Fusheng Wang · Gang Luo
Chunhua Weng · Arijit Khan
Prasenjit Mitra · Cong Yu (Eds.)

LNCS 9579

# Biomedical Data Management and Graph Online Querying

VLDB 2015 Workshops, Big-O(Q) and DMAH
Waikoloa, HI, USA, August 31 – September 4, 2015
Revised Selected Papers

Springer

# The Study of the Compatibility Rules of Traditional Chinese Medicine Based on Apriori and HMETIS Hypergraph Partitioning Algorithm

Miao Wang[1], Jiayun Li[2], Li Chen[2], Yanjun Huang[2], Qiang Zhou[1], Lijuan Che[1], and Huiliang Shang[2(✉)]

[1] Shanghai University of T.C.M., No. 1200, Cailun Road, Shanghai 201210, China
miawang2004@hotmail.com, john-zh@163.com, chelijuanl52003@yahoo.com.cn

[2] Department of Electronic Engineering, Fudan University, No. 220, Handan Road, 200433 Shanghai, China
{jiayunlill, li_chenl2, yjhuangl2, shanghl}@fudan.edu.cn

**Abstract.** One of the major research contents carried by scholars of Traditional Chinese medical science (TCM) is to discover the compatibility rules of herbs to increase the efficacy in treating certain syndromes. However, up to now, most of the compatibility rules of herbs are based on empirical analyses, which make them hard to study. Since concepts of Big Data and machine learning have been popularized gradually, how to use data mining techniques to effectively figure out core herbs and compatibility rules becomes the main research aspect of TCM informatics. In this paper, the hypergraph partitioning algorithm HMETIS based on Apriori is applied to exploit and analyze clinical data about lung cancer. The result shows that all 15 Chinese herbs obtained by the algorithm accord with the core concepts of the treatment of lung cancer by experienced TCM doctors, namely replenishing nutrition, clearing heat-toxin, resolving phlegm and eliminating pathogenic factors.

**Keywords:** Data mining · Compatibility rules of herbs · Apriori algorithm · Hypergraph · Community partitioning · HMETIS

## 1 Introduction

### 1.1 The Value of the Study on the Compatibility of Chinese Herbs

The culture of traditional Chinese medical science (TCM) has a long history. In the lengthy course of the clinical medical practice, a large number of scholars of TCM gradually realize the intricacy of disease mechanism. As the therapeutic effect of one kind of Chinese herb to diseases with complicated pathogenic factors is often unsatisfactory, scholars have attempted to use several types of herbs in combination [1]. With the increasing popularity of herb combination in the field of TCM, it has been found out that some groups of combined Chinese herbs may strength treatment effects,


F. Wang et al. (Eds.): Big-O(Q) and DMAH 2015, LNCS 9579, pp. 16–31, 2016.
DOI: 10.1007/978-3-319-41576-5_2

while other groups may have counteractive or even side effects [1]. On the basis of the above facts, scholars through the ages have gathered rich clinical experience, and tracked out a relatively mature compatibility system of Chinese herbs [2, 3].

At the present stage, most compatibility rules, which are supported by partial TCM theories, are concluded by famous veteran doctors of TCM through practice. However, the substantive characteristics of these rules, such as unity, dialectic, objectivity, etc., are too abstract and subjective to be quantified and inherited [4, 5]. Since database, artificial intelligence, and mathematical statics are developing rapidly, data mining has become an important method to interdisciplinary research. Therefore, applying data mining techniques to the compatibility study of Chinese herbs, is a significant research direction which means TCM theories are able to be combined with data techniques. Through data mining, the core herbs of certain syndromes can be figured out from the TCM clinical database, subsequently disclosing the compatibility rules concealed in the data, and promoting the dissemination and popularization of these rules ultimately.

## 1.2 Basic Knowledge About Data Mining

Data mining is a crucial step in the discovery of database knowledge [6], which occurs in an electronic databank, and aims at discovering the hidden patterns in the data by statistics, machine learning, expert systems, etc. [7]. According to the knowledge types obtained by data mining algorithms, the data mining system can be classified into three categories: classification, clustering, and association rule. The analysis in this paper will be primarily based on the association rule, a kind of algorithm regularly used in shopping habits analysis to promote sales of commodities, or pointed advertising placement. The main idea of the association rule algorithm is to capture the relations among different data items from large amounts of data. The next paragraph explains the algorithm in detail.

Let $I$ be a set ($I = \{I_1, I_2, \ldots, I_m\}$). Let $T$ be the set of transactions ($T = \{T_1, T_2, \ldots, T_n\}$) where each transaction $T_i$ is a non-empty subset of the set $I$. An association rule is an expression of the form $A \Rightarrow B$, where $A$ and $B$ are elements or element groups of $I$, and are called the antecedent as well as the consequent respectively. The support of a rule indicates the ratio of the occurrence numbers of $A$ and $B$ in the same transaction to the total transaction number $n$, and is defined as:

$$Support = \frac{|A \cup B|}{|T|} \tag{1}$$

The confidence measures the probability of the presence of $B$ in the same transaction under the condition where $A$ has appeared in a certain transaction. The confidence is defined as:

$$Confidence = \frac{|A \cup B|}{|A|} \tag{2}$$

The association rule algorithm aims to discover the rules satisfying the support and confidence thresholds [8] 327–404. Nevertheless, since a practical database contains an extremely large number of items, it is impossible to search for the required rules by traversal and enumeration [9, 10]. In order to reduce some unnecessary calculation of support and confidence, and to enhance the efficiency of the association rule algorithm, Agrawal and Srikant advanced the Apriori algorithm in 1994 [9, 10]. The algorithm finds out the frequent itemsets meeting the requirements, and then seeks for the association rules fulfilling the confidence threshold in these itemsets. The Apriori algorithm avoids needless calculation and judgment, and greatly elevates the efficiency, making the application of association rules in large-scale database possible.

### 1.3 Related Work on the Compatibility Rules of Chinese Herbs Using Data Mining

Jingsheng Shang, Lisheng Hu et al. [11] did research on the compatibility rules of Banxia Xiexin decoction using data mining. In their paper, they utilized the frequency statistics to analyze the rules, and found the distribution characteristics of the herbs in the decoction.

Liang Ye et al. [12] conducted research on the compatibility rules of Siwu decoction treating dysmenorrhea by the association rule algorithm, studied the involved herbs via the Hypothesis Testing approach, and then concluded the herbs closely related to the prescription.

As for using the complex network to study the compatibility rules, Xuezhong Zhou et al. [13] synthesized and modeled the clinical symptoms and diagnosis results. Complex network can integrate and visualize TCM information, as well as provide a platform for researchers to discover the compatibility rules of Chinese herbs.

Runshun Zhang et al. [14] also applied the complex network to the research on the compatibility rules of herbs treating the disharmony of liver and spleen. After studying the node centricity of various herbs and the side centricity of herb pairs, they gained the core herbs which could cure the disharmony.

Research on the compatibility rules of Chinese herbs based on the Apriori algorithm is mainly conducted by setting a support threshold as well as a confidence threshold, and finding association rules meet these requirements. However, when encountering a large database with huge numbers of herbs and clinical cases, we may obtain quite sophisticated association rules. Due to the fact that the Apriori algorithm only figures out frequent itemsets according to the support threshold, we cannot acquire some frequent itemsets when the threshold is relatively high. By lowering the support threshold, we may not lose these frequent itemsets, but at the same time, many disturbances are introduced.

The visualization of the compatibility of Chinese herbs is usually achieved by using network to express the relations between herbs, and analyzing the compatibility with the centricity of complex network and community detection algorithms. Generally, in a complex network, herbs are presented as vertices and the appearance number of two herbs in the same case represents the weight of edges, thus inevitably causing the loss of information about herbs and prescriptions [14].

### 1.4 Our Research

The paper applies a hypergraph partitioning algorithm based on the Apriori algorithm to the discovery of core herbs treating lung cancer. The experimental data come from 952 effective clinical lung cancer cases of Longhua Hospital Shanghai University of Traditional Chinese Medicine, and we obtain 15 kinds of core herbs, including Chinese Sage Herb, Coastal Glehnia Root, Radix Asparagi, Radix Asteris, etc. After the comparison with literature on TCM, it is shown that the core herbs accord with the main ideas, namely replenishing nutrition, clearing heat-toxin, and resolving phlegm, during the treatment of lung cancer.

The paper is divided into four parts. The first part is the introduction which presents the major research directions as well as the commonly-used algorithms in the domain of data mining, and analyzes the meaning and progress of the research on the compatibility of Chinese herbs. The second part applies the Apriori algorithm to the construction of the hypergraph. On the basis of the hypergraph, the HMETIS algorithm is utilized to partition it for the purpose of getting the compatibility rules of core herbs. The third part is the experimental process, including the introduction of the data set, the presentation of the experimental results, and the analyses of these results. The fourth part makes a conclusion, pointing out the remaining problems in the experiment and indicating the further research direction for enhancement in the future.

## 2 Data Mining on the Compatibility Rules of Core Herbs with Apriori and Hypergraph Partitioning Algorithm

### 2.1 Apriori Algorithm

**Basic Ideas.** The Apriori property: Any subset of a frequent itemset is also frequent [6, p. 333].
k-itemset: A set with $k$ elemental items is called a k-itemset [6, pp. 328–329]

**Principle of the Apriori Algorithm.** In the first part *Introduction*, the paper preliminarily states the association rules and the concepts of support as well as confidence. However, if the association rules meeting the support and confidence thresholds are searched for by traversing the database, the time complexity will be extremely high. Therefore, this method is impractical in a large-scale database. Through calculation, as for a database containing $d$ itemsets, the total number of association rules may be [6, p. 331]:

$$R = 3^d - 2^{d+1} + 1 \tag{3}$$

For example, there will be 3,484,687,250 association rules if an actual database has 20 transactions (itemsets). Generally, a databank of TCM includes hundreds of or thousands of medical cases, so that it is unfulfillable to calculate the support and confidence of all the possible association rules via directly traversing a huge database.

In order to lower the time complexity and avoid the unnecessary calculation of support and confidence, the mining of association rules is divided into two sub-problems [6, pp. 331–391]:

(1) The mining of frequent itemsets: Find out all the frequent itemsets having support greater than the threshold. According to the Apriori property, any subset of a frequent itemset is also a frequent itemset. Therefore, if any subset of an itemset is infrequent, the itemset is absolutely not a frequent itemset. The Apriori algorithm operates the data mining process by several steps: The algorithm firstly obtains all the frequent 1-itemsets. Then, it combines different frequent (k-1)-itemsets to get candidate k-itemsets by iteration. Finally, Apriori finds out frequent k-itemsets due to the support threshold from the candidates.
(2) The mining of association rules: As for each frequent itemset $L$, generate all the subsets of it. As for each subset $S$, calculate the corresponding confidence. If the confidence satisfies the confidence threshold, generate the association rule $S \Rightarrow L - S$, where $L - S$ represents a set made up of elemental items belonging to $L$, but not belonging to $S$.

**Process of the Apriori Algorithm.** The Apriori algorithm captures the frequent itemsets with support greater than the threshold, and then finds out the association rules meeting the confidence requirement. The algorithm separates the mining of the association rules into two steps, averting the calculation of unnecessary support and confidence, reducing the time complexity, and enormously enhancing the algorithm efficiency. Figure 1 is the flow chart of Apriori:

Illustration: freq 1itemset—frequent itemsets with one elemental item;
freqKitemset—frequent itemsets with K elemental items.

## 2.2 Hypergraph and Hypergraph Community Partitioning

**Basic Ideas of Hypergraph.** Suppose $V$ is a set of vertices and $E$ is a set of hyperedges, any hyperedge $e$ included in $E$ is a subset of $V$.

$$\forall e_{e \in E} \subseteq V \tag{4}$$

Therefore, $G = \{V, E, W\}$ represents a hypergraph, and $W$ is the weight of hyperedge. Generally speaking, a hypergraph consists of vertices and hyperedges which connect two or more vertices [15, 16].

A graph made up of vertices and edges can only represent relations between two vertices, but there are many complicated relations among multiple objects in real life, which are difficult to express with simple graphs, such as the cooperation between several authors and the classification of documents. In some cases where complicated relations other than pairwise relations exist, a hypergraph is useful in describing these relations.

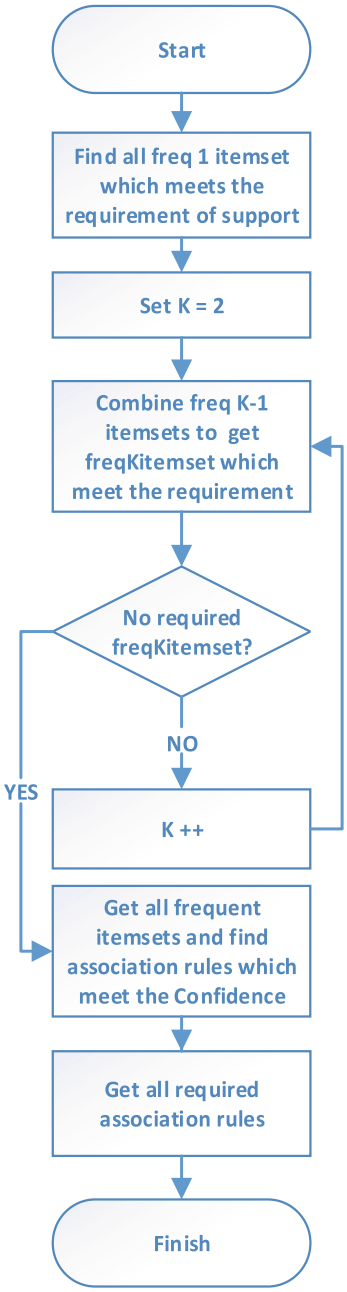
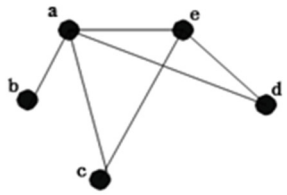



**Fig. 1.** Flow chart of the Apriori algorithm finding association rules

Figure 2 shows the differences and connections between a hypergraph and an ordinary graph.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 1 | 1 | 1 | 1 |
| b | 1 | 0 | 0 | 0 | 0 |
| c | 1 | 0 | 0 | 0 | 1 |
| d | 1 | 0 | 0 | 0 | 1 |
| e | 1 | 0 | 1 | 1 | 0 |

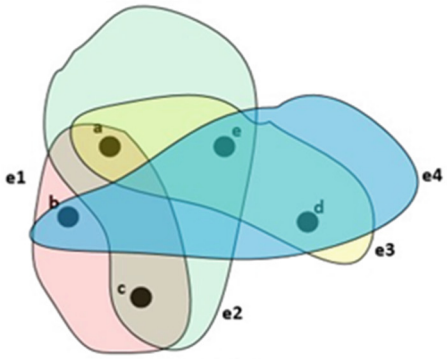
A. Adjacency matrix of graph G




B. Graph G

| | |
|---|---|
| e1 | (a, b, c) |
| e2 | (a, c, e) |
| e3 | (a, d, e) |
| e4 | (b, d, e) |

C. Hyperedge of hypergraph G'




D. Hypergraph G'

**Fig. 2.** Differences and connections between a hypergraph and an ordinary graph

**Basic Ideas of Hypergraph Community Partitioning.** The hypergraph community partitioning is to divide a hypergraph into two or more internally closely connected subgraphs (or communities). In the realm of graph partitioning, Kernighan–Lin is a heuristic partitioning algorithm with time complexity of O(n2 log(n)). The algorithm separates vertices into two sets, so that the weights of the edges connecting two parts are minimized [40]. Fiduccia and Mattheyse proposed a heuristic partitioning algorithm using linear time complexity, improving the efficiency by avoiding unnecessary search and calculation [17]. Vazquez et al. [18] advanced a hypergraph partitioning algorithm based on Bayesian. Samuel R. Bulò et al. [19, 20] put forward a hypergraph clustering algorithm in a game-theoretic approach, converting the hypergraph clustering problem into a non-cooperative game cluster issue.

HMETIS is a hypergraph partitioning algorithm based on a multilevel structure, so that it can produce communities with high quality and have low time complexity. The algorithm also operates fast when partitioning large-scale hypergraphs, so it can be used in a series of practical problems including Very Large-Scale Integration (VLSI). The core of the algorithm is to find hyperedge cuts with least weights, so that a hypergraph is divided into closely connected communities [21, 22]. In this paper, we

apply the HMETIS algorithm to the partitioning of the hypergraph constructed from frequent itemsets so as to find core herbs treating lung cancer.

### 2.3 The Construction of Hypergraph and the Assessment of Hypergraph Community Partition

In this paper, we apply a hypergraph partitioning algorithm HMETIS based on Apriori to the discovery of core herbs in Traditional Chinese Medicine (TCM). Main processes of the algorithm include: inputting data to construct a hypergraph, partitioning the hypergraph, assessing the division results, choosing the closely connected communities, and acquiring core herbs.

**Construction of Hypergraph.** According to the definition of hypergraph, each hyperedge consists of two or more vertices. The frequent itemsets found by the Apriori algorithm meet the support threshold and have certain relevance, so we choose these frequent itemsets to construct hyperedges. However, support alone cannot reflect the associations of two or more elemental items comprehensively. For example, some kinds of supplementary herbs usually appear in TCM prescriptions, so that pairs made up of these herbs and other herbs generally have high support. Nevertheless, it does not mean that the supplementary herbs have strong compatibility with others or are qualified to become core herbs in TCM prescriptions. To avoid the disturbance from quite frequent items, we set the average of confidence of association rules within a frequent itemset as the weight of a hyperedge. For instance, the average confidences of all frequent itemsets acquired by the Apriori algorithm are presented in Table 1.

**Table 1.** Average confidences of frequent itemsets acquired by the Apriori algorithm

| Frequent itemset ID | Frequent itemset | Average confidence |
|---|---|---|
| e1 | (A,B,C) | 0.4 |
| e2 | (A,C,D,E) | 0.6 |
| e3 | (B,C,D) | 0.3 |
| e4 | (A,D,E) | 0.8 |

According to the frequent itemsets and the average confidences, we construct the hypergraph model, shown in Fig. 3. If no association rules in one frequent itemset meet the confidence threshold, then delete the hyperedge.

Next, we use JAVA to read the TCM diagnosis data from Excel, apply Apriori to find frequent itemsets which meet the requirements, present the frequent itemsets in the form of hyperedges, calculate the average of confidence in association rules as the weight of the hyperedges, and finally write the result of hypergraph in a txt file.

**Hypergraph Community Partitioning and the Assessment.** When we invoke HMETIS command lines in JAVA, the main parameter we should set is the community number after the partition. By calculating the inner connectivity closeness index Fitness

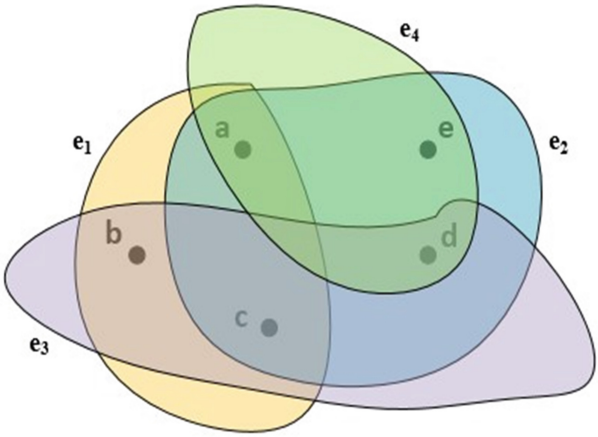



**Fig. 3.** Hypergraph constructed by frequent itemsets

of each partitioned community in the hypergraph, we constantly modify the parameter to achieve better results. The index Fitness is calculated as [21]:

$$fitness = \frac{\sum_{e \subseteq C} Weight(e)}{\sum_{|e \cap C| > 0} Weight(e)} \tag{5}$$

Through the calculation of Fitness in every partitioned community, we can get an average Fitness. By modifying the community number, we can get a partially optimized partitioning solution. After determining the best community number, we use the HMETIS algorithm to partition the hypergraph and output the community (or subgraph) with the largest Fitness. The flow chart of the partitioning algorithm is shown in Fig. 4.

## 3 Data and Results

### 3.1 Data Source

The experimental data come from clinical lung cancer diagnosis records of Longhua Hospital Shanghai University of Traditional Chinese Medicine. The original data have 1000 cases. After the deletion of duplicated and ineffective data, the total number of the available cases is 952. Each case contains a patient ID, symptoms, and herbs. Table 2 is the basic statistics results of the data.

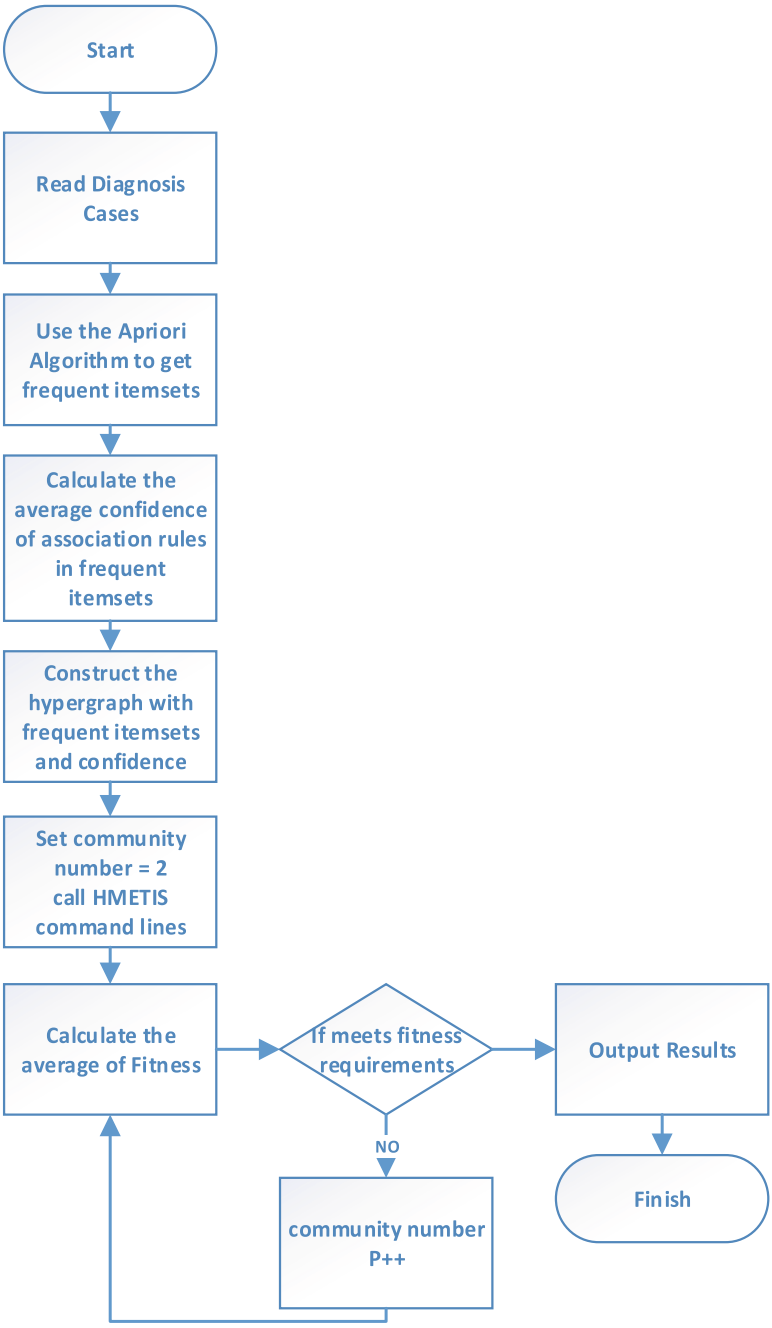



**Fig. 4.** Flow chart of the partitioning algorithm

**Table 2.** Statistics results of the lung cancer data

| Item name | Quantity |
|---|---|
| Diagnosis cases | 952 |
| Total kinds of symptoms | 77 |
| Total kinds of herbs | 348 |

### 3.2 Process of the Experiment

The whole experiment is based on JAVA, which mainly includes: Apriori algorithm class, community partitioning class, file reading and writing class, and other supplementary classes.

**Apriori Class.** This class is used to calculate the support of candidate itemsets to select frequent itemsets, and figure out the required association rules by the confidence threshold. Then, the average confidences of the association rules in the frequent itemsets are computed and serve as the weights of the hyperedges. In this class, main parameters that should be given are the support threshold and confidence threshold. When we set the support threshold, if the threshold is too low, some frequent itemsets with weak associations will be introduced as disturbance. In addition, in that case, the number of frequent itemsets will be largely increased, so that the hypergraph constructed is too dense for the HMETIS partitioning algorithm to achieve satisfactory results. However, if the support threshold is too high, the frequent itemsets are too few so that we may lose association rules we are interested in. After several experiments, we finally set the support threshold as 0.18, and the confidence threshold as 0.9. Table 3 shows the number of hyperedges with the increase of the support threshold.

**Table 3.** The relationship between the number of hyperedges and the support threshold

| Support threshold | Number of hyperedges |
|---|---|
| 0.1 | 3404 |
| 0.12 | 2177 |
| 0.15 | 1114 |
| 0.17 | 725 |
| 0.18 | 593 |
| 0.2 | 457 |
| 0.25 | 243 |

From Table 3, it can be seen that with the decrease of the support threshold, the number of hyperedges and the density of hypergraph are largely increased.

**Community Partitioning Class.** Community partitioning class is used to modify parameters and call the HMETIS command lines. The class partitions the hypergraph and selects the community with the finest connectivity as the output. As for determining parameters, we find it most satisfying when the support threshold is 0.18, the

confidence threshold is 0.9, and the community (subgraph) number is 3. With these parameters, the HMETIS function is used to separate the hypergraph.

### 3.3 Results

After processing the lung cancer diagnosis records, we obtain 593 frequent itemsets (hyperedges), the maximum of which includes 7 kinds of herbs. Use the HMETIS algorithm to partition the hypergraph, and the biggest connectivity closeness index Fitness of the community is 0.470588. A total of 15 kinds of herbs treating lung cancer are found in this community, listed in Table 4.

**Table 4.** Core herbs treating lung cancer acquired by the hypergraph partitioning

| Herb | Herb |
|---|---|
| Chinese Sage Herb | raw oysters |
| Coastal Glehnia Root | Raw Radix Astragali |
| Radix Asparagi | Fruit of Fiverleaf Akebia |
| Radix Asteris | Chickens Gizzard-membrane |
| Coix Seed | Bulb of Thunberg Fritillary |
| Spica Prunellae | Herba Selaginellae Doederleinii |
| Radix Ophiopogonis | Spreading Hedyotis Herb |
| almond | |

### 3.4 Result Assessment

In order to evaluate the core herbs gained from the experiments, we refer to the medical records from Pro. Jiaxiang Liu and other prescriptions on treating lung cancer from famous veteran doctors of TCM.

According to the establishment mechanism of lung cancer, Liu considers the treatment should be focused on replenishing nutrition, clearing heat-toxin, and resolving phlegm.

Herbs that help to nourish stomach can replenish the patients' nutrition, and enhance their immunity. Herbs helping to eliminate pathogenic factors can remove pathogens away from the body, but such effects cannot be achieved when patients are rather weak. Hence, these two kinds of herbs may be combined to attain the ultimate goal [23].

The 15 kinds of herbs obtained by the hypergraph community partitioning algorithm are analyzed in the following paragraphs on the basis of our reference to the medical records from Liu.

**Herba Selaginellae Doederleinii, Chinese Sage Herb, Spreading Hedyotis Herb.** Herba Selaginellae Doederleinii, Chinese Sage Herb, Spreading Hedyotis Herb have the effect of clearing heat-toxin. According to *National Herbs Compilation*, Herba Selaginellae Doederleinii can treat Inflammations and cancers [24]. *Ben Jing* describes

that Chinese Sage Herb also has the effect of removing blood stasis. Besides, experiments show Spreading Hedyotis Herb regulates immunity and resists malignant tumors [25, 26].

Research [25] also found that the combination of Chinese Sage Herb and Herba Selaginellae Doederleinii had a synergy, strengthening the effect of clearing heat-toxin, as both herbs targeted the lung.

**Spica Prunellae, Raw Oysters, Fruit of Fiverleaf Akebia.** Phlegm results from the obstruction in primary and collateral channels due to the invasion of pathogens and toxics [25].

In light of clinical diagnosis cases, Spica Prunellae is effective in preventing tumors from spreading [27]. Raw oysters are mainly used to dissipate phlegm and resolve masses [28]. Fruit of Fiverleaf Akebia nurtures liver, kidney, stomach, and spleen, mainly easing the pain caused by the stagnation of the vital energy circulation.

With the combination of these three herbs, their medical effect of dissipating phlegm is enhanced. In addition, treating heat as well as dampness stasis and nurturing body organs are beneficial to resolving phlegm, so these three herbs are widely applied together in curing phlegm accumulation and some kinds of tumors [25, 28] 62.

**Coastal Glehnia Root, Radix Asparagi, Radix Ophiopogonis.** Liu considers the physical weakness is the cause of lung cancer, so the key points should be the cultivation of immunity and the nurture of body. By replenishing nutrition, the tumors may be dissolved [25].

Coastal Glehnia Root, Radix Asparagi, and Radix Ophiopogonis belong to the herbs with nurturance. Radix Ophiopogonis can cure weakness and nurture the lung [28] 16, while Radix Asparagi can promote the production of body fluid and moisten the lung [28] 9.

The combined use of Radix Ophiopogonis, Radix Asparagi, and Coastal Glehnia Root can enhance the effect of building up health. Radix Asparagi and Radix Ophiopogonis are also usually used together when curing sore throat and cough. The combination of Coastal Glehnia Root and Radix Ophiopogonis is necessary when pulmonary functions are regulated [25, 29].

**Raw Radix Astragali, Chickens Gizzard-membrane, Coix Seed.** Raw Radix Astragali, Chickens Gizzard-membrane, and Coix Seed are usually utilized together to nurture spleen and stomach [28], strengthening the body, replenishing nutrition, and subsequently helping the treatment of lung cancer.

**Radix Asteris, Bulb of Thunberg Fritillary, Almond.** Radix Asteris, Bulb of Thunberg Fritillary, and almond all have the effect of arresting the cough and clearing the lung-heat [30].

However, according to *Ben Cao Jing Shu*, Radix Asteris is mild in medicine property, so it is not suitable to be applied alone. With the company of Radix Asparagi and Radix Ophiopogonis, which are cool in medical property, their properties are counteracted, so the ideal effects of herbs are achieved.

Generally speaking, the core herbs obtained through the algorithm basically accord with the conclusions on treating lung cancer of Pro. Jiaxiang Liu and other experienced

TCM experts. The purposes of replenishing nutrition, clearing heat-toxin, and resolving phlegm can also be seen in the literature on TCM, suggesting our results have a certain reference value.

# 4 Conclusion and Outlook

## 4.1 Conclusion

In this paper, the hypergraph community partitioning algorithm is used to exploit 952 cases of clinical lung cancer diagnosis records from Longhua Hospital Shanghai University of Traditional Chinese Medicine.

Firstly, frequent itemsets and association rules are obtained using the Apriori algorithm. These frequent itemsets meeting the support and confidence requirements serve as hyperedges; the average confidence of association rules in a frequent itemset serves as the weight of the corresponding hyperedge; those association rules which do not meet the confidence requirement are removed.

After the construction of the hypergraph, the HMETIS algorithm is used to partition it. The degree of internal connectivity closeness of one subgraph is evaluated by calculating the index Fitness, and higher average of the indexes of all the subgraphs is acquired by modifying the parameters in the HMETIS algorithm.

From data mining, we find compatibility rules with 15 kinds of core herbs: Chinese Sage Herb, Coastal Glehnia Root, Radix Asparagi, Radix Asteris, Coix Seed, raw oysters, Raw Radix Astragali, Fruit of Fiverleaf Akebia, Chickens Gizzard-membrane, Bulb of Thunberg Fritillary, Spica Prunellae, Herba Selaginellae Doederleinii, Radix Ophiopogonis, Spreading Hedyotis Herb, and almond.

By comparing the core herbs achieved from the algorithm with diagnosis records and academic materials of TCM, it proves that the herbs accord with the conclusions on treating lung cancer of Pro. Jiaxiang Liu and other experienced TCM experts. The purpose of replenishing nutrition, clearing heat-toxin, and resolving phlegm can also been seen in the literature on TCM, suggesting there is a certain reference value in our results.

## 4.2 Outlook

In our paper, 15 core herbs treating lung cancer are concluded through the hypergraph community partitioning algorithm, which are substantially consistent with the diagnosis results of Pro. Jiaxiang Liu and other well-known TCM experts. However, the support threshold, the partitioned community number, and other parameters in the algorithm are obtained by testing during several experiments, with subjectivity and uncertainty to some extent.

In addition, the paper only considers whether herbs appear in clinical diagnosis records, but do not consider how the dosage affects the efficacy and the compatibility rules of herbs. Besides that, the assessment of the herbs is largely dependent on the existing literature on TCM and the therapeutic experience from physicians, lacking objective and theoretical evaluation criteria.

Based on the above points, the directions of improvement are:

1. Exploring methods of adjusting parameters to avoid subjectivity and instability caused by the artificial setting; increasing training data sets, so that the optimal parameter combination is able to be summarized.
2. Gathering the information on the ratios of herbs; using the Apriori algorithm or other complex network algorithms to discover the relations between the efficacy and the ratios, and to find more accurate compatibility rules between Chinese herbs.

**Acknowledgment.** This work was supported by National Natural Science Foundation of China (Grant No. 61301028); Natural Science Foundation of Shanghai China (Grant No. 13ZR1402900); Doctoral Fund of Ministry of Education of China (Grant No. 20120071120016).

## References

1. Yejun, C.: Data mining Technology and its application in Traditional Chinese Medicine. Zhejiang University (2003)
2. Zhou, X., Liu, Y., et al.: Research on compound drug compatibility of complex network. Chin. J. Inf. Tradit. Chin. Med. **15**(11), 98–100 (2008)
3. Meng, F., Li, M., et al.: Mining the medication law of ancient analgesic formulas based on complex network. J. Tradit. Chin. Med. **54**(2), 145–148 (2013)
4. Zhang, B.: Research on data-mining technology applied traditional Chinese prescription compatibility based on association rules. J. Gansu Lianhe Univ. (Nat. Sci.) **25**(1), 82–86 (2011)
5. Zhou, X., et al.: Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif. Intell. Med. **48**(2), 139–152 (2010)
6. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining, vol. 1. Pearson Addison Wesley, Boston (2006)
7. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
8. Han, J., Kamber, M., Pei, J.: Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, San Francisco (2006)
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of 20th International Conference Very Large Data Bases VLDB, vol. 1215, pp. 487–499 (1994)
10. Tseng, V.S., et al.: Efficient algorithms for mining high utility itemsets from transactional databases. IEEE Trans. Knowl. Data Eng. **25**(8), 1772–1786 (2013)
11. Shang, J., Lisheng, H., et al.: Data mining of the law of compatibility of medicines and application of banxia xiexin decoction. J. China-Jpn. Friendship Hosp. **19**(4), 227–229 (2005)
12. Ye, L., Fan, X., et al.: Association among four-drug decoction and the like for dysmenorrhea at all times. J. Nanjing Univ. Tradit. Chin. Med. (Nat. Sci.) **24**(2), 94–96 (2008)
13. Zhou, X., Liu, B.: Network analysis system for traditional Chinese medicine clinical data. In: 2nd International Conference on Biomedical Engineering and Informatics, BMEI 2009, pp. 1–5. IEEE (2009)

14. Zhang, R., Zhou, X., et al.: Study on compounding rules of Chinese herb prescriptions for treating syndrome of liver and spleen disharmony by scale-free network. World Sci. Technol.-Modernization Tradit. Chin. Med. **12**(6), 882–887 (2010)
15. Yu, J., Tao, D., Wang, M.: Adaptive hypergraph learning and its application in image classification. IEEE Trans. Image Process. **21**(7), 3262–3272 (2012)
16. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. **49**(2), 291–307 (1970)
17. Fiduccia, C.M, Robert M.M.: A linear-time heuristic for improving network partitions. In: 19th Conference on Design Automation. IEEE (1982)
18. Vazquez, A.: Finding hypergraph communities: a bayesian approach and variational solution. J. Stat. Mech. Theor. Exp. (2009)
19. Chakraborty, A., Saptarshi, G.: Clustering hypergraphs for discovery of overlapping communities in folksonomies. In: Mukherjee, A., Choudhury, M., Peruani, F., Ganguly, N., Mitra, B. (eds.) Dynamics on and of Complex Networks, vol. 2, pp. 201–220. Springer, New York (2013)
20. Bulò, S.R., Marcello, P.: A game-theoretic approach to hypergraph clustering. Adv. Neural Inf. Process. Syst. **35**, 1571–1579 (2009)
21. Li, Y.: An entropy-based algorithm for detecting overlapping communities in hyper-networks. Sci. Technol. Eng. **13**(7), 1856–1859 (2013)
22. Han, E.-H., et al.: Clustering in a high-dimensional space using hypergraph models. In: Proceedings of Data Mining and Knowledge Discovery (1997)
23. Church, K.W., Patrick, H.: Word association norms, mutual information, and lexicography. Comput. linguist. **16**(1), 22–29 (1990)
24. Lin, X.: Experience of Professor Xixiang Liu in treating lung cancer. Inf. Tradit. Chin. Med. **12**(4), 36–37 (1995)
25. Liu, J., Pan, M., et al.: Clinical study of Jin Hu Kang oral liquid for treating non-small cell lung cancer. Tumor **21**(6), 463–465 (2001)
26. Ji, W.: Professor LIU Jia-xiang's experience in the treatment of lung cancer with Chinese drug pair. Chin. Arch. Tradit. Chin. Med. **28**(6), 1154–1156 (2010)
27. Shiyun, Z., Jinnan, Z.: Experimental study on treatment of compound hedyotic diffusa in tumor-burdened mice. Pract. Clin. J. Integr. Tradit. Chin. West. Med. **09**, 81–83 (2014)
28. Wang, P., Zhang, S.: Research advances of the anticancer mechanisms of prunella vulgaris. Shandong Sci. **23**(2), 38–41 (2010)
29. Yan, P.: Traditional Chinese Medicine Dispensing Technology. Chemical industry press (2006)
30. The State Administration of Traditional Chinese Medicine《Zhonghua Bencao》.Shanghai: Shanghai science and technology publishing house (1999)