AN ANALYZING ALGORITHM BASED ON LEARNING AND SEARCHING IN CHINESE MEDICAL BIG DATA

LUO Jie, ZHOU Ziyang, SONG Qiaolin and QIU Qin'nan
*LI Zhimin

Medical Technology College, Zhejiang Chinese Medical University, Hangzhou, China, 310053

ABSTRACT

We propose an improved analysis algorithm based on learning and searching methods in the field of Big Data to optimize TCM information management. We use TF-IDF theory in document clustering to cluster the name of Chinese Medical prescripts, construct word bag model, and combineuniversal hash with perfect hash,in order to establish a special key-value hashing band between Chinese Medical Data and diseases.

KEY WORDS:

TCM, Chinese Medical Massive Data, Big Data, Hashing, Index, Mapping

1. INTRODUCTION

1.1. TCM Big Data

With the dramatic development of information technologies, the memory forms of massive data generated by traditional Chinese medicine(TCM) clinical courses is transforming from paper to digital methods gradually. At the same time, the scale of TCM diagnosis and treatment data shows explosive growth. How to deal with contemporary TCM big data by information technologies has become a mainstream task. Facing with a large TCM database, manual retrieval of ancient Chinese medicine books is a very heavy work, which not only consumes a lot of time and manpower, but also easily lead to errors and omissions.

In the existing medical query system, the query system based on single table database retrieval and the single coding query system are more common. These two systems usually search the whole table index in retrieval, so the overhead of retrieval and update is extremely high when facing with mass data. Another shortage is that they can only target at one set of specific TCM data retrieval at a time, and have poor effect in data analysis and statistics, which greatly hinders TCM data mining and big data processing.

For mass data analysis and processing, there are two mainstream solutions in the field of information technology:

1) Adopt memory database and optimize compilation environment. Such schemes load relevant data into memory, thereby reducing I/O overhead. However, due to the strict database ACID

Dhinaharan Nagamalai et al. (Eds) : COSIT, AIAPP, DMA, SEC - 2019
pp. 21–32, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90203

principle (atomicity, consistency, isolation, and durability), there will be unnecessary overhead and limitations for some applications with weak consistency requirements.

2) MapReduce framework[4]. The MapReduce framework consists of Map function and Reduce function. In this framework, key-value mapping is always used as a key operation to achieve good scalability and fault tolerance. But at the same time, because of the frequent I/O operations, the overhead of Map function is enormous when facing with massive amounts of data that need to be processed in real time. At present, the mainstream solutions tend to improve hardware and the schemes of memory framework. But there are few studies aiming at optimizing in the level of algorithm.

1.2 Research Process Framework

Based on the idea of machine learning, this article proposes an improved analysis algorithm to process TCM big data which is a particular but generic object. This algorithm is able to learn TCM massive data by utilizing document clustering algorithm and generate hashing weight value to map and search TCM data in hash table. In addition , we compare and evaluate the efficiency between original index model and improved model as the scale of data increasing dramatically.
In the second part of this article, we introduce the hashing framework and original index model. In the third part, we propose an improved index model and. In the fourth part, we describe the details of the process of improved hashing method. In the fifth part, we compare the efficiency and overhead between two methods and evaluate them. The limitation and future improvements are discussed in the sixth part and conclusion is given in the last part.

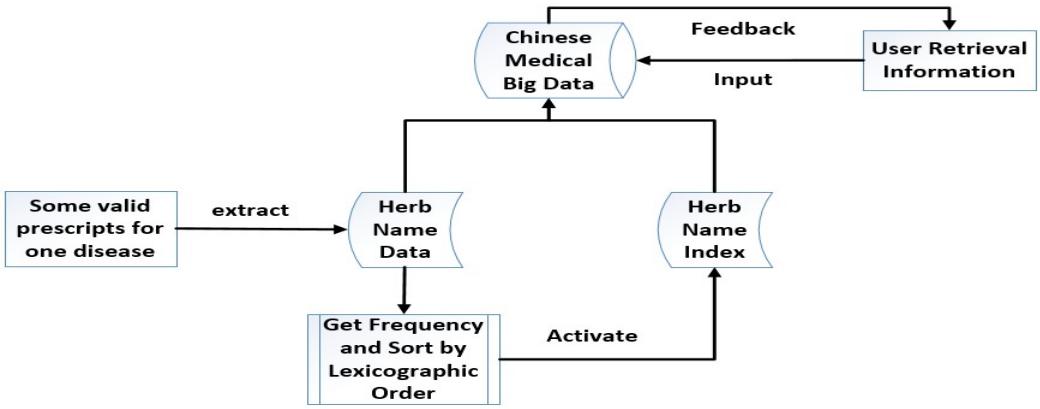
2. INDEX INITIALIZATION

Existing TCM big data includes massive traditional Chinese medicine(TCM) literature and a large number of TCM materials. And due to the complexity of TCM diagnosis and treatment, as well as a variety of TCM materials, there are tens of millions of actual clinical prescriptions for the treatment of diseases. So we select only one of many diseases and conduct a simulation experiment aiming at TCM materials and prescriptions of this disease.

2.1. Index Initialization Model

This article sort data by frequency dictionary in Universal System Index, that is, the index is generated and continuously allocated memory units according to the occurrence frequency of TCM materials.

Flow Chart 1. Index initialization model.

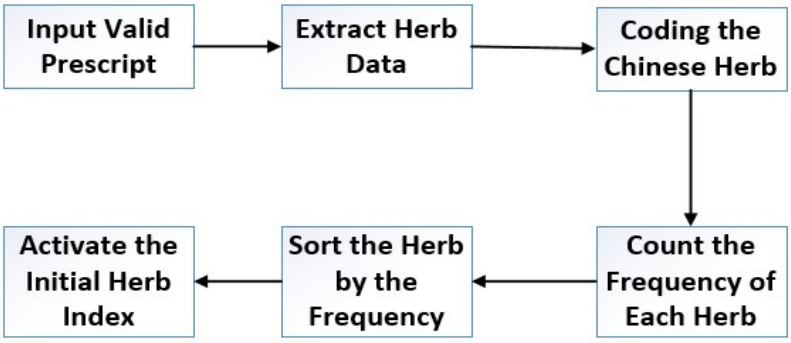


2.2. Modelling Process and Results

Each clinical prescription includes the data of TCM materials and dosages. The name of TCM materials are consisted by Chinese characters and some of them are rarely-used Chinese characters, so we code all materials by numbers and ensure each code of materials is distinct. And then the TCM materials are sorted to generate an initialization index and stored sequentially in database.

A basic index address allocation scheme of TCM database can be obtained to store massive data of TCM in order of frequency. In this way, the system can maintain and use the index to retrieve TCM massive data.

Flow Chart 2. Process of Index Initialization.



3. AN IMPROVED LEARNING MODEL

In this part, we provide an improved analyzing method to excavate TCM data. We use the same test data in this method to compare the efficiency and overhead with the original index method.

3.1. Task Object

The scale of massive TCM data not only is exponential, but will also expand with the development in the relevant fields. So the first step in this task is to design a learning system, classifying and clustering disordered TCM data to obtain scientific frequency of TCM materials.We illustrate the learning task as follows:

Task Framework 1.

Task of TCM prescription leaning
Task T: classify the materials of valid TCM prescription
Performance standard P: percentages of materials in this prescription and others
Training experience E: valid prescription data in training set

We choose one of many diseases as task object. Prescriptions for this disease vary in material dosages depending on the disease's subdivision and severity. Many doctors tend to use aliases of herbs in their prescriptions as well as omitting or elaborating the preparation methods of TCM. So we select 20 clinical prescriptions in the same category as experimental object. Then we extract and code the Chinese name and dosage of materials, generating 10 separate documents and the matrix of the name of TCM materials. In this way, a scientific TCM dictionary can be built based on document clustering.

3.2. Bag of Words - a Model of Prescription

In the modelling process, a valid prescription can be seen as a Bag of Words[18].The name of components in prescription can be seen as term vectors. In this way, we can utilize TF-IDF(term-frequency&inverse-document-frequency) to learn and analysis prescription data.

TF-IDF is a data statistics method used to analyze the importance of words in documents[19]. TF(term-frequency) refers to the frequency with which the word appears in this document. IDF(inverse-document-frequency) refers to the frequency with which the word appears in other documents[20]. The main idea of TF-IDF is: if a word appears more frequently in this paper and less frequently in other documents, it indicates that the word has a good degree of discrimination, that is, the word is more important to this document[21].

Referring to this idea, it can be considered that if a TCM material appears repeatedly in multiple prescriptions of different categories, it is unlikely to play a key role in the diagnosis and treatment of this disease, that is, it is a auxiliary ingredient or tonic maintenance type. But if the TF value of this material is large, its auxiliary proportion in such diseases can be approximately considered to be large.

We quantitatively count the dosages of TCM materials appearing in single prescription. The dosages of TCM materials are taken as the quantitative criteria of frequency to generate valid prescription document. Finally, TF-IDF frequency statistics are added into this document to construct TCM data dictionary which is shown as blow:

Table 1. TCM data dictionary（excerpts）

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
八月扎	0	0	0	0	0	12	0	0	15	0
白芍	0	0	0	0	0	0	0	0	12	0
白术	0	0	0	0	0	0	15	0	12	0
败酱草	20	0	0	0	0	0	0	0	30	30
半枝莲	0	0	20	0	0	0	0	0	0	0
薄荷	0	9	0	0	0	0	0	0	0	0
北沙参	0	0	15	0	0	0	0	0	0	0
北秫米	0	0	0	0	0	0	0	30	30	0
蝉衣	6	9	0	0	0	0	0	0	0	0
炒稻芽	30	15	0	0	30	0	0	30	0	0
炒黄连	0	0	0	0	0	0	0	6	0	0
炒黄芩	0	15	15	0	0	0	0	0	0	0
炒竹茹	15	0	0	0	0	0	0	15	0	0
陈皮	12	12	0	0	0	12	0	0	0	0
赤芍	0	0	0	15	0	12	15	0	0	0
川朴	0	12	0	0	0	0	0	0	0	0
川朴花	12	0	0	0	9	0	0	0	0	0
淡附子	0	0	0	10	0	0	10	0	0	0

Note: In this table, 0 means the corresponding material does not appear in this prescription. And the name of TCM materials are presented in Chinese to avoid ambiguities in translation.

The TCM data dictionary can lay a foundation to the universal hashing. The TF-IDF value in this table can be calculated as follows:

Formula 1.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Formula 2.

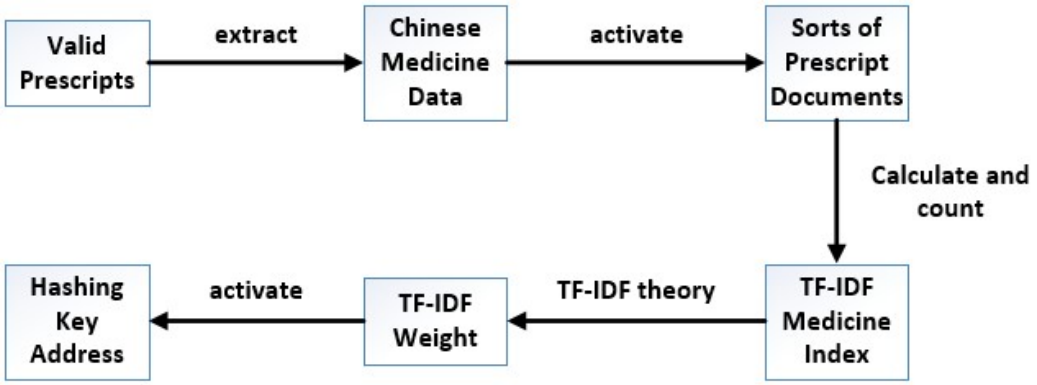
$$idf_i = \log \frac{|P|}{\left|\{j : t_i \in p_j\}\right|}$$

In formula 1, $n_{i,j}$ means the times that materials appear in prescription p_j ; $\sum_k n_{k,j}$ is the sum times that materials appear in all prescriptions. In formula 2, |P| is the sum of prescriptions in this table; $|\{ j : t_i \in p_j \}|$ means the number of prescriptions which contains material t_i . When a kind of material is not in a prescription, the corresponding value would be 0, which could result in a program error. So we add 1 to the denominator of the second formula to avoid that.

3.3. Algorithmic Learning Flow Chart

As we can see in this flow chart, we extract TCM prescription data and generate TF-IDF value.

Flow Chart 3. TCM data learning.



4. IMPROVED HASHING METHOD

After generating TF-IDF value, we use hashing method to further improve the algorithm next.

4.1. Universal Hashing Representation

We can infer from the TF-IDF value that, for a branch of a certain kind of disease, TCM material with high IDF value plays an important role in the prescription. TCM material with high TF value plays an auxiliary role in the treatment of the whole disease category. For the initial classification of TCM produced by TF-IDF, we need to map them into continuous memory space. The higher the TF-IDF value of TCM material, the higher the priority of its memory address. Therefore, we choose universal hashing to implement this task.

We first select a very large prime number so that every possible key values are smaller than the prime number, and then we construct a preliminary universal hash function as following:

Function 1.

$$h_{a,b}(k) = ((ak + b)\%\, p)\%m$$

The m in this formula above could be adjusted according to experimental results. And the cluster function of hashing is as follows:

Function 2.

$$H(p,m) = \{h_{a,b} : a \in Z_p^* 且 b \in Z_p\}$$

In this cluster function, each Z_p can be mapped to Z_m and we can adjust the value of m to reduce the possibility of collisions in hashing.

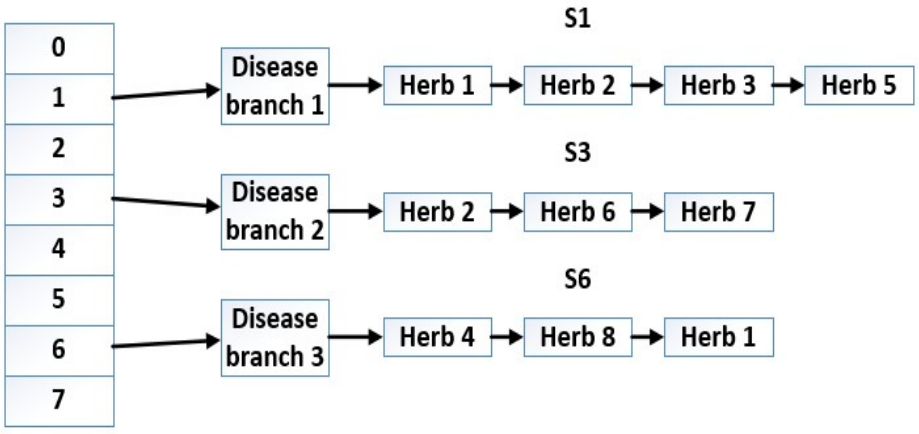
4.2. Improved Perfect Hashing

Next, we introduce the perfect hashing to improve the universal hashing. Full hashing is efficient because it has a time complexity of O(1). For the massive data of TCM, we use two-level perfect hashing, and in each level, we use universal hashing. In the first level of perfect hashing, the hash function h is selected from a universal hash cluster to hash n keywords into m slots; next, in the second level of hashing, the size of the hash table is the square of the number of keys hashed into a particular slot[22] for which can reduce the conflicts in hash and can carry out the second hash depending on TF-IDF values.

With this improved perfect hashing, we use TCM material with a high weight value as the key, and put the rest of the disease information into the corresponding slot. When retrieving, we can quickly use this connection to search drug information or disease information.

4.3. The Structure of Hashing

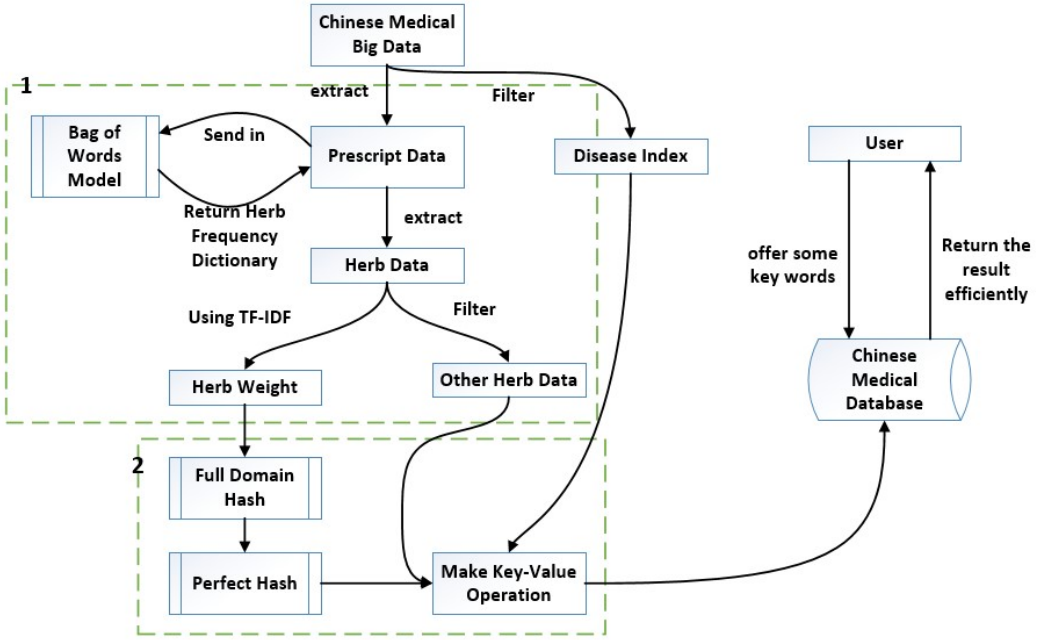
Model 1. memory structure of perfect hashing.



In case of conflicts, the chain address method is adopted for horizontal address expansion, in which the chain structure is in the order of weight value, so that the distance of pointer movement can be saved in querying to achieve a efficient retrieval.

The model is mainly divided into two modules: word bag processing module and hash module. In the word bag processing module, TCM materials are classified depending on prescriptions they are in and the corresponding TF-IDF values are calculated. In hash module, the data of materials, prescriptions and diseases is assigned memory addresses.

Flow Chart 4. process of improved method.



5. EVALUATION

5.1. Analysis About Index Initialization Model

Index Initialization Model is a statistical word frequency structure. When maintaining data, we chose dichotomy to insert the new data of TCM material to avoid breaking the overall alphabetical order, so the pure time complexity is O(log n). This is called pure time complexity because we also need to manipulate the amount of data inserted and other processing. Since word frequency needs to be counted and reordered, the time complexity of the whole structure[23] is shown in the following table:

Table 2. Time complexity of Index Initialization.

	Average case	Worst case
query	$O(\log n)$	$O(\log n)$
insert	$O(\log n)$	$O(\log n)$
delete	$O(\log n)$	$O(\log n)$

The time overhead of word frequency is relatively ideal , but what we should pay attention to is that this structure only construct an index. In this way, the information in the database needs to be invoked again during the query. Since each query depending on the index, the system have to switch between the index and the database to input and output. Therefore, IO overhead would be extremely large when facing with actual TCM mass data.

As for space overhead, although the space complexity of Index is O(n), the scale of index would expand as the amount of data continues to grow. In this case, system have to allocate additional memory space for index, which is a large waste of space overhead.

5.2. Analysis About Improved Algorithm

The improved algorithm is divided into word bag and hashing, so we need to analyze these two parts respectively.

Firstly, the bag of word operates on matrix through the fast matrix algorithm and the idea of dimensionality reduction. The time complexity is O(tkn) where t is the number of matrix iterations, k and n are the rows and columns of the matrix after dimension reduction, respectively[24]. It doesn't seem to be as efficient as index initialization, but what we should pay attention to is that, the bag of word is a training set generated from a small part of TCM data. Analyzing and summarizing the overall rules of TCM data from a small part saves a lot of overhead. In addition, the generated machine learning framework can be used for over time.

Secondly, in hash module, the time complexity depends on hash function. The time complexity of hash function is linear in text selecting, that is, the best time complexity is O(1) and the worst case is O(n). As for space overhead, the memory space needed is in the level of O(n) because the whole hash structure can be embedded in the structure of TCM database.

Overall, combined with these two modules, the improved algorithm has lower overhead and higher efficiency when retrieving not only on time but also on space.

5.3. Initializes Index Model Evaluation

The initialization index model is established according to the occurrence frequency of TCM materials. The results of sample frequency data are shown in the figure below(In order to avoid ambiguity in translation, the names of TCM data are all in Chinese):

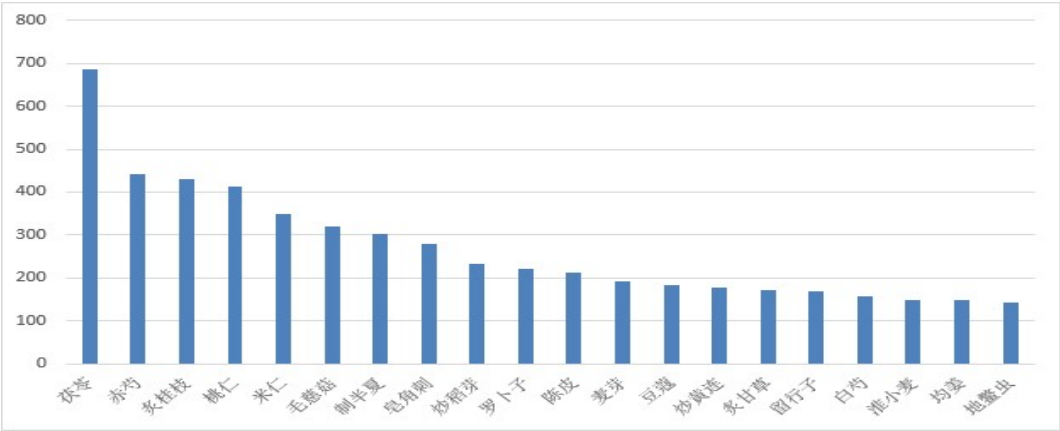


Figure 1. Part of the frequency table of TCM materials corresponding to a disease.

According to this figure, the frequency of Fuling is relatively high, but the efficacy of Fuling mainly plays a nourishing role, and the efficacy of primary treatment is not completely matched with the disease data selected in this paper. At the same time, the selection of medicinal materials is affected by many factors, such as price, efficacy, and patients' physical condition. Therefore, the initialization index method lacks rationality in the data analysis of TCM, and the logical

correlation between the data in the model structure is poor, which leads to the disorder of the logical classification and retrieval, resulting in redundant overhead.

5.4. Improved Algorithm Evaluation

In the initial learning analysis, the improved algorithm introduces tf-idf theory in the document clustering model, so that the TCM materials can be directly classified from the prescriptions and allocated with memory space continuously. So logical addresses are more centralized when retrieving so that corresponding main disease of treatment can be better excavated.

In the perfect hashing, the only possible problem is hash collision. In this case, we refer to a theorem: If n keywords are stored in a hash table and the size of this hash table m is n2 ,using a hash function H randomly selected from a universal hash function class, the probability of collision is less than 1/2 [26]. If the hash function H is selected randomly, the collision probability of any two key words is 1/N. If we set X is a random variable, when m=n2 , the expectancy of collision is as the formula below [27]:

Formula 3.

$$E[X] = \binom{n}{2} \times \frac{1}{n^2} = \frac{n^2 - n}{2} \times \frac{1}{n^2} < \frac{1}{2}$$

In this way, the collision probability of the whole perfect hash function will be less than 50%. For massive data, it can complete a better hash mapping.

5.5 Retrieval Overhead Comparison

Here we compare the time overhead between two methods, initialization index and improved perfect hashing, as the scale of data increasing.

Table 3. Experimental results.

Scale of data（10^n）	Initialization index(s)	Improved hashing(s)
1	0.187	0.186
2	0.234	0.230
3	1.026	0.503
4	1.233	0.729

As we can see from this table, improved perfect hashing method based on TF-IDF has more noticeable efficiency especially when the scale of data increasing dramatically.

6. LIMITATIONS AND IMPROVEMENTS

6.1 Limitations

TCM theory has a long history. For many difficult diseases, many regions still retain a large number of folk prescriptions, most of which lack sufficient data and theoretical support. In addition, due to the differences in patients' clinicopathological symptoms and personal physical conditions, doctors can adjust the classical prescriptions in the level of dose and material type

according to the actual situation and personal experience. Therefore, there are some difficulty in hashing and retrieval of these complex diseases and TCM data. Currently, electronic data do not account for the majority of the TCM massive data. How to unify the non-electronic data and electronic data, and how to ensure the high accuracy and efficiency of hash analysis to electronic data after it is promoted are also great challenges to the algorithm structure.

6.2 Future Improvements and Studies

We can introduce machine learning and data mining to analyze the intricate TCM semantics, names of prescriptions, diseases and materials. Through more accurate semantic analysis and clustering, we can further process data, mine and utilize its rules, aiming at the automation of TCM data processing.

7. CONCLUSIONS

This paper deeply studies the traditional Chinese medicine(TCM) massive data and propose an improved algorithm for the research and retrieval of TCM massive data. Firstly, this paper summarizes the knowledge of hashing, machine learning and other related fields, and proposes a simple initialization index algorithm to generate the processing results and take it as the control group. And then, TF-IDF theory in the field of literature clustering is used to preliminarily classify names of TCM materials and generate hash weight values. Then, the universal hash and perfect hash are used to allocate and generate key-value addresses for TCM data, so as to achieve reasonable analysis and efficient retrieval. Last but not least, the structure, logic and overhead of the two algorithms are evaluated.

Our experimental results and evaluations show that this improved algorithm is feasible in TCM data modelling and mining. It provide a method based on perfect hashing to construct TCM data dictionary which can be retrieved and managed efficiently, giving a reference to future TCM big data mining and learning. Based on this method, we will continue our research on the modernization and digitalization of TCM big data.

This thesis has done the following preparation work: 1) Foundation. Consult and study relevant data structure and data processing knowledge, and verify the correctness of the algorithm through the proof. 2) Analysis and extract data from particular experiment object, TCM prescription text, building data structure model. 3) The control experiment, through the concrete data, has carried on the actual processing to the algorithm and verified the algorithm validity. 4) Evaluation. The two algorithms are evaluated in different aspects, and the conclusion that the improved algorithm is better is obtained.

REFERENCES

[1] RenT, XiaoX, Liu X, Sun Y. Study on knowledge acquisition of TCM prescription[J]. Chinese Journal of Information of TCM,2012, 19(7):24-27.

[2] Zhou C. Chinese traditional medicine culture in overseas development status and reflection. World journal of integrated traditional Chinese and western medicine, 06(8), 733-733.

[3] Li H. (2006). Characteristics and enlightenment of the development of traditional Chinese medicine in foreign countries. Chinese journal of traditional Chinese medicine, 21(6), 359-361.

[4] J.Dean and S.Ghemawat. MapReduce, simplified data processing on large clusters. In Communications of the ACM. V.51.n.1, January 2008

[5] http://en.wikipedia.org/wiki/Hash_table, Hash table

[6] Ma R, Jiang H, &ZhangQ. (2008). An improved method for fast hash table lookup. Computer engineering and science, 30(9), 66-68.

[7] Bertoni, G. , Daemen, J. , Michaël Peeters, & Assche, G. V. . (2014). Sakura: A Flexible Coding for Tree Hashing. Applied Cryptography and Network Security. Springer International Publishing.

[8] Lu, Y. , Prabhakar, B. , & Bonomi, F. . (2006). Perfect hashing for network applications. IEEE International Symposium on Information Theory. IEEE.

[9] Freitas, A. T. . (2003). Introduction to algorithms. Resonance,1(9), 14-24. Page 132.

[10] Mark L. Krotoski.Co-author, Using "Digital Fingerprints" (or Hash Values) for Investigations and Cases Involving Electronic Evidence, United States Attorneys' Bulletin, Vol. 62

[11] Freitas, A. T. . (2003). Introduction to algorithms. Resonance,1(9), 14-24. Page 138.

[12] Fredman, M. L. , Fredman, M. L. , Fredman, M. L. , Fredman, M. L. , Komlos, J. , &Komlos, J. , et al. (2008). Storing a sparse table with O(1) worst case access time. 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982). IEEE.

[13] Hagerup, T. , & Tholey, T. . (2001). Efficient minimal perfect hashing in nearly minimal space. Symposium on Theoretical Aspects of Computer Science. Springer-Verlag.

[14] Fox, E. A. , Chen, Q. F. , & Heath, L. S. . (1992). A faster algorithm for constructing minimal perfect hash functions. ACM.

[15] Chen kejia. (2011). Review of machine learning technology in social network analysis. Journal of nanjing university of posts and telecommunications (natural science edition), 31(3), 83-89.

[16] Klix, F. . (2010). Michalski, r. s. / carbonell, g. / mitchell, t. m. (eds.), machine learning. an artificial intelligence approach. ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik, 65(11), 568-568.

[17] Han, J. . (2005). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc.

[18] http://en.wikipedia.org/wiki/Bag_of_words_model

[19] Rajaraman, Anand, Ullman, Jeffrey David.Mining of Massive Datasets: Recommendation Systems.

[20] http://en.wikipedia.org/wiki/Tf-idf, TF-IDF

[21] http://baike.baidu.com/link?url=VYoxFr1UxhS94IGqYSmSybYQuEs4NPKcYrYng2N4ru1c 1I_MUbI5BxwvDs-S15pX5PkFiSacCHapHpZaH6k2Nq, Baidu Baike, tf-idf theory

[22] Lu, Y. , Prabhakar, B. , & Bonomi, F. . (2006). Perfect hashing for network applications. IEEE International Symposium on Information Theory. IEEE.

[23] http://en.wikipedia.org/wiki/Red%E2%80%93black_tree

[24] Xu, W. , Liu, X. , & Gong, Y. . (2003). Document Clustering Based On Non-negative Matrix Factorization. DBLP.

[25] C?linescu, & Gruia. (2016). 1.61-approximation for min-power strong connectivity with two power levels. Journal of Combinatorial Optimization, 31(1), 239-259.

[26] Zhang min, gao xiao-hong, sun xiao-meng, xu jia-tong, li xiang-yan, & shi yanjie, et al. (2008). Pharmacological action and research progress of poria cocos. Journal of beihua university (nature), 9(1), 63-68.

[27] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein INTRODUCTION TO ALGORITHMS, Hashing, 2006

AUTHORS

LUOJie, Master of engineering; associate professor in Medical Technology College, Zhejiang Chinese Medical University; member of computer society of China.

ZHOUZIyang, Bachelor of engineering,undergraduate in Medical Technology College, Zhejiang Chinese Medical University.

SONG Qiaolin, attending doctor of traditional Chinese medicine, master's degree, research direction: clinical teaching management of traditional Chinese medicine, clinical tumor of integrated traditional Chinese and western medicine

QIU Qin'nan, Bachelor of engineering,undergraduate in Medical Technology College, Zhejiang Chinese Medical University.

Li Zhimin, associate professor,majoring in medical data processing and analysis. Corresponding author of this article.