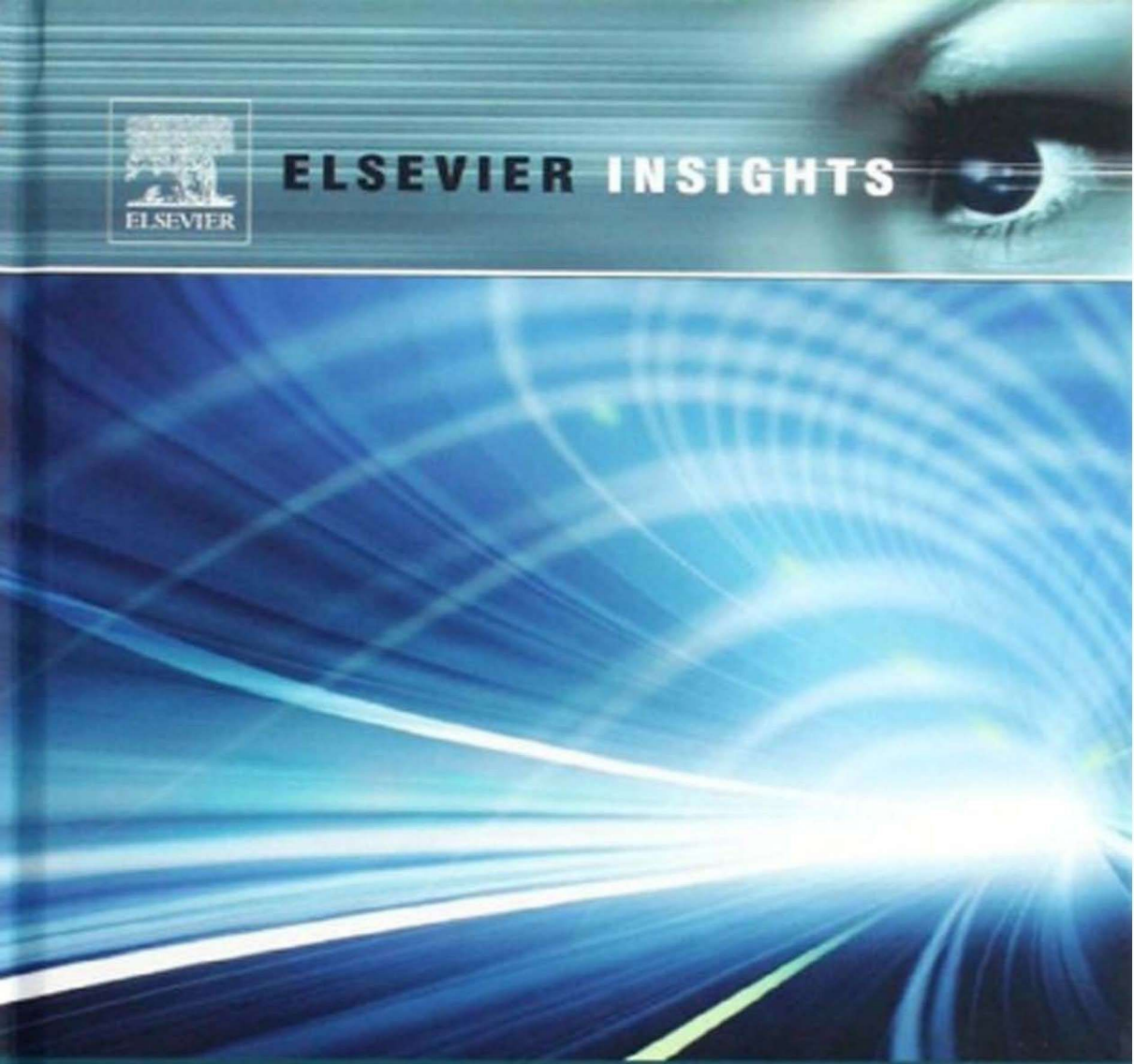

ELSEVIER INSIGHTS

MODERN COMPUTATIONAL APPROACHES TO TRADITIONAL CHINESE MEDICINE

Edited by
ZHAOHUI WU • HUAJUN CHEN • XIAOHONG JIANG

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

现代计算技术与中医药信息处理＝Modern Computational Approaches To Traditional Chinese Medicine：英文／吴朝晖，陈华钧，姜晓红著. —杭州：浙江大学出版社，2012.8

ISBN 978-7-308-08457-4

Ⅰ.①现… Ⅱ.①吴…②陈…③姜… Ⅲ.①中国医药学—英文 Ⅳ.①R2

中国版本图书馆 CIP 数据核字（2011）第 030545 号

现代计算技术与中医药信息处理
吴朝晖　陈华钧　姜晓红　著

责任编辑	黄娟琴　陈静毅
出版发行	浙江大学出版社
	（杭州市天目山路 148 号　邮政编码 310007）
	（网址：http://www.zjupress.com）
	ELSEVIER INC
	网址：http://www.elsevier.com
排　　版	杭州中大图文设计有限公司
印　　刷	浙江印刷集团有限公司
开　　本	710mm×1000mm　1/16
印　　张	15.75
字　　数	516 千
版 印 次	2012 年 8 月第 1 版　2012 年 8 月第 1 次印刷
书　　号	ISBN 978-7-308-08457-4（浙江大学出版社）
	ISBN 978-0-12-398510-1（ELSEVIER INC）
定　　价	120.00 元

Preface

Traditional Chinese medicine (TCM) is both an ancient and a living medical system using fully developed theoretical and practical ideas. In China, traditional medicine accounts for around 40% of all health care delivered. TCM is recognized as an essential component of Chinese culture, and the preservation and modernization of this Chinese cultural heritage is prioritized in the Chinese government's planning program.

TCM has an independently evolving knowledge system, which is expressed mainly in the Chinese language. TCM knowledge discovery and knowledge management have emerged as innovative approaches for the preservation and utilization of this knowledge system. It aims at the computerization of TCM information and knowledge to provide intelligent resources and supporting evidence for clinical decision making, drug discovery, and education.

Specifically, the expansion of TCM practice results in the ongoing accumulation of more and more research documents and clinical data. The major concern in TCM is how to consolidate and integrate the data, and enable efficient retrieval and discovery of novel knowledge from the massive data. Typically, this requires an interdisciplinary approach involving Chinese culture, modern health care, and life sciences. For example, in order to map a global network of herb–drug interactions revealing drug communities, explicit knowledge should be integrated from a plurality of heterogeneous data resources in health care and the life science domain, including electronic health records, literature databases, and domain knowledge databases.

Additionally, TCM knowledge is commonly available in the form of ancient classics and confidential family records, which are disparate among people and organizations across geographical areas. Novel knowledge integration and discovery approaches are thus required to link data across database and organizational boundaries so as to enable more intuitive queries, search, and navigation without the awareness of these boundaries.

The goal of exploring effective methods for knowledge discovery and management in TCM in this book is to provide a systematic interface which can bridge the linguistic gap, cultural gap, and methodological gap between TCM and Western science, by extracting intelligent resources from physicians' theoretical and practical knowledge and applying computational approaches to promote the automatic progress of clinical decision making, drug discovery, and education.

This book compiles a number of recent research results from the Traditional Chinese Medicine Informatics Group of Zhejiang University. This book reports

systematic approaches for developing knowledge discovery and knowledge management applications in TCM. These approaches feature in the utilization of the modern Semantic Web and data mining methods for more advanced data integration, data analysis, and integrative knowledge discovery. Driven by the heterogeneous distribution and obscure literature of TCM data, these methods and techniques mentioned in this book aim to analyze and understand such huge amounts of data in a controllable style and turn this into integrated knowledge. The digested knowledge could thus be used to promote systemized knowledge discovery and mining, so that TCM experts, physicians, even ordinary people, can gain excellent experience of effective knowledge acquisition. For example, a large-scale TCM domain ontology is utilized to improve the quality of search and query, and to interpret statistically important patterns in those reported approaches. Semantic graph mining methodology is developed for discovering interesting patterns from a large and complex network of medical concepts. The platform and underlying methodology has proved effective in cases such as personalized health care with TCM characteristics, TCM drug discovery, and safety analysis. This book can be a reference book for researchers in TCM informatics. Generally, the topics in this book cover major fundamental research issues, track current challenges, and present core applications. Specifically, we mainly make important contributions to TCM knowledge discovery and management from the following aspects:

1. TCM Data Mining

TCM is a completely dependent discipline, and is a complementary knowledge system to modern biomedical science. Due to diverse and increasing biomedical data, it is difficult to obtain effective information for applications from such massive data. Different forms of TCM data also hinder integration of information sources from different disciplines. Data mining techniques provide flexible approaches to uncovering implicit relationships in these data sources.

Chapter 2 assumes that related genes of the same syndrome will have some biological functional relationships, and thus constitute a functional gene network. We generated syndrome-based gene networks from 200,000 syndrome-gene relations, in order to analyze the functional knowledge of genes from the syndrome perspective. The primary results suggest that it is worthy of further investigation. In Chapter 3, a path-finding algorithm in the context of complex networks was designed to detect network motifs. Performance evaluation is made to learn about the data block size, node number, and network bandwidth in considering the MapReduce-based path-finding performance. In addition to the architectural level of data mining methods, Chapter 5 presents a unified Domain-driven Data Mining platform, which carries out data mining applications by resource orchestration through web services from a variety of heterogeneous intelligence resources and data. The effectiveness of this platform has been proved by a series of in-use applications in the TCM domain. Semantic associations are complex relationships between resource entities, which is a topic studied in Chapter 10 to explore and interpret the knowledge assets of TCM. A case study is demonstrated that discovers and integrates relationships and interactions on TCM herbs from distributed data

sources. Chapter 13 presents a novel approach, which utilizes node and link types together with the topology of a semantic graph to derive a similarity graph from linked datasets. Semantic similarity is calculated through semantic similarity transition during the process of generating a similarity graph.

2. TCM Knowledge Discovery and Retrieval

Confronted with the increasing popularity of TCM and the huge volume of TCM data, there is an urgent need to explore these sources effectively so as to generate useful knowledge, by the techniques of knowledge discovery and retrieval. Knowledge discovery is one proper methodology for analyzing such heterogeneous data.

Chapter 1 provides readers with a perfect overview of knowledge discovery in TCM, including knowledge discovery in a database (KDD) for the research of Chinese medical formula, Chinese herbal medicine, TCM syndrome research, and TCM clinical diagnosis. Chapter 4 attempts to investigate data-quality issues particularly in the field of TCM. Three data-quality aspects are highlighted as key dimensions, including representation granularity, representation consistency, and completeness, and practical methods and techniques are proposed to handle data-quality problems. In order to achieve seamless and interoperable e-Science for TCM, Chapter 6 presents a comprehensive approach to building dynamic and extendable e-Science applications for information integration and service coordination of TCM. The semantic e-Science infrastructure uses domain ontologies to integrate TCM database resources and services, and delivers a semantic experience with browsing, searching, querying, and knowledge discovery for users. Chapter 11 introduces an in-use application deployed at the China Academy of Traditional Chinese Medicine (CATCM), in which over 70 legacy relational databases are semantically interconnected by a shared ontology, providing semantic query, search, and navigation services to the TCM communities. Chapter 12 proposes a probability-based semantic relationship discovery method, which combines a TCM domain ontology and more than 40,000 relative publications so as to uncover hidden semantic relationships between resources. A probabilistic RDF model is defined and used to store semantic relations identified as uncertain and assigned with a probability.

3. TCM Knowledge Modeling

Knowledge representation is a primary step in understanding the nature of diverse domains and conducting useful applications, especially in scientific fields such as biology, economics, and medicine. A lot of knowledge modeling techniques are proposed to solve different levels of knowledge representation obstacles so as to construct an applicable infrastructure.

As a complete knowledge system, TCM researches into human health care via a different approach compared to orthodox medicine. In Chapter 7, a unified traditional Chinese medical language system (UTCMLS) is developed through an ontology approach, which will support TCM language knowledge storage, concept-based information retrieval, and information integration. It is a huge project which was collaborated on by 16 distributed groups. Moreover, unlike Western Medicine,

knowledge in TCM is based on inherent rules or patterns, which can be considered as causal links. Chapter 8 presents a semantic approach to building a TCM knowledge model with the capability of rule reasoning using OWL 2, a kind of web ontology language defined by the W3C consortium. The knowledge model especially focuses on causal relations among syndromes and symptoms and changes between syndromes. The evaluation results suggest that the approach clearly displayed the causal relations in TCM and shows great potential in TCM knowledge mining. The on-demand and scalability requirement ontology-based systems should go beyond the use of static ontology and be able to self-evolve and specialize in the domain knowledge. Chapter 9 refers to the context-specific portions from large-scale ontologies like TCM ontology as sub-ontologies. A sub-ontology evolution approach is proposed based on a genetic algorithm for reusing large-scale ontologies.

For a short overview, the book is specifically organized as follows: Chapter 1 gives an overview of the progress of knowledge discovery in TCM; Chapter 2 introduces a specific text mining application that integrates TCM literature and MEDLINE for functional gene networks analysis; Chapter 3 introduces a novel approach that utilizes a MapReduce framework to improve mining performance with an application of network motif detection for TCM; Chapter 4 discusses the data-quality issue for knowledge discovery in TCM; Chapter 5 reports on a service-oriented mining engine and several case studies from TCM; Chapter 6 elaborates on a systematic approach to TCM knowledge management based on Semantic Web technology; Chapter 7 introduces a large-scale ontology effort for TCM and describes the unified traditional Chinese medical language system; Chapter 8 reports an approach to modeling causal knowledge for TCM using OWL 2; Chapter 9 discusses the ontology evolution issue as related to TCM web ontology; Chapter 10 proposes an ontology-based technical framework for hypothesis-driven Semantic Association Mining, which allows a knowledge network to emerge through the communication of Semantic Associations by a multitude of agents in terms of hypotheses and evidence; Chapter 11 describes a Semantic Web approach to knowledge integration for TCM; Chapter 12 introduces a probabilistic approach to discovering semantic relations from large-scale traditional Chinese medical literature; Chapter 13 presents methods of analyzing semantic linked data for TCM.

This book is the result of years of study, research, and development of the faculties, Ph.D. candidates, and many others affiliated to the CCNT Lab of Zhejiang University. We would like to give particular thanks to Xuezhong Zhou, Peiqin Gu, Xiangyu Zhang, Yuxin Mao, Xiaoqing Zheng, Yi Feng, Yu Zhang, Chunyin Zhou, Tong Yu, Jinhua Mi, Yang Liu, Junjian Jian, Sen Liu, Mingkui Liu, Hao Shen, Jinhuo Tao, and many others who have devoted their energy and enthusiasm to this book and relevant projects.

We would also like to give particular thanks to our long-term collaborator: the China Academy of Chinese Medical Science (CACMS). We would like to thank Hongxin Cao, director of CACMS, and Baoyan Liu, vice director of CACMS, who gave us ongoing strong support over the past 10 years. Also, we are grateful to Meng Cui, the director of the Institute of TCM informatics of CACMS, and all of

his kind colleagues. Without their strong support, this book could not have been finished.

In addition, the work in this book was mainly sponsored by the “973” Program (National Basic Research Program of China) of the Semantic Grid initiative (No. 2003CB317006); the National Science Fund for the Distinguished Young Scholars of China NSF Program (No. NSFC60533040); and the Program for New Century Excellent Talents in University of the Ministry of Education of China (No. NCET-04-0545). The work was also partially supported by the National Program for Modern Service Industry (No. 2006BAH02401); “863” Program (National High-Tech Research and Development Program of China) (Nos. 2006AA01A122, 2009AA011903, 2008AA01Z141); the Program for Changjiang Scholar (IRT0652); the NSFC Programs under Grant No. NSFC61070156, NSFC60873224, Important Programs of Zhejiang Sci-Tech Plan (No. 2008C03007).

Zhaohui Wu
Zhejiang University, Hangzhou, China
November 2011

Contents

1 Overview of Knowledge Discovery in Traditional Chinese Medicine[1]

1.1 Introduction

As a complete medical knowledge system other than orthodox medicine, traditional Chinese medicine (TCM) has played an indispensable role in health care for Chinese people for thousands of 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 came to realize the limitations of reductionism and began to lay more emphasis on 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, 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 (CHM) and acupuncture worldwide is a good indication of the public interest in TCM [2–6].

Countless TCM practices and theoretical research over thousands of years accumulated a great deal of knowledge in the form of ancient books and literature. In China, the domestic collection of ancient books about TCM published before the 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 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, databases, statistics, and machine learning, the idea of KDD

[1] Reprinted from Feng Y, Wu Z, Zhou X, Zhou Z, Fan W. Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. Artif Intell Med 2006;38:219–36. © 2006 Elsevier B.V., with permission from Elsevier.

Modern Computational Approaches To Traditional Chinese Medicine. DOI: http://dx.doi.org/10.1016/B978-0-12-398510-1.00001-7

came into being in the late 1980s. The most prominent definition of KDD was proposed by Fayyad et al. [7] in 1996. 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], DM is the knowledge extraction step in the KDD process, which also involves the selection and preprocessing of appropriate data from various sources and proper interpretation of the mining results. Typical DM methods include concept description, association rule mining, classification and prediction, clustering analysis, time-series analysis, text mining, and so forth [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 of medicine by Altman [10]. As Roddick et al. [11] indicate, the application of KDD to medical datasets is a rewarding and highly challenging area. Due to the ever-increasing accumulation of biomedical data and the pressing demand to explore these resources, the methods of knowledge discovery have been widely applied to analyze medical information over 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 research. 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 chapter 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 there are covered in this review. For each work, the KDD methods used in the study are introduced, as well as corresponding results. In particular, 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, and the association between the M-cholinoceptor blocking drug and Solanaceae discovered by

association rule mining. The existing work in KDTCM demonstrates that the usage of KDD in TCM is both feasible and promising. Meanwhile, it should be noted 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 chapter is arranged as follows. The prerequisite for applying KDD is the digitalization of the vast amount of data. Thus, an overview of currently available TCM data resources is first presented in Section 1.2. Subsequently, the review of KDTCM work is presented in four research subfields in Section 1.3, including KDD for the research of Chinese medical formulae (CMF), KDD for the research of 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 1.4, and the future directions for each subfield are also presented. Finally, we conclude in Section 1.5.

1.2 The State of the Art of TCM Data Resources

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

As a significant part of complementary and alternative medicine (CAM), literature reporting TCM issues can be found in the main CAM databases, such as CAM on PubMed (Complementary and Alternative Medicine subset of PubMed), AMED (Allied and Complementary Medicine Database), CISCOM (Centralized Information Service for Complementary Medicine), and CAMPAIN (Complementary and Alternative Medicine and Pain Database). A more comprehensive list of TCM databases can be found in Ref. [17]. Currently, the primary data resources specific to TCM include 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 and Development Center, a subsidiary of the Intellectual Property Publishing House of the State Intellectual Property Office (SIPO) of China. More than 19,000 patent records and over 40,000 TCM formulae published from 1985 to the present are contained in CTCMPD [18]. TradiMed Database was built by the Natural Product Research Institute at Seoul National University, Republic of 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]. The TCM chemical database was developed by the National Key Laboratory of Bio-chemical 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 the TCM-online Database System is the largest TCM data collection in the world. The prototype of TCM-online was first built in the late 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 multidatabase query system TCMMDB [21] 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 the Traditional Chinese Medical Literature Analysis and Retrieval System (TCMLARS), Traditional Chinese Drug Database (TCDBASE), and Database of Chinese Medical Formula. TCMMDB was replaced by the Grid-based system TCM-Grid [22] in 2002, which provides more powerful functions, such as dynamic registration, binding, and associated navigation. The TCM-Grid system was further extended to a semantic-based database Grid named DartGrid in 2002. At present, these databases are available as the TCM-online Database System via web site [23] 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.

1.2.1 Traditional Chinese Medical Literature Analysis and Retrieval System

The bibliographic system TCMLARS [25] has two versions. So far, the Chinese version contains over 600,000 TCM periodical articles, while the corresponding number reaches 92,000 in the 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 to MEDLINE, such as title, author, journal title, publication year, and abstract. Besides, some fields specifically existing in TCM are also included, such as pharmacology of Chinese herbs, ingredients and dosage of formulae, drug compatibility, and acupuncture and Tuina points. TCMLARS is considered an important new asset in the literature review and meta-analysis of CHM by McCulloch et al. [26]. It also serves as a significant data resource for KDTCM, especially for the methods based on text mining.

1.2.2 Figures and Photographs of Traditional Chinese Drug Database

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