Design and Development of the Bilingual Platform of Chinese Cultural Lexicon Based on ElasticSearch Technology

Tiantian Liang[1,a*], Longxue Yu[2,b], Ying Chen[1,c]

[1]School of Foreign Languages, Shenyang Jianzhu University, Shenyang, China
[2]Nanjing Wanggan Zhicha Information Technology Co., Ltd, Nanjing, China

[a*]teacherltt@126.com
[b]13591995090@163.com
[c]8902935@qq.com

Abstract. College English education shoulders the responsibility of cultivating international communication talents, and faces the challenge of revolutionizing teaching modes. The bilingual platform of Chinese cultural lexicon was designed and implemented in such a social background with ample culture-specific words and expressions to serve to improve students' interpretation capability of Chinese culture and integrate intelligent technology into the traditional language teaching. ElasticSearch technology was applied to the development of the bilingual platform. Based on the ElasticSearch inverted indexing, a distributed file storage and a retrieval analysis engine were established to realize a full-text search system with high efficiency and accuracy. The platform prioritizes users' needs and offers great convenience and pleasant using experience with its easy access and rich contents.

Keywords: ElasticSearch; inverted indexing; bilingual platform; Chinese culture interpretation.

1 Introduction

China's stories and voices take up a significant part of China's discourse system for global communication. The translation and dissemination of this system, embodying Chinese characteristics, are crucial for introducing and explaining China to the world. Recently, the ability to interpret this discourse system has emerged as a vital topic in higher education, aiming to enhance both "the language competency" and "the cultural interpretation capability". The international information dissemination capacity is a critical component of national language capacity[5], essential for effectively conveying China's story and culture in foreign languages and for disseminating theoretical ideas and research findings across various social fields and academic disciplines[2]. From this perspective, talents in the new era need to develop their multiliteracies, including scientific literacy, media literacy, cultural literacy and critical literacy[8].

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_40

Cultivation of talents with international communication capability has become an essential task for foreign language courses in colleges and universities.

With the rapid development of information technology, the integration of intelligent education and traditional education has gradually established a “ubiquitous” learning mode which enables users to conduct teaching and learning activities at anytime, anywhere, and anyway by taking advantage of digital content, physical surroundings, mobile devices, pervasive components, and wireless communication [6]. Its extensive application has greatly inspired the language teaching by providing diversified and innovative teaching methods. The college English course now has more access to various learning or teaching platforms on the way to the cultivation of talents with multi-competence and multiliteracies.

The bilingual platform of Chinese cultural lexicon (the platform) focuses on Chinese cultural vocabulary, exploring the cultural connotations and significance of the vocabulary with Chinese characteristics via English-Chinese bilingualism. The platform aims to enhance language proficiency and improve culturally critical thinking skills by providing a scaffold for the interpretation of the Chinese culture-specific vocabulary. As a helper for training international communication talents, this content-rich and easy-to-use platform effectively serves the foreign language teaching in higher education, enriching digital teaching resources and supplementing blended teaching methods.

2 Demand Analysis

We conducted a survey among a total number of 742 second-year college students. The survey mainly includes three aspects: the students’ Chinese cultural literacy, the acceptance of the integration of Chinese culture into college English learning, and the current situation of the integration, especially the teaching phases the Chinese culture has been integrated in. The survey results are shown in figures 1-3 respectively.

According to Figure1, 23% of the students admit that under the trend of economic and cultural globalization, they have a tendency to blindly worship the foreign cultures; 39% of the students admit that they lack the critical thinking awareness and skills in cross-cultural communication; 48% and 40% of the students acknowledge that they lack the industrious and enterprising attitude towards life and the conviction and passion for Chinese culture respectively.

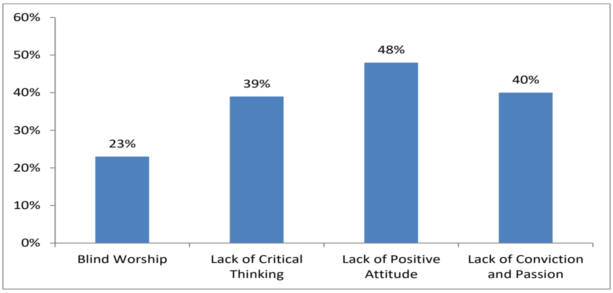


Fig. 1. Survey results of Chinese cultural literacy

Based on Figure 2, we can conclude that most students believe the integration of Chinese culture into the language learning is important and necessary, with 38% of high acceptance and 42% of acceptance. However, there are still a small amount of students who are unaware of the cultural significance in learning a foreign language.

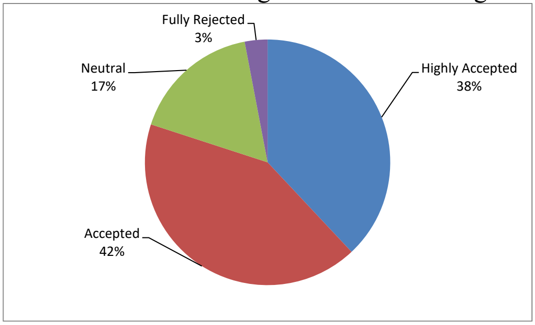


Fig. 2. Survey results of level of acceptance

As shown in Figure 3, cultural integration is mainly completed in the in-class teaching phase, especially in the sections of lead-in and text analysis, in which college English teachers exert a high initiative in digging deeply into the teaching material for the implicit cultural elements and realizing the integration through various classroom activities. However, the pre-class and after-class phases, in which the students' motivation in learning can be greatly enhanced, have been ignored.

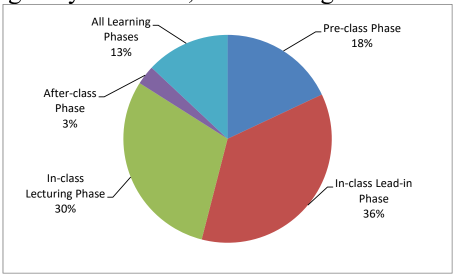


Fig. 3. Survey results of integration phases

Responding to the teaching and learning needs, the bilingual platform of Chinese cultural lexicon was designed and developed. The platform assists the students to learn languages from a cross-cultural perspective, which offers a medium for them to further understand the Chinese virtues and national spirit conveyed by the vocabulary rich in Chinese culture, so as to improve their Chinese cultural literacy. Also, realized as a WeChat mini program, the platform is more easily accepted by these young language learners. WeChat mini programs nowadays are especially prevalent among the youth, for they offer a greatly convenient channel to connect users with services, eliminating the need for download and memory usage, and bypassing the differences between iOS and Android systems compared with traditional applications[7]. The platform, simple in operation, caters to both students and teachers. By applying the platform to the teaching phases before and after class, teachers can complete the teaching cycle with Chinese culture integration, stimulate students' learning interest and improve their cross-cultural awareness.

3 Design and Implementation

3.1 Design Based on ElasticSearch

ElasticSearch (ES) is an inverted index server based on Lucene, providing a distributed file storage and a retrieval analysis engine. Lucene is an excellent retrieval tool that enables the full-text retrieval in all scenarios[1]. Integrated with Django WEB backend's multi-process services, ES offers an efficient and precise full-text search system for multiple users. It rebuilds a forward index into an inverted index and transforms the mapping of the file ID to the keywords related to it to that of the keywords to the file ID, which means each keyword corresponds to a series of files where this keyword appears[11]. The full-text search system, based on ES technology, addresses issues like low search efficiency, poor search quality, and single keyword matching in high-concurrency internet scenarios[9]. ES, developed in Java and released as open-source under the Apache license, has been widely applied to cloud computing and search ranking. What contributes more to ES popularity is its capability of real-time second-level searches on massive data and compatibility with multi-language development. Besides, ES enjoys a body of advantages, such as efficiency, stability, reliability, distributed storage, and etc.

Using ES as the core technology for data storage and retrieval, the design of this platform realizes a distributed search engine with a capability of real-time analysis by employing an inverted index programming. More specifically, the system design splits the index into multiple sharding, each with multiple replicas. Sharding is the cornerstone of distributed storage, and its purpose is to partition large indexes and distribute data throughout the cluster[4]. Every data node in the cluster can carry one or more sharding, coordinating and handling various operations, and realizing automatic load balancing through configuration files, thus meeting users' real-time search needs. Through this engine, users can quickly obtain a list of documents containing the searched word, completing the bilingual matching. The design framework of the platform is shown in Figure 4.

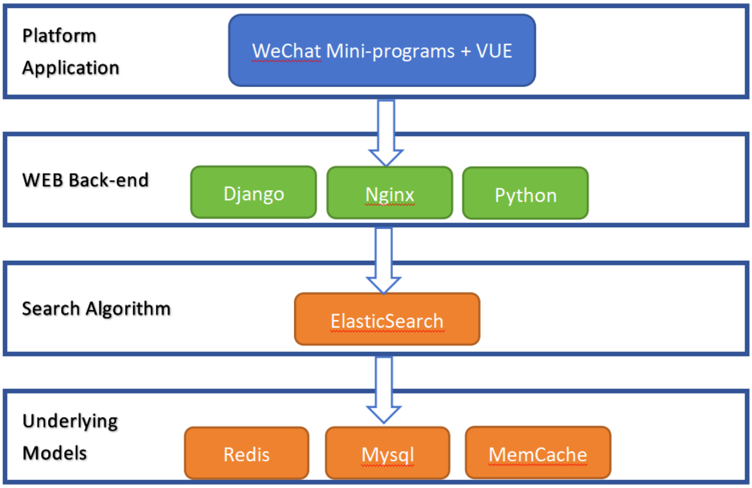
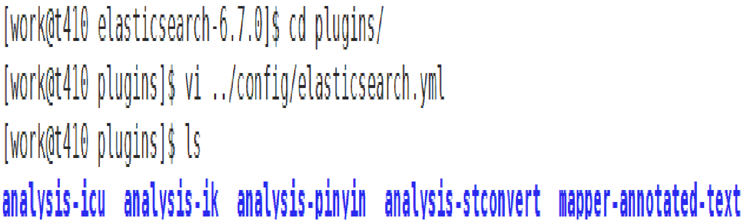


Fig. 4. Design framework based on ES

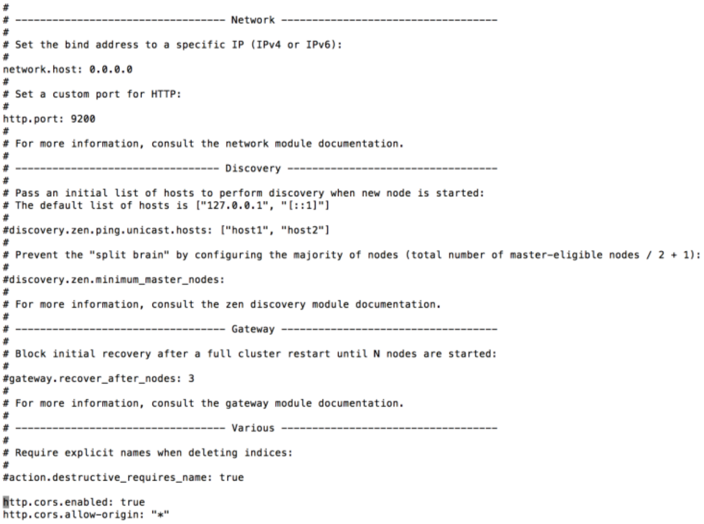
(1)Data Preparation: a total of 90 words and expressions with Chinese cultural information are extracted from related works and documents, each of which is matched with English translation within the realm of China's discourse system for global communication. The contexts and scenarios where each word or expression appeared are queried, and the specific contextual paragraphs are offered in bilingual versions, thus supplementing the culture-related background knowledge. The rich cultural contexts are further enhanced by audio and video files which are intended to facilitate the understanding of the cultural connotations. The platform is adaptable for multicultural education and English-Chinese translation, and allows the incorporation of data resources for iterative updates.

(2)Data Input and Storage: the prepared bilingual data have to be annotated before inputting. The ES segmentation plugin is configured (the configuration file is shown in Figure 5) and the sharding is stored across multiple nodes of the ES cluster (the configuration file is shown in Figure 6), providing an annotated data source for subsequent retrieval.



```
[work@t410 elasticsearch-6.7.0]$ cd plugins/
[work@t410 plugins]$ vi ../config/elasticsearch.yml
[work@t410 plugins]$ ls
analysis-icu  analysis-ik  analysis-pinyin  analysis-stconvert  mapper-annotated-text
```

Fig. 5. The configuration file of ES segmentation plugin

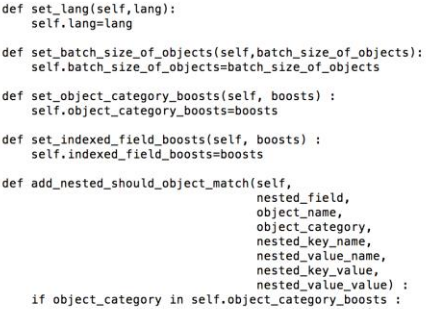


```
#
# ---------------------------------- Network -----------------------------------
#
# Set the bind address to a specific IP (IPv4 or IPv6):
#
network.host: 0.0.0.0
#
# Set a custom port for HTTP:
#
http.port: 9200
#
# For more information, consult the network module documentation.
#
# --------------------------------- Discovery ----------------------------------
#
# Pass an initial list of hosts to perform discovery when new node is started:
# The default list of hosts is ["127.0.0.1", "[::1]"]
#
#discovery.zen.ping.unicast.hosts: ["host1", "host2"]
#
# Prevent the "split brain" by configuring the majority of nodes (total number of master-eligible nodes / 2 + 1):
#
#discovery.zen.minimum_master_nodes:
#
# For more information, consult the zen discovery module documentation.
#
# ---------------------------------- Gateway -----------------------------------
#
# Block initial recovery after a full cluster restart until N nodes are started:
#
#gateway.recover_after_nodes: 3
#
# For more information, consult the gateway module documentation.
#
# ---------------------------------- Various -----------------------------------
#
# Require explicit names when deleting indices:
#
#action.destructive_requires_name: true

http.cors.enabled: true
http.cors.allow-origin: "*"
```

Fig. 6. The configuration file of sharding storage

(3)Data Retrieval and Output: At the front-end of the mini program, users conduct keyword searches via an input box. The keywords typed by the users are automatically recognized as either Chinese or English and then retrieved based on ES inverted indexing of the corresponding Chinese and English data sharding. The results, including the top relevant Chinese and English data and the most relevant highlighted words, are returned to the users in a list format through the mini program. The code snippets of ES data retrieval are shown in Figure 7.



```
def set_lang(self,lang):
    self.lang=lang

def set_batch_size_of_objects(self,batch_size_of_objects):
    self.batch_size_of_objects=batch_size_of_objects

def set_object_category_boosts(self, boosts) :
    self.object_category_boosts=boosts

def set_indexed_field_boosts(self, boosts) :
    self.indexed_field_boosts=boosts

def add_nested_should_object_match(self,
                                   nested_field,
                                   object_name,
                                   object_category,
                                   nested_key_name,
                                   nested_value_name,
                                   nested_key_value,
                                   nested_value_value) :
    if object_category in self.object_category_boosts :
```

Fig. 7. The code snippets of ES data retrieval

3.2 Design Implementation

The platform features multiple translation engines and matching systems, along with artificial intelligence-based fuzzy matching techniques, for the realization of relevance translation and bilingual whole-sentence translation. Under the premise of correct grammar, whole sentences are input and translated with high accuracy and precision. The main interface of the mini program displays the latest entries in the database, allowing users to choose their learning content autonomously.

Users are supposed to scan a QR code with WeChat to add the mini program, and then enter the platform. Presented with a “Login Window”, users can log in to the main interface of the platform using their WeChat linked accounts. The layout of the main interface is simple and concise, with the search bar at the top and the four main modules - “Search”, “History”, “Contributions”, and “Me” - displayed at the bottom, providing an easy-to-use user experience.

3.2.1 “Search” Module.

The platform enables bilingual searching. Users type the keywords of the desired search entries into the search bar. The back-end detects the text language type and sends the entered keywords for processing and analysis through an Http request. Once the back-end receives the user’s request, the coordinating nodes will transmit the request to the primary or secondary sharding of one or more data sharding for query. The results from each sharding are then aggregated at the coordinating node and returned to the user end. The core retrieval function sets different query weights according to the fields of title and content in the ES index, retrieving the most relevant documents based on the TF-IDF algorithm in ES. TF-IDF, a statistical algorithm, calculates the TF-IDF value of each word in the query within a document, based on which the matching degree between the document and the search query can be determined[3]. The TF-IDF formula is shown as follows:

$$\mathrm{Idf}\,(t) = 1 + \log\,(\mathrm{numDocs} / (\mathrm{docFreq}+1))$$

ES will evaluate the results of each matching query condition by performing score calculation, the higher the score, the higher the correlation[10]. To put it in another way, by considering word frequency, inverse document frequency, and word length, a score is obtained, and the document with the highest score will be returned to the user. Users can view all related translations, related scenarios, audio or video contexts, and cultural connotations of each entry. This feature allows users to learn comprehensively and their understanding of the cultural implications will be deepened as well.

In addition to free text search, the platform also implements voice search. With built-in voice recognition software, AI recognition technology accurately identifies users’ speech and converts it into text, facilitating entry searches in various usage environments.

On the right side of the user interface, there is a “Favorites” option, allowing users to autonomously save words and expressions for later study and review. The “favorites” are stored in user’s corresponding data form and can be queried and returned through the back-end interface.

3.2.2 “History” Module.

The platform automatically records all the entries that a user has searched. Users can freely view previously browsed entries in the mini program, including entry information, all related texts, audio or video files, and the exact time of search, which facilitates the review of past information.

3.3.3 “Contributions” Module.

After logging in via WeChat, users can submit their collected vocabulary entries through the “Contributions” module. Once approved, these entries can be shared with all users of the mini program, achieving the purpose of information sharing. This module design effectively exerts the users’ initiative, sparks their learning interests, and enhances their self-learning ability. Also, by multi-user sharing, the number of entries can be increased, which in turn expand the database. However, only approved entries will be shared with other users. Taking the security of this part into full consideration, the system designs a user and entry permission module in the database form, which can strictly isolate the unapproved entries from the general users.

3.3.4 “Me” Module.

In this module, users can view all their saved entries, the entries they have contributed, and the browsing history after logging in. This module design provides users with great convenience to record their learning process and accumulate learning data, essentially establishing a small learning portfolio for users themselves. This module not only meets users’ learning needs and enhances user experience but also strengthens their confidence for continuous use of the platform. At the same time, it also provides foundational data support for further expansion and development of the platform.

4 Conclusions

From the perspective of technical application, the bilingual platform of Chinese cultural lexicon is developed based on ElasticSearch technology in which the inverted index functions as a key technique to implement a distributed file storage and a retrieval search engine. Utilizing Redis, MySQL, and MemCache, it establishes a foundational model for quick and accurate search service. In terms of user experience, the platform, leveraging the WeChat mini program, offers easy access and use. The clear and concise interface design takes users’ needs as a priority, which has led to a positive user experience during the trial phase. Regarding future applications, this platform effectively serves the educational goals and teaching needs of cultivating international communication talents in colleges and universities. It enriches educational resources and modernizes learning methods, delving into the deeper cultural meanings through language learning, thus fostering students’ critical thinking and cultural interpretation capability. Students can also record their learning process through the platform, establish their own e-learning portfolios, and gradually develop their autonomous learning abilities.

Acknowledgment

The 2022 Liaoning Provincial Department of Education Research Project “Research on the Interpretive Capability of Foreign Discourse with Chinese Characteristics from the Perspective of New Liberal Arts” (Project Number: LJKMR20220925) ; The 2023 Higher Education Science Research Plan Project “Research on the Construction and Enhancement Strategies of Ideological and Political Competency Model for Foreign Language Teachers in Colleges and Universities” (Project Number: 23WYJ0413) ; The 2022 Liaoning Province Education Science “14th Five-Year Plan” Project “Study on Integrating Moral and Professional Education into College Foreign Language Courses for the Cultivation of International Communication Talents” (Project Number: JG22DB569).

References

1. Fan Liu, “Research and implementation of scientific and technological resources retrieval system based on ElasticSearch,” Modern Computer, vol. 27, 2021, pp. 93-100.
2. Jigang Cai, “On the cultivation of international information dissemination capacity: discursive competence and translation capacity,” China University Teaching, No. 1-2, 2023, pp. 19-24.
3. Lei Tao, Chenyang Su, Zhengdan Li, Jingwen Zhu, Yuzhi Zhang, “Educational resource search strategy based on ElasticSearch and semantic similarity matching,” Frontiers of Data and Computing, vol. 4, 2022, pp. 50-62.
4. Mingkai Li, Li Wen, “Design and implementation of knowledge based retrieval engine system based on ElasticSearch,” Software, vol. 44, 2023, pp. 184-186.
5. Qiufang Wen, “International information dissemination capacity, national discourse capability and national language capacity-- a strategy of developing ‘Two Types of Personnel’ for enhancing international information dissemination capacity,” Journal of Hebei University (Philosophy and Social Sciences), vol. 47, 2022, pp. 17-23.
6. Robledo L.A.C. & Ayala A.P., “Ubiquitous learning: a systematic review,” Telematics and Informatics, vol.35, 2018, pp. 1097-1132.
7. Rongshan Xing, Fangjun Kuang, “Design of WeChat mini-programs,” Computer Era, No.8, 2018, pp. 9-12.
8. Thwaites, T, “Multiliteracies: a new direction for arts education”, ACE Papers, No.13, 2003, pp.14-29.
9. Wenkai Liang, Hongling Tu, Jiahuan Chen, “A study based on ElasticSearch full-Text retrieval technology,” China Science and Technology Information, vol. 18, 2021. pp.82-84+87.
10. Yuanhe Dong, Yan Jia, Yong Zhu, Enze Li, Xianhong Xue, “Research on information retrieval method based on ElasticSearch distributed search engine,” Journal of Hebei Normal University (Natural Science), vol. 43, 2023. pp. 56-61.
11. Zhiqin Ma, Xuehua Liao, Wei Deng, Wenchao Xiao, “Research on distributed ElasticSearch based on similar content comparison algorithm,” Computer & Digital Engineering, vol. 48, 2020, pp. 2843-2849.