Article

Constructing a Dynamic Monitoring and Analysis Platform for the International Mainstream Media Corpus on China’s Image

Weihua Hu * , Wendi Wang , Chenghao Qu , Ruitian Li

Posted Date: 17 October 2024

doi: 10.20944/preprints202410.1364.v1

Keywords: mainstream media; China's Image; dynamatic mornitoring; platform constructing; sentiment analysis

Article

Constructing a Dynamic Monitoring and Analysis Platform for the International Mainstream Media Corpus on China's Image

Weihua Hu [1,*]; Wendi Wang [2]; Chenghao Qu [1] and Ruitian Li [2]

[1] Xi'an Polytechnic University
[2] Xi'an International Studies University
* Correspondence: huwh@xpu.edu.cn

Abstract: This article reports on the development of a big data processing platform that uses hot words（politics, economy, military, culture and society） reflecting the national image as the search object and the key time points as the time domain of data resources. The platform crawls (in real time), continuously expands and dynamically monitors a corpus of mainstream media covering China in developed countries (including the United States, Britain, Germany, France and Japan) and in developing countries (such as Brazil, India and South Africa). In this study, the principles of platform development and operation and the software and hardware operating environment were determined.

Keywords: mainstream media; China's image; dynamic monitoring; platform construction; sentiment analysis; natural language processing

1. Introduction

China's image is the recognition and evaluation of China's politics, economy, society, culture and other aspects by the domestic and international public, media and government (Hu & Li, 2017). It is also the interaction and care of self-cognition and others' cognition and expectation (Zhang & Wang, 2017). The 19th National Congress of the Communist Party of China (CPC) proposed to establish the image of a major country that is cooperative, inclusive, responsible and accountable, and takes a win-win position. The development vision of jointly promoting the construction of a community with a shared future for humankind was proposed. Today's international relations are complex, and China faces many difficulties in its comprehensive development and international exchanges. Against this backdrop, it is of great significance to timely monitor the China-related coverage in the international mainstream media, gain insight into the concentration of the visible or hidden coverage on China's image, and grasp the international public opinion and its diachronic changes for formulating and adjusting foreign policies promptly by the relevant departments.

With the continuous improvement in China's comprehensive national strength, the international community has paid increasing attention to China, and the influence of the international community on all aspects of China has become increasingly evident. Therefore, research on China's image in the international media has increasingly attracted the attention of scholars and relevant national departments. With the help of CiteSpace, Hu and Xu (2022) conducted a scientific knowledge map analysis of the research on China's image in China-related coverage in overseas mainstream media as collected by Chinese Social Sciences Citation Index (CSSCI) core journals from 2001 to 2020. This analysis revealed that Chinese scholars' research on China's image in the China-related coverage in overseas media in the past 20 years has yielded fruitful results, but there are also obvious limitations in the research methods adopted in previous research. Specifically, the academic articles in this field mostly adopted the framework theory combined with (critical) discourse analysis to explore the changes in China's image-building in the international mainstream media and the underlying causes. However, from a practical point of view, this method was used to objectively describe China's image in the international mainstream media in combination with specific problems in a specific period and a specific field. In contrast, articles that use statistical analysis of data and various measurement methods for quantitative research and visual presentation are clearly insufficient. There is thus a great lack of research on diachronic dynamic analysis and detection of China's international image by using corpus, statistics, visual presentation and discourse analysis. Therefore, it is necessary to build a dynamic monitoring and analysis platform for the China-related coverage in international mainstream media on China's image to make up for the shortcomings of previous research in data quantification and provide scientific methodological support for dynamically monitoring and evaluating China's image as reported by international mainstream media. The platform can fully use automatic data mining technology (Wang, 2016) to search the high-frequency core words in the reporting field and monitor the China-related coverage in the selected media in real time, thereby continuously, comprehensively and objectively analyzing the focuses and diachronic changes in the China-related coverage in international mainstream media.

The construction of a dynamic monitoring and analysis platform for the international mainstream media corpus of China's images is a highly professional and complex work combining computer science, machine learning, corpus retrieval and other technologies. Platform construction is directly related to the automatic retrieval effect of the platform in the future use process and the potential for big data research and analysis. Considering its importance, this paper aimed to introduce in detail the main links involved in constructing the dynamic monitoring platform for the international mainstream media corpus on China's image from the technical point of view. These links included determining the platform operating environment; designing the corpus crawling path; automatic data classification, organization, cleaning and storage methods; and developing corpus analysis software based on big data. First, a brief review and summary of the current research on China's image in international mainstream media were conducted to further emphasize the significance of building a dynamic monitoring platform for the international mainstream media corpus on China's image.

2. Literature Review

2.1. Research on China's Image in the International Mainstream Media

In recent years, research on China's image in the international mainstream media has received increasing attention from academics. This research has mainly focused on five aspects. The first aspect has researched the cognition of China's image in the international mainstream media during a certain period. Zhang Ying (2011) analyzed the New York Times' construction of China's national image during the Obama administration, and concluded that the New York Times has been more objective in its coverage aiming at China since the Obama administration. In a study examining the coverage of China in the New York Times and Los Angeles Times between 1992 and 2001, Peng (2004) found that while the coverage of China increased significantly over time, the overall tone remained negative. The study also found that stories presented in political frames and ideological frames were more likely to be unfavorable and no significant differences were found between the two newspapers. He and Chen (2012) analyzed Newsweek's reports on China from 2009 to 2010 and found a tendency of political generalization in its China image construction. The second area of focus is the study on the tendency of China image of mainstream media in a certain geographical area. Zhang and Cameron (2003) found that American mainstream newspapers tend to portray a negative image of China by analyzing the China-related reports of The New York Times, The Washington Post and The Los Angeles Times. As the United States is the largest dominant player in international public opinion, the tendency of China's image in multiple domains (positive-neutral-negative) and China's national image as a multi-ethnic country presented in various mainstream media reports have attracted scholars' constant attention (Zhou and Zheng, 2010; Gao and Jia, 2016, among others). Lan and Luo (2013) analyzed the British mainstream media's construction of China's image, pointing out that their reports on economic development are more objective and neutral, while their reports on politics or emergencies, as well as "human rights issues" are ideologically biased and one-sided. Zhang and Wu (2017) utilized critical discourse analysis to examine the representation of China in

English-language newspapers in China versus the UK. Their study revealed that China Daily emphasized the positive attributes of the One Belt, One Road Initiative, portraying China as a peace-loving nation and an emerging global economic power, while Financial Times presented mixed and conflicting images of China, depicting it as both having a significant impact on the global economy and as an authoritarian state and geopolitical threat. Huan (2023) examined the representation of China in Australian media discourses and found that the conventional binary perception of China–Australia relations as either an opportunity or a threat is oversimplified. According to the study by Huan and Deng (2021), the portrayal of China in South Africa's mainstream English-language newspapers transcended the traditional and often oversimplified dichotomy of partner or predator, recognizing the complexities, contradictions, and changing dynamics of Sino-SA relations. Li (2017) summarized the ways and characteristics of the Japanese mainstream media's portrayal of China and its tendency to maliciously distort and negatively focus on China. Shen and Wu (2013) analyzed the reports on China by three regional mainstream media in Russia and found that the image of China reported by them was somewhat distant from the actual image of China and what was described in the joint statements/communiqués of the two countries. The third area is the study of China's image shaping in international mainstream media in a certain field. Wang (2009) and scholars such as Wang and Han (2010) analyzed the image of China's economy in overseas mainstream media by focusing specifically on economic reports related to China's industries, China's manufacturing, and the three rural issues. Xia (2012) interpreted the backward and negative image of China in the Western media by analyzing the New York Times' coverage of the Forbidden City theft case. Pan and Dong (2017) used reports related to the 2016 China-Russia joint military exercises to scrutinize the negative Chinese military image constructed by the U.S. mainstream news media. Tang (2018) analyzed the portrayal of China in three US mainstream newspapers between 2008 and 2010, and revealed a predominantly negative representation. This representation is characterized by participant roles such as the Persuaded, the Criticized, the Labeled, the Contained, the Punished, the Helped, and the Praised, collectively contributing to a negative thematic emphasis, particularly in the domains of economy and trade. Wang (2018) conducted a corpus-assisted critical discourse analysis of news reports on air pollution published from 2008 through 2015 by China Daily. The study emphasized the submissive role of the official Chinese press under the CPC's strict censorship system and the role of international sporting events, such as the 2008 Beijing Olympic Games and the 2010 Guangzhou Asian Games, in raising awareness of environmental issues among both the public and the government. Teo and Xu (2021) conducted a critical discourse analysis comparing the portrayal of China's Belt and Road Initiative (BRI) in China Daily (CD) and New York Times (NYT). They uncovered divergent discourses between the two newspapers, with CD depicting the BRI as a collaborative project aimed at unifying and benefiting member countries, while NYT constructed the BRI as a geopolitical threat to the waning global influence of the US. Fourth, in a study of the causes of China's image shaping in international mainstream media, Zhang and Chen (2014) examined the political and cultural factors that interfered with the construction of China's image by British and American broadsheets in their coverage of the geopolitical conflict—Diaoyu Islands incident. Xu and Wang (2016) focused on the New York Times' ten-year China-related reports and analyzed the changes of China's image in the discourse construction of the western media in recent years and the causes behind it. Fifth, research has also taken a linguistic perspective on the study of China's image in international mainstream media. Zhong (2013) used critical discourse analysis theory to study the three-dimensional negative image of China constructed by the intertextuality of "Ye Shiwen's discourse" in British mainstream newspapers. Chen and Chen (2017) used corpus linguistics to analyze the changes of China's image in the eyes of Western media with news discourse from two different time periods of The Guardian. Hu and Li (2022) conducted a corpus-based analysis comparing English translations of the Report on the Work of the Government with the State of the Union Address. Additionally, studies from a linguistic perspective have contributed to our understanding of China's image in international mainstream media. Yang and Van Grop (2023) conducted an extensive analysis of diverging interpretations of the Belt and Road Initiative (BRI) through frames, shedding light on core claims constituting China's discursive legitimation of the BRI,

doi:10.20944/preprints202410.1364.v1

the differences between Chinese officials and foreign political-media elites, the continuity or change in foreign governments' positions on the BRI, and the increasing critical coverage by foreign elite media.

2.2. Research on Platforms for Opinion Mining

Tsirakis et al. (2017) proposed a platform for real-time opinion mining, named "PaloPro," to address the challenges faced by companies in processing huge volumes of streamed data from social media and other sources. This platform utilized prior research and real application services to meet the needs for brand monitoring in near real-time and providing insights on data in motion. Hartmann et al. (2018) demonstrated that compared to traditional Support Vector Machines (SVM) and Linguistic Inquiry and Word Count (LIWC), Random Forest (RF) and Naive Bayes (NB) exhibited superior performance in automatically classifying social media text, particularly in three-class sentiment classification and with small sample sizes. In the context of evaluating China's national image across various international media platforms, the methodologies and insights provided by Nandwani and Verma (2021) on sentiment analysis and emotion detection offered a promising avenue for deepening our understanding of the underlying attitudes and sentiments conveyed in media texts. Their work highlighted the capability of sentiment analysis to discern the polarity of texts—identifying whether the sentiments are positive, negative, or neutral towards subjects such as nations, administrations, or individuals. This approach could potentially augment current research methodologies by providing quantitative data on the sentiment bias in media coverage of China, thus offering a nuanced perspective on its image construction. Furthermore, the application of advanced analytical methodologies, such as the automated public opinion monitoring mechanism described by Karamouzas, Mademlis, and Pitas (2022), which utilizes Natural Language Processing algorithms to quantify text polarity, offensiveness, bias, and figurativeness in social media content, could significantly enrich the current research landscape on China's image in international media. The study by Chaudhary et al. (2021) demonstrated the use of big data analytics to predict consumer behavior on social media platforms by analyzing diverse and high-volume data collected from platforms such as Facebook, Twitter, LinkedIn, YouTube, Instagram, and Pinterest. Ramya and Sivakumar (2021) proposed a comprehensive approach involving sentiment analysis and the identification of influential users in social media platforms. By utilizing techniques like weighted partition around medoids with artificial cooperative search (WPAM-ACS) and fuzzy deep neural network (FDNN), they addressed challenges in sentiment classification, especially in languages lacking adequate resources for NLP models. Additionally, the survey conducted by Adak et al. (2022) comprehensively reviewed various aspects related to the infosphere, including collaborative systems like Wikipedia, the role of the infosphere in facilitating scientific citations and interdisciplinary research, and challenges related to governance, such as addressing rising hateful behavior, abusive behavior, bias, and discrimination in online platforms and news reporting. Although research on China's image in the international mainstream media have generated insights, there are still several shortcomings. First, **the existing research covers a short time span, and there is a lack of dynamic monitoring and analysis of China's image**. The limited time span has broken the continuity of the diachronic evolution of the research on China's image in the international mainstream media, making it impossible to grasp the dynamic changes in China's image in real time. Second, **the media and regional scope that have been studied are limited, and most studies have focused on the mainstream media in several developed countries in Europe and the United States**. There is little research on important media in Germany, Russia, Japan and other countries. The voice of mainstream media in Asian and African countries is ignored. In particular, the mainstream media lack the recognition of China's image in developing countries and are less representative due to the imbalance of the media's regional choice. Third, **the existing research focuses on topics in the political and economic fields with little attention paid to other topics**. With the enhancement of China's role as a responsible great power, it is difficult to present the complex composition and evolution of China's image from a single dimension. With respect to China's image, research must be extended to other dimensions, such as society, culture, military, diplomacy, climate and geography. Fourth, **most of**

the existing studies focus on analyzing the causes at the macro level, with few targeted strategy suggestions. Most of the studies describe only the image of China that is presented by international mainstream media and analyze the macro causes, such as ideology, cultural background and international relations. These studies lack operational and targeted strategy suggestions for improving China's international image. Fifth, **the interdisciplinary and quantitative level of existing research is insufficient, and the scientificity of research methods must be improved**. The disciplines involved in the research include politics, communication and linguistics. However, communication and linguistics are not sufficiently connected in terms of disciplinary theories and methods, thus greatly limiting the breadth and depth of the study. Wang (2016) proposed to develop a large diachronic corpus to compensate the deficiency of such research by providing constantly increasing objective descriptions and analyzable data.

In conclusion, the methods and tools used for researching China's image in international mainstream media should be further enriched. The country and regional scope of media sources and the related disciplines should be expanded to include comprehensive linguistics, communication, sociology, political science, journalism, and statistics. A comprehensive analysis of qualitative and quantitative, diachronic and synchronic, language and framework methods with a long-time span should be carried out based on big data. However, the prerequisite for achieving this goal is the construction of a dynamic monitoring and analysis platform for the international mainstream media corpus on China's image. With hot words as the search object and key time points as the horizon, this platform is technically characterized by dynamic monitoring and continuous expansion of the corpus. This platform can deepen the previous research on the causes of China's image-building in the international mainstream media, the analysis of China's image in the international media, and the combination of China-related public sentiment and opinion. Additionally, this platform can provide dynamic and quantitative information support for timely grasping the focuses and diachronic changes and comprehensively understanding the international mainstream media corpus on China's image in the future. Since the construction of a dynamic monitoring platform for the international mainstream media corpus on China's image is a highly technical work, the platform development principles, operating environment and how to realize media corpus crawling, automatic storage and corpus analysis during platform construction were introduced from the technical perspective in the following section.

3. Platform Construction and Operation

3.1. Construction Principles and Software and Hardware Operating Environment

According to the practical, rational, secure, extensible, maintainable and open principles for platform construction, the Python developing language was adopted under the Model-View-Template (MVT) architecture to build the operating framework of Django (Huang, 2019). For hardware operation, a 1.8 GHz Intel Core 2 or higher equivalent Advanced Micro Devices (AMD) processor was required, whose memory capacity was Windows 7:2 GB or Windows 10:4 GB, and hard disk capacity was above 15 GB. For software operation, PyCharm was used as the developing platform, MySQL was used to provide back-end service support, Redis and MongoDB as the nonrelational database, native HTML5 to design front-end pages and ECharts technology to render data.

3.2. Principles and Methods of Corpus Crawling

3.2.1. Principles of Corpus Data Crawling

The system incorporated the following 10 international mainstream media platforms in the real-time crawling platform of the corpus: *The New York Times* (US), *The Times* (UK), *Frankfurter Rundschau* (DE), *Le Monde* (FR), *The Moscow Times* (RUS), *Asahi Shimbun* (JP), *The Times of India* (IN), *The Australian* (AU), *The Citizen* (ZA) and *Folha de Sao Paulo* (BR). With the Scrapy architecture, the crawler system analyzed the crawled data on the web page in real time. Scrapy is an application framework that is coded to crawl website data and extract structural data (Cui, 2018). The framework can be applied to a range of programs, such as data mining, information processing and historical data saving. Scrapy was initially designed for web page crawling (more exactly, network crawling), but the framework can also be applied to retrieve data returned by an application programming interface (API) (such as Amazon Associates Web Services) or as a general network crawler. Scrapy has been widely used for data mining, monitoring and automated testing.

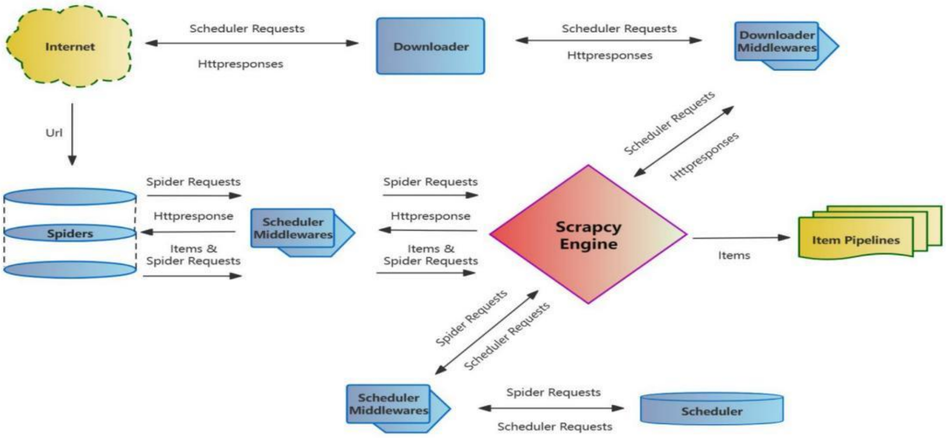


Figure 1. Architecture of the Scrapy crawler.

Figure 1 presents the architecture of the Scrapy crawler. As shown in Figure 1, the Scrapy crawler undergoes three stages:

First, the Uniform Resource Locator (URL) is obtained to generate the request.

Second, the Spider sends the request to the Scheduler through the engine. The Scheduler places the request in the queue for the Downloader to crawl. Once the web page downloading is completed, the response will be sent.

Third, the Spider receives the response sent by the Downloader to analyze and save the data.

To further improve Scrapy's crawling efficiency, the download speed of the Downloader should be increased in these three steps. By extending the middleware of the Scheduler and Downloader, one crawling Queue is always maintained to realize the shared Queue crawling. In this manner, after a queue is dispatched by the Scheduler, other Schedulers do not repeat the Request to achieve synchronous crawling by multiple Schedulers (Figure 2).

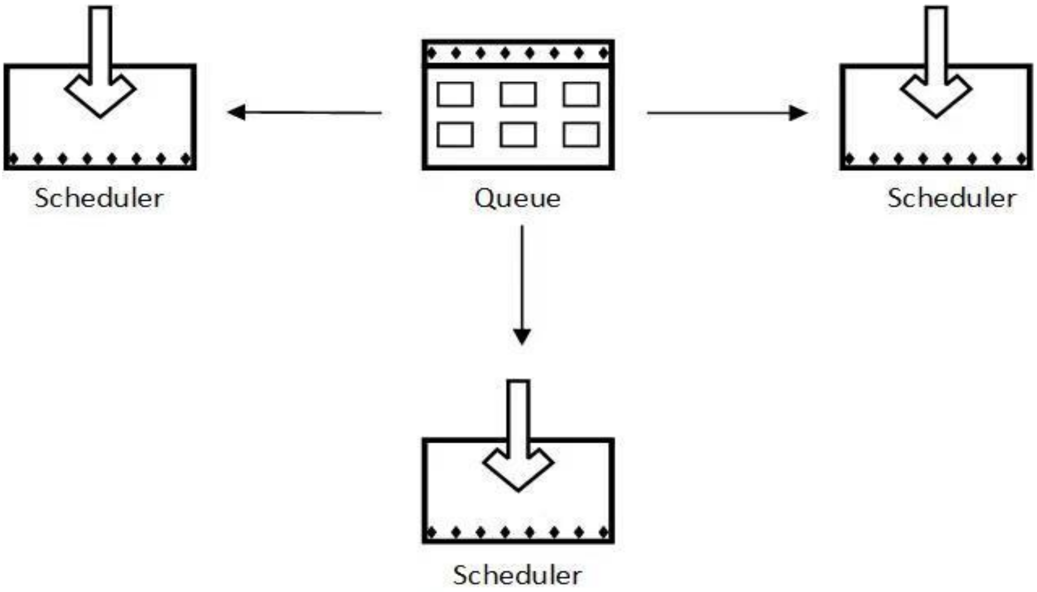


Figure 2. Scheduling of the distributed crawler.

3.2.2. Acquisition and Deduplication of Corpus Data

(1) Maintenance of the Crawling Queue

In view of the data crawling efficiency, Redis based on memory storage was adopted by the system. Redis supports various data structures (List, Set, Sorted Set, *etc.*) with simple access operation. Each data structure supported by Redis enjoys its own advantages in storage. Several methods are listed, such as lpush (inserting one or more values into the list head), lpop (removing and returning to the first element in the list), rpush (inserting one or more values into the list end) and rpop (removing and returning the last element in the list). These methods can be adopted by the system to realize a first-in, first-out crawling queue or a first-in, last-out crawling queue. There are unordered and nonrepetitive elements in the Set, thus making it easy to realize a random-ordered and nonrepetitive crawling queue. Sorted Set is represented by scores, and the Scrapy Request can also be used to control the priority scheduling queues. Based on the needs of specific crawlers, different queues can be flexibly selected.

(2) Data Deduplication

For distributed crawlers, if the set of each crawler is used to deduplicate, the set is still maintained independently by each host and cannot be shared. If different hosts generate the same Request, they can only be deduplicated individually.

To realize deduplication, Scrapy's built-in fingerprint set also must be shared. Redis happens to have the collected structure of saved data, and the system can adopt the Redis set as the fingerprint set to achieve sharing the deduplication set based on Redis. As shown in Figure 3, after a new Request is generated by a host, the Request's fingerprint will be compared with the set. If the fingerprint already exists, then this Request is a duplicate; otherwise, the fingerprint of this Request should be added to the set. The same principle can be used to realize the distributed Request deduplication for different storage structures.

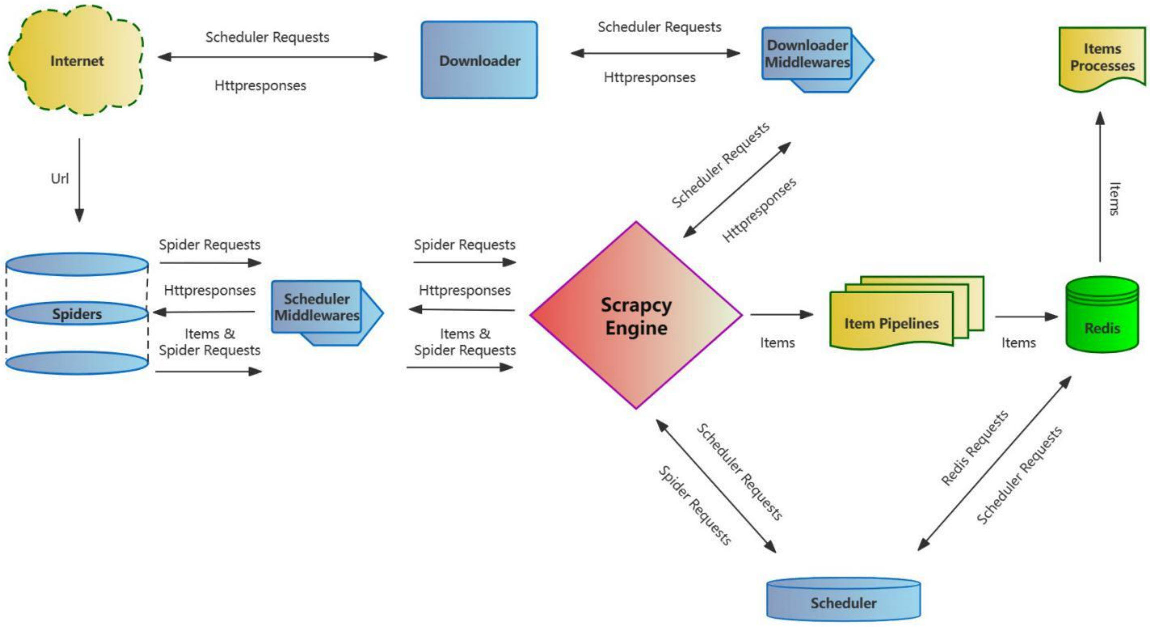


Figure 3. Architecture of the distributed crawler.

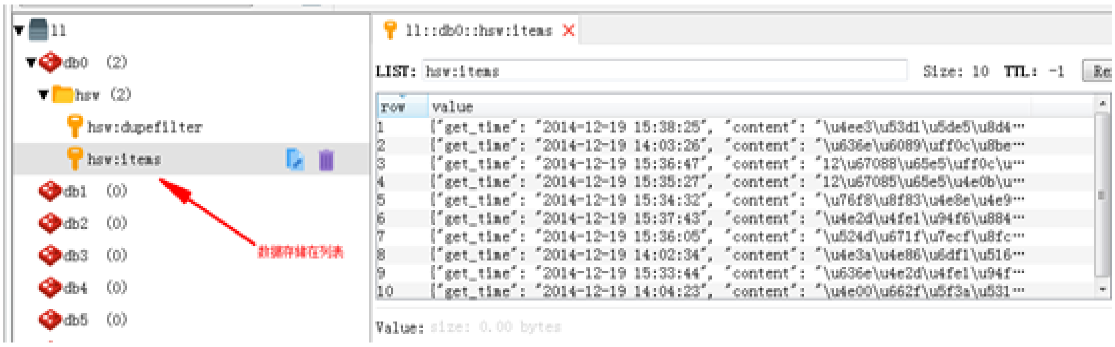
(3) Interruption Prevention

The Request queue in Scrapy is placed in memory when the crawler is running. When the crawler's operation is interrupted, the space of this queue is released, and the queue is broken. Therefore, once the crawler's operation is interrupted, the next crawling is a completely new process.

To continue the crawling after an interruption, the system is required to save the Request in the queue so that the next crawling can retrieve the previous crawling queue by directly reading the stored data. The previous crawling queue can be retrieved by specifying a storage path for the crawling queue in Scrapy; this storage path can be identified by the JOB_DIR variable, with the following command: scrapy crawl spider-s JOB_DIR=crawls/spider. The system of Scrapy actually saves the crawling queue locally, and the second crawling can directly read and retrieve the queue. As the queue itself is saved as the database, even if the crawler suffers an interruption, the Request in the database still exists, and the crawler will continue to crawl from the last interruption next time.

(4) Data Storage

Deployed as mentioned above, all Slaver servers start to crawl data, and the crawled data are saved in the Redis database with the shared request queue, request fingerprint set and data queue together with the Redis database (Figure 4), which can be used for subsequent data analysis.



```
LIST: hsv:items                                Size: 10   TTL: -1
row  value
1    {"get_time": "2014-12-19 15:38:25", "content": "\u4ee3\u53d1\u5de5\u8d4…
2    {"get_time": "2014-12-19 14:03:26", "content": "\u636e\u6089\uff0c\u8be…
3    {"get_time": "2014-12-19 15:36:47", "content": "12\u67088\u65e5\uff0c\u…
4    {"get_time": "2014-12-19 15:35:27", "content": "12\u67085\u65e5\u4e0b\u…
5    {"get_time": "2014-12-19 15:34:32", "content": "\u76f8\u8f83\u4e8e\u4e9…
6    {"get_time": "2014-12-19 15:37:43", "content": "\u4e2d\u4fe1\u94f6\u884…
7    {"get_time": "2014-12-19 15:36:05", "content": "\u524d\u671f\u7ecf\u8fc…
8    {"get_time": "2014-12-19 14:02:34", "content": "\u4e3a\u4e86\u6df1\u516…
9    {"get_time": "2014-12-19 15:33:44", "content": "\u636e\u4e2d\u4fe1\u94f…
10   {"get_time": "2014-12-19 14:04:23", "content": "\u4e00\u662f\u5f3a\u531…
Value: size: 0.00 bytes
```

Figure 4. Redis storage.

3.3. *Automatic Classification and Organization of Corpus*

3.3.1. Design of Corpus Database

The news data captured by the crawler system should be organized and saved in the database to facilitate data query, corpus analysis and front-end data display in the system subsequently. The design of the database plays an important role in the overall information system development and design. The efficient operation of the system depends on the merits and demerits of the database design. Before constructing a system, therefore, the database should be designed according to the specific needs and the functions of the system to meet the application needs of different users. Following the corpus data design specification, the organized data are saved in a MySQL database for level 1 cache of the crawled data into Redis and MongoDB. According to the specification of corpus data, the data are escaped, cleaned and archived after further cleaning.

(1) Corpus E-R diagram (Figure 5)

Name: news_table

Description: The structure of the processed database after being acquired by the crawler.

Definition: The database fields include news no. (id), title, state, newspaper, date, author, language, word count and classification tag.

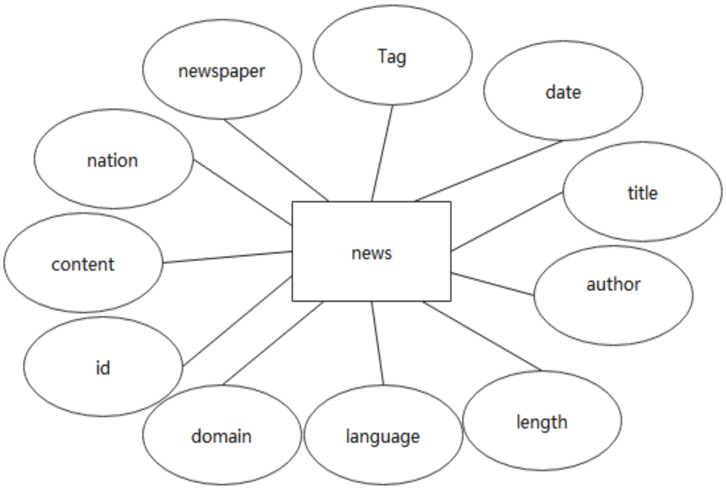


Figure 5. Database E-R diagram.

Each field in the database corresponds to the marking information of the corpus. The field of "content" stores the content of the whole corpus, and the retrieval corpus takes all the fields as retrieval objects. "Tag" is an important field that shows the classification of corpus content. The "Tag" field is independently set as a foreign key structure, as a separate sheet, representing the information on politics, economy and other dimensions.

(2) Processing Flow of Generated Corpus Data (Figure 6)

The data crawling of each website is conducted in accordance with the crawler data architecture. After the data are cleaned by the data cleaning algorithm, the result of extracting the keywords and text similarity can be saved in the database. The Django framework is used for back-end data management and data transmission, while ECharts is used for front-end data flow display.

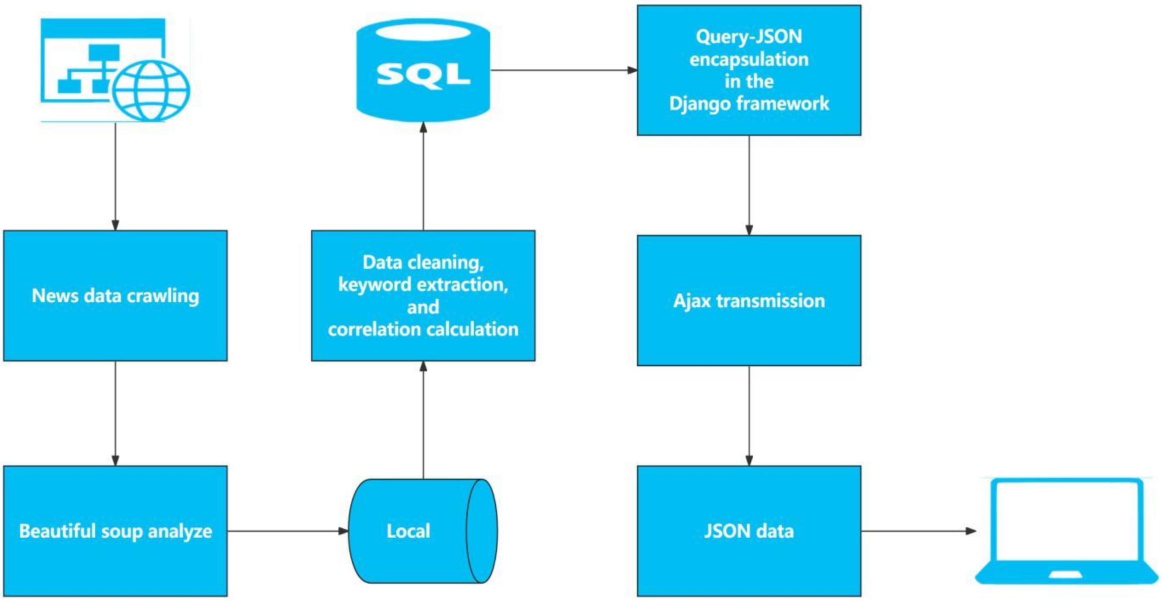
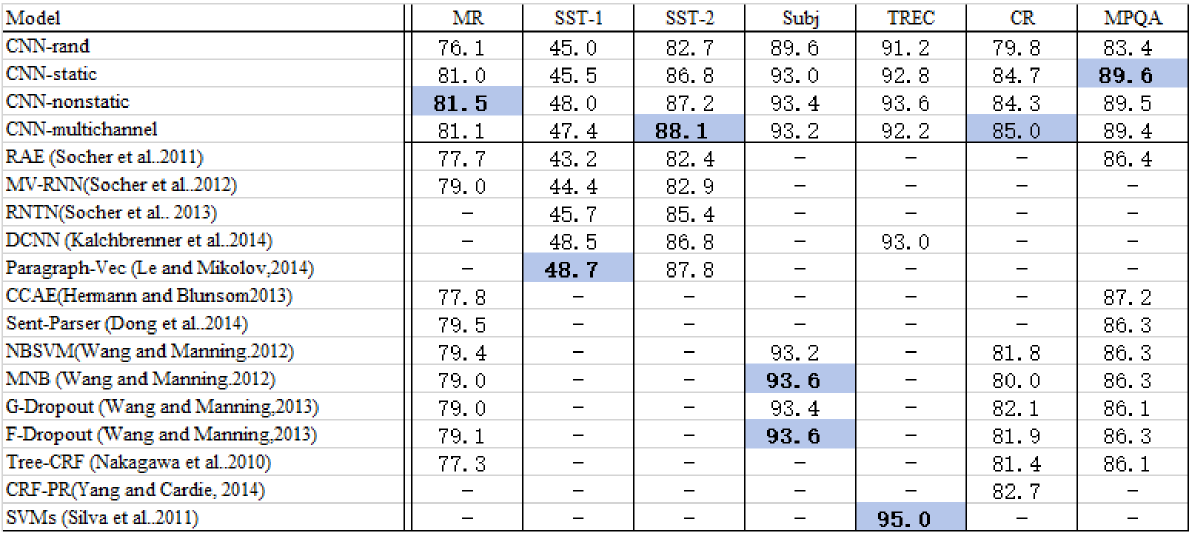

SQL
Query-JSON encapsulation in the Django framework
News data crawling
Data cleaning, keyword extraction, and correlation calculation
Ajax transmission
Beautiful soup analyze
Local
JSON data

Figure 6. Data processing flow.

3.3.2. Classification Tag Generation

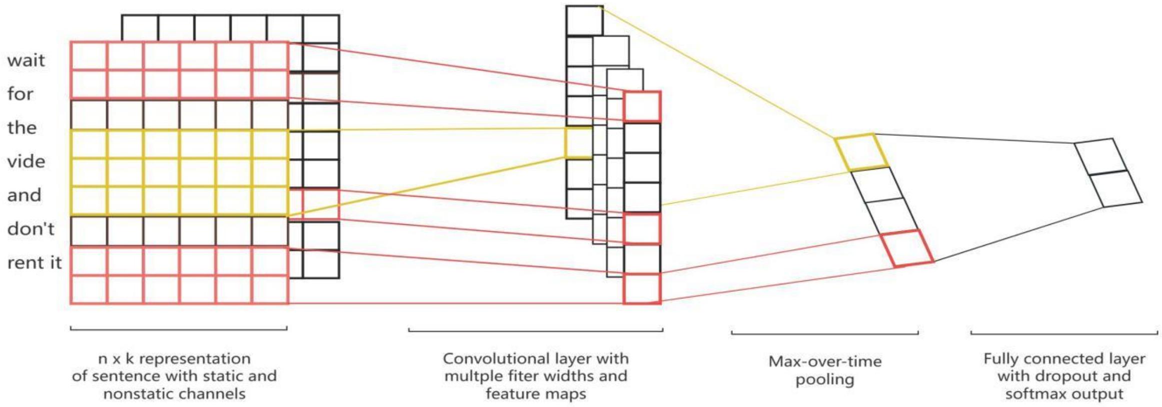
In the experiment involving classification tags, various methods were tried. Among these methods, the simplest was the tag searched according to the preset keywords, but the effect of this method was not ideal. The ideal effect is to generate the model through machine learning (Zhou, 2017) and deep learning (Saito, 2020) algorithms and then obtain the model with a higher generalization effect through convergence based on adjustment of the parameters of the model.



Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-nonstatic	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	85.0	89.4
RAE (Socher et al..2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN(Socher et al..2012)	79.0	44.4	82.9	–	–	–	–
RNTN(Socher et al.. 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al..2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov,2014)	–	**48.7**	87.8	–	–	–	–
CCAE(Hermann and Blunsom2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al..2014)	79.5	–	–	–	–	–	86.3
NBSVM(Wang and Manning.2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning.2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning,2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning,2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al..2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR(Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVMs (Silva et al..2011)	–	–	–	–	**95.0**	–	–

Figure 7. Comparison of running efficiencies of different model algorithms.

More classification algorithms (such as TextRank, the word vector model and the text classification model) and the existing open-source software were tried to generate the classification tags (Figure 7), and the training was also carried out on the datasets provided by open competitions. The algorithm follows the principle of continuously optimizing the results of training, reducing the errors of the algorithm and improving the algorithm’s generalization effect. In this process, the parameters of the model must be continuously debugged, and the iteration layer of the model must be reasonably improved. We optimized the parameters with the existing model, thereby achieving a better training effect. However, this model was obtained through a public dataset, which differed significantly from the original dataset, and it took a long time to obtain the model from the original training set; thus, we tried to improve the effect by optimizing the model while training.



wait
for
the
vide
and
don't
rent it
n x k representation of sentence with static and nonstatic channels
Convolutional layer with multple fiter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 8. CNN text classification algorithm structure Yoon Kim *Convolutional Neural Networks for Sentence Classification* (EMNLP 2014 Conference).

doi:10.20944/preprints202410.1364.v1

At present, the similarity analysis is conducted mainly by a classification tagging algorithm based on the TextRank algorithm, which can extract keywords and abstract from the text. This algorithm can split long sentences in the original abstract into several sentences, filter out deactivated words in the sentences and keep the words of the specified word class, thus obtaining the set corresponding to the sentences (Figure 8).

The keywords and abstracts split by TextRank are compared with the preanalyzed keywords. With the help of a text similarity algorithm, such as FastText model, the essays can be classified. Three groups of tags with high weight are acquired and saved in the corpus after the classified results are compared.

3.3.3. Corpus Organization

The crawled corpus data were classified and archived by tags. A separate data table was prepared for tags to store topics, such as politics, economy, society, culture, military, diplomacy, technology, people's livelihood and geography. If a new topic needed to be created, the topic field would be added by the administrator logging into the background management system. Three keywords, namely, topic, country and newspaper, were taken as the main retrieval points for corpus organization, and the retrieval navigation bar was set up by developers. The existing corpus data were separated, and a training set and a sample set were built at a 7:3 ratio to continuously regress the error and improve the recognition accuracy of the model. A search statistics system was constructed based on the specimen corpus obtained after machine training in each dimension, with the key time points as the horizon, to count the concentration of media coverage in each period. Coverage bias was displayed through word frequency, semantic web and other dimensions, and coverage with nine dimensions was drawn into a graph according to the results.

3.4. Cleaning and Storage of Corpus

The main function of the system is to automatically collect data on news websites through crawlers, generate corpus data in a standardized format and store them in the database. Many technologies were involved in generating the corpus, and the data from the news websites were presented mainly in the HyperText Markup Language (HTML) in the browser. Generally, the program simulated a browser to capture the HTML structure corpus of news websites while crawling news data. Upon obtaining the original website data, the original website should be analyzed manually. Each website adopts different technologies due to the complex original website data and the inconsistent format. For the simplest news websites, the data are displayed on the websites *via* HTML tags, while the data will be transmitted encrypted for some websites that adopt anticrawling measures. No matter how the data were crawled, as long as the information sources were different, the crawled data needed to be further processed. Conventional processing is to analyze the website format and extract the data content on the websites. However, data extracted by such a method usually contain special characters, hyperlinks and tags; consequently, the original corpus must be cleaned. In the automatic corpus cleaning stage, two main links were designed (Figure 9).

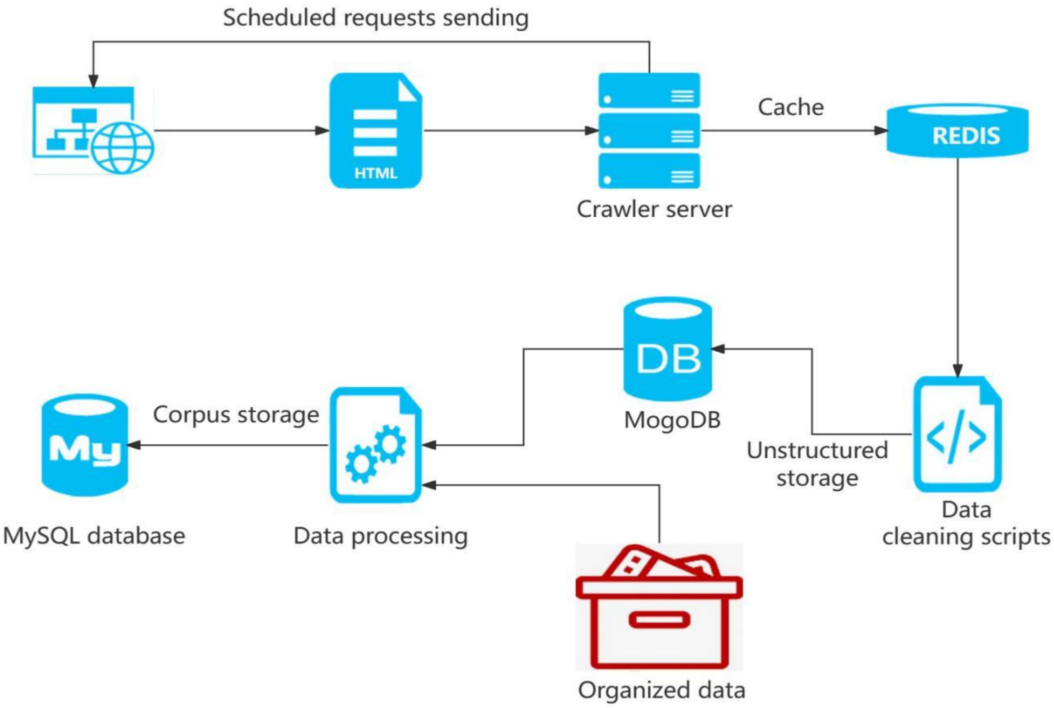


Figure 9. Corpus data processing flow.

3.4.1. Cleaning of Crawler Data

Most crawler systems contain a data cleaning process, which is to store the data in an unstructured database after the crawlers obtain the data and then to calibrate the data format; and clean, transliterate and dispose of the illegal characters, data, hyperlinks, tags, *etc*., from the website data. The data crawled was mainly in JavaScript Object Notation (JSON), which was stored directly in MongoDB. Redis conducted a data cache of the distributed queue, defined some data stored within the specified period and transferred the data into MongoDB. The numerous crawled data determined a high degree of concurrency of the program at this stage. The program operating results should prevail.

3.4.2. Calibration of Corpus Data

The crawled corpus data were cleaned and stored in MongoDB, and as MongoDB is structured as a column store and there are strict specifications in the format of the corpus, the read data must be processed by a special program during the transliteration process. In this paper, the regular expression tool of Python is used to generate corresponding corpus labels based on the *k-v* results of MongoDB and stored in the MySQL database. In addition, each piece of corpus data was saved by the content field in the news-table, which helped quicken the search.

3.4.3. Input of Existing Corpus Data

The corpus data crawled from the international mainstream media coverage on China from 2010 to 2019 were saved in .txt format. The corpus content specified in the txt file was extracted and stored in the corresponding database field by creating a separate file reading program. The method was to use regular expressions for overall recognition and to read document data for input in multiple threads.

3.5. Research and Development (R&D), Embedding and Operation of Corpus Analysis Software

3.5.1. R&D of Corpus Analysis Software

(1) Topic Recognition

To demonstrate the effect of text topic recognition clearly, the topic model was adopted to classify and reduce the dimensions of the texts. A topic model is a statistical one that clusters the implicit semantic structure of a corpus in an unsupervised learning manner. By comparing the existing topic models, the LDA2Vec model (a joint training topic model based on deep learning) was used to recognize the topic of this platform. Combining the advantages of latent Dirichlet allocation (LDA) global prediction and local prediction of the word vector model, LDA2Vec extends the skip-gram negative sampling loss function, jointly trains document vectors and word vectors to complete the context vector prediction of pivot words and obtains word vector representations and topic vector representations containing topic information. Additionally, this model can produce sparse interpretable document vectors, allowing for an easier understanding of the topic recognition results.

The specific process is as follows:

First, using the Word2Vec word vector model, the word vectors of words in the fusion document set were generated from the preprocessed corpus, which was input as part of the model. Python Gensim was used to encapsulate the Word2Vec model. The gensim.models.Word2Vec of Python Gensim was used to train the word vectors of the skip-gram model in Word2Vec.

Second, the topic-word distribution matrix and document weight were generated through the LDA model as another part of the input of LDA2Vec, and similarly, the preprocessed corpus was input to the LDA model as the dataset for training. There are encapsulations for the LDA model in toolkits, such as Gensim and Sklearn in Python. Considering the calculation of perplexity evaluation indicators in the later experiment, the LDA model was trained based on the Sklearn.

Finally, the obtained word vectors and document vectors were input into LDA2Vec for fusion training and topic extraction. The topics were ranked in descending order of probability of occurrence. Topic words showing the top 10 probabilities were selected under each topic to elucidate the implicit semantics of each topic more clearly and accurately. pyLDAvis was used to visually display the results of topic identification, allowing for more intuitive observation and analysis of hot topic results.

(2) Entity Recognition

As an information extraction technique, entity recognition can obtain entity data, such as person names and location names from text data. A named entity recognition method based on multiple features was adopted in this system to fully discover and utilize the contextual features and the internal features of the entity. Morphology includes the following situations: any character or word in the dictionary is in a separate category, while person names (Per), abbreviations of person names (Aper), location names (Loc), abbreviation of location names (Aloc), organization names (Org), time words (Tim) and number words (Num) are each defined as a separate category. The functions of morphological features and part-of-speech features were comprehensively utilized to establish entity recognition models according to the structural characteristics of different entities, which were superior in recognition performance and system efficiency.

(3) Keyword Extraction

The TextRank algorithm can be used to extract keywords and summaries (key sentences) from texts. As the Python implementation of the TextRank algorithm, TextRank4ZH can extract the summaries of Chinese and English articles and has been widely used because of its simplicity and effectiveness.

The TextRank4ZH algorithm split the original texts into sentences, filtered out stop words (optional) in each sentence and kept only words with specified parts of speech (optional), from which a collection of sentences and a collection of words could be obtained.

Each word acted as a node in PageRank. The window size was set as *k*, assuming that a sentence consists of the following words in turn:

w1, w2, w3, w4, w5, ..., wn

where w1, w2, ..., wk, w2, w3, wk+1, w3, w4, wk+2, *etc.* are all windows. There is an undirected and unweighted edge between the nodes corresponding to any two words in a window.

Based on the composition graph above, the importance of each word node can be calculated. The most important words can be used as keywords.

(4) Analysis of Text Similarity

To figure out the correlation degree between the text contents and the topics, the correlation was calculated with the text similarity calculation method.

The three currently most popular similarity calculation methods were analyzed as follows:

Editing Levenshtein distance: Although its computational complexity is high, it is outstanding in actual scenes and has the most accurate similarity calculation.

Cosine similarity: The computational complexity is high. However, since the data sparsity is too high in practical applications, the cosine similarity calculation will produce misleading results.

MinHash: Both SimHash and Minhash have sensitivity properties that are not found in general Hash methods (Hash LSH that is locally sensitive belongs to the Hash function), MinHash and SimHash will result in close Hash results of two similar documents.

After analysis and comparison, the editing Levenshtein distance algorithm performing best in the actual text was adopted to calculate the similarity.

3.5.2. R&D and Embedding of Corpus Analysis Software

The most widely used corpus analysis software includes Wordsmith and Antconc, which are often used as independent application software. However, the research and development (R&D) of this platform was based on the B/S architecture and deployed and used in the Linux environment. Since the compatibility between the platform system and these commercial software systems could not be solved, the R &D team decided to develop customized corpus statistics and analysis software through machine learning algorithms to realize the corpus analysis function. Based on the functions of existing commercial corpus analysis software, the following functions were developed and realized on the platform: the research, judgment and labeling of the collected corpus on the part of speech, syntax and concentration of coverage topics, the extraction of high-frequency words, collocated lexical chunks and keyword tables, statistics of type-token ratio, standard type-token ratio, average sentence length, structural capacity, visual analysis of emotional tendency and semantic prosody tendency, *etc.* At present, as the technology of natural language processing (NLP) based on deep learning (Goldberg, 2018; He, 2020) has been well developed, it is feasible to analyze and recognize multilingual part-of-speech and count word frequency through it. Moreover, it is more suitable for a platform based on Python language development, which makes preliminary preparations for further deep learning algorithms in the later stage.

3.5.3. Operation of Corpus Analysis Software

Based on the requirements of the machine learning algorithm experiment, the crawled corpus was preprocessed according to the bibliographic information, including news titles, summaries, keywords, authors and texts. Titles, summaries and texts were integrated as the basic corpus. The text was preprocessed with English stop words and customized stop words, and the missing values were processed with a regular matching method. In addition, the data in the dataset were divided into a training set and a test set at a ratio of 7:3 and were sent to the model training, with 70% for training the model and 30% for predictions. Then, the text processing process was improved based on the feedback of the topic modeling results. After the training, the machine can classify the text with a high concentration of a certain topic in politics, economy, military, culture, social and other topics for later corpus analysis.

3.6. Detection and Statistics of Coverage Concentration of Dimensions of China's Image and of Each Dimension

From the perspective of the topic model, the evolution of dimension of China's image and the concentration of each dimension in the China-related coverage in 10 international mainstream media platforms, including *The New York Times* (USA), *The Times* (UK) and *Asahi Shimbun* (JP), in the past 10

doi:10.20944/preprints202410.1364.v1

years were analyzed. The topic mining effect of each topic in the past ten years and the changes in topic intensity over time were explored, and hot topics and topics with rising and falling intensity were analyzed and evaluated. Furthermore, the hot topics of coverage in the dimension were found out by measuring topic intensity, and the topics with decreasing intensity were analyzed.

First, during the analysis and processing of the text content, stop words should be removed correctly to ensure the accuracy of the analysis results. Then, the LDA topic model was realized through Python-based Sklearn to train the optimal number of topics, and the text content was analyzed by year according to the optimal number of topics, with 2000 iterations. The topic intensity recognized in each natural year and the word probability under each topic were recorded, and topics with an intentional tag were labeled based on the words. Topic intensity over 0.02 was set as a hot topic, and hot events were recognized. The change in the intensity of the same topic between different years was measured to obtain topics of increasing intensity, that is, hot topics, and the topic trend was visualized. The topics that declined in intensity were analyzed in the same manner.

Given that the number of topics directly influences the effect of recognized topics, it is necessary to preset the number of topics *K* before constructing the LDA model. A low value of *K* would lead to broadly recognized topics, which does not reflect the core essence of the document, while a high value leads to the recognized topics being too detailed and unsystematic. The generalization ability of the model should be evaluated by effective means for analyzing topic evolution, namely, the well-recognized "perplexity indicators" (calculated from corpus classification tags and high-frequency topic words). The lower value indicated the stronger generalization ability of the model. Generally, when the *K* value in which the number of topics keeps increasing while the perplexity value drops marginally was selected, some obvious inflection points result. To reflect the degree of difference between topics and ensure the effect of topic recognition, an average topic similarity indicator was added to measure the generalization ability of the model. The number of topics *K* was determined by integrating the perplexity and average topic similarity indicators to ensure a stronger generalization ability of the model and thus improve the semantic effect of recognized topics.

There was no similar direct calculation method for the time characteristics of topics. The time characteristic was measured by the release time of coverage, because the coverage released on the same day contributes the greatest to the topic heat, while the contribution of past coverage decreases over time. By analyzing the time characteristics of hot topics and using the topic heat calculation method with customized thresholds, the topic heat was comprehensively quantified. The most popular topic was selected from the topic collection to display to the users. Additionally, according to the dynamic changes in the topic content and time, the development trend of topics could be traced, and those topics that were no longer popular were eliminated, ensuring the high quality of hot topic collections.

In terms of the evolution path of topic concentration, by calculating the topic similarity in adjacent periods and determining the changes in topics before and after adjacent periods, the evolution path of topic content was constructed. If there was a strong correlation between the local topics and the global topics, starting from the strongly correlated local topics, the forward topic and the backward topic query were performed consecutively to create the evolution path of the global topics. Without a strong correlation between the local and global topics, a local topic that had a high correlation with the global topic in each period was selected to construct a topic evolution path based on the correlation between local topics in adjacent periods.

3.7. Extraction and Analysis of Statistical Results

The system supervised training based on the provided corpus data. By continuing to iterate the number of cycles, the training effect was ensured to be better to improve the generalization effect of the program. In the experimental stage, combined with the deep model, various toolkits of NLP were used to allow the training effect to continue to regress. The ultimate goal was to control the convergence of the generalization results to a controllable range and to provide reliable results.

4. Conclusion

In this paper, the construction process of the dynamic monitoring and analysis platform for the international mainstream media corpus on China's image was introduced in terms of the three technical functions, namely, media corpus crawling, automatic storage and corpus analysis. With hot words as the search object and key time points as the horizon, the big data processing platform with dynamic monitoring can realize real-time corpus crawling of coverage on China from mainstream media in major developed and developing countries and continuously expand the corpus. The research on China's image in international media coverage based on this platform can compensate the shortcomings of previous studies in terms of quantification and diachronic analysis due to the limited amount of text. Relying on the powerful corpus capacity and automatic keyword retrieval function of the platform, the focus and diachronic changes in China-related coverage in international mainstream media can be dynamically monitored and grasped, and diachronic and synchronic changes and concentration tendencies of the various dimensions of China's image in developed and developing countries can also be tracked and examined continuously to improve the depth and breadth of future research. Since the value judgment and emotional choice of media readers' comprehensive impression of the country are deeply affected by the mirror image and construction of a national image in international media coverage, a full understanding of the international media attention areas and degrees about China and the image perception changes can provide the prerequisites for the Chinese government to adjust its diplomatic strategy promptly and actively exert its discourse power in the international arena. In the future, the research team will demonstrate the actual case of the dynamic monitoring platform, including the retrieval effect of keywords in specific fields, the operability of analyzing the focus and tendency of media coverage by using retrieval data reports, and the reliability of semantic analysis through platform-embedded corpus analysis software. The team will continue to improve and solve technical operation problems discovered during the actual use of the platform.

Disclosure statement: The authors report there are no competing interests to declare.

References

1. Adak, S., Chakraborty, S., Das, P., Das, M., Dash, A., Hazra, R., Mathew, B., Saha, P., Sarkar, S., & Mukherjee, A. 2022. Mining the online infosphere: A survey. *WIREs Data Mining and Knowledge Discovery, 12*(5), e1453.
2. Chaudhary, K., Alam, M., Al-Rakhami, M. S., et al. 2021. Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. *Journal of Big Data, 8*, 73.
3. Chen, J. S., & Chen, X. J. 2017. Changes in China's national image through the eyes of Western media—A corpus-based study. *Journal of Guangdong University of Foreign Studies, (9).*
4. Cui, Q. C. 2018. *Python3 Web Scraping Hands-On Development Practice*. Beijing: Posts & Telecom Press.
5. Gao, W. H., & Jia, M. M. 2016. Analysis of the reporting framework of China's multi-ethnic state image in American mainstream media. *Journalism University, (4).*
6. Goldberg, Y. 2018. Neural network methods in natural language processing. Beijing: China Machine Press.
7. Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing, 36*(1), 20-38.
8. He, H. 2020. *Introduction to natural language processing.* Beijing: Posts and Telecom Press.
9. He, Z. W., & Chen, X. X. 2012. The generalization of politics: The bias of American media in constructing China's image—A content analysis of articles on China in Newsweek (2009-2010). *Contemporary Communication, (1).*
10. Hu, K., & Li, X. 2022. The image of the Chinese government in the English translations of Report on the Work of the Government: A corpus-based study. *Asia Pacific Translation and Intercultural Studies, 9*(1), 6-25.
11. Hu, K. B., & Li, X. 2017. Corpus-based study of translation and China's image: Connotations and implications. *Foreign Language Research, (4).*
12. Hu, W. H., & Xu, Y. J. 2022. Twenty years of research on the image of China in the international media——a scientific knowledge mapping analysis based on CiteSpace. *Technology Enhanced Foreign Language Education, (4).*
13. Huan, C. P. 2023. China opportunity or China threat? A corpus-based study of China's image in Australian news discourse. *Social Semiotics.*
14. Huan, C. P., & Deng, M. M. 2021. Partners or predators? A corpus-based study of China's image in South African media. *African Journalism Studies, 42*(3), 34-50.

15. Huang, Y. X. 2019. *Development practice of Django Web*. Beijing: Tsinghua University Publishing House.
16. Karamouzas, D., Mademlis, I., & Pitas, I. 2022. Public opinion monitoring through collective semantic analysis of tweets. *Social Network Analysis and Mining, 12*, 91.
17. Lan, J., & Luo, R. 2013. British mainstream media's perception of China's image. *Hubei Social Sciences, (8).*
18. Li, J. 2017. A review of studies on China's image in Japanese mainstream media. *Communication and Copyright, (3).*
19. Nandwani, P., & Verma, R. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining, 11*, 81.
20. Pan, Y. Y., & Dong, D. 2017. Study on the discourse strategies of American mainstream news media in constructing China's image and major country relations: A case study of the coverage of the 2016 China-Russia joint military exercise. *Journal of Xi'an International Studies University, (3).*
21. Peng, Z. J. 2004. Representation of China: An across time analysis of coverage in the New York Times and Los Angeles Times. *Asian Journal of Communication, 14*(1).
22. Ramya, G. R., & Bagavathi Sivakumar, P. 2021. An incremental learning temporal influence model for identifying topical influencers on Twitter dataset. *Social Network Analysis and Mining, 11*, 27.
23. Saito, Y. 2020. *Deep learning from scratch.* Beijing: Posts and Telecom Press.
24. Shen, Y., & Wu, G. 2013. China's image in Russian regional media: A case study of reports by "Gubernia Daily", "Business World", and "Ural Politics Web". *Russian, East European & Central Asian Studies, (1).*
25. Tang, L. 2021. Transitive representations of China's image in the US mainstream newspapers: A corpus-based critical discourse analysis. *Journalism, 22*(3), 804-820.
26. Teo, P., & Xu, H. 2021. A comparative analysis of Chinese and American newspaper reports on China's Belt and Road Initiative. *Journalism Practice.*
27. Tsirakis, N., Poulopoulos, V., Tsantilas, P., & Varlamis, I. 2017. Large scale opinion mining for social, news and blog data. *Journal of Systems and Software, 127*, 237-248.
28. Wang, G. 2018. A corpus-assisted critical discourse analysis of news reporting on China's air pollution in the official Chinese English-language press. *Discourse & Communication, 12*(6), 645-662.
29. Wang, K. F. 2016. Construction of a novel diachronic multiple corpora. *Chinese Social Sciences Today.*
30. Wang, X. L., & Han, G. 2010. "Made in China" and national image communication—A content analysis of 30 years of reports in American mainstream media. *International Journalism, (9).*
31. Wang, Z. Q. 2009. The economic image of China from the perspective of Germany's Die Zeit (2004—2009). *German Studies, (4).*
32. Xia, F. 2012. The discursive construction of other countries' images—The image of China in The New York Times' reporting on the theft case in the Forbidden City. *Journalism and Communication, (15).*
33. Xu, M. H., & Wang, Z. Z. 2016. The "change" and "unchange" of China's image in Western media discourse. *Modern Communication, (12).*
34. Yang, H., & Van Gorp, B. 2023. A frame analysis of political-media discourse on the Belt and Road Initiative: Evidence from China, Australia, India, Japan, the United Kingdom, and the United States. *Cambridge Review of International Affairs, 36*(5), 625-651.
35. Zhang, K., & Chen, Y. L. 2014. Differential analysis of China's image construction in geopolitical conflict reporting: A case study of The Times and The New York Times' reporting on the Diaoyu Islands incident. *Contemporary Communication, (4).*
36. Zhang, K., & Wang, C. Y. 2017. Analysis of the national image model under the dimension of time and space—Based on the perspective of cognitive interaction. *Journalism and Communication, (5).*
37. Zhang, L. J., & Wu, D. 2017. Media representations of China: A comparison of China Daily and Financial Times in reporting on the Belt and Road Initiative. *Critical Arts, 31*(6), 29-43.
38. Zhang, J., & Cameron, G. T. 2003. China's agenda building and image polishing in the US: Assessing an international public relations campaign. *Public Relations Review, (1).*
39. Zhang, Y. 2011. Textual analysis of The New York Times' perception and shaping of China's national image during the Obama administration. *Journal of International Relations, (5).*
40. Zhong, X. 2013. Intertextuality and China's image: "Ye Shiwen discourse" in British national newspapers. *Modern Communication (Journal of Communication University of China), (6).*
41. Zhou, Y., & Zheng, M. 2010. Image China: The image of China in American mainstream newspapers. *International Journalism, (12).*
42. Zhou, Z. H. 2017. *Machine learning.* Beijing: Tsinghua University Publishing House.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.