Engineering, 2013, 5, 1-6
http://dx.doi.org/10.4236/eng.2013.510B001 Published Online October 2013 (http://www.scirp.org/journal/eng)

Construction and Application of the Multidimensional Table for Knowledge Discovery in Ancient Chinese Books on Materia Medica

Rui Jin[1,2], Qian Lin[3*], Jun Zhou[1], Boyu Sun[1], Bing Zhang[1*]
[1]School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing, China
[2]Department of Pharmacy, Beijing Shijitan Hospital, Beijing, China
[3]School of Mathematical Science, Peking University, Beijing, China
Email: *lillianlin@pku.edu.cn, *zhangbing6@263.net

Received August 2012

ABSTRACT

Knowledge discovery, as an increasingly adopted information technology in biomedical science, has shown great promise in the field of Traditional Chinese Medicine (TCM). In this paper, we provided a kind of multidimensional table which was well suited for organizing and analyzing the data in ancient Chinese books on Materia Medica. Moreover, we demonstrated its capability of facilitating further mining works in TCM through two illustrative studies of discovering meaningful patterns in the three-dimensional table of *Shennong*'*s Classic of Materia Medica*. This work might provide an appropriate data model for the development of knowledge discovery in TCM.

Keywords: Multidimensional Table; TCM; Herbal Medicine; Data Mining; Knowledge Discovery

1. Introduction

Data mining and knowledge discovery, as incremental adopted information technologies in biomedical science, have shown great promise in the field of Traditional Chinese Medicine (TCM) for years. Based on a different view toward human life and disease, TCM has developed a distinct medical system for diagnosis and treatment during thousands of years, which has accumulated a large number of medical and pharmaceutical data [1]. In the past years, it has been increasingly adopted as an important complementary healing therapy around the world [2] and has attracted researchers among different areas to mine the "knowledge gold" buried in TCM data mountains [3-5]. Thus, data mining techniques are believed to be able to bridge the gap between the availability of large amounts of data and the difficulty of obtaining novel knowledge about TCM, especially the medical theory such as yin-yang and five elements.

Learning rich dialectical thoughts from the ancient Chinese philosophies, TCM views the world and human body as a whole and analyzes their relationship with yin-yang and five elements theory. These theories build a universal foundation for the specific theories related to the diagnosis and treatment, such as syndrome differentiation theory, Zang Fu theory, and Chinese herbal medicine theory [1]. Among them, the Chinese herbal medicine theory (herbal property, compatibility, the multiple effectiveness of herbal medicine, etc.) is believed to be a breakthrough in TCM modernization, which is worthy of further investigation. Thus, Chinese Herbal Medicine Informatics (CHMI) has arisen gradually [6-8] and ancient Chinese books on Materia Medica, the conventional media storing the information of medicinal herbs, are always the preferred materials for study.

Shennong'*s Classic of Materia Medica* (*SCMM*), also known as *Shennong Bencao Jing*, is among the great classics of herbal pharmacology and the earliest extant one. The book collects 365 kinds of Chinese medicines and involves many aspects of medicines such as alias, qi and flavor, efficacy and their origins. More than 170 kinds of diseases are discussed, including diseases of internal medicine, surgery, gynecology, pediatrics, etc. [9]. Since many of the recorded herbs are still used in TCM therapies currently, *SCMM* has received sufficient attention in modern research. However, due to the ancient Chinese vocabularies, expert data cleansing and integration are needed for accessibility to modern researchers.

Moreover, to be more effective and valuable, the credibility of the data source and the contribution to new knowledge acquisition are required in the process of data mining. For knowledge discovery in TCM, three aspects of data quality should be highlighted to improve data

[*]Corresponding author.

credibility including representation granularity, representation consistency and completeness [10]. Another key issue is the transformation from data mining results generated by the computer into novel TCM knowledge. As a solution, the two-cycle model was provided by Wang in 2008 [11] who has attached importance to the collaboration of medical researchers and data mining researchers.

In this paper, we intended to establish a kind of multidimensional table to manage herbal information contained in Materia Medica books, as well as to permit data to be easily accessed and analyzed. Taking *SCMM* for example, we constructed the three-dimensional table that presented the major aspects of herbs including herbal qi, herbal flavor and herbal efficacy. Furthermore, we applied the three-dimensional table of *SCMM* to mining novel knowledge related to Chinese herbal theory. This framework might provide as a helpful tool for information management and understanding in TCM.

The rest of the paper is organized as follows. Section 2 described the process of constructing the multidimensional table for appropriate organization of information contained in *SCMM*. Section 3 presented two application examples involving association rules mining and clustering analysis. Finally we provided the conclusions in Section 4.

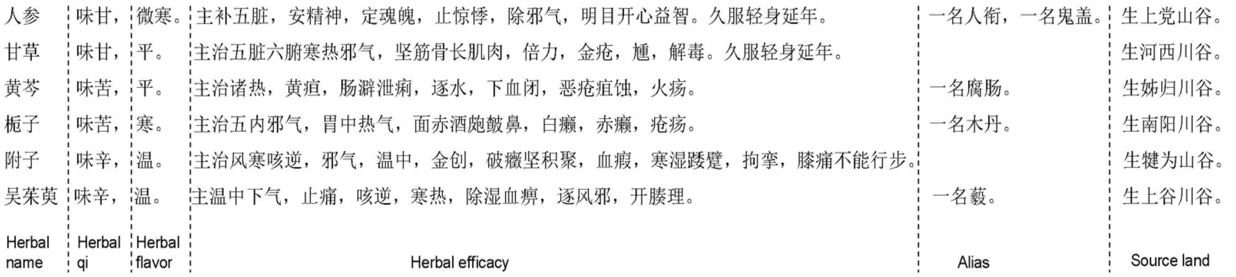
2. The Construction of Multidimensional Table

Ancient Chinese materia medica books are among the most important resources of TCM for data mining, which constitute the foundations of CHMI. As a practical manual of TCM drug therapy, the information about herbal name and botanical origins recorded in the book guarantees the fit medicinal herbs, while the information about herbal property and efficacy reflects the direct experiences of TCM practitioners on clinical drug use [12,13]. Actually, in the view of data management, the text in these books shares common features of semi-structured data, which contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data [14]. For example, *SCMM* is composed of 365 medicine records prepared by classical Chinese words. Each record is written in accordance with the record format which can be divided into six parts including herbal name, herbal qi, herbal flavor, herbal efficacy, alias and source land (**Figure 1**). The first four parts which formed the main body of Chinese herbal theory were selected in this work to construct the multidimensional table.

2.1. Table Structure

A multidimensional table is a multidimensional array consisting of records (rows) and fields (columns), which is suited for organizing and analyzing the data in ancient Chinese books on Materia Medica. In the data table of *SCMM*, each row represented a single herbal medicine. Herbal qi, herbal flavor and herbal efficacy, which are among the most significant parameters to define the clinical performances of medicinal herbs, were employed as fields. In addition, each of the first two fields could be split into five categories due to its structured data model. However, the field of efficacy presented as semi-structured text, would be split into a determinate number of categories after appropriate data integration. Therefore, the resulting table would have three dimensions, since each categorized variable represented one dimension. The ultimate data model was shown in **Table 1**, which also contained a unique identifier (Herb ID) and herbal name. The concrete information of each dimension is as follows:

1) *Herbal qi dimension:* It is the structured data which has five attributes (equivalent to categories in the field in this paper) including cold, cool, neutral, warm and hot. Only one attribute can serve as the marker for each herb in this dimension.



人参	味甘，	微寒。	主补五脏，安精神，定魂魄，止惊悸，除邪气，明目开心益智。久服轻身延年。	一名人衔，一名鬼盖。	生上党山谷。
甘草	味甘，	平。	主治五脏六腑寒热邪气，坚筋骨长肌肉，倍力，金疮，尰，解毒。久服轻身延年。		生河西川谷。
黄芩	味苦，	平。	主治诸热，黄疸，肠澼泄痢，逐水，下血闭，恶疮疽蚀，火疡。	一名腐肠。	生姊归川谷。
栀子	味苦，	寒。	主治五内邪气，胃中热气，面赤酒皰皶鼻，白癞，赤癞，疮疡。	一名木丹。	生南阳川谷。
附子	味辛，	温。	主治风寒咳逆，邪气，温中，金创，破癥坚积聚，血瘕，寒湿踒躄，拘挛，膝痛不能行步。		生犍为山谷。
吴茱萸	味辛，	温。	主温中下气，止痛，咳逆，寒热，除湿血痹，逐风邪，开腠理。	一名藙。	生上谷川谷。
Herbal name	Herbal qi	Herbal flavor	Herbal efficacy	Alias	Source land

Figure 1. Herbal medicine records in Shennong's classic of materia medica.

Table 1. Data model of SCMM.

Column 1	Column 2	Column 3	Column 4	Column 5
Herb ID	Herb name	Herb nature	Herb flavor	Efficacy

2) *Herbal flavor dimension:* It is the structured data which has five attributes including pungent, sweet, sour, bitter and salty. Only one attribute can serve as the marker for each herb in this dimension.

3) *Herbal efficacy dimension:* It is the semi-structured data which can be divided into a finite number of attributes after data integration. Several attributes can serve as the markers for each herb in this dimension.

2.2. Data Preprocess

Since most of the ancient Chinese Materia Medica books are prepared by classical Chinese and provided with different versions, data preprocess (e.g. data cleaning, data integration and annotation) is indispensable for ensuring data quality. In this work, regarding to synonyms of efficacy terms in Classical Chinese, some ancient and contemporary references including *Zhu Bing Yuan Hou Lun* [15], *Internal Medicine of TCM* [16], *Surgery of TCM* [17], *Obstetrics and Gynecology of TCM* [18] and two proofreading and annotation books for *SCMM* [19,20] were employed to achieve representation consistency. Finally, 196 items were acquired for attributes in efficacy dimension. Thus, semi-structured data records presented in **Figure 1** can be converted into a data table shown in **Table 2**.

After the selection of defined attributes in three dimensions separately, a kind of three-dimensional table was constructed in an Excel file format. The row of the table represented the information of a single herbal medicine. The medicine was located in the table using Boolean values whose expression was evaluated to 0 if the medicine did not have the corresponding attribute, 1 if it have (**Table 3**). Taking ginseng for example, the value of the cell identified by the row of ginseng and the column (attribute) of cool was 1 while other values in this dimension were 0 because the herbal qi of ginseng was cool.

3. The Application of Multidimensional Table

Above all, the digitization of information in ancient Chinese materia medica books was achieved appropriately

Table 2. Data table of herbal medicine records in SCMM.

Herb ID	Herb name	Herbal qi	Herbal flavor	Herbal efficacy
1	Radix Ginseng	Cool	Sweet	Tonifying the middle qi, Nourishing essence-spirit, Settling soul and spirit, Tranquilizing, Removing pathogenic qi, Improving vision, Enhancing the wisdow, Promoting longevity
2	Radix Glycyrrhizae	Neutral	Sweet	Removing pathogenic qi in Zang and Fu, Strengthening muscles and bones, Tonifying qi, Curing war wounds, Removing toxicity, Promoting longevity
3	Radix Scutellaria	Neutral	Bitter	Clearing heat, Treating jaundice, Curing diarrea, Removing water retention, Curing amenorrhea, Treating sore and ulcer, Treating unhealed sore
4	Fructus Gardenia	Cold	Bitter	Removing pathogenic qi in Zang, Clearing heat, Treating sore and ulcer, Curing leprosy, Relieving reddened complexion, Treating acne erythematosa
5	Radix Aconiti Carmichaeli	Hot	Pungent	Warming the middle qi, Removing pathogenic qi, Relieving cough with dyspnea, Curing aggregation-accumulation, Curing impediment disease and wilting disease, Curing war wounds
6	Fructus Evodiae	Hot	Pungent	Warming the middle qi, Relieving cough with dyspnea, Curing cold and heat, Treating fixed impediment and blood impediment, dispersing wind pathogen, Relieving pain, Releasing the exterior

Table 3. An example of the three-dimensional table.

Herb ID	Herb name	Herbal qi					Herbal flavor					Herbal efficacy[a]				
		Cold	*Cool*	*Neutral*	*Warm*	*Hot*	*Sour*	*Bitter*	*Sweet*	*Pungent*	*Salty*	*Promoting longevity*	*Removing pathogenic qi*	*Warming the middle qi*	*Clearing heat*	*Curing war wounds*
1	Radix Ginseng	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0
2	Radix Glycyrrhizae	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1
3	Radix Scutellaria	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
4	Fructus Gardenia	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0
5	Radix Aconiti Carmichaeli	0	0	0	0	1	0	0	0	1	0	0	1	1	0	1
6	Fructus Evodiae	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0

[a]Five attributes in herbal efficacy dimension were chosen for display.

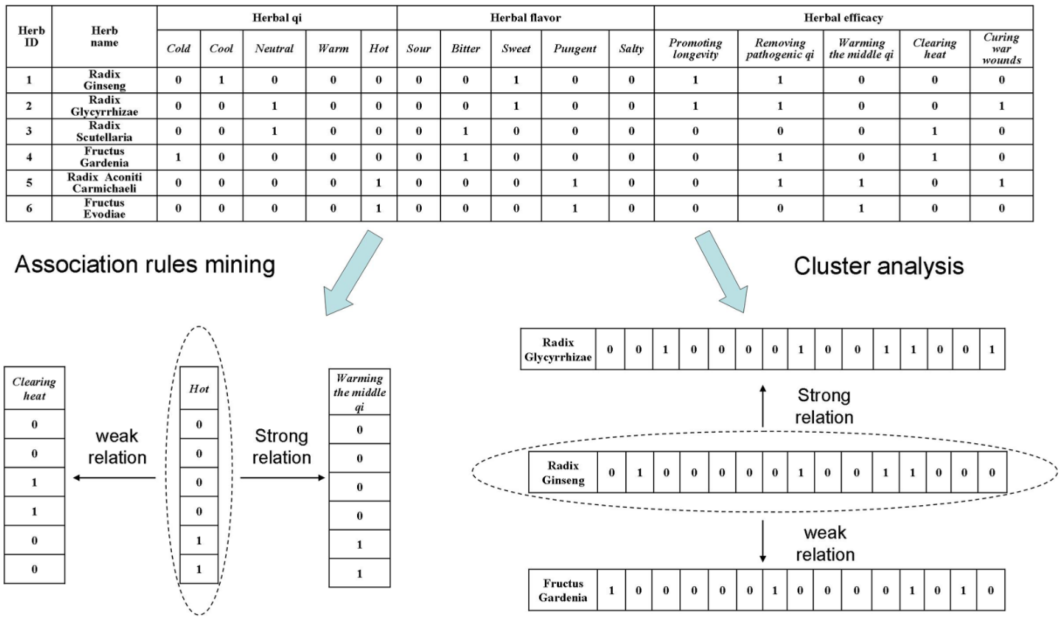
by the multidimensional table, which could facilitate further data mining works. The complete three-dimensional table of *SCMM* consisted of 365 herb records, including 5 attributes in herbal qi dimension, 5 attributes in herbal flavor dimension and 196 attributes in herbal efficacy dimension. Then, two data mining researches, an association rules mining [21] and a cluster analysis [22], were implemented to search for correlations between attributes and between records respectively (**Figure 2**). They would contribute to the acquisition of novel knowledge about Chinese herbal theory.

3.1. Association Rules Mining

In this section, frequent patterns and valued association rules between attributes in the dimension of herbal qi/flavor and herbal efficacy were mined. These kinds of association rules demonstrated the strong relations between herbal property and herbal efficacy, promoting the understanding of Chinese herbal theory. Setting the proper parameters, we acquired 115 strong association rules by the Apriori algorithm (**Table 4**), which presented the evidence to discriminate the qi/flavor of the medicinal herb with specific efficacy. As we can see, some efficacy attributes in **Table 2** were among them such as promoting longevity, clearing heat, warming the middle qi, etc.

3.2. Cluster Analysis

In this section, a classification study was implemented by using semi-supervised incremental clustering algorithm. Calculating the jaccard's index of similarity between every two herb records, we first selected the micro-clusters whose members had exceptionally close correlations. Then a *k*-nearest neighbor algorithm (k = 3) was used to



Herb ID	Herb name	Herbal qi					Herbal flavor					Herbal efficacy				
		Cold	*Cool*	*Neutral*	*Warm*	*Hot*	*Sour*	*Bitter*	*Sweet*	*Pungent*	*Salty*	*Promoting longevity*	*Removing pathogenic qi*	*Warming the middle qi*	*Clearing heat*	*Curing war wounds*
1	Radix Ginseng	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0
2	Radix Glycyrrhizae	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1
3	Radix Scutellaria	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
4	Fructus Gardenia	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0
5	Radix Aconiti Carmichaeli	0	0	0	0	1	0	0	0	1	0	0	1	1	0	1
6	Fructus Evodiae	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0

Clearing heat	*Hot*	*Warming the middle qi*
0	0	0
0	0	0
1	0	0
1	0	0
0	1	1
0	1	1

Radix Glycyrrhizae	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1
Radix Ginseng	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0
Fructus Gardenia	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0

Figure 2. Two data mining studies based on the three-dimensional table of SCMM.

Table 4. Strong association rules.

Form	Strong association rules	
	Total number	*Examples*
Qi ⇒ Efficacy	1	Neutral ⇒ Promoting longevity
Flavor ⇒ Efficacy	3	Sweet ⇒ Tonifying qi
Qi∧Flavor ⇒ Efficacy	18	Sweet∧Neutral ⇒ Tonifying qi
Efficacy ⇒ Qi	38	Warming the middle qi ⇒ Hot; Promoting longevity ⇒ Neutral; Removing toxicity ⇒ Neutral; Clearing heat ⇒ Cold
Efficacy ⇒ Flavor	46	Warming the middle qi ⇒ Pungent; Relieving cough with dyspnea ⇒ Pungent; Nourishing essence-spirit ⇒ Sweet; Removing water retention ⇒ Bitter
Efficacy ⇒ Qi∧Flavor	9	Warming the middle qi ⇒ Pungent∧Hot; Resolving hard mass in stomach and intestine ⇒ Bitter∧Cold

classify the rest of the herbs. The results showed that 253 herbal medicines were reasonably classified as 14 types such as sort of invigoration, clearing heat, diuresis, treating impediment disease and treating gynecological disease, while the other 112 medicines were classified into 112 individual types. The same high similarity to different known types might be the main reason for those individual herbs. **Table 5** showed the major clusters involving more than 10 herbs.

4. Conclusion

Data mining is a promising technology which can be applied in analyzing vast amounts of TCM data for investigating novel knowledge. In this paper, we provided a kind of multidimensional table that was suited for the data in ancient Chinese materia medica books, in order to assist researchers to manage the data in an efficient way. Moreover, we also introduced two illustrative studies of mining meaningful patterns in the three-dimensional table of *SCMM*. The results provided evidence that the multidimensional table could facilitate data mining works in TCM.

Table 5. Representative clusters.

Type	Clusters		
	Total number	***Examples***	***Efficacy***
1	105	Radix ginseng Radix Glycyrrhizae	Invigoration
2	50	Radix Scutellaria Fructus Gardenia	Clearing heat
3	30	Rhizoma Ligustici Wallichi Radix Angelicae Sinensis	Treating gynecological disease
4	15	Rhizoma Podophyllum Scolopendra Subspinipes	Treating strange diease caused by ghost
5	12	Nidus Vespae Calculus Bovis	Treating fright palpitation
6	11	Folium Pyrrosiae Semen Plantaginis	Diuresis
7	10	Radix Aconiti Carmichaeli Fructus Evodiae	Treating impediment disease

5. Acknowledgements

This work is supported by China 973 project (2007-CB512605), the Scientific Research Innovation Team of Beijing University of Chinese Medicine (2011-CXTD-14) and Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment (second author).

REFERENCES

[1] X. Z. Zhou, Y. H. Peng and B. Y. Liu, “Text Mining for Traditional Chinese Medical Knowledge Discovery: A Survey,” *Journal of Biomedical Informatics*, Vol. 43, 2010, pp. 650-660. http://dx.doi.org/10.1016/j.jbi.2010.01.002

[2] National Center for Complementary and Alternative Medicine, “The Use of Complementary and Alternative Medicine in the United States,” 2008.

[3] Y. Feng, Z. H. Wu, X. Z. Zhou, Z. M. Zhou and W. Y. Fan, “Knowledge Discovery in Traditional Chinese Medicine: State of the Art and Perspectives,” *Artificial Intelligence in Medicine*, Vol. 38, 2006, pp. 219-236. http://dx.doi.org/10.1016/j.artmed.2006.07.005

[4] S. Lukman, Y. L. He and S. C. Hui, “Computational Methods for Traditional Chinese Medicne: A Survey,” *Computer Methods and Programs in Biomedicine*, Vol. 88, 2007, pp. 283-294.

[5] S. Li, B. Zhang, D. Jiang, Y. Y. Wei and N. B. Zhang, “Herb Network Construction and Co-Module Analysis for Uncovering the Combination Rule of Traditional Chinese Herbal Formulae,” *BMC Bioinformatics*, Vol. 11, Suppl 11, No. S6, 2010, pp. 1-12,. http://dx.doi.org/10.1016/j.cmpb.2007.09.008

[6] Y. Y. Cheng, X. H. Fan and H. B. Qu, “Discussion on the Establishment and Development of Chinese Herbal Medicine Informatics,” *Chinese Journal of Information on TCM*, Vol. 10, No. 2, 2003, pp. 84-92. http://dx.doi.org/10.1186/1471-2105-11-S11-S6

[7] Z. P. Ding, J. H. Wang and Y. J. Qiao, “Explanation of Chinese Herbal Medicine Informatics,” *Chinese Journal of Information on TCM*, Vol. 10, No. 4, 2003, pp. 92-94.

[8] R. Fang, “Progress in TCM Informatics,” *Chinese Journal of Information on TCM*, Vol. 16, No. 1, 2009, pp. 2-7.

[9] ChineseCultureOnline.org, “Chinese Medicine Book: Shennong Emperor’s Classic of Materia Medica,” 2013. http://www1.chinaculture.org/library/2008-01/31/content_26874.htm

[10] Y. Feng, Z. H. Wu, H. J. Chen, T. Yu, Y. X. Mao and X. H. Jiang, “Data Quality in Traditional Chinese Medicine,” *Proceedings of IEEE Symposium on BioMedical Engineering and Informatics* (*BMEI* 2008), IEEE Press, 2008, pp. 255-259. http://dx.doi.org/10.1109/BMEI.2008.268

[11] H. Wang and S. H. Wang, “Medical Knowledge Acquisition through Data Mining,” *Proceedings of IEEE Symposium on IT in Medicine and Education* (*ITME* 2008), IEEE Press, 2008, pp. 777-780.

[12] B. Zhang, Z. J. Lin, H. Q. Zhai and J. M. Huang, “Research of Chinese Medicine Property Theory Based on the ‘Three-Element’ Hypothesis,” *China Journal of Chinese Materia Medica*, Vol. 33, No. 2, 2008, pp. 221-223.

[13] R. Jin, B. Zhang, X. Q. Liu, S. M. Liu, X. Liu, L. Z. Li, Q. Zhang and C. M. Xue, “Study of Biological Performance of Chinese Materia Medica with Either a Cold or Hot Property Based on the Three-Element Mathematical Analysis Model,” *Chinese Journal of Integrative Medicine*, Vol. 9, No. 7, 2011, pp. 715-724. http://dx.doi.org/10.3736/jcim20110704

[14] Wikipedia (Semi-Structured Data), 2013. http://en.wikipedia.org/wiki/Semi-structured_data

[15] Y. F. Chao, “Treatise on the Pathogenesis and Manifesta-

tions of Diseases. One Hundred Classics of Traditional Chinese Medicine. Edited by China Association of Chinese Medicine," Huaxia Publishing House, Beijing, 2008.

[16] Y. Y. Wang and Z. L. Lu, "Internal Medicine of Traditional Chinese Medicine," People's Medical Publishing House, Beijing, 1999.

[17] X. H. Tan and D. M. Lu, "Surgery of Traditional Chinese Medicine," People's Medical Publishing House, Beijing, 1999.

[18] M. R. Liu and W. X. Tan, "Obstetrics and Gynecology of Traditional Chinese Medicin," People's Medical Publishing House, Beijing, 2001.

[19] G. G. Gu, "Shennong's Classic of Materia Medica, Annotated by P. J. Yang," Academy Press, Beijing, 2007.

[20] Z. J. Shang, "Annotations of Shennong's Classic of Materia Medica," Academy Press, Beijing, 2008.

[21] R. Jin, Q. Lin, B. Zhang, X. Lin, S. M. Liu, Q. Zhao and X. L. Liu, "A Study of Association Rules in Three-Dimensional Property-Taste-Effect Data of Chinese Herbal Medicines Based on Apriori Algorithm," *Chinese Journal of Integrative Medicine*, Vol. 9, No. 7, 2011, pp.794-803. http://dx.doi.org/10.3736/jcim20110715

[22] R. Jin, B. Zhang, C. M. Xue, S. M. Liu, Q. Zhao and K. Li, "Classification of 365 Chinese Medicines in Shennong's Materia Medica Classic Based on a Semi-Supervised Incremental Clustering Method," *Chinese Journal of Integrative Medicine*, Vol. 9, No. 6, 2011, pp. 665-674. http://dx.doi.org/10.3736/jcim20110614