METHODOLOGICAL REVIEW

Knowledge discovery in traditional Chinese medicine: State of the art and perspectives

Yi Feng a,*, Zhaohui Wu a, Xuezhong Zhou b, Zhongmei Zhou a, Weiyu Fan b

a AdvancEd Computing aNd sysTeM (CCNT) Lab, College of Computer Science, Zhejiang University, Hangzhou 310027, PR China
b China Academy of Traditional Chinese Medicine, Beijing 100700, PR China

Received 25 January 2006; received in revised form 4 July 2006; accepted 7 July 2006

ARTIFICIAL INTELLIGENCE IN MEDICINE
http://www.intl.elsevierhealth.com/journals/aiim

KEYWORDS
Traditional Chinese medicine; Knowledge discovery; Data mining

Summary

Objective: As a complementary medical system to Western medicine, traditional Chinese medicine (TCM) provides a unique theoretical and practical approach to the treatment of diseases over thousands of years. Confronted with the increasing popularity of TCM and the huge volume of TCM data, historically accumulated and recently obtained, there is an urgent need to explore these resources effectively by the techniques of knowledge discovery in database (KDD). This paper aims at providing an overview of recent KDD studies in TCM field.

Methods: A literature search was conducted in both English and Chinese publications, and major studies of knowledge discovery in TCM (KDTCM) reported in these materials were identified. Based on an introduction to the state of the art of TCM data resources, a review of four subfields of KDTCM research was presented, including KDD for the research of Chinese medical formula, KDD for the research of Chinese herbal medicine, KDD for TCM syndrome research, and KDD for TCM clinical diagnosis. Furthermore, the current state and main problems in each subfield were summarized based on a discussion of existing studies, and future directions for each subfield were also proposed accordingly.

Results: A series of KDD methods are used in existing KDTCM researches, ranging from conventional frequent itemset mining to state of art latent structure model. Considerable interesting discoveries are obtained by these methods, such as novel TCM paired drugs discovered by frequent itemset analysis, functional community of related genes discovered under syndrome perspective by text mining, the high proportion of toxic plants in the botanical family Ranunculaceae disclosed by statistical analysis, the association between M-cholinoreceptor blocking drug and Solanaceae revealed by association rule mining, etc. It is particularly inspiring to see some studies connecting TCM with biomedicine, which provide a novel top-down view for func-

* Corresponding author. Tel.: +86 571 87951647; fax: +86 571 87953079.
E-mail address: fengyi@zju.edu.cn (Y. Feng).

0933-3657/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2006.07.005
1. Introduction

As a complete medical knowledge system other than orthodox medicine, traditional Chinese medicine (TCM) plays an indispensable role in the health care for Chinese people for several thousand years. The holistic and systematic ideas of TCM are essentially different from the thinking modes based on Reductionism in Western medicine. With the development of modern science, people come to realize the limitations of Reductionism, and begin to lay more emphasis on the systematic thinking patterns, such as Systems Biology [1]. Based on the methodology of holism, TCM plays a unique role in advancing the development of life science and medicine. Meanwhile, with the dramatic increase in the prevalence of chronic conditions, the chemical medicines cannot totally satisfy the needs of health maintenance, disease prevention, and treatment. Human health demands the large-scale development and application of natural medicines, to which TCM experiences and knowledge can contribute a lot. The ever-increasing use of Chinese herbal medicine and acupuncture worldwide is a good indication of the public interest in TCM [2–6].

Countless TCM practices and theoretical researches in thousands of years accumulated a great deal of knowledge in the form of ancient books and literatures. In China, the domestic collection of the ancient books about TCM published before Xinhai Revolution (1911) reaches 130,000 volumes. Besides, thousands of studies on TCM treatments are published yearly in journals all around the world. There were more than 600,000 journal articles during the period of 1984–2005. With such a vast volume of TCM data, there is an urgent need to use these precious resources effectively and sufficiently. Besides, the last decade has been marked by unprecedented growth in both the production of biomedical data and the amount of published literature discussing it. Thus, it is an opportunity, but also a pressing need to connect TCM with modern life science.

Knowledge discovery in databases (KDD) is one proper methodology to analyze and understand such huge amounts of data. As an interdisciplinary area between artificial intelligence, database, statistics, and machine learning, the idea of KDD came into being in the late 1980s. The most prominent definition of KDD was proposed by Fayyad et al. [7] in

Conclusions: Existing studies demonstrate that KDTCM is effective in obtaining medical discoveries. However, much more work needs to be done in order to discover real diamonds from TCM domain. The usage and development of KDTCM in the future will substantially contribute to the TCM community, as well as modern life science.

© 2006 Elsevier B.V. All rights reserved.
In that paper, KDD was defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." This definition may also be applied to "data mining" (DM). Indeed, in the recent literature of DM and KDD, the terms are often used interchangeably or without distinction. However, according to classical KDD methodologies [7], data mining is the knowledge extraction step in KDD process, which also involves the selection and preprocessing of appropriate data from various sources, and proper interpretation of the mining results. Typical data mining methods include concept description, association rule mining, classification and prediction, clustering analysis, time-series analysis, text mining, etc. [8]. During the last two decades, the field of KDD has attracted considerable interest in numerous disciplines, ranging from telecommunications, banking and marketing to scientific analysis. It is also the case within medical environments. The discipline of medicine deals with complex organisms, processes and relations, and KDD methodology is particularly suitable to handle such complexity [9]. Besides, the advent of computer-based patient records (CPRs) and data warehouses contribute greatly to the availability of medical data and offer voluminous data resources for KDD. Also, the need to increase medical knowledge of human beings pushes researchers to carry out knowledge discovery, not only in CPRs and clinical warehouses, but also in biomedical literature databases. The creation of new medical knowledge with DM techniques is listed as one of the 10 grand challenges to medicine by Altman [10]. As Roddick et al. [11] indicates, the application of KDD to medical datasets is a rewarding and highly challenging area. Due to the ever-increasing accumulation of biomedicine data and the pressing demand to explore these resources, the methods of knowledge discovery are widely applied to analyze medical information during the decades. Reviews of KDD in the medical area from different perspectives can be found in refs. [9,11–16]. However, the topic of knowledge discovery in TCM (KDTCM) is not covered in these reviews.

Considering the fast-growing number of researches carried out on KDTCM, it is also necessary and helpful to provide an overview of recent KDTCM researches. As a complementary medical system, TCM is quite different from Western medicine, both in practice and in theory. In view of the high domain-specificity of KDD technology, it is more necessary to gain an insight into KDTCM. Motivated by these needs, this review paper focuses on the introduction and summarization of existing work about KDTCM. Because a great amount of KDTCM work is reported only in Chinese literature, the literature search is conducted in both English and Chinese publications, and the major KDTCM studies published in these materials are covered in this review. For each work, the KDD methods used in the study are introduced, as well as corresponding results. Particularly, some studies with interesting results are highlighted, such as novel TCM paired drugs discovered by frequent itemset analysis, the laboratory-confirmed relationship between CRF gene and kidney YangXu syndrome discovered by text mining, the high proportion of toxic plants in the botanical family Ranunculaceae discovered by statistical analysis, the association between M-cholinceptor blocking drug and Solanaceae discovered by association rule mining, etc. The existing work in KDTCM demonstrates that the usage of KDD in TCM is both feasible and promising. Meanwhile, it should be noticed that the TCM field is still nearly a piece of virgin soil with copious amounts of hidden gold as far as KDD methodology is concerned. To ease gold mining in this field, the future directions of KDTCM research are also provided in this article based on a discussion of existing work.

The rest of this paper is arranged as follows. The prerequisite of applying KDD is the digitalization of the vast amount of data. Thus, an overview of currently available TCM data resources is firstly presented in Section 2. Subsequently, the review of KDTCM work is presented in four research subfields in Section 3, including KDD for the research of Chinese medical formula (CMF), KDD for the research of Chinese herbal medicine (CHM), KDD for TCM syndrome research, and KDD for TCM clinical diagnosis. Based on a discussion of these KDTCM studies, the current state and main problems of KDTCM work in each subfield are summarized in Section 4, and the future directions for each subfield are also presented. Finally, we conclude in Section 5.

2. State of the art of TCM data resources

Data availability is the first consideration before any knowledge discovery task could be undertaken. In this section, we introduce the current state of TCM data resources, especially those data resources focusing on TCM particularly.

As a significant part of complementary and alternative medicines (CAM), literature reporting TCM issues can be found in main CAM databases, such as CAM on PubMed (Complementary and Alternative Medicine subset of PubMed), AMED (Allied and Complementary Medicine Database), CISCOM (Centralised Information Service for Complementary Medicine), CAMPAIN (Complementary and Alternative...
A more comprehensive list of TCM databases can be found in ref. [17]. Currently, the primary data resources specific to TCM includes China TCM Patent Database (CTCMPD) [18], TradiMed Database [19], TCM chemical database [20], and TCM-Online Database System [21]. CTCMPD has been established by Patent Data Research & Development Center, a subsidiary to the Intellectual Property Publishing House of State Intellectual Property Office (SIPO) of China. More than 19,000 patent records and over 40,000 TCM formulas published from 1985 to present are contained in CTCMPD [18]. TradiMed Database was built by the Natural Product Research Institute at Seoul National University, Korea. Based on various Chinese and Korean medical classics, TradiMed represents a combination of traditional medicine knowledge and modern medicine. So far, TradiMed contains information of 3199 herbs, 11,810 formulae, 20,012 chemical compositions of herbs, and 4080 diseases [19]. TCM chemical database was developed by National Key Laboratory of Biochemical Engineering at The Institute of Process Engineering, Chinese Academy of Sciences. This database contains detailed information of 9000 chemicals isolated from nearly 4000 natural sources used in TCM and provides in-depth bioactivity data for many of the compounds [20].

In this section, we place our emphasis on the TCM-Online Database System. To the best of our knowledge, currently TCM-Online Database System is the largest TCM database collections in the world. The prototype of TCM-Online was firstly built in the later 1990s. In 1998, the advanced Computing and System (CCNT) Lab in the College of Computer Science in Zhejiang University and China Academy of Traditional Chinese Medicine (CATCM) began to collaborate in building the scientific databases for TCM, and established a unified web accessible multi-database query system TCMMDB [22] that integrates 17 branches in the whole country. Through the input from nearly 300 scientists from more than 30 colleges, universities and academies of TCM, this system has already integrated more than 50 databases, including Traditional Chinese Medical Literature Analysis and Retrieval System, Traditional Chinese Drug Database, Database of Chinese Medical Formula, etc. TCMMDB was replaced by the grid-based system TCMMDB [22] in 2002, which provides more powerful functions, such as dynamic registration, binding, associated navigation, etc. The TC-Grid system was further extended to a semantic-based database grid named DartGrid in 2002. In present, these databases are available as TCM-Online Database System via website [21] and CD-ROM versions. Besides, a large-scale ontology-based Unified TCM Language System (UTCMLS) [24] has been developed to support concept-based information retrieval and information integration since 2001. All these efforts help to realize the organization, storage, and sharing of TCM data, which provide a feasible environment for the effective implementation of KDD technology.

Today, the TCM-Online Database System integrates more than 50 TCM-related databases. The main databases are listed as below.

### 2.1. Traditional Chinese medical literature analysis and retrieval system (TCMLARS) (Chinese/English)

The bibliographic system TCMLARS [25] has two versions. So far, the Chinese version has contained over 600,000 TCM periodical articles, while the corresponding number reaches 92,000 in English version of TCMLARS. The source material for the database is drawn from about 900 biomedical journals published in China since 1984. The main fields included in TCMLARS are similar with MEDLINE, such as title, author, journal title, publication year, abstract, etc. Besides, some fields specifically existing in TCM are also included, such as pharmacology of Chinese herbs, ingredients and dosage of formula, drug compatibility, acupuncture and Tuina points, etc. TCMLARS is considered as an important new asset in literature review and meta-analysis of Chinese herbal medicine by McCulloch et al. [26]. It also serves as a significant data resource for KDTCM, especially for the methods based on text mining.

### 2.2. Traditional Chinese Drug Database (TCDBASE) (Chinese/English)

This database also has Chinese and English versions. The Chinese version contains over 11,000 records, while the English version contains 545 records. Each record represents a single herb, or mineral drug, or other natural medicine, and provides the cited information. The data is derived from Chinese Materia Medica Dictionary, Thesaurus of Chinese Herbs, Chinese Medicinal Materials, Manual of Composition and Pharmacology of Common Traditional Chinese Medicine, etc. Main contents of TCDBASE include drug name, original source of medical materials, collection and storage of medical materials, parts of the plant/animal for medicinal use, chemical composition, physical/chemical properties, processing methods, dosage form, pharmaceutical techniques, pharmacokinetics, toxicology, compatibility of medicines, efficacy, adverse reactions and their treatment, etc. Such contents of TCDBASE make it an important collection for CHM research, as well as the KDTCM research related to CHM.
2.3. Database of Chinese Medical Formula (DCMF) (Chinese)

This database contains more than 85,000 formulae. Each record represents a single prescription and provides the cited information. Data is derived from modern publications such as *Pharmacopoeia of the People’s Republic of China, Chinese Medical Formula Dictionary*, etc. The main contents of DCMF include formal name, efficacy, indication, usage, precaution, adverse reactions and its treatment, ingredients, modification of the prescription, dosage form and specifications, preparation, storage, compatibility of medicines, chemical composition, toxicology, etc. DCMF collects a comprehensive clinical cases using combinatorial medicines in thousands of years, thus is particularly worth in-depth analysis by KDD technology, especially the approaches based on frequent itemset analysis and association rule mining.

2.4. Database of Chemical Composition from Chinese Herbal Medicine (DCCCHM) (Chinese)

This database contains over 4500 records. Each record represents a single chemical composition and provides the cited information. Data is derived from *Active Compositions of Chinese Herbal Medicine, Pharmacology of Traditional Chinese Medicine, Chinese Herbal Medicine*, etc. Main contents include formal name, chemical name, physical/chemical properties, molecular formula, chemical formula, origin, pharmacological action, efficacy, toxicity, adverse reactions, chemical category, functional category, etc. Although the information of 3D chemical structure is not included in this database, DCCCHM could still be used as an important data resource for CHM research and drug discovery from CHM.

2.5. Clinical Medicine Database (CLINMED) (Chinese)

This database contains information on more than 3500 diseases. The source material for the database was drawn from authoritative reference books and teaching materials of Chinese and western medicine. The main contents of CLINMED include name of diseases, disease classification code by western medicine/TCM, disease name definition by western medicine/TCM, western medicine/TCM etiology, pathology, pathological physiopathology, pathogenesis, diagnosis, western medicine therapy, treatment of TCM, Chinese herbal medicine therapy, acupuncture and moxibustion, massage, integrated therapy of western medicine and TCM, etc. CLINMED reflects the understanding and experience of treating diseases in today’s China, involving both western medicine treatment and TCM therapy. As an effective avenue to combine Western medicine and TCM, multilevel knowledge discovery could be carried out based on the integration of data in both fields, and CLINMED could contribute in this trend.

2.6. TCM Electronic Medical Record Database (TCM-EMRD) (Chinese)

TCM Electronic Medical Record (EMR) contains the data of TCM clinical practice on inpatient and outpatient. Clinical daily practice takes a vital role for TCM research and theory refinement, because unlike modern medicine, almost no bench-side studies are performed in TCM. China Government has initiated several important programs since 2002 to collect the clinical TCM EMR data and the clinical data warehouse TCM-EMRD is built for potential decision support applications and TCM knowledge discovery. Currently, TCM-EMRD contains more than 3500 EMRs of inpatients. The main contents of TCM-EMRD include TCM Diagnosis, Chinese Medical Formula, the conceptual description of symptoms, etc. TCM-EMRD is especially ideal for clinical KDTCM studies because KDD is considered as the major purpose when establishing this database. Thus, special considerations are given to issues related to knowledge discovery, such as the quality of data, the standardization of terms, etc.

Other data resources within TCM-Online Database System include Database of Tibetan Medicines, Medical News Database, OTC Database, Database of State Essential Druggery of China, Database of Medical Research Awards in China, Database of Medical Product in China, TCM Pharmaceutical Industry Database, China Hospitals Database, Database of Paired Drugs (DPD), etc. The TCM-Online Database System, as well as other digitized TCM data resources, serves as the available data sources in various KDTCM researches, which can be seen in the following review section.

3. Review of KDTCM researches

3.1. Knowledge discovery for CMF research

One distinguishing feature of TCM lies in its emphasis on the usage of combinatorial medicines, in the form of CMF. A CMF is composed of selected drugs and suitable doses based on syndrome differentiation for etiology and the composition of therapies in accordance with the principle of formulating a TCM
prescription. Except for very few single drugs of all the prescriptions used clinically, the great majority of them are compound drugs consisting of two or more drugs. The reasons are that the potency of a single drug is usually limited, and some of them may produce certain side effects or even toxicity. But when several drugs are applied together, ensuring a full play of their advantages and inhibiting the disadvantages, they will display their superiority over a single drug in the treatment of diseases. In this way, TCM uses processed multi-component natural products in various combinations and formulations. Due to the great diversity of candidate drugs to form a compound prescription, hundreds of thousands of Chinese medical formulae have been accumulated over thousands of years.

To use combinatorial medicines properly, one key issue is to realize the combination rules of multiple drugs. Compared with western biomedicine, countless TCM practices in thousands of years accumulate numerous cases of combinatorial medicines as the form of formulae. These TCM formulae are valuable resources for research of drug compatibility. Besides, due to the tradition of usage of combinatorial medicines, the analysis of combination rules of multiple drugs exhibits much more significance in TCM than western medicine. Therefore, the CMF research attracts more data miners than other subjects.

A breakthrough of herbal combination rule research lies in the analysis of paired drugs, which usually means a relatively fixed combination of two CHMs. In generalized definition of paired drugs, the number of medicines in paired drugs can be extended to three, four, etc. According to TCM theory, such combinations can increase their medical effectiveness, or reduce the toxicity and side effects of some drugs. A number of paired drugs have already been induced from practice and included in the Database of Paired Drugs as a part of TCM-Online Database System recently. However, such kind of valuable combinations could be revealed more and deeper. When two drugs are frequently used in combination with each other in practice, they are more likely to be paired drugs. Therefore, frequent itemset analysis and association rule mining could be used to discover paired drugs [27–31]. As the most classical example of data mining, association rule mining aims at searching for interesting relationships among items in a given data set. It can serve as one proper KDD method to analyze combination patterns of multiple drugs in TCM. The first step in association rule mining, also the core issue in association rule mining, is to find frequent itemsets, which means each of these itemsets will occur at least as frequently as a pre-deter-

mined percentage in whole dataset. This method can be used to find frequent herb co-occurrences.

In 2002, the classical association rule method was used by Yao et al. [27] to study 106 formulae for treating diabetes. The results indicate that different experts have similar ideas and principles for treating this disease, which helps to reveal the scientific rule in the composition of the TCM formulae for diabetes. This work also demonstrates that KDD is a powerful tool for acquiring knowledge and expressing TCM knowledge in an understandable way. In 2003, Jiang et al. [28] used basic frequent itemset analysis and association rule mining to analyze 1355 formulae for the treatment of spleen-stomach diseases from the section of Prescriptions, Dictionary of Traditional Chinese Medicine. The results of KDTCM in ref. [28] are found to correspond basically with the rules and characteristics of the compatibility of spleen-stomach formulae in TCM. Another association mining on spleen-stomach formulae was carried out by Li et al. [29] using an effective algorithm based on a novel data structure named indexed frequent pattern tree. However, the number of formulae used in experiment was not reported in that literature. To provide a powerful assist for TCM experts and KDTCM participants, Li et al. [30] developed a CMF data mining system named TCMiner in 2004. Several efficient algorithms were implemented for frequent itemset/association rule mining in TCMiner.

The largest frequent itemset analysis in TCM field was carried out by He et al. [31] in 2004. In this study, the DCMF database in TCM-Online Database System, which contains 85,989 formulae collected in thousands of years, was used as the data source to discover potential paired drugs. To improve performance, FP-growth algorithm was applied to find frequent herb co-occurrences in DCMF without candidate generation. Besides, an important preprocessing step was carried out, that is, to remove *Radix Glycyrrhizae* from each formula. This is because *Radix Glycyrrhizae* is frequently used in all kinds of formulae (it can decrease or moderate medicinal side-effects or toxicity and regulate actions of all other herbs in one formula). This research is also noteworthy in that all of the generated paired drugs are compared with DPD. The results of discovered top-15 herb co-occurrences with highest frequency are presented in Table 1.

<table>
<thead>
<tr>
<th>Herb</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Radix Glycyrrhizae</em></td>
<td>15,000</td>
</tr>
<tr>
<td><em>Angelica Sinensis</em></td>
<td>10,000</td>
</tr>
<tr>
<td><em>Ginseng</em></td>
<td>8,000</td>
</tr>
<tr>
<td><em>Astragalus</em></td>
<td>7,000</td>
</tr>
<tr>
<td><em>Cordyceps</em></td>
<td>6,000</td>
</tr>
</tbody>
</table>

The frequency and content of these herb co-occurrences are listed in leftmost and rightmost columns, respectively. The middle column indicates whether these discovered herb co-occurrences exist in DPD. Those records with the field “Yes” substantiate the existing paired drugs statistically. More significantly, as is shown in Table 1, many frequent herb
Knowledge discovery in TCM

Table 1: Top-15 discovered herb co-occurrences with highest frequency

<table>
<thead>
<tr>
<th>Frequency of herb co-occurrences</th>
<th>Exist in DPD</th>
<th>Discovered frequent herb co-occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>5127</td>
<td>Yes</td>
<td>Radix Ginseng, Rhizoma Atractylodis Macrocephalae</td>
</tr>
<tr>
<td>4428</td>
<td>Yes</td>
<td>Radix Angelica Sinensis, Radix Ginseng</td>
</tr>
<tr>
<td>4062</td>
<td>Yes</td>
<td>Radix Angelica Sinensis, Rhizoma Chuanxiong</td>
</tr>
<tr>
<td>3523</td>
<td>No</td>
<td>Radix Angelica Sinensis, Rhizoma Atractylodis Macrocephalae</td>
</tr>
<tr>
<td>3049</td>
<td>Yes</td>
<td>Radix Ginseng, Radix Astragali</td>
</tr>
<tr>
<td>2853</td>
<td>No</td>
<td>Radix Ginseng, Poria</td>
</tr>
<tr>
<td>2760</td>
<td>No</td>
<td>Radix Angelica Sinensis, Radix Sapshnikoviae</td>
</tr>
<tr>
<td>2688</td>
<td>Yes</td>
<td>Rhizoma Atractylodis Macrocephalae, Poria</td>
</tr>
<tr>
<td>2678</td>
<td>Yes</td>
<td>Radix Sapshnikoviae, Rhizoma Notopterygii</td>
</tr>
<tr>
<td>2596</td>
<td>Yes</td>
<td>Radix Angelica Sinensis, Radix Astragali</td>
</tr>
<tr>
<td>2522</td>
<td>No</td>
<td>Radix Ginseng, White Poria</td>
</tr>
<tr>
<td>2321</td>
<td>Yes</td>
<td>Rhizoma Atractylodis Macrocephalae, Pericarpium Citri Reticulatae</td>
</tr>
<tr>
<td>2210</td>
<td>Yes</td>
<td>Radix Angelica Sinensis, Radix Paooniae Alba</td>
</tr>
<tr>
<td>2142</td>
<td>No</td>
<td>Radix Ginseng, Radix Sapshnikoviae</td>
</tr>
<tr>
<td>2127</td>
<td>No</td>
<td>Radix Sapshnikoviae, Rhizoma Chuanxiong</td>
</tr>
</tbody>
</table>

co-occurrences that do not exist in DPD could also be revealed, such as Radix Angelicae Sinensis and Rhizoma Atractylodis Macrocephalae. Such combinations are highly likely to be paired drugs, which are worthy of further analysis and verification by TCM experts. Actually, the above process of knowledge discovery essentially simulates what TCM researchers did to generalize novel paired drugs in the past. The only distinction is that past discovery is based on human observations and experiences, while KDTCM participants utilize powerful computers and efficient algorithms instead to accomplish this task, which greatly quickens the disclosure of new paired drugs and promotes the development of compatibility rule research.

With the deeper understanding of frequent itemsets and association rules in TCM background, researchers realize that it is not enough to discover patterns using conventional methods. Although itemsets with very high frequency always indicate significant discoveries in many disciplines, this is not the case in TCM formulae, just as the condition of Radix Glycyrrhizae mentioned above. To overcome this limitation, Zeng et al. [32] introduced a new support (maximum support) into association mining. Based on this new bi-support, as well as a bitmap matrix technique to improve efficiency, a new association mining algorithm named BM_DB_Apriori was proposed in ref. [32]. The experimental results on 1060 formulae show that BM_DB_Apriori is much faster and more accurate than the baseline algorithm Apriori.

Another improvement is the consideration of correlated patterns in TCM. In the basic approaches of frequent itemset/association mining, we aim to find itemset/rule with minimum support and confidence. However, for TCM formulae, the interesting itemsets/rules always have low support but high confidence. Besides, considering a paired drug A and B, if A and B frequently coexist in formulae, but the existence of A cannot increase the likelihood of the existence of B, it shows that B is very likely a general drug used to moderate side-effects or regulate actions of all other herbs in many kinds of formulae (e.g., Radix Glycyrrhizae). As far as drug compatibility research is concerned, this is not what we are interested. To remedy this situation, Zeng et al. [33] proposed a new method to discover bidirectional TCM association rules in 2005. Similar idea was presented and formalized by Zhou et al. [34] in 2006, and the concept of both associated and correlated pattern was presented in that paper. A new interesting measure corr-confidence was proposed for rationally evaluating the correlation relationships. This measure not only has proper bounds for effectively evaluating the correlation degree of patterns, but also is suitable for mining long patterns. The experimental results on 4643 TCM formulae data demonstrate that the mining considering both association and correlation is a valid approach to discovering both associated and correlated patterns in TCM formulae.

One interesting extension of frequent itemset mining in CMFs appeared in 2005 [35]. Considering that most CHMs are natural products, the knowledge of botanical taxonomy could be introduced to analyze CMFs. Based on frequent itemset mining of paired drugs, as well as the mapping relation between CHM and botanical family, Zhou et al. [35] performed experiments to discover family combination rules of CHM in formulae. Over 18,000 CMFs from TCMLARS were collected and five subtypes of these CMFs were chosen according to the efficacy such as promoting blood circulation for removing
blood stasis, invigorating spleen and replenishing qi, etc. The discovery process was carried out for each subtype, and many family combination rules got revealed. For example, *Radix Astragali—Radix Angelicae Sinensis* is a typical CHM pair with the efficacy of invigorating spleen and replenishing qi, and *Umbelliferae—Legaminosae* builds the core of CMFs with this efficacy, because the support of *Umbelliferae—Legaminosae* combination in CMFs with this efficacy is as high as 180%. Compared to CHM combination rules, family combination rules give a higher level of knowledge about the drug combinations, which would be of help to clinical CMF prescription practice and new drug development.

Both frequent itemset and association rule mentioned above can be classified as linear rule model. To explore and present the pattern of combinatorial medicines in different perspectives, researchers began to introduce other models. Considering that undirected graphical models is suitable for discovering causal relationships and associations between variables, Deng et al. [36] proposed a structural learning algorithm named Information-Based Local Optimization (IBLO) to analyze CHMs in 554 formulæ for apoplexy patients. These formulæ were collected from books about historic prescriptions of Chinese medicine. A graph including 40 most important herbs in these formulæ was obtained by IBLO algorithm, from which we can see which CHMs are frequently used together. This method is noteworthy for TCM in that it extends the pattern of combinatorial medicines research from linear rule model to graph model.

The studies mentioned above share a common feature, i.e., they generally analyze CMFs without the consideration of herb dosage. The reason behind it might be that in TCM literature, especially in historical literature, the dosage information of each herb in one formula is quite fuzzy. In some cases, the dosage is a range value (e.g., 20–30 g). In many other cases, the dosage information is even missing. New knowledge discovery approaches are needed to analyze formulæ with such fuzzy dosage information in historical literature.

One way of formulæ research with dosage is to study the dispensing ratio of different CHMs within one formula. Suppose a combination pattern of multiple CHMs is found to have therapeutic effects for a certain condition, the next issue is to determine optimal dispensing ratio of these CHMs. This task can be performed with the help of knowledge discovery techniques. In 2003, Xiang [37] proposed a three-stage voting algorithm for mining optimal herb dispensing ratio in formulæ. This method was applied to obtain optimal *Radix Salviae Miltiorrhizae—Radix Notoginseng* dispensing ratio in the formulæ to treat cardiopathy based on an animal model (dogs). Ten test groups including seven ratio groups (10/6, 10/3, 10/1, 1/1, 10/0, 0/10, and 1/10) and three comparison groups (groups of western medicine, model of compound, and pseudo-surgery) were used in the experiments. The latter was designed to test the cardio-index of all samples. After extracting features about the therapeutic effects from time series samples, the three-stage voting algorithm works as follows: the preliminary vote generates the features of each test group and cardio-index by each sample; the metaphase vote obtains the value of the therapeutic effect by each feature of the given group and index; the final vote mines the optimal dispensing ratio. An optimal dispensing ratio 10/6 was finally obtained in the experiment, showing the effectiveness of this method.

Confronted with the voluminous amount of TCM literature, historically accumulated and recently published, text mining is another group of knowledge discovery methods which can be used in KDTCM. Text mining is defined as the discovery of novel, previously unknown information, by automatically extracting information from different textual resources. For research of CMF, this technique was applied by KDTCM participants [35,38,39] to extract knowledge of herbs and formulæ from TCM literature. Cao et al. [38] developed an ontology-based system for extracting knowledge of CHMs and CMFs from semi-structured text. In this work, ontologies of CHMs and CMFs were developed, consisting of a set of classes and their relations. In addition to offering a terminology for describing classes and their instances, the attributes of a class also play the role of knowledge place-holders for the class and its instances, and they are to be filled in during the knowledge extraction process. To perform knowledge acquisition, an executable knowledge extraction language (EKEL) was proposed to specify knowledge-extracting agents based on the guidelines of the frame-oriented ontologies, and a support machine was implemented to execute EKEL programs. The authors reported that the system successfully extracted knowledge of more than 2710 CHMs and 5900 CMFs. A limitation of this work lies in the requirement of semi-structured text as input. A more general method of text mining, based on the idea of bootstrapping, was utilized by Zhou and co-workers [39] in 2004. Given a small set of seed words and a large set of unlabeled data, this approach can automatically iterate to extract the objective patterns and new seeds from free texts. Based on this bootstrapping method, as well as other components (e.g., TCM literature DB, data mining...
module), the authors developed a text mining system called MeDisco/3T (Medical Discover for Traditional Treatment inTelligence) [35]. In practice, MeDisco/3T was used in the family combination rule experiment [35] mentioned above, and the bootstrapping method was used as an significant step to extract CMF names, CHM components and efficacy descriptions from TCMLARS with high precision.

3.2. Knowledge discovery for CHM research

Apart from CMF research, knowledge discovery techniques could also be used in CHM research. Existing work in this aspect can be further divided into two subfields: KDD for the research of CHM characteristics and KDD for the research of CHM chemical compositions.

3.2.1. KDD for the research of CHM characteristics

As one of the core contents in TCM knowledge system, the completeness of theory of CHM characteristics largely determines the accuracy and scientificity of forming a CMF, and also the effectiveness and safety of this therapy. The therapeutic effects and side effect/toxicity of drugs are often two sides of the same coin. The toxicity of drugs must be thoroughly understood in order to ensure safety, not only for orthodox medicine, but also for complementary and alternative medicine. Chen and Giu [40] applied computer-based statistical analysis to study the relevant factors of toxicity for 3906 kinds of common used CHM. Besides toxicity, seven other aspects of CHM data were collected, including nature and flavor, meridian tropism, processing method, botanical family, chemical compositions, medicinal part, and pharmacological effects. The relations between these factors and CHM toxicity were analyzed by statistical analysis, yielding some meaningful results. Among 1119 CHMs with toxicity information recorded in literature, only 3.3% is found to have high toxicity, providing an evidence for the safety of most CHMs. However, 59.7% of the CHMs with hot nature exhibit toxicity in varying degrees, showing that CHM toxicity is somehow related to the hot nature. It is also found that the botanical family of CHM is closely related to its toxicity. Among 108 kinds of CHMs belonging to Ranunculaceae, the proportion of CHMs with toxicity in varying degrees reaches as high as 42.6%. Moreover, 29.7% of Ranunculaceae CHMs exhibits high toxicity. Other families of CHMs with relatively large proportions of toxic ones include Araceae (14.2%) and Euphorbiaceae (11.3%). As for the factor of chemical compositions, alkaloid is noteworthy in that 71.8% of CHMs with alkaloid ingredients exhibit toxicity. This number is by far largest compared with other groups of chemical compositions in CHMs.

Another statistical analysis about CHM characteristics was performed by Yang [41] in 2005. In the study, 417 CHMs were collected from Chinese Pharmacopeia, from which 101 blood pressure-reducing CHMs were further selected. A comparison was made between these two groups of CHMs (101 group and 417 group) in aspects of four natures, five flavors and meridian tropism. The results show that most blood pressure-reducing CHMs are characterized by pungent nature and bitter flavor, mainly acting on liver and gallbladder meridians. Such knowledge is beneficial to the discovery of effective phytochemical components and the research of pharmacological properties.

The relation between CHM efficacy and other characteristics is a hot topic in CHM research. The usage of knowledge discovery approaches can partly help this research. Yao et al. [42] applied artificial neural network (ANN) and decision tree to classify 54 deficiency-nourishing CHMs into four groups (four subtypes of deficiency-nourishing efficacy) in 2004. Two quantification methods of CHM characteristics were used in the experiments, including two-valued quantification and multi-valued quantification. The classification results of KDD methods were compared with the records in TCM teaching materials as the evaluation. It is found that the classification accuracies of ANN equals decision tree (98.11%) for multi-valued quantification, but ANN (96.13%) outperforms decision tree (87.10%) for two-valued quantification. The results show that different quantification methods of CHM characteristics have certain influence on the prediction of efficacy, and ANN outperforms decision tree in this task. However, the sample size in this experiment seems too small.

Besides statistical analysis and classification, clustering is another type of KDD approaches used in researches of CHM characteristics. Based on characteristics features, Zhou et al. [43] applied clustering to group 28 CHMs for relieving exterior syndrome in 2004. By this clustering method, the authors obtained quite a few identical results to the CHM theory. Although the sample size was also small (28 CHMs), the methodology is still worth recognition. Another larger scale of clustering was carried out by He et al. [44] in 2004. In this KDTCM study, TCDBASE in the TCM-Onkine Database System was used the as data source, which contains information of more than 11,000 CHMs. Different from the former experiment, this was an efficacy-based clustering, and the similarity of values in attribute Efficacy was used to evaluate the closeness between CHMs.
The clustering algorithm applied in the experiment was agglomerative approach, also known as the bottom–up approach. As a hierarchical clustering method, this algorithm starts with each object forming a separate group. It successively merges the objects or groups close to one another into larger and larger clusters, until certain termination conditions. By this clustering, the large number of CHMs in TCDBASE were clustered into different groups based on efficacy. Such kind of categorization is highly important for the research of CHM characteristics and the analysis of effective chemical components in CHM, because similar efficacies of different CHMs often indicate high closeness of characteristics, as well as the chemical ingredients contained. Furthermore, these clustering results can serve as valuable references for selection of substitute drug in formula-forming and design of novel CMFs.

3.2.2. KDD for the research of CHM chemical compositions

The substances that exhibit therapeutic effects are the chemical compositions of CHM. However, because the medicinal powers of CHM come from the cooperative effects of multiple active ingredients, the conventional research methodology that targets at only single ingredient is not enough. One probing method for this problem is to analyze CHM by knowledge discovery techniques at the level of chemical element. In 1998, Qi et al. [45] applied factor analysis and clustering to study the relationships between CHM characteristics and the contents of trace elements in CHM. Forty-two kinds of trace elements in 105 CHMs were determined and factor analysis was carried out. The results indicate that a 10-factor model can interpret the correlation of these trace elements. Besides, the information of trace elements was used in a hierarchy clustering to classify 105 CHMs into different nature groups. Compared with traditional knowledge of CHM nature, the trace-element-based clustering achieved the accuracy of 78.1%, showing that trace elements are related to the nature of CHMs to a certain degree. A following study of trace elements was conducted by Qi et al. [46] in 2003. This time the information of trace elements was also used in a hierarchy clustering, but aiming at classifying 10 CHMs for treating exterior syndromes into different efficacy groups. The accuracy of classification in the experiment is 90%, partly revealing the correlation between efficacy and the amount of trace elements in CHMs.

The modern technique of fingerprint enables us to takes thousands of chemical compositions into account simultaneously, thus providing huge data resources for knowledge discovery. Due to its specificity, integrity, stability and quantifiability, the fingerprint method has been successfully used in identification and quality evaluation of traditional Chinese medicinal materials. This is an application where knowledge discovery approaches can contribute a lot. Typical data mining methods used in this field include primary component analysis (PCA), ANN, fuzzy clustering, etc. [47]. The main drawback of these methods is that extracted features are needed from original information to form the feature space. For fingerprint data, the dimensions of the feature space are up to 1000 after discretization, and the sample set is usually small. ANN is easy to suffer from overfitting on such small train set, while PCA method is more dependent on the statistical information after discretization. To solve this problem, a novel method based on nearest neighbor (NN) and genetic algorithm was presented by Zhang et al. [48] in 2004. A new measure, named corresponding-peak distance, was proposed to calculate the distance between samples directly for NN classifier, and genetic algorithm was used to optimize the parameters of NN classifier. Experiments on HPLC data of Radix Ginseng indicate that this hybrid method can effectively identify the medicine material of different harvest time or habitats.

Another KDTCM study with regard to CHM chemical compositions was conducted by Lu et al. [49] in 2005. TCM chemical database [20] was used as the data source in this study, and association rule mining was carried out to analyze the relations between CHM efficacy, botanical family, bioactivity of chemical compositions, and pharmacology of CHM extracts. Some meaningful rules were obtained in the experiments. For example, a bidirectional association rule was found between M-cholinoceptor blocking drug and Solanaceae. This rule indicates a large proportion of plants in Solanaceae (21.37%) contain ingredients with M-cholinoceptor blocking activity, and meanwhile, a large proportion of M-cholinoceptor blocking drugs (15.69%) can be found from plants in Solanaceae. Another interesting rule was found between Ranunculaceae and anti-hypertension drug. Such kind of rules found by knowledge discovery techniques is valuable for drug discovery process.

3.3. Knowledge discovery for research of TCM syndrome

Syndrome (zheng) is one of the core issues in TCM; it is a holistic clinical disease concept reflecting the dynamic, functional, temporal and spatial morbid status of human body. Considering the systematic knowledge accumulated from thousands of years’
TCM clinical practice, which is valuable hypothesis for modern biomedicine research, it is of great importance to deepen the study of TCM syndrome. One idea of syndrome research in this post-genomic era is to study syndrome at molecular level. As an effective avenue to combine Western medicine and TCM, knowledge discovery could be undertaken based on the integration of literature and clinical data in both fields.

In a probing research connecting TCM with modern biomedicine, Wu et al. [39] proposed a text mining approach to identifying the gene functional relationships from MEDLINE based on TCM knowledge stored in TCM literature. The TCM literature used in this study comes from TCMLARS, which contains more than 600,000 articles from 900 biomedical journals published in China since 1984. There materials were treated by a simple but efficient bootstrapping method to extract syndrome—disease relationships. Besides, the term co-occurrence was used to identify the relationships between disease and gene from MEDLINE. Then the authors got the syndrome—gene relationships by one-step inference, i.e., to compute the genes and syndromes with the same relevant disease. The underlying hypothesis behind this research was that the relevant genes of the same syndrome would have some biological interactions.

The relationship of kidney YangXu syndrome and related genes was taken as an example by the authors because kidney YangXu syndrome is an important syndrome involving caducity, neural disease and immunity etc. Moreover, through an experimental study of kidney YangXu syndrome at molecular level, Shen [50] found that kidney YangXu syndrome was associated with the expression of CRF (C1q-related factor). This laboratory-confirmed relationship, as well as other novel syndrome—gene relationships, was found revealable by the text mining method in ref. [39].

The authors compiled about 2.5 million disease relevant PubMed citations and 1,479,630 human gene relevant PubMed citations in local database drawn from online MEDLINE. Meanwhile, about 1100 syndrome—disease relationships were obtained from TCM literatures published in 2002. Considering the bi-lingual issues, before searching MEDLINE, Chinese disease name was translated partially automatically to formal English disease name according to the TCM headings database, and manual check by TCM terminological expert was conducted when no TCM headings of disease existed. Filtering by total co-occurrence is above 10 or the number of relevant diseases is above 2, 72 genes related to kidney YangXu syndrome were discovered, in which CRH (corticotropin releasing hormone) was included. By querying gene databases, it was found that CRF was an alias name of CRH. Thus, it means the relationship between kidney YangXu syndrome and CRF had already been discovered by the method. Furthermore, suppose the alias relation between CRF and CRH was not known, the authors showed that it was also capable to confirm the relationship between CRF and kidney YangXu syndrome by the next several steps. First, some important genes in kidney YangXu syndrome were selected, including CRP (C-reactive protein, pentraxin-related), CRH, IL10 (interleukin 10), ACE (angiotensin I converting enzyme), PTH (parathyroid hormone), and MPO (myeloperoxidase). Second, the PubGene [51] was searched for subset network using each of the above genes, and the corresponding subset networks were shown as Fig. 1. The third step was analyzing the extracted knowledge. Now suppose that we do not know CRF is a relevant gene of kidney YangXu syndrome. By analyzing the six subset networks in the left part and the CRF subset network of Fig. 1, we may conclude that CRF is somewhat relevant to kidney YangXu syndrome, because the subset networks, which reassembled with the gene nodes such as IL10, CRAT, CRF/CRH/ACE/MPO/PTH, etc., constitute possible functional gene communities that contribute to kidney YangXu syndrome. No existing literature reporting the relationship between CRF and kidney YangXu syndrome was used to generate the novel knowledge. It is exciting that this simple demonstration has shown the primary text mining results will largely decrease the labor in molecular level syndrome research.

This work proposes a tool for the TCM researchers to rapidly narrow their search for new and interesting genes of a specific syndrome. Meanwhile, the work gives specific functional information to the literature networks and divides large literature networks to functional communities (e.g., the community containing IL10, CRAT, CRF/CRH/ACE/MPO/PTH, etc., genes for kidney YangXu syndrome), which cannot be identified in the current PubGene. Moreover, through syndrome perspective, the study gives an approach to having a subset selection or explanation of giant literature based gene network. The results demonstrate that syndrome can give a novel top—down view of functional genomics research, and it could be a promising research field while connecting TCM with modern life science using text mining and other knowledge discovery methods.

3.4. Knowledge discovery for TCM clinical diagnosis

Diagnosis is of crucial importance in any medical or healing system that works on the body. Unlike Wes-
tern medicine, diagnostic methods in traditional Chinese medicine include four basic methods (called *si zhen* in Chinese): inspection, listening and smelling, inquiry, and palpation. The case history, symptoms, and signs gained through those four diagnostic methods are analyzed and generalized to find the causes, nature, and interrelations of the disease, and to provide evidence for further differentiation of syndromes. Although used in clinical practice for thousands of years, these diagnostic methods seem subjective and unreliable at present time. Thus, researchers begin to apply modern techniques to improve the objectivity of these methods. Knowledge discovery is one of the techniques that could contribute greatly in this process.

Among the diagnostic methods mentioned above, the examination of tongue is one of the most important approaches for getting significant evidences in diagnosing the patient’s health conditions. However, the clinical competence of tongue diagnosis was largely determined by the experience and knowledge of the physicians adopting the tongue diagnosis, and was easily influenced by environmental factors. Therefore, it is necessary to build an objective diagnostic standard for tongue diagnosis. In the last decades, researchers have been developing various methods and computer-based systems to solve this problem. A recent review of studies in pattern recognition of tongue image was undertaken by Yue and Liu [52] in 2004. By analyzing these researches, we can find many cases where image-based knowledge discovery methods got used. A typical study involving knowledge discovery approaches was conducted by Pang et al. [53] in 2004. In this work, two kinds of quantitative features, chromatic and textural measures, were extracted from tongue images by digital image processing techniques, and Bayesian networks were employed subsequently to model the relationship between these quantitative features and diseases. Experiments on a group of 455 in-patients affected by 13 common diseases, as well as other 70 healthy volunteers, show that the diagnosis accuracy based on the previously trained Bayesian networks is up to 75.8%. In another study in 2005, Ying et al. [54] proposed eight characteristic quantity variables of tongue manifestations, including color, shape, wetness—dryness, etc. These features were treated by *k*-means clustering and ANN to predict the diagnostic result. An experiment was performed on 49 patients with cerebrovascular diseases and 39 health people, and the classification accuracies

---

**Figure 1** The six subset networks of each selected genes (left part of the vertical line). Gene CRAT, which in all of the six subset networks, may be a novel relevant gene of *kidney YangXu* syndrome. The right part gives the subset network of the already known relevant gene CRF, and the subset network of CRAT.
for $k$-means clustering and ANN based on the newly presented eight characteristics reach 87.5 and 92%, respectively.

Besides the diagnostic information obtained through tongue diagnosis, other measures, including all kinds of symptoms and signs, are also considered as potential important features in TCM diagnosis. Unlike Western medicine in which interventive are performed only after crisis arises, TCM is characterized by diagnosing diseases at earlier stages and subsequently preventing greater problems from occurring by adjusting the imbalance in the body in time. Therefore, symptoms that are not characteristic features for certain diseases in biomedicine background, are also used in TCM diagnosis in many cases. The rationality of this diagnostic idea is partly supported by a recent study conducted by Lu et al. [55] in 2005. In this work, correlation between CD4, CD8 cell infiltration in gastric mucosa, *Helicobacter pylori* infection and symptoms in 62 patients with chronic gastritis, were analyzed by logistic regression analysis and $k$-means clustering. The results indicate that an assemblage of eight non-digestive related symptoms (such as heavy feeling in head or body, thirst, cool limbs with aversion to cold, etc.) could increase the predicted percentage of CD4 and CD8 cell infiltration in gastric mucosa, including lower CD4 infiltration by 12.5%, higher CD8 infiltration by 33.3%, and also non-*H. pylori* infection by 23.6%. This study also demonstrates that the subjective symptoms could play an important role in the diagnosis and treatment of diseases.

Apart from the studies of subjective symptoms in TCM diagnosis, some novel objective parameters for syndrome differentiation were also proposed by researches, and usually these new parameters were further treated by knowledge discovery methods to establish the diagnoses. In a study conducted by Deng et al. [56] in 1996, the contents of 14 trace elements in hair of 163 cases of rheumatoid arthritis were used as the objective parameters in syndrome differentiation. Dynamic clustering was carried out based on these indexes, and the diagnostic results were compared with clinical diagnoses. Great consistency in this comparison was observed in the experiment, and the accuracy of clustering-based method reached as high as 95.70%.

The studies above show the power of subjective symptoms and objective parameters in TCM diagnosis. However, this does not mean that all subjective symptoms and objective parameters should always be taken into account simultaneously. A recent study undertaken by Li et al. [57] in 2005 showed that the selection of symptoms can influence the accuracy of diagnosis in 209 patients with *H. pylori* infection. Thirty-five clinical presentations were observed in the experiment, including digestive symptoms (such as appetite, stomachache, nausea, etc.), general status (such as complexion, stool, etc.), spirit and psychological status (such as sleep, emotion, etc.), and pathogenic factors (such as smoking, alcohol, etc.). These parameters were analyzed by statistical methods and $k$-means clustering. In the experiment, the diagnostic accuracy based on 35 clinical presentations was 65.7%. It could be improved by 5.7% when only the assemblage of digestive symptoms was engaged, or by 8.6% when the pathogenic factors, general status and tongue observation were combined. The diagnostic accuracy could be decreased when only the general symptoms were engaged, or when the pathogenic factors were accompanied with some common digestive symptoms. We can conclude from this study that it should be cautious on the selection of symptoms in diagnosis.

Researches in the last decades showed that knowledge discovery methods combined with clinical experiments could greatly help to determine this selection. Existing KDD methods used in this task include factor analysis [58], clustering [58], rough set [59], Bayesian networks [60], latent class analysis [61], etc. In an original research conducted by Zhang et al. [58] in 2005, the methods of factor analysis and clustering were combined in the diagnosis of TCM syndromes in 310 patients with posthepatic cirrhosis. The information of conventional four diagnoses was collected by the method of clinical epidemiological research, and these features were reduced and grouped by factor analysis. Based on these obtained common factors, clustering was carried out subsequently to form different groups of syndromes. The results indicate that the TCM syndromes in 287 of the 310 patients (92.58%) could be classified, and the 287 cases could be divided into 7 categories of syndromes, showing the effectiveness of combing factor analysis and clustering. Rough set is another effective knowledge discovery method for attribute reduction, thus is suitable for the selection of symptoms in TCM diagnosis. In 2001, Qin et al. [59] applied rough set in the diagnosis of rheumatoid arthritis. The results show that the diagnostic accuracy of rough set for rheumatoid arthritis is greatly higher than that of fuzzy set. Because of its powerful capacity of handling uncertainty, Bayesian network has become an attractive tool able to model the dependence and independence relationships among the variables in the domain with network. In a self-learning expert system constructed for TCM diagnosis in 2004, a novel hybrid learning algorithm GBPS* based on Bayesian networks was proposed by Wang et al. [60]. This efficient algorithm was applied to discover...
the dependence and independence relationships among symptoms and essential symptoms (named key-elements), and the results were represented by directed acyclic graphs. This explicit representation of Bayesian networks helps to gain an insight into TCM knowledge and explain the diagnosis and treatment of TCM experts. Besides Bayesian networks, this expert system also included other modules to extract knowledge from clinical data automatically, such as the module for mining frequent sets of key-elements. This data-driven nature made this system distinguishing from the rule-based expert systems developed previously.

Another interesting knowledge discovery study in TCM diagnosis was conducted by Zhang and Yuan [61] in 2006. Unlike other researches which focused on the selection of distinguishing symptoms, Zhang aimed at discovering the latent variables hidden in the TCM diagnosis. Considering that latent variable detection might lead to scientific discovery [62], Zhang proposed hierarchical latent class (HLC) models [63] to discover latent structures, and applied HLC models to discover the latent variables in TCM diagnosis of kidney deficiency syndromes [61]. In this study, 2600 cases of the elderly older than 60 years were investigated to collect 67 symptoms related to kidney deficiency syndromes. The HLC model was subsequently learned by a search-based algorithm proposed in ref. [64]. The resultant model was found to match the relevant TCM theories well, and diagnosis based on the model produced conclusions consistent with those by a group of experts. This indicates that latent structure models can help greatly in building an objective statistical foundation for TCM diagnosis.

4. Discussion and future directions

The previous sections aim at presenting a whole picture of state of the art of KDTCM researches. In this picture, a number of future research directions in KDTCM need to be highlighted. In this section, a discussion of these previous KDTCM studies is given, as well as a summary of the future directions in KDTCM.

Due to the preference and more understanding of combinatorial medicines, TCM community provides a large number of real-world cases using multiple drugs as the form of CMF over thousands of years. With the digitalization of these data and the development of computer science, it is very natural for the researches to introduce knowledge discovery techniques, especially the methods based on frequent itemset analysis and association rule mining, to discover compatibility rules of medicines in CMF data. These KDTCM researches are inspiring, with plenty of room for improvement. The extension from conventional methods to correlated pattern mining and graph model learning adapts to the features of CMF analysis better, however, most of these studies still stay at the level of analyzing formulae without the consideration of CHM dosage. The researches of optimal dispensing ratio for CHMs are noteworthy, but new knowledge discovery approaches are still needed to analyze formulae with the fuzzy dosage information in historical literature. Moreover, thorough experiments combined with knowledge discovery methods, the action target of different drugs in a formula could also be analyzed by observing how the dosage change of some drug could influence the curability of related syndrome and symptoms.

Besides dosage, current KDTCM researches in CMF also suffer from the exclusion of some other important attributes of CMF, such as efficacy and indication. As a complex drug system, a CMF involves a large number of effective ingredients, which cooperate with each other and contribute to the final holistic therapeutic effects. Meanwhile, there are complicated relations among ingredients, efficacy, indication, and other elements within this system. A variety of important relationships between these attributes in CMF, such as the correlation between ingredients and indication, between efficacy and indication, and between ingredients and adverse reactions, are still waiting to be disclosed with the aid of knowledge discovery methods.

As for the research of CHM characteristics, the relations between medicinal effects and toxicity, as well as the connections among four natures, five flavors, meridian tropism, botanical family and other elements, are of vast importance to TCM community. Existing KDTCM studies demonstrate that it is very helpful to apply knowledge discovery approaches in this subfield. For example, the strong connections between some botanical families and their toxicities (e.g., Ranunculaceae) discovered by knowledge discovery methods [40] are highly beneficial to toxicological research. Processing is the preparation of crude medicinal materials according to TCM theory, which mainly serves as the functions of reducing side effect/toxicity, and promoting therapeutic effects. Currently, there are few KDTCM researches conducted with regard to processing methods. We believe that the future KDTCM studies in processing methods could help us to gain an insight into the traditional wisdom of TCM in treatment of herb toxicity.

Compared with other subfields, KDTCM studies in chemical compositions of CHM is a bit preliminary. The analysis of trace elements by knowledge dis-
covery methods seems interesting, but is by far not enough. Due to its specificity, integrity, stability and quantifiability, the fingerprint technique has been successfully used in identification and quality evaluation of traditional Chinese medicinal materials, in which various knowledge discovery approaches are used to construct the classes. However, it is worth noting that fingerprint technique also lays the foundation for data mining of spectra-effect relationship. Through analyzing the relativity of the variables generated by fingerprint and the ones related to CHM characteristics and effects, it is promising to aid the discovery of active ingredients in CHM with known and unknown pharmacological actions. To derive new chemical drugs from CHM, the main approach is to extract potential effective ingredients from Chinese herbs as lead compounds, and then design new chemical drugs by structural modification on the basis of these lead compounds. However, during the discovery of lead compounds from natural products, the conventional random screening method suffers from low efficiency. For CHM, the ingredients with similar efficacy always have resemblance in active group. Thus, by QSAR analysis and related methods, knowledge discovery techniques could be utilized to seek promising active groups with regard to some therapeutic effect, which are beneficial to directional screening. However, there are few KDTCM researches currently in this aspect. With the increasing availability of high quality 3D chemical structure databases and the development of KDTCM techniques, we believe that knowledge discovery will contribute greatly to the drug discovery in CHM.

Although the number of KDTCM studies in syndrome research is relatively small, it is very encouraging to see some studies connecting TCM with modern biomedicine in this subfield. In spite of the differences on methodology and technology, modern biomedicine and TCM share the same research subject, i.e., the diseases phenomena of human body. Thus, there are a lot of connections between biomedicine and TCM. For instance, on the one hand, the syndrome research could be carried out at molecular level. On the other hand, the syndrome could give a novel top—down view of functional genomics research, which is particularly required in this post-genomic era. It is proved that text mining could contribute greatly in this multidisciplinary research. Based on the integration of literature in both Western medicine and TCM, the syndrome—gene relationships, including some laboratory-confirmed relationships (e.g., CRF—kidney YangXu syndrome), are found discoverable by text mining. Moreover, through syndrome perspective, current KDTCM study gives an approach to having a subset selection or explanation of giant literature based gene network. As protein—protein interactions are central to most biological processes, the systematic identification of all protein interactions is considered as a key strategy for uncovering the inner workings of a cell. Thus, one aspect of future KDTCM work could focus on studying the protein—protein interactions using the knowledge of TCM syndrome.

In spite of the interesting KDTCM work above in functional genomics research, it should also be noticed that for TCM itself, currently the standard of classification hierarchy for specific syndromes and even the definitions for some syndromes have not reached consensus yet. Thus, the standardization and in-depth study of specific syndrome has become an urgent demand for development and research of TCM. This problem can also be partly solved by knowledge discovery approaches. For instance, information extraction could be carried out on the vast amount of TCM literature, which helps to discover the most frequently used terms to describe syndromes in practice. All variety of classification and clustering methods could be applied on syndrome data, which could provide useful guidance on how to form a classification hierarchy for syndrome. More KDTCM researches in the study of syndrome standardization can be expected in the future.

Apart from the above KDTCM directions mainly with regard to basic medical research, knowledge discovery techniques could also contribute to clinical TCM research. In practice, various kinds of data mining methods have already been used in TCM clinical diagnosis, including factor analysis, clustering, ANN, rough set, Bayesian networks, latent class analysis, etc. Among these methods, Bayesian networks have several advantages. The main reasons are as follows. First, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention, which is particularly suitable for analyzing the complicated features obtained by TCM diagnostic methods. Second, prior knowledge in TCM can be easily represented in Bayesian networks. Third, it is found [65] that diagnostic performance with Bayesian networks is often surprisingly insensitive to imprecision in the numerical probabilities. As an extension of conventional Bayesian networks, HLC model can serve as a better tool to discover latent structures behind diagnostic observations in TCM. As is reported in ref. [61], currently the problem of HLC model lies in the performance, which is only capable to deal with 35 variables in 67 candidates. In the future, this bottle-neck can be expected to be removed with the development of related algorithms.
Another aspect of clinical research in TCM is the clinical evaluation. So far there are hardly any knowledge discovery studies undertaken in this subfield. The reason probably lies in the unavailability of high-quality TCM clinical data sources in the past. However, the situation has been improving in recent years. The establishments of related real-world clinical databases, such as the TCM-EMRD mentioned in section 2, will be of great help to promote the usage of knowledge discovery in TCM clinical evaluation in the future.

5. Conclusion

With a history that spans thousands of years, TCM provides Chinese people with effective health undertakings. As a medical system based on holistic idea, which is totally different from orthodox medicine, TCM is a field worthy of in-depth analysis and research. What KDD is good at is searching the huge volume of data for meaningful patterns and knowledge, which is otherwise an almost impossible task to accomplish manually. Thus, KDD is a necessary technology to be applied in analyzing the great amounts of data in TCM. Based on an introduction to the current state of TCM data resources, this paper provide an overview of knowledge discovery research in TCM in recent years. Major studies in four subfields of KDTCM are reviewed, including KDD for CMF research, KDD for CHM research, KDD for TCM syndrome research, and KDD for TCM clinical diagnosis. The methods used in these studies are introduced, and some interesting results are highlighted. Finally, in a discussion section based on existing KDTCM studies, the current state and main problems of KDTCM work in each subfield are summarized, and the future directions for each subfield are also presented.

The journey of KDTCM in recent years is inspiring, resulting with considerable meaningful discoveries (such as the laboratory-confirmed relationship between CRF gene and kidney YangXu syndrome, the high proportion of toxic plants in Ranunculaceae, the association between M-cholinoceptor blocking drug and Solanaceae, etc.). However, it should be noticed that the KDTCM achievement currently is preliminary. For knowledge discovery, the TCM field is still nearly a piece of virgin soil with copious amounts of hidden gold. Here, the questions arise: where is the gold, what kind of gold is hidden, and how can we mine for the gold? This paper aims at answering these questions based on a review and analysis of existing KDTCM researches, as well as a discussion of future directions. As is indicated by Roddick et al. [11], mining over medical, health or clinical data is arguably the most difficult domain for the KDD field. As a huge non-linear complicated system, human body always involves a large amount of mutual influence and dynamic balance among complex factors, which makes it extremely difficult to fully unravel the inner mystery of life phenomenon. The advantage of TCM lies largely in the holistic thinking pattern and its preference for multiple-component therapy based on natural products. However, this also increases the complexity for KDTCM. Besides, currently the data of TCM still suffer from high individuality, ambiguity and incompleteness. All these bring new problems and challenges when traditional KDD methods are applied to TCM field. These problems are partly solved by the studies reviewed in this paper. However, much more work needs to be done. Considering the ever-increasing volume of TCM data and the pressing demand to extract knowledge from these resources, we believe that the usage and development of KDTCM in the future will substantially contribute to the TCM community, as well as modern life science.

Acknowledgements

We gratefully acknowledge all the researchers from the Institute of Information on Traditional Chinese Medicine, China Academy of Traditional Chinese Medicine for the TCM databases and discussion of TCM topics. This research is partly supported by National Basic Research Priorities Programme of China Ministry of Science and Technology (No. 2005DKA32400), subprogram of China 973 project (No. 2003CB317006), National Science Fund for Distinguished Young Scholars of China NSF program (No. NSFC60533040), and also a grant from Program for New Century Excellent Talents in University of Ministry of Education of China (No. NCET-04-0545).

References

Knowledge discovery in TCM 235


