The Chinese Knowledge Graph on Domain-Tourism

Weizhen Zhang, Han Cao(✉), Fei Hao, Lu Yang, Muhib Ahmad, and Yifei Li

School of Computer Science, Shaanxi Normal University, Xi'an 710119, Shaanxi, China
{zhangweizhen, caohan, fhao, lu-yang, liyifei}@snnu.edu.cn

Abstract. Tourism plays an increasingly important role in people's daily life. However, in the era of big data, tourists find, it is difficult to acquire useful knowledge of travel information on the Internet. Knowledge Graph describes the real-world concepts, entities and their relationships in a structured form. Which is capable of mapping the Internet information in the form that is closer to the human cognition. Thus, provides the ability to organize, manage and understand the vast travel information available on the Internet. In order to promote the sharing of tourism knowledge and culture in China, we are committed to constructing the Chinese knowledge graph on domain-tourism, in which knowledge is obtained from the existing Chinese encyclopedia knowledge graph and unstructured web pages. In addition, we provide a storage scheme of RDF triples data in the graph database and build the tourism knowledge application platform based on it.

Keywords: Knowledge graph · Tourism · RDF

1 Introduction

Recently, with the development of Semantic Web, knowledge representation has evolved from early first-order logic (e.g., FOIL [1], which learns probabilistic Horn clauses.), production rule to RDF and OWL. Based on these representations, a large amount of structured knowledge in different fields is generated and published by building a common ontology base [2]. Knowledge graph based on semantic web, which is further divided into two categories: general-purpose knowledge graph (GKG) and domain-specific knowledge graph (DKG). General-purpose includes WordNet [3], DBPedia, YAGO [2], Freebase, etc. The study of Chinese knowledge graph can be traced back to HowNet project in China. In the industry, there are OpenKG.CN, Baidu Zhixin, Sogou Zhicube. The academic circle includes XLore, zhishi.me [4], and CN-DBpedia [5]. However, there are few Chinese domain-specific knowledge graph. Especially, there is still shortage of Chinese knowledge graph on tourism domain, which seriously hinders the development and inheritance of Chinese

J. J. Park et al. (Eds.): MUE 2019/FutureTech 2019, LNEE 590, pp. 20–27, 2020.
https://doi.org/10.1007/978-981-32-9244-4_3

tourism culture. Therefore, we are committed to build a Chinese domain specific tourism knowledge graph. The key contributions of this work are concluded as follows:

1. We have studied the neural network for word vector representation model in natural language processing (NLP), and these models use to generate Chinese word vectors to achieve the purpose of entity alignment.
2. After acquiring tourism knowledge, The Chinese domain-tourism ontology is constructed by protégé and save it as RDF triples.
3. We map the tourism knowledge to the ontology, then the RDF triples data built and stored in the Neo4j graph database. And finally based on the graph database a tourism knowledge application platform built.

The rest of this paper is organized as follows. In Sect. 2, the construction of Chinese domain-tourism knowledge graph is described. Section 3 describes the construction of the domain-tourism ontology, the storage of tourism ontology and the knowledge application platform developed on it. Finally, we conclude the paper and point out further work in Sect. 4.

2 Construction of Chinese Domain-Tourism Knowledge Graph

Figure 1 shows the development architecture of our Chinese domain-tourism knowledge graph. It mainly includes three parts: acquisition of tourism knowledge, knowledge fusion and knowledge completion. And we develop the Chinese domain-tourism knowledge application platform based on it.

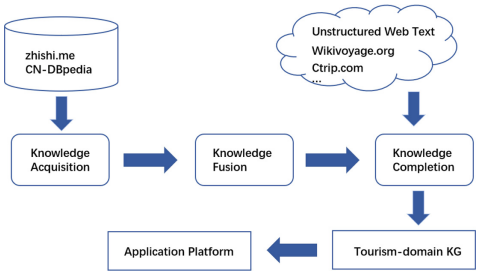

Fig. 1. Development architecture of Chinese domain-tourism Knowledge graph.

2.1 Knowledge Acquisition

The key data sources of this work are as follows: 1. Zhishi.me[1]: Zhishi.me provides dump download. This study uses turtle format data (RDF triple format). We obtain the links of tourism-related entities by searching for keywords such as travel, attractions,

[1] http://openkg.cn/dataset/zhishi-me-dump.

and scenic spots from the official website of zhishi.me. Then we extract the relevant knowledge from the downloaded data. The main knowledge extracted includes: abstract, category, infobox property and label. 2. CN-DBpedia is a large-scale general-purpose structured knowledge base. Its official website provides free data download service. We extract the travel-related knowledge from the structured triplet data that we downloaded.

2.2 Knowledge Fusion

Through knowledge acquisition, we obtain the entity, relationship and entity property information from data. After obtaining these goals, we need to integrate them to combine the properties, texts, and relationship information refer to same entity in different knowledge bases. The work we have done in this step focuses on entity alignment in knowledge fusion. After the entity alignment, we fuse the relevant information of the same entity and finally save it as triple.

Entity Alignment. Entity alignment refers to the process of judging whether there are other entities in the knowledge base with the same meaning, that is, whether one or more entities in the knowledge base points to the same object in the real language context [6]. The entity in the knowledge base may have multiple expressions, such as: different entity names of "terracotta army" and "terracotta warrior", and these different entities map to the same entity through entity alignment.

We have done vast study on word vector representation model in NLP, by using these models to calculate the cosine distance among entities in the semantic space, that is, the semantic similarity to get the purpose of entity alignment.

Skip-Gram Model. The Skip-gram model predicts the context through the target word, and the goal is to find word representation that is useful for predicting surrounding words in a sentence or a document [7]. More detail, given a sequence of training words: $w_1, w_2, w_3, \ldots, w_T$, the purpose of the Skip-gram model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{0<|j|\le c}\log p(w_{t+j}|w_t) \tag{1}$$

where c denotes the size of the training window. The primary Skip-gram defines $p(w_{t+j}|w_t)$ using the softmax function:

$$p(w_0|w_1) = \frac{exp(v'^{T}_{wo}\, v_{w_I})}{\sum_{w-1}^{W} exp(v'^{T}_{w}\, v_{w_I})} \tag{2}$$

where v_w and v'_w represent the input and output vector representations of w respectively, and W is the total number of words. Mikolov also proposes using Hierarchical softmax or negative sampling category to improve the Skip-gram model [8].

Dataset. The web pages in Sogou-T[2] and Chinese Wikipedia dump[3] as the text corpus are used. Sogou-T is provided by a Chinese commercial search engine, which contains 2.7 billion words in total and Chinese Wikipedia dump contains approximately 2 billion words.

Settings. During training, the dimensions of word are dynamically selected ranging from 100 to 500 randomly. And we set the training iterations to be 100, 150 or 200. The word frequency less than 5 in the corpus is discarded, the Hierarchical Softmax algorithm is used as the improvement of the Skip-gram model, and for learning rate η, its initial value is 0.025.

Word2vec[4] is an efficient toolkit of Google to obtain word vectors that released in 2013. The implementation of the Skip-gram model in this paper is based on the toolkit.

In this paper, total 7 groups of experiments are carried out, and we select 95 tourism-related entities as sample words. The word vector (200 dimensions) of 61 entities is obtained through the training experiment with 200 training iterations. By calculating and obtaining the most similar entities, 36 entities aligned, with an accuracy rate of 59%. To exemplify the learned representations, in below we show the most similar entities for some sample entities in below Table 1. It proves that we have obtained the same pointed entity in the real environment, that is, the effect of entity alignment. For instance, "terracotta army" and "qin terracotta army", "terracotta warrior" all points to the same entity in real life, similarity "temple" and "monastery", "shrine" referring to the same entity. The CBOW model is similar to the SG model, so we are not repeating it again.

Table 1. Word Similarity result when trained using the Skip-gram model.

Entity	兵马俑(terracotta army)	度假(vocation)	旅游(tourism)	寺庙(temple)
Top Similar Entities	秦兵马俑(qin terracotta army)	渡假(holiday)	观光旅游(sightseeing tour)	寺院(monastery)
	陶俑(terracotta figurine)	度假胜地(vacation spot)	旅游业(tourism industry)	佛寺(buddhist temple)
	秦俑(terracotta warrior)	度假村(resort)	特色旅游(characteristic tourism)	庙宇(shrine)
	俑(tomb figure)	度假地(holiday resort)	观光(sightseeing)	该寺(the temple)
	秦始皇陵(qin shi huang mausoleum)	休闲(leisure)	旅游观光(tour and sightseeing)	道观(taoist temple)
	俑坑(figurine pit)	旅游胜地(tourist attraction)	生态旅游(ecotourism)	宫观(taoist shrine)
	秦始皇(qin shi huang)	避暑(prevent sunstroke)	旅游景点(tourism spot)	该庙(the shrine)
	铜车马(bronze chariot and horse)	旅游(tourism)	自驾游(self-driving tour)	庙(temple)
	汉墓(hantomb)	胜地(famous spot)	旅游区(tourist area)	寺观(taoist temple)
	古墓(ancient tomb)	疗养(recuperation)	红色旅游(red tourism)	清真寺(mosque)

Directional Skip-Gram Model. The DSG model of Tencent AI Lab is based on word vector training model of SG. On the basis of the co-occurrence relationship of words in the text window, an additional position vector is added to represent the relative position

[2] https://www.sogou.com/labs/resource/t.php.

[3] https://dumps.wikimedia.org/zhwiki/.

[4] https://code.google.com/p/word2vec/.

of target words in the given context, to improve the accuracy of semantic representation of word vectors. DSG model proposes an additional directional softmax function:

$$g(w_{t+j}, w_t) = \frac{exp\left(\delta^T_{w_{t+j}} v_{w_t}\right)}{\sum_{w=1}^{W} exp\left(\delta^T_{w_{t+j}} v_{w_t}\right)} \tag{3}$$

To measure the context w_{t+j} is more biased to the left or right of the target word w_t, where W denotes the total number of words in the corpus. v_w represents the word vector, and δ represents the direction vector of each word in the context [9]. Moreover, the g function has similar update strategy to the negative sampling technique: increase the probability of positive samples while reduce the probability of negative samples.

The final softmax function of DSG model is added to formula (2): $f(w_{t+j}, w_t) = p(w_{t+j}|w_t) + g(w_{t+j}, w_t)$.

We use the open source word vector published by the Tencent AI Lab. Using Chinese Wikipedia redirection (synonym) entity pairs as test data, total 22797 pairs, by computing the cosine distance of entity pairs to characterize the result of the entity alignment. In our experiment, the accuracy of entity alignment results is obtained by setting different thresholds of similarity. From the experimental results, when the threshold of similarity is set to 0.65, the accuracy rate of entity alignment as high as 91.67%. However, as the threshold continues to increase, the accuracy rate of entity alignment reduced, and when the threshold is set to 0.85, the accuracy rate becomes reduced to 45.26%.

BERT Model BERT addresses the Generative Pre-trained Transformer unidirectional constraints by proposing a new pre-training target. Model named "masked language model" (MLM) [10]. The masked language model masks randomly some of the tokens from the input, and predict the original vocabulary id of the masked word based only on its context [11]. Furthermore, BERT adds an additional sentence-level continuous prediction task on the basis of bidirectional language model, aiming at predicting whether the text at both ends of the input is continuous text. As a result, BERT representations are fine-tuned with just one additional output layer to create state-of-the-art models for a variety of tasks, without substantial task-specific architecture modifications [11].

We achieve entity alignment by using Google's pre-training BERT (BERT-Base, Chinese) model based on character level to get the word vector of entities. 22,797 pairs of Chinese Wikipedia entity pairs are extracted by us and HIT IR-Lab Tongyici Cilin (Extended) dataset, that are used for experimental data. When the thresholds are set to 0.65, 0.70, 0.75, 0.80 and 0.85, respectively, the accuracy of entity alignment is more than 99%. We also conduct a comparative experiment with the three models mentioned above. For a fair comparison among different models, we use same dimension size for all word embeddings, and discard the rare words that appeared less than 5 times in the training corpus. For learning rate η, its initial value is 0.025. The window size and negative samples are both set to 5. The experimental results are shown in Table 2.

Table 2. Accuracy of using different models in entity alignment. CWEP represents Chinese Wikipedia entity pairs dataset and HTTTCL represents HIT IR-Lab Tongyici Cilin (Extended) dataset.

	SG		CBOW		DSG		BERT	
0.65	0.58583	0.808177	0.522655	0.741696	0.902535	0.916743	0.99698	0.998947
0.7	0.482811	0.727069	0.439567	0.664587	0.808676	0.85599	0.99547	0.997894
0.75	0.369222	0.623766	0.347652	0.566432	0.767733	0.767733	0.997456	0.997456
0.8	0.259698	0.500329	0.246911	0.446793	0.690393	0.634031	0.994425	0.997061
0.85	0.151336	0.355442	0.150288	0.301582	0.358107	0.452603	0.987805	0.996008
	CWEP	HITTCL	CWEP	HITTCL	CWEP	HITTCL	CWEP	HITTCL

From the experimental results, compared with SG and CBOW models, DSG model adds an additional position vector to represent the relative position of target words in the given context, which improved the accuracy of semantic representation of word vectors. BERT model through pre-train deep bidirectional representations by jointly conditioning on both context in all layers, obtains state-of-art results on accuracy of semantic representation of the word vector. All these ideas have reference significance for us to study entity alignment.

2.3 Knowledge Completion

In the process of knowledge completion, the work we have done is mainly to complete the infobox properties of entities. Infobox properties is a form of knowledge, which is used to summarize the features of entities [12]. As the properties extracted from zhishi.me and CN-DBpedia are limited, we extract the properties related to tourism entities from the unstructured network text to complete the infobox properies. We use regular expressions in Python scripts to extract entity properties information from tourism-related sites (e.g. wikivoyage), mainly by making rules manually.

3 Construction of Chinese Domain-Tourism Ontology and Application Platform

To acquire structured knowledge, it is necessary to construct the ontology. By summarizing the concepts, properties and relations in the data, we determine the category structure of the domain-tourism ontology, combined with the ontology construction method of Stanford university, “seven-step method”, and use the ontology editing tool Protégé to complete our Chinese domain-tourism ontology construction.

Furthermore, the ontology saves in RDF format. RDF data usually appears in form of triples (S, P, O), i.e., (subject, predicate, object). An entity is usually described by multiple triplet information, so an entity triplet information can form an RDF directed subgraph, each triple is represented in the graph as “node-edge-node” relationship. Neo4j is a graph-oriented database whose primitives are nodes, relationships, and attributes. In this paper, we create a mapping between the labels of RDF data and the Neo4j database, store the RDF data in the Neo4j graph database.

Finally, we developed a domain-tourism knowledge application platform based on knowledge graph. The main functions are as follows: (1) Visualize the urban tourism heat and color depth represents the heat of tourism; (2) Semantic search: combine the knowledge base to parse the entities entered by users, so that users' intentions can be more accurately understood, and entity overview can be given in the form of knowledge cards; (3) Providing tourism-domain knowledge service, in which the specific scenic spots will display the deep knowledge information such as history, culture, geography, climate, sightseeing route, food, accommodation, etc.

4 Conclusion and Future Work

In this work, after information extraction, we introduce neural network for representation models to get the word vectors of the tourism entities, which are used to achieve the purpose of entity alignment. Combined with other knowledge graph technology we build the domain-tourism knowledge graph and develop a knowledge application platform.

This paper is beneficial exploration and practice from tourism information service to tourism knowledge service. In the future, in-depth research will be conducted on the aspects of question answering, accurate tourism entity links and automatic construction of tourism ontology, and recommendation systems.

Acknowledgements. This work is supported by the primary research and development plan of Shaanxi Province "Research on Characteristic Module Design of SOTA Platform for Intelligent Tourism" (NO: 2018SF-361), the National Key R&D Program of China under grant No. 2017YFB1402102, and the Fundamental Research Funds for the Central Universities (GK201806012).

References

1. Quinlan, J.R., Cameron-Jones, R.M.: Foil: a midterm report. In: Proceedings of ECML (1993)
2. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706. ACM (2007)
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
4. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi.me: weaving chinese linking open data. In: International Conference on the Semantic Web. Springer, Heidelberg (2011)
5. Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., et al.: CN-DBpedia: a never-ending chinese knowledge extraction system. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham (2017)
6. Hao, Y., Zhang, Y., He, S., Liu, K., Zhao, J.: A joint embedding method for entity alignment of knowledge bases. In: Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data. Springer, Singapore (2016)

7. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: International Conference on Machine Learning (2008)
8. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. Adv. Neural. Inf. Process. Syst. **26**, 3111–3119 (2013)
9. Song, Y., Shi, S., Li, J., Zhang, H.: Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In: NAACL 2018 (Short Paper)
10. Taylor, W.L.: Cloze procedure: a new tool for measuring readability. Journal. Bull. **30**(4), 415–433 (1953)
11. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bi-directional transformers for language understanding (2018)
12. Wu, T., Gao, C., Qi, G., et al.: KG-Buddhism: the Chinese knowledge graph on Buddhism (2017)