Building A Natural Disaster Knowledge Base Expert System based on the Distributed and Incremental Crawling Technology

To cite this article: Qiujin Chen et al 2020 IOP Conf. Ser.: Earth Environ. Sci. 435 012024

View the article online for updates and enhancements.
Building A Natural Disaster Knowledge Base Expert System based on the Distributed and Incremental Crawling Technology

Qiujin Chen¹,a, Yuanrong He¹,b, QunSu¹ and Tingting He²

¹Big Data Institute of Digital Natural Disaster Monitoring in Fujian, Fujian Xiamen 361024;
²The Civil Engineering and Architectural Association of Fujian, Fujian Xiamen 361022;

¹892725709@qq.com; bheyuanrong@126.com

Abstract. China is one of the countries with the most natural disasters in the world. Natural disasters seriously endanger people's safety and economic situations. Therefore, it is very necessary to strengthen the prediction and research on the characteristics and the causes of various natural disasters. Based on the Internet distributed and incremental crawling technology, this system crawls the knowledge literatures, disaster news and professional disaster data from domestic and overseas on natural disasters, as well as drawing on the advanced knowledge management concept, the captured data will be displayed in the form of knowledge tree. Integrating the authoritative global typhoon live maps and orthophotos which acquired by UAV to better monitor and predict typhoons; This system determined that the SSM framework is used as the system framework, combined MySQL and Redis database to store data stably and efficiently. In addition, the Echarts component can be used to analyze and visualize the crawled data. Big data analysis techniques such as knowledge graphs and related algorithms for literature analysis, data cleaning to better data store and display. In the end, it will enable the researchers to more easily and efficiently carry out the knowledge sharing, collaborative cooperation and interactive communication, which has the certain research significance and use value.

1. Introduction
With the global climate has changed and the environment has damaged, the various disaster events occur frequently. China is the country with the most natural disasters in the world [1], which is affecting about 200 million people every year. Therefore, research on the causes, predictions, and trends of disasters is of great significance and urgent need. The monitoring and evaluation of natural disasters needs to be based on the accumulation of rich and reliable historical background disaster data[2]. With the advent of the "Internet informationization" era, today's technology level is booming, how to quickly obtain data related to natural disasters and facilitate research for the scientific research institutions from the massive data of the Internet is still worthy of further research, so that to better serve the disaster prevention and mitigation work. In recent years, massive data and natural language processing have increasingly become important research directions in the field of computer science, deep learning and artificial intelligence. Foreign research on knowledge base was first carried out on the basis of research expert decision system. With the maturity and development of artificial
intelligence technology application, the research on the knowledge base has been further developed[3], the knowledge base is an important collection of knowledge for storing, organizing and processing knowledge as well as providing the knowledge services[4]. The construction of the knowledge base could accelerate the flow of knowledge and information, promote the sharing and exchange of knowledge, and facilitate the collaboration and communication between knowledge users and organizations. Knowledge graph is a knowledge base which used by Google to enhance its search engine function. Using the knowledge graph form analysis to study natural disaster research hotspots, trends, research institutions, etc. can help people grasp the frontier and situation of current research with an intuitive and unique perspective [5].

In this background, in order to improve the efficiency of browsing, accessing the relevant news and literature information, and realize the research analysis and prediction of the status quo of natural disasters, as well as the construction and sharing of various professional knowledge in the scientific research institutions, research institutes, and enterprises. The natural disasters knowledge base system comprehensively used the Internet crawler technology, network development technology, database management technology, data analysis technology and knowledge base related algorithms. On the one hand, it can monitor disaster data from all over the world in a timely and real-time manner, and create a more advanced laboratory research environment for researchers which was focused on research in the field of natural disasters, improve the efficiency of laboratory researchers, increase the knowledge accumulation of users, and the work synergy of the knowledge base; On the other hand, it can also enable users to participate in the disaster assessment work to better serve the disaster research.

2. Research status at home and abroad
Since the first expert system DENDRAL[6] was introduced at Stanford University in the United States in 1965, through 20 years of research and development until the mid-1980s, various expert systems have been spread across various professional fields, which has greatly promoted the development of artificial intelligence applications; From the early 1980s to the early 1990s, the expert system experienced a golden period of ten years; In the early 1980s, Bonczek et al.[7] proposed a preliminary model combining the decision support system with the expert system as an intelligent decision support system(IDSS); Then introduced the artificial intelligence technology into the field of geosciences, GIS adopts the empirical model of expert system[8]; The development of GIS to decision support systems is the Space Support System (SDSS). At present, the expert system has become one of the three frontiers in the field of artificial intelligence development. It is generating a new discipline—knowledge engineering. At the 5th International AI Conference in 1977, Professor Feigenbaum proposed the concept of “knowledge engineering”. In China, as early as 1986, there was a research on the knowledge base of meteorological experts. However, there are not many data on natural disasters of system specifications. Therefore, there is no professional knowledge base of disaster experts. As early as 2003, Hong Kong University of Science and Technology began to build the earliest IR(Institutional Knowledge Base) in China. IR is a natural extension of the digital library and the development of knowledge base applications. In 2016, Zhang Bin et al.[9] concluded that the current research at home and abroad mainly focuses on the construction of knowledge base, knowledge extraction, knowledge indexing, knowledge representation, knowledge organization, knowledge retrieval, knowledge discovery and knowledge base integration. Compared with the domestic research base of knowledge base, the research subject is more diversified, the practice details are more specific, and the knowledge base theory, technology and tools are more mature in research depth. In 1996, Grigorios Antoniou[10] explored some of the anomalies in the non-monotonic knowledge base, such as redundancy, contradictions, and lack of knowledge. He believed that some traditional verification methods could be used to detect some anomalies in the non-monotonic knowledge base; In 2002, R. Crow[11] firstly proposed the concept of institutional knowledge base. In the same year, the United States established the first institutional knowledge base. In 2003, Richard C. Hicks[12] proposed the development and verification of knowledge base management system for business rules and expert systems. It provides a complete lifecycle environment, and builds an expert
system based on the research of life cycle and framework structure; For the research of the knowledge base model, Zou Junhua[13] proposed to use the tree structure to organize and represent knowledge in 2009; The International Open Access Knowledge Base(COAR) serves as a global open access knowledge platform. In 2016, the “Next Generation Institutional Knowledge Base” project was launched to build an institutional knowledge base into a distributed academic exchange platform based on a global network infrastructure.

The emergence of the web crawler technology has greatly improved the collection speed and efficiency of information data. The first crawler was born in 1993 and it was written by Massachusetts Institute of Technology student Ma Hugh Gray[14], which provided a solid foundation for the development of search engines; In 1998, Sergey and Lawrence[15] used a large-scale search engine as an example to analyze the mechanism of Google's indexing of web pages on massive web pages, and analyzed the use of link analysis algorithm PageRank as a user behavior model to generate better results in search engines; Domestic research on crawler technology, such as in 2014 Yang Yang[16] and others based on the web crawler technology to realize a system of intelligent search for literature resources and grasping the key information functions, and used the ontology method to classify and identify the crawled information; In 2019, Wang Lei and Liu Xiaodan [17] studied the web crawler system framework which was based on the Scrapy, as well as the MongoDB and Redis databases inside the crawler system respectively store the data crawled by the crawler to ensure the speed and quality of the information.

3. System design
The construction of the natural disaster knowledge base expert system is designed to help researchers understand the recent disaster information and analyze the disaster data so that the relevant prevention plans and forecasts can be made. And a natural disaster big data management and analysis monitoring system will be established. It is aimed at providing the collaboration service system and the file data storage sharing platform, which integrates the multi-source data such as remote sensing, geospatial, UAV, GNSS, 3D point cloud, social network, etc., and realized the collection and archiving of professional data through the cloud computing, big data distributed storage technology.

3.1. The design of system framework
SSM framework set, which was the system's overall framework. It integrates several of the most powerful frameworks in the current web development environment, Spring, Spring MVC, and MyBatis. The page sends a request to the Controller, the Controller calls the business layer to process the logic, and the logical layer sends a request to the persistence layer, the persistence layer interacts with the database, and then returns the result to the business layer, the business layer sends the processing logic to the controller, which in turn calls the view to present the data in the presentation layer. The specific work logic is shown in figure 1:

![Figure 1. The SSM framework set working logic](image)

(1) Spring is an open source and lightweight application development framework, which was
designed to simplify the application development and reduced the inter-program coupling. Spring's core ideas are IOC (control inversion) and AOP (face-oriented). The principle of IOC is that the programmer no need to explicitly create an object, the bean will be injected into the corresponding class in a passive form when needed, thus implementing the corresponding logic; AOP can supervise and control a certain type of object to achieve the function of a certain module expansion. Spring not only creates and maintains the dependencies relationships between objects, reduces the degree of coupling between objects in order to maintain code, but also encapsulates and simplifies the common APIs.(2)Spring MVC is a lightweight web framework based on Java's request-driven type of MVC design pattern, which uses the MVC idea to decouple the WEB layer. The specific business logic execution flow is: the browser sends a request to the controller, the controller receives the request to delegate the request to the business object for processing, and then returns the result to the browser by using the view rendering.(3)MyBatis is a persistence layer framework based on Java, which was originated from Apache's open source project iBatis, which is more lightweight than traditional hibernate. The persistence layer framework which was provided by iBtis includes SQL Maps and Data Access Objects(DAO).MyBatis's JDBC encapsulation makes the underlying operations of the database transparent. Its advantages are mainly two points: First, it can freely control SQL to improve the access efficiency of the database; Second, the use of xml configuration to organize the management of SQL for easy expansion and post-maintenance optimization.

3.2. The technical route of system
The implementation process of the specific technical route of the system is as follow: Firstly, The system develops the authoritative and professional websites (Baidu News, Zhihu News, WeChat. News, Sogou News, Sina News, Phoenix News,etc., and the authoritative academic websites such as CNKI, Weipu Journal, Wanfang Data, Research Gate and China Science and Technology Papers Online, etc.) to be crawled as the data source; Sencondly, building a distributed server crawld system for distributed and incremental crawling of Internet-related data, and setting up the keywords related to natural disasters (such as earthquakes, tsunamis, typhoons, floods, droughts, etc.) for automatic crawling, and the API gateway is built to update the date conveniently and quickly. The data what obtained by the crawler technology was initially de-duplicate. The Redis database stores the data efficiently, then performs the resolving content, cleaning, text extraction and other operations, and the obtained metadata was translated, classified, and semantically understood. Finally, the structured data was stored in the My SQL database. At the same time, the keywords are associated with each nodes of the knowledge tree, and the structured data can be displayed in the system through the knowledge tree.

3.3. The key technologies

3.3.1 The distributed and incremental crawling techniques. Distributed crawler technology, that is, the multiple crawlers are deployed on the multiple hosts at the same time. The crawler task multi-threaded to perform the collaborative crawling to increase the concurrency. The prerequisite for the collaborative crawling is the shared crawl queue. The key to distributed crawlers in the system is loading balancing. The most critical function of the distributed crawler system is "concurrent crawling". It is very important to maintain the crawler queue for sharing. Usually, the Redis database first-in first-out mode is selected to implement the random sorting and the crawl queue is not repeated. Incremental crawling collects information intermittently and re-crawls for updating data after a certain period of time[18].Incremental crawlers are in the process of continuous crawling, and regularly update the crawled web pages to reflect the changes in web pages in a timely manner. The most important task of the incremental crawler is the "judging the de-duplication". The key steps of the work are:(1)Before the request is sent, it is judged whether the URL has been crawled;(2)After parsing the content, determining whether the content has been crawled;(3)When writing to the storage medium, it is judged whether the content already exists in the medium.
3.3.2 The combination storage of MySQL and Redis database. Considering that the massive data related to natural disasters which were crawled from the Internet is prone to data duplication and redundancy, this system integrates the advantages of both relational database and cache database to complete the design of the database, thereby improving the efficiency and quality of users' access to data information. This system creates several tables in the MySQL database to store the relevant attribute information of the crawled news, documents and other data information. MySQL database is limited to a fixed table structure, which is not easy to extend, and the query efficiency is susceptible. Redis is mainly used to build a high-performance in-memory database. Its high-speed access speed is suitable for the knowledge base expert system, which requires a large number of data browsing queries. A relational database can be used to store data for a long time and stable storage. The cache database can store the initial URL with a large amount of data[19]. The specific storage rule is to request data directly from the Redis database when the expert system initiates the request, and realizing all data related to natural disasters in a very short time (several milliseconds) and submit it to the expert system for processing, and finally store the valid data information in MySQL database.

3.3.3 The data cleaning and de-duplication. It need to clean and deduplicate these data, thus reducing the phenomena such as data redundancy and invalidation. The easiest way to deduplicate is to save all the visited URLs and their corresponding content, then re-crawl and compare them after a period of time, and then decide whether you need to overwrite. This method will consume a lot of resources, it is obviously not practical. Usually the method of deduplication is to give the URL or its content an identifier, that is, the data fingerprint. When the amount of data is not large (several hundred or thousands), use the characteristics of small functions or collections to achieve the goal of deduplication. If the amount of data is large enough, the value of setting the data fingerprint (identification) is reflected in the space saving, and the Bloom Filter algorithm can be introduced as a means of deduplication. Bloom Filter is a typical and efficient space-time random data structure that is often used to monitor whether a URL has been accessed in a web crawler. The Bloom filter can be thought of as an extension to the bit-map. The Bloom filter is designed as a bit array of elements with N, initially with all bits set to zero. The concrete idea is that when an element is added to a collection, the element are set to 1. While searching it only need to be seen if these points are all 1 (about) that would know if there is any in the collection; If these points have 0, the checked elements must not be there; If they are all 1, the detected elements are likely to exist.

4. System implementation
The detailed functional module implementation of the system as shown in figure 2:

Figure 2. The main function structure diagram of this system
4.1. The implementation of front-end of the system

The front end of the system is intended for users including experts, scholars, researchers, teachers and students. In the query information module, this system displays the structured data such as the titles of the literature, keywords, journal name, publication time, download volume, citation volume and the original website. Users have the functions such as viewing, downloading, and playing the article summaries. Users can search for news and professional literature knowledge by searching for the keywords related to natural disasters or by selecting the different sources of information. As shown in figure 3(a) and (b), there are the literature and news information module of the user side respectively.

![Figure 3. The query information function module](image)

In the typhoon information module, the information which was crawled from the global typhoon monitoring website was stored in the local database, as well as structured display in the system, which specifically showing the typhoon number, name, time of occurrence, top time, duration, original website and other information. The users can understand the typhoon live situation and authoritative information in all aspects from the original website. The typhoon module also displays the orthophoto maps of typhoon disasters which were collected by UAV these years, so that users can understand the typhoon information and disasters situation more intuitively and detailed; This module can realize the statistics of disasters damage in various areas after typhoon(including fallen trees, fallen street lights, collapsed houses, etc.), display them in the form of a bar graph or a line chart and the results are output in a table Excel or text TXT format. This module provides the researchers with real and reliable typhoon information data. The typhoon data display interface is as shown in figure 4(a), and figure 4(b) shows the local information interface of the typhoon module.

![Figure 4. The typhoon live situation module](image)

One of the most intuitive functions which were provided by the system is that the extraction, analysis and structured the massive crawled data and visualizes it to the users. By using Echarts component technology, the results of data analysis are visualized and delivered to the users intuitive and easy-to-interactive information in the form of charts (columns, line charts, pie charts, etc.). The visual analysis interface of the system mainly includes two aspects: One is to display the total amount of literature and the author's publication situation in the field of natural disasters in recent years in the form of a histogram/line chart/pie chart; The second is to show the search situation of keywords in the field of natural disasters through the knowledge graph, as well as the co-occurrence of the organization and the author. As shown in figure 5(a) and (b), there are the data analysis visualization interface and the author co-occurrence diagram respectively.
4.2 The implementation of system background

The system background can be seen as a knowledge base management and resource monitoring system within the organization (colleges, research institutes, enterprises). The system administrator has five main management functions: The management of data resources, search conditions, crawler engine, each user's information and permissions, and various log information of the system. The system administrator mainly manages two kinds of data information: The news information and literature materials, the administrator can manually "hide" the invalid or repeatedly crawled data. As well as “favorites” the important articles to display in red font for viewing conveniently. The system is mainly searched by the keywords and retrieved by the knowledge trees. The advantage of retrieving data by keywords is search refinement, fine-grained search, narrowing the search range, and thus accurately and efficiently retrieving the accurate data. The administrators can not only set, add and delete the keywords, but also add and delete the existing knowledge tree roots and nodes. It can also import a custom complete knowledge tree with Excel format, and assign the different keywords to different knowledge tree nodes, so that the keywords and knowledge trees can be accurately matched. This function indicates that the system can customize the knowledge tree of different professional fields according to the users’ requirements, thus implementing the customized design knowledge base expert system, which fully reflects the flexibility and scalability of this system.

The crawl engine management module contains three sub-functions: The force crawl management, the days settings of crawling and the real-time monitoring of crawling. The function module can realize the function of monitoring and querying the core business logic of the expert system operation, thereby ensuring the visualization and the monitorable management of the entire process of crawling the data to the display, so that the system administrator can find the problem and maintain the system. Thereby improving the security, reliability and maintainability of the entire knowledge base expert system.

In the user management module, the system administrator strictly controls the access rights of various roles. The ordinary users(students) have functions such as browsing and downloading, and voice playback the summaries of news and literature; The advanced users (professors, teachers, etc.) also have the function of setting keywords and custom the knowledge trees according to their own research directions; The administrator owned all permissions, in addition, they can set the crawl time and tasks, edit the users’ information and view the system work logs. The system administrator can view the users’ IP address and login time for logging in to the system at different times. This information allows administrators to understand the usage frequency of this system, the time of use, the common IP and other information, the administrators can collect these data to analyze the knowledge base that is custom-tailored for the common users.

5. Conclusion and prospect

This system is convenient for users to quickly view the whole, as well as browse, download, manage
and analyze the disaster data; It can also provide the disaster emergency departments with tools to collect and organize the disaster information, and realizing the knowledge sharing, knowledge transferring and collaboration within the research institutions. In addition, there are still some problems in the construction of this system, which mainly reflects in the two aspects of data mass storage pressure and effective reliability of data[20]. On the one hand, crawling data through the distributed server and the incremental crawling technology, the crawler program starts the service continuously at the regular intervals, which will cause the database to store massive data pressure phenomenon; On the other hand, due to the richness of Internet resources, only by setting "keywords" to crawl the relevant data will appear repeated, redundant and even the irrelevant data. In response to the above two major problems, the system has also developed the emergency response plans, such as the use of high-performance equipment, backup servers, database backup, data backup; and manual de-duplication, algorithm de-duplication and other methods.

In summary, the natural disaster knowledge base expert system has realized the functions of intelligent knowledge base management information and document information, and it can analyze and refine the existing professional knowledge data. In the future, we will further optimize and upgrade the system on this basis, knowledge reasoning. The artificial intelligence prediction of disasters needs the further research. The knowledge reasoning is conducive to scientific research personnel to use knowledge to predict and avoid disasters.

Acknowledgments:
This research was supported by the Open Fund Programs of Big Data Institute of Digital Natural Disaster Monitoring in Fujian (grant no. NDMBD2018004) And as well as supported by the Education and Scientific Research Project (Type A) for Young and Middle-aged Teachers of Fujian Province (Project No.JAT160360).

References
[1] MA Rongyong, ZHANG Donglong. Risk assessment of flood disasters in Nanning City[J]. Journal of Natural Disasters, 2017,26(3):200-206.
[2] He Jianbang, Tian Guoliang. Construction and Application of Information System for Monitoring and Evaluating Major Natural Disasters in China[J]. Natural Disaster Journal, 1996, 5(03): 1-7.
[3] Lin Na. Exploration of the construction model of enterprise knowledge base [J]. Science and Technology Information, 2018 (17): 130-131.
[4] Zhang Bin, Wei Biao, Hao Qi. Review and Comparison of the Research Status of Knowledge Bases at Home and Abroad[J]. Library and Information Science, 2016(3): 15-25.
[5] Wang Ping, Liu Tao, Du Ping, et al. Knowledge Mapping of China's Disaster Risk Research from 2000 to 2017[J]. Journal of Natural Disasters, 2019, 28(04): 169-177.
[6] Yin Wensheng. Expert System for Problem Analysis and Decision Making[J]. Journal of Computer Applications, 2008(12): 3645-3649.
[7] Zhang Jianting, Qiu Youliang. A review of the application of artificial intelligence and expert systems in geosciences[J]. Progress in Geography, 1998(01): 44-51.
[8] Shu Huicui. The role of artificial intelligence in decision support systems and expert systems[J]. Computer Knowledge and Technology, 2009, 06(18): 4897.
[9] Zhang Bin, Wei Biao, Hao Qi. Review and Comparison of Research Status of Knowledge Base at Home and Abroad [J]. Library and Information Knowledge, 2016(3): 15-25.
[10] Grigoris Antoniou. Integrity and rule checking in nonmonotonic knowledge bases[J]. Knowledge-Based Systems, 1996(9): 301-306.
[11] Wang Haihua. Analysis of the development status and trend of institutional knowledge base [J]. Gansu Science and Technology, 2019, 13 (07): 73-75.
[12] Richard C. Hicks. Knowledge base management systems - tools for creating verified intelligent systems [J]. Knowledge-Based Systems, 2003 (16): 165-171.
[13] Zou Junhua, Liu Qingtang. Knowledge Element Model for Knowledge Abstract and Fusion...
[14] Zong Jingxin. Design and Implementation of High Performance Web Crawler System [J]. Communication World, 2019 (04): 78-79.

[15] Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual web search engine [J]. Computer Networks and ISDN Systems, 1998: 107-117.

[16] Yang Yang, Li Xiaofeng, Zhao He. Research and Implementation of Literature Retrieval System Based on Web Crawler[J]. Computer Technology and Development, 2014(11): 35-38.

[17] Wang Lei, Liu Xiaodan. Design and Implementation of Scrapy-based Web Crawler System Framework[J]. Microcomputer Applications, 2019(04):48-50.

[18] Feng Fei. Design and implementation of distributed web crawler system based on Python[J]. Electronic Technology and Software Engineering, 2018, 23:6-6.

[19] Chen Meng. Design and implementation of Sybase-based Sina News crawler system [J]. Modern Information Technology, 2018, 07: 111-112.

[20] Dong Xiang, Xiao Lanxi, Du Xianzong et. Network-based analysis and processing system for seismic disaster collection in Shandong Province[J]. North China Earthquake Sciences, 2007(03):6-10.