Application of Big Data Technology in Extracting Information Analysis of Traditional Chinese Medicine

Jing Sun and Xinxin Meng*
Shaanxi University of Chinese Medicine, Shaanxi, China

*Corresponding author e-mail: 505054595@sntcm.edu.cn

Abstract. With the advent of the Internet, big data has penetrated into all aspects of life. In such an environment, Chinese medicine needs not only inheritance, but also advance with the times, innovative thinking and technology. This article takes big data processing and analysis as the goal, and analyzes the application scenarios of big data in Chinese medicine analysis. The model adopts APRIORI algorithm and text clustering, which improves the application of big data technology in the analysis of information extraction of traditional Chinese medicine. The test results of traditional Chinese medicines show that the use of basic big data to process the data, combined with the K-Means text clustering calculation method, the scientific conclusions obtained in the medical and health industry application rate has increased by 18%, these conclusions provide a certain degree for Chinese medicine manufacturers Help to improve production efficiency and product quality. However, the basic large database of Chinese medicine in my country has not been established, which makes the use of big data unable to be fully implemented.

Keywords: Big Data Technology; Chinese Medicine Information Analysis; Text Clustering Algorithm; Association Rule Algorithm

1. Introduction
By accelerating the process of global economic integration, international competition has intensified, and science and technology have become particularly important in economic development. With policy encouragement and financial support from governments of various countries, various scientific research institutions and enterprises have stepped up scientific research and innovation due to the needs of the times and the pressure of survival. The research and application fields of science and technology have been greatly expanded, and technological innovation has sprung up. Burst out in various fields [1-3]. Therefore, in order to change the status quo of China's traditional Chinese medicine industry, which is large but not strong, and to enhance the status of China's traditional Chinese medicine industry in the world, it is necessary to improve the technological research and development capabilities of producing high-quality Chinese medicine. Advanced TCM technology research and development capabilities are the foundation and key to the production of TCM. The prerequisite for improving the level of high-quality TCM research and development technology is to strengthen the experimental analysis of TCM.
Big data, as a currently popular technical field, has attracted widespread attention and research. This article discusses the concept, characteristics and development status of big data at home and abroad, and analyzes the basic and cutting-edge technologies of big data, including data collection and perception, data storage and processing, analysis, visualization, security and privacy protection, Huang Y pointed out The latest research directions of these technologies are summarized, and the technical and strategic challenges they face are summarized [4-6]. Finally, the essence of the technology is analyzed, which provides important guidance for big data research and engineering applications [12].

The main innovation of this paper is that the data has good compatibility, scalability, high performance and high real-time. Because the ETL process in the data warehouse in this article is implemented based on triggers and temporary table spaces, and the application of memory database technology and column storage technology to achieve high-speed data access, memory database technology can greatly improve database read and write performance, and the list storage technology is aimed at The data warehouse application scenario of writing less reading and more applications can reduce the addressing time of column calculation data and improve the efficiency of column statistics calculation [7-11]. The ETL design based on trigger technology can trigger the data extraction process when the source database table changes, thereby ensuring the real-time nature of source data acquisition. Analysis and calculation are moved down to the storage layer to reduce computing time while reducing the amount of data transmission, thereby ensuring the real-time performance of the data analysis interface. This article uses BO data display tool combined with HANA platform in the data display layer to realize rapid refresh of tax reports.

2. Big Data Technology in Chinese Medicine Extraction Information Method

2.1. Text Clustering Technique Method

Clustering refers to dividing data into several groups based on descriptive information about the objects in the data and their relationships. These groups are called clusters. The purpose is to make the objects in the group the same, but the objects in different groups are different. The greater the similarity within the group, the greater the difference between the groups, and the better the clustering effect.

As a type of clustering, text clustering uses text similarity information to aggregate text into different clusters. Different from traditional clustering, because text is a semi-structured data form, it is impossible to perform mathematical calculation and analysis on the data form, so text-specific information needs to be processed. By segmenting the text, extracting the features of the text, and then calculating the similarity (or distance) of the text to establish the similarity (distance) matrix of the text. On this basis, the text is clustered by a specific algorithm.

Due to the particularity of text, text clustering is more complicated than general clustering. Before clustering, the text needs to be processed specifically to make the text a kind of structured data that can represent the information between the texts. In order for the computer to cluster the text, we need to represent the non-numerical text into a text representation model that the computer can understand.

2.2. APRIORI Algorithm

According to the algorithm rules of the control organization, many algorithms have been developed, such as Apriori algorithm, Apriori Hybrid algorithm, FP Growth algorithm, ECLAT algorithm, etc. These algorithms are due to the frequent item basis of the data processed by the association rules or based on the dimensionality of the data in the rules. Perform classification, etc. Apriori algorithm is a very influential one-dimensional first-level algorithm for Boolean association rules. It is considered to be one of the most classic algorithms for joining rules. The Apriori algorithm was first proposed by Agrawal and Srikant in 1993.
The prior algorithm mainly searches for frequent items in all transaction databases through multiple iterations, that is, selects items that support no less than the set minimum condition value, and constructs frequently used items in the next step to minimize trust.

First, find out all frequent 1-itemsets and record them as 1 L. On the basis of 1-itemsets, the candidate set of 2L is basically all items appearing as its candidate sets, that is, through 1-itemsets producing 2- The item set continues, and its support must be greater than the threshold set by the support. Repeat the above steps until no more k objects (k L) are found. In the process of giving birth to established candidates, you will immediately delete those candidates by establishing your own important system.

The Apriori algorithm first traverses and calculates the support of each item set in order to determine the frequent 1-item set. After determining the 1-item set, continue to calculate the 2-item set until the (k-1)th frequent found The item set 1 k L and the Apriori-gen function generate a candidate item set, so the production of the candidate item set is the core of the entire algorithm, which is generated by the Apriori-gen function. Next, scan the database to calculate support for the candidate units specified in K C. Because the entire Apriori algorithm is computationally large and requires a lot of data to be scanned, many scholars have improved the Apriori algorithm, such as the later proposed AprioriHybird algorithm, FP growth algorithm, etc., these algorithms have been optimized on the basis of the Apriori algorithm And improvement.

Suppose \( E = \{e_1, e_2, e_3, \ldots, e_n\} \) is a collection of data items, and \( F = \{t_1, t_2, t_3, \ldots, t_n\} \) is a transaction database, where each transaction \( Y \) is a subset of the data item set, that is, each transaction \( t \) of \( t \subseteq E \) has an identifier \( E t \) associated with it. When the subset of \( X \) is satisfied, the thing \( t \) is called the set containing \( X \).

For example, in the extraction of traditional Chinese medicine information, the extraction process is a and the solid content is b, then the association rule R can be used to express: R: ab. The important concepts in association rules are support and confidence, as shown in formula (1) and formula (2) respectively:

\[
\text{sup port}(a \Rightarrow b) = \text{sup port}(a \cup b)
\]

\[
\text{confidence}(X \Rightarrow Y) = \frac{\text{sup port}(a \cup b)}{\text{sup port}(a)}
\]

In practical applications, only associations with higher support and trust can be used as useful association rules. The minimum support represents the lowest importance of the data item in the statistical sense, and only the data item set that meets the minimum support can appear in the association rule, which is called the universal object set; the minimum confidence indicates the minimum reliability of the association rule. The rules that meet the minimum support and minimum confidence are called strong rules. The task of mining association rules is to detect all frequent object sets and destroy all powerful rules in the transaction database D.

3. Information Extraction Experiment of Big Data Technology in Chinese Medicine

3.1. Big Data Technology in Chinese Medicine Information Extraction

In the extraction of traditional Chinese medicine, due to the need for a lot of quality inspection data, it uses the iTleai1s algorithm to cluster analysis on the seven attributes of boiling liquid relative density, detection temperature, solid content and relative density, detection temperature, solid content and spray drying humidity. To obtain the above quality inspection indicators. The data distribution thus reasonably optimizes the current standards of the enterprise. Now take the detection index extracted from the traditional Chinese medicine Jiangya avoidfeng tablets as an example for illustration, and the results are shown in Table 1.
Table 1. Extraction of the test index preparation solution in production

|                 | Mixing liquid | Boiled liquid | Spray drying moisture/% |
|-----------------|---------------|---------------|------------------------|
| Relative density | 1.07-1.2      | 1.07-1.2      |                       |
| Detection       | 56-62         | 56-62         |                       |
| Solid content%  | 31            | 17-22         | ≤ 8.1                  |

According to the 7 attributes of the quality inspection data of traditional Chinese medicine extracted in Table 2, the initial parameters of k-means clustering are used to compare the advantages and disadvantages. The parameter range of the number of clusters is 2 to 4. The parameter corresponding to the minimum value of goodnessIndex is selected as the Nepalese The initial parameters of a mean clustering algorithm. Through calculation, if accurate clustering is to be obtained, the division of the goodnessIndex value of the quality factor should be small, so the k-means clustering result can be obtained by substituting the parameters into the corresponding function according to the best clustering data.

4. Big Data Technology in Chinese Medicine Information Analysis

4.1. Quality Inspection Data of Traditional Chinese Medicine Extraction

Since Figure 1 uses the best initial parameters selected by the system as the initial parameters, the k-means clustering algorithm is used to cluster the 7 quality inspection indicators and the results are obtained, so the operator can study based on the above results to be more in line with the production process. According to the quality testing requirements, if a certain interval has the largest amount of parameter data, the interval can be used as the reference value of the quality inspection index for data analysis.
Based on the clustering results, first sort the clusters according to their proportions, and extract ten large clusters that account for more than 90% of the total, and then select several high-frequency features from each cluster words are analyzed, and the final topic words are obtained by analyzing the connection between the characteristic words and the meaning of the characteristic words.

4.2. Discretized Continuous Data Analysis

Through data discretization, the current association rule discovery data is processed, and the data is input by the way that the user freely drafts the number of groups and divides each attribute with equal width. In the traditional Chinese medicine extraction companies, in order to obtain a better blending liquid and solid content, the improved FP-growth algorithm can be used. Before generating the FP number, the transaction data sequence is sorted in descending order of common items, so as to search for public under ideal conditions Prefix, thereby reducing search time. According to this algorithm, all frequent item patterns can be quickly mined, and the mining efficiency can be significantly improved. Discretized continuous data division interval is shown in Figure 2.
4.3
2.5
3.5
2.4
4.4
1.8
Lot number 1 Lot number 2 Lot number 3
Value

Mixed liquid solid content Liquid intake Concentrate

Figure 2. Discretized continuous data divided into intervals

It can be seen from the Figure that batch number 1 has a trend of declining and increasing, batch number 2 has a slow growth trend first, and batch number 3 has a rapid growth first and then a sharp decline. From this picture, we can conclude that the solid content, liquid intake and concentrated liquid of different batch numbers show different growth and decline trends. This allows me to have a good understanding of the batch numbers required under different circumstances.

5. Conclusions

Although this paper has made certain research results on the text clustering technique and APRIORI algorithm, there are still many shortcomings. There are still many in-depth contents worth studying in the method of extracting Chinese medicine information based on big data. There are still many steps in the process of extracting information that have not been covered because of space and personal ability. In addition, the actual application effect of the improved algorithm can only be compared with the traditional model from the level of theory and simulation.

References
[1] Wang Y. Research on internet application of big data in traditional Chinese medicine based on ancient Chinese medicine books[J]. Boletin Tecnico/Technical Bulletin, 2017, 55(10):114-119.
[2] Liu Y, Sun M, Yao H, et al. Reflections on the application status of big data technology in traditional Chinese medicine[J]. Chinese Journal of Evidence-Based Medicine, 2018, 18(11):1180-1185.
[3] Wang, X, Y, et al. STUDY ON RESOURCES INTEGRATION OF TRADITIONAL CHINESE MEDICINE DIGITAL LIBRARY BASED ON BIG DATA[J]. Basic & clinical pharmacology & toxicology. 2016, 118(Suppl.1):73-74.
[4] Liu K J, Cao Y D, Hu Y, et al. Application status and development of big data in medical education in China[J]. TMR Medical data mining, 2019, 002(003):P.118-125.
[5] Liu J, Zhai X, Liao X. Bibliometric analysis on cardiovascular disease treated by traditional Chinese medicines based on big data[J]. International Journal of Parallel Emergent and
Distributed Systems, 2019(1):1-17.

[6] Qian-wen, Ma, Yong, et al. Effectiveness of co-treatment with traditional Chinese medicine and letrozole for polycystic ovary syndrome: a meta-analysis[J]. Journal of Integrative Medicine, 2017, 02(v.15):23-29.

[7] Chen X, Wu Q. Research on the Method of Big Data Collecting, Storing and Analyzing of Tongue Diagnosis System[J]. IOP Conference Series Materials Science and Engineering, 2018, 322(6):062013.

[8] Yu Y D, Xiu Y P, Li Y F, et al. To Explore the Mechanism and Equivalent Molecular Group of Fuxin Mixture in Treating Heart Failure Based on Network Pharmacology[J]. Evidence-based Complementary and Alternative Medicine, 2020, 2020(3):1-10.

[9] When Big Data Comes Into TCM Clinic Research—Analysis of Professor Ouyang Chensheng’s Experience in Treating Lung Cancer Based on Data Mining[J]. Traditional Chinese Medicine, 2020, 09(4):332-341.

[10] Rong P J, Wang Y, Xu N G. Brain science promotes the development of acupuncture in treating brain diseases[J]. Zhen ci yan jiu = Acupuncture research / [Zhongguo yi xue ke xue yuan Yi xue qing bao yan jiu suo bian ji], 2019, 44(12):859-862.

[11] Shi J, Chen G. From solar terms to medical terms (Part II): Some implications for traditional Chinese Medicine[J]. Science China Earth Sciences, 2017, 60(11):1-8.

[12] Huang Y, Yu Z, Xie C, et al. Study on the Application of Electric Power Big Data Technology in Power System Simulation[J]. Proceedings of the Csee, 2015, 35(1):13-22.