Machine Learning and Artificial Intelligence
A.J. Tallón-Ballesteros and C. Chen (Eds.)

doi:10.3233/FAIA200774

Research on the Construction of Cutting Edge Technology Monitoring System Based on Multi-Source Heterogeneous Data

Qiang XIAO[1], Siming TAN and Shengfeng YU
Qingdao Institute of Science and Technology Information, Qingdao, China

Abstract. With the rapid development of the Internet and big data, the data resources in various industries and technical fields are constantly emerging and growing. How to monitor and identify the effective data information in the massive big data has become one of the key contents of the current scientific and technological information work. This paper designs and implements the advanced technology monitoring system based on multi-source heterogeneous data. It comprehensively uses information collection technology, database technology and big data mining technology to realize the accurate monitoring, acquisition and analysis of multi-source heterogeneous data. It reveals the coupling relationship of technologies, people and institutions in different fields and the future technology development trend, and finally visualizes in various states It provides a reference for the strategic decision-making of relevant government departments, and provides efficient and convenient research tools and methods for scientific research institutes and enterprises.

Keywords. Cutting edge technology, multi-source heterogeneous data, monitoring, system, construction

1. Introduction

Cutting edge technology is a leading and exploratory major technology in the core or key technology field. It represents the latest development trend of high-tech in the world. It not only plays an important leading role in the formation and development of emerging industries in the country and even the world in the future, but also plays a positive role in promoting the technological upgrading and R & D infrastructure construction of enterprises.

With the rapid development of science and technology, a new round of industrial technology revolution will be triggered in the fields of information, manufacturing, biology, new materials and energy. In the national science and technology innovation plan of the 13th five-year plan, China has clearly pointed out that it is necessary to strengthen the early warning of the trend of industrial change and major technologies, strengthen the prediction of the turning point of disruptive technology replacing

[1] Corresponding Author: Qiang Xiao, Qingdao Institute of Science and Technology Information, Qingdao, China; E-mail: qd82898286@163.com.

traditional industries, and timely lay out the research and development of frontier technologies in emerging industries. Therefore, it has become the focus of scientific and technological workers and managers to quickly and accurately understand and master the state-of-the-art technology situation, so as to help the government and enterprises to formulate science and technology development strategy, and improve the national high-tech research and development ability and the international competitiveness of the industry.

Some government departments or research institutions in some countries have established corresponding strategic decision support systems to provide intellectual support for their competitiveness and sustainable development in some technical fields. The U.S. Department of defense has established a number of information technology analysis centers (IACS) to provide information analysis services for managers and decision makers of the Department of defense by using databases and intelligence analysis tools. Japan's Nomura institute takes consulting and knowledge services, system integration services and decision-making services for the government as its core business, and takes the construction of information technology service platform as an important means to provide high-quality services to users. The British Institute of international strategy also has a corresponding Intelligence Analysis Department [1].

In recent years, scholars have begun to pay attention to data sources and identification methods of cutting-edge technologies. For example, scientific and technological media data are used as data sources of frontier technology identification, and neural network model is used in citation analysis process. This study holds that the biggest characteristic of Internet information is that there are many kinds of data sources and rich contents, and the data are not limited by countries and geographical space. These characteristics make the acquisition of data more convenient, and the data objects used for the monitoring of frontier technologies are more multi-source, which is more conducive to improving the accuracy of monitoring results of frontier technologies. The corresponding data analysis methods need to be more in-depth data fusion and systematic analysis methods [2].

In this study, we propose and design a cutting-edge technology monitoring system, which is a research method and tool platform. Using network information collection technology, database technology and data mining technology, it provides research services such as information automatic collection, data depth calculation, knowledge intelligent discovery and visualization, and realizes real-time monitoring, automatic collection and automatic collection of the latest technology trends and scientific research achievements Intelligent analysis can accurately grasp or predict the development trend of frontier technology in specific fields, study and judge its development path, provide decision-making reference for the government's scientific research plan and major project layout in specific fields, and provide research methods and tools for scientific research institutes and enterprises to carry out scientific research.

2. Structure Design of Monitoring System

The overall structure design of advanced technology monitoring system is divided into three layers: data acquisition layer, analysis storage layer and application display layer. As shown in the figure 1 as below:

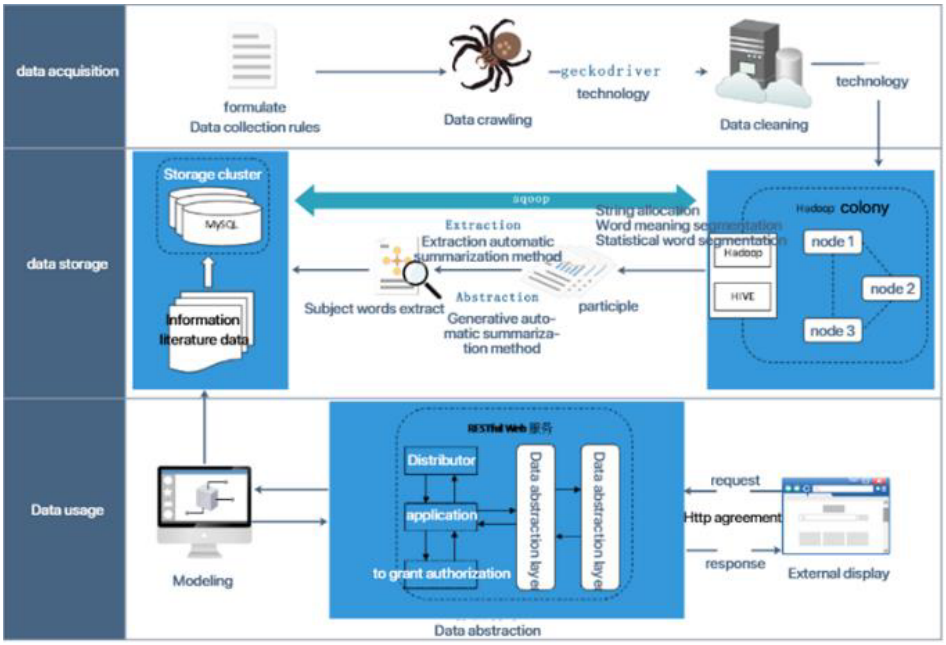

Figure 1. Overall design of the system.

2.1. Data Acquisition Layer

The data acquisition of this system should meet three characteristics: comprehensiveness, multidimensional and high efficiency. The system uses distributed web crawler, and the bottom layer of crawler is geckodriver technology. Geckodiver can complete remote web content crawling. The system can collect dynamic data related to cutting-edge technology according to the set rules.

Data cleaning includes: 1) data De duplication: only one copy of the same paper or patent obtained from different data sources can be retained. The paper title and the author can be used to judge whether the paper is duplicate, and the patent title and the inventor / inventor can be used to judge whether the patent is repetitive; 2) data filling: in the process of data De duplication, if it is found that the duplicate data of subsequent papers / patents contains the previous data (3) data standardization: if the abbreviation or nonstandard elements in the paper / patent data are found, standardization shall be carried out. For example, if the address in China is written as Chi, it shall be standardized as China; if the unit abbreviation is abbreviated, it shall be changed to the full name.

2.2. Analysis Storage Layer

The system needs to extract effective information from multi-source heterogeneous data. The entity extraction method based on the deep learning model BilSTM-CRF is used to add the manually extracted features such as word segmentation, indexing and word frequency statistics into the model to describe the semantic information of each Chinese character more fully, and extract new technical keywords from the technical dynamic data as far as possible In the future, which technologies may become hot spots provide

data support. The dynamic data storage engine of the system implements permanent storage of the effective data collected by the web crawler, which can be used for various needs analysis of the system or shared to the third party. At the same time, the data storage should consider the periodic update of the data.

2.3. Application Display Layer

This system uses RESTful architecture to realize the access of display chart to statistical analysis data. In this way, the complexity of the system and the coupling degree of the system can be reduced.

This system has statistical analysis algorithm library, all of which can be independent of the line algorithm program, and the algorithm library records the meaning of each statistical analysis algorithm and the executable SQL statement. The client uses the general web request such as GET/ POST to visit the statistical analysis algorithm program based on HTTP protocol, and the server receives the request from the display side through the distributor Parameters automatically forward the request to the corresponding statistical analysis algorithm program in the algorithm library; the statistical analysis algorithm program calls the SQL statements stored in the database, and then executes to obtain the statistical analysis data. After obtaining the data, the statistical analysis results are submitted to the format converter, and the format converter converts the statistical analysis result set into JSON format and then transmits it to the display terminal. The application display end can use echarts graphic display component, which contains rich graphics such as fishbone diagram, hot word graph, relationship diagram and so on. After receiving the data in JSON format, echarts displays the conclusion through the specified graph.

3. Function Design of Monitoring System

The function design of advanced technology tracking and monitoring system is divided into four parts: (1) technology dynamic analysis; (2) expert dynamic analysis; (3) mechanism dynamic analysis; (4) large screen visualization centralized display.

3.1. Technical Dynamic Analysis

First of all, the dynamic analysis of technology is to display the dynamic information of the latest technology development in a specific field, and at the same time, it should have the translation function for the collected foreign language information. Secondly, it analyzes the development process of technology in specific fields, excavates important node figures and events, and displays the development trend of technology in various forms. Then there is dynamic information analysis, which is used to show the geographical distribution, source type and ranking of all collected information. The last is hot word analysis. Through clustering analysis of hot words (or key words), a high-frequency hot word atlas is formed, and the correlation between technical dynamic information and hot words is established, and relevant information such as hot word retrieval technology, organization and expert dynamics is realized.

3.2. Dynamic Analysis of Experts

According to the matching degree of the given domain keywords, the domain expert database is formed. According to the data collection, the basic information of the experts' institutions, research hotspots, technical cooperation and academic achievements is displayed. At the same time, according to information sources, journals, patents and other information, the distribution of key experts in countries, industries and disciplines is analyzed.

3.3. Dynamic Analysis of Mechanism

According to the matching degree of keywords in the given field, the Organization database can be classified according to the attributes of universities, scientific research institutes and enterprises. Display the research direction, research hotspot, key experts, cooperation network and research dynamic information of each institution. Cooperation network analysis and regional distribution of all scientific research institutions in specific fields can be carried out.

3.4. Large Screen Visual Display

According to the actual needs, it can realize various functional requirements, such as dynamic tracking of new technology, hot word map, new technology trend prediction, expert dynamic, organization dynamic, regional distribution of technology dynamic, etc., and conduct statistical analysis and display in the form of column chart, broken line chart, fan chart, relationship diagram, fishbone diagram, hot word chart, etc.

4. Application of Analysis and Mining Technology

There are two main tasks for data analysis and mining in this system: one is to extract knowledge from the acquired new technology information to form a new technology industry knowledge map, which includes the relationship among new technology, institutions, scholars and events. Second, according to the knowledge map of new technology industry, we can extract the new rules or relationships among the elements of the industry, and provide data basis for the trend analysis, statistical analysis and other modules.

So in general, information analysis and mining includes two sub modules: knowledge extraction and knowledge mining. In addition, keyword extraction is also a key function of the system, which needs to capture new technology keywords according to the information captured. Its function flow is shown in Figure 2

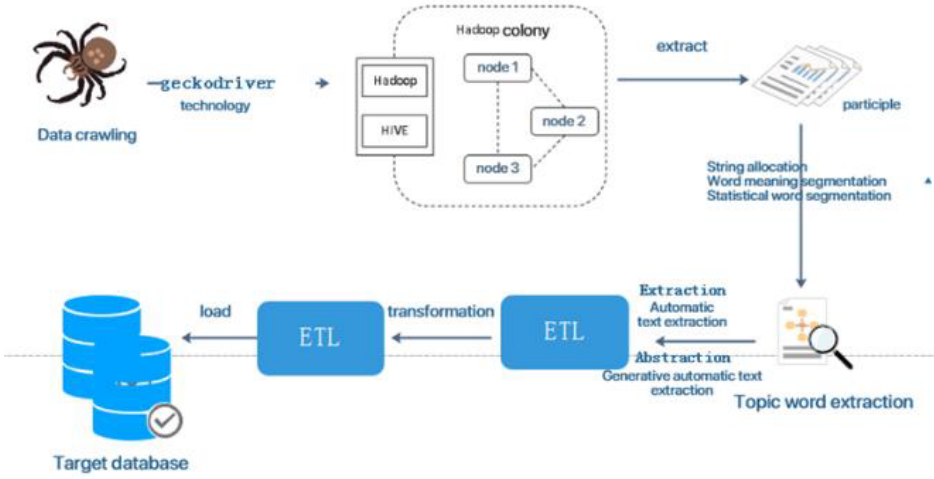

Figure 2. System function flow chart

4.1. Knowledge Extraction

The advanced technology tracking and monitoring system needs to extract effective information from structured, semi-structured and pure text data from the Internet and various thematic databases to form knowledge (structured data) and store it in the knowledge map. A lot of hidden information can be extracted from the dynamic data of advanced technology.

This system uses the method of entity extraction based on the deep learning model BilSTM-CRF. New named entities will appear in the dynamic information of cutting-edge technology, such as information source type, organization name, expert name, etc., and the core of rule-based naming extraction method is rule-making. Therefore, once a new naming entity appears, it will consume a lot of time and energy to update the rules manually, in other words, its formulation and the portability of the rules is poor. Therefore, in order to solve the deficiency of rule-based name extraction, the system adopts the entity extraction method based on the deep learning model BilSTM-CRF. On the basis of this model, this scheme adds the manually extracted features such as word segmentation, indexing and word frequency statistics into the model to describe the semantic information of each Chinese character more fully. Through the above process, BilSTM-CRF algorithm can extract new technical keywords from the captured dynamic data to the maximum extent, and provides data support for predicting which technologies may become hot technologies in the future.

4.2. Extraction of Structured and Semi-Structured Data

In the cutting-edge technology monitoring system, structured data mainly refers to the archived literature data and patent data, which can be transformed into RDF or other forms of knowledge base content. For example, a common W3C recommended mapping language is R2RML (RDB2RDF).

In the cutting-edge technology monitoring system, semi-structured data mainly refers to the data captured from major science and technology news websites, official websites of authoritative scientific research institutions, microblogs and science and

technology columns. This kind of data itself has a certain structure, but it needs further extraction and sorting to get the data. Web page data extraction is generally generated by wrapper, and the process of extracting information is shown in the following figure 3:

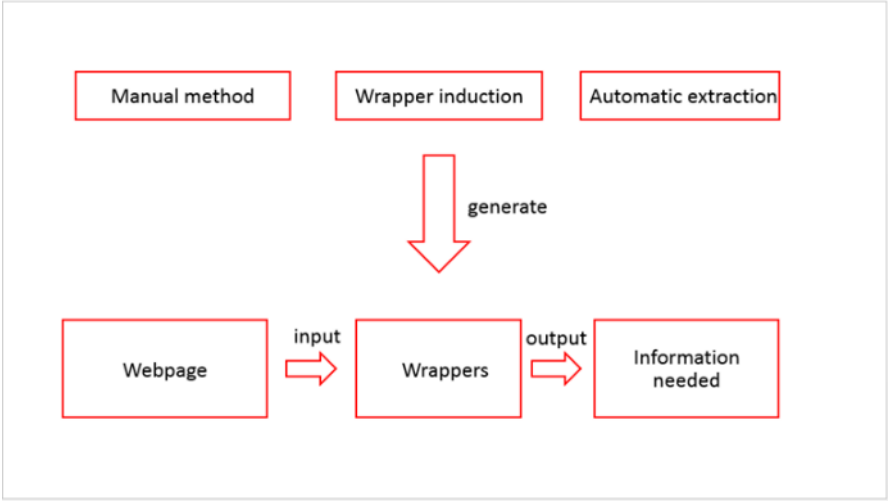

Figure 3. Information extraction process.

4.3. Keyword Extraction

Keyword extraction has important applications in dynamic data retrieval, automatic summarization, text clustering / classification and so on. Keyword extraction algorithms are generally divided into supervised and unsupervised.

Two unsupervised learning algorithms, TF-IDF algorithm and TextRank algorithm, are adopted in the advanced technology tracking and monitoring system.

① TF-IDF algorithm

TF-IDF is a numerical statistical method, which is used to reflect the importance of a technical dynamic hot word to a news, paper or patent. Its main algorithm is: if a word appears frequently in the same document, the TF of the word is high; if the word rarely appears in other documents, the IDF of the word is high. In this case, it is considered that the word has a good classification ability [3], and it is also proved to be a hot word.

② TextRank algorithm

An important feature of this algorithm is that it can analyze a single document and extract keywords without a corpus. The algorithm is as follows:

The first step is word segmentation and part of speech tagging. The effective data collected into the database are divided into sentence patterns, word segmentation and part of speech tagging are carried out sentence by sentence. At the same time, stop words are filtered, and only nouns, verbs, adjectives and other specified part of speech words are retained [4]. The second step is the construction of candidate keyword graph. The co-occurrence relationship is used to construct the edge between any two points in the keyword graph. There are edges between the two nodes only if their corresponding words appear together in the same technical dynamic [5,6,7]. In the third step, according to the formula of TextRank, the weight of each node is propagated iteratively until it converges. The node weights are sorted in reverse order to get the most important words as candidate

keywords [8]. Finally, these keywords are marked in the original text. If they form adjacent phrases, they are combined into multi word keywords [9,10].

4.4. Patent Feature Extraction

The method of artificial intelligence is used to extract the features of data and analyze patent features intelligently, such as rule-based clustering algorithm, text clustering algorithm. The patent keywords are extracted by automatic summarization method and rule-based method. According to the suggestions of domain experts, domain dictionaries or professional websites, the keywords are filtered and refined to extract patent features.

5. Conclusion

The strategic significance of big data technology is to be able to obtain useful information from massive data and apply it to practice after professional processing and mining analysis. The traditional methods and means of scientific and technological information analysis have been unable to meet the needs of the current world development situation. To complete the monitoring of new technologies from massive multi-source heterogeneous big data, it is necessary to build an accurate, efficient and stable intelligent monitoring system, and give full play to the decision-making and consulting ability of think tanks through big data analysis and artificial intelligence assistance. At the same time, the diversity of data sources puts forward higher requirements for the rapid integration and analysis ability of data. Therefore, this paper proposes the overall process structure and functional architecture of the new technology tracking and monitoring system based on multi-source heterogeneous data, and explores the establishment of a set of intelligent, automatic and more perceptive new technology tracking and monitoring system, so as to strengthen the support and leading role of new think tanks in scientific and technological innovation, and play a more obvious role.

References

[1] Tan ZY, Wang Q, Cang HY, Rao YH, Yang NH, Nie L. Construction of monitoring and analysis platform for frontier information of scientific and technological development. Sci. Res., 2010(02): 37-43
[2] Zeng W, Li H, Fan YF, Liu GY, Li R, Xu Z. Research on the identification system of science and technology frontier in the open source information environment. Information Studies: Theory & Application, 2019(07): 34-38
[3] Zhu J, Li HW, Peng X, Zhao WY. Research on the correlation between function and authority of Android application system. Comp. Appl. Software, 2014 (10): 33-39
[4] Zhao Z. Research on the effect of text vectorization on text classification. China excellent master's thesis full text database, 2018 (01)
[5] Chen Z, Zheng S. Research on chemical emergency information extraction based on multi algorithm fusion. Comp. Digital Engin., 2018 (2): 6-6
[6] Zhao MY. English short text measurement method based on part of speech and keywords. Chinese excellent master's thesis full text database, 2018, (01)
[7] Xiang ZX. Design and implementation of cloud health information platform based on distributed crawler. China excellent master's dissertation full text database, 2018, (01)
[8] Li MX. Design and implementation of an interactive mobile user interest discovery system. China excellent master's dissertation full text database, 2016, (11)
[9] Wang Q. Analysis and research on microblog public opinion in emotional context. China excellent master's thesis full text database, 2018, (02)

[10] Wang ZW, Qiu HP, Sun Y, Ke DL. Intelligent recommendation technology for military information service. Command Contr. Simul, 2019 (04): 120-125