge because of unsuitable rules or algorithm precision issues. The knowledge validation was executed by ICH experts and knowledge engineers to correct these errors.

The massive entity-relationship triples were input to knowledge base. Knowledge inference was executed to look up new links among the entities of the 24 solar terms. The knowledge networks of the 24 solar terms were expanded and enriched in the process of knowledge inference. Finally, the 24 Solar Terms knowledge base was constructed, and the knowledge graph was developed based on the above methods.

### 5.7. The 24 solar terms knowledge graph visualization

Finally, we developed the visual presentation of the 24 Solar Terms knowledge graph using Force-Directed Graph. The entity and entity relationship of the 24 Solar Terms were presented using the knowledge graph visualization, as shown in Fig. 8. Each node represents the entities of 24 Solar Terms including conceptual thing, physical thing, person, organization, time, location and activities, e.g., “Spring begins”, “The rains”, “Insects awaken”, “Vernal equinox”, “Clear and bright”, spring welcome, catch the autumn festival, Twenty-four Solar Terms protection league, the inheritor of the 24 Solar Terms. The connection between nodes represents the relationship between the entities. For instance, the 24 Solar Terms is carried out by China Agricultural Museum. Spring motivates certain activities, such as welcome spring, and the 24 Solar Terms are a type of folk custom. When the mouse points to each node, the entity name and entity type are presented to the users. To indicate the knowledge graph clearly, the entity of 24 Solar Terms is presented using node combined text. Fig. 9 shows part of the visualization of the knowledge graph for the 24 Solar Terms using node with text description.

## 6. Conclusions

In this study, we attempted to construct an ICH knowledge graph. At first, we constructed the ICH domain ontology to normalize the knowledge of intangible cultural heritage. Next, the ICH data was tagged by manual annotation according to the ICH ontology/schema. It was a very heavy job for our team to manually annotate a large number of 24 Solar Terms text data. The deep learning methods were used in NLP technology to extract the entity and relation from ICH text data. We gathered all of the obtained knowledge of intangible cultural heritage to build the ICH knowledge base and develop the knowledge graph. The knowledge graph of the 24 Solar Terms was developed to provide users with a comprehensive knowledge system. The knowledge graph could also assist knowledge management for cultural departments and related institutes. The greatest challenge was in the data collection of intangible cultural heritage. Because there was no related database, the team spent considerable time collecting the massive data of the 24 Solar Terms. There is rich and diverse intangible cultural knowledge in China, and our team only extracted a portion of it. We will constantly enrich the ICH data in the next stage. The collected data will be published to ensure that more people can participate in building the data sets and sharing them in the future.

## Acknowledgments

We would like to thank all of the teachers and students who helped for the completion of our work. Professor Xucheng Yin and Bowen Zhang provided technical support for ICH entity-relation extraction. Funding: This work was supported by the Ministry of Education Humanities and Social Sciences Research Youth Fund Project [Grant number 15YJCZH034].

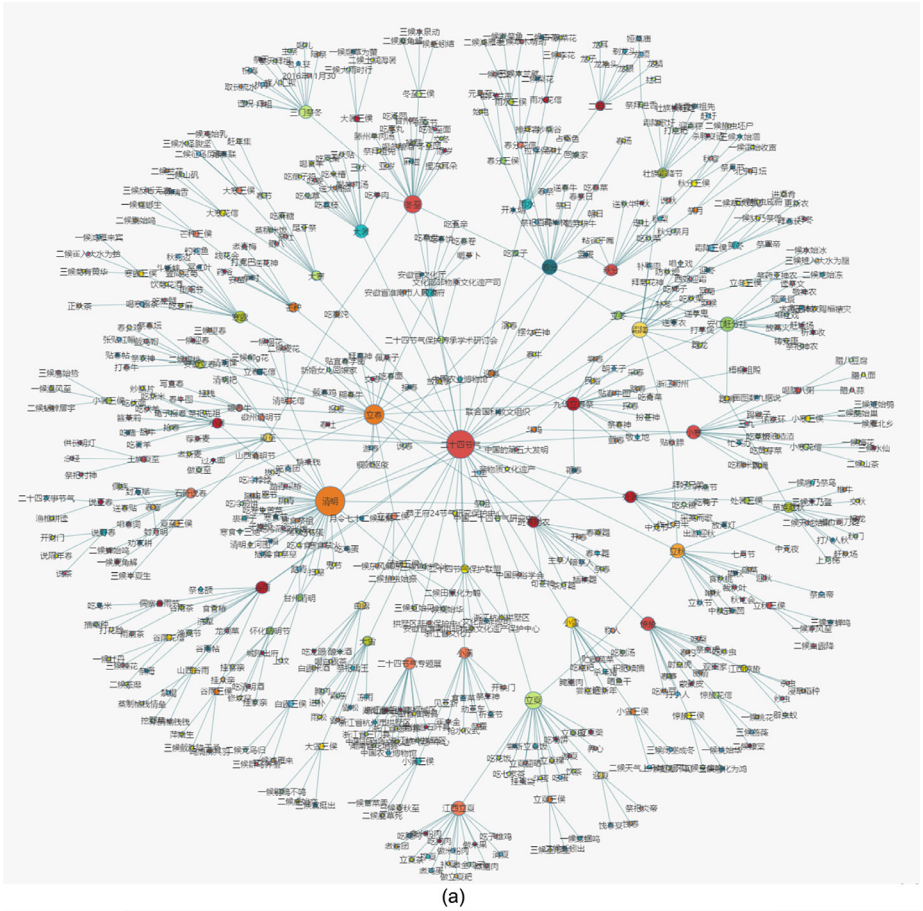


**Fig. 9.** Part of the knowledge graph visualization for the 24 Solar Terms using node with text description.

## References


[1] U. Headquarters, R. Paris, Convention for the safeguarding of the intangible cultural heritage, Unesco Org. (2003).
[2] The State Council of the People's Republic of China, Notifications by the State Council on the publication of the first batch of national intangible cultural heritage list, http://www.gov.cn/zwgk/2006-06/02/content_297946.htm, (2006) (Accessed 12 January 2017).
[3] The State Council of the People's Republic of China, Notifications by the State Council on the publication of the second batch of national intangible cultural heritage list, http://www.gov.cn/zwgk/2008-06/14/content_1016331.htm, (2008) (Accessed 12 January 2017).
[4] The State Council of the People's Republic of China, Notifications by the State Council on the publication of the third batch of national intangible cultural heritage list, http://www.gov.cn/zwgk/2011-06/09/content_1880635.htm, (2011) (Accessed 12 January 2017).
[5] The State Council of the People's Republic of China, Notifications by the State Council on the publication of the fourth batch of national intangible cultural heritage list, http://www.gov.cn/zhengce/content/2014-12/03/content_9286.htm, (2014) (Accessed 12 January 2017).
[6] A. Castiglione, F. Colace, V. Moscato, F. Palmieri, CHIS: A big data infrastructure to manage digital cultural items, Future Generation Computer Systems 86 (2018) 1134–1145.
[7] F. Colace, M.D. Santo, L. Greco, A. Chianese, V. Moscato, A. Picariello, CHIS: cultural heritage information system, Int. J. Knowl. Soc. Res. 4 (4) (2013) 18–26.
[8] F. Colace, M.D. Santo, V. Moscato, A. Picariello, F.A. Schreiber, L. Tanca, PATCH: A Portable Context-Aware ATlas for Browsing Cultural Heritage, Data Management in Pervasive Systems, Springer Publishing Company, New York, 2015, pp. 345–361.
[9] G. Zhang, Y. Yang, X. Zhai, W. Huang, J. Wang, Public cultural big data analysis platform, IEEE Second International Conference on Multimedia Big Data(April 20-22), Taipei, 2016, pp. 398–403.
[10] G. Zhang, W. Jian, W. Huang, H. Su, L. Zhi, Y. Qi, S. Ye, Big data collection and analysis framework research for public digital culture sharing service, IEEE International Conference on Multimedia Big Data (Apr 20-22), Beijing, 2015, pp. 196–199.
[11] Opinions on Speeding up the Construction of Modern Public Cultural Service System, http://www.gov.cn/xinwen/2015-01/14/content_2804250.htm, (2015) (Accessed 7 May 2016).
[12] Intangible cultural heritage, http://www.ihchina.cn/, (2006) (Accessed 12 December 2016).
[13] P. Dömel, Webmap: a graphical hypertext navigation tool, Comput. Netw. Isdn. Syst. 28 (95) (1994) 85–97.
[14] A. Singhal, Introducing the knowledge graph: things, not strings, The official Google blog, 2012 http://googleblog.blogspot.com/2012/05/introducing-knowledge-graphthings-not.html (Accessed 01 November 2016).
[15] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Vanconver, 2008, pp. 1247–1250.
[16] J. Pujara, H. Miao, L. Getoor, W.W. Cohen, Ontology-aware partitioning for knowledge graph identification, The Workshop on Automated Knowledge Base Construction, 71 ACM, 2013, pp. 19–24.
[17] H. Arnaout, S. Elbassuoni, Effective searching of rdf knowledge graphs, J. Web Semant. (2017).
[18] V. Fionda, C. Gutierrez, G. Pirrò, Building knowledge maps of web graphs, Artif. Intell. 239 (C) (2016) 143–167.
[19] T. Gruber, Ontology, in: L. LIU, M.T. ÖZSU (Eds.), Encyclopedia of Database Systems, Springer , Boston, MA, 2009, pp. 1963–1965.
[20] N.F. Noy, Ontology Development 101: A Guide to Creating Your First Ontology: Knowledge Systems Laboratory, Stanford University, 2001.
[21] The CIDOC CRM, http://www.cidoc-crm.org/. (Accessed on 12 October 2017).
[22] L.M. Surhone, M.T. Tennoe, S.F. Henssonow, CIDOC Conceptual Reference Model, Archive2 Official 40 (5) (2010) 212.
[23] T. Messaoudi, P. Véron, G. Halin, L.D. Luca, An ontological model for the reality-based 3d annotation of heritage building conservation state, J. Cult. Heritage (2017) 1–13.
[24] R.A. Haubt, P.S.C. Taçon, A collaborative, ontological and information visualization model approach in a centralized rock art heritage platform, J. Archaeol. Sci. Rep. 10 (2016) 837–846.
[25] Y. Yang, G. Zhang, J. Wang, S. Ye, J. Hu, Public cultural knowledge graph platform, IEEE, International Conference on Semantic Computing, 2017, pp. 322–327.
[26] N. Crofts, M. Doerr, T. Gill, The CIDOC conceptual reference model a standard for communicating cultural contents, Cultivate Interact. 9 (2003) 1–14.
[27] J. Unbehauen, S. Hellmann, S. Auer, C. Stadler, Knowledge Extraction from Structured Sources[C]//Search Computing, Springer Berlin Heidelberg, 2012, pp. 34–52.
[28] J. Cowie, W. Lehnert, Information extraction, Commun. ACM 39 (1) (1996) 80–91.
[29] L.F. Rau, Extracting company names from text, artificial intelligence applications, 1991, Proceedings, In: Seventh IEEE Conference on IEEE Xplore, 1991, pp. 29–32.
[30] X. Liu, S. Zhang, F. Wei, M. Zhou, Recognizing named entities in tweets, Proceedings of Acl, 1 2011, pp. 359–367.
[31] A. Jain, M. Pennacchiotti, Open entity extraction from web search query logs, Proc of the 23rd Int Conf on Computational Linguistics, Stroudsburg, PA:ACL, 2010, pp. 510–518.
[32] Y. Kim, Convolutional neural networks for sentence classification, in: 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2014), 1746–1751.
[33] N. Kalchbrenner, E. Grefenstette, P.A. Blunsom, Convolutional neural network for modelling sentences, Eprint Arxiv. (2014) 1.
[34] P. Wang, J. Xu, B. Xu, C. Liu, H. Zhang, F. Wang, H. Hao, Semantic clustering and convolutional neural network for short text categorization, Proceedings ACL, 2015, pp. 352–357.
[35] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, X. Wang, Modeling mention, context and entity with neural networks for entity disambiguation, International Conference on Artificial Intelligence, IJCAI, 2015, pp. 1333–1339.
[36] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, International Conference on Learning Representations (ICLR), 2016, pp. 1–13.
[37] Emma Strubell, Patrick Verga, David Belanger, Andrew McCallum, Fast and accurate entity recognition with iterated dilated convolutions, EMNLP, 2017, pp. 2670–2680.
[38] T.H. Nguyen, R. Grishman, Relation extraction: perspective from convolutional neural networks, Workshop on Vector Modeling for NLP, 2015, pp. 39–48.
[39] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, Attention-based bidirectional long short-term memory networks for relation classification, Meeting of the Association for Computational Linguistics, 2016, pp. 207–212.
[40] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, Meeting of the Association for Computational Linguistics, 2016, pp. 2124–2133.
[41] C. Lee, LSTM-CRF models for named entity recognition, IEICE Trans. Inf. Syst. E100–D (4) (2017) 882–887.
[42] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, Comput. Sci. (2014).
[43] http://www.crownpku.com. (Accessed 12 August 2017).
[44] UNESCO inscribes China's '24 Solar Terms' on ICH list, http://www.scio.gov.cn/32618/Document/1533265/1533265.htm, (2016) (Accessed 12 October 2017).
[45] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Comput. Sci. (2013) 1–12.
[46] T. Mikolov, Q.V. Le, I. Sutskever, Exploiting similarities among languages for machine translation, Comput. Sci. (2013).