Journal of Biomedical Informatics 43 (2010) 650–660

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

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

ARTICLE INFO

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

Keywords:
Text mining
Traditional Chinese medicine
Review

ABSTRACT

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.

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].

* Corresponding author. Fax: +86 10 51840526.
E-mail addresses: xzzhou@bjtu.edu.cn, lzxzhou@gmail.com (X. Zhou), y.h.peng@bradford.ac.uk (Y. Peng), liuby@mail.cintcm.ac.cn (B. Liu).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.01.002

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 TM information sources, the related TCM TM 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 2009. 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, Shennong'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 inter-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 interdependent (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 analysis 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

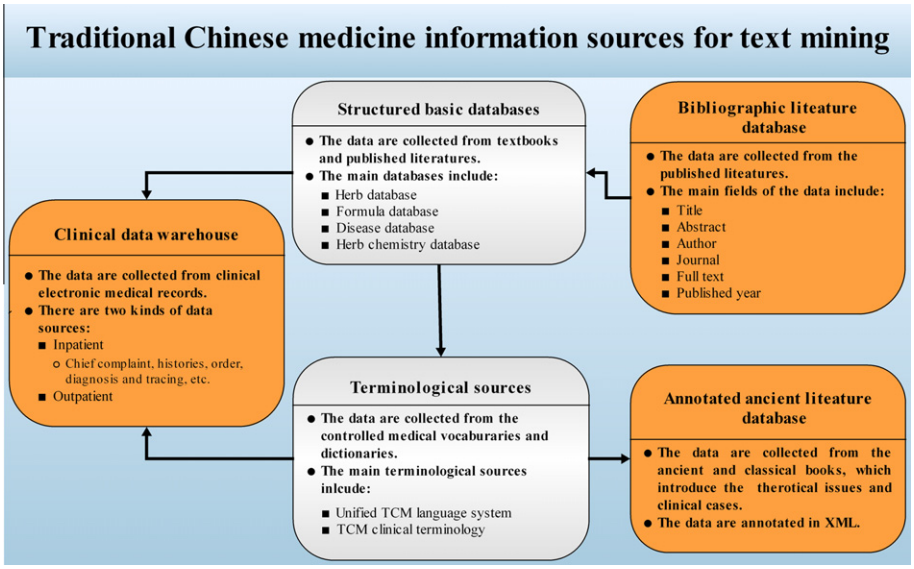

Fig. 1. The general view of the related information sources for TCM text mining. The direction of lines represents the supporting relationships between two information sources, for example, the terminology sources are used for annotating the ancient literature database. The TCM information sources consist of five main components: (1) bibliographic literature database, (2) ancient TCM literature database, (3) clinical data warehouse, (4) terminology databases and (5) structured basic databases.

relationships, syndrome–formula relationships and symptom–herb relationships) from a broad spectrum of heterogeneous TCM databases. A few review papers have introduced the TCM information sources for TCM data mining [70], the computational methods [59] and the domain databases [71]. Fig. 1 shows the main information sources that can be potentially useful to TM applications in the TCM domain.

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 database[1], 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 za 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

[1] http://www.cintcm.com/e_cintcm/version.htm.

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 TM the 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 (UTCMLS), has been developed for this purpose [87]. The UTCMLS 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*

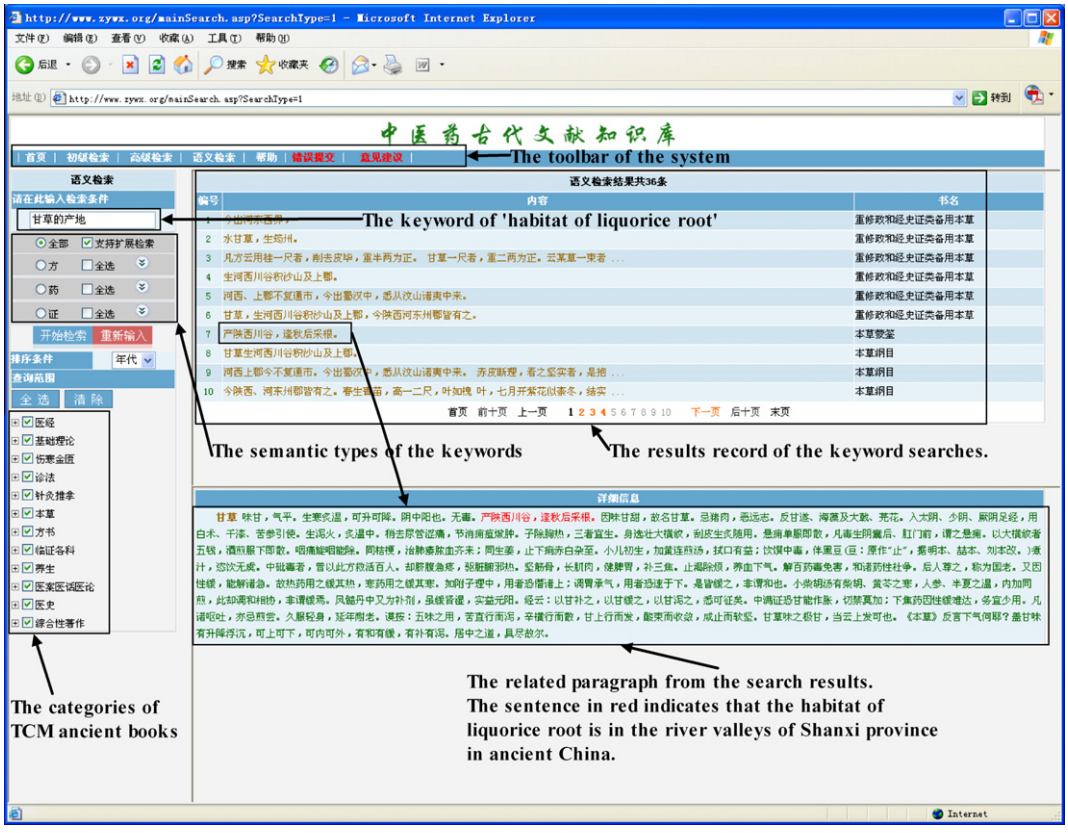

Fig. 2. Query results for the keyword '*habitat of liquorice root*'. The lower right part of the window displays a paragraph that discusses the nature and flavor, habitat, harvesting application of *liquorice root*. The sentence in red describes the *habitat of liquorice root*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Subject Headings (TCM MeSH) [88] and *Chinese Library Classification (4th ed.)* [89], have been used as sources for the UTCMLS ontology development. The UTCMLS has the main top-level concept related categories, semantic types, semantic relationships, and 14 essential sub-ontologies (e.g. basic TCM theories, formula, Chinese herbal medicine, acupuncture, prevention and geography) to build up the skeleton of the TCM terminological framework, which can be browsed online [90]. The UTCML system has become a significant terminological information source for TCM literature-searching and for supporting heterogeneous database integration [91]. Recently, several applications, such as the e-science platform [92], have been built on the basis of the UTCML terminology system. By February 2009, the UTCMLS system contained 122,675 concepts and 302,375 terms.

Other clinical terminological systems, such as ICD-10 [93], SNOMED-CT [94], clinical terminology for TCM diagnosis and treatment [95], are also used in TCM clinical practice. These terminology systems focus on the standardization of modern biomedical named entities, such as disease, symptom and sign. To support structured data entry for daily TCM clinical operations, a TCM clinical terminology (this is not available online) [96] has been developed with a terminology classification framework corresponding to the content of TCM EMRs. This terminology system has gathered over 160,000 clinical terms. Both UTCMLS and TCM clinical terminology provide significant terminology support for the ontology-based IE in TCM clinical texts and bibliographic literature.

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. Jenssen et al. [11] 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://tcm.lifescience.ntu.edu.tw/). TCMGeneDIT provides a tool for understanding the roles of herb components in producing prescribed effects, and the understanding of therapeutic 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

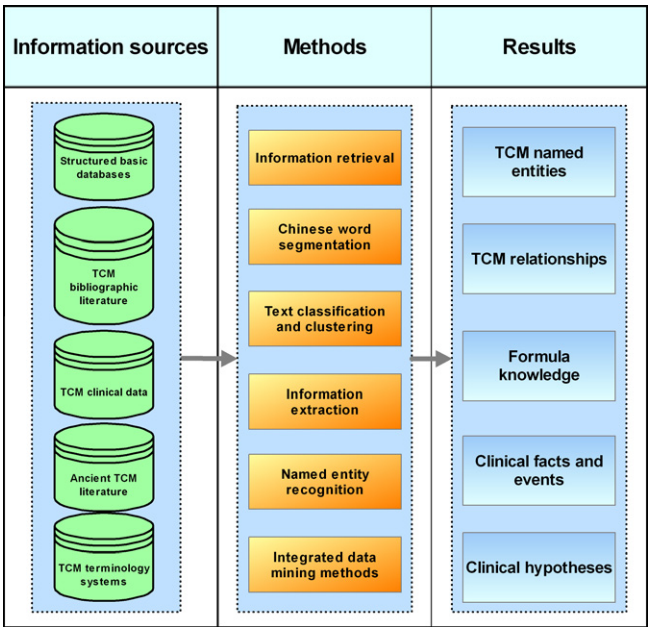

Fig. 3. Our view of the TCM TM framework. By taking the TCM information sources, such as the TCM bibliographic literature, TCM clinical data and the ancient TCM literature as the main data sources, the TM tasks include the recognition of TCM named entities and the relationships of those entities, the extraction of the constituent herb information (e.g. herb name and herb dosage) in formula, and the discovery of novel clinical facts and events.

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 (2007RC072), Program of Beijing Municipal S&T Commission, China (D08050703020804), National Key Technology R&D Program (2007BA110B06) 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

[1] Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press; 2006.
[2] Hearst MA. Untangling text data mining. Proc Assoc Comput Linguist 1999;37:3–10.
[3] Wikipedia (information retrieval). Available from: http://en.wikipedia.org/wiki/Information_retrieval. Accessed: 20 October 2009.
[4] Wikipedia (information extraction). Available from: http://en.wikipedia.org/wiki/Information_extraction. Accessed: 20 October 2009.
[5] Berry MW, Castellanos M. Survey of text mining II: clustering, classification, and retrieval. 1st ed. Springer; 2007.
[6] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30:7–18.
[7] Swanson DR. Complementary structures in disjoint science literatures. In: Proceedings of the 14th ACM SIGIR. New York: ACM Press; 1991. p. 280–9.
[8] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 1997;91(2):183–203.
[9] Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Brief Bioinform 2005;6(3):277–86.
[10] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005;74(2–4):289–98.
[11] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006;7:119–29.
[12] Yetisgen-Yildiza M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 2006;39(6):600–11.
[13] de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inform 2002;67(1–3):7–18.
[14] Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol 2003;10(6):821–55.
[15] Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005;6(1):57–71.
[16] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform 2007;8(5):358–75.
[17] Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. Bioinformatics 2002;18(12):1553–61.
[18] Yandell MD, Majoros WH. Genomics and natural language processing. Nat Rev Genet 2002;3(8):601–10.
[19] Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. Genome Biol 2005;6:224.
[20] Ganiz M, Pottenger WM, Janneck CD. Recent advances in literature based discovery. JASIST; 2006. Available from: http://www.dimacs.rutgers.edu/~billp/pubs/JASISTLBD.pdf. Accessed: 20 October 2009.
[21] Cakmak A, Ozsoyoglu G. Discovering gene annotations in biomedical text databases. BMC Bioinform 2008;9:143.
[22] Tsuruoka Y, McNaught J, Ananiadou S. Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinform 2008;9(Suppl. 3):S2.
[23] Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinform 2008;9(Suppl. 3):S3.
[24] Jenssen T-K, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001;28:21–8.
[25] Zhou D, He Y. Extracting interactions between proteins from the literature. J Biomed Inform 2008;41(2):393–407.
[26] Seki K, Mostafa J. Discovery implicit associations between genes and hereditary diseases. Pac Symp Biocomput 2007:316–27.
[27] Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinform 2008;9:207.
[28] Rinaldi F, Schneider G, Kaljurand K, Hess M, Andronis C, Konstandi O, et al. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. Artif Intell Med 2007;39(2):127–36.
[29] van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet 2006;14(5):535–42.
[30] Scherf M, Epple A, Werner T. The next generation of literature analysis: integration of genomic analysis into text mining. Brief Bioinform 2005;6(3):287–97.
[31] Hoeglund A, Blum T, Brady S, Donnes P, Miguel JS, Rocheford M, et al. Significantly improved prediction of subcellular localization by integrating text and protein sequence data. Pac Symp Biocomput 2006:16–27.
[32] Gabow AP, Leach SM, Baumgartner WA, Hunter LE, Goldberg DS. Improving protein function prediction methods with integrated literature data. BMC Bioinform 2008;9:198.
[33] Aragues R, Sander C, Oliva B. Predicting cancer involvement of genes from heterogeneous data. BMC Bioinform 2008;9:172.
[34] Dudley J, Butte AJ. Enabling integrative genomic analysis of high impact human diseases through text mining. Pac Symp Biocomput 2008;13:580–91.
[35] Li S, Wu LJ, Zhang ZQ. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics 2006;22:2143–50.
[36] Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol 2008;4:189.
[37] Anonymous. The inner canon of emperor Huang. Beijing: Chinese Medical Ancient Books Publishing House; 2003.
[38] WHO Regional Office for the Western Pacific. WHO international standard terminologies on traditional medicine in the western pacific region; 2007.
[39] Robinson N. Integrated traditional Chinese medicine. Complement Ther Clin Pract 2006;12:132–40.
[40] Stone. Chen Zhu interview: China's modern medical minister. Science 2008;1748a–9a.
[41] Tang J-L, Liu B, Ma K-W. Traditional Chinese medicine. Lancet 2008;372(9654):1938–40.
[42] The 2007 National Social Statistical Data of China. Available from: http://www.stats.gov.cn/tjsj/qtsj/shtjnj/2007/. Accessed: 20 October 2009.
[43] Harmsworth K, Lewith GT. Attitudes to traditional Chinese medicine amongst western trained doctors in the People's Republic of China. Soc Sci Med 2001;52(1):149–53.
[44] Barnes PM, Powell-Griner E, McFann K, Nahin RL. Complementary and alternative medicine use among adults: United States, 2002. Adv Data 2004;343:1–19.
[45] National Center for Complementary and Alternative Medicine. The use of complementary and alternative medicine in the United States; 2008.
[46] Flaws B, Sionneau P. The treatment of modern western medical diseases with Chinese medicine: a textbook and clinical manual. 2nd ed. Blue Poppy Press; 2005.
[47] Konkimalla VB, Efferth T. Evidence-based Chinese medicine for cancer therapy. J Ethnopharmacol 2008;116:207–10.
[48] Rao JK, Mihaliak K, Kroenke K, Bradley J, Tierney WM, Weinberger M. Use of complementary therapies for arthritis among patients of rheumatologists. Ann Intern Med 1999;131:409–16.
[49] Wang Z-Y, Chen Z. Acute promyelocytic leukemia: from highly fatal to highly curable. Blood 2008;111:2505–15.
[50] Wang L, Zhou G-B, Liu P, Song J-H, Liang Y, Yan X-J, et al. Dissection of mechanisms of Chinese medicinal formula Realgar-Indigo naturalis as an effective treatment for promyelocytic leukemia. PNAS 2008;105:4826–31.
[51] Diener HC, Kronfeld K, Boewing G, Molsberger A, Tegenthoff M, Trampisch HJ, et al. Efficacy of acupuncture for the prophylaxis of migraine: a multicentre randomised controlled clinical trial. Lancet Neurol 2006;5:310–6.
[52] Bensoussan A, Talley NJ, Hing M, Menzies R, Guo A, Ngu M. Treatment of irritable bowel syndrome with Chinese herbal medicine: a randomized controlled trial. JAMA 1998;280:1585–9.
[53] Tsang IK. Establishing the efficacy of traditional Chinese medicine. Nat Clin Pract Rheumatol 2007;3:60–1.
[54] Xie W, Gao X. The idea and approach of the mutual complementarity and integration of TCM and western medicine. J Chendu College TCM 1994;17(2):14–7 [in Chinese].
[55] Chen K, Lu A, Chen S, Wei B, Lu W, Mu D, et al. Survey on the developing status of integrative Chinese and western medicine. Zhongguo zhong xi yi jie he za zhi 2006;26(6):485–8 [in Chinese].
[56] Tu B, Johnston MF, Hui KK. Elderly patient refractory to multiple pain medications successfully treated with integrative East–West medicine. Int J Gen Med 2008(1):3–6.
[57] Liu B, Hu J, Xie Y, Wen W, Wang R, Zhang Y, et al. Effects of integrative Chinese and western medicine on arterial saturation in patients with Severe Acute Respiratory Syndrome. Chin J Integr Med 2004;10(2):117–22.
[58] World Health Organization. SARS: Clinical trials on treatment using a combination of Traditional Chinese medicine and Western medicine; 2004.

[59] Lukman S, He Y, Hui S-C. Computational methods for traditional Chinese medicine: a survey. Comput Methods Progr Biomed 2007;88:283–94.
[60] Tang Y, Geng E, Dang Y. Acupuncture and moxibustion. Academy Press; 1999.
[61] Liu Z (Chief Editor). Basic theories of traditional Chinese medicine. China: Higher Education Press; 2007 [written in both Chinese and English].
[62] Nanjing University of TCM (Compiled). Basic theory of traditional Chinese medicine (newly compiled practical English–Chinese library). Shanghai University of TCM Press; 2002.
[63] Bailey EJ. African American alternative medicine: using alternative medicine to prevent and control chronic diseases. 1st ed. Praeger Publishers; 2002.
[64] Xu X (Chief Editor). The English–Chinese encyclopedia of practical traditional Chinese medicine. Higher Education Press; 1991.
[65] Maciocia G. The foundations of Chinese medicine: a comprehensive text for acupuncturists and herbalists. Churchill Livingstone; 1997.
[66] Wikipedia (five phases theory). Available from: http://en.wikipedia.org/wiki/Wu_Xing. Accessed: 20 October 2009.
[67] Liu B, Wang Y. The study on concept of syndrome, symptom and the relationships. J Trad Chin Med 2007;48(4):293–6, 298 [in Chinese].
[68] Lean MEJ, Mann JI, Heok JA, Elliot RM, Schofield G. Translational research: from evidence-based medicine to sustainable solutions for public health problems. BMJ 2008;337:a863.
[69] Chen K, Song J. Clinical study by way of combining diseases with differentiation of their syndromes, an important mode in study on combination of Chinese traditional and western medicine. World Sci Technol Modern TCM Mater Med 2006;8(2):1–5 [in Chinese].
[70] 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(3):219–36.
[71] Cui M. The state-of-the-art survey of traditional Chinese medicine domain databases building. Chin J Inform TCM 2004;11(3):189–92 [in Chinese].
[72] Fan W, Tong Y, Pan Y, Shang W, Shen J, Li W, et al. Traditional Chinese medical journals currently published in mainland China. J Altern Complement Med 2008;14(5):595–609.
[73] The TCM online database searching website. Available from: http://cowork.cintcm.com/engine/windex.jsp. Accessed: 20 October 2009.
[74] Liu Y, Sun Y. China traditional Chinese medicine (TCM) patent database. World Patent Inform 2004;26(1):91–6.
[75] China TCM Patent Database. Available from: http://218.240.13.195/tcm_patent/englishversion/login/index.asp. Accessed: 20 October 2009.
[76] Liu C. A knowledge representation method of ancient Chinese medical literature based on the concept of knowledge unit. In: Proceedings of the 3th International Congress on Traditional Medicine; 2004. p. 368–9 [in Chinese].
[77] Yang J. The research on the method of ontology-based construction of the thesaurus of traditional Chinese medicine classics. Ph.D. thesis. China Academy of Chinese Medical Sciences; 2008.
[78] Ancient TCM literature search website. Available from: http://bencao.cintcm.ac.cn. Accessed: 20 October 2009.
[79] Liu B, Zhang H, Ni W. The speciality of TCM electronic medical record system. Med Inform 2004;17(1):9–11 [in Chinese].
[80] Yu X, Ling C. The research status and prospect of TCM electronic medical record. Hosp Admin J Chin PLA 2006;13(7):612–4 [in Chinese].
[81] Zhou X, Liu B, Wang Y, Zhang R, Li P, Chen S, et al. Building clinical Data warehouse for traditional Chinese medicine knowledge discovery. BMEI 2008;1:615–20.
[82] Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med 2009. doi:10.1016/j.artmed.2009.07.012.
[83] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
[84] Oracle data minder website. Available from: http://www.oracle.com/technology/products/bi/odm/odminer.html. Accessed: 20 October 2009.
[85] BusinessObjects website. Available from: http://www.sap.com/solutions/sapbusinessobjects/index.epx. Accessed: 20 October 2009.
[86] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 2008;15(1):87–98.
[87] Zhou X, Wu Z, Yin A, Wu L, Fan W, Zhang R. Ontology development for unified traditional Chinese medical language system. Artif Intell Med 2004;32(1):15–27.
[88] Wu LC (Chief Editor). Chinese traditional medicine and materia medical subject headings. Beijing: Chinese Medical Ancient Books Publishing; 1996.
[89] Editing Committee of Chinese Library Classification. Chinese library classification. 4th ed. Beijing: Beijing Library Press; 1999.
[90] The UTCMLS semantic classes browsing website. Available from: http://search.cintcm.com/sw/. Accessed: 20 October 2009.
[91] Mao Y, Wu Z, Chen H, Xu Z. Context-based web ontology service for TCM information sharing. In: Proceedings of IEEE international conference on web services (ICWS'05); 2005. p. 699–705.
[92] Chen H, Mao Y, Zheng X, Cui M, Feng Y, Deng S, et al. Towards semantic e-Science for traditional Chinese medicine. BMC Bioinform 2007;8(Suppl. 3):S6.
[93] World Health Organization. ICD-10: International statistical classification of diseases and related health problems (10th Revision, Version for 2007). Available from: http://www.who.int/classifications/apps/icd/icd10online/. Accessed: 20 October 2009.
[94] SNOMED-CT website. Available from: http://www.ihtsdo.org/snomed-ct/. Accessed: 20 October 2009.
[95] General administration of technology supervision of the People's Republic of China. National standard: clinic terminology of traditional Chinese medical diagnosis and treatment-(diseases, syndromes and therapeutic methods), GB/T 16751.1∼3-1997, Beijing, 4 March 1997.
[96] Guo Y, Liu B, Li P, Zhou X. Ontology and standardization of the traditional Chinese medicine terms. Chin Arch TCM 2007;7(25):1368–70 [in Chinese].
[97] TCM online website. Available from: http://www.cintcm.com. Accessed: 20 October 2009.
[98] Tradimed website. Available from: http://www.tradimed.com. Accessed: 20 October 2009.
[99] Chen X, Zhou H, Liu YB, Fang JF, Li H, Ung CY, et al. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. Br J Pharmacol 2006;149:1092–103.
[100] Xu M, Dominic TS, Xie J, Dai W, Chen L, Deng B, et al. A database on treating drug addiction with traditional Chinese medicine. Addiction 2007;102(2):282–8.
[101] Wu Z, Zhou X, Liu B, Chen J. Text mining for finding functional community of related genes using TCM knowledge. In: Proceedings of the 8th PKDD, LNAI 3202. Berlin: Springer-Verlag; 2004. p. 459–70.
[102] Peng Y, Zhang X. Integrative data mining in systems biology: from text to network mining. Artif Intell Med 2007;41(2):83–6.
[103] Zhou X, Liu B, Wu Z, Feng Y. Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. Artif Intell Med 2007;41:87–104.
[104] Zhou X, Liu B, Wu Z. Text mining for clinical Chinese herbal medical knowledge discovery. Lect Notes Comput Sci 2005;3735:396–8.
[105] Li S, Zhang ZQ, Wu LJ, Zhang XG, Li YD, Wang YY. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. IET Syst Biol 2007(1):51–60.
[106] Cao C, Wang H, Sui Y. Knowledge modeling and acquisition of traditional Chinese herbal drugs and formulae from text. Artif Intell Med 2004;32(1):3–13.
[107] Fang Y-C, Huang H-C, Chen H-H, Juan H-F. TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complement Altern Med 2008;8:58.
[108] Zhou X, Cui M, Wu Z. Traditional Chinese medical subject heading extraction based on text mining. Chin J Inform TCM 2003;10(1):71–4 [in Chinese].
[109] Li Y, Li S, Lu A. Comparative analysis via data mining on the clinical features of western medicine and Chinese medicine in diagnosing rheumatoid arthritis. Zhongguo zhong xi yi jie he za zhi 2006;26(11):988–91.
[110] Shen ZY. The continuation of kidney study. Shanghai: Shanghai S&T Publishers; 1990 [in Chinese].
[111] Wilkinson D, Huberman BA. A method for finding communities of related genes. Proc Natl Acad Sci 2004;101(Suppl. 1):5241–8.
[112] Agrawal R, Srikant R. Fast algorithms for mining association rules. VLDB 1994:487–99.
[113] Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform 2005;6(3):239–51.
[114] Gao JF, Li M, Wu A, Huang C-N. Chinese word segmentation and named entity recognition: a pragmatic approach. Comput Linguist 2005;31(4):531–74.
[115] Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, et al. Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. Genome Biol 2008;9(Suppl. 2):S1.
[116] Caporaso GJ, Deshpande N, Fink JL, Bourne PE, Cohen KB, Hunter L. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. Pac Symp Biocomp 2008;13:640–51.
[117] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus — semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl. 1):i180–2.
[118] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009;42:950–66.
[119] Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinform 2006;7:356.
[120] Liu B, Zhou X. Design and practice of wet-dry approach in clinical research of traditional Chinese medicine. World Sci Technol Modern TCM Mater Med 2007;9(1):85–9 [in Chinese].
[121] Liu B. Evidence-based medicine and the modernization of traditional Chinese medicine and pharmacology. Chin J Evidence Based Med 2001;1(1):3–4.
[122] Liu C, Wang Y, Ouyang F. Advances in artemisinin research. Prog Chem 1999;11(1):41–8 [in Chinese].
[123] Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol 2006;24(12):571–9.
[124] Nicholas D, Ritchie M. Literature and bibliometrics. London: Clive Bingley; 1978.
[125] Xie Q, Cui M, Pan YL. Thinking on TCM literature evaluation methods and techniques based on mass information. Zhongguo zhong xiyi jie he za zhi 2007;27(8):753–6.
[126] Li T-Q, Wang G, Wang L, Mao B. Clinical trials of traditional Chinese medicine in China: status and evaluation. Chin J Evidence Based Med 2005;5(6):431–7 [in Chinese].

Glossary

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.