A Regional Industry Intelligence Business Platform based on Adaptive Clustering

Junjie Liu[1], Danlin Cai[2], Daxin Zhu[3] and Siyu Huang[4]

[1]Network Information Center, Quanzhou Normal University, Quanzhou, China
[2]Fujian University Laboratory of Intelligent Computing and Information Processing, Quanzhou, China
[3]Fujian Provincial Key Laboratory of Data-Intensive Computing, Quanzhou, China
[4]College of Mathematics and Computer Science, Quanzhou Normal University, Quanzhou, China

Received: August 3, 2020. Revised: September 22, 2020. Accepted: October 13, 2020.
Published: October 19, 2020.

Abstract—How to grasp comprehensive, timely, effective and accurate business competitive intelligence has become an urgent and critical issue for regional industrial clusters. Therefore, a Industrial Internet Platform for Regional Economic based on Adaptive Clustering called IIPRE is developed. The multi thread oriented extraction technology is used to collect the business intelligence data of specific industries. The data is clustered by rule-based machine learning, and the business information data model is used for analysis. Finally, the visualization report is generated by big data visualization software. The system uses Intelligent Retrieval technology to automatically complete the functions from acquisition to processing automatically, to generate data information to meet the application requirements, uses Automatic Classification Technology to provide automatic classification function combining machine-based automatic learning and rule-based information, and uses Personalized Display Technology to provide customizable personalized display pages for individuals, to organize and adjust the hot information, thematic information, clustering, related words, early warning, statistics and other information released by the system. The construction of the platform provides enterprises with comprehensive, timely, effective and accurate business competitive intelligence services, improves the strategic planning, competitive intelligence acquisition and industrial information sharing capabilities of regional industrial clusters, and will achieve tremendous economic and social benefits.

Keywords—Web crawler, ETL, Big data of business competitive intelligence, Service platform, Adaptive Clustering.

I. INTRODUCTION

Along with with the advent of globalization, mobile Internet and e-commerce era, regional industrial clusters need to establish international brands and occupy a key position in the global industrial chain. How to grasp the comprehensive, timely, effective and accurate business competitive intelligence has become an urgent key issue for regional industrial clusters.

The construction of enterprise business competitive intelligence system needs to involve multiple business modules such as stimulated transaction support, knowledge base system construction, market monitoring alarm, personalized real-time request, etc., which has complexity, large investment and non-real-time characteristics. Therefore, the threshold of traditional business competitive intelligence system construction is too high, the quality of service provided is limited, and the implementation effect and time are uncontrollable, which can not enhance the overall competitiveness of industrial clusters.

This paper studies the construction of the "big data cloud service platform for business intelligence of regional industrial clusters" by means of big data and cloud computing technology, so as to provide customized, template and real-time business competitive intelligence analysis service for enterprise customers in regional industrial clusters. In the face of different sectors of industrial cluster, in-depth investigation and positioning are carried out according to the personalized needs of customers. On the basis of general big data platform, we customize the visual demand template and analysis report template for customers, locate and extract huge amounts of open information sources, and develop an intelligent data mining tool suitable for regional industrial clusters, which improves the strategic planning, competitive intelligence acquisition and industrial information sharing capabilities of regional industrial clusters[1,2].

Industrial big data is an important field in the big data industry. From the perspective of the overall layout of big data industry, the economy of the eastern coastal areas of China is relatively developed in the current regional layout of big data industry. Under the background of the emerging industries of intelligent manufacturing and industrial Internet, with the development of new generation communication technologies such as 5G and quantum communication, the industrial Internet will also make great progress with the digitization of machines, the ubiquity of industrial network and the improvement of cloud computing ability. The generation of massive industrial big data will be the inevitable result, and the innovation based on industrial big data is the main part of the new industrial revolution Driving force.

II. RELEVANT RESEARCH AND TECHNOLOGY

A. Web Crawler

Web crawler is a code program that can automatically start from a URL to access other URLs related to it according to the

key fields and targets that have been set, and collect the required value data from each Internet page[3].

As shown in Fig.1, the function of a web crawler is to download relevant data from Internet pages and then provide data for various systems. Many well-known Web search engine systems are called Web-based data collection search engine systems, such as GOOGLE Company. From this we can see the importance of web crawlers. In addition to text information, Internet pages also contain hyperlinks. Web crawlers collect web pages by virtue of these hyperlinks, and this process is very similar to spiders crawling and foraging on spider webs, so it is also known as the web spider[4].

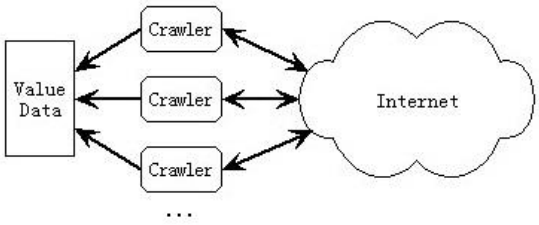

Fig.1. Simple web crawler concept graph

The method of web crawler is used to crawl the network information according to the pre-determined lexicon. The web crawler simulates the browser to access the web page URL. Users can automatically obtain the necessary data without looking through it one by one. The crawler uses Python language, which is an object-oriented interpretative language with concise grammar and supports dynamic input. The position of the first character in each line determines each module. It is an ideal scripting language on most operating system platforms, especially for fast application development [5].

B. Data Cleaning

Data cleaning is to extract valuable information from a large number of raw data, mainly to deal with the input of various data, that is, to transform data into information. Its basic goal is to extract valuable and meaningful data from a large amount of data that is uneven and difficult to understand. The native data of irregular pages crawled from the network can not meet our basic requirements for data processing, so it needs to be preprocessed and converted into relatively regular data needed for our future work. Therefore, data cleaning here actually refers to the basic preprocessing of data to facilitate our later analysis[6]. There are many ways to process data. Which one to adopt depends on the structure of the processing equipment, the working mode and the spatial and temporal distribution of data. Data processing requires software support, including programming languages for programming programs and compilers, file and database systems for managing data, and application packages.

In general, there are repeated fields, noise and higher dimensions in raw and unprocessed data. Our approach is to select appropriate attributes from the raw data as attributes of text mining. The principle of selecting attributes is to give the attributes a clear meaning and ensure that they are unique attributes, remove stop words, remove negligible fields and so on.

C. Data Processing

Network data has the characteristics of huge amount of data and complex data structure. In this case, it is not easy to mine a large amount of data and data of different types and structures accurately and efficiently[7], and it is also very difficult for industry data analysis and detection. To analyze centralized and representative data, effective clustering algorithm in data mining algorithm is often used to improve the accuracy of industry data analysis. Clustering algorithm is a data mining method that divides data into different clusters according to attribute values. The data is divided into several different data modules, and the data with high similarity are concentrated in the same cluster through the algorithm, so that the data in the set has certain characteristics. Big data environment is used to achieve the high efficiency of data analysis. The attributes of data sets processed by clustering are different among different sets.

Our approach is to randomly select K objects as the initial clustering center in all objects. An optimal K value can be found through many experiments according to canopy algorithm. The attributes of each object have a clear identity. The distance between all objects and K object is calculated. According to the clustering comparison between all objects and k value, the clustering is divided. A new clustering is formed according to the clustering comparison value, and the center of all objects in the new clustering is averaged until the function converges, that is, the distance classification between all objects and the selected k value is the optimal solution to achieve the mean of the best accuracy.

III. SYSTEM DESIGN

A. Laser Source and Laser Frequency Stabilization System

The big data service cloud platform for business intelligence of regional industrial cluster includes three parts: network information acquisition distributed data basic service module, massive data industry intelligence retrieval and processing module, and intelligence intelligent display module.

The technical route of this project is shown in Fig. 2.

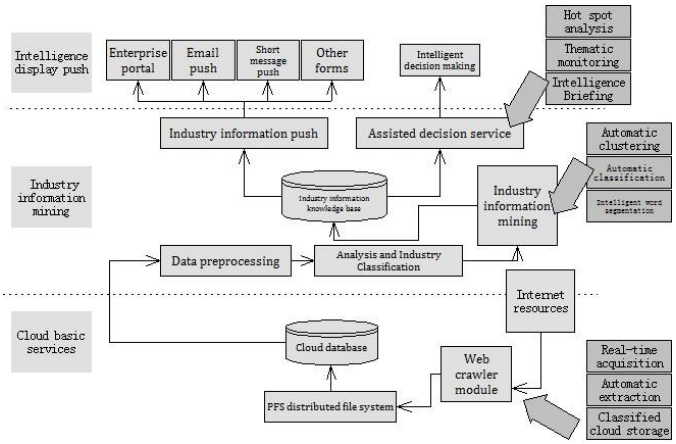

Fig.2. Structure graph

Our main work includes as follows.

(a)Distributed data acquisition engine, which extracts a large number of domain business intelligence data sources, supports vertical retrieval based on specific industry domains and meta-search based on various Internet general search engines, and supports data acquisition in various structured, semi-structured and unstructured forms. Our design idea is to use a code program which can automatically extract web pages to judge whether the web pages content is related to the subject according to the keywords, and then consider whether to download or not. Then on this basis, we continue to visit and download other pages until the requirements are met[9].

(b)The information collected above is processed intelligently on the platform of large data analysis, including automatic classification, automatic clustering, automatic summary, automatic keywords, intelligent word segmentation, information extraction, information filtering, automatic elimination of duplicated data, similarity retrieval and so on. Through the above data mining technologies, deep-seated associations among various kinds of information can be found, and the data with great business value hidden in the mass data can be automatically and precisely mined to guide the decision-making and management of enterprise operation.

(c)Through intelligent visualization technology, various kinds of visualization reports are further generated, including information navigation, information early warning, hot spot analysis, dissemination analysis, time trend analysis, thematic monitoring, information briefing, information push, statistical analysis, information retrieval, etc. Finally, they are pushed to enterprise users through client software to provide comprehensive, timely, effective and accurate business competitive intelligence services for enterprises.

(d)Intelligent retrieval strategy mode based on immediate return: in the process of information collection, breadth first strategy is usually used to traverse. However, if there are too many branches of a single page during retrieval, it will directly lead to inefficiency and even resource exhaustion. The probability of this problem is particularly frequent in the case of massive data, which is a difficult problem to solve in this project. In the project, the retrieval strategy mode based on immediate return value evaluation is adopted to avoid such problems. The relevance degree of keywords and links is used as the basis of value evaluation. The improved fish swarm algorithm is used to calculate the correlation degree, and the situation with high correlation degree is given priority to improve the collection efficiency.

(e)Fusion of semantic rule classification technology: SVM classifier and KNN classifier are both good classifiers at present. However, for large-scale web data sets, SVM classifier needs more examples and longer training time. If KNN selects too many feature words, it will lead to high vector dimension, increase calculation cost, make distance calculation inaccurate and affect classification accuracy. Therefore, we combine the two classifiers to construct a multi classifier engine, which achieves better classification performance than the above two separate classifiers.

IV. PLATFORM IMPLEMENTATION

A. Distributed Data Acquisition

Fig.3 shows the process of data crawling. We configure the development environment of python 3.5 + pyCharm, install MySQL database and Navicat for MySQL. pymysql is installed to connect MySQL in Python, and finally the crawler code is written.

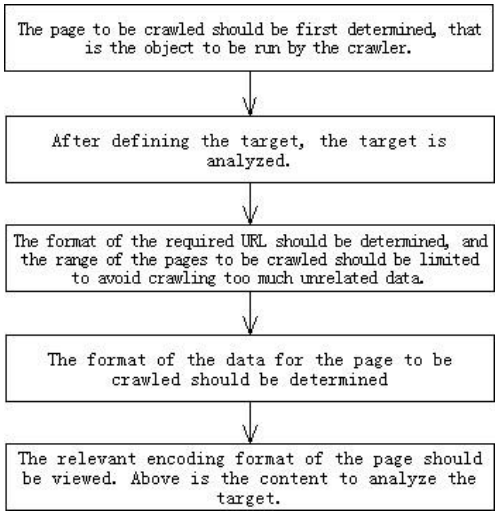

Fig.3. Data crawling process

The database "shuju" should be created in Navicat for MySQL. The Python crawler code should be written in accordance with the above objectives. The first step is to connect to the database and insert the crawled data: Whether the table "datas" exists should be checked. If it does not exist, the table is created. Then, the webpage construction should be analyzed, and the regular expressions should be written to extract the needed information from the webpage source code. The extracted information should be added to the dictionary, and then the data in the dictionary should be stored in the local MySQL database. Finally, XPath is used to get the data correctly and completely, and the data crawling is started.

B. Data Cleaning and Storage Module

There is a lot of noise in the native data collected from HDFS, which is also called irregular data. The format of the data can not meet our basic requirements for data processing, so it needs to be preprocessed.

Before data analysis, we must first process the data into numerical format as far as possible, eliminate the string content contained in it, eliminate invalid records, and transform it into more regular data needed for our later work. That is, basic pre-processing of data should be carried out according to different business needs in order to facilitate our later statistical analysis. We mainly clean three kinds of data: repetition, missing and exception.

(a).Duplicated data

If the actual business does not need duplicated data, duplicated data can be deleted directly, and python can handle it directly with commands.

```
drop_duplicates( )
```

(b).Data missing

The missing is the null value, which is usually handled by filling, and this is done according to the actual business needs. Generally speaking, when the number of null values is small,

one of the continuous values can be filled, such as the average, median, etc. When the number of null values is large, accounting for more than 50%, mode can be used to fill. When the number of null values accounts for the majority, there is no need to use the original data at this time. You can make your own data and generate an indicative dummy variable to participate in the subsequent modeling.

The processing method in python is as follows:

df.apply(lambda col:sum(col.isnull())/col.size)

df.col1.fillna(df.col1.mean())

(c).Data exception

Data exception refers to the value that differs greatly from other values in the data, and some are called outliers, such as those more than 150 in age. The abnormal data will seriously interfere with the results of the model and make the conclusions unreal or biased. Therefore, it is necessary to remove these noise values. The commonly used methods are cap method, binning method, clustering method and so on.

We parse the corresponding IP address of the collected information, and divide its request into method, request_url, http_version and so on for the convenience of other statistics. The process of data cleaning is mainly to write MapReduce program, which is divided into three basic processes: Mapper, Reducer and Job. The main sentences are as follows:

Yarn jar data-clean-1.2-SNAPSHOT-jar-with-dependencies.jar cn.xpleaf.dataClean.mr.job.CleanJob hdfs://ns1/input/data-clean/access/2016/10/23 hdfs://ns1/output/data-clean/access

The distributed way after data cleaning makes the whole information acquisition process have the characteristics of high efficiency, high scalability, reliability and low cost. Data storage adopts the strategy of using both relational and full-text databases. Relational databases store small amounts of data but need to read and write frequently. Full-text databases store large amounts of data and need full-text retrieval of web data. Indexes can be created to support fast retrieval and positioning. The data stored in the full-text database is the main part of the business intelligence big data platform, and the amount of data is huge. Therefore, the system uses a professional distributed massive data storage scheme, which provides global namespace through Pfs distributed file system, aggregates 512 data nodes into a disk space of 120 PB, which is provided to the big data computing layer through NFS and CIFS protocols. Five data storage modes are provided, including distribution mode, strip mode, replication mode, distribution-replication mode and strip-replication mode.

C. Data Mining and Display Module

Aiming at the problem that hierarchical clustering algorithm is difficult to achieve parallel computing, this paper adopts parallel hierarchical clustering algorithm based on data partition, focusing on solving the parallel problem of hierarchical clustering algorithm. The parallel hierarchical clustering is used to introduce the data partition into the traditional hierarchical clustering algorithm. The data partitioning algorithm can be used after optimizing the vertical partition algorithm based on vector component group feature statistics, and reasonably using the sorting characteristics of Mr programming model and the secondary sorting technology to efficiently select the merging points, and realize the effective clustering of redundant industrial Internet data.

The main intelligent processing functions involved in the analysis and mining of information data include text categorization. The system supports content-based and rule-based automatic classification. Content-based classifier takes classified corpus text as input and generates content-based automatic classification template. Rule-based automatic classification provides a rule generator for classification. Rule writing satisfies logic operations such as “and, or, non, and xor”. At the same time, the system implements the structure of classification tree and supports hybrid classification method based on content and rules. To discover and associate the relevance of network information, similarity document association needs to realize automatic similarity processing based on text content. The threshold of document similarity and the size of retrieval result set are set freely by the similarity calculation method based on text content. The automatic duplicate checking of text should be realized. Automatic clustering and hotspot information analysis system realizes automatic clustering function to facilitate users to discover the hotspots and the evolution of hotspots over time. Advanced automatic word segmentation system and similarity algorithm are studied, and category keywords are given for each category. This technology can directly provide analysis report and briefing service for relevant departments. Intelligence presentation part, combining with knowledge base, builds graphical information functions, including hot spot analysis, trend analysis, thematic analysis, dissemination analysis, sensitive information analysis and association analysis based on the information obtained from rich semantic analysis technology, model and analysis algorithm in the previous state, aiming at users’ needs of paying attention to information and business analysis, in the form of various histograms, pie charts and broken line graphs. Then, they are targeted and customized to customers. Finally, through the client software, they are pushed to enterprise users. The algorithm steps of automatic clustering are as follows.

(a)Select the initial k category centers u1, u2, ……uk.

(b)Each sample is marked as the nearest category to the category center, that is

$$label_i = \arg\min_{1 \le j \le k} \|x_i - u_j\| \tag{1}$$

(c)Each category center is updated to the mean of all samples belonging to that category.

$$u_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_i \tag{2}$$

(d)Repeat the last two steps until the change in the category center is less than a threshold.

The flow chart of the algorithm is as follows.

K-means({$x1$, $x2$,…,xn}, k) Set the number of clusters k

1. ($s1$,$s2$,…,sk)←SelectRandomSeeds({$x1$,$x2$,…xn},k)
   Random Selection of K centers
2. 2 for k ← 1 to K

3. do μk ← sk
4. while stooping criterion has not been met
5. do for k ← 1 to k
6. do ωk ← { }
7. for n ← 1 to N
8. do j← argmin |μj - xn|
9. ωj ← ωj ∪{Xn} (Each record Xn is grouped into the cluster nearest to its central point.)
10. for k ← 1 to K
11. do μk ← 1/|ωk | ∑X∈ ωk (Update the central point of each cluster)
12. return { μk1,…, μk }

The time complexity of the algorithm is O (The number of central points * the number of data sets * the number of iterations).

The final effect we achieved provides a customizable personalized display page for enterprises, which can organize and adjust a variety information, such as the hot spot information, thematic information, clustering, related words, early warning, and statistics released by the system, and can add, delete and drag modules. They can edit and customize their own display page by dragging.

V. CONCLUSIONS

Through the latest big data and cloud computing technology, the project uses Internet thinking to build "big data service for business intelligence of industrial clusters". The construction of the project plays a great role in enhancing the core competitiveness of local industrial clusters. It will play a key strategic role in guiding the government to give full play to the location and industrial advantages of industrial clusters, and will achieve great economic and social benefits. Key technologies adopted in the project include:

(a)Intelligent retrieval technology: The system will take Internet data acquisition as the main line, study multi-threaded concurrent data acquisition technology and intelligent retrieval strategy, seamlessly integrate all key technologies into the system, automatically complete the functions from acquisition to processing, and finally generate data information to meet the application requirements. The whole process is reasonable and efficient.

(b)Automatic classification technology: It can provide automatic classification function combining machine-based automatic learning and rule-based information. The system adopts automatic classification technology, which can classify and manage according to topics, keywords, sources, etc. The information can also be classified by statistics or rules to create proprietary classification models. The classification adopts tree structure, which can be managed and maintained without restriction of series.

(c)Personalized display technology: It can provide customizable personalized display pages for individuals, which can organize and adjust the hot information, thematic information, clustering, related words, early warning, statistics and other information released by the system. Using Portal-like technology, modules can be added, deleted and dragged. By dragging, users can edit and customize their own display pages to facilitate their work. Different users define different permissions so that they can see information related to their own permissions.

ACKNOWLEDGEMENTS

The study is supported by Natural Science Foundation of Fujian Provincial Science and Technology Department (NO.2018J01558) and Fujian University Laboratory of Intelligent Computing and Information Processing, Fujian Provincial Big Data Research Institute of Intelligent Manufacturing.

REFERENCES

[1] Danlin Cai, Junjie Liu, Taisheng Zeng, Daxin Zhu. Credible Recommendation Mechanism in Collaborative Filtering Recommendation System. Revista de la Facultad de Ingeniería, 2017.10.

[2] Danlin Cai, Daxin Zhu, Lei Wang, Xiaodong Wang. A Simple Linear Space Algorithm for Computing a Longest Common Increasing Subsequence. IAENG International Journal of Computer Science, 2018.8.

[3] WiseNut Search Engine White Paper (2001).http://www.wisenut.com/pdf/ WISEnutWhitePaper.pdf

[4] Daxin Zhu, Xiaodong Wang. An Efficient Dynamic Programming Algorithm for a New Generalized LCS Problem. IAENG:International Journal of Computer Science, 2016.

[5] Lin LI．Design and Implementation of Network Crawler System Based on Python．Information Communication, 2017.

[6] Big Data Acquisition, Cleaning and Processing: A Complete Case of Using MapReduce for Offline Data Analysis, https://blog.csdn.net/ting_163/article/details/80048886

[7] Zhixian Zheng, Ming Wang. Application of K-means Clustering Algorithms Based on Large Data in Network Security Detection．Journal of Hubei Second Normal University, 2016．

[8] Data Cleaning in Data Processing. https://blog.csdn.net/mochou111/article/details/81744846.

[9] Junjie Liu, Daxin Zhu, Danlin Cai, Siyu Huang. Design and Implementation of Big Data Cloud Service Platform for Business Intelligence of Regional Industrial Clusters.Journal of Basic & Clinical Pharmacology& Toxicology, 2019.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US