Study of New Materials Design based on Hadoop

Wu Jun[1,2,a], Huang Zhixiong[1,b]

[1]School of Materials Science and Engineering, Wuhan University of Technology, Wuhan 430070, China

[2] School of Computer Science, Hubei University of Technology, Wuhan, 430068, China

[a] wujun@whut.edu.cn, [b] lavazza@foxmail.com

Abstract. With the rapid development of information technology, the scientific research shows that the data mining and other information technology could be used in the design of new materials. It is explicit that Intelligent Materials research focuses on using physical and chemical principles combined with computer techniques such as Big Data, Cloud computing and Intelligent modeling and simulation to solve chemical problems. In this paper, based on the cluster based outlier algorithm as the main body, this paper discusses the definition New Materials research In the Hadoop cloud platform, and the parallel processing of Map-Reduce model. The performance this model of new material was established by using the method of Map-Reduction provided the basis for the performance optimization.

Keywords: Materials Science, New Materials Design, Computational Methods, Data Mining, Hadoop

1 Introduction

In recent years, the scientific research shows that the data mining and other information technology could be used in the design of new materials. On one hand this new research method would help to make the new materials much more intelligent and useful than the traditional method, on the other hand it can get more ideal material with less experiment, and achieve the result with half the effort. From the first International Conference on Computer Aided Design of New Materials in 1990 to the Intelligent Materials developed in Big data time, it is explicit that Intelligent Materials research focuses on using physical and chemical principles combined with computer techniques such as Big Data, Cloud computing and Intelligent modeling and simulation to solve chemical problems[1-4].

Nowadays there are some international journals about this interdisciplinary research, such as Modeling and Simulation in Materials Science and Engineering, Computational Materials Science and so on. Nongnuch Artrith and Alexander Urbanb[5] have worked about Machine learning interpolation of atomic potential energy surfaces enables the nearly automatic construction of highly accurate atomic interaction potentials. Mansouri Iman and Ozbakkaloglu Togay[6] studies the ability of artificial neural network (ANN), adaptive neuro fuzzy inference system (ANFIS), multivariate adaptive regression splines (MARS) and M5 Model Tree (M5Tree) techniques to predict ultimate conditions FRP-confined concrete. Performances of the proposed models are also compared with those of the existing conventional and evolutionary algorithm models, which indicate that the proposed ANN, ANFIS, MARS and M5Tree models exhibit improved accuracy over the existing models. Cabaleiro Manuel, Riveiro Belen, and etc[7] have researched how to represent an important lack of material in the structural member. All the changes in the cross section of the beam must be considered in any kind of strength calculation and the protocols for the structural health monitoring. The results obtained demonstrate that the algorithm proposed is adequate for the automatic analysis of the geometrical properties in the section of timber beams with a lack of material or irregular section.

2 Computer Aided Material Design

Computer Aided Material Design is divided into two parts: First - Principle Calculations and Data Mining.

In the 1980s with the new material exploration and research when First - Principle Calculations method played an important role, it is important to discover those new materials, such as High temperature superconducting material, Super hard material, Nano material, Artificial low dimensional quantum structure materials. The first principle calculation is also called as Calculated based on the design of quantum theory. And its basic method has the solid quantum theory and the quantum chemistry theory. Especially suitable for the calculation and design of materials for atomic scale, Nano scale engineering, materials for many devices, and materials for electronic devices. Main research task in the aspect of material surface and interface would be as following work: reveal occur in the physical connotation of material surface and interface phenomena; how to use first principles method to calculate, design of surface and interface of physical chemical and dynamical processes. At present, the most powerful theoretical method is molecular dynamics simulation and Monte Carlo simulation. The key of the technique lies in the accurate calculation of the interaction potential between atoms.

However Data Mining is so different as First-Principle Calculations which be viewed as inductive method. Data mining is mainly defined as ‘‘a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data’’, or ‘‘the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner’’. As such, data mining and knowledge discovery are typically considered knowledge intensive tasks. Thus, knowledge plays a crucial role in Intelligent Materials research areas[8-10]. Through the data mining method would sum up the law and the method to obtain the required materials. In this paper, based on the Hadoop cloud platform, The performance prediction model of new material was established by using the method of data reduction and support vector machine or artificial neural network, which provided the basis for the performance optimization.

3 Data Mining and Hadoop

Calculation and design from the theory, although it can provide many useful clues, but the complexity of the actual material makes people tend to give priority to the induction method. Has been widely used in computational materials design, artificial neural network in data mining, pattern recognition, genetic algorithm, regression analysis and other algorithm, in recent years, the international has formed a group of computer software based on these methods. On the basis of experimental, using neural network method, the material mechanical properties and alloy composition are used as input of the network, material of other alloy composition and heat treatment temperature as the output of the network established mathematical model which can reflect the inherent law of the experimental data, in order to realize the optimization design of materials. The data mining method can also be used in the formulation design and control, forming, glazing, drying, sintering process and parameter control in ceramic industry, so as to achieve the goal of controlling and improving the quality of products.

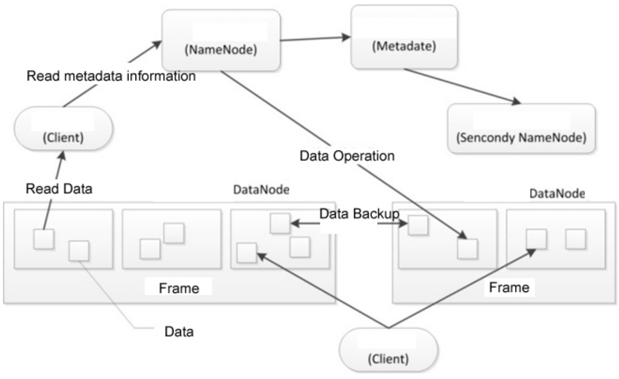

Fig1 Hadoop Architecture

Hadoop distributed file system (HDFS) is the bottom Hadoop file system, is designed by Google open source implementations of distributed file system GFS. HDFS Hadoop as one of the core components, and Hadoop HDFS data in blocks of storage in the cluster files. HDFS by superior fault tolerance, which can run on a large number of common storage hardware, to meet the data storage requirements, and has very strong extensibility, HDFS architecture figure. HDFS is a file system that can provide massive data storage and high throughput data access to the application, at the same time, she and the common file system, can be very convenient to achieve the file operation. Thanks to design architecture of HDFS master-slave (Master/slave), a HDFS cluster contains a node name (name node) and a series of data nodes (data node), which acts as a master role name node mainly responsible for the HDFS file system management, and accept the request sent by the client. As data node is slave role, is the main function of the data file is stored in HDFS by a file cut into one or more data block, the data block is storage in one or more data node, and the generated logical file storage structure on the name node nodes. HDFS file system in order to achieve high fault tolerance of the system, will be a number of copies of each data block to be stored in a number of different data nodes.

4 Map-Reduce Model and Experiment

Map-Reduce Hadoop as an open source implementation of Map-Reduce Google, is a high reliability, high fault tolerance capability of parallel computing software framework. Map-Reduce based applications can be run in large clusters of parallel processing large data sets. Simplifies concurrent programming model provides application programming interface (API) to the user through the map reduce, make not familiar with parallel computing users can easily develop map reduce application, and can reduce the amount of repeated work, map reduce execution flow chart is as follows Fig 2.

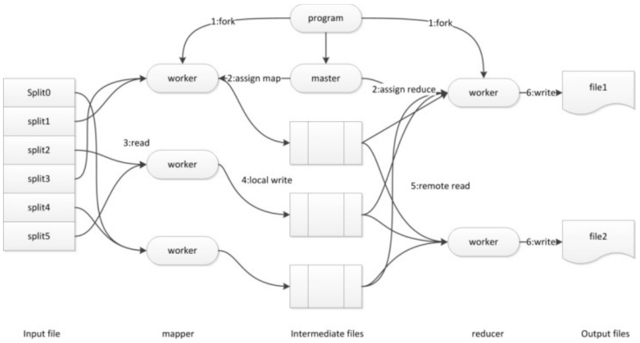

Fig 2 Map-Reduce structural work

Map-Reduce mainly consists of two core operations: mapping (Map) and protocol (Reduce). The mapping (Map) method is used to get a set of key value pairs (key/value) mapping into a new set of key value pairs (key/value). Protocol (Reduce) is used to ensure that all the keys to share the same key group. By running the Map-Reduce computing framework and Hadoop distributed file system (HDFS) on the same node, the cluster can use the network bandwidth to achieve efficient task scheduling. In the construction of the cipher text index structure should be considering the characteristics of HDFS file system, establishes the cipher text index compression and for small files for constructing reasonable with large files and the common keyword composition large inverted list for a reasonable segmentation, make full use of the characteristics of the Hadoop cluster computing, to achieve safe and efficient indexing, storage and query.

Acknowledgments

The research is supported by the National Science Foundation (NSF) under Grants (No.61502155, Hubei Natural Science Foundation under Grants (No. 2014CFB590）, Natural Science Foundation of Hubei University of Technology under Grant (No. BSQD13039), Wuhan University of Technology Hubei Key Laboratory of Transportation Internet of Things Foundation under Grants (No.2015ⅢI015-A03).

References

1. Rong Yu,Qi Zhang,Qian Zhan. Softest elastic mode governs materials hardness. Chinese Science Bulletin. 2014(15)
2. A.L. Ivanovskii. Mechanical and electronic properties of diborides of transition 3 d –5 d metals from first principles: Toward search of novel ultra-incompressible and superhard materials. Progress in Materials Science . 2011 (1)
3. Minghai Yao,Miao Qi,Jinsong Li,Jun Kong. A novel classification method based on the ensemble learning and feature selection for aluminophosphate structural prediction. Microporous and Mesoporous Materials . 2014
4. Qunyi Wei,Xiaodong Peng,Xiangguo Liu,Weidong Xie. Materials informatics and study on its further development. Chinese Science Bulletin . 2006 (4)

5. Nongnuch Artrith, Alexander Urbanb. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2. Computational Materials Science.2016(114): 135-150.
6. Mansouri Iman and Ozbakkaloglu Togay. Predicting behavior of FRP-confined concrete using neuro fuzzy, neural network, multivariate adaptive regression splines and M5 model tree techniques. Materials and Structures/Materiaux et Constructions, p 1-16, January 5, 2016
7. Cabaleiro Manuel, Riveiro Belen. Algorithm for the analysis of the geometric properties of cross-sections of timber beams with lack of material from LIDAR data. Materials and Structures/Materiaux et Constructions, 1-14, December 30, 2015
8. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Advances in knowledge discovery and data mining, in: American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996, pp. 1–34.
9. D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, 2001.
10. C. Bizer, T. Heath, T. Berners-Lee, Linked data—the story so far, Int. J. Semant.Web Inform. Syst.5 (3) (2009) 1–22.
11. Amazon Elastic MapReduce, http://aws.amazon.com/elasticmapreduce/(2013).
12. The Apache Hadoop Framework, http://hadoop.apache.org (2013).
13. Aridhi, S.,d'Orazio, L. ,Maddouri,M.,MephuNguifo,E. Density-based data partitioning strategy to approximate large-scale sub graph mining. Inf.Syst.2015vol.48, 213–223.
14. Kang, U.,Faloutsos,C.,2013.Biggraphmining:algorithmsanddiscoveries.SIGKDD Explor.Newsvol.14(2),29–36
15. Aridhi S, Lacomme P, Ren L, et al. A MapReduce-based approach for shortest path problem in large-scale networks. Engineering Applications of Artificial Intelligence, 2015, 41:151-165.