Research on the Method of Big Data Collecting, Storing and Analyzing of Tongue Diagnosis System

To cite this article: Xiaowei Chen and Qingfeng Wu 2018 IOP Conf. Ser.: Mater. Sci. Eng. 322 062013

View the article online for updates and enhancements.

Related content
- Big Data in industry
  T S Latinović, D M Preradović, C R Barz et al.
- Next Generation Workload Management System For Big Data on Heterogeneous Distributed Computing
  A Klimentov, P Buncic, K De et al.
- Analysis of Computer Network Information Based on "Big Data"
  Tianli Li
Research on the Method of Big Data Collecting, Storing and Analyzing of Tongue Diagnosis System

Xiaowei Chen¹, Qingfeng Wu²*
¹,²School of software, Xiamen University, Xiamen, China

*Corresponding author e-mail: qfwu@xmu.edu.cn

Abstract. This paper analyzes the contents of the clinical data of tongue diagnosis of TCM (Traditional Chinese Medicine), and puts forward a method to collect, store and analyze the clinical data of tongue diagnosis. Under the guidance of TCM theory of syndrome differentiation and treatment, this method combines with Hadoop, which is a distributed computing system with strong expansibility, and integrates the functions of analysis and conversion of big data of clinic tongue diagnosis. At the same time, the consistency, scalability and security of big data in tongue diagnosis are realized.

1. Introduction
As one of the main objective indexes of TCM clinical syndrome diagnosis, tongue diagnosis is emphasized by oriental medicine and western medicine. In the process of the development of clinical medicine, the research of tongue diagnosis has developed from simple naked eye observation, which is easily influenced by external environment, doctors' clinical experience and other subjective and objective factors, to a multidisciplinary clinical and practical research [1]. The sources of clinical data of tongue diagnosis are increasing all the time, including not only structured data but also includes unstructured data and semi-structured data with large capacity and fast growth rate. Obviously, the way to collect, store and analyze the clinical data of tongue diagnosis is also in urgent need of updating. This paper combines with the cloud computing technology and big data technology, puts forward the method of Hadoop, which is a distributed computing system to collect, store and analyze tongue diagnosis data with strong expansibility.

2. Relevant Background

2.1. The necessities and challenges of big data processing in tongue diagnosis
The traditional mode of tongue diagnosis and treatment of TCM is based on the personal experience and it can’t meet the needs of development, and the establishment of a big data collection, storage and analysis system to process the clinical data of tongue diagnosis intelligently is meaningful in the following aspects:

- Classified the traditional description of TCM as quantitative and standardized objective expression.
- The disease diagnosis model based on classification algorithm can be established to provide thoughts for the diagnosis of specific diseases.
- The objective expression of tongue image can provide basis for objective diagnosis of TCM and evaluation of curative effect of sub-health.

However, there are still several challenges during the process of the tongue diagnosis:
• Data collection: patients’ information is scattered, the data is heterogeneous and of high complexity, the procession of data collection is difficult [2].

• Data storage and management: Medical data with complex structure and diverse sources have higher requirements on the storage capacity, instantaneity, concurrent access, security and privacy.

• Data analysis: (1) The manual labeling of the large-scale tongue image data is time-consuming and labor-consuming [3], (2) Clinical data collected from tongue diagnosis are large and lack of integration, which is difficult to process and query.

The tongue diagnosis is based on the principles of “the four diagnosis” and the essence of “Syndrome differentiation and treatment” [4]. Only when taking the theory of TCM syndrome differentiation as the core, establishing a large sample of disease and syndrome combination of tongue database and exploring the specific relationship between tongue image data and disease diagnosis, can we make use of big data of tongue diagnosis effectively to realize the function of intelligent decision and the modernization of tongue diagnosis [5].

2.2. Hadoop technology
This paper combines Hive data warehouse software with Hadoop technology to collect, store and analyze the big data of tongue diagnosis.

Hadoop is a distributed computing system with strong expansibility, which is very suitable for the collection, storage and analysis of tongue diagnosis data. Its core module includes two parts: HDFS and MapReduce.

HDFS is a distributed file system. The storage capacity of this system can be expanded, and it has the characteristics of high reliability and high performance-to-price ratio [6], which is suitable for distributed computing and storing unstructured data. MapReduce is a technology providing parallel computing functions [7]. They are co-deployed to form a single cluster [8]. Hive is the software that can be used to query and manage large-scale data sets of distributed systems (such as HDFS) conveniently. At the same time, it can be converted to MapReduce program underlying to improve query efficiency.

3. Architecture Design of data Collection, Storage and Analysis System based on Hadoop Technology
The general framework for the collection, storage and analysis of clinical big data of tongue diagnosis is shown in Figure 1.
3.1. Collection of big data of tongue diagnosis
Firstly, this system develops ETL (Extract, Transform and Load) module based on Sqoop technology to realize the migration of structured data from RDBMS (Relational Database Management System) to Hadoop. Secondly, it conveys the semi-structured data and unstructured data through Hadoop Common.

In the integration of the data collection of tongue diagnosis, to ensure the consistency of the data, three main tasks are carried out: (1) The related semi-structured data, such as the detailed information of tongue diagnosis, are transformed to a unified data format, (2) Review and clean the parsed data, which contains the procession of checking the format of the data and the logic of the value, correcting errors, removing data which are noise or repetitions, and converting the data to information which is convenient for statistical analysis [7], (3) The interpolation strategy is given to the missing value due to the objective factors such as network failure and power outage, and the data complement is carried out.

3.2. Storage and management of big data of tongue diagnosis

3.2.1. Storage architecture
In this system, cloud storage technology is introduced to meet the need of the storage capacity, availability and performance, and prevent data missing caused by the change of local single storage. The Hadoop storage cluster is chosen as the data storage center, and HDFS is used as the physical unit of big data storage in tongue diagnosis. Then the Hive data warehouse is built on the Hadoop cluster, and the logical management and high-speed access of multi-type big data are realized on the basis of Hive in HDFS. As the result, the efficiency of storing and managing big data of tongue diagnosis is improved. The strong fault tolerance and reliability of big data storage and management of tongue diagnosis are realized.

3.2.2. Cloud storage architecture
In order to prevent the big data of tongue diagnosis from leaking, it is necessary to provide encryption services. As shown in Figure 2, the system is divided into three modules: (1) security client module, (2) secure and reliable transmission module, (3) server module. The server module is the system that provides the cloud storage service.

![Fig.2 Cloud Storage System Architecture with Function of Encryption](image_url)

3.2.3. Data masking
In order to ensure the security of cloud data, data masking is required. Data masking is a technical mechanism to transform and modify sensitive data under given rules and policies, which can address the use of sensitive data in non-trusted environments such as cloud platforms to a large extent [10].
This paper deals with the data masking in three aspects: identification and management of sensitive data before use, protection in use, audit and data trace after use.

- The identification of sensitive data are realized with data feature learning and natural language learning, and the sensitive static data should be stored and managed well [10].

- According to the level of the instantaneity of data, carry on the dynamic data masking or the static data masking [11]. At the same time, the sensitive data leakage warning, blocking and other functions should be implemented [10].

- Once data leakage occurs, the source of the leak needs to be found in time, so that the leak can be checked and filled [12].

3.2.4. Data scalability

It is important to deal with data rightly when more and more data are generated to be processed. So the following is a method named parallel power iteration clustering (p-PIC) to do with it:

After the master processor marks the indexes of the beginning and ending of the remote data chunk, every processor does the following calculation on the basis of the input file.

\[ Ai(r, c) = \frac{x_{rc}}{||x||^2} \quad \text{for } r \neq c \text{ from the input} \]

\[ Ai(r, :) = \frac{\sum Ai(r, c)}{\text{row sum}} \quad \text{for } c = r \text{ from the input} \]

Then a similarity sub-matrix is produced. The master processor collects these data and puts them into an overall row sum. Every processor interacted with the master processor updates vectors through matrix–vector multiplication.

3.3. Analysis technique of big data of tongue diagnosis

The analysis module of tongue diagnosis includes traditional analysis methods to maximize the value of data. It also contains technologies like parallel computing and data mining algorithms to deal with big data. Precise analysis is of vital importance to improve care, lower costs and save lives [7]. The followings are the applications of several methods.

- The semi-supervised algorithm based on graph learning is used to realize the automatic semantic annotation of tongue image quickly and effectively [3].

- The automatic classification of tongue color and coating color in tongue image analysis can be realized by neural network classifier based on LVQ (Learning Vector Quantization) or SVM (Support Vector Machine) [14].

- Fuse and reconstruct the tongue image to integrate multi-modal medical images to simplify the analysis of the tongue diagnosis [15].

- Another functions: Big data analysis module of tongue diagnosis also includes providing visual interface, providing OLAP (Online Analytical Processing) multidimensional analysis tools for analysis to form multidimensional reports and automatically generating retrieval system that suits users’ preference.

4. Conclusion

With the field of tongue diagnosis of TCM has entered the era of big data, what researchers have to do is to keep pace with the times and promote medical modernization. In this paper, the method based on Hadoop technology of data collection, storage and analysis system of TCM tongue diagnosis has strong advantages in the era of new technology environment, hoping to provide some references for scholars to study medical data.

Acknowledgement

This work was supported by NSFC (No. 61402387, No. 61402390); the Key Program of Science and Technology of Fujian Province of China (No. 2014H0044); Science and Technology Guiding Project of Fujian Province of China (No.2015H0037, No.2016H0035);Enterprise Technology Innovation Project of Fujian Province; the Education and Research Project of Middle and Young Teacher of Fujian Province of China (No.JA15018); the Overseas Study Scholarship of Fujian Province; Science and Technology Project of Xiamen, China (No. 3mu502Z20153026).
References

[1] Z. Qingwen, “Research on Portable Integrated Auxiliary diagnosis and treatment equipment based on Pulse diagnosis and tongue diagnosis Information extraction and recognition”, Beijing: Beijing University of Chinese Medicine, 2007.

[2] S. Aiting, Z. Haiping, “The Development Prospect of the Application of Big data Technology in Medical Field”, China Management Informationization, Vol.20, No.19, October 2017, PP-193-195

[3] L. Hanqing, L. Jing, “Image Annotation Based on Graph Learning”, Chinese Journal of Computers, Vol.31, No.9, September 2008, PP-1629-1639

[4] J. Yiwu, C. Jianzhong, Z. Henghong, M. Jianzhong, “Computerized TCM tongue diagnosis system”, Chinese Journal of Integrated Chinese and Western Medicine, Vol. 20, No.2, February 2010, PP-66-68+80

[5] F. Chenhua, T. Zhipeng, “Application of Modernized Tongue Diagnosis in Clinical Research”, Chinese Journal of Information on Traditional Chinese Medicine, Vol.23, No.6, June 2016, PP-119-122

[6] C. Jirong, L. Jiajin, “Reviewing the big data solution based on Hadoop ecosystem”, Computer engineering&Science, Vol.35, No. 10, October 2013, PP-25-35

[7] Wullianallur.R, Viju.R, “Big data analytics in healthcare: promise and potential”, Health information science and systems, Vol.2, No. 2, February 2014, PP-3-12

[8] Hashem.IAT, Yaqoob.I, Anuar. N.B, Mokhtar .Salimah, Gani .A, Khan .S.U, “The rise of “big data” on cloud computing: Review and open research issues”, Information Systems, Vol. 47, No.1, January 2015, PP-98-115

[9] P. Fubin, “Research and Implementation of the Secure Cloud Storage System Based on Hadoop”, Sichuan: University of Electronic Science and Technology of China, 2013

[10] C. Tianying, C. Jianfeng, “Intelligent data masking system in big data Environment”, Communications Technology, Vol.49, No.7, July 2016, PP-915-922

[11] Q. Hongming, L. Huan, “Discussion on Data Masking Oriented to Big Data Application for Operators”, mobile communication, Vol.39, No.13, July 2015, PP-17-20+24

[12] L. Minghui, Z. Ni, Z. Yunyong, H. Kun, G. Xue, Q. Dalin, “Research on Sensitive Data Protection Technology on Cloud Computing”, Telecommunications Science, Vol. 30, No.11, November 2014, PP-2-8

[13] W. Yan, U. Brahmakshatriya, Y. Xue, M. Gilder, B. Wise, “p-PIC: parallel power iteration clustering for big data”, Journal of Parallel and Distributed Computing, Vol.73, No.3, March 2012, PP-352-359.

[14] W. Aiming, Z. Zhongxu, S. Lansun, “Study on the Classification method of tongue Color and Coat Color in automatic Analysis of tongue Image of traditional Chinese Medicine”, Beijing Biomedical Engineering, Vol. 19, No.3, September 2010, PP-136-142

[15] Y. Yaqing, T. Xuelong, Y. Chunhong, “Classification and present situation of Medical Image Registration methods”, Journal of Chongqing University, Vol. 26, No.8, August 2013, PP-114-118