EasyChair Preprint

№ 4013

Construction of Multi-modal Chinese Tourism Knowledge Graph

Xie, Qinghua Wen, Hailong Jin, Zhenhao Dong, Lei Hou, Hongyin Zhu and Juanzi Li

August 9, 2020

Construction of Multi-modal Chinese Tourism Knowledge Graph

No Author Given

No Institute Given

Abstract. This paper proposes a construction process of **M**ulti-modal **C**hinese **T**ourism **K**nowledge **G**raph (MCTKG). In order to maintain the semantic consistency of heterogeneous data sources, the tourism ontology is constructed by semi-automatic method, and the data sources are aligned to the tourism ontology. The construction process of MCTKG includes following modules: ontology construction, entity alignment, tourism route automatic generation, sharing platform establishment. In ontology construction, semi-automatic fine-tuning operations are carried out, which optimize and simplify the concept hierarchy relationship abstracted from the obtained data resources, and some new concept hierarchy relationships are manually defined according to the actual application scenarios. In entity alignment, we adopt a method based on n-gram distance to align entities from different sources, and fuse and cross-validate their attributes. In addition, based on the concepts of attraction and tourism style in the knowledge graph, we propose a tourism route generation algorithm, which could automatically schedule the tourism routes by incorporating the characteristics of attraction and tourists' demands. Lastly, a sharing platform is established, which provides open data access and query interface.

Keywords: Tourism ontology · Entity alignment · Knowledge graph · Route generation.

1 Introduction

Knowledge graph has been widely used in various domain including finance, medicine and e-commerce. However, in domain of tourism, it is not mature enough. Although the development of the Internet has given birth to enormous multi-source tourism information and some websites are organized, information heterogeneity is still the bottleneck of users' information acquisition and many downstream applications, so it is urgent to build a tourism knowledge graph.

Many companies, such as Ctrip, have built their own knowledge graph applications. Chinese tourist attractions knowledge graph launched by the Institute of automation, Chinese Academy of Sciences, which extracted from Baidu Encyclopedia and Interactive Encyclopedia and can be used for geography, life and entertainment. Different from these existing knowledge graph of tourism field, the tourism knowledge graph constructed in this paper focuses on the travel

itinerary and attraction information that tourists really care about, covering the knowledge of food and abundant image modal information, which can directly serve the needs of tourists and assist tourists in the planning of tourism itinerary. In the process of constructing MCTKG, mainly faces on the following challenges.

1) Tourism ontology construction. At present, there is no mature and available tourism ontology. Therefore, we need to consider the actual needs and the conceptual hierarchy of the data and reconstruct it.

2) Semantic information extraction. Since there are many different data sources that constitute MCTKG, the information extracted from these heterogeneous data sources needs a series of complex preprocessing processes, such as data filtering, structuring and semantic alignment.

3) Entity alignment. For the entities from different data sources, the aligned entity pairs are fused with their attribute values.

4) Tourist routes generation. Considering the practical application needs, it is necessary to generate tourism routes reasonably according to the attraction information. The generated routes are defined into MCTKG.

Based on the above analysis, we propose a construction process of MCTKG and study the key technologies. For realizing the query and sharing of knowledge in the knowledge graph, this paper also constructs the application platform of MCTKG, and opens the data access and query interface.

2 Related Work

Since the 1990s, semantic web related technologies have begun to flourish, and ontology technology has become a research hotspot. A number of excellent ontology knowledge bases have begun to emerge, such as DBpedia [14] and WordNet [15], which marking the maturity of semantic web technology and entering the stage of practical application. However, since the construction of ontology knowledge base is a very complex and time-consuming systematic project, its progress is relatively slow and has become one of the bottlenecks in the development of ontology technology. Therefore, it is urgent to research and construct various ontology knowledge bases.

Internationally, with the release of Google knowledge graph, the construction and application of knowledge graph has attracted extensive attention from academia and industry. Chun Lu et al. [18] constructed a rich world scale travel knowledge graph from existing large knowledge graphs namely Geonames, DBpedia and Wikidata. The underlying ontology contains more than 1200 classes to describe attractions.

The research on ontology construction technology in China is still in its infancy. In terms of domain ontology construction, although there have been many important research results, such as Sogou's Knowledge Cube and Baidu' s Zhixin. GDM Laboratory of Fudan University designed a Chinese knowledge graph [19] for book reading. Weiwei Wang et al. [20] constructed the bilingual films knowledge graph, including the construction of films ontology base, entity link, entity matching, and built the application platform and open data

access interface. Shijia E et al. [21] proposed an end-to-end automatic construction scheme of Chinese knowledge graph based on Chinese encyclopedia data, and developed a user-oriented Chinese knowledge graph system. But in general, there are few fields involved, which can not meet the needs of practical application in terms of scale and quality. Especially in the tourism domain with wide application prospects, there is no high-quality knowledge base in Chinese.

3 The Process of Constructing MCTKG

The basic construction process of MCTKG includes five steps, as shown in Figure 1.

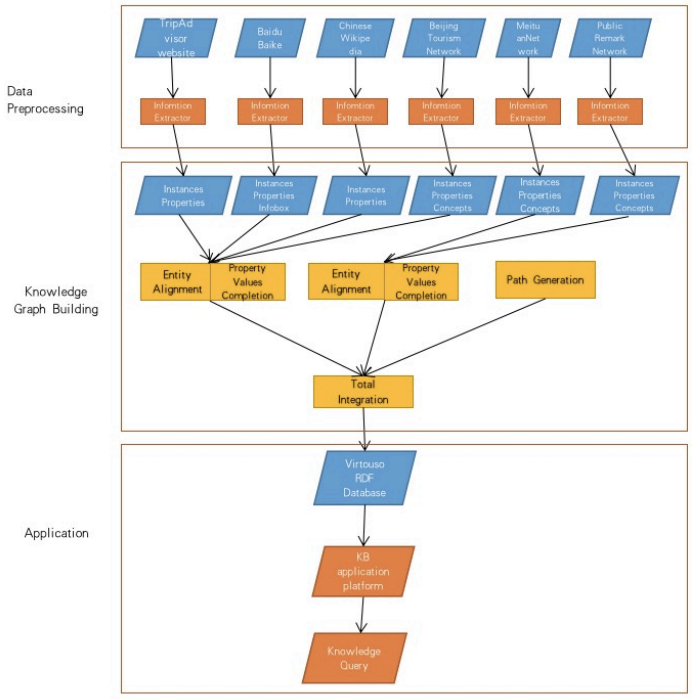

Fig. 1. Pipeline of the MCTKG.

1) Structured knowledge extraction. Extract structured knowledge from various data sources, and align these knowledge to the constructed tourism ontology.

2) Tourism ontology construction. The tourism ontology is constructed semi-automatically through optimizing and simplifying the concept hierarchy relationship abstracted from data resources, and combining with the actual needs and application scenarios of tourists.

3) Entity alignment. The equivalent entities from heterogeneous data sources are aligned. For aligned entity pairs, their attribute values are fused with each other, which can solve the problem of missing attribute values to a certain extent, and finally realize the knowledge fusion of different data sources.

4) Tourism route automatic generation. Based on the concepts of attraction entity and tourism style in knowledge graph, this paper proposes an algorithm for automatic generation of tourism route. The route generated by the algorithm is defined in the knowledge base, and recommended to tourists as a recommended route in the needs of tourists.

5) Sharing platform of MCTKG. Build such a platform to realize data visualization and help tourists access and query the knowledge base.

3.1 Various Heterogeneous Data Sources

The data needed for the construction of tourism knowledge graph are semi-structured data obtained from major tourism websites on the Internet, mainly including Beijing tourism website, Baidu Encyclopedia, Wikipedia, TripAdvisor, meituan website and dianping website.

Beijing tourism website is the largest tourism portal in Beijing, covers all the tourist attractions, hotels, hot springs and resorts in Beijing and its surrounding areas. It is the first choice for Beijing travel to learn about tourism information. We extract a total of 1574 attraction entities, as well as 16 attributes such as the address, level and ticket information of attractions.From the extraction results, there are many entities, which exist a serious problem of missing attribute values.

Baidu Encyclopedia is the largest Chinese encyclopedia at present. In recent years, Baidu Encyclopedia data, especially in tourism information, has been significantly improved in terms of scale and quality. Tourism information is relatively rich, which can be used as an effective supplement to the data source of Beijing tourism website. We use the list of 1574 attractions extracted from Beijing tourism website to extract the attribute information of these attractions from Baidu Encyclopedia, including historical evolution, important events, cultural relics collection and other attributes. In addition, we also extracted 942 attributes of each attraction from the infobox of Baidu Encyclopedia, and selected only 7 attributes from these attributes after filtering Properties of the point entity. To harvest knowledge of image modal instead the text modal information from Baidu Images, we use python selenium webdriver to simulate user scroll, click actions etc by using the entity name extracted from Beijing tourism website. We extract the image modal information of each attraction from Baidu pictures and add it to the knowledge graph. Among them, low quality images are filtered out by manual filtering.

TripAdvisor is a foreign tourism review website, which provides tourism related information about the world's attractions and their nearby hotels and restaurants. At present, we have obtained 1632 attractions. Considering the low quality of some attractions, we selected some attribute information of attractions that appeared in the list of attractions to the knowledge base.

Wikipedia is a multi-language encyclopedia collaboration project based on Wiki technology. It is a network encyclopedia written in many languages. Wikipedia contains a wealth of museum information. We extract the attribute information of museum entities.

Meituan website and dianping website are two of China's leading local life information and trading platforms. They can not only provide users with information services such as merchant information, consumer reviews and consumer discounts, but also provide transaction services such as group purchase, restaurant reservation. In order to meet the dining needs of tourists and show the culture of Beijing time-honored brands for tourists, we aligned the entities of time-honored stores and got a total of 1771 time-honored store entities and 16 attributes such as store score, store address, contact phone number, average price, etc.

For each data source, we mainly get the source code of the web page through crawler tools, and then use regular expressions to extract the entity and entity attribute information from the source code.

3.2 Ontology Construction

Ontology is an abstract model for describing the objective world, which gives a clear definition of concepts and their relations in a formal way. The common approaches are divided into three types: semi-automatic construction, automatic construction and manual construction. For the tourism knowledge graph to be constructed in this paper needs to be able to directly participate in solving the actual problems in the tourism scene, so, it is difficult to adapt to this demand through automatic construction technology. Therefore, this paper uses the approach of semi-automatic ontology construction, through comprehensive consideration of various factors to complete the ontology construction. This paper introduces the steps of constructing tourism domain ontology.

1) Ontology and terms are abstracted from obtained data resources. According to the data extracted from Beijing tourism website, we can get the concept hierarchy relationship of tourist attractions, and then we can get the concept classification system of attractions by optimizing and simplifying these concept hierarchical relations.

2) Combined with the practical problems to be solved and the specific application scenarios, a new concept hierarchy is added. Traditional ontology construction methods often do not consider the final application scenarios and practical problems to be solved in the construction stage. We combine the actual problems, and define the hierarchical relationship of the concepts of tourism style, time-honored brand, time-honored stores, dishes and so on.

3) Ontology iterative evolution. The construction of ontology is a complex process. After the completion of the construction, it is necessary to observe the effectiveness and rationality of the ontology from various aspects, and to iterate and optimize the ontology according to the actual feedback.

The illustration of concept classification system is shown in Figure 2.

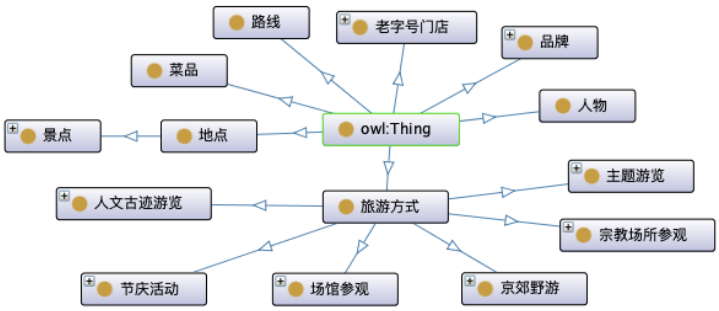

Fig. 2. Illustration of Concept Classification System.

3.3 Structured Knowledge Extraction

Structured knowledge extraction is a process of extracting knowledge from heterogeneous data sources, analyzing data of various formats, unifying semantics and structure, and roughly includes the following two modules.

1) Web page analysis. This module mainly obtains the source code of web page through crawler tool, and then extracts entity and entity attribute information from source code by regular expression.

2) Attribute value completion. The main task of this process is to align entities from heterogeneous data sources, and then complete the attribute values of the same attributes of the two entities after alignment, so as to achieve the purpose of knowledge fusion.

After the above two steps, the source data is transformed into structured JSON format data.

3.4 Entity Alignment

Due to the serious problem of missing attribute value exists in entities obtained from Beijing tourism website, we have carried out entity alignment for realizing the knowledge reuse and fusion of different heterogeneous data sources. For attraction entities, the entities obtained from Wikipedia, TripAdvisor, Baidu Encyclopedia and other data sources are aligned with the attraction entities obtained from Beijing tourism website, and the attribute values of aligned entity pairs are fused. To solve this problem, we adopt an entity alignment algorithm based on n-gram distance to align entities between different data sources.

Definition (N-gram Distance) Firstly, we express two entities A and B from different data sources into their n-gram forms, and obtain two sets S and T. then we calculate their n-gram distances. The n-gram distances are defined as follows.

$$n\_gram\_distance = |GN(S)| + |GN(T)| - 2 * |GN(S) \cap GN(T)| \quad (1)$$

Where $|GN(S)|$ is the size of the n-gram set of entity A, $|GN(T)|$ is the size of the n-gram set of entity B, and $|GN(S) \cap GN(T)|$ represents the number of identical element pairs in set S and set T.

The pseudo code of entity alignment algorithm based on n-gram distance is shown in Algorithm 1.

Algorithm 1 Entity Alignment Algorithm Based on N-gram Distance

Input: – Entity set obtained from Beijing Tourism Website A={a1, a2, ..., an}
– Entity set obtained from other data resource B={b1, b2, ..., bm}
– Distance threshold Q

Output: – Matching dictionary d

```
function Entity_Alignment(A, B, Q)
    for i = 0 → n do
        min_distance ← INT_MAX
        n ← len(A[i])
        while 1 < n do
            for j = 1 → m do
                S ← n_gram_set(B[j], n)
                T ← n_gram_set(B[j], n)
                distance ← n_gram_distance(S, T)
                if distance < min_distance then
                    min_distance ← distance
                end if
                if min_distance < Q then
                    break
                end if
                n ← n − 1
                if n! = 1 then
                    d[A[i]] ← j
                end if
            end for
        end while
    end for
    return d
end function
```

Evaluation We extract 1574 and 1149 attraction entities from Beijing tourism website and Baidu Encyclopedia respectively. In order to align these entities, we use an n-gram distance based alignment algorithm, in which the threshold Q is set to half of the sum of set s and set t. In order to analyze the performance of this algorithm, a relatively simple entity alignment method is adopted: if two attraction entity names are identical or one attraction entity name contains another attraction entity name, they are considered to be aligned. The entity alignment results are shown in Table 1.

The data needed for the construction of tourism knowledge graph are semi-structured data obtained from major tourism websites on the Internet, mainly including Beijing tourism website, Baidu Encyclopedia, Wikipedia, TripAdvisor, meituan website and dianping website.

Table 1. Entity Alignment Result.

Alignment algorithm	Number of matches	Number of errors	Accuracy
Simple approach	1082	37	96.69
Based on n-gram	1112	7	99.37

From Table 1., compared with the simple string matching algorithm, the accuracy of the n-gram-based entity alignment algorithm has been significantly improved. This is because the algorithm based on n-gram can effectively reduce the similarity of misaligned entities. For example, two entities named " 东宫影剧场" and " 东宫影剧院" can not be aligned by using simple entity alignment algorithm, but they can be successfully aligned by using n-gram based algorithm.

3.5 Automatic Generation Algorithm of Tourist Routes

Considering the actual needs of tourists for tourism route recommendation, we need to generate specific tourism routes according to the attraction entities obtained from various data sources and add them into the knowledge graph. In order to automatically generate tourism routes, we propose an algorithm for automatic generation of tourism routes. For each travel style, several routes are generated. The pseudo code of the algorithm is shown in Algorithm 2.

Evaluation In order to evaluate the quality of the routes generated by the algorithm, we compare the routes generated by the algorithm with those obtained from various data sources. That is, we use the approach of scoring these routes according to some evaluation indicators to judge whether the routes generated automatically are as satisfactory as those obtained from the data sources, including Cellular website, Qunar, Douban and other tourism websites.

In order to make the scoring result more reasonable and accurate, according to the proportion of the total number of the two types of routes, we generated 10 groups of route data. Each group of data randomly selected 6 and 30 routes from the online routes and automatically generated routes, that is, each group of data contains 36 tourism routes.

We recruited some experienced tourists to score these routes. The score ranges from 1.0 to 5.0, and the evaluation indicators include the rationality of travel arrangement, the rationality of time planning and the comfort of tourism experience. The final scoring results are shown in Table 2.

From the scoring results, we can see that the overall score of routes generated by the automatic generation algorithm and the scores of popular routes obtained

Algorithm 2 Automatic generation algorithm of tourist routes

Input: – Travel style set: A={A1, A2, ..., Am}
– Candidate attractions set corresponding to each travel style, Bi={C1, C2, ..., Cn} (i=1, 2, ..., m)
– Weight dictionary d: Each attraction level corresponds to a weight, The higher attraction level is, the greater the weight is, and the highest weight is 10
– Number of cycles threshold Q

Output: – List of one day tour route: One_day_tour_routes
– List of half day tour route: Half_day_tour_routes

function GENERATING_TOURIST_ROUTES(A, B, Q, d)
  for $i = 0 \rightarrow m$ **do**
    $candidate \leftarrow []$
    $count \leftarrow 0$
    for $j = 0 \rightarrow n$ **do**
      if $d[B[j]] > 7$ **then**
        $candidate.append(B[j])$
      end if
    end for
    while $count < Q$ **do**
      $count \leftarrow count + 1$
      $Attractions \leftarrow np.random.choice(candidate, np.random.choice([2, 3, 4, 5, 6]))$
      $T \leftarrow calculate\_total\_time(Attractions)$
      if $T > 6$ **and** $T <= 10$ **then**
        $Route \leftarrow Permutation\_and\_CalculateShortestDistance(Attractions)$
        $One\_day\_tour\_routes.append(Route)$
      end if
      if $T >= 3$ **and** $T <= 6$ **then**
        $Route \leftarrow Permutation\_and\_CalculateShortestDistance(Attractions)$
        $Half\_day\_tour\_routes.append(Route)$
      end if
    end while
  end for
  return $One\_day\_tour\_routes, Half\_day\_tour\_routes$
end function

from tourism websites are slightly lower, but the score gap is not too large, which indicates that the automatically generated tourism routes can meet the needs of tourists to a certain extent.

3.6 MCTKG Sharing Platform

Knowledge graph is a kind of network of knowledge relationship constructed by information visualization technology. The purpose of establishing this knowledge graph sharing platform is to display MCTKG from the perspectives of concept, instance and attribute, Figure 3 shows the sharing platform of the knowledge graph. The website is developed based on React open source framework and uses Virtuoso as database server, which mainly provides two functions: (1) basic

Table 2. Route Scoring Results.

Route Number	Score of Network Routes	Score of Generated Routes
1	4.12	4.08
2	3.88	3.79
3	4.20	4.06
4	4.20	3.88
5	4.15	3.96
6	4.08	3.98
7	4.93	4.89
8	3.96	3.97
9	4.22	4.02
10	4.00	3.95

information of tourism ontology, which can provide various statistical information of concept hierarchy and knowledge base, (2) data query interface, includes SPARQL terminal query interface, classified index query interface and composite query interface.

Fig. 3. The interface of MCTKG Sharing Platform.

In addition to the information displayed on the shared platform, Table 3 shows the comprehensive statistical data of the knowledge base. And Table 4 shows number of instances as well as image links of each concept.

Table 3. Overall Statistics.

Item	Quantity
Number of triples	390746
Number of concepts	238
Number of entities	47718
Number of properties	54
Number of imageLinks	187954
Number of routes	232

Table 4. Statistics of Each Concept.

Concept	Number of instances of the concept	Number of ImageLinks
景点	1975	39564
老字号门店	1771	11564
品牌	60	1799
路线	232	0
菜品	42430	482514
旅游方式	22	0

4 Conclusion

In this paper, a construction process of MCTKG based on multiple heterogeneous data sources is proposed, and the main problems and challenges encountered in the whole process and the solutions are described.

In fact, the construction of ontology knowledge base is a long-term, systematic complex work, which needs continuous improvement. And there are still many aspects require to be improved in construction process of MCTKG. For example, (1) seeking higher quality and more completed data sources to complete the missing attribute values of some attraction entities. (2) Establish the relational link between the concept of people and attractions. (3) The automatic updating mechanism of knowledge base is established. The method proposed in this paper has a certain reference significance for the construction of domain ontology knowledge base which needs to integrate multiple heterogeneous data sources and entity alignment in specific domain.

In general, MCTKG is a high-quality RDF tourism ontology knowledge base which integrates six heterogeneous high-quality data sources, and fills the gap of Chinese tourism ontology knowledge base in China. The knowledge base provides an important basis for the mining and utilization of tourism related information, and also has an important significance to expand the international influence of Chinese tourism information.

References

1. Liu, J.: Research on the Construction and Application of Knowledge Graph in Tourism Domain[D]. 2019.
2. Xuan, T.: Building Medical Ontology Base and Its Semantic Applications[D]. Chengdu: University of Electronic Science and technology, 2013.
3. Zhao, X., Qiu, L., Zhao, T.: Construction Technology of Ontology Knowledge Base in Multiple Minority Languages[J]. Journal of Chinese Information Processing, 2011, 25(4).
4. Zhang, W., Zhu, Q.: Research on Construction Methods of Domain Ontology[J]. Library and Information, 2011(01):16-19.
5. Xu, P.: Research and Implementation on Construction Method of Knowledge Graph in Tourism Domain[D].
6. Du, Y., Wu, Y.: Research on Constructing the Knowledge Graph Based on Microblog[J]. Journal of Xihua University(Natural Science Edition), 2015, 000(001):27-35..
7. Hu, F.: Chinese Knowledge Graph Construction Method Based on Multiple Data Sources[D]. Shanghai: East China University of science and technology, 2015.
8. Pan, J Z.: Knowledge extraction from Chinese wiki encyclopedias[J]. Journal of Zhejiang University-Science C(Computers & Electronics), 2012(04):268-280.
9. Li, Q.: Research on The Construction of Tourism Domain Ontology[D]. Zhengzhou University, 2015.
10. Wang, S.: Research on the construction of tourism destination Ontology[D]. Xiangtan University, 2016.
11. Song, W.: Research on Entity Alignment for the Medical Field[D]. Harbin Institute of Technology, 2018.
12. Su, J., Wang, Y., Jin, X., Li, M., Cheng, X.: Knowledge Graph Entity Alignment with Semantic and Structural Information[J]. Journal of Shanxi University(Natural Science Edition), 2019, 42(01):23-30.
13. Guan, S., Jin, X., Wang, Y.: Self-learning and embedding based entity alignment[J]. Knowledge and Information Systems, 2019.
14. Lehmann, J., Robert, I., Max, J.: Dbpedia—a large- scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal, 2014, 5: 1–29.
15. Miller, G A.: WordNet: a lexical database for English. Communications of the ACM, 1995, 38(11): 39–41.
16. Liu, Q., Li, Y., Duan H., Liu Y., Qin, Z.: Knowledge Graph Construction Techniques[J]. Computer research and development, 2016, 53(3): 582-600.
17. Hu, f.: Chinese Knowledge Graph Construction Method Based on Multiple Data Sources[D]. East China University of Science and Technology, 2015.
18. Lu, C., Laublet, P., Stankovic, M.: Travel Attractions Recommendation with Knowledge Graphs[C]// European Knowledge Acquisition Workshop. Springer International Publishing, 2016.
19. Xiao Yanghua, Zhang kezun, Wang, W.: A method of constructing knowledge map of reading domain for books: China, CN103488724A [P]. 2014.01.01.
20. Wang, W., Wang, Z., Pan, L., Liu, Y., Zhang, J.: Research on the Construction of Bilingual Movie Knowledge Graph[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 25-34.
21. E, S., Lin, P., Xing, Y.: Automatical construction of Chinese knowledge graph system[J]. Computer application, 2016, 36(4): 992-996.