Methodological Review

Text mining for traditional Chinese medical knowledge discovery: A survey

Xuezhong Zhou\(^a\), Yonghong Peng\(^b\), Baoyan Liu\(^c\)

\(^a\) School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
\(^b\) School of Computing, Informatics and Media, University of Bradford, BD7 1DP, UK
\(^c\) China Academy of Chinese Medical Sciences, Beijing 100700, China

A R T I C L E   I N F O

Article history:
Received 18 October 2008
Available online 13 January 2010

Keywords:
Text mining
Traditional Chinese medicine
Review

A B S T R A C T

Extracting meaningful information and knowledge from free text is the subject of considerable research interest in the machine learning and data mining fields. Text data mining (or text mining) has become one of the most active research sub-fields in data mining. Significant developments in the area of biomedical text mining during the past years have demonstrated its great promise for supporting scientists in developing novel hypotheses and new knowledge from the biomedical literature. Traditional Chinese medicine (TCM) provides a distinct methodology with which to view human life. It is one of the most complete and distinguished traditional medicines with a history of several thousand years of studying and practicing the diagnosis and treatment of human disease. It has been shown that the TCM knowledge obtained from clinical practice has become a significant complementary source of information for modern biomedical sciences. TCM literature obtained from the historical period and from modern clinical studies has recently been transformed into digital data in the form of relational databases or text documents, which provide an effective platform for information sharing and retrieval. This motivates and facilitates research and development into knowledge discovery approaches and to modernize TCM. In order to contribute to this still growing field, this paper presents (1) a comparative introduction to TCM and modern biomedicine, (2) a survey of the related information sources of TCM, (3) a review and discussion of the state of the art and the development of text mining techniques with applications to TCM, (4) a discussion of the research issues around TCM text mining and its future directions.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

The current flood of text documents is increasing the demand for new data mining methods for text data processing. Text mining (TM) or knowledge discovery in text, which aims at extracting structured information or discovering novel knowledge (e.g. producing a scientific hypothesis) from large volumes of textual media (e.g. literature, emails and documents) using data mining, machine learning, statistics, and natural language processing (NLP) techniques [1,2], is a hot research topic. Various methods, such as information retrieval (IR [3]), information extraction (IE [4]), text classification & clustering (TC) and topic detection, have been put together for TM applications [5].

In the biomedical science fields, there has been an unprecedented growth in both biomedical experimental data and the amount of published literature during the past decades, which makes it difficult even for biomedical researchers to track new information and knowledge in their own specific field. This results in the loss of novel hypotheses which are buried in the data mountains. TM of the published biomedical literature (e.g. MEDLINE, the annotations of Swiss-Prot and GenBank) has shown great promise for closing the gap between the availability of large amounts of data and the difficulty of obtaining new knowledge for biomedical research. Biomedical TM has thus become one of the hot topics in both bioinformatics and the TM fields [6–12]. The related methods, applications and tools of biomedical TM have been intensively reviewed [9,11,13–20]. The core tasks of biomedical TM are to recognize the biomedical named entities (e.g. genes, proteins, diseases and drugs) [21–23], expose the inter-related relationships of these biomedical entities [10,24–28], and find novel scientific knowledge or hypotheses among the biomedical entities extracted from the biomedical literature [20].

For biomedical TM research, MEDLINE is one of the distinguished biomedical literature databases with more than 17 million records in 2009. Pioneering studies in medical knowledge discovery from MEDLINE have demonstrated the great potential for extracting innovative knowledge from the literature [6,7]. Furthermore, recent studies demonstrated the promise of combining different biomedical data sources, such as expression, sequence and literature data, by means of integrative data mining, to generate useful knowledge [12,29–36].
As a system of healing and treatment, Traditional Chinese Medicine (TCM) has a long history in Chinese society [37]. The philosophy of TCM very much reflects the classical Chinese belief that the life and activity of individual human beings has an intimate relationship with the environment. In TCM, the general principle of health and the ultimate goal of treatment are to maintain the balance of yin and yang [38] inside the human body. TCM defines a different methodology and approach for disease diagnosis and treatment, which has been widely accepted in China [39–41]. The reported data of the National Bureau of Statistics of China in 2007 [42] shows that there are 2720 TCM hospitals (see Glossary) and 123,760 TCM clinicians (including physicians and apothecaries) in China. In 2007, the number of inpatients in TCM hospitals reached 6,930,000 and the number of visits to outpatients and emergency cases is about 2210 million. Even the doctors trained in modern western medical programs in China consider that Chinese herbal medicine is safe and would like to use them to supplement western medicine in treating patients with chronic or intractable illness [43].

In the past decades, TCM has been increasingly adopted as a complementary medical therapy around the world [44,45]. Actually, TCM has been successfully applied to the treatment of various complex diseases [46], such as cancer [47], rheumatoid arthritis [48], promyelocytic leukemia [49,50], migraine [51] and irritable bowel syndrome [52], and its effectiveness has been validated in modern clinical or laboratory studies. However, establishing a practical and rational efficacy assessment system is a vital issue if TCM is to be widely accepted and used [53].

It is widely accepted in China that TCM and modern biomedicine are mutually beneficial and complementary in generating an understanding of the body and of disease phenomena [39,54]. It is hoped that the integration of TCM and modern medical therapies will provide great possibilities for developing novel methods of disease treatment [55,56]. One example is the integrated use of TCM and modern medicine in the treatment of SARS, which has proved to be more effective than the use of modern therapies alone [57,58].

Since 2003, the Chinese government has initiated several significant scientific programs to modernize TCM in the hope of making significant progress in improving our understanding of the human body and the treatment of chronic disease. One of these programs is the digitization of TCM ancient literature data, clinical data and research publications. The digitized data provides the basis for the application of advanced information technologies to modernize TCM. Various computing and statistical methods have been used in TCM clinical studies, clinical decision support, and TCM knowledge discovery [59]. In this paper, we intend to provide a review and discussion of the basic knowledge of TCM, the related TCM information sources, the related TCM work and the research issues for the future development of TCM TM.

The rest of the paper is organized as follows. Section 2 describes the methods used for searching the literature and the selection of articles used in this review. Section 3 presents a brief introduction to TCM and a comparison of TCM and modern biomedicine. We introduce the relevant information sources, related research and future directions for TM in TCM in Section 4. Finally, we present a discussion of the results and our conclusions in Sections 5 and 6, respectively.

2. Methods and scope

This article surveys the state-of-the-art work of TM in TCM and the related data sources over the period from the beginning of 1999 to the beginning of 2008. We performed a keyword query in the CNKI (China National Knowledge Infrastructure, http://www.global.cnki.net/) full-text database, which is one of the biggest databases of Chinese journals and academic publications, to acquire the relevant articles published in Chinese. International publications were selected from the journals of Elsevier, Springer, ACM and IEEE. The PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) bibliographic database was used to perform the bibliographic queries for: ‘text mining or information extraction or text classification or text clustering in traditional Chinese medicine’ and ‘database in traditional Chinese medicine’.

We focus on publications that are concerned with TM techniques for the processing of TCM data. In addition, publications on the TCM databases are also selected. In order to introduce practical opinions and the basic theories of TCM, we also refer to popular textbooks and up-to-date international publications in the international journals (e.g. JAMA, PNAS and The Lancet).

3. Background: a brief introduction to traditional Chinese medicine

As two different methodologies to view the phenomena of human life and disease, TCM and modern biomedicine have developed two distinct medical systems for diagnosis and treatment. TCM views the human body and the environment as two parts of a system in which they interact with each other. It embodies rich dialectical thoughts from the ancient Chinese philosophies.

3.1. Basic TCM concepts and theories

It is known that the basic theories of TCM were formed more than 2000 years ago [37]. Many distinguished classical books (e.g. Huangdi’s Classic of 81 Medical Problems, Treatise on Cold-Induced and Miscellaneous Diseases, Shenmeng’s Herbal, The Pulse Classic and Treatise on Cause and Symptoms of Diseases) in Chinese have been written to decipher the basic TCM theories and concepts. The basic TCM concepts and theories include qi, yin-yang, five phases, the human body channel system, zang fu, organ and syndrome (see Glossary), etc. In this subsection, a brief introduction to the most fundamental TCM theories is presented, including the qi theory, yin-yang theory and the five phases theory. The interested reader could refer to the bilingual or English textbooks [38,60-65] for more information.

3.1.1. Qi theory

The theory of qi has been considered as one of ancient Chinese philosophical thought, which explains the origin, development and variation of things in the universe [61]. Qi originally refers to clouds and gases in the sky. Later on, by daily observation and theoretical abstraction, the qi is considered as the common origin of all things in the universe, and the most fundamental, primitive and disposable substance [62]. It is also generally believed that movement of qi is the basis for variations in all things, and the qi is the medium by which all things are inter-related with one another.

In TCM, it is considered that the qi is the fundamental substance which constitutes the human body, and the regular movement of qi is essential to maintain the human life. Thus, various pathological changes in the human body are attributed to abnormality in the qi [62]. The essence of TCM diagnosis is to understand where the flow of qi has been disturbed and, once known, the aim of treatment is to re-balance the harmonious flow of qi [63].

There are many specific methods in TCM to manipulate, augment and balance the flow of qi including regulating the emotions, moderating the diet, balancing work and rest, prescribing Chinese herbal medicine, acupuncture therapy, etc. [61]. The common
target is to ensure sufficiency of qi and smoothness in qi movement by regulating and nourishing qi.

3.1.2. Yin-yang and five phases theories

The theory of yin and yang holds that the world is a material wholeness and the result of the unity and opposition of yin qi and yang qi. The interaction between yin and yang is fundamental for the occurrence, development and change of things [61]. There are four main aspects of yin and yang relationships, namely the unity of opposites, waxing and waning, interdependence and inner-transformation [64]. This means that yin and yang are two opposites in a unified system, when one is waxing, the other is waning and vice versa. On the other hand, yin and yang are inter-dependent (i.e. one cannot exist without the other) and can be transformed into the other.

TCM considers that yin and yang always exist in the human body, and the human body suffers from illness when an imbalance of yin and yang exists. The core treatment principle of TCM is thus to restore the proper balance of yin and yang. It has been said that all Chinese medical physiology, pathology, and treatment have been developed based on yin and yang [65].

In TCM theories, five phases or five elements (Wu Xing in Chinese) refer to metal, wood, water, fire and earth. The doctrine of five phases was used to illustrate the nature of things and the relationships between them, based on their properties, movements and interactions [64]. It is considered that there are two cycles of balancing: a generating cycle and an overcoming cycle [66].

The yin-yang and five phases theories are the fundamental theories in TCM, which build a universal infrastructure for the specific theories including the diagnosis related theories, such as zang fu theory, syndrome differentiation (see Glossary) theories, pathology and pathogeny theories, and the treatment related theories like the therapeutical principle, herb prescription compatibility and herb nature, etc.

3.2. Diagnostics and treatment

TCM diagnostics is based on the overall observation and analyses of human symptoms. Four basic diagnostic skills and procedures are used in TCM, namely inspection, olfaction and auscultation, interrogation and palpation [65]. Based on these four skills, TCM practitioners acquire the essential clinical information about the disease, and provide the evidence and prerequisite information for diagnosis. TCM diagnostics has one distinct kind of diagnosis named syndrome, which is the outcome of the analysis of the symptoms.

The methods of TCM treatment include Chinese herbal medicine, acupuncture, moxibustion [38,60], massage, food therapy, physical exercise, etc. [41]. In China, Chinese herbal medicine is considered as the primary therapeutic form of internal medicine. Rather than being prescribed individually, herbs are combined into formulae (Chinese medical formulae, see Glossary) to meet the specific needs of individual patients according to their corresponding syndromes.

The primary principle of TCM diagnostics and treatment is the bian zheng lun zhi (see Glossary), which forms a unified procedure to prescribe effective therapies for individual patients [61,64].

3.3. The scientific differences between TCM and modern biomedicine

Liu and Wang [67] outlined five main scientific differences between TCM and modern biomedicine, namely the start-points of research, the objects of research, the modes of research, the methodologies and the theoretical characteristics.

TCM regards the patient as a whole functional life system in the context of the social and natural environments, and takes the functional information (e.g. symptoms, signs and behaviors) of the patient (or healthy person) and the environment as the research object. The functional information at the holistic level is indeed complicated and diverse. For example, two different patients with common chronic diseases (e.g. type 2 diabetes mellitus) often manifest completely different and diverse symptoms. TCM assimilates Chinese philosophical theories, which argue that there exists a unified law, called the theory of yin-yang and five phases, in the universe. Based on this, TCM practitioners attempt to grasp complicated patient information in practical clinical operations. As there is still a gap between different TCM theories, in both abstract and conceptual representation, and clinical procedures, the clinical effectiveness of TCM is actually influenced by the competence of the TCM practitioners.

Moreover, the research mode applied in TCM is clinically-based. TCM clinical operations are innovative procedures conducted on real-world patients with empirical reasoning and deduction based on TCM doctrines. TCM physicians make specific diagnoses and prescribe formulae based on general TCM theories and their empirical knowledge (including personal experience and empirical knowledge from the published literature). As a result, clinical practice provides the most important knowledge source for TCM research, especially the records of the daily clinical efforts of TCM physicians. Mining the clinical data in both the ancient and the contemporary literature has great promise to generate clinical knowledge to fill the gap between TCM theories and clinical practice.

In contrast, modern biomedicine focuses on the phenomena of the structures and substances of the human body. The reductionist, analytical and differential methods used in modern biomedicine are primarily aimed at analyzing the structure and substance of human body. In modern biomedicine, biomedical knowledge discovered in the experimental research plays a dominant role in clinical practice. The discovery obtained in experimental research help modern biomedicine practitioners tackle most of the common clinical cases. However, the transformation of the discoveries from the ‘bench’ into sustainable solutions for public health delivered at the ‘bedside’ has become significant issue for the development of modern biomedicine [68]. Furthermore, as characterized by the nature of experimental work, it is difficult for modern biomedicine practitioners to deal with new or intractable diseases that have no clinically evaluated symptoms recorded in the existing experimental studies.

In conclusion, TCM has many advantages in clinical practice, particularly the knowledge of the phenotypic regularities of the human body and the interaction between the human body and the natural environment. The large volume of TCM clinical data and published clinical literature provides a significant data source for the discovery of new knowledge. Modern biomedicine, on the other hand, has its advantages in experimental practice and the availability of large volume of micro-level data about the structure of human body. The integration of TCM and modern biomedicine is becoming possible with the increasing support of advanced computing and informatics technologies [69].

4. Results

In this section, a review of the related TCM information sources for TM and the related TCM TM research is presented. The potential applications and future development of TM in TCM is discussed.

4.1. TCM information sources for text mining

TM in TCM concerns the extraction, analysis and visualization of hidden knowledge (e.g. TCM named entities, symptom–syndrome
4.1.1. TCM bibliographic literature databases

As shown in Fig. 1, the TCM bibliographic literature databases that contain the citations for journal articles in TCM are one of the main data sources for TM applications. They are manually curated and maintained by various TCM libraries. There are about 149 TCM journals published in mainland China, 132 of them are academic journals and the others are general or popular science journals [72]. To meet the need for TCM bibliographic literature information services, the Institute of Information on TCM of the China Academy of Chinese Medical Sciences developed a Traditional Chinese Medical Literature Analysis and Retrieval System (TCMLARS) in the 1980’s. The TCMLARS has accumulated over 800,000 references and abstracts which include Chinese herbal medicine, acupuncture, qigong, and Chinese massage and health promotion. The database is available at the website [73] for registered users. The source material for the database is around 900 biomedical journals published in China since 1984 [70]. The structure of TCMLARS is similar to MEDLINE, and contains fields including paper title, author, journal title, the year of publication and abstract. In addition, it contains several fields that are specifically designed for TCM, including the pharmacology of Chinese herbs, ingredients and the recommended dosage of formula, drug compatibility, acupuncture, etc. TCMLARS has been categorized into several subsets according to specific diseases: tumors, diabetes, AIDS and geriatric diseases, in order to facilitate the data searches. Currently, about 60,000 records are being added to TCMLARS each year. TCMLARS also has an English version database1, which contains about 68,000 records and provides an English keyword query facility using specific data fields, such as title, abstract and pathogenesis.

The China TCM patent database (CTCMPD) is another bibliographic literature database [70,74]. The CTCMPD has been established by the Patent Data Research & Development Center, which is a subsidiary of the Intellectual Property Publishing House of the Chinese State Intellectual Property Office (SIPO). More than 22,000 patent records published from 1985 to the present have been included in the CTCMPD [75].

It is necessary to mention that only very limited TCM bibliographic literature data has been included in the international databases, such as MEDLINE, EMBASE and BIOSIS. For example, among 149 TCM journals currently published in mainland China, only 10 TCM journals (e.g. Zhongguo zhong yao zu zhi, Journal of traditional Chinese medicine and Chinese medical journal) were indexed by the MEDLINE [72]. It can be seen that the Chinese TCM bibliographic literature databases in mainland China are the main information sources for TM applications.

4.1.2. Annotated ancient TCM literature database

The ancient TCM literature database is presented in a semi-structure linked to that of the ancient TCM books. Because most of the TCM literature is prepared by classical Chinese words, the manual annotation of these ancient books is a challenging task. TCM experts decipher the data sentence by sentence. Liu [76] developed an indexing method based on the knowledge elements of TCM to support the annotation of ancient Chinese medical literature with tags in structured XML documents. This method focuses on the indexing of the principal TCM classifications, such as formula, herb, symptom and syndrome, based on domain ontology [77]. The indexing is a semi-automatic process, which starts with the manual extraction of the terminological tags allocated by TCM experts and their representation in XML format. The experts read through the TCM book, and mark the text segments to capture the independent information of a TCM term or concept like liquorice root. The text segments are then tagged with different XML labels. For example, if a paragraph discusses the information of a herb like liquorice root, then an XML node of herb classification is inserted in the XML document. Furthermore, in the annotated paragraph, the attributes for herb classification, such as nature and flavor, channel entry, harvesting, processing, synonym, explanatory terms, identification and habitat are manually inserted in the
appropriate locations of the XML document. As a result, a TCM book is represented in an XML document with various nodes of semantic classifications and attributes.

One recent task has been the annotation of the ancient TCM *materia medica* books. By 2008, over 260 ancient TCM books with more than 60 million Chinese words have been annotated. A web system (as shown in Fig. 2) using keywords to search the database has been developed and is available for free at the website [78]. Compared with the original ancient TCM literature that is represented entirely by classical Chinese text, the annotated ancient TCM literature provides an important data source for TM applications with the paragraphs and sentences marked with semantic labels by TCM experts.

4.1.3. TCM clinical data warehouse

Daily clinical practice plays a vital role in TCM research to support the refinement of TCM theories. It has been recognized that the electronic medical record (EMR) for both inpatients and outpatients is a significant data source for TCM research [79]. Free-text EMR data has been collected in TCM hospitals in the major cities (e.g. Beijing, Shanghai and Guangzhou) of China for over ten years [80]. Since 2002, a clinical data warehouse (CDW) has been developed for the integration and management of structured TCM EMR data [81,82]. By 2007, the CDW had collected data from about 20,000 inpatients with conditions of diabetes, coronary heart disease and stroke from 10 TCM hospitals and TCM wards in the western medical hospitals in Beijing. In addition, more than 20,000 outpatient data instances have been recorded including the outpatient data from 20 high-experienced TCM physicians in Beijing.

The CDW has the TCM clinical information model, a physical data model and a multidimensional data model to manage the clinical data. Besides including a tool to perform the preprocessing and integration of clinical data, the CDW platform integrates data mining components, such as Weka [83], Oracle data miner (ODMiner) [84], and the business intelligence tool (BusinessObjects) [85]. This leads to the implementation of a TCM clinical intelligence platform that provides an effective infrastructure for online analytical processing (OLAP) and TCM clinical knowledge sharing.

The main contents of the CDW include TCM diagnosis, symptoms and formula. Compared to the clinical data of modern biomedicine, TCM clinical data contains distinct information components such as TCM symptoms and signs, syndrome, formulae and herbs, which form the core elements of TCM clinical data. Because the structured data entry is an additional task for TCM physicians, an automatic tool is required for the extraction of structured data from free-text data. TM methods, such as IE and named entity recognition (NER, see Glossary), could be used for this purpose [86], however, in order to maintain patient confidentiality, the clinical data warehouse information is not publicly available on the web.

4.1.4. TCM terminology systems

Due to the various expressions, synonyms and phrases used in the clinical literature, it is challenging to perform NER tasks for TCM data. To get reliable discovery results, it is necessary to develop a standardized terminology system that has a systematic definition of medical concepts with an appropriate hierarchical structure. A medical ontology framework, called the unified traditional Chinese medical language system (UTCMSL), has been developed for this purpose [87]. The UTCMSL proposes an effective organization framework for the TCM terminological sources to support the construction of the linguistic knowledgebase and concept-based information retrieval. Hundreds of TCM based terminologies and vocabularies such as the *Traditional Chinese Medical*
4.1.5. Structured basic databases

The basic TCM information sources, such as formula databases, the Chinese traditional patent medicine (CTPM) database, herb (e.g. Chinese materia medica, Tibetan herb and Mongolia herb) databases, the herb ingredient database and the disease database, are developed and managed by several organizations in China. This data is represented in a structured relational database, and it is mainly collected from the reference books, published literature and publicly available data sources. Cui [71] has reviewed the research of structured basic databases in China before 2004. TCM Online [70,73,97] is one of the major TCM basic information resources, which comprises more than 20 structured basic databases, including formula databases, herb databases, herb ingredient databases, a TCM organization database, TCM OTC prescription database and a TCM news database. Most of the structured basic databases are in Chinese, but several bilingual and English structured basic databases, such as TradiMed [98] and TCM-ID [99], have also been developed.

The structured basic databases are important for TCM knowledge discovery. For example, the TCM herb related databases (e.g. herb database, formula database and herb ingredient database) have demonstrated great potential for chemical drug discovery [99]. The sizes of the basic structured databases (e.g. the herb ingredient database, and CTPM) that are manually curated from the published literature are increasing rapidly, at the same time as new data is being generated by scientific research [100]. Other than the free-text TCM literature, the structured basic databases provide well-formed data sources for traditional data mining applications since most of the data fields (e.g. herb name, herb alias and herb taxonomy) are strictly structured. However, there still exist free-text or semi-structured data fields like the herb constituents of formula, herb or formula efficacy, clinical studies of formula, and the pharmacological effects of herbs, which need IE or NER to extract the TCM named entities and relationships.

4.2. TCM text mining related research

Compared with the intensive research work and the immense publicly accessible bibliographic literature data of modern biomedical science, the development of TM for TCM is still in the early stages. Interesting TM work [101–109] in TCM to date has been focused on literature-based discovery and information extraction. In the following, we introduce some representative research in these areas.

4.2.1. Integrative mining of TCM literature and MEDLINE for functional gene networks

One of the pioneer studies of TCM text mining was presented by Wu et al. [101], in which a TM approach was developed for the identification of functional relationships among genes cited in MEDLINE papers based on TCM knowledge, syndrome and disease association. The TCM literature, which was from TCMLARS, was processed by a bootstrapping method to extract the syndrome–disease associations. In addition, the term co-occurrence was used for the identification of disease–gene associations from MEDLINE. As a result, the relationships between syndromes and genes were identified in common relevant diseases. The underlying hypothesis was that the genes related to the same syndrome would have a certain degree of biological interaction. The kidney-yang deficiency syndrome (KYD syndrome, see Glossary) and the related genes were specifically investigated. The study was able to identify one of the related genes of the KYD syndrome: CRF (C1q-related factor), which was previously found in the experimental study by Shen [110].

This study had been enhanced by a TM system, called MeDisco/3S. The MeDisco/3S is an integrative data mining system which aims to uncover the functional relationships among genes from MEDLINE and TCM bibliographic literature [101–103]. Based on the TCM literature (about 500,000 records), a complex literature-based gene network was developed, for which syndrome was used to automatically associate related genes. The syndrome–gene relationships discovered are based on (i) the syndrome–disease associations extracted from the TCM literature and (ii) the disease–gene associations extracted from MEDLINE. By means of bubble bootstrapping, the MeDisco/3S system extracted about 200,000 syndrome–gene relationships to generate the syndrome-based gene networks. The syndrome-based gene networks enable the functional annotation of genes to be analyzed from a syndrome perspective. By investigating the gene network of the KYD syndrome and the functions of the relevant genes, such as CRH (corticotropin releasing hormone), PTH (parathyroid hormone), PRL (prolactin), BRCA1 (breast cancer 1, early onset) and BRCA2 (breast cancer 2, early onset), it was demonstrated that genes related to the same syndrome have a degree of biological functional relationship, these are then clustered as a functional network module. Jensen et al. [111] and Wilkinson and Huberman [111] constructed similar literature-based gene networks although only the gene co-occurrences were considered in their work.

4.2.2. Herb prescription knowledge modeling and acquisition from text

Cao et al. [106] developed an ontology-based system for extracting knowledge about the TCM herbs and formulae from semi-structured text. They developed the herb and formula ontologies from 7 knowledge sources, including textbooks, codices, encyclopedias and dictionaries. The two ontologies consist of a set of classes and their relationships, and formal axioms for constraining the interpretation of those classes and relationships. Based on the defined ontologies and the canonical description of herb and formula texts, an executable knowledge extraction language (EKEL) was developed which assists in the extraction of knowledge from the herb and formula texts. The system has been tested on several herb and formula textual sources. A knowledge-base of more than 2710 herbs and 5900 formulae was constructed. The other work regarding the automatic extraction of formula knowledge from the TCM bibliographic literature is the MeDisco/
3T system [104]. The MeDisco/3T system iteratively extracts new TCM names and patterns using a small initial set of formula names to act as seeds. The MeDisco/3T system was able to extract correctly over 95% of the formula names. Based on the extracted formula names, heuristic rules were used to extract the constituent herb information from the semi-structured abstracts of the literature. With more than 18,000 formulae extracted, the final step was to discover interesting herb pairs and herb family combinations by means of an association rule mining algorithm, i.e. the Apriori algorithm [112].

4.2.3. The gene network analysis of the Cold and Hot syndromes in the context of the neuro-endocrine-immune network by literature mining

As discussed in the previous sections, syndrome is the basic element and the key concept in TCM theory. Syndrome could be considered as the abstraction and classification of disease, based on patient manifestations (e.g. symptom and sign) and TCM theories. To find the molecular-level associations of syndrome is one of the significant tasks of modern TCM studies. Li et al. [105] report their work on the Cold and Hot syndromes in the context of the neuro-endocrine-immune (NEI) network. In their study, a gene network is constructed with the assistance of the TCM disease database and MEDLINE query. It was found that hormones are predominant in the Cold syndrome network, while immune factors are predominant in the Hot syndrome network, and these two networks are connected by the neuro-transmitters. Furthermore, the herbal-treatment experiments on the rat model of collagen-induced arthritis revealed that the corresponding herbal treatments affect the hub nodes of the Cold and the Hot syndrome gene networks. This illustrates the feasibility of gaining a better understanding of syndrome based on the NEI network.

4.2.4. TCMGeneDIT: a database for associated herbal medicine, gene and disease information using text mining

Fang et al. [107] present a database, TCMGeneDIT, providing associations between Chinese herbal medicines, genes, diseases, effects and ingredients, and the relations between herb effects and effecters from a vast amount of biomedical literature (i.e. PubMed). The protein–protein interactions and biological pathways from the public databases (e.g. HPRD, KEGG) were also used to explore the action of genes associated with the curative effects of Chinese herbal medicine. The names of Chinese herbal medicines, genes, MeSH disease, Chinese herbal medicine ingredients and effects were used to annotate the literature corpus. The annotated literature corpus was then used to find various associations including the associations: (herb, gene), (herb, disease), (herb, gene, disease), (herb, ingredient), (herb, effect) and (gene, ingredient). Also, a rule-based information extraction method was used to extract the relation between Chinese herbal medicine effects and effecters from the literature by using part-of-speech tagging and noun-phrase chunk identification. Thereafter, the discovery of association was based on co-occurrences of terms and t-statistics testing. Transitive associations were inferred according to Swanson’s closed discovery model [7]. A web-based searching system has been developed to enable users to search for related associations and networks (http://tcmlife science.ntu.edu.tw/). TCMGeneDIT provides a tool for understanding the roles of herb components in producing prescribed effects, and the understanding of therapeutical mechanisms involving Chinese herbal medicine and gene interactions.

4.3. The future development of TCM text mining

The previous work indicates a promising future for TCM TM. However, it is obvious that TM of TCM is still in the early stages. Substantial TM methods need to be developed for the NER of diseases, symptoms, herbs and therapeutic terms, and for the discovery of relationships among the TCM named entities. Furthermore, in the future, it is necessary to develop a systematic evaluation strategy for the extraction of the diamonds out of the large-scale TCM data. As shown in Fig. 3, the clinical literature data, including contemporary clinical literature published in journals and conferences, the ancient literature recorded in the form of historical clinical cases and theoretical comments, and free-text clinical data, are the main TCM information sources for TM research. Medical ontologies are the prerequisites for advanced TM applications [113]. Hence, future TM applications in TCM should integrate the terminology systems (e.g. UTCMLS and TCM clinical terminology) in order to make semantic-rich and high quality discoveries.

One of the main objectives of TM in TCM is to help generate scientific hypotheses and clinical guidelines for practical diagnoses and treatments. To achieve this aim, it is essential to extract the clinical facts and events from the data. There are two important kinds of TCM knowledge which should be extracted by TM methods: the relationships of the TCM named entities (e.g. syndrome–symptom relationship, disease–syndrome relationship and herb–symptom relationship), and the constituent herb information of formula in TCM.

There is an urgent need for persistent and informed data processing tasks to extract from the TCM literature both the specific TCM named entities (e.g. herb, formula, syndrome, symptom and disease) and their relationships among these entities. To reduce the manual labor, appropriate IE and NER methods are needed to automatically extract the structured data, and to assist in the data curation tasks. One difference between TM in modern biomedicine and TM in TCM is the additional and indispensable preprocessing step for NER. Chinese word segmentation (see Glossary) [114] is needed to automatically segment sentences into words, since the Chinese language has no single-word boundaries. In addition, to improve the quality and efficiency of the TM process, IR and TC would be indispensable in order to facilitate the data searching and filtering of the TCM literature. As the insights and hypotheses
are most likely to be found by integrating multiple TCM data sources, the development of integrative data mining methods would be a promising step for TCM TM.

Actually, evaluation of the TM methods and systems is a complicated but significant task necessary to get practical results [16,115,116]. The annotated biomedical corpora, such as BioCreative [115], GENIA [117] and CLEF [118], are useful resources for promoting the development of TM methods [119]. There is, as yet, no study on this issue in TM of TCM. Constructing the annotated TCM corpora (e.g. annotated ancient TCM literature) would be significant for benchmarking the performance and usefulness of the TM methods.

5. Discussion

TCM as a distinct medical discipline has many information and data sources, including data being generated in practical clinical processes and research activities. Different to modern biomedical science, TCM does not involve general experimental practice in the laboratory. TCM clinical practice is a kind of clinical experiment, in which novel prescriptions for individualized patients are tested and evaluated. Manual induction of empirical knowledge from the daily clinical practices is one of the approaches available for the distillation of TCM knowledge. It is important to develop a new TCM clinical research framework that focuses on the acquisition, management and analysis of TCM clinical data [120]. TM is a feasible solution for the extraction of structured data and the discovery of regularities from free-text TCM clinical data and literature, and it will help TCM practitioners to utilize efficiently data collected from clinical practice and to promote the development of TCM from the collected experience of individuals into evidence-based medicine (EBM) [121].

Different kinds of clinical data, such as EMR in TCM hospitals and clinical cases recorded in the ancient textbooks, are the main TCM knowledge sources for the generation of appropriate clinical hypotheses. Extraction of the clinical facts and events from that clinical data is therefore significant for TM in TCM. The TCM literature and free-text clinical data (mainly TCM EMR data) constitute the core data sources for TM. The EMR in TCM records the detailed clinical events (e.g. manifestations, diagnoses, prescriptions and curative effects) of every medical case, while the contemporary bibliographic literature data contains a summary of the clinical facts. Therefore, the EMR data and the bibliographic literature data form the two important complementary data sources, and are valuable for integrative TM applications in TCM. Besides the theoretical literature, such as ‘The Inner Canon of Emperor Huang’ and ‘Treatise on Cold Pathogenic and Miscellaneous Disease’, most of the ancient literature discusses clinical cases or knowledge about clinical prescriptions. The information in ancient literature has helped TCM scientists to develop new ideas for diagnosis and treatment and has continuously enriched TCM knowledge. One example is the compound called artemisinin from the sweet wormwood herb which was discovered by Chinese scientists in the 1970s [122]. This success actually originated from the traditional texts ‘Handbook of Prescriptions for Emergency’, which records that the sweet wormwood herb could treat malaria. Facilitating the search for knowledge embedded in the ancient literature by TM methods promises to produce exciting research.

Furthermore, as two particular complementary knowledge sources, TCM information sources (e.g. TCM clinical data) and modern biomedicine data (e.g. MEDLINE) could be further integrated to promote the TM-based systems biology research [123]. Because TCM mainly studies macro-level phenomena and the functional state of the human system, and modern biomedicine focuses on micro-level knowledge and the structural substance of the human body, the integrative analysis of these data sources would provide a unique knowledge source. By using TM methods, it is possible to integrate the macro-level clinical data obtained in TCM clinical practices and the micro-level experimental data obtained in modern biomedical science. This will contribute to the connection of the functional systems and the structural systems of the human body, which will provide scientists with significant insights to make breakthroughs in medical and life sciences. For this work, the existence of TCM terminologies in both Chinese and English (e.g. herb names, symptom names, syndrome names and disease names) would be indispensable because of the bilingual data involved. Although there is some work on the international standardization of TCM terminology [38], the translation of TCM terms from Chinese to English is still a challenge. The inconsistent translation of the TCM terms would become an obstacle for integrated TM applications.

NER is one of the key steps for TM in TCM. Various terminologies which originated in the ancient TCM literature have become a major obstacle for this task. Furthermore, the segmentation of Chinese words is another challenge that needs be addressed before the NER is performed. The ancient TCM literature which is written using ancient Chinese words and sentences poses a real challenge for NER tasks due to the very different syntax and phrasing of ancient Chinese. To enhance the quality and efficiency of TM in TCM, it is necessary to integrate the information about the terminology and the structured basic databases to standardize the terminology concepts and terms.

Due to the size of the TCM bibliographic literature database and concerns for the quality of the contemporary TCM published literature, an initial and careful evaluation of the TCM literature is an important step. Besides manual analysis, bibliometric methods [124,125], such as citation analysis and content analysis, are appropriate to evaluate the quality of TCM literature. Although the current clinical trials in TCM are still low in quality from the EBM perspective [126], the clinical data and events from the observational studies and the case reports are actually of good quality. Thus, it is important for TM in TCM, first to identify and extract the factual data, such as the symptoms, syndromes and herb prescriptions. With these symptoms, syndromes and herb prescriptions consistently defined, the large-scale factual data extracted from the literature through data mining and statistical methods would become the most valuable knowledge source for the generation of high quality clinical evidence.

6. Conclusion

In this paper, we introduce the basic theories of TCM and discuss the differences between TCM and modern biomedicine from the perspectives of methodology and general approach. TCM uses medical theories originating in the ancient philosophy of China but the research mode of TCM is clinically-based. Most of the misunderstandings around TCM will probably be clarified through the analysis of its extensive clinical data and its published literature. The lack of common operational procedures (e.g. the clinical guidelines) for TCM physicians to deal with real-world clinical cases is the main issue that undermines the effectiveness of TCM clinical work. The application of TM methods promises to provide significant help for the generation of common operational clinical procedures.

We provide a description of the content and structure of the information sources that are relevant to TM in TCM. This is intended to assist researchers to use the data in an efficient way. The TCM literature, including the bibliographic literature and the annotated ancient literature, is the general information source for TM studies. The clinical data, represented in free-text, are also
important data sources. Both of them require intensive research to extract the TCM named entities and structured information.

The development of TM techniques provides an opportunity to reveal buried data. Clinical data and the related publications are the core knowledge sources for TCM. The free-text and natural language representation (particularly in ancient Chinese) of these data have been the main obstacles for large-scale data analysis. The exposure of this buried data will make it possible to gain important understanding about disease and the human life system at the holistic level.

Acknowledgments
This work is partially supported by the S&T Foundation of Beijing Jiaotong University (2007RC02), Program of Beijing Municipal S&T Commission, China (D08050703020804), National Key Technology R&D Program (2007BA110806) and China 973 Project (2006CB504601). Dr. John Baruch made English proof-reading for this paper, for which we would like to express our sincere thanks to him.

References
[32] WHO Regional Office for the Western Pacific. WHO international standard terminologies on traditional medicine in the western pacific region; 2007.


Bian zheng lun zhi: The main diagnosis and treatment principle of TCM. TCM diagnosis is performed, based on the overall observation of human symptoms, to differentiate the syndromes of patients. Appropriate treatments like formula, acupuncture are prescribed according to the syndromes of the patients.

Chinese medical formula: Also called fufang in Chinese, which is a kind of TCM therapy with herbs as ingredients. The organization of the herbs is based on the holistic philosophy of TCM and keeps to the rules of drug synergism and compatibility. We simply use ‘formula’ to represent it in the article.

Chinese word segmentation: A process of dividing a string of written language in Chinese into its component words since Chinese sentences have no single-word boundaries. It is often a non-trivial task.

Information extraction (IE): A technique to extract structured information automatically (e.g. named entities, facts and events) from unstructured documents. The IE process often involves natural language processing and machine learning methods to deal with the large amounts of free-text data.

Information retrieval: The science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web.

Kidney-yang deficiency syndrome: One of the syndromes of TCM, which is an important syndrome involving diseases, such as caducity, neural disease and immunity, and has the related symptoms, including sore lower back, copious pale urine, poor appetite and infertility.

Moxibustion: A therapeutic procedure involving burning materials (usually moxa) to apply heat to certain points or areas of the body surface for curing disease through regulation of the function of meridians/ channels and visceral organs.

Named entity recognition: One of the subtasks of IE that seeks to automatically extract the domain-specific named entities, such as drugs, products, locations, genes and diseases, from the unstructured documents.

Syndrome: Also called pattern in TCM. It is the main TCM diagnosis result, which has a summary and theoretical analysis of the manifestations (e.g. symptoms and signs) of patients. There are several hundred common syndromes in TCM.

Syndrome differentiation: The unique diagnostic method in TCM. By comprehensive analysis the manifestations of patients, syndrome differentiation is to classify the patterns of maladjustment in the body through determining the nature, location, mechanism and tendency of the maladjustment. The eight principles of syndrome differentiation are the fundamental methods of TCM and include yin/yang, cold/heat, deficiency/excess and exterior/interior.

TCM hospital: A hospital taking TCM as its main clinical approach. It also integrating western medical approaches in the clinical practices.

Yin and yang (in traditional Chinese medicine): The general descriptive terms for the two opposite, complementary and inter-related cosmic forces found in all matter in nature. The ceaseless motion of both yin and yang gives rise to all changes seen in the world.