Text Mining for Clinical Chinese Herbal Medical Knowledge Discovery

Xuezhong Zhou[1], Baoyan Liu[1], and Zhaohui Wu[2]

[1] China Academy of Traditional Chinese Medicine, Beijing, 100700, P.R. China
{zxz, liuby}@mail.cintcm.ac.cn
[2] College of Computer Science, Zhejiang Univeristy, Hangzhou, 310027, P.R. China
wzh@cs.zju.edu.cn

Abstract. Chinese herbal medicine has been an effective therapy for healthcare and disease treatment. Large amount of TCM literature data have been curated in the last ten years, most of which is about the TCM clinical researches with herbal medicine. This paper develops text mining system named MeDisco/3T to extract the clinical Chinese medical formula data from literature, and discover the combination knowledge of herbal medicine by frequent itemset analysis. Over 18,000 clinical Chinese medical formula are acquired, furthermore, significant frequent herbal medicine pairs and the family combination rule of herbal medicine have primary been studied.

1 Introduction

Recently, text mining has attracted great attention in the biomedical research community [1,2,3] due to the large amount of literature and TextBases (e.g. Medline) have been accumulated in the biomedical fields.

Traditional Chinese Medicine (TCM) has been a successful approach for Chinese health practice since several thousand years ago. It is significant to study the compositional rule of Chinese Herbal Medicine (CHM) since CHM has been a novel basis of new drug development. The TCM bibliographic database, which contains over one half million records from 900 biomedical journals published in China since 1984[1]. This paper aims to discover knowledge from TCM literature with regard to clinical CMF (Chinese Medical Formula) component CHM combination. We follow the approach suggested in [4] to extract the structured objective information and then apply the traditional data mining algorithms. We develop a text mining system called MeDisco/3T(**Me**dical **Disco**ver for **T**raditional **T**reatment in**T**elligence) to mine the CHM knowledge from TCM literature. Firstly, MeDisco/3T extracts structured CMF information (e.g. CMF name, CHM components and efficacy description) from literature based on bootstrapping method [5]. Secondly, it uses frequent itemset algorithm to analyze the data.

2 MeDisco/3T Text Mining System

Fig 1 depicts the framework of MeDisco/3T. There are three main steps to be processed in MeDisco/3T.

[1] http://www.cintcm.com

A. Hoffmann, H. Motoda, and T. Scheffer (Eds.): DS 2005, LNAI 3735, pp. 396 – 398, 2005.
© Springer-Verlag Berlin Heidelberg 2005

(1) Iterative extracts the CMF names from literature when provided with a handful of CMF seed name tuples.
(2) Extracts the CHM components and efficacy descriptions data according to the extracted CMF names. Some simple heuristic rules are used in this procedure since the abstracts are semi-structured, because most of them are delimited by special word labels such as "Approaches", "Objectives" and "Results" etc..
(3) Conducts various kinds of data mining algorithms based on the clinical CMF database, currently, we only perform the simple frequent itemset analysis.

It is clearly that MeDisco/3T will produce two important results, namely a database of novel clinical CMFs and support of classical data mining studies on CHM.

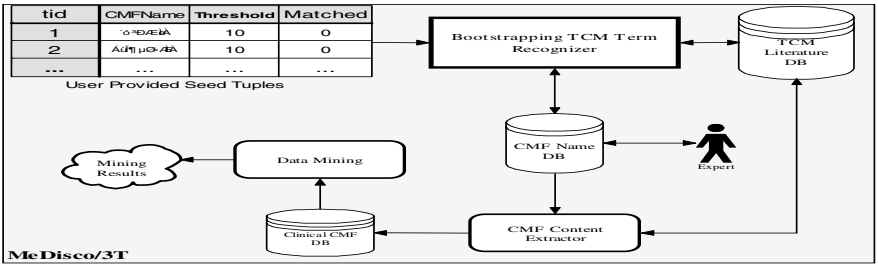

Fig. 1. MeDisco/3T text mining system

3 Main Results

We extract and identify 18,213 CMFs (from the year of 2000 to 2003 of TCM literature database) with different CHM composition to have a frequent itemset CHM analysis. The average name extraction precision by bootstrapping method[2] is high over 95%. MeDisco/3T performs a preprocessing procedure to transform the extracted data to a completely structured form, which is suitable for data mining algorithms. Where after, it apply the Apriori algorithm to analyze the frequent CHM pairs and CHM family combination characteristics in clinical CMF using.

All the clinical CMF used in TCM can be classified by its efficacy. Exactly, One CHM can be used for different efficacy in different CMF. We have chosen five different types of CMFs according to the efficacy such as HuoXueHuaYu 活血化瘀, BuZhongYiQi 补中益气 etc.. The 10 frequent CHM pairs and family combinations of the above two CMF types are listed in Table 1. Due to the page limit, the results of the other three CMF types are not depicted. It indicated from the experiment results that there exist many important CHM pairs and family combinations with different efficacy. For example, 黄芪 - 当归 is a typical CHM pair with BuZhongYiQi efficacy, and 伞形科 - 豆科 builds the core of CMF with BuZhongYiQi efficacy, because the

[2] Zhou, X., Text Mining and the Applications in TCM. PhD thesis, College of computer science, Zhejiang University, 2004,12,8. The thesis has a detail description of bootstrapping method used in MeDisco/3T.

support of 伞形科 - 豆科 combination in CMF with BuZhongYiQi efficacy is 180%. This knowledge will surely help to clinical CMF prescription practice and new drug development.

Table 1. The 10 top frequent CHM and its family combinations of efficacy BuzhongYiQi and HuoXueHuaYu, Supp(Family/CHM) represents the support of frequent family/CHM combination. For convenience, the CHM name is in Chinese, but all can be referred from the online databases on http://www.cintcm.com for the Latin or English names.

Family	BuZhongYiQi Supp(Family/CHM)	CHM	Family	HongXueHuaYu Supp(Family/CHM)	CHM
伞形科 - 豆科	1.8/0.3	当归 - 黄芪	伞形科 - 菊科	1.2/0.36	川穹 - 当归
桔梗科 - 豆科	1.1/0.26	甘草 - 党参	伞形科 - 唇形科	1.1/0.31	红花 - 桃仁
毛茛科 - 豆科	0.96/0.24	当归 - 白术	伞形科 - 豆科	1.03/0.30	桃仁 - 当归
菊科 - 豆科	0.96/0.23	党参 - 黄芪	伞形科 - 豆科	0.95/0.29	红花 - 当归
姜科 - 豆科	0.80/0.23	当归 - 柴胡	伞形科 - 蔷薇科	0.93/0.29	赤芍 - 当归
伞形科 - 菊科	0.79/0.22	白术 - 黄芪	伞形科 - 伞形科	0.77/0.28	川穹 - 桃仁
豆科 - 豆科	0.74/0.22	白术 - 党参	伞形科 - 姜科	0.69/0.26	川穹 - 红花
蔷薇科 - 豆科	0.70/0.22	甘草 - 黄芪	唇形科 - 豆科	0.67/0.26	川穹 - 赤芍
伞形科 - 桔梗科	0.63/0.21	升麻 - 黄芪	伞形科 - 芍药科	0.65/0.24	丹参 - 当归
伞形科 - 豆科	0.63/0.21	升麻 - 当归	唇形科 - 菊科	0.59/0.24	赤芍 - 桃仁

Acknowledgements

This work is partially supported by Scientific Breakthrough Program of Beijing Municipal Science&Technology Commission under grant number H020920010130.

References

1. Blagosklonny M.V., Pardee A.B., Unearthing the gems. Nature, 2002, 416(6879). 373.
2. Jenssen T.K., Lagreid A., Komorowsk J. et al., A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, vol 28,2001.pp: 21-28.
3. Bunescu R., Ge R., Rohit J.K. et al, Learning to Extract Proteins and their Interactions from Medline Abstracts. Tom Fawcett, Nina Mishra eds.,Proc. of ICML-2003 on Machine Learning in Bioinformatics, Menlo Park :AAAI Press, 2003, pp:46-53.
4. Nahm U.Y., Mooney R.J., A Mutually Beneficial Integration of Data Mining and Information Extraction. AAAI-2000, Austin, TX, pp: 627-632, July 2000.
5. Brin, S., Extracting Patterns and Relations from the World Wide Web. WebDB Workshop at EDBT-98. 1998.