Short Communication Industrial Engineering & Management 2021
Vol.10 No.1

A Big Data Knowledge Computing Platform for Intelligence Studies- Wen Yi, Chinese Academy of Sciences, China

Wen Yi
[1] Chinese Academy of Sciences, China

Intelligence studies is a method of using modern information technology and soft science research methods to form valuable information products by collecting, selecting, evaluating and synthesizing information resources. With the advent of the era of big data, the core work of information analysis with data is facing enormous opportunities and challenges. How to make good use of big data in an effort to solve the problem of big data, optimize and improve the traditional intelligence studies methods and tools, innovation and research based on big data are the key issues that need to be studied and solved in current intelligence studies work.

Through the analysis of intelligence studies methods and common tools under the background of big data, we sort out the processes and requirements of the intelligence studies work under big data environment, design and implement a universal knowledge computing platform for intelligence studies, which enables intelligence analysts to easily use all kinds of big data analysis algorithms without writing programs (http://www.zhiyun.ac.cn). Our platform is built upon the open source big data system of Hadoop and Spark. All the data are stored in the distributed file system HDFS and data management system of Hive. All of the computational resources are managed with Yarn and each of the submitted task is scheduled with the workflow scheduler system Oozie. The core of the platform consists of three modules: data management, data calculation and data visualization.

The data management module is used to store and manage the relevant data of intelligence studies, which consists of four parts: metadata management, data connection, data integration and data management. The platform supports the import and management of multi-source heterogeneous data, including papers, patents from ISI, PubMed, etc., and also supports the data import with API of MySQL, Hive and other database systems. The platform has more than 20 kinds of data cleaning and updating rules, such as search and replace, regular cleaning, null filling, etc., and also supports users to customize and edit the cleaning rules.

The data calculation module is used to store and manage the big data analysis algorithm and intelligence analysis process, and provides a user-friendly GUI for users to create customized intelligence analysis process, and the packaged process can be submitted to the platform for calculation and obtain the calculation results of each step. In the system, a task is formulated as a directed acyclic graph (DAG) in which the source data flows into the root nodes. Each node makes operations on the data, generates new data, and sends the generated data to its descendant nodes for conducting further operations. Finally, the results flow out from the leaf nodes. The data visualization module is used to visualize the results of intelligence analysis and calculation, including more than ten kinds of visualization charts such as line chart, histogram chart, radar chart and word cloud chart.

Practice has proved that the platform can well meet the requirements of intelligence studies in various fields in the era of big data, and promote the application of data mining and knowledge discovery in the field of intelligence studies.

This work is partly presented at 7th International Conference on Big Data Analysis and Data Mining, July 17-18, 2020 Vienna, Austria