Summer 7-26-2022

Exploratory Research on Knowledge Graph Construction and Attribute Value Extraction for Large-scale Textual Data of Tourism Products

Jiayi Xu
School of Economics and Management, China University of Geosciences, Wuhan, 430074, China

Shuang Zhang
School of Economics and Management, China University of Geosciences, Wuhan, 430074, China

Zhen Zhu
School of Economics and Management, China University of Geosciences, Wuhan, 430074, China, zhuzhen2008@gmail.com

Lincan Zou
School of Economics and Management, China University of Geosciences, Wuhan, 430074, China

Mengting Yang
School of Economics and Management, China University of Geosciences, Wuhan, 430074, China

Short Research Paper

Exploratory Research on Knowledge Graph Construction and Attribute Value Extraction for Large-scale Textual Data of Tourism Products

Jiayi Xu, Shuang Zhang, Zhen Zhu, Lincan Zou, Mengting Yang*

School of Economics and Management, China University of Geosciences, Wuhan, 430074, China

Abstract: Complexity of tourism products exists in route design and service combination provided by suppliers. Generating reasonable rules by utilizing large-scale textual data of tourism products will be an effective way to explore imitation and competition from product-product relationships thus observing how they respond to consumer demands and market changeable. In this sense, constructing a tourism product knowledge graph (TPKG) will be a key data strategy for a travel agency or e-commerce platform. This paper constructs a knowledge graph of tourism products with seven feature dimensions and creates a structured model to imagine the service details and features. BiLSTM-CRF was used to extract entities from large-scale textual data, while entity-related attributes and attribute values were extracted from Baidu Baike. Furthermore, possibility of correlation between entities using the link prediction algorithm was checked, and entity disambiguation was accomplished. Finally, TPKG was visualized using Neo4j graph to show the validity of ontological structure for tourism products. Our paper provides a new process and method to uncover the imitation and competition relationship among tourism products from a nuanced particle perspective.

Keywords: knowledge graph, tourism products, named entity recognition, attribute value extraction

1. INTRODUCTION

Under the rapid development of the Internet and high transparency of information, opportunities and challenges coexist in the e-commerce platform ecosystem, which allow enterprises to continuously accelerate updates and iterations of products by learning or imitation, launching new products or services in line with the overall market trends. The above indicates that the procedure of product innovation evolution determines the competitiveness of enterprises on e-commerce platforms in market. For tourism products, combining the route design and service is considered as essential indicators to reflect their product characteristics. Therefore, it makes significant to applicably use large-scale product data information in e-commerce platforms to present it in certain rules and analyze its concrete content, which provides the relationship between products (imitation, competition, etc.) from a micro perspective and allows us to detect product evolution characteristics, variation trends of market, etc. from the macro view as well.

In addition, with the development of artificial intelligence and the popularization of the information explosion and big data, big data analysis and proceeding have also become an important subject currently. In this context, knowledge graph, as a semantic network with strong expression capacity and modeling flexibility, has played an essential role in various fields. It was primarily proposed by Google in 2012[1] in order to significantly improve the performance of search engines, such as extracting entities, attributes, and the correlates from massive web data. The knowledge graph displays information in the form of various Subject-Predicate-Object (SPO) ternary from the reality. In some case, it will also be referred to as an extensive knowledge base, widely used in intelligent retrieval, recommendation system, policy analysis, etc.

In tourism, most of the existing research aims to establish knowledge graphs of the whole tourism field, which are mainly for constructions of recommendation systems or intelligent Question Answering platforms. Xu

* Corresponding author. E-mail address: zhuzhen2008@gmail.com

has established a knowledge graph in Chinese tourism [2]; Liu has built a knowledge graph from the perspective of tourists to assist tourists in itinerary planning[3]. However, previous studies have not investigated the construction of knowledge graphs in tourism products. Tourism products are composed of a series of parts widely recognized as five elements: physical location, service, friendliness, free choice, and participation, proposed by Smith S. in 1994[4].

In travel products, large-scale routes generally have longer travel time, which means their products have more content and complexity. Before the COVID-19 outbreak, the number of outbound tourisms in China has increased year by year since 2016. In 2019, the outbound tourism reached to 170 million, and the outbound travel expenditure reached to 127.5 billion dollars. Moreover, Europe and the United States with their high degree of development and completed social public facilities, which are symbols of high quality in outbound travel, are the most popular destinations for Chinese in recent years. Therefore, this study has selected outbound travel to Europe and the United States as the analysis objects and extracted the data from March 2017 to December 2019 on the Ctrip platform. Firstly, we put forward the ontology structure in tourism products with seven dimensions and constructed the ontology. Secondly, we extracted entities from massive data using BiLSTM-CRF and defined 8 relations. Next, we extracted attribute values from Baidu Encyclopedia to enrich our knowledge graph. Finally, the entity disambiguation was carried out by link prediction algorithm, then visualizing the knowledge graph.

We summarized two major innovations in our research:

- This paper proposes an ontology construction method for tourism products, showing the characteristics of tourism products in seven dimensions.
- This paper extracts entity from the massive product data of Ctrip with the content from Baidu Encyclopedia to build a TPKG.

2. ONTOLOGY CONSTRUCTION OF TOURISM PRODUCTS

2.1 Data collection

The data was collected from large-scale holiday-making products provided in Ctrip. Due to the pandemic ahead of the 2020 Lunar New Year holiday, the boom index of China's outbound tourism market began to decline sharply. To avoid the impact of the outbreak on outbound travel, this study took the whole data of outbound tourism to Europe and the United States from March 2017 to December 2019. We selected the data of the first day in each month, with a whole piece of 169,598, and totally left 169457 after pre-processing and removing repeated data from product IDs in each issue, among which are 85,443 outbound travel product data in Europe and 84,014 in the United States. Figure 1 shows information about a holiday-making product with destination of the United States on Ctrip, where the information of the title, subtitle, *服务保障*(service support), *供应商*(supplier), and *产品卖点*(selling points) of the product were extracted. The main contents of the tour route and the services provided have been fully reflected in this part of the information.

Moreover, the TPKG was also enriched by collecting entity data from Baidu Encyclopedia which was extracted based on Ctrip. Figure 2 shows an entry introducing the information about *察里津诺庄园*(Charizino Manor), where we only extracted the location information about *地理位置*(geographic location) and the content about description since TPKG was mainly for indicating product features. In this case, the entity *察里津诺庄园* (Charizino Manor) extracted before would have two attributes and attribute values: <*察里津诺庄园*(Charizino Manor), location, *莫斯科南部*(Southern Moscow)> and <*察里津诺庄园*(Charizino Manor), description, *著名的宫廷建筑群，典型的哥特式风格建筑*(famous palace architecture and typical Gothic architecture)>. Due to the small number of entities requiring attribute value and the increased manual intervention when filtering attribute values, we employed a combination of manual collection and web crawlers to extract these attribute values.

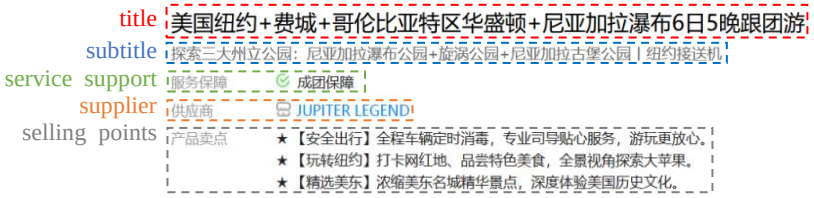

Figure 1. The page of a tourism product in the Ctrip website

Figure 2. The page of *察里津诺庄*(Charizino Manor) in Baidu Encyclopedia

2.2 Ontology construction

The concept of ontology was first proposed by scholars in the philosophical field, whose purpose was to describe the objective existences in the world systematically. In recent decades, it has been applied to the computer domain and played an increasingly significant role in the fields of artificial intelligence, computer language, etc. The definition of ontology given by Gruber of Stanford University is widely accepted, expounded as "An ontology is an explicit specification of a conceptualization"[5]. The knowledge graph can be seen as a knowledge base for a specific domain, so the construction of domain ontology usually requires the cooperation of experts from related research and the field of ontology construction. We asked experts in related fields to complete our ontology construction and referred to the dimensional classification proposed in previous studies. In 1991, six elements of tourism experience, including food, accommodation, touring, shopping, and entertainment, were put forward by *Strategic Research Report on China's Tourism Economic Development*, headed by Sun Shangqing. From tourists' perspective, Wu Heng et al.[6] classified the content of travel notes into five categories: superior tourism landscape resources, rich tourism entertainments, perfect tourism service and facilities, specialty food and high openness degree. Table 1 summarizes the knowledge classification of previous studies on the ontology construction process in tourism. We drew on the previous studies and combined the contents mentioned in the text data of tourism products. Considering to reflect the relationship between each product and their overall characteristics, this paper divides the knowledge into seven categories: destinations, international transport, tourist attractions, specialty food, accommodation, tourism services and facilities, tourism entertainments and shopping. Destination and international transport are the basic attributes of the product, which means that these attributes cannot reflect the characteristics and innovation of the product. On the contrary, the five remaining categories can measure the product's characteristics and innovation. Hence they were classified as the core attributes. The ontology structure of tourism products is shown in Figure 3.

Table 1. Historical Research on Ontology Construction in Tourism Field

Ontology	knowledge category
Tourism domain[7]	Accommodation, activity, destination
E-Tourism[8]	Activity, tourist attraction, destination, service Provider, Event, Service, Site
Tourism domain[9]	Figure, transportation, entertainment, literature and art, scenic spots, folk customs, tourism objectives, service agencies, tourist routes, geographical location, accommodation
Tourism domain[3]	Location, province, city, attractions, restaurants, hotels, cuisine

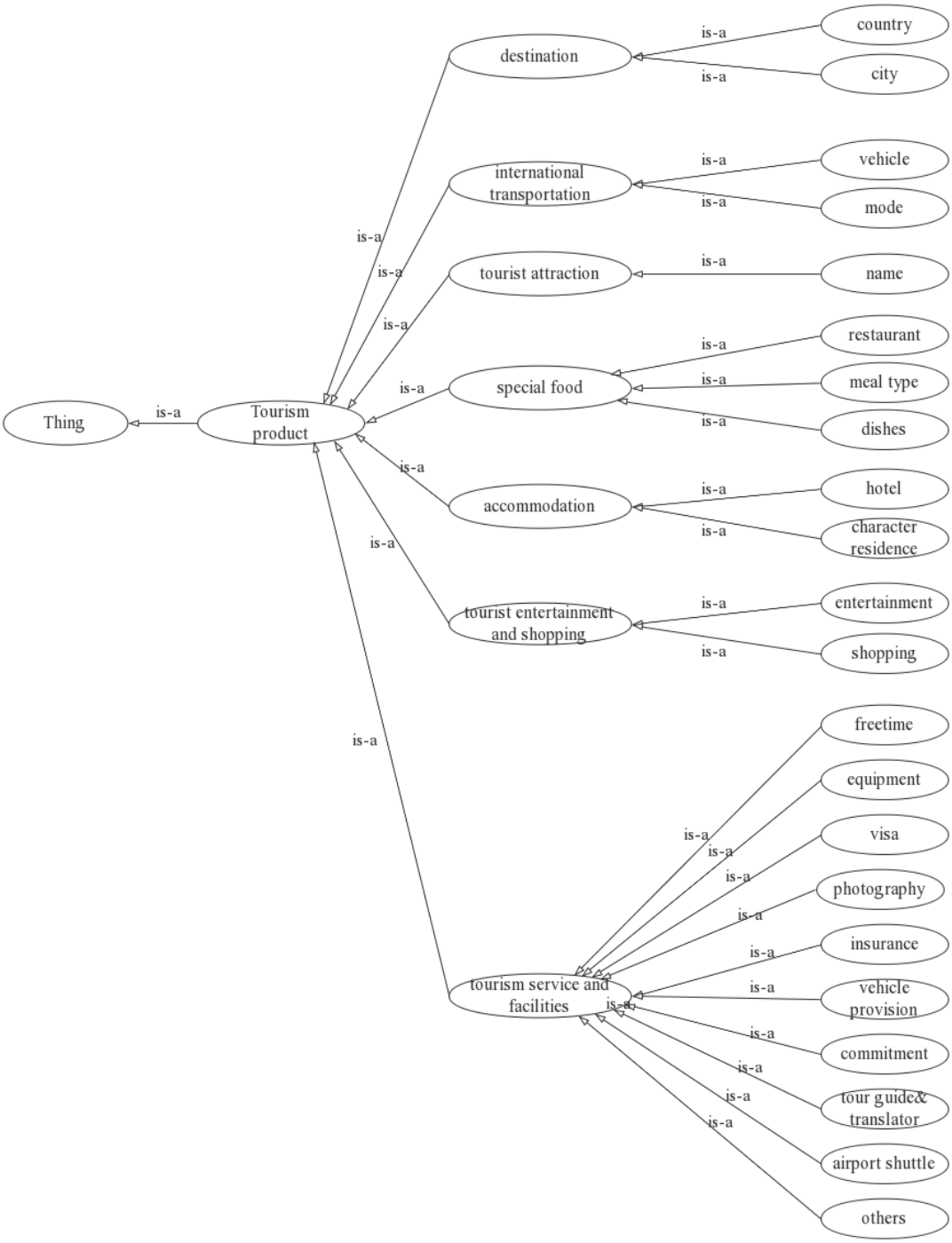

Figure 3. Ontology structure of tourism products

3. KNOWLEDGE GRAPH CONSTRUCTION

3.1 Relationship definition

A fact is represented by a triplet in knowledge graph. Relation extraction aims to construct the triplet of two entities and their relation. In the tourism product, the product is the central entity, only relationships between products and other types of entities are needed. Therefore, 8 relations were defined in this paper:

(1) HasProduct: the relationship between the supplier and the product, in which the supplier provides the product to consumer.

(2) StayAt: the relationship between the product and the hotel or specialty residence, in which the passenger temporarily resides in this hotel or characteristic residence.

(3) CityInclude: the relationship between the product and the city, where the product involves the tour of this city.

(4) CountryInclude: the relationship between the product and the country, in which the country is the destination of the product.

(5) ServicesInclude: the relationship between a product and a service, in which the product will provide the certain service.

(6) TripModeIs: the relationship between the product and the mode of international transportation, in which consumers will arrive in destination by this mode of transportation.

(7) TourIn: the relationship between the product and the attraction, in which the product contains the tour of this attractions.

(8) EntertainmentInclude: the relationship between the product and shopping or entertainment, which means that the product involves some activities or items related to entertainment or shopping.

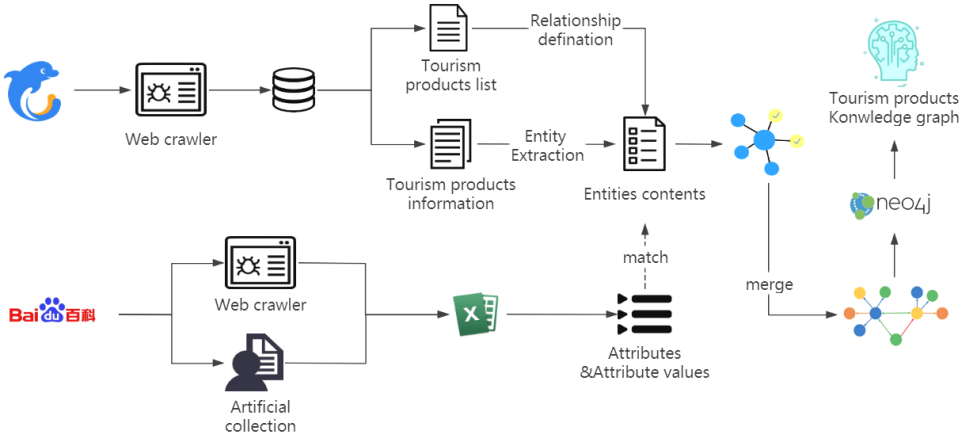

Figure 4. The framework of TP KG construction

3.2 Entity extraction and attribute value extraction

The named entity recognition (NER) was first proposed in MUC-6[10] and aims to identify the entities from text such as organizations, names, places, etc. Training data should be prepared to extract entities of tourism products. The BIO[11] schema is a standard annotation method for NER and was used to label the entities of training data in this paper. We used seven dimensions of tourism products proposed in section 2.2 as our data annotation categories. Due to the different amounts of product data every month, we randomly selected about 8% of the monthly data and got 13,385 pieces for training. To prevent the effect of subjective factors from the process of data annotation, we sampled the annotation results of the annotation personnel, in which 1,000 pieces of annotated data were selected randomly to evaluate the accuracy after the annotation. The accuracy rate of annotation was proved to be more than 90%. BiLSTM-CRF was one of the state-of-art methods for NER, combined with a bidirectional LSTM network and a CRF network[12], utilizing the past input features and sentence-level tag information future input features. BiLSTM has a good performance in processing long-distance text information but is useless in the dependence between adjacent tags, while CRF can cover this shortcoming by obtaining an optimal prediction sequence through the relationship between adjacent tags. BiLSTM-CRF was used to extract entities in this paper. In addition, considering every piece of product data is discrete, we input each product data as a separate sentence for model training after deleting the special symbols.

F1 value was taken to be an evaluation index to assess our model, which was the harmonic mean of precision p and recall r, as shown in (1), (2), and (3). After training, each category of entities reached 60% in F1 value, while practically being more than 80% was significant to improve the annotation data in future work.

Moreover, extracted entities classified as the "tourism services and facilities" category exist in many entities with similar semantics. To reduce the number of entities and connect as same nodes as possible between product entities, entities in this category were converted into 0-1 variables. For example, for the sub-dimension of "free time", entities having similar semantics such as *2 日自由活动*(2 days free activity), *充足自由时间*(sufficient free activity time), etc. were replaced with *自由活动*(free time).

As shown in Figure.5, the text of the product with ID 5912959 contains the entity of *2 日自由活动*(2-day free activity). We replaced it with *自由活动*(free time), indicating that the product includes the service content of free time.

$$p = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$r = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{F} = \frac{2}{\frac{1}{p}+\frac{1}{r}} \tag{3}$$

The extraction of attribute values was mainly for the entities we have extracted, which could enrich the TPKG. Although the product text data have illustrated the product's relevant information, attributes should be added to sufficiently delineate more details, particularly for attractions or accommodations. For example, for *察里津诺庄园*(Charizino Manor), we extracted *莫斯科南部*(Southern Moscow) and *著名的宫廷建筑群，典型的哥特式风格建筑*(famous palace complex, typical Gothic architecture) as its attribute value of the location and description respectively. As shown in Figure 4, we produced a list of them based on the original data and applied a combination of web crawlers and artificial collection to sort out all attributes and attribute values to match this list.

3.3 Knowledge merging and entity linking

Given the task of entity alignment based on the entity extraction results of tourism products, similarity and intimacy could achieve entity disambiguation by identifying diversity of the majority entities. Therefore, the cosine similarity and link prediction algorithms were introduced to distinguish the same entities with different names, in which the cosine similarity was for text vector comparison, used in information retrieval and text mining, the link prediction algorithm could better capture semantic dependencies on our task, as a graph data based on the closeness measurement of common neighbors between nodes mining algorithm. The larger the nodes (or entities), the higher the intimacy value. To enhance the entity alignment results, we set thresholds for similarity and intimacy between entities to evaluate the performance of algorithms.

The calculation method of similarity and intimacy are shown as equation (4) and (5), where two products a and b are regarded as two vectors in the m-dimension product space (or the k-dimension space in case of reduced representation). Additionally, $N(u)$ is the set of nodes adjacent to node u.

$$cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|^2 * \|\vec{b}\|^2} \tag{4}$$

$$A(x,y) = \sum u \in N(x) \cap N(y) \frac{1}{log^{|N(u)|}} \tag{5}$$

3.4 TPKG visualization

The knowledge graph is presented in the form of a ternary. All the ternaries were extracted from the merged data, including <entity, relationship, entity> and <entity, attribute, attribute value>. We used the Neo4j graph

database for knowledge representation, displayed by an example of a product, as shown in Figure 5. It suggests that in addition to extracting and displaying the original product data information from the Ctrip, we have also added relevant attribute values extracted from Baidu Encyclopedia to the entities *黄石公园*(Yellowstone National Park) and *威基基海滩*(Waikiki Beach) .

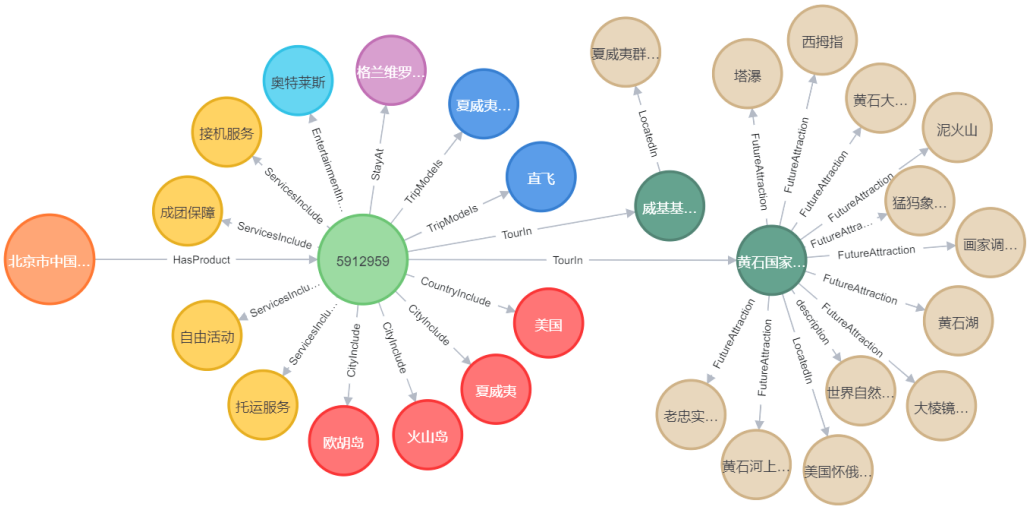

Figure 5. This figure is part of the TPKG we built, using the product whose ID is "5912959" as an example.

4. CONCLUSION

In previous research, knowledge graphs have always been established for the whole tourism field as a knowledge base to optimize the recommendation system or a question-answering platform. There hasn't established knowledge graphs for tourism products in China before. In this study the web crawlers were compiled to crawl the large-scale outbound travel data from the Ctrip and preprocess the data first. Then, ontology construction rules for tourism products were presented based on data and research requirements, including ontology structure and relationship definitions, covering food, accommodation, touring, shopping, entertainment, etc. In knowledge extraction, BiLSTM-CRF was employed for entity extraction, in which we used web crawlers to crawl semi-structured and unstructured data for entities related to attractions in Baidu Encyclopedia, expanding and improving the knowledge graph by combining with manual extraction of attribute values. Finally, we employed implemented entity linking and knowledge merging via the link prediction algorithm and utilized the neo4j graph database to visualize the ternary we have built, composed of <entity, relationship, entity> and <entity, attribute, attribute value>. The constructed knowledge graph provided a new possible relation of imitation and competition among tourism products from a nuanced particle perspective.

5. FUTURE WORK

This paper constructs a TPKG based on some outbound tourism product data on Ctrip. As the basis for related research in tourism products, further research will focus on two aspects in the future. First, based on a preliminary attempt to construct a TPKG, a more elaborated knowledge graph construction framework can be established in the future, including the named entity recognition and expansion of corresponding attribute values and domain knowledge base, etc. Second, more analysis methods can be employed to mine the relationship between products, such as imitation and competition between products. Moreover, in the entire product field, we can explore and summarize the evolutionary process of products and identify the degree of product innovativeness and the market response by combining the product with the market.

ACKNOWLEDGEMENT

This study was supported by the National Training Programs of Innovation and Entrepreneurship for Undergraduates [Grant Numbers 202110491015].

REFERENCES

[1] Singhal A. (2012).Official Google Blog: Introducing the Knowledge Graph: things, not strings.

[2] XU Fu. (2016).Research and Implementation on Construction Method of Knowledge Graph in Tourism Domain.MA Thesis.Beijing: Beijing Institute of Technology(in Chinese)

[3] LIU Jiyuan. (2019).Research on the Construction and Application of Knowledge Graph in Tourism Domain.MA Thesis.Hangzhou:Zhejiang University(in Chinese)

[4] Smith S. (1994).The tourism product. Annals of tourism research, 21(3):582-595.

[5] Gruber T R. (1993).A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199-220.

[6] WU Heng, CHEN Yan-ling. (2017). Study on Information of Tourists'Destination Selections Based on UGC and Text Mining—Taking Honeymoon Travel Notes From Ctrip as an Example.Information Science, 35(1):5.(in Chinese)

[7] Knublauch H. (2004).Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with. annex xvii and.

[8] Mili H. et al. (2011) E-Tourism Portal: A Case Study in Ontology-Driven Development. In: Babin G., Stanoevska-Slabeva K., Kropf P. (eds) E-Technologies: Transformation in a Connected World.MCETECH 2011. Lecture Notes in Business Information Processing, vol 78. Springer, Berlin, Heidelberg,76-79.

[9] LI Qingsai. (2015).Research on the Construction of Tourism Domain Ontology.MA Thesis.Zhengzhou:Zhengzhou University(in Chinese)

[10] Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

[11] Kim T. (1999).Representing Text Chunks. proc of eacl.

[12] Huang Z , Wei X , Kai Y. (2015).Bidirectional LSTM-CRF Models for Sequence Tagging. Computer Science.