基于文本挖掘技术分析类风湿性关节炎、强直性脊柱炎、溃疡性结肠炎和哮喘中医用药规律*

□丁晓蓉** （中国中医科学院中医临床基础医学研究所 北京 100700）
吕毅斌 （江中药业股份公司 南昌 330006）
王志飞 郑 光 郭洪涛 姜 森 吕爱平**
（中国中医科学院中医临床基础医学研究所 北京 100700）

摘 要：方剂可以概括为中医辨证论治体系的数据集合，文本挖掘技术可将隐含在数据中的配伍规律以可理解方式进行表述。本研究选择中国生物医学文献数据库为载体，类风湿性关节炎（RA）、强直性脊柱炎（AS）、溃疡性结肠炎（UC）和哮喘（Asthma）4种自身免疫性疾病为研究对象，采用数学和统计的方法，获取高频率、协同出现的关键药对，探寻中医治疗这4种疾病的用药规律。结果表明，采用这种计算方法总结出的4种疾病常用中药规律与其病机是相符的，且黄芪作为这4种疾病共同使用的常用中药，可能作用于自身免疫性疾病的特异性病理靶标。

关键词：自身免疫性疾病 文本挖掘 中药 配伍规律
doi: 10.3969/j.issn.1674-3849.2010.05.031

对已有的中医药数据进行分析挖掘是进一步丰富中医药信息资源的重要内容。文本挖掘是从非结构化的文本中发现潜在的概念以及概念间的相互关系，是指从大量文本数据中提取出可理解的、未知的、最终可用的知识的过程[1]。文本挖掘技术作为知识获取的有力工具，可以将隐含在数据中的配伍规律以可理解方式进行表述，为核心处方的提取提供技术支持。它能以线性和非线性方式解析数据，且能进行高层次的知识整合，又善于处理模糊和非量化数据。大量的中医药临床报道积累了丰富的文本数据，为文本挖掘提供了充分的数据条件。

收稿日期：2010-10-13
修回日期：2010-10-15
* 科学技术部“十一五”国家支撑计划项目（2006BAI04A10）：基于二次临床研究的中医药治疗类风湿性关节炎的临床评价，负责人：吕爱平；国家科学技术部创新方法学专项（2008IM020900）：中医药科学方法研究-中医药科学方法总论研究，负责人：吕爱平；国家自然科学基金杰出青年项目（30825047）：中西医结合基础——疾病证候分类研究，负责人：吕爱平；国家自然科学基金项目（30902000）：基于肾阳虚关节炎大鼠壮骨关节丸毒性反应评价研究，负责人：吕诚；第二批中国博士后科学基金特别资助金（200902184）：基于文本挖掘和复杂病证生物网络构建技术探索异病同治法中“证”的生物学特征，负责人：丁晓蓉。
** 联系人：丁晓蓉，中西医结合博士后，副研究员，主要研究方向：疾病证候分类；吕爱平，博士生导师，研究员，主要研究方向：疾病证候分类，E-mail：lap64067611@126.com。

类风湿性关节炎(RA)、强直性脊柱炎(AS)、溃疡性结肠炎(UC)和哮喘(Asthma)在现代医学分类系统中均属于自身免疫性疾病,有着相同的病理基础,在临床中也会采用相同的治疗药物,比如 TNF-a 抑制剂治疗 RA 和 UC。虽然当代中医治疗大多数是建立在病证结合的基础上,但辨证依然在指导谴方用药中起重要作用。在本项研究中,我们以中医临床文献数据库为载体,应用文本挖掘技术,对治疗这 4 种疾病的方药进行抽提分析,以期从海量的数据中发现这 4 种疾病的谴方用药规律,并对比其相同点和不同点。

一、材料和方法

1. 文本数据收集

首先,登录中国生物医学文献数据库*(Chinese BioMedical Literature Database,CBM)在主题检索下检索关键词"类风湿性关节炎"、"强直性脊柱炎"、"溃疡性结肠炎"和"哮喘"。经过检索,出现款目词、主题词、命中文献数,合并检索主题词,分别得到文献 13994 篇、3507 篇、8705 篇和 36278 篇(检索日期:2010 年 8 月 29 日)。

2. 文本数据处理

我们提取的信息主要是机标关键词(包括核心和非核心两种类型,以下简称关键词)。将收集来的数据,整合到一个平面文件(后缀 TXT)里面,以 ANSI 编码格式保存。然后,利用专有的文本提取工具(正申请软件著作权),对下载的非结构化的 TXT 文本数据进行信息提取,保存成格式化的、便于数据库(Access)和大型数据库(Microsoft SQL Server,SQL)处理的格式。提取出来的数据,首先存入 Access 数据库,作为我们下一步数据处理的材料,然后导入 SQL 中进行下一步的挖掘分析。

3. 数据挖掘以及分析

根据上述生成的 Access 数据库,将"结果"数据表导入 SQL 中,以"Table_Initial"为表名称,针对"序号"和"机标关键词"进行处理。为了方便处理,我们将"序号"和"机标关键词"两个字段分别用 PMID(类似于 PubMed 里面的字段名)和 DescriptorName(类似于 PubMed 里面的字段名)来表示。

我们假设,在同一篇文章中出现的关键词,在关键词这一抽象层面上,部分反映整篇文章的信息。并且,就某一篇具体的文献来说,相关的关键词之间存在着"共同出现"这一基本事实。这种协同出现不是随机的,而是蕴含有一定的意义[2],尤其是在以很高的频率、协同出现的关键词对,在一定的程度上,反映了全国乃至世界科研工作者对它们的重视程度。更重要的是,针对目前的文本挖掘技术来说[2~4],这些协同出现的关键词,是很好的基础素材。

基于上面的分析,我们要做的第一步,就是从初始数据表(Table_Initial)中构造针对每一篇文献共同出现的关键词对。就此,我们构造了关键词组合算法,来实现这一工作。该关键词组合算法的核心,就是对同一篇文献中出现的关键词进行配对,然后去除冗余的关键词对,然后将构造出来的关键词对,输出到关键词对数据表中(DN_pairs),供以后分析使用。

经过关键词组合算法的构造,我们得到名为 DN_pairs 的数据表。经过观察,我们发现数据表 DN_pairs 存在大量相同的关键词对,这些冗余的数据,对于数据分析来说,大部分属于噪音,对此,我们将相同的关键词儿对进行合并处理,只保留它们出现的频数。这一工作,我们构造关键词对频数统计的算法来实现。该统计算法,将关键词对以及其出现的频数,输出到名为 DN_pairs_frqcy 的数据表中。在数据表 DN_pairs_frqcy 内,所有的关键词对,都只出现一次,并且都有一个出现的频数(Frequency)。

4. 数据的可视化

根据"数据挖掘以及分析"中得到的数据表,我们抽出不同频数的关键词对,用 Cytoscape 2.7 进行可视化处理。根据中药间相关频次逐步抽提,将数据分为 1~7 层(频数由低到高),我们抽取每个病的中间层和最高层进行分析。

二、结　果

1. 治疗 RA 的中药使用规律

从图 1A 中可以看出,黄芪、桂枝、芍药、知母、当归、威灵仙、川芎、地龙、全蝎、防己是中医治疗 RA 的常用药物。此外,白芍和雷公藤、独活和羌活、乳香和没药、川乌和草乌分别成对出现,也是中医治疗 RA 的常用药物组成。从图 1B 中可以看出,芍药、知母和

* http://sinomed.cintcm.ac.cn/index.jsp.

桂枝是治疗 RA 的核心药物。

2. 治疗 AS 的中药使用规律

从图 2A 中可以看出，地龙、麻黄、当归、杜仲、虎杖、桂枝、骨碎补、鹿角、土鳖虫、黄芪是中医治疗 AS 的常用药物。此外，鸡血藤和青风藤，川芎、红花和防风分别成对出现，也是中医治疗 AS 的常用药物组成。从图 2B 中可以看出，麻黄、鹿角胶和当归，川芎和红花是中医治疗 AS 的核心药物。

3. 治疗 UC 的中药使用规律

从图 3A 中可以看出，黄芪、党参、木香、甘草、白术、茯苓、白芍、柴胡、黄连、白头翁、黄芩、黄柏、地榆是中医治疗 UC 的常用药物。从图 3B 中可以看出，黄芪、党参、白术、黄连和甘草是中医治疗 UC 的核心药物。

4. 治疗 Asthma 的中药使用规律

从图 4A 中可以看出，麻黄、杏仁、苏子、射干、陈皮、半夏、细辛、五味子、甘草、黄芩、白果、桔梗、葶苈子、石膏、桂枝、干姜、甘遂、黄芪、僵蚕和款冬花是中医治疗哮喘的常用药物。从图 4B 中可以看出，麻黄、杏仁和甘草是中医治疗哮喘的核心药物。

三、讨 论

自身免疫性疾病(Autoimmune Disease，AID)是指机体免疫系统产生了针对自身成分的自身抗体(Autoantibody) 或自身反应性 T 淋巴细胞 (Autoreactive T lympho-cyte)，并与相应的自身成分产生了过强的免疫应答，使机体组织细胞发生病理损害，表现出一系列临床症状。RA、AS、UC 和 Asthma 在现代生物医学分类中均属于自身免疫性疾病，因此在病理上都会出现这类疾病的共性反应—过强的免疫应答。这种过强的免疫应答由于作用的靶器官不同，产生的病理损害也会不同。当组织器官的病理损害和功能障碍仅限于抗体或致敏淋巴细胞所针对的某一器官时，称之为器官特异性自身免疫疾病，例如 UC；而由于抗原抗体复合物广泛沉积于血管壁等原因导致全身多器官损害，称之为非器官特异性自身免疫疾病，例如 RA、AS 和 Asth-ma。这种病理变化的共性和差异性，决定了这 4 种疾病在应用现代医疗手段治疗时既有相同点，又有不同处。

从文本挖掘分析的结果可以看出，黄芪在这 4 种疾病中都是被经常使用的中药。在中医理论中，黄芪具有补气固表、托毒生肌、养血助阳、止汗利尿等功效。现代研究证实，黄芪提取物能够有选择性的改变 Th1/Th2 细胞因子的分泌模式，增加 Th1 细胞 IL-4 的表达，降低 Th2 细胞 INF-γ 的表达[5]。此外，黄芪提取物可能通过抑制 IL-6，PGE2 的生物合成起到抗

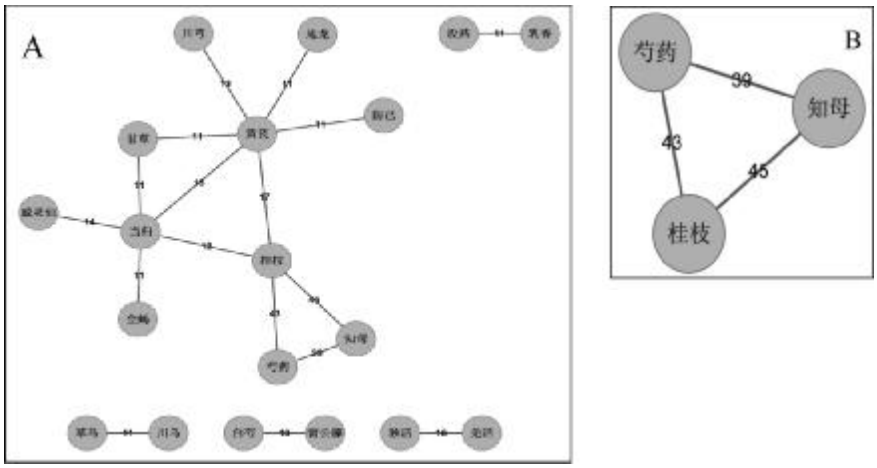

图 1 治疗 RA 常用中药频数关系图

注：A 为中间层数据，B 为第高层数据。

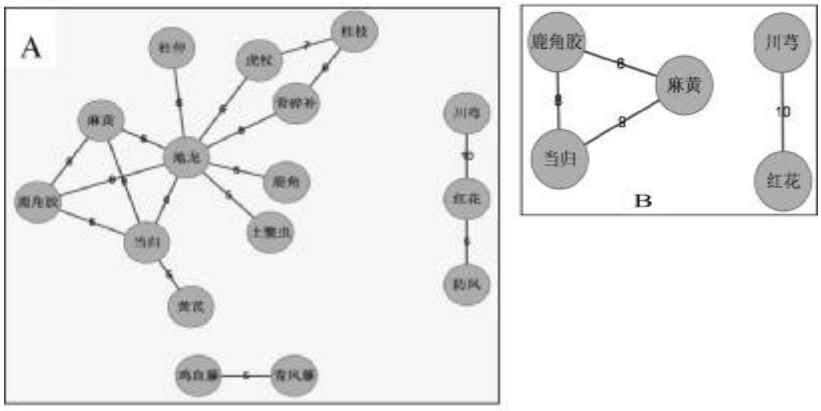

图 2 治疗 AS 常用中药频数图

注：A 为中间层数据，B 为第高层数据。

感染的作用[6]。黄芪在这四种疾病的治疗中同时出现，提示黄芪可能作用于自身免疫性疾病的特异性病理靶标。那么,随着自身免疫性疾病病理生理研究的深入，黄芪的相关药理研究也可以开拓出新的领域。

此外,从图 1A 和图 2A 中,我们也可以看到,黄芪和当归这一对药，是 RA 和 AS 这两种疾病都比较经常使用的药物组合，而在 UC 和 Asthma 中却没有出现;在图 1A、图 2A 和图 4A 中,黄芪和桂枝、地龙同时在 RA、AS 和 Astham 中出现,而在 UC 中没有出现；在图 2A 和图 4A 中，黄芪和麻黄同时在 AS 和 Asthma 中出现,而在 RA 和 UC 中没有出现;在图 3A 和图 4A 中，黄芪和黄芩同时在 UC 和 Asthma 中出现,而在 RA 和 AS 中没有出现。那么,这种药对应用的偏好,能否延伸到自身免疫性疾病分类的研究,即能够采用相同药对治疗的疾病，可能存在一类相同的病理变化，这种相同的病理变化可使自身免疫性疾病的分类进一步细化,进而又影响到临床治疗。

传统的中医谴方用药是在辨证治疗思维模式的指导下确立的,形成证-法-方-药的治疗体系。其中,方剂是中医辨证论治的完整体现，可以概括为这一治疗体系的数据集合。目前已有学者利用数据（文本)挖掘技术,对中医用药规律进行总结,例如刘杰等[7]运用无尺度网络等数据挖掘方法,从中药的功效、种类、剂量、配伍关系等方面,探讨肺癌中医证候特征及治疗特点，总结出肺癌中医治疗以扶正培本为主要原则;周鲁等[8]基于中医文献数据库分析了治疗高血压常用中药复方的用药规律，指出利用计算机技术进行大规模的数据统计,是寻找中药复方治疗高血压用药规律的有效方法;杨洪军等[9]基于《中医方剂大辞典》的 10 余万首方剂，选出 502 首以头痛为主治的方剂，通过计算中药出现的频次,总结出辛味药在头痛用药中出现频次最高,而这一结果与“通则不痛”的疼痛病机认识十分贴切。是进一步分析我们的研究结果可以发现,中药治疗 RA(图 1B)以温经通脉(桂枝)、养血止痛(芍药)和滋阴清热(知母)药物最为常用;治疗 AS(图 2B)以祛风散寒(麻黄)、补血活血(当归)、补益精血(鹿角胶)和活血化瘀、行气止痛(川芎和红花)药物最为常用;治疗 UC(图 3B)以补中益气(黄芪和党参)、健脾益气燥湿(白术和甘草)、清热解毒燥湿(黄连）药物最为常用；治疗 Asthma(图 4B)以宣肺平喘(麻黄)、祛痰止咳利肺气(杏仁和甘草)药物最为常用。RA 的主要病机是气血阴液不足，风寒湿邪及痰瘀阻络,AS 的主要病机为肾虚督寒、痰瘀阻络,UC 的主要病机为肺脾功能失调、痰瘀阻络,而 Asthma 的主要病机为痰饮伏肺,我们关于这四种疾

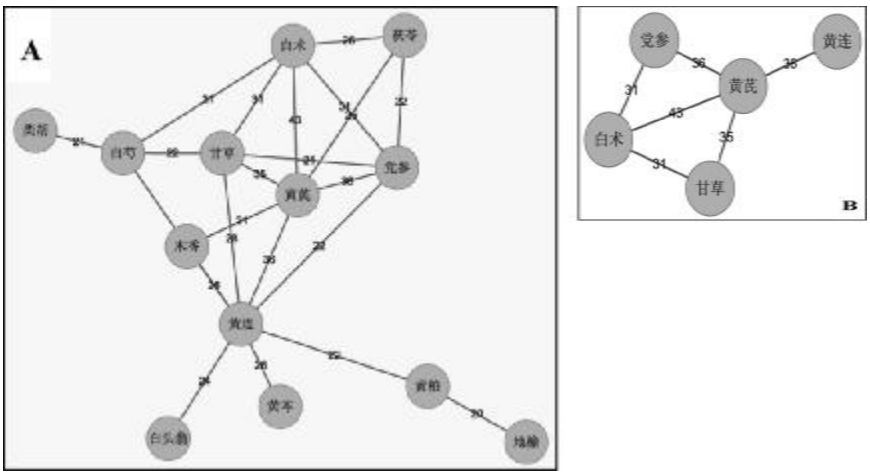

图 3　治疗 UC 的常用中药频数关系图

注:A 为中间层数据,B 为第高层数据。

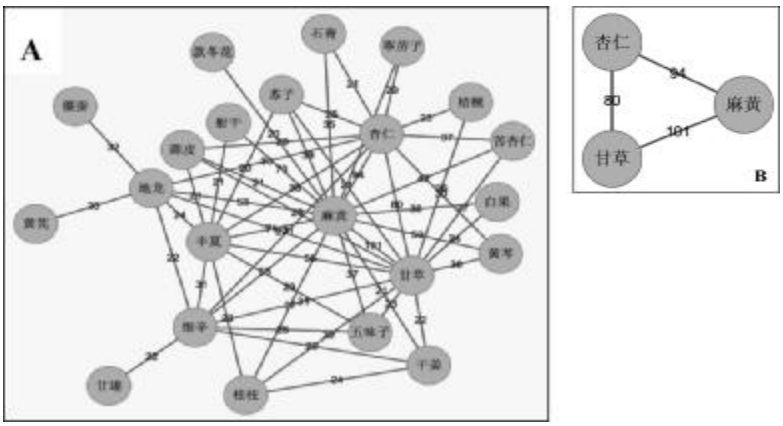

图 4　治疗 Asthma 常用中药频数关系图

注:A 为中间层数据,B 为最高层数据。

病用药规律的研究结果与其传统病机认识不谋而合。

中医处方的组成及配伍规律，受医生对疾病及证候的认识，对药性的理解及前人的经验等多个因素的影响。其加减变化规律，反映了临床证候的多样性及辨证论治的复杂性。尽管本项研究所显示的文本挖掘结果，只是代表了一种文献偏好，不能代表实际临床应用的全部情况。但是，应用数学和统计的方法，通过对这种文献偏好的分析，在某种程度上也可以为相关的临床基础研究提供思路。

参考文献

1 Feldman R, Dagan I. "Knowledge discovery in textual databases (KDT)" Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95) Montreal: AAAI Press, 1995:112~117.

2 Andrea Campagna, Rasmus Pagh. Finding associations and computing similarity via biased pair sampling. 2009 Ninth IEEE International Conference on Data Mining, 2009:61~70.

3 Jeffrey W Seifert. Data mining: An overview. CRS Report RL31798, 2004.

4 Brigitte Mathiak, Silke Eckstein. Five steps to text mining in biomedical literature. In Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, held in Conjunction with ECML/PKDD in Pisa, Italy, 2004, 24:47~50.

5 Kang H, Ahn K S, Cho C, et al. Immunomodulatory effect of Astragali Radix extract on murine TH1/TH2 cell lineage development. Biol Pharma Bulletin, 2004,27(12):1946~1950.

6 Shon Y H, Kim J H, Nam K S. Effect of Astragali radix extract on lipopolysaccharide-induced inflammation in human amnion. Biol Pharma Bulletin, 2002, 25(1):77~80.

7 刘杰，林洪生，侯炜，等. 利用数据挖掘方法对肺癌中医药治疗特点的初步研究. 世界科学技术-中医药现代化，2009，11(5):753~757.

8 周鲁，付超，蔡鑫. 治疗高血压常用中药复方用药规律分析. 中医杂志，2005，46(5):351~354.

9 杨洪军，王永炎. 头痛方剂用药规律研究. 中国中药杂志，2005，30(3):226~232.

To Analyze the Regularity of Traditional Chinese Medicine Herbs Application for Rheumatoid Arthritis, Ankylosing Spondylitis, Ulcerative Colitis and Asthma with Text Mining Technique

Ding Xiaorong[1], Lv Yibin[2], Wang Zhifei[1], Zheng Guang[1], Guo Hongtao[1], Jiang Miao[1], Lv Aiping[1]

(1. Institute of Basic Research In Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China;
2. Jiangzhong Pharmaceutical Co., Ltd., Nanchang 330006, China)

Abstract: Prescription can be acted as data set of zheng differentiation-treatment system. The implied regularity of this data set can be presented with text mining technique. In this research, we select Chinese BioMedical Literature Database as vector to analyze the Chinese herb medication regularity of Rheumatoid Arthritis, Ankylosing Spondylitis, Ulcerative Colitis and Asthma. The high frequency and synergism herbs pattern are obtained. The results show that the medication regularity get with this computational method are in accordance with the pathogenesis of these four diseases. Besides, Huang Qi (Radix Astragali) is included in these four diseases, which could affect the specific pathological target of autoimmunity disease.

Keywords: Autoimmunity disease, Text mining, Herb medicine, Medication regularity

（责任编辑：李沙沙，责任译审：吕爱平）