Text Mining and Analysis of Treatise on Febrile Diseases Based on Natural Language Processing

Kai Zhao¹, Na Shi², Zhen Sa³, Hua-Xing Wang⁴, Chun-Hua Lu⁵, Xiao-Ying Xu⁶
¹School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, ²School of Life Science, Beijing University of Chinese Medicine Beijing 100029, China

Abstract

Objective: With using natural language processing (NLP) technology to analyze and process the text of “Treatise on Febrile Diseases (TFDs)” for the sake of finding important information, this paper attempts to apply NLP in the field of text mining of traditional Chinese medicine (TCM) literature. Materials and Methods: Based on the Python language, the experiment invoked the NLP toolkit such as Jieba, nltk, gensim, and sklearn library, and combined with Excel and Word software. The text of “TFDs” was sequentially cleaned, segmented, and moved the stopped words, and then implementing word frequency statistics and analysis, keyword extraction, named entity recognition (NER) and other operations, finally calculating text similarity. Results: Jieba can accurately identify the herbal name in “TFDs.” Word frequency statistics based on the word segmentation found that “warm therapy” is an important treatment of “TFDs.” Guizhi decoction is the main prescription, and five core decoctions are identified. Keyword extraction based on the term “frequency-inverse document frequency” algorithm is ideal. The accuracy of NER in “TFDs” is about 86%; latent semantic indexing model calculating the similarity, “Understanding of Synopsis of Golden Chamber (SGC)” is much more similar with “SGC” than with “TFDs.” The results meet expectation. Conclusions: It lays a research foundation for applying NLP to the field of text mining of unstructured TCM literature. With the combination of deep learning technology, NLP as an important branch of artificial intelligence will have broader application prospective in the field of text mining in TCM literature and construction of TCM knowledge graph as well as TCM knowledge services.

Keywords: Knowledge discovery, natural language processing, text mining, traditional Chinese medicine literature, treatise on febrile diseases

INTRODUCTION

“Treatise on Febrile Diseases (TFDs)” was written by Zhang Zhongjing in the late Eastern Han Dynasty. The history of this book is more than 2000 years. The theory of traditional Chinese medicine (TCM) syndrome differentiation and treatment originates from this book which records the beginning of Chinese decoctions. The effectiveness of TCM is mainly from the “TFDs.” Ancient and modern skilled doctors of TCM are always famous for their expertise in using the therapeutic methods of “TFDs.” Sun Simiao, who is known as the king of Chinese medicine, spoke highly of “TFDs” in his masterpiece “Valuable Prescriptions for Emergency.” He said that “The therapy of Zhang Zhongjing has a special magical power. As long as, I carry it out, it has never failed.”[1] “The Summary of Complete Collection in Four Treasures” mentioned that “as long as you have a smattering of the treatment in this book, you can bring the dying back to life.”[2] The description of disease in “TFDs” is different from the definition and classification of diseases in modern medicine. The angle of view on understanding diseases in this book was generated from oriental philosophy, which was very different from western philosophy and provided a completely different perspective on research and treat illnesses. It is precisely because of the different methods of observation and research of diseases that the book agglomerates special therapies of ancient Chinese sage doctors which distinguished from modern medicine. This book is a masterpiece that summarized the common theories of TCM in later ages. It is a valuable resource for the study of traditional Chinese medicine.
clinical experiences. It recorded rules of diseases through observing the performances and developments of diseases by ancient Chinese doctors and incorporated the complete contents of principle-method-recipe-medicines of TCM. Most of the treatment methods recorded in this book have been clinically validated for thousands of years in ancient China. “TFDs” defined and treated diseases according to symptom complex. At the very beginning, ancient Chinese doctors observed that some certain Chinese herbs had a special effect on eliminating certain syndromes. Then, the doctor gave the corresponding Chinese herb to the patient according to his symptom. However, the doctor only could give a single Chinese herbal medicine to treat one symptom of the patient at one time. The ancient Chinese doctors afterward observed that some symptoms often accompanied with each other. For example, in some cases, symptoms appeared in groups, such as a patient who got a fever may commonly accompany by aversion to cold, who vomited may commonly accompany by another symptom of diarrhea, so that a single Chinese herb may not work well very in this situation. Therefore, ancient Chinese doctors began to mix a variety of Chinese herbs. Although every single Chinese herb could once eliminate a specific symptom, a soup with different functions, which were made from a mixture of Chinese herbal medicines, could eliminate the complex syndromes performed by patients. After long-term clinical practice, many syndromes were found to be repeatedly occurred among the population. The Chinese herbal formulas, which could eliminate the specific syndromes and were suitable for the majority of people, were formed. “TFDs” recorded the therapeutic prescriptions on specific syndromes, which have never changed in the human body for thousands of years and were repeatedly performed. Thus, the prescriptions created at that time are still worked well today. Correspondence of prescription and syndrome is a basic theory of “TFDs.” Prescription and syndrome differentiation is a simple and efficient method to use the treatment principles of “TFDs.” According to the syndrome, the patient can be treated with the corresponding prescription. In the consequences, the diagnosis and treatment process can be simplified into a matching problem, which means that there is a high correlation between the prescription and the syndrome. All the while, exploring and utilizing the valuable clinical information from “TFDs” is a research hotspot in the field of TCM.

Natural language processing (NLP) is an important subdomain of artificial intelligence and is also one of the most difficult problems in this field. It mainly about the theory and method of automatically analyzing and characterizing human natural language using computational technology. NLP technology is mainly used in the field of big data such as intelligent questioning and answering, public opinions analysis, semantic understanding, machine translation, and knowledge graph. The application of NLP in the medical field has increasingly become a tendency. A recent research suggests that scientists can build a medical clinical decision support system by using text input when applying NLP. Based on the International Business Machines Corporation Watson, IBM researchers have used NLP and supervised learning techniques to automatically identify disease status. There are also researchers using NLP technology for information extraction and knowledge discovery from clinical cases. The TCM literature is a huge treasure of medical knowledge. In recent years, as the TCM literature has more and more become electronic version, it is a difficult task on how to use modern computer technology to quickly and easily discover new therapies and knowledge from the vast number of the TCM literature. This paper is mainly about the basic techniques of processing and analyzing the text of “TFDs” by means of NLP, and then obtaining some rules of treatments and prescriptions in the “TFDs.” At the end of this paper, the text categorization is an important application in the field of NLP, we calculate text similarities of “Understanding of Synopsis of Golden Chamber (USGC)” versus “Synopsis of Golden Chamber (SGC)” and versus “TFDs” based on the previous text mining and processing, which illustrates the possibility of further application of NLP technology in the field of TCM literature text mining.

MATERIALS AND METHODS

Source of literature
Zhao Kaimei’s version of “TFDs” in Song Dynasty, which is the most thoroughly studied version at present.

Processing strategy
This paper is based on Python 3 which invokes NLP toolkit such as Jieba, nltk, gensim, and sklearn, meanwhile combined with Excel and Word software to implement words segmentation, stop words removing, words frequency statistics, and other operations on the text of “TFDs.”

Inclusion and exclusion criteria
Incorporating the chapter from “Differentiation of pulse and syndrome related to Taiyang disease and treatment” to “Differentiation of yin and yang, examination of pulse and treatment of syndrome” in “TFDs.” In total 247 clauses which include prescriptions after removal of the soil melon roots, limonitum pill, the honey-fried recipe, the pig bile recipe, the powder of muskmelon pedicel, Shaohui powder which are no longer appropriate from “TFDs” are selected.

Data analysis
The basic processes of NLP include word segmentation, removing stop words, keyword extraction, and named entity recognition (NER).

Word segmentation
Word segmentation is the primary task of NLP technology. First of all, using Python’s Jieba module to implement Chinese word segmentation for the text of “TFDs.” Jieba is a free open source Chinese word process module particularly developed for Python. It is mainly based on the directed acyclic graph, dynamic programming to find the maximum path, hidden Markov models, Viterbi and other algorithms, which can achieve word segmentation, part-of-speech tagging, and
Keyword extraction. After the word segmentation, loading the Chinese stop word list.

Word frequency statistical analysis
This paper compares the results of two methods, namely word frequency statistics by Python and searches word frequency by Word software.

1. Word frequency statistics by Python: using Python to count word frequency and sorting the results, manually selecting the names of herbal medicines and words describing the syndrome, then sorting the statistical results by Excel
2. Search word frequency with word: using Word software to search frequency of 56 syndrome words given by Python, and comparing the difference with Python word frequency statistics
3. Word frequency statistics of prescription clauses: Using Excel software to analyze the word frequency of 107 prescriptions in “TFDs.” Invoking nltk toolkit in Python to draw a word frequency discrete distribution graph.

Keyword extraction
Term “frequency-inverse document frequency” (TF-IDF) is called the inverse text frequency index, which can reflect the importance of a word in an article. It is often used as a rating method for search engine and has a wide range of applications in the field of NLP.[13] The equations of TF-IDF were illustrated as follows:

\[ tf_{ij} = \sum_{d \in D} n_{ij} \]
\[ idf_i = \log \left( \frac{|D|}{|\{ j : t_i \in d_j \}|} \right) \]
\[ TF-IDF = TF \times IDF \]

\( n_{ij} \) is the number of occurrences of the word in the file \( d_j \), and the denominator is the sum of the occurrences of all the words in the file \( d_j \). \( |D| \) is the total number of files in the corpus. \( |\{ j : t_i \in d_j \}| \) indicates the number of files containing the word \( t_i \) (i.e., the number of files \( n_{ij} \) ≠ 0). If the word is not in the corpus, it will result in a denominator of zero, so in general use \( 1 + |\{ j : t_i \in d_j \}| \).

TF-IDF is high when a word appears frequently in a document; the same word meanwhile rarely appears in other documents in the corpus; in other words, this word is representative.

This paper uses TF-IDF algorithm in Jieba to extract keywords from the text of “TFDs.”

Named entity recognition
NER refers to the identification and extraction of entities with specific meanings in the text, such as names of people, places, and terminology. NER is a basic step of NLP. Previous research suggested that using the method of conditional random field to study NER for “TFDs.”[14] We use Jieba Chinese processing toolkit with TextRank algorithm to implement NER operation on the text of “TFDs” after the word segmentation. TextRank algorithm can be explained by the following equation.

\[ WS(V_i) = (1 - d) + d \sum_{j \in \text{In}(V_i)} \frac{W_{ij}}{\sum_{j' \in \text{Out}(V_i)} W_{ij'}} WS(V_j) \]

\( W_{ij} \) is the weight of the edge of the nodes \( V_i \) to \( V_j \) in the figure. \( d \) is still the damping coefficient, which represents the probability of pointing from a node to any other node in the graph, which is generally 0.85. In \( \text{In}(V_i) \) and \( \text{Out}(V_i) \) are also similar to PageRank, which are the set of nodes pointing to node \( V_i \) and the set of nodes pointed to by the edge starting from node \( V_i \).

Text similarity analysis
Calculating the similarity between two pieces of text is an important application of NLP, such as in information retrieval and recommendation systems. In this paper, we use this method to calculate the similarity between the texts of “TFDs” and “SGC, ” and compare the similarity of “USGC” versus “SGC” and versus “TFDs, ” “USGC” is an important reference book for the “SGC” written by You Yi in Qing Dynasty.[15] In theory, it should be more similar to the text of “SGC.” First of all, loading the three texts of “The TFDs,” “SGC ” and “USGC ” into Python, implementing word segmentation, removing stop words, setting up TF-IDF model and calculating the similarity between “TFDs” and “SGC.” Then we use two different models, TF-IDF and LSI,[16] to calculate the similarity respectively between the “USGC” and “The TFDs” and “SGC,” comparing the differences of the two results.

Results
Analysis of word frequency statistics
The results of word segmentation show that Jieba can completely and accurately identify the name of TCM herbs in “TFDs.” Herbs with frequency >10 are shown in Figure 1. The statistical results of herbs found that Gan Cao has the highest frequency among all the herbs, then followed by...
Gui Zhi. The top five herbs were precisely the composition of Guizhi decoction. The top 11 herbs with frequency >40 are almost warm drugs in TCM properties. Cold drugs such as Huang Qin, Zhi Zi, and rhubarb just begin to appear after that. We can draw a conclusion that warm drugs are relatively more important than cold drugs in “TFDs.” Fu Zi, which is a hot drug, is ranked fifth. The frequency of Fu Zi is almost as same as jujube and peony. Guizhi decoction, which is a warm decoction, is the most important prescription. It is enough to come to a conclusion that “warm therapy” in TCM is much more significant than other treatment methods in “TFDs.”

Python gives 56 syndrome-like words with a frequency >10, partly as shown in Figure 2. “Cold damage” has with the highest, which is up to 99 times. The term “cold damage” mainly refers to exogenous affections and miscellaneous diseases as well as their syndromes caused by wind and cold. Another translation of this book is called “On Cold Damage,” which indicates that cold-evil is the main pathogenic factor in the completion date of the book and explain the reason why using large quantities of warm drugs, which also confirms that Guizhi decoction, a warming prescription, is the main prescription in the book.

Because the results of the two-word frequency statistics methods, Python and Word, are nonnormal distribution, we implement paired nonparametric test. The result shows that P < 0.0001, indicating that there is significant difference between the two methods. That is, the result of searching word frequency by Word software is higher than the result of Python afterword segmentation. The reason is that although Jieba is currently the most common Chinese word segmentation toolkit, there is still some inaccuracy when using Jieba to segment ancient Chinese text such as “TFDs.”

Figure 3 shows that some discrete values are relatively large. Most of the large value points represent the syndromes described by a single Chinese word, which can bring about ambiguity and inaccuracy of segmentation. For example, Jieba recognizes Chinese Words “Fa Han” “Han Chu,” which may mean “sweating” and “sweats out.” Jieba segments them into different words so that the frequency of searching a single Chinese word “Han” in Word software is higher than the result of Python. Jieba cannot segment the syndrome words in “TFDs” very well because of polysemy. Then, we remove the points whose difference value are >100, and compare the text cosine similarities of these two methods.

Cosine similarity is used to calculate the degree of similarity between the two texts. The greater the similarity, the smaller the difference between individuals. The principle of this algorithm is to measure the difference using cosine of the angle between two individuals in vector space, which is the reliable index of similarity to measure strings.[17] This paper directly calls the sklearn toolkit in Python to calculate the cosine similarity between two sets of data, which are the frequency by Python and frequency by Word software. Although the result of nonparametric test shows that the statistical word frequency by Python and searching word frequency by Word software are different. However, the value of many points in Figure 3 is equal. The result of cosine similarity calculation is 0.85, which indicates that the two methods are very similar. This result is of the practical reference value.

As shown in Figure 4, through the statistical analysis of clauses contained decoctions, clauses of Guizhi decoction are the most, followed by Da–Cheng-Qi decoction, Xiao Chai Hu Tang, Sini decoction, and Mahuang decoction, which shows that these five decoctions are the key prescriptions in “TFDs.” Zhang Zhongjing spends a lot of efforts to illustrate them in his book. The relationship between decoctions and herbs was demonstrated in Figure 5, which showed that Guizhi decoction and Gan Cao were used most frequently. The size of nodes represents the count of each entity. The pink nodes represent the 17 decoctions mentioned above.

In Figure 6, “BMZBZ (Bing Mai Zheng Bing Zhi)” is the Chapter subtitles of the book, which appears 10 times in total. They can be used as a topic marks for dividing each chapter. As shown in the lexical dispersion plot, the content of Taiyang disease accounts for about 60% of the book. It can be found that “Guizhi Decoction” is mainly distributed and intensively occurred in the chapter of Taiyang disease. Guizhi decoction appears 52 times in the book as Figure 7
shown, which is 2 times than other four decoctions. Guizhi decoction almost appears in every chapter of the book. Thus, we can draw a conclusion that Guizhi decoction is the most significant decoction in “TFDs.” Almost one-half of the books is discussed around it. The discussion of Guizhi decoction distributes throughout the book.

As shown in Figure 6, we can also conclude that Da Chengqi decoction mainly appears in the chapter of Yangming disease; a few appear in the chapter of Shaoyin disease. Xiaoachaihu decoction mainly appears in the second part of the chapter of Taiyang diseases and appears only once in the chapter of Shaoyang disease. Sini decoction appears 5 times in the chapter of Jueyin disease and 2 times in the chapter of Shaoyin disease. Mahuang decoction mainly appears in the first and second part of the chapter of Taiyang diseases. However, the location of Xiaoachaihu decoction and Sini decoction in the book is different from we usually think. The efficacy of Xiaoachaihu decoction is eliminating pathogenic factors located in Shaoyang channel which runs between the exterior and interior portions of the body. It is generally believed that Xiaoachaihu decoction should be mainly discussed in the chapter of Shaoyang diseases. On the contrary, it is actually largely appears in the chapter of Taiyang diseases. One reason should be that it is necessary and easy to illustrate the differences of Xiaoachaihu decoction between Guizhi and Mahuang decoction in one chapter. The functions of these three decoctions are easy to confuse. In sum, as mentioned above, we found out five key prescriptions in “TFDs.”

**Comparison between keywords extraction and word frequency statistic**

The top 10 words with the highest frequency are listed in the first column of Table 1; the top 10 keywords extracted by the TF-IDF algorithm are listed in the second column.

It can be seen that the results of the two methods are still very different. TF-IDF algorithm has removed some words which occupy high frequency but are not unique in “TFDs,” such as “illness,” “boil,” and “Zhi.” What’s more, the importance of more representative words such as “Gui Zhi” and “Wen Fu” has increased.

The gensim toolkit is called in Python to calculate texts similarity of the top 50 words respectively given by the two methods. Gensim is a free open source Python toolkit that can

| Word frequency | TF-IDF |
|----------------|--------|
| 三两 three Liang | 三两 three Liang |
| 淫Zi | 一升 one Sheng |
| 二两 two Liang | 二两 two Liang |
| 甘草 Gan Cao | 甘草 Gan Cao |
| 病 illness | 桂枝 Gui Zhi |
| 煮boil | 汤主 Tang Zhu |
| 炙Zhi | 風服 Wen Fu |
| 一升one Sheng | 去皮 Qu Pi |
| 桂枝Gui Zhi | 伤寒 cold damage |
| 伤寒 cold damage | 三服 three times |

TF-IDF: Term frequency-inverse document frequency
automatically extract the semantics of documents, including the function to calculate text similarity. The result of similarity is 0.85583997, which indicates that the similarity between the two texts is higher, but there is a slight difference. According to the analysis above, TF-IDF algorithm is more reasonable when extracting keywords than word frequency statistics, because it removes unnecessary words in “TFDs.”

**Result of named entity recognition**

Part of the results is shown in Figure 8. Jieba toolkit with Textrank algorithm is used to implement NER operation on the text of “TFDs” after word segmentation. The entity that obviously recognized incorrect such as “Chu Zhe” is eliminated manually. It is found that there are 7 apparently incorrect entities among the top 50 entities, that is, the accuracy of recognition is 86%.

**Result of text similarity analysis**

The result of similarity calculated by TF-IDF model between “TFDs” and “SGC” is 0.803, indicating that the two texts are very similar. The results of the similarity of “USGC” versus “The TFDs” and “USGC” versus “SGC,” which are, respectively, calculated by TF-IDF model and LSI model, are shown in Table 2. Ideally, USGC should be more similar to SGC than with TFD, for the reason that USGC is the reference book of SGC; the content of this book is much about SGC.

As demonstrated above, the results calculated by TF-IDF model are approach, which means this model cannot classify the two texts distinctly. The reason is that training corpus is not enough. The results calculated by LSI model show that “USGC” is much more similar to “SGC” than to “TFDs,” which indicates that the effect of LSI model is better. Therefore, in the classification task of TCM literature, the LSI model is more effective than TF-IDF model.

**Discussion**

In this paper, we first use NLP technology to carry out a series of operations to the text of “TFDs” such as corpus loading, word segmentation, removal of stop words, word frequency calculation, keyword extraction, and NER. We can summarize and analyze rules of therapy in “TFDs” through word frequency statistics, which is one of the most important purposes of applying NLP to process and mine TCM literature.

**Financial support and sponsorship**

Nil.

**Conflicts of interest**

There are no conflicts of interest.
References

1. Yi L, Li-Li D. Transmission and exploration of beiji qianjin yaofang on zhang zhong-jing’s academic idea. J Tianjin Univ Tradit Chin Med 2014;33:196-8.

2. Yanfing L. The charm of treatise on febrile diseases. Natl Physician Forum 1996;04:13.

3. Fangfang W, Jiaxu C, Ming S, Yajing H, Qixia P. Development and prospect on formula-based syndrome differentiation. J Beijing Univ Tradit Chin Med 2017;40:103-6.

4. Lingxiu C, Lin Z. Study on correlation of prescriptions with syndromes. China J Tradit Chin Med Pharm 2016;31:3166-9.

5. Cambria E, White B. Jumping NLP curves: A review of natural language processing research [review article]. IEEE Comput Intell Mag 2014;9:48-57.

6. Reyes-Ortiz JA, González-Beltrán BA, Gallardo-López L. Clinical Decision Support Systems: A Survey of NLP-Based Approaches from Unstructured Data, 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), Valencia; 2015. p. 163-7.

7. Alemzadeh H, Devarakonda M. An NLP-based Cognitive System for Disease Status Identification in Electronic Health Records. 2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Orlando, FL; 2017. p. 89-92.

8. Souili A, Cavallucci D, Rousselet F. Natural language processing (NLP) a solution for knowledge extraction from patent unstructured data. Procedia Eng 2015;131:635-43.

9. Qi Y, Rubin Z, Haiyang Z, Fengzhi C. Research overview of treatise on febrile disease edition. J Changchun Univ Chin Med 2015;31:635-7.

10. Qingyu L. Differentiation of doubts in zhang zhong-jing’s author’s prefacetreasise on febrile diseases. KnwolAncientmed Lit 2000;03:235.

11. Lan L. Research on Chinses Word Segmentation Method Based on Word Embedding[D]. Harbin Engineering University, 2017.

12. Peng KH, Liou LH, Chang CS, Lin CC. IEEE 2015 24th Wireless and Optical Communication Conference (WOCC) Taipei, Taiwan (2015.10.23-2015.10.24)) 2015 24th Wireless and Optical Communication Conference (WOCC) Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts; 2015. p. 9-14.

13. Xia T, Chai Y. An Improvement to TF-IDF: Term Distribution Based Term Weight Algorithm(C)\Second International Conference on Networks Security. IEEE Computer Society, 2010.

14. Hongyu M, Qingyu X, Hong C, Qinggang M. Automatic identification of TCM terminology in shanghan lun based on conditional random field. J Beijing Univ Tradit Chin Med 2015;38:587-90.

15. Rui Z, Xuemei L, Xiaoping F. On the academic thought of jinkuiyaoexindian. Chin Arch Tradit Chin Med 2010;28:2174-5.

16. Zhang W, Yoshida T, Tang X. A comparative study of TF IDF, LSI and multi-words for text classification. Expert Syst Appl 2011;38:2758-65.

17. Tata S, Patel JM. Estimating the selectivity of tf-idf, based cosine similarity predicates. ACM SIGMOD Rec 2007;36:75-80.

18. Yongyi W, Zhimei W. Construction of Chinese medical literature corpus and its crital issues in the top-down design. West J Tradit Chin Med 2018;31:62-5.

19. Bin C, Feng W, Shiyu L. Design of intelligent chat robot system for Chinese medicine. Comput Knowl Technol 2019;15:174-5, 185.

20. Jia H, Jian-Qiang D, Bin N, Wang-Ping X, Ji-Gen L. Research on the application of intelligent of question-answering system in medical field. Med Inform 2018;31:16-9.

21. Yong X, Shuanggui T, Shaowu S. Thought on building and development of traditional Chinese medicine informatization in China. J Med Inform 2019;07:127.

22. Tengyuan C, En Z, Zhen S. Corpus-based international standardization feasibility study of TCM terminology. Glob Tradit Chin Med 2018;11:538-42.

23. Xiaofang W, Cheng L. A corpus-based study of english translation of TCM terminology. Sci Technol Vision 2019;09:2156.