Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7589727, 8 pages
https://doi.org/10.1155/2022/7589727

Research Article

Construction of Online English Corpus Based on Web Crawler Technology

Yanfei Qi[1,2]

[1]*School of Foreign Languages, Hanjiang Normal University, Shiyan 442000, China*
[2]*University of Perpetual Help System DALTA, Las Piñas 1740, Philippines*

Correspondence should be addressed to Yanfei Qi; qiyanfei@hjnu.edu.cn

Received 26 May 2022; Revised 16 July 2022; Accepted 25 July 2022; Published 28 August 2022

Academic Editor: Jun Ye

Using 100 random questionnaires, 89 people said they were increasingly dependent on the Internet for information, while the remaining 11 said they were not dependent much. Nowadays, the development of Internet technology is more and more mature, and the scale of the Internet is more and more large. At the same time, with the gradual deepening of global economic integration, English has become one of the indispensable language methods for international communication and cooperation. The development of network technology has been applied more and more widely in the process of English teaching; especially, the construction, research, and practical application of corpus have ushered in a broad development prospect. Based on web crawler technology, this paper focuses on the construction of web English corpus, which lays a foundation for English learning. Experiments show that crawler technology can effectively solve the collection and recognition of big data in English corpus.

1. Introduction

In recent years, with the in-depth development of my country's education reform, Chinese higher education is developing in the direction of international trade, communication, and information technology. The transformation and development of higher education has begun to pay attention to the use of information network media. The university websites of Chinese universities are gradually becoming more and more important. At the same time, most college websites have both Chinese and English versions. College English websites have become an important medium for their outreach, exchange programs and partnerships, finding different students and countries, and raising their profile globally [1]. There is no doubt that English language websites play an important role in international educational exchanges and participation, international school participation, and international student education. At present, most Chinese universities have opened English websites. Since then, good translation, good publicity, good communication, website construction, and internationalization of college English websites have also become new concepts studied by scholars. In this context, the development of online English corpus is particularly important. Therefore, this paper focuses on the development of online English corpus with the help of web crawler technology. On this basis, a web crawler graph is first created [2], as shown in Figure 1.

2. Literature Review

Buts and Jones said that before the big data revolution, the government and enterprises could not save all data for a long time, nor could they efficiently manage and analyze such a huge data set [3]. Ukraine said that under the traditional technology, the data storage is limited, the management is backward, and the cost is expensive [4]. In a big data environment, the most powerful new process is the collection, cleansing, and analysis of multiple files to ensure efficient, flexible, and efficient operations. According to Wang and He, from the government, industry, and all walks of life, big data becomes important for them to see new ideas and provide self-help [5].

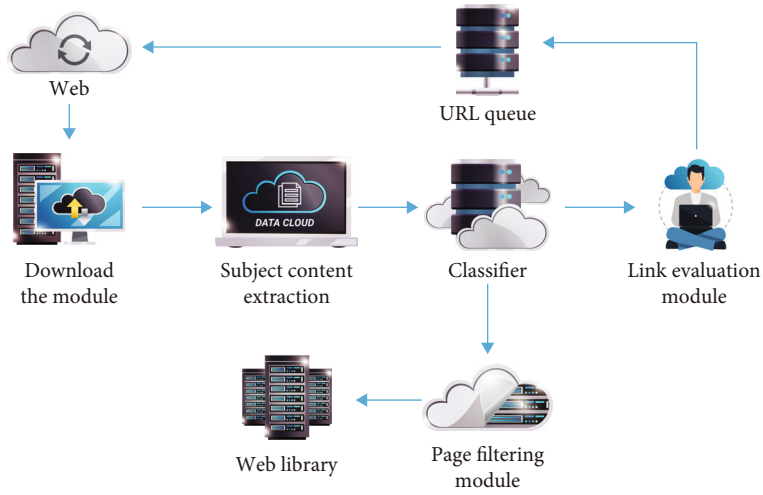

Figure 1: Architecture diagram of topic web crawler.

At present, Azazil has few research results on the application of corpus to college English websites [6]. Elgibreen et al., a corpus-based comparative analysis of English profiles of professors on English websites of Chinese and American universities, conducted a study on English profiles of Chinese and American professors from the perspective of evaluation theory [7]. In order to study the distribution of evaluation resources in the introduction discourse of English website professors in Chinese and American universities, Zhang et al. built a corpus and made a comparative analysis of the distribution of evaluation resources in the three subsystems of attitude, judgment, and grade difference under the framework of evaluation theory [8]. Kim and Davies selected 18 English profiles of Chinese and American university websites and built a micro corpus to study the public selection of high-frequency words of English profiles of Chinese and American university websites based on corpus [9]. Ding et al. conducted a corpus-based study on the core theme words of English profile of Chinese and American college websites. From minimal translation selections to full texts, these studies have started using the tools used to develop college English websites and completed some studies [10].

This paper makes a preliminary discussion on the translation of school website from the perspective of translation work. There are great differences in language and culture between English and Chinese college web pages. A full understanding of these differences is of great significance to the English translation of profiles. It is proposed that interpretative addition, modification or reorganization, and zero translation can be used as effective strategies for the translation of web profiles. Using the methods of case study and comparative study, this paper analyzes the English language news updated by four "985" colleges and universities in a city in 2021 and discusses the difficulties and solutions encountered in the network communication of college English websites. The audiences of college English websites mainly include students from various countries, overseas media, and brother colleges [11]. English websites must meet the needs of the above three groups of people in order to attract audiences and achieve the expected communication effect; this paper analyzes the problems existing in the translation of college English website propaganda and puts forward countermeasures and suggestions for the construction of college English websites from the following two aspects: on the one hand, through the comparison of parallel texts, analyze the similarities and differences in language and structure of global college English websites, and find out possible language translation errors, cultural translation errors, and functional translation errors; on the other hand, taking the text of college English website as the corpus, this paper investigates the translation initiator, translator, and audience and makes appropriate adjustments to the content and presentation of the source language, so as to make the translation meet the needs of the audience and make the college English website really play the role of external publicity.

3. Method

3.1. Overview of Web Crawler Technology. A web crawler (also known as a web spider and web bot) is a program or script that receives information from around the world through certain websites under certain laws. Web crawlers can skip a website's standard hyperlinks to search and store information. It starts with one page of the website, reads the content of the page, looks at other hyperlinks in the page, and then sees the next page through these hyperlinks [12]. Continue until all web pages on the Internet are captured. The broad classification of web crawler technology is shown in Figure 2.

Crawling means moving slowly in one direction. Technically speaking, web crawlers are tools used for data collection in search engines. They are called web crawlers, web spiders, or web robots. With the continuous development of technology, web crawlers are becoming more and more mature, which has gone beyond the definition of just a tool for search engines to collect data [13]. Generally speaking, a basic web crawler should have a set of seed URLs as input and a set of

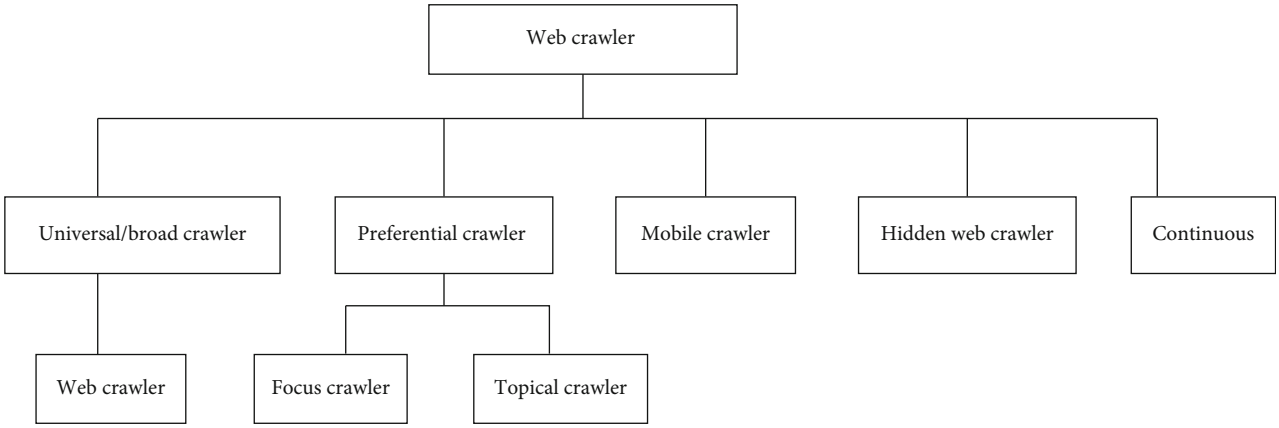

Figure 2: Classification of web crawlers.

crawled web pages as output. The specific workflow is as follows:

(1) Firstly, several initial URLs are selected as the starting position of web crawler according to the target

(2) Put the seed URL in the URL line to get it

(3) Read the URL through the URL line, resolve the DNS, and get the IP of the host. To pull the web page relative to the URL, identify the desired page, or remove the URL from the web page and place it on a line with the entry, and provide an entry for the URL in the URL field

(4) Compare the URL extracted from the web page with the crawled URL to remove the duplicate. Finally, put the deduplicated URL into the URL to be crawled, enter the cycle, and stop the web crawler when the stop condition is reached

3.2. Framework Comparison. Web crawler is the core tool for big data industry to obtain data [14]. Generally, when it comes to the collection task with small amount of data and simple capture logic, you can use self-made web crawler code for data collection. However, when it comes to the collection task with large amount of data and complex capture logic, if we continue to use the way of self-made web crawler for data collection, it will greatly increase the code development time, thus increasing the cost and time of the project. At present, there are many web crawler frameworks in the market, and different programming languages correspond to multiple web crawler frameworks. Table 1 gives some open source web crawler frameworks corresponding to programming languages in the current market.

Although there are many web crawler frameworks in the market at present, there are not many open source web crawler frameworks that are popular with developers and often used. The specific comparison can be seen in Table 2.

Scrapy is also an open source web crawler framework developed by python programming language [15]. Scrapy framework crawler has extremely powerful functions, high crawling efficiency, many related extension components, and very high degree of configurability and scalability and can flexibly customize crawling data. Scrapy can be used to download websites and extract output files from files without any issues on the page (Scrapy also provides users with options (links more like lxml) and cannot complete HTML code and simply delete files) and does a good job. For data mining, monitoring and automated testing, Scrapy can be easily customized as needed. Scrapy can be easily customized as needed. It supports the generation of crawler files from built-in definition templates, speeds up the creation of crawler code, and ensures that the code remains unchanged in large projects. In addition, scrapy also provides a variety of data export formats (JSON, CSV, XML, etc.), which can facilitate the connection with the database and data transmission in the project pipeline. The scrapy web crawler framework community has a large number of people and complete documents. It can deal with almost all current anticrawling websites. It is the most widely used web crawler framework in Python at present, as shown in Figure 3.

GitHub sets three options for open source project code: watch, star and fork, which can generally be used to indicate the activity and attention of the open source project code. The following table shows the comparison of watch, star and fork data of GitHub, an open source web crawler framework commonly used in the market so far (in descending order of star number, see Table 3) [16].

3.3. Algorithm Comparison. The most commonly used feature selection algorithms are data frequency, data gain, mutual data, and access statistics.

3.3.1. Document Frequency. The number of documents containing a word in corpus training is the frequency data for that word. The basic idea of this method is that words with low frequency often carry little information, so they cannot distinguish the categories well. Therefore, words with low frequency can be deleted, which can not only reduce the feature dimension but also improve the accuracy of classification.

Table 1: Web crawler framework corresponding to different programming languages.

Language	Web crawler framework
JAVA	Apache Nutch, webmagic, Heritriz3, WebCollector, craw1er4j, Spiderman, SeimiCrawler, jsoup-Gecco, and htmlunit
Python	Scrapy, pyspider, Newspaper, and Crawley
PHP	cola, Portia, python selenium, QueryList, phpspider, and PHPCrawl
Go	Beanbun, php selenium
C#	SmartSpider, Abot, xet, AngleSharp, HtmlAgilityPack, and CsQueryopen-source-sear ch-engine.Cobweb
C/C++	upton, Spidr, and Larbin
Ruby	wombat
node.js	node-crawler

Table 2: Comparison of web crawler frameworks.

Web crawler framework	Programing language	Describe
Apache Nutch	Java	It can collect all the contents of the website (general crawler and whole web crawler), regardless of the accuracy of collection and analysis. It is suitable for web search engines. However, nutch's crawler customization ability is relatively weak, modular design, and strong scalability; rich extraction page APIs. Support multithreading and distributed crawling. Support JS dynamic rendering page crawling. There is no framework dependency and can be flexibly embedded into the project.
Webmagic	Java	
Webcollector	Java	The Java crawler framework, which does not need configuration and is convenient for secondary development, can realize a powerful crawler with only a small amount of code. Support distributed.
Hretrix3	Java	The extensibility is enhanced to facilitate users to realize their own crawl logic. The biggest feature of the lightweight single machine open source crawler framework based on Java is simplicity. In addition, it also supports multithreading and proxy and can filter duplicate URLs.
Crawler4j		
scrapy	Python	A fast, simple, efficient and extensible web content capture framework developed entirely based on Python is used to extract the required data from the website. Scrapy has a wide range of uses and can be used for data mining, monitoring, information processing, and automated testing. Using scrapy, you can easily modify it according to your needs (scrapy is available).

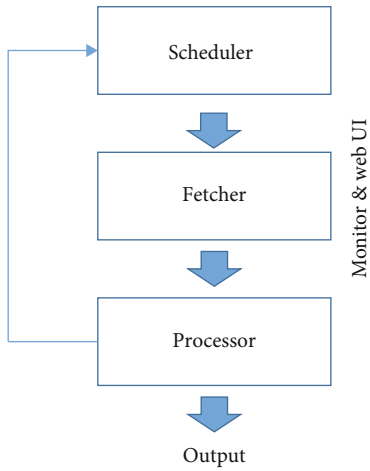

Figure 3: Pyspider architecture data flow diagram.

3.3.2. Information Gain. Information gain (IG) is to calculate the difference between the amount of information carried by the system when a feature appears and does not appear. For text classification, the difference between document frequency with and without feature word t represents the IG value of feature word t. IG value adopts the following formula, as shown in

$$\mathrm{IG}(t) = -\sum_{i=1}^{n} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{n} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{i=1}^{n} P(C_i|\bar{t})\log P(C_i|\bar{t}). \quad (1)$$

3.3.3. Chi Statistics. Chi statistics is often called square statistics, which is used to test whether two variables are independent. On the premise that the two variables are independent of each other, the deviation degree between the actual observed value and the theoretical value of the sample is calculated and expressed as chi value [17]. The larger the chi value, the two variables tend to be correlated; on the contrary, the two variables tend to be independent. The correlation between feature words and document categories can also be measured in this way. First, assume that the entry is independent of a category. The larger the chi value of the entry calculated on this basis, the greater the deviation between the result and the assumption, and the more relevant the entry is to the category. Therefore, the process of feature selection in this method is to calculate the chi value of each entry

Table 3: Comparison of GitHub data of web crawler framework.

Project	Language	Watch	Star	Fork
Scrapy	Python	1840	31956	7573
pyspider	Python	888	12865	3163
webmagic	JAVA	809	7730	3395
Colly	Go	219	7164	536
Pholcus	Go	441	5209	1331
node-crawler	node.js	256	4555	732
crawler4j	JAVA	307	3429	1719
WebCollector	JAVA	329	2294	1324
Apache Nutch	JAVA	245	1895	1135
QueryList	PHP	67	1469	250
Gecco	JAVA	133	1466	606
heritriz3	JAVA	174	1413	596
wombat	Ruby	53	1143	113

and category and sort it from large to small, and the top value is the feature. The calculation formula of chi value of word t for category CI is shown in

$$\mathrm{Chi}(t, C_i) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B) + (C + D)}. \quad (2)$$

In the model, N represents all the data in the body, A represents the data that contains t and is Ci, B represents the data that contains t but does not contain Ci, C represents the number of data that does not contain t but belongs to Ci, and D represents no word t or data from Ci.

3.3.4. Mutual Information. In data theory, interpersonal data (MI) refers to the amount of data provided by two events. The greater the amount of mutual information, the greater the correlation. The mutual information calculation formula of word t and category Ci is shown in

$$\mathrm{MI}(t, C_i) \approx \log \frac{A \times N}{(A + C)(A + B)}. \quad (3)$$

The meaning of variables in formula (3) is consistent with that in formula (2).

3.4. Final Feature Extraction. Based on the candidate feature set, extract the final feature set. The specific steps are shown in Figure 4.

(1) Calculate the information gain value. The information gain value of each feature in the candidate feature set is calculated in the training corpus

(2) Get features. According to the decreasing order of information gain value, select some of the top features

(3) Get the final feature. Add the features obtained from the entity information to the feature set obtained in step (2), and take this feature set as the final feature set

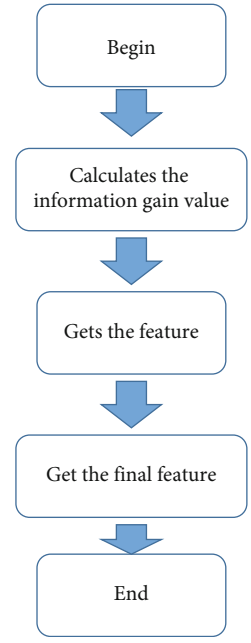

Figure 4: Final feature extraction process.

3.5. Classifier Based on Naive Bayesian Algorithm. The classifier is constructed by naive Bayesian algorithm [18]. Suppose that each instance a is represented by a feature set, and class c takes a value from a finite set C. A training example set and a test example $(a1, a2, \cdots, am)$ are provided.

The target of the instance A to be classified is to obtain the class tag c (a) of the instance $(a1, a2, \cdots, am)$, as shown in

$$c(a) = \arg\max_{c \in C} P(a_1, a_2, \cdots, a_m | c) P(c). \quad (4)$$

What we need to do now is to estimate the two probability values in equation (4) based on the training example set. Naïve Bayes classifiers assume that attribute values are conditionally independent of each other when a class tag is given. That is, the joint probability is exactly the product of each individual

feature probability. The specific formula is shown in

$$P(a_1, a_2, \cdots, a_m|c) = \prod_{j=1}^{m} P(a_j|c). \quad (5)$$

Substituting into formula (4), the classification formula of naive Bayesian classifier can be obtained, as shown in

$$c(a) = \arg\max_{c\in C} P(c) \prod_{j=1}^{m} P(a_j|c). \quad (6)$$

It can be simply calculated by calculating the occurrence frequency of different classes and eigenvalue combinations in the training example set. The specific formula is shown in

$$P(c) = \frac{\sum_{i=1}^{n} \delta(c_i, c)}{n}, \quad (7)$$

$$P(a|c) = \frac{\sum_{i=1}^{n} \delta(a_{ii}, a_j)\delta(c_i, c)}{\sum_{i=1}^{n} \delta(c_i, c)}. \quad (8)$$

Obviously, this approach leads to an underestimation of the results when the value of the zero-frequency property is present. In more severe cases, some values will be 0, making all numbers calculated by equation (6) to be 0. Laplace estimation is often used for smoothing to avoid the above problems. Equations (7) and (8) are rewritten, as shown in

$$P(c) = \frac{\sum_{i=1}^{n} \delta(c_i, c) + 1}{n + n_c}, \quad (9)$$

$$P(a|c) = \frac{\sum_{i=1}^{n} \delta(a_{ii}, a_j)\delta(c_i, c) + 1}{\sum_{i=1}^{n} \delta(c_i, c) + n_j}. \quad (10)$$

The workflow of naive Bayesian classifier based on entity link is as follows:

(1) The feature extraction method based on entity link is used for feature extraction [19]

(2) According to the obtained feature set, a naive Bayesian classifier is constructed and trained

(3) Preprocess the web page captured by the crawler, including topic information extraction, word segmentation, and other preprocessing, and then, quantify the web page

(4) The classifier is used to recognize the theme of the web page after vectorization processing. If the web page belongs to the theme class, the web page is saved to the theme page library; otherwise, the page is discarded [20]

4. Results and Analysis

For the evaluation of subject recognition effect, three indexes are mainly used: accuracy (P), recall (R), and F value. Accuracy is the proportion of the number of texts related to the topic accurately identified; recall rate is the ratio of the number of texts related to the topic accurately identified to the number of texts related to all topics in the training set; F value is a comprehensive evaluation index [21]. Suppose: in the training corpus, the text related to the subject and determined to be related to the subject is a, the number of texts unrelated to the subject but determined to be related to the subject is B, and the number of texts related to the subject but not determined to be related to the subject is C; then, the calculation formula of the three evaluation indexes is shown in

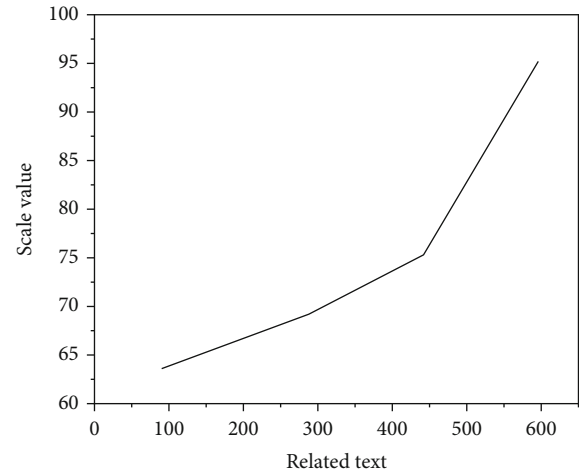

Figure 5: Experimental results (broken line).

Accuracy:

$$P = \frac{a}{a + b}. \quad (11)$$

The recall rate is shown in

$$R = \frac{a}{a + c}. \quad (12)$$

The F value is shown in

$$F = \frac{2 \times (P \times R)}{P + R}. \quad (13)$$

A total of 1443 military (587) and nonmilitary (856) articles were selected from Sogou news corpus as training corpus. The method based on entity link proposed in this chapter is used to construct naive Bayesian classifier for experiment [22]. The experimental results are shown in Figure 5.

From the experimental results, compared with the traditional naive Bayesian classifier, the introduction of entity link technology to improve it can achieve better results. As the entity link-based topic recognition algorithm can achieve good recognition effect, it lays a technical foundation for the construction of English corpus, because only efficient and accurate corpus recognition can build English corpus with richer content and more complete functions.

Table 4: Specific classification.

Common language	School profile, subject introduction, academic resume of professors, news activity reports, and introduction of institutions and departments.
Special terms	Organization department name (department, institute, institution, etc.), course name, major name, discipline name, and position name.

Corpus collection is the first step in the construction of a special English corpus for college English websites. In order to connect with international famous universities as soon as possible, improve the internationalization of Chinese university English website construction, standardize the classification system of Chinese university English website construction, and further improve the information content of Chinese university English website, we should focus on the current international first-class university English website construction, learn advanced experience, collect relevant corpus, and classify as a whole. So as to improve the Chinese expression of Chinese English and Chinese College English websites.

Use Baidu Encyclopedia, Wikipedia, and global college websites to collect information on the top ten colleges and universities in the United States. Classify and sort out the source language (English) texts of global university websites. Taking the English texts of global university websites (including website introduction, school brochures, and English version of teaching materials) as the research object, classify and sort out the corpus of Chinese and American university websites according to common terms and special terms, and construct "English original university website corpus" and "Chinese university English website corpus," respectively. This paper takes the special English of global college English websites as the research object, which mainly includes the Chinese and English versions of special English in two categories and ten subcategories of global college English websites. It mainly includes the Chinese and English versions of special English in two categories and ten subcategories of global college English websites. Each group of special texts should be no less than 10,000 words (English), and each special character text should be no less than 2,000 words (English). The specific classification is shown in Table 4.

From the results, a major problem in the development of English websites in Chinese colleges and universities is that there are too many Chinglish and Chinglish languages, and the language is incorrect. There are many news publicity materials and few practical application materials. There are many propaganda terms for Chinese audiences, but few actual contents for international scholars and international students. In particular, the construction of English websites of subordinate colleges and departments of Chinese universities, English websites of research institutions, and English websites of international students lags far behind that of the official website of Chinese universities.

On the basis of corpus collection, this paper describes the language characteristics of the English texts of "English original university website corpus" and "Chinese university English website corpus," respectively, and summarizes the unique language characteristics of the two kinds of corpus. Based on the two corpora established by the project, the world original vocabulary and vocabulary with Chinese characteristics are produced. The characteristics of vocabulary, syntax, and stylistic structure of college English websites are studied. Then, according to the characteristics of vocabulary, syntax, and stylistic structure of English websites, this paper explores typical sentence patterns and translation skills in the construction of college English websites. Through the comparative study, we find the similarities and differences between the English text of college websites translated from Chinese and the original English text and then put forward the similarities and differences between the two functions, so as to provide a real and objective basis for the subsequent improvement of the language quality of college English websites and the translation research of college English websites. Through corpus description and text analysis, we find that there are major differences between American college English official websites and Chinese college English websites in language style, discourse structure, vocabulary syntax, cultural connotation, and so on.

Generally speaking, the discourse structure of Chinese college English websites is influenced by Chinese discourse, focusing on parataxis and paying attention to the integrity, richness, and literary grace of the discourse. The English discourse of American college websites embodies the characteristics of English, focusing on hypotaxis and focusing on the form and logic of the discourse. Taking the above general introduction of the Tsinghua University and Princeton University as an example, it is not difficult to see that Tsinghua University advances linearly in chronological order, from the initial historical evolution to the introduction of the current school development in the middle and then to the final expression of the future vision and objectives of the school. On the whole, it is close to the structure of relevant Chinese texts on the school Chinese website, and the length is relatively long. The official website of the Princeton University first makes an overall introduction to the school in simple language to give readers a clear impression, and then, the detailed information classification in the website makes it easy for readers to find the practical information they want to know. In other words, a whole long text is divided into several short sub texts, which are introduced by classification, and the language introduced is mostly phrases and short sentences, which are concise and easy to understand.

5. Conclusion

The collection and recognition of corpus data is the foundation of English corpus construction, and web crawler technology has obvious advantages in data collection. Through experiments, this paper proves that web crawler technology can effectively identify and collect English corpus data, thus solving the core technical problems of English corpus construction. In addition, the construction of online English corpus based on web crawler technology is feasible, and it is very efficient in the collection and identification of corpus materials.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the fund project: The Phased Achievements of the Provincial Teaching Research Project (Research on School-Based Construction and Practice of Online and Offline Hybrid "First-Class Courses" of College English. No. 2020699) of Higher Education in Hubei Province in 2020 and Key Project of Science Research Plan of Hanjiang Normal University (Research on Information Literacy Based on Mobile Assisted Language Learning, No. XJ2020a01).

References

[1] A. E. Goldberg and T. Herbst, "The nice-of-you construction and its fragments," *Linguistics*, vol. 59, no. 1, pp. 285–318, 2021.

[2] J. Li, "Design, implementation, and evaluation of online English learning platforms," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5549782, 11 pages, 2021.

[3] J. Buts and H. Jones, "From text to data: mediality in corpus-based translation studies," *MonTi Monografías de Traducción e Interpretación*, vol. 13, no. 13, pp. 301–329, 2021.

[4] Z. Ukraine, "English detached adjectival constructions with an explicit subject: a quantitative corpus-based analysis," *Journal of Linguistics/Jazykovedný casopis*, vol. 72, no. 2, pp. 465–474, 2021.

[5] R. Wang and J. He, "Social gender construction in political context: a corpus-based study of lexical differences across genders," *Linguistics and Literature Studies*, vol. 8, no. 3, pp. 114–124, 2020.

[6] L. Azazil, "Frequency effects in the l2 acquisition of the catenative verb construction – evidence from experimental and corpus data," *Cognitive Linguistics*, vol. 31, no. 3, pp. 417–451, 2020.

[7] H. Elgibreen, M. Faisal, M. A. Sulaiman, S. Abdou, and M. Algabri, "An incremental approach to corpus design and construction: application to a large contemporary Saudi corpus," *IEEE Access*, vol. 9, pp. 88405–88428, 2021.

[8] J. Zhang, T. Zou, and Y. Lai, "Novel method for industrial sewage outfall detection: water pollution monitoring based on web crawler and remote sensing interpretation techniques," *Journal of Cleaner Production*, vol. 312, no. 1–4, p. 127640, 2021.

[9] J. B. Kim and M. Davies, "English what with absolute constructions: a construction grammar perspective," *English Language and Linguistics*, vol. 24, no. 4, pp. 637–666, 2020.

[10] H. Ding, Y. Chen, and L. Wang, "College English online teaching model based on deep learning," *Security and Communication Networks*, vol. 2021, Article ID 8919320, 11 pages, 2021.

[11] A. H. Meftah, M. Qamhan, Y. Seddiq, Y. A. Alotaibi, and S. A. Selouani, "King Saud University emotions corpus: construction, analysis, evaluation, and comparison," *IEEE Access*, vol. 9, pp. 54201–54219, 2021.

[12] N. Li, X. Jin, and Y. Li, "Identification of key customer requirements based on online reviews," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 1984, pp. 1–14, 2020.

[13] K. Thirugnanasambanthan, "A new approach to web crawling — dhekts crawler in comparison with various crawlers," *Indian Journal of Science and Technology*, vol. 14, no. 19, pp. 1580–1586, 2021.

[14] Z. Jiang, C. Chi, and Y. Zhan, "Research on medical question answering system based on knowledge graph," *IEEE Access*, vol. 9, pp. 21094–21101, 2021.

[15] G. Liu, S. Fei, Z. Yan, C. H. Wu, and J. Zhang, "An empirical study on response to online customer reviews and e-commerce sales: from the mobile information system perspective," *Mobile Information Systems*, vol. 2020, Article ID 8864764, 12 pages, 2020.

[16] N. A. Ismail, N. I. Ramzi, E. Su, and M. Razak, "Webometric analysis of institutional repositories of Malaysian public universities," *DESIDOC Journal of Library & Information Technology*, vol. 41, no. 2, pp. 130–139, 2021.

[17] U. K. Balajisaravanan, K. Karthick, S. Rajkumar, M. Murali, and N. Selvanathan, "Design of a personalized domain specific web crawler," *International Journal of Advanced Science and Technology*, vol. 29, no. 7, pp. 12162–12167, 2020.

[18] Y. Li, H. Wei, Z. Han, J. Huang, and W. Wang, "Deep learning-based safety helmet detection in engineering management based on convolutional neural networks," *Advances in Civil Engineering*, vol. 2020, Article ID 9703560, 10 pages, 2020.

[19] Y. Hao, X. Yan, J. Wu, H. Wang, and L. Yuan, "Multimedia communication security in 5G/6G coverless steganography based on image text semantic association," *Security and Communication Networks*, vol. 2021, Article ID 6628034, 12 pages, 2021.

[20] M. Li and J. Zhang, "Integrating Kano model, AHP, and QFD methods for new product development based on text mining, intuitionistic fuzzy sets, and customers satisfaction," *Mathematical Problems in Engineering*, vol. 2021, Article ID 2349716, 17 pages, 2021.

[21] Y. Yao, D. Hu, C. Yang et al., "The impact and mechanism of fintech on green total factor productivity," *Green Finance*, vol. 3, no. 2, pp. 198–221, 2021.

[22] Z. Zhai, X. Chen, Y. Zhang, and R. Zhou, "Decision-making technology based on knowledge engineering and experiment on the intelligent water-fertilizer irrigation system," *Journal of Computational Methods in Sciences and Engineering*, vol. 21, no. 3, pp. 665–684, 2021.