Overview of cross-language retrieval technology based on knowledge graph

Shengyin Zhu [1,a], Xiangzhen He [1,b,*] and Honzhi Yu [1,c]
[1] Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University Lanzhou, Gansu 730000, China
[a]87961192@qq.com, [b]5967148 @qq.com ,[c]yuhongzhi@hotmail.com
*corresponding author

***Abstract*—The "sootc" e-commerce platform of One Belt and One Road provides users with a clearer, more intuitive and comprehensive information retrieval after the introduction of knowledge graph. Besides, it not only implements semantic retrieval, but also provides a better retrieval experience in terms of user personalized recommendation and scene-based real-time information service. With the ideas and methods of multi-language knowledge graph, the authors have carried out experiments and analyze the construction of semantic search functions of Chinese, Tibetan, Mongolian, English and other cross-language systems. Through the analysis of data, we take regions as the mapping entity, and realize the structure of "entity--- relationship--- entity". Each language is an entity, and entities are corresponding. Each language entity takes the triple of "entity ---attribute--- value" as the basic expression of fact. Then all stored data constitute a huge multi-language entity relationship network of commodity information, forming a multi-language knowledge graph of commodity information.**

Keywords—knowledge graph, cross-language retrieval, multi-language

1. INTRODUCTION

The knowledge graph, based on the knowledge base of relational data, by labeling the data, identifying the relationship and constructing the underlying knowledge structure network, strives to organize the complex and huge knowledge in a systematic and orderly way. The knowledge graph is of irreplaceable importance in the era of big data.

In order to make users find new commodity information in other languages more quickly and more conveniently, the commodity names of various languages are constructed to be entity relationship and specific information is additional information of entity. What's more, entities are classified and entity relationships are analyzed, so it makes it possible that as long as users input information of one language in "sootc" website, they can retrieve all language commodity description information supported by the system, which effectively eliminates obstacles among different languages, and provides convenience for users to retrieves multilingual commodity information resources. Although cross-language retrieval can let users get more and more complete information, it is just the keyword retrieval rather than the user intention retrieve. In the current environment of intelligent retrieval, such results will inevitably affect the user experience. However, the research of cross-language retrieval system based on knowledge graph enables the search to have a better understanding of the users' needs, and can provide users with more intelligent, accurate and human-oriented results. How do we construct knowledge graph based on cross language retrieval system? This paper describes an experiment and analyzes on the construction of semantic retrieval function of cross- language systems such as Chinese, Tibetan, Mongolian, Uygur and English, which provides barrier free retrieval for different language users. After the introduction of knowledge graph to the "sootc" e-commerce platform, it provides clearer, more intuitive and comprehensive information retrieval to users. Therefore, in the face of the disadvantages of traditional information retrieval, cross language semantic retrieval method has become the focus of our research to meet the high-level retrieval needs of different language users

II. RESEARCH STATUS AT HOME AND ABROAD

A.Cross-language retrieval

Cross language retrieval was proposed by Prof. G. Salton in 1973. Its main function is to retrieve documents related to the source language from the document set represented by the target language. When users input a certain language, it can retrieve other language information supported by the system, which can eliminate the barriers between different languages and become the link for users to find all language resources and communication.

B. The knowledge graph

The knowledge graph was first proposed by Google in 2012, which is mainly used to optimize the query and retrieval technology of its search engine, and reason the related knowledge. The first large-scale cross-language knowledge graph "Xlore" in China was first proposed by Tsinghua University. Academician Lu Ruqian of the Chinese Academy of Sciences put forward the concept of "zhijian". Professor Wang Fanjin of Shanghai Jiaotong University has studied and constructed the most popular knowledge graph of Chinese at present. The characteristics of the research mentioned above are rich application knowledge and wide fields, and it can provide users with certain knowledge search and Q & A services. However, the research on how to construct knowledge in the field of multilingual e-commerce through technology is still lacking.

III. DATA SOURCES AND RESEARCH METHODS

A. Data source

The data of this experiment comes from the plan supported by national science and technology Integration and Application Demonstration of Key Technologies of Multilingual Online Trading Platform for Agricultural Products with National Characteristics. With the application

of multilingual information processing technology, the project sorts out the characteristic agricultural product resources along one belt and one road area, studies multi language retrieval technology and builds multi-language knowledge graph management system, which provides a bridge of information communication for agricultural products and customers in all regions along the economic belt and promotes cross- language and trans-regional trades.

B. Research methods

The research method focuses on solving the problems, such as language recognition, vocabulary segmentation, construction of knowledge graph entities, sorting of retrieval results. By constructing management database of knowledge graph and analyzing knowledge graph, identity of named entity, knowledge fusion, entity link, the connectivity of knowledge graph entity relations and other technologies are achieved. And then by using the method of query translation, the multi-language is mapped to the query words of another language, so that the query words and the retrieved documents are in the same feature space, and the user search is transformed from a single keyword search to a multi-entity relationship composite search. Eventually, the selective results of multiple languages are presented to provide users with multi-dimensional, all-round multilingual information retrieval function for characteristic agricultural products.

IV. CONSTRUCTION OF CROSS-LANGUAGE RETRIEVAL SYSTEM BASED ON KNOWLEDGE GRAPH

Cross-language retrieval based on knowledge graph is to optimize the results of traditional retrieval. The retrieval system should have a complete and systematic knowledge system, the knowledge of which has multi-dimensional and cross semantic relationship. In the process of searching target words or sentences, users can not only effectively gain accurate retrieval results, but also find some relevant goods and service information according to the semantic relationship, obtaining more humanized recommendation retrieval service [Figure 1].

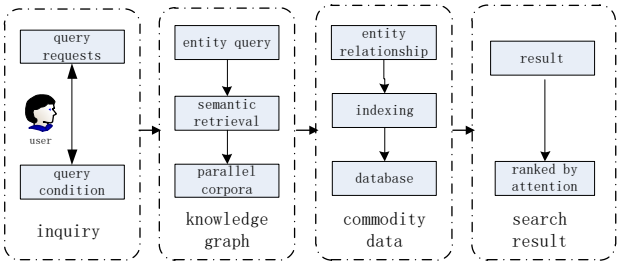

Figure 1 knowledge graph structure of cross language retrieval

A. Parallel corpus

In the research about cross-language information retrieval, in order to facilitate the training of parallel corpus such as machine translation technology and statistical methods and extract the translation relationship between multi-language words or phrases, we will establish a multi-language parallel corpus with paragraph level alignment [Figure 2]. In cross -language information retrieval, corpus is a very important basic data resource, so we need to build a retrieval framework based on multi-language cross-language information. The framework can extract the semantic representation of the same semantic object from the multi-language parallel corpus, construct linear or non-linear representation of multi-language correspondence and carry out cross-language retrieval, cross- language text classification and cross- language text clustering.

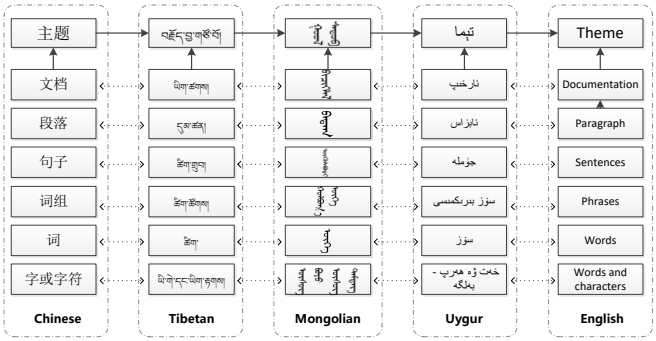

Figure 2 parallel corpus correspondence of Chinese, Tibetan, Mongolian, Uyghur and English

B. Multilingual text clustering

Cross-language text clustering is the process of dividing multi-language documents into a specific number of clusters. The cross language text clustering in this study is actually multilingual document, so we call it multilingual document clustering, and the documents expressed in different languages are similar. The significant difference between it and single language text clustering is that the clustering objects and documents in production cluster are all multilingual. Text clustering is mainly to solve the operation of cross-language semantics. Because documents expressed in different languages are mapped in different language spaces, and most single language text clustering methods cannot effectively establish cross-language semantic relationships. The main method of text clustering is translation method, and the method based on translation mainly translates words or documents, but it has strict requirements on the quality of translation.

C. Acquisition of knowledge graph

In the design of knowledge graph ontology, we describe the product itself in multi-language, and get the language needs of consumers through ID. Another core work is to collect data of relevant information. We have made a detailed description of multilingual titles, pictures, details, multilingual evaluation management platform, model, purchase point, scene and other information. This requires that the named knowledge recognition system has the recognition ability across large-scale entity types, supports the basic data of e-commerce, the natural language problem of human-computer language interaction and links the identified entities to the knowledge graph.

D. Knowledge fusion of multilingual trading platform

The knowledge fusion of knowledge graph of multi-language trading platform mainly involves the knowledge fusion of different commodities, large-scale clustering, large-scale entity chain index and large-scale level classification technology. Level classification needs to classify the target commodities into thousands of commodity level 1 and level 2. The difficulty lies in the subdivision and confusion of categories, as well as the generation and disambiguation of training data.

The purpose of large-scale clustering is to fuse the information of unified data source first. The core of large-

scale entity chain is to sort the candidate entities of knowledge graph and associate the new entities with the target recognition of knowledge graph, thus new knowledge is integrated into the knowledge graph. The integration of new knowledge into engineering involves the mapping and standardization of different data attribute names and attribute values, which requires the construction and excavation of large-scale e-commerce vocabulary.

E. Application of knowledge graph

Knowledge graph is a structured semantic knowledge base, which is used to describe concepts and their relationships in the physical world in the form of symbols. Its basic unit is "entity--- relationship--- entity" triplet, and entity and its related attribute value. Entities are connected by relationship to form a network of knowledge structure. Through knowledge graph, web can be transformed from web link to concept link, and support users to search by subject instead of string, to realize semantic retrieval. Our search engine can feedback structured knowledge to users in a graphical way, so users can accurately locate and acquire knowledge in depth without browsing a large number of web pages.

The knowledge graph of multi-language transaction is a general shopping guide for a commodity display library in different languages. The so-called shopping guide is to make it easier for consumers to find what they want. If the buyer inputs "special products in high altitude areas", the system will extract the semantic key points "specialty", "Qinghai", "Tibet" and other key words through grammatical word analysis so as to help the buyer search for the right goods. In order to make the search easier, the system also learned a lot of industry norms and national standards, such as organic, low sugar and natural. In addition, it has the advantage of keeping pace with the times. The system can also identify the recent hot words from the information of public media and professional communities, track the changes of hot words, and confirm whether they are hot words by operation. This is also why when the buyer is inputting "travel", he or she can find words like "beef jerky, coarse grain biscuits, nuts, etc.", which are high calorie foods with strong sense of satiety, and meet the requirements of travel food.

TABLE 1 2018-2019 HIGH-FREQUENCY KEY WORDS SEARCHED IN THE "SOOTC" DATABASE

serial number	high-frequency words	key words
1	1558	甘肃特产（Gansu specialty）
2	1513	牦牛肉干（Yak dried beef）
3	1477	新疆特产（Xinjiang specialty）
4	1456	兰州百合（Lanzhou lily）
5	1416	鲜牛肉（fresh beef）
6	1398	青海黑枸杞 （Qinghai black goji berry）
7	960	宁夏羊肉（Ningxia mutton）
8	759	枸杞（goji berry）

TABLE 2 2018-2019 HIGH-FREQUENCY MULTILINGUAL SEMANTIC RETRIEVAL IN "SOOTC" DATABASE

serial number	high-frequency words	key words
1	691	ལན་ཙོའུ་ཡི་ལྗགས་འབྲུང། （Lanzhou lily）
2	663	རྨ། （tea）
3	512	گازىر（melon seeds）
4	510	ཞུན་འབྲས་སྐམ་ངོ་（raisin）
5	497	（Gansu specialty）
6	495`	بيڭى كالا گوشي（fresh beef）
7	412	（Yak dried beef）
8	294	nut free shipping

[Table 1] shows "special products" platform mainly searches by region and category, while the frequency of a specific product search is less. It can be seen that consumers don't know much about regional specialties but have strong desire to find new featured products. Therefore, when making semantic recommendation, it is necessary to be targeted and recommend products with regional representation. [Table 2]represents the search frequency of four other languages except Chinese. It can be seen that the search frequency of users of other languages is mainly based on agricultural products with regional characteristics.

Through the analysis of data, we take regions as the mapping entity, and realize the structure of "entity--- relationship--- entity". Each language is an entity, and entities are corresponding. Each language entity takes the triple of "entity ---attribute--- value" as the basic expression of fact. Then all stored data constitute a huge multi-language entity relationship network of commodity information, forming a multi-language knowledge graph of commodity information. For example, "Lanzhou Lily" in Tibetan, Mongolian, Uygur, Chinese and other languages is taken as an entity to establish the entity relationship between the same entity and different languages. Users can easily retrieve different language entities of "Lanzhou Lily" through entity association when using a language retrieval. At this time, a ternary expression of "Lanzhou Lily" and "entity ---attribute --- value" in Chinese is established. When users use Chinese "Lanzhou Lily" as the key word for retrieval, they take Chinese as the initial relationship entity and get information in different languages through the relationship network.

V. CONCLUSION

This research is an attempt of knowledge graph in the application of multi-language e-commerce platform retrieval, and has acquired some research results. The idea, process and method of this study lay a theoretical foundation for the subsequent large-scale commodity retrieval, and finally strive to play a greater role in the research of cross- language retrieval system based on knowledge graph.

In this paper, Chinese, Tibetan, Uyghur, Mongolian and English are the research objects. In the current development trend of semantic retrieval, in order to achieve cross-language semantic retrieval, a cross- language retrieval system based on knowledge graph is proposed. Under the background of "One Belt and One Road ", this study is not only conducive to eliminating the unfavorable factors in the stability and rapid development of ethnic areas, but also beneficial to the realization of "multi integration" architecture of the multi-ethnic language information service field in the digital era. Meanwhile, it provides some basic theoretical research for the national strategic arrangement of digital information management involving ethnic languages.

REFERENCES

[1] Mai Shuping, an analysis of cross language information retrieval technology [J]. Chinese Journal of Medical Library and information. 2008

[2] Sun Xiaoxin. Construction of subject knowledge map based on latent semantic analysis [D]. Central China Normal University, 2013.

[3] Singhal A. introducing the knowledge graph: things, not strings [J]. Official Google Blog, may, 2012.

[4] Zhang Jing, Tang Jie. Focus of next generation search engine: knowledge map [J]. Communication of China computer society, 2013, 9 (4).

[5] Carlson a, betteridge J, kisiel B, et al. Toward an architecture for never ending language learning [C] / / AAAI. 2010, 5:3.

[6] Liu Yanlu, Yu Yong. Knowledge representation framework for Semantic Web [J]. Journal of Shanghai Jiaotong University, 2002, 36 (9): 1309-1311.

[7] Liang Xiujuan. A review of the study on of scientific knowledge graph [J], Library Journal, 2009,06:58-62