Research on Intelligent Retrieval Model of Multilingual Text Information in Corpus

Ri-han Wu[1(✉)] and Yi-jie Cao[2]

[1] School of Chinese Language and Literature, Northwest Minzu University, Lanzhou 730030, China
wurihan21322@yeah.net
[2] School of Ethnology and Sociology, Northwest Minzu University, Lanzhou 730030, China

Abstract. Cross language information retrieval focuses on how to use the query expressed in one language to search the information expressed in another language. One of the key problems is to adopt different methods to establish bilingual semantic correspondence. In recent years, topic model has become an effective method in machine learning, information retrieval and natural language processing. This paper systematically studies the cross language retrieval model, cross language text classification method and cross language text clustering method. Without the help of cross language resources such as machine translation and bilingual dictionaries, it can effectively solve the many to many problem of Vocabulary Translation in CLIR and the problem of partial decomposition of unknown words. The experimental results on the cross language text classification evaluation corpus established in this paper show that the performance of cross language and single language text classification on the bilingual topic space constructed by this method is close to or better than that of single language classification on the original feature space, and the performance of cross language text clustering is close to or better than that of single language document clustering.

Keywords: Corpus · Language · Information retrieval

1 Introduction

Language and text information processing technology is a technology that transforms the language processing used by humans in interactive communication into machine language that computers can understand. It is a model and algorithm framework that uses language ability as the research object. It involves linguistics and computers. Cross-cutting research areas of science. In the "Internet+" era, the innovation and breakthrough of this technology can not only promote the development of human-machine intelligence, bring about a revolution in computing technology, but also enable humans to further understand their own thinking and language, and pay more attention to language teaching and learning [1]. Information technology is rapidly penetrating into all levels of the economy and society and profoundly changing people's work and lifestyle. At the same time, it also brings new challenges to the improvement of the quality of information

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 416, pp. 26–40, 2022.
https://doi.org/10.1007/978-3-030-94551-0_3

technology application of the entire nation. The competition in modern society is not only manifested in politics, economy, military, etc., but also in informatization, that is, the competition between information system and the right to speak [2]. Information technology has brought new development opportunities to the development of language itself. As the carrier of information, language information construction will be put on the national agenda, and the development of information technology will in turn promote the development of linguistics itself [3]. The current period is not only an important period of strategic opportunities for my country's modernization drive, but also an important period of strategic opportunities for the construction of language informatization. It is necessary to enhance the awareness of opportunities and the overall situation, and make full use of contemporary information technology to carry out language work in the new era.

The experimental results on the cross-language text classification evaluation corpus established in this paper show that the cross-language and single-language text classification performance completed on the bilingual topic space constructed by this method is close to or better than the single-language classification of the original feature space.

2 An Intelligent Retrieval Model of Multilingual Text Information in Corpus

2.1 Corpus Multilingual Text Information Feature Collection

A key issue in information retrieval is a variety of semantic representations of a description object, which are expressed as polysemous words, synonyms and synonyms in language form [4]. Even if the words or phrases contained in the query appear in the document, they may indicate another meaning because of the different context. The rich semantic representation of natural language [5] increases the difficulty of the IR system to retrieve and query related documents. In cross language information retrieval or multi language information retrieval, queries and documents are expressed in different languages. In addition to the semantic combination of words or phrases in a single language, there are also cross language semantic combinations, which makes it more difficult to retrieve related documents [6]. The overall strategic structure of language information construction is shown in the following figure. The implementation of the overall strategic plan relies on the construction of the information chemistry department and the construction of standards. The application support platform is based on the construction of a resource library, and the service system is built on the application support platform. It is not appropriate to pursue informatization construction in one step. It is necessary to combine advancement and practicability, make overall planning, and implement step by step. Based on this, the corpus multilingual text information database model is constructed, as follows (Fig. 1):

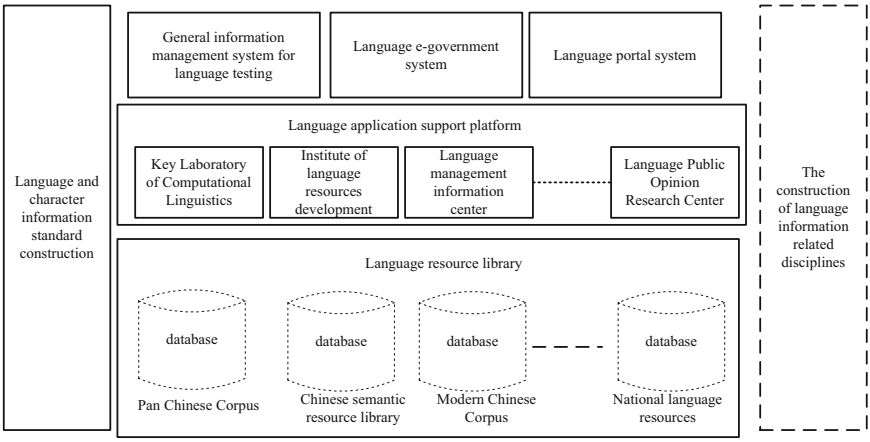

Fig. 1. Corpus multilingual text information database

The construction of language and text resource base is the foundation of information strategy research. The construction of resource base will change from word annotation database, tree database to semantic knowledge base; from text resource base to multimedia and multimodal resource base integrating voice, video and text [7]. It is not only necessary to establish a balanced resource database, but also to pay attention to the establishment of a multi-language simultaneity data database [8]. It is not only necessary to establish a domestic language resource monitoring corpus, but also to pay attention to the establishment of a language resource monitoring resource database in the entire Chinese-speaking region. It is not only necessary to establish a written language resource database, but also to pay attention to the establishment of a spoken language and dialect resource database. In addition, it is necessary to strengthen the sorting and inventory of existing databases, avoid duplication of construction, and take the road of intensive development.

Preserve, monitor, analyze and utilize language phenomena and language resources through the construction of a resource database, provide data services for social language and writing applications, provide experimental objects for scientific research, and provide decision-making basis for the formulation of national language and writing guidelines and policies [9, 10]. The object of language description "cross-language information retrieval" is expressed in five languages, and the form of expression is to describe the object using different language symbol systems endowed with rich semantics. Essentially, it is a multi-view representation formed by "cross-language information retrieval" on different language symbol systems [11, 12]. It can be seen from this that in natural language, for the same semantic object, using different languages (or languages) for text representation is essentially a different representation of the object, and different views formed from different sides (different languages). (Multilingual documentation). Based on this, the corresponding relations of different semantic granularities of the parallel corpus are displayed, as follows (Fig. 2):

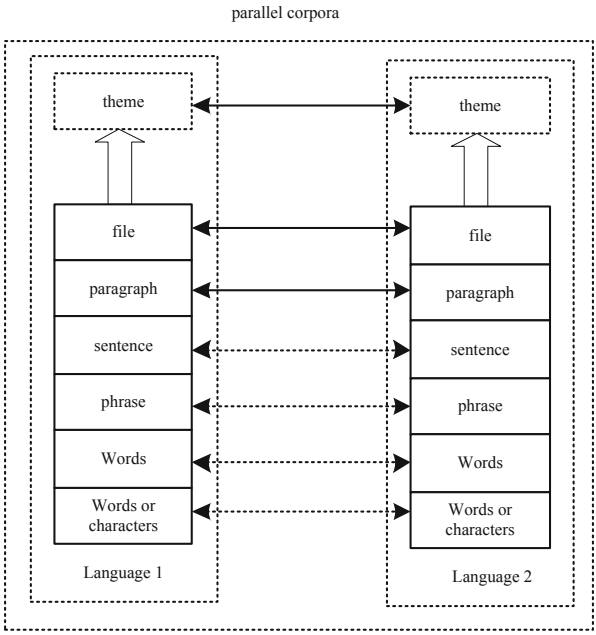

Fig. 2. Correspondence of different semantic granularity of parallel corpus

As shown in the figure, in the bilingual parallel corpus, it is divided according to the actual semantic granularity in the document set. From small to large, the granularity can be divided into words (or characters), words, phrases, sentences, paragraphs and documents. In the parallel corpus, there are semantic correspondences at each level of semantic granularity. Generally speaking, except for the document and paragraph levels, there is no strict bilingual semantic equivalence at the other levels. In the figure, there is no strict bilingual semantic equivalence relationship indicated by a dashed arrow, and a strict bilingual semantic equivalence relationship is indicated by a solid arrow [13]. For bilingual words or characters, the semantically equivalent forms include time, number, date and other language-independent content. Because of the differences in language expression, it is more difficult to align words, phrases, and sentences. However, there are many identical expressions in the languagesof the same language family, such as English and French. They can achieve the translation correspondence between words and phrases relatively easily through homology matching methods. If the corpus is aligned with sentences or words (phrases), the semantic correspondence between words, phrases, and sentences in bilinguals can be easily extracted in cross-language information retrieval and machine translation.

2.2 Corpus Multilingual Text Information Evaluation Algorithm

There are many commonly used evaluation indicators for the performance of information retrieval systems. In addition to the use of these evaluation indicators, the cross-language information retrieval system also uses the ratio relative to the performance of a single

language as an evaluation indicator. The most basic evaluation indicators are precision rate and recall rate. The accuracy rate is the ratio of the number of related documents in the number of retrieved documents:

$$\text{Accuracy} = \frac{\text{Number of related documents in search results}}{\text{Number of documents retrieved}} \tag{1}$$

The recall rate is the ratio of the number of documents retrieved in the number of related documents:

$$\text{Recall} = \frac{\text{Number of documents retrieved}}{\text{Number of related documents}} \tag{2}$$

The relationship between the precision rate and the recall rate is that when the precision rate increases, the recall rate decreases; conversely, when the recall rate increases, the precision rate decreases. In order to fully evaluate the performance of the model, the average precision rate at the 11-point recall rate is usually used. When the recall rate is 0, the precision rate is obtained by interpolation. Another widely used evaluation index is the average accuracy rate, MAP, whose calculation method is:

$$\text{MAP} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\text{pr}(d_{ij}) \tag{3}$$

$$\text{pr}(d_{ij}) = \begin{cases} \frac{r_i}{n_i} & \text{if } n_i \le \text{MAX} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Among them, n is the ranking of the Chinese documents in the search ranking results of the documents related to the query J, 1n is the number of related documents found with a ranking of n, M is the number of queries, and MAX is the ranking threshold (the TREC evaluation generally takes a value of 1000). Calculate the top n documents in the document retrieval ranking results to get the PN, which mainly reflects the direct evaluation of the user on the retrieval results. The correlation degree is obtained by calculating the similarity of the vector. The commonly used similarity calculation method is the angle cosine, which is defined as follows:

$$\text{sim}(\overline{D}, \overline{Q}) = \frac{\text{pr}(d_{ij})\overline{D} \cdot \overline{Q}}{|\overline{D}| \times |\overline{Q}|} \tag{5}$$

Among them:

$$|\overline{D}| = \sqrt{\sum_{i=1}^{n}(d_i)^2} \tag{6}$$

Further construct a simple probability model and a binary independent retrieval model. The BIR model assumes that the terms are independent of each other. The documents are sorted according to the odds ratio:

$$\text{Score}(Q, D) = \log\frac{P(\text{rel}|D, Q)}{P(\text{irrel}|D, Q)}$$

$$= \log \frac{P(D|Q,\ \text{rel}\ )P(\text{rel}|Q)}{P(D|Q,\ \text{irrel}\ )P(\text{irrel}|Q)}$$
$$= \propto \log \frac{P(D|Q,\ \text{rel}\ )}{P(D|Q,\ \text{irrel}\ )} \tag{7}$$

The document D is represented as a collection of independent binary events xn, and x = 1 represents the term x, which does not appear in the document. Then, from the formula:

$$\begin{aligned} &\text{Score}(Q, D) \propto \sum_{x_i \in D} \log \frac{P(x_i = 1|Q, rel)^{x_i}(1 - P(x_i = 1|Q, \text{rel}))^{1-x_i}}{P(x_i = 1|Q,\ \text{irrel}\ )^{x_i}(1 - P(x_i = 1|Q,\ \text{irrel}\ ))^{1-x_i}} \\ &= \sum_{x_i \in D} x_i \log \frac{P(x_i = 1|Q, \text{rel})(1 - P(x_i = 1|Q,\ \text{irrel}\ ))}{P(x_i = 1|Q,\ \text{irrel}\ )(1 - P(x_i = 1|Q, rel))} + Const \end{aligned} \tag{8}$$

Given such a document sample set, a contingency table for each term t can be calculated. Assuming that ni is the total number of documents in the sample set and R is the number of related documents, then (Table 1):

Table 1. Co-occurrence contingency table of information terms

Name	Relevant	Irrelevant	Total
Number of documents containing t	r_i	$n_i - r_i$	n_i
Number of documents without t	$R - r_i$	$N - n_i - (R - r_i)$	$N - n_i$
Total	R	N − R	N

The following probabilities can be derived from the table:

$$P(x_i = 1|Q, rel) = \frac{r_i}{R} \tag{9}$$

$$P(x_i = 1|Q,) = \frac{n_i - r_i}{N - R} \tag{10}$$

2.3 Realization of Multilingual Information Retrieval in Corpus

Based on the overall framework construction method described above, after we construct a bilingual topic space from a bilingual parallel corpus, we can achieve specific cross-language retrieval tasks. Without loss of generality, suppose the query expressed in language L1 retrieves the document set expressed in language L2. As shown in the figure, the general process of cross-language retrieval is (Fig. 3):

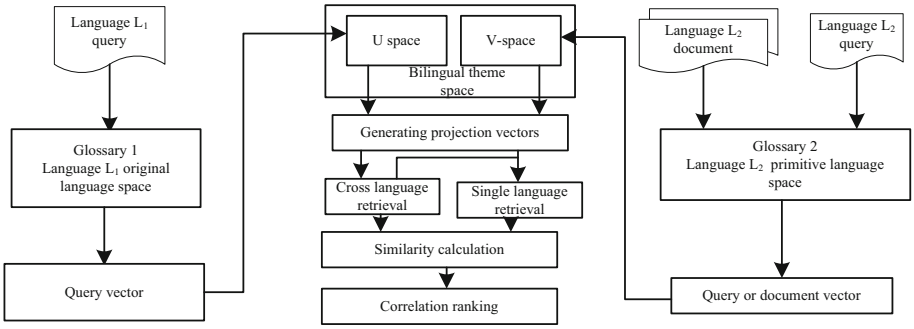

Fig. 3. The general process of cross-language retrieval under the overall framework

Perform simple similarity (such as angle cosine) or retrieval model calculation of similarity to query vector and document vector. According to the similarity calculated in the previous step, the retrieved documents are sorted based on the relevance ranking evaluation function, and the relevance ranking document list is returned. Information retrieval can be simply described as: According to the user's information needs, a query string is constructed and submitted to the information retrieval system. Then the system

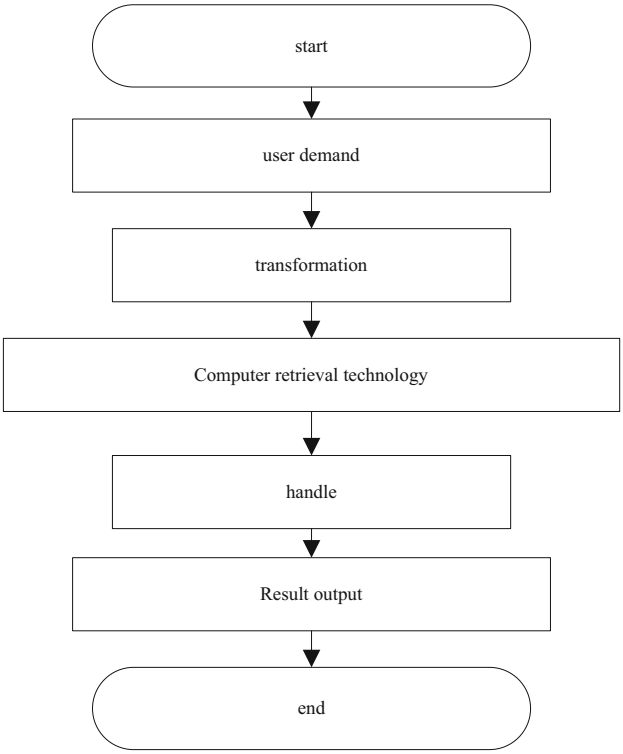

Fig. 4. Language and text information recognition steps

retrieves the document information related to the user's query from the relatively stable unstructured or semi-structured text data set, and sorts it according to the relevance from high to low. Finally, the sorted retrieval results are returned to the user, which is also called document retrieval. The identification steps of database information retrieval technology are shown in the figure (Fig. 4):

Information retrieval technology adopts the conversion technology of demand recognition between computer and user, involving retrieval structure, programming language, personalized demand, agent intelligence, information filtering, machine translation and other technologies. As can be seen from the figure, information retrieval is based on the user's information, enter the question in the computer, and then match the retrieved question with the identifier between the storage. The core is mainly to meet the needs of the retrieval subject, and then calculate the identified retrieval expression. The search expression can use the position operator and the restriction operator to combine and match the key words of the question, thereby determining the concept and position of the search key words, expressing the accurate content of the user's question, and ensuring the accuracy of the search accuracy.

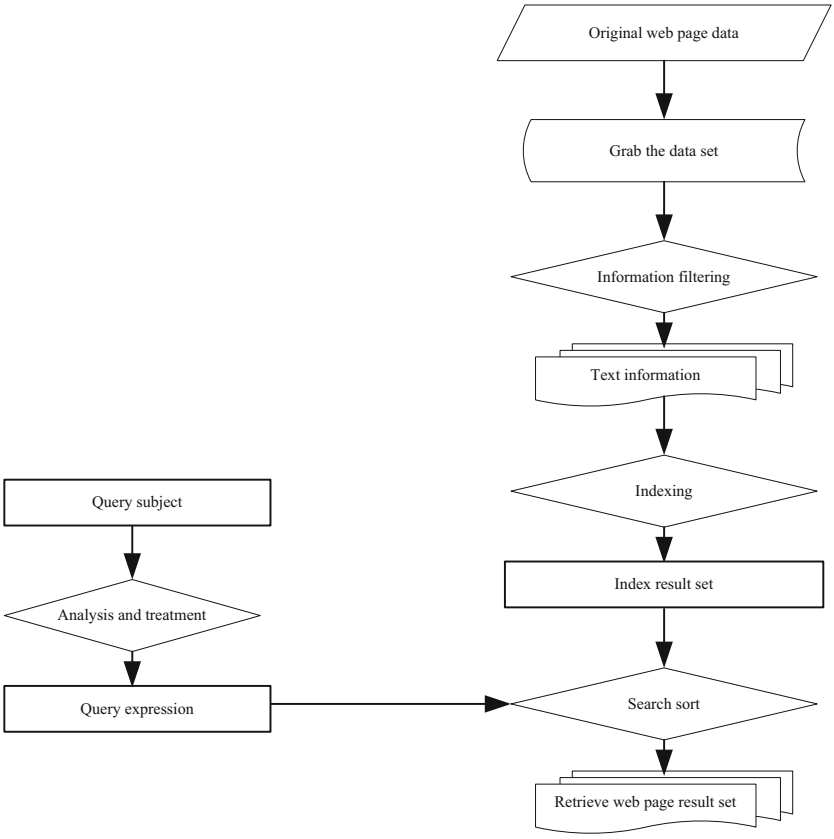

Fig. 5. Language and text information retrieval steps

Information retrieval includes clarifying users' information needs, information retrieval methods and techniques, and whether they can meet users' information needs. Among them, clarifying user information requirements, that is, user query, is a prerequisite for information retrieval, which is equivalent to when we perform a task, we must first clearly understand the requirements of the task, so as not to deviate from the direction. The method and technology of information retrieval is a means used in order to better achieve the goal during information retrieval, including some models and methods of information retrieval. Whether the information needs of users can be met is to evaluate the results of information retrieval to see how well it matches the needs of users. The higher the matching degree, the closer to the information needs of users. The basic process of general information retrieval is shown in the figure (Fig. 5):

In cross-language information retrieval, corpus is a very important basic data resource. As far as we know, there is no Chinese-English bilingual parallel corpus for cross-language information retrieval and evaluation. Regarding bilingual parallel documents as two views of the same semantic content, and assuming that they share the same semantic information, a cross-lingual information retrieval framework is proposed. The framework can extract the semantic representations of the same semantic object at various representation levels from the bilingual parallel corpus, construct a topic space representing linear or non-linear bilingual correspondence, and perform cross-language retrieval, multi-language retrieval, cross-language text classification, and cross-language text classification. Language and text clustering, etc., and lay the foundation for subsequent research content. When using the statistical language model for information retrieval, the query likelihood scoring method is defined as a conditional probability, that is, the statistical query Q is in the model M. Probability of generation under this condition. Therefore, the model of the query likelihood scoring method can be expressed as follows:

$$\text{Score}(Q, D) = p(Q|M_D) \tag{11}$$

In order to build a cross-language text classification corpus, firstly, some Chinese and English documents are clustered using K-Means algorithm to determine the document category. The SVM classification model is used to train the documents with the marked categories, and the unmarked documents are marked with categories. Perform manual category marking for proofreading and adjustment according to the content of the document. In order to establish a cross-language information retrieval corpus, the query content is determined according to the content description of the document set, and a standard query set is established. Based on the established query set, the probabilistic retrieval model is used to retrieve relevant documents to provide a preliminary basis for judging the relevance of documents. According to the retrieval results of the BM25 model, the document relevance is manually judged. Specific steps are as follows (Fig. 6):

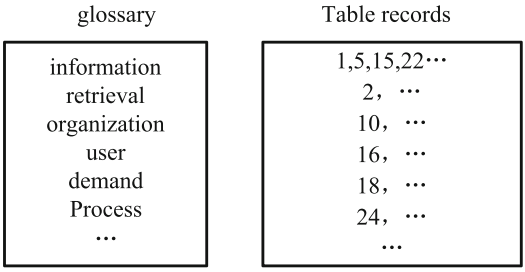

Fig. 6. Indexing process of text information files

Because a record table involves every word that appears in the text and its position in the text, the record table occupies a relatively large space. Retrieval using an inverted index is usually divided into three steps: vocabulary search, record table search and record table operation, which can do operations such as inserting, deleting, and updating documents. When searching with an inverted index, you only need to search the inverted table to obtain the documents required by the user. This search method is much faster than matching all documents linearly, thereby better guaranteeing the information retrieval effect.

3 Analysis of Results

In order to verify the actual application effect of the corpus multilingual text information intelligent retrieval model, the experimental environment is set up to ensure the detection effect. The operating environment of the experiment is:

Software Environment:
Operating system: Windows732 bit; Development tool: MicrosoftVisualStudio2008; Development language: C++; Search tool: Lemur.
Hardware environment:
System Model: LENOVOThinkCentreM8200T; processor: Intel(R)CoreM)Core 15650; Memory: 4 GBDDR2 experimental environment settings as shown in the table (Table 2).

Table 2. Experimental parameter settings

Parameter	Parameter
Host memory	4 GB
CPU frequency	3.50 GHz
Operating system	64 bit
Program running platform	Visual studio

The experimental reference database uses data collected randomly from 2000 networks, and the above data is repeatedly loaded 20 times to reach a database with a scale

of 40,000. The experimental data are randomly selected from the database. The data used in the word sense disambiguation module of this article comes from the training corpus released by the 2009 National Statistical Machine Translation Conference. The corpus contains 67288 double sentences and their corpus. Through comparative experiments on the information retrieval system, a series of test data are obtained, and the reasons why the statistical language model is better than the traditional information retrieval model are analyzed, and the three common smoothing techniques in the statistical language model are also compared and analyzed. And lists the results of several common evaluation indicators (Table 3).

Table 3. IFIDF common index evaluation results

Evaluation index	TFIDF evaluation results
map	0.3309
ndcg	0.6064
R-prcc	0.3465
bpref	0.8732
P10	0.2904

In order to generate a more intuitive result image, Lemur's own GUI program is used to graphically represent the evaluation results, and the graphical interface part is developed with Java/Swing. The following figure shows the recall rate based on Lemur's own corpus, query subject and standard set-accuracy rate and the accuracy rate of the top N results returned by the system for the query (Fig. 7).

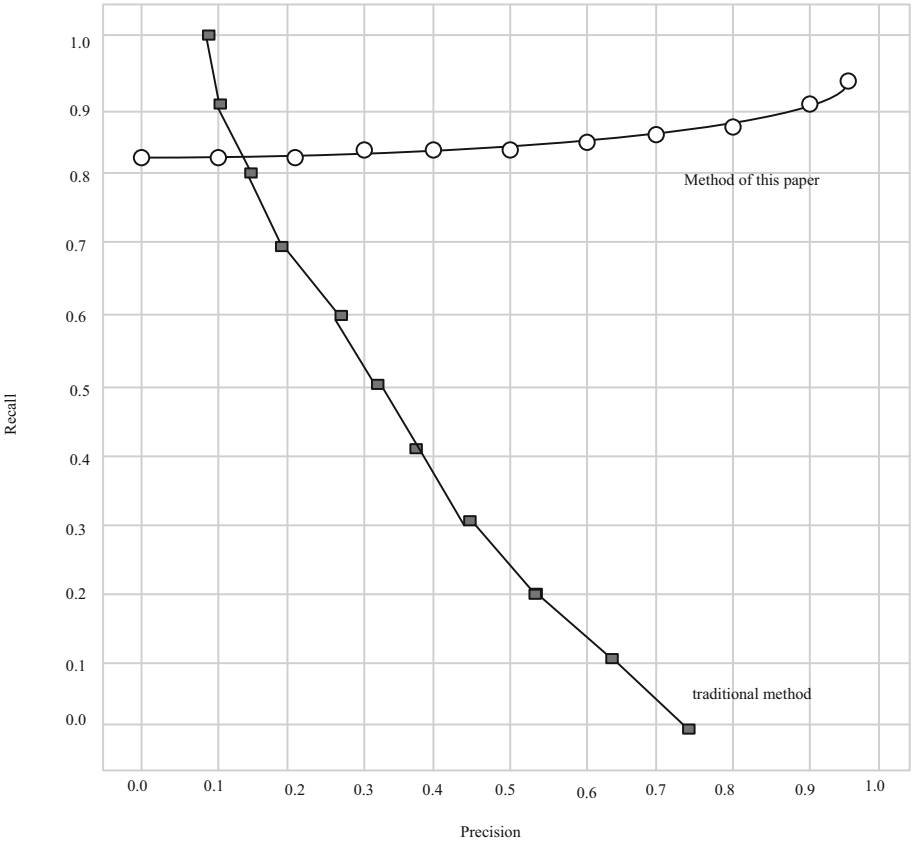

Fig. 7. Comparison of experimental monitoring results

Information retrieval evaluation can evaluate the pros and cons of different technologies and the impact of different factors on the system, thereby promoting the continuous improvement of the research level in this field. In the experiment, the three models of vector space model, probability model and statistical language model as well as the three smoothing methods of Jelinek-Mercer, Absolute Discounting and Dirichlet Prior in the statistical language model were retrieved in the corpus CWT200G, using MAP, P@N and R -Precision is the evaluation index. When searching, use the homepage as the entry point, and require the top ten results to include as many sites as possible. Therefore, the experimental results after treatment are shown in the table (Table 4):

Table 4. Retrieval model evaluation results

Evaluation model Evaluation index	TFIDF	OKPAI	KL-JM	KL-Abs	KL-Dir
MAP	0.1504	0.1387	0.1655	0.1728	0.1810
P@10	0.1182	0.1418	0.1768	0.1843	0.2098
P@20	0.0991	0.1136	0.1551	0.1693	0.1777
R-Precision	0.0446	0.0492	0.0603	0.0619	0.075

The traditional retrieval system is further compared with the recall rate of the system designed in this paper. In this experiment, segmentation, indexing and retrieval are all carried out by block processing. Therefore, in the retrieval process, each piece of data is retrieved, and the first 200 records are retained for each query, and then all the results are combined, and the results are sorted again according to the correlation score, and then the top 200 records are selected, so the most relevant results are basically included in the 200 records. The results are shown in the table (Table 5).

Table 5. Comparison results of recall rates of the two systems

Name	Traditional retrieval method	Article retrieval method
Decimation number	2165	2987
Correct number	1600	2987
Retrieval accuracy /%	52%	100%
System recall rate /%	60%	100%
Retrieval time / ms	2260	2100

For a large-scale corpus, it is impossible to list the accuracy of each query, so the calculation of the recall rate is not very accurate. Therefore, this article selects the accuracy as the evaluation index, which is simple and intuitive. For a given set of queries, the overall accuracy of the language model proposed in this article is higher than the previous two models. It is obvious that the retrieval performance of the model in this paper is better.

4 Concluding Remarks

The construction of language informatization is a basic work related to the long-term development of national informatization and information industry. Combined with the development status and practical application of language informatization, this paper makes a Macro Thinking on the construction of language informatization, and further discusses the key points, tasks and specific measures of the construction of language

informatization in China. The construction of language informatization should closely focus on the information application needs of national economic and social development, adhere to the principle of demand traction and technology promotion, and constantly contribute to the construction of a harmonious language life, the improvement of national cultural quality, the enhancement of the soft power of national sustainable development, and the promotion of all-round economic and social progress.

There is a large semantic gap between the low-level data features and the high-level semantic features of multimodal data sets, which makes the retrieval accuracy and retrieval quality need to be further improved. Most of the research focuses on constructing the neighbor structure based on the category label information and the similarity matrix, and maintaining semantic consistency. Perhaps a more perfect algorithm suitable for this field can be proposed based on the particularity of cross-modal data.

Fund Projects. 1. This paper is supported by the Postgraduate Research and Innovation Project of Northwest Minzu University , the name of the project is "A study on the names and descriptions of the things of mongolian <qin ding li fan yuan ze li>" (project number is Yxm202009), which is the phased achievement of this project.

2. This paper is supported by "the Fundamental Research Funds for the Central Universities-A Study on the Influence of Ecological Protection Policy of Qilian Mountains on Surrounding Herdsmen from the Perspective of Ecological Safety", which number is 31920190133.

References

1. Zou, J., Kanoulas, E.: Towards question-based high-recall information retrieval: locating the last few relevant documents for technology-assisted reviews. ACM Trans. Inf. Syst. **38**(3), 1–35 (2020)
2. Kim, H., Cha, M., Kim, B.C., et al.: Part library-based information retrieval and inspection framework to support part maintenance using 3D printing technology. Rapid Prototyping J. **25**(3), 630–644 (2019)
3. Kanwal, S., Malik, K., Shahzad, K., et al.: Urdu named entity recognition: corpus generation and deep learning applications. ACM Trans. Asian Lang. Inf. Process. **19**(1), 8.1-8.13 (2020)
4. Rascon, C., Ruiz-Espitia, O., Martinez-Carranza, J.: On the use of the AIRA-UAS corpus to evaluate audio processing algorithms in unmanned aerial systems. Sensors **19**(18), 3902 (2019)
5. Rojc, M., Mlakar, I.: A new unit selection optimisation algorithm for corpus-based TTS systems using the RBF-based data compression technique. IEEE Access **7**(10), 1 (2019)
6. Mishra, S., Soni, D.: Smishing detector: a security model to detect smishing through SMS content analysis and URL behavior analysis. Futur. Gener. Comput. Syst. **108**(10), 803–815 (2020)
7. Oh, I., Kim, T., Yim, K., et al.: A novel message-preserving scheme with format-preserving encryption for connected cars in multi-access edge computing. Sensors **19**(18), 3869–3870 (2019)
8. Deng, N., Deng, S., Hu, C., et al.: An efficient revocable attribute-based signcryption scheme with outsourced unsigncryption in cloud computing. IEEE Access **8**(10), 42805–42815 (2020)
9. Ferdinando, D.M., Sabrina, S., Salvatore, S.: A lightweight clustering–based approach to discover different emotional shades from social message streams. Int. J. Intell. Syst. **34**(7), 1505–1523 (2019)

10. Jia, Z., Jafar, S.A.: On the asymptotic capacity of X-secure T-private information retrieval with graph-based replicated storage. IEEE Trans. Inf. Theory **66**(10), 6280–6296 (2020)
11. Bhattacharya, P., Goyal, P., Sarkar, S.: Using communities of words derived from multilingual word vectors for cross-language information retrieval in indian languages. ACM Trans. Asian Lang. Inf. Process. **18**(1), 1.1-1.27 (2019)
12. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
13. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mobile Netw. Appl. **26**, 1891–1903 (2021)