Advances in Engineering Research, volume 119
2nd International Conference on Automatic Control and Information Engineering (ICACIE 2017)

Research on Tibetan-Chinese Cross-language Information Retrieval System Base on E-commerce Platform

Fucheng Wan
Key Laboratory of National Language Intelligent Processing
Northwest University for Nationalities
Lanzhou Gansu, China
306261663@qq.com

Lin Zhu
Department of Computer Science
Guangdong University of Science & Technology
Dongguan Guangdong, China
Loda0622@163.com

***Abstract*—Based on the research of the information retrieval system of Tibetan and Chinese, this paper uses the bilingual dictionary and query extension method to extend the semantic aspect of the query word, and establishes the document Index model. The test system is built on the basis of this model, and the results of recall and precision are achieved.**

Keywords—e-commerce, information retrieval, Tibetan, cross-language

I. INTRODUCTION

Cross-language information retrieval (cross-language Information retrieval, CLIR) is the process by which users construct questioning query information in a language (called the source language) to retrieve a document set that conforms to the user's requirements in another language (called the target language) [1].

In layman's terms, cross-language information retrieval allows the language used to query the keyword to be different from the language used to retrieve the target document, and any documents related to the information to be retrieved by the query keyword are retrieved whether the language they are using is consistent with the query language. With the rapid growth of network coverage and the rapid development of society, online shopping has become one of the main life links of modern people, E-commerce has also enter into a rapid development. In the emerging situation of this new technology and mode, the electric commerce has been developing in the direction of diversification. Because of the limited language and the rich diversity of information, it is not only the need of the society but also the market.

Of the research of cross language information retrieval, the core content is to solve a series of problems brought by the different language of querying and retrieving documents. Therefore, first of all, we should establish a corresponding translation relationship between the query type and the retrieval document, map the words of one language to another language to make them in the same language characteristic space, so that the information retrieval between different languages can be converted to single language information retrieval. Then these problems could be solved by using the related technology of single language information retrieval.

Based on the research of translation technology, query extension method, cross language information retrieval model, document indexing, and various technical methods, this paper constructs a simple feasible information retrieval system of Tibetan and Chinese language.

II. EASE OF USE

A. Query translation based on bilingual dictionaries

Translation is the channel through which the language for the query is connected with the language for the document and can communicate with it. Before the translation, the query type and the document are preceded by a series of preprocessing, such as word segmentation and entry. There are many translation methods, such as based on knowledge translation, machine translation technology, dictionary translation, corpus-based translation methods and so on [2]. These translation methods have their own advantages and disadvantages. Based on the characteristics of e-commerce, this paper mainly deals with the research of cross language information retrieval in the field of e-commerce platform, which is obtained by using the Tibetan-Chinese bilingual Dictionary for entry-control translation, and dictionaries of Tibetan-Chinese dictionary (upper, middle and lower), Amdo Oral dictionary, Tibetan homonym Dictionary, and Western In addition, some information related to E-commerce and some new words about 120,000 Tibetan and Chinese translation entries. Examples of dictionaries are shown in table 1 below:

TABLE I. EXAMPLES OF TIBETAN-CHINESE TRANSLATION DICTIONARIES

Tibetan	Chinese
ལྷག་མཉེ་སྐམ་པོ།	Dry Lily
ལན་ཐུའི་ལྷག་མཉེ་སོ་མ།	Lanzhou Fresh Lily
ཐུ་ཞའི་ཏན་འབྲུམ་སྐམ་པོ།	Turpan raisins
ཁག་ཚ་ལང་ཏན་འབྲུམ་སྐམ་པོ།	Black Currant raisins
ཤིན་སྟལ་ཀུན་འབྲུམ་སྐམ་པོ།	Fragrant Imperial raisins
འཁང་འབས།	Lycium
འཁང་འབས་ནག་པོ།	Black Wolfberry
སྨད་ཤ་སྐམ་པོ།	Beef Jerky
སྨད་ཤ་རྡོག་པོ།	Beef Grain
མཚོ་སྔོན་ཕྱི་ཤོར་ཤ་སྐམ་པོ།	Qinghai Beef Jerky
ནོར་ཤ་སྐམ་པོ།	Yak Jerky
ཀུ་ཤུ།	Apple
ཞི་ཁ་ཀུ་ཤུ།	Qixia Apple
ལེཁགས་ཡུལ་ཀུ་ཤུ།	Le wild Apple
ཆནྡ་ཡང་ཀུ་ཤུ།	Qingyang Apple
ཀླུ་ཁྲོན་ཀུ་ཤུ།	LuoChuan Apple
ཞི་ཡུལ་དཔལ་རིལ།	Xinjiang JuJube
ཧོ་ཐེན་དཔལ་རིལ།	Hetian JuJube
མཐོ་སྐང་དཔལ་རིལ།	Plateau JuJube
ཚོང་ཁའི་རིན་གོང་།	Market price
ཚོང་ཁའི་རིན་གོང་།	Mall Price
མ་གཞིའི་རིན་གོང་།	Original price
ཚོང་རྟགས་དཔྱད་བརྗོད།	Product Rating
ཚོང་རྟགས་དཔྱད་བརྗོད།	Commodity evaluation
དཔྱད་བརྗོད་ཀྱངས།	Article comments
རྒྱ་ཤེས་ཚང་ཁ་ཤུགས་འགོད།	Scan two-dimensional code
ཁ་པར་ཕྱི་ཉམ་རས་ཚོ།	Shopping on the phone
སྒུལ་འཚོང་།	Distribution
དུས་ཚད་རིན་གཅོག	Limited discount
དངོ་པད།	Straight

B. Query extension processing

The query extension is optimized based on the original query retrieval of user input. Using computer linguistics, information science and other technologies, the function is to add or find a similar query, to form a more consistent user's query intention of the new query sequence [3] through a number of methods and strategies to the original query search words related to the combination of words, word items. For cross-language information retrieval has to undergo a translation process, it is necessary to determine the degree of matching between the translated retrieval words and the document, rather than just the matching of the words and keywords in the document, so the application of this technique in the cross language information retrieval is more important than the single language information retrieval.

The technology of query extension is divided into two kinds. One of them is global analysis method, without consideration of query and the document returning, in which case the initial query retrieval is extended and reconstructed. The other is the local analysis method, which is to modify the initial query retrieval by the initial matching document of the query. In this paper, the pseudo correlation feedback method is used to solve the query extension in Cross-language information retrieval. Since the process of translation must be experienced in cross language information retrieval, the query extension method is used before translation process and after it, as shown in Fig. 1. In this way, the conceptual information and semantic information of query retrieval are further strengthened.

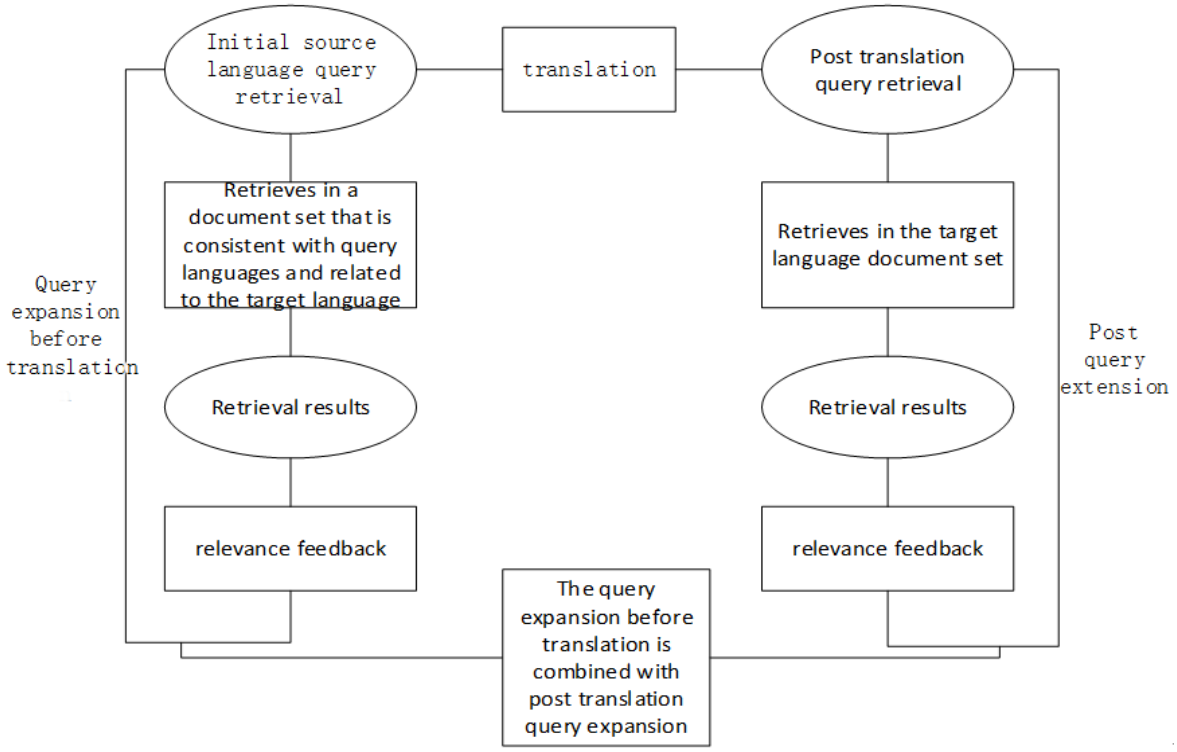

Fig. 1. Diagram of Query extension method

III. Document Index

Index is an effective means to speed up the query process, which directly affects the time users wait for the results of the search. When the document is processed by word segmentation and verb normalization, the index of the retrieved document is stored to prepare for the next library retrieval.

A. Weight Mapping

In some dictionary-based query translation, the user-input query keyword is processed as a retrieval key. Each keyword is regarded as a separate vector in the vector space model, and then each keyword was weight assigned, so as to calculate the weight value of each document. With the cosine angle formula between two vectors used to calculate the similarity between the query type and document, the document scoring algorithm is designed to sort the document.

B. Index Build

Through each keyword item in the vector space, the inverted index is mapped by mapping different text points to the same key item [4]. An inverted index in Tibetan text is shown in Fig. 2:

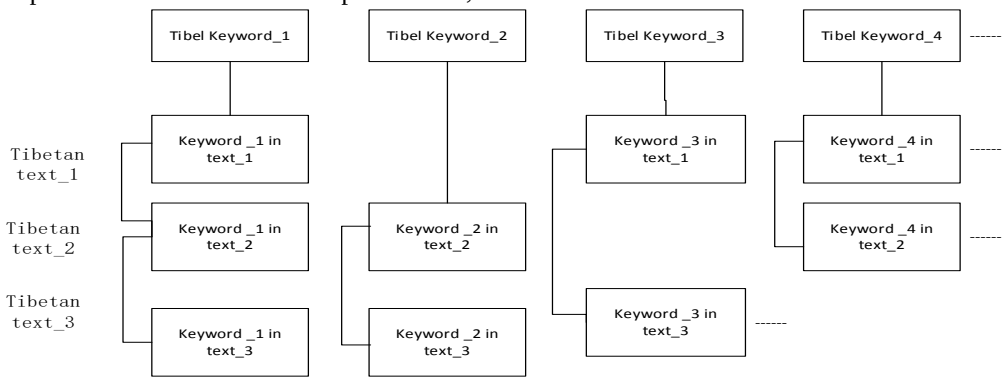

Fig. 2. Inverted index schematic

Then by the hash table (Hashtable) the keyword items and text are mapped to the data store, and inverted index is established with key items in the text, in which value is the text linked list that contains this keyword item. If the word ID is the word item ID and also the value key for the index entry, when it enters the system, a list that identifies the document as DocId would be gotten, the contents of which are the information that appears in each document.

IV. System and Experiment Design

A. System Design

Based on the theoretical research of the Tibetan and Chinese information Retrieval system, this paper designs and realizes the Tibetan-Chinese Cross language information retrieval system for the agricultural products in the electronic commerce platform. The development technology languages used are: Jsp+javabean+servelet+jdbc+mysql. First step is to establish a connection to the database. With JavaBean platform and Wxbean's management, the platform and any function established for a connection in it can set and get a function of the values of each property and constructs a function that retrieves an expression based on a field and a search term. The user's retrieval request is then obtained. Second, retrieves the records that meet the requirements from the database, establishes a view based on each index table in the database, whose field is the set of the fields in the established Index table, and then retrieves the view. Finally it would display a list of records and each record, with a loop to display each record (field) in one page, and to provide a link in a field, clicking to display the specific content of the record.

B. Experimental Design

1) Retrieval evaluation

The most commonly used two evaluation criteria for evaluating the performance and effectiveness of a retrieval system are: accuracy P (precision, also known as the precision ratio) and recall rate R (Recall, also known as recall ratio), as shown in the following formula:

$$P = \frac{\text{returns the number of documents related to the result}}{\text{the total number of documents returned}}$$

$$R = \frac{\text{Returns the number of related documents in the result}}{\text{total number of related documents}}$$

2) Experimental content design

In order to verify the function validity of the Tibetan-Chinese Cross-language Information Retrieval system, the experiment content is designed as follows by setting up reference experiment:

1 Single Language Information Retrieval: Chinese query retrieval type, Tibetan-Chinese mixed Document Set (abbreviation: Chinese-Tibetan), evaluation of search results;

2 Cross-language information retrieval: Tibetan Query retrieval type, Tibetan and Chinese mixed Document set (abbreviation: Tibetan-Tibetan Han), evaluation of search results;

3 Single Language information retrieval: Tibetan Query retrieval type, Tibetan and Chinese mixed Document set (abbreviation: Tibetan-Tibetan Han), evaluation of search results;

4 Cross-language information retrieval: Chinese query retrieval type, Tibetan-Chinese mixed Document Set (abbreviation: Chinese-Tibetan), evaluation of search results.

Among them, the test (1) and (3) is to carry on the single language information retrieval. When carrying on the single language information retrieval, whether the query retrieval type is Chinese or Tibetan, the result can only retrieve the document language as same as question. The experiment (1) is a reference experiment of Experiment (2), which examines the cross-language information retrieval of Tibetan query retrieval to Chinese document set. The Experiment (3) is a reference experiment of experiment (4), which investigates the cross language information retrieval of Chinese query retrieval to Tibetan document set. The cross-language information retrieval is a combination of the original query retrieval and the translated search after the query retrieval is translated, and the retrieval is done. At this time documents in two languages can be retrieved. The language barrier is broken, the search scope is enlarged, and the information quantity of retrieval results is increased. The experimental results of Tibetan-Chinese Cross language information retrieval are shown in table 2:

TABLE II. Experimental results of cross-language information retrieval in Tibetan and Chinese

Test serial Number	Test content	Recall R	Precision P
(1)	Chinese-Tibetan	0.6203	0.7358
(2)	Tibetan-Tibetan-Chinese	0.8102	0.7064
(3)	Tibetan-Tibetan-Chinese	0.5706	0.7128
(4)	Chinese-Tibetan	0.7801	0.7012

V. Conclusion

Based on the actual situation of Tibetan information processing, this paper studies the technology of Tibetan and Chinese Cross language retrieval. Combined with the characteristics of E-commerce platform, it translate the query by the dictionary based query translation method to the use of query expansion before translation and query expansion after translation query to query expansion, establish an index for the document, finally designed a simple and feasible information retrieval system of Tibetan and Chinese language. This paper realizes a simple Tibetan-Chinese cross-language information Retrieval system, makes full use of the resources that can be used in combination with the characteristics of E-commerce platform, but there are still some problems not solved. As there are some restrictions on the Tibetan language resource pool, the relevant techniques for the processing of Tibetan languages are still being studied

Acknowledgements

This work is supported by "the National Natural Scien ce Foundation of China(Grant: 61602387)", and "the Fund amental Research Funds for the Central Universities(31920 170155)"

REFERENCES

[1] Rahimi R, Shakery A, King I. Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework[J]. Information Processing & Management, 2016, 52(2):299-318.

[2] Shashirekha H L, Gashaw I. Dictionary Based Amharic-Arabic Cross Language Information Retrieval[C]//International Conference on Advances in Computer Science and Information Technology. 2016:49-60.

[3] Alonso M A, Doval Y, Vilares M. Studying the effect and treatment of misspelled queries in Cross-Language Information Retrieval[J]. Information Processing & Management, 2016, 52(4):646-657.

[4] Rupnik J, Leban G, Grobelnik M. News across languages - cross-lingual document similarity and event tracking[J]. Journal of Artificial Intelligence Research, 2015, 55(1):283-316.

[5] Nie J. Cross-Language Information Retrieval[J]. Synthesis Lectures on Human Language Technologies, 2003, 2(1):19-24.

[6] Pemawat V, Saund A, Agrawal A. Hindi - English based cross language Information Retrieval system for Allahabad Museum[C]// International Conference on Signal and Image Processing. IEEE, 2010:153-157.

[7] Ballesteros L, Croft W B. Phrasal translation and query expansion techniques for cross-language information retrieval[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1997:84-91.

[8] Sadat F, Yoshikawa M, Uemura S. Cross-Language Information Retrieval Using Multiple Resources and Combinations for Query Expansion[M]// Advances in Information Systems. Springer Berlin Heidelberg, 2002:114-122.

[9] Danielyan T, Zuev K, Anisimovich K, et al. Cross-language text classification[J]. 2017.

[10] Prettenhofer P, Stein B. Cross-language text classification using structural correspondence learning[C]// ACL 2010, Proceedings of the, Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. DBLP, 2010:1118-1127.

[11] Prettenhofer P, Stein B. Cross-Lingual Adaptation Using Structural Correspondence Learning[M]. ACM, 2011.