

利用数据挖掘技术探索类风湿关节炎与糖尿病“同证”的科学基础*

□郭洪涛** (中国中医科学院中医临床基础医学研究所 北京 100700)
(上海中医药大学 上海 201203)

郑光** 张弛 吕诚 查青林 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:数据挖掘可以从海量数据中抽取有用的信息,本文用数据挖掘方法对类风湿关节炎与糖尿病“同证”的科学基础进行探索。首先,从中国生物医学文献数据库中下载中医药治疗类风湿性关节炎及糖尿病的相关数据,保存于 Access 数据库中,然后导入 SQL Server 对数据进行进一步处理并得到挖掘结果。针对得到的结果,用 Cytoscape 软件对结果进行可视化处理,进而得到中医治疗类风湿性关节炎及糖尿病的用药规律网络图。从图中可以发现,黄芪、桂枝、当归、川芎、防己等为中药治疗类风湿性关节炎常用核心药物。黄芪、当归、丹参、川芎、麦冬、山药、玄参等为中药治疗糖尿病常用的核心药物。同时发现中药治疗两病共同的核心用药:黄芪、当归。中医在治疗类风湿性关节炎及糖尿病时,均采用补气活血化痰的治疗方法。因此,根据数据挖掘的结果,可以证明两种疾病存在共同的证候特点是有科学基础的。

关键词:异病同证 类风湿性关节炎 糖尿病 数据挖掘 可视化

doi: 10.3969/j.issn.1674-3849.2010.05.029

数据挖掘是指从数据库中挖掘出潜在的, 前所未有的, 且很可能非常有用的信息的过程^[1]。Grupe 及 Owrang^[2]则认为数据挖掘是从已存在的数据库中挖掘出专家仍未知的新事实。有许多学者视数据库的知识发现 (Knowledge Discovery in Database, KDD) 为

数据挖掘的同义词, Fayyad 等^[3]则视数据挖掘为基于数据库的知识发现中的一个步骤, 而数据库的知识发现为从大量数据中选取合适的数据, 进行数据处理、转换等工作, 再进行数据挖掘与结果评估等一系列过程。

随着大量中医药文献的不断产生, 我们有理由相信, 这些数据中可能存在大量有用而未被发现的信

收稿日期: 2010-10-13

修回日期: 2010-10-15

* 科学技术部“十一五”国家科技支撑计划项目(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项目(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; 国家自然科学基金项目(81072982): 隶属度和模糊模拟数学方法在类风湿关节炎分类中的应用基础研究, 负责人: 郑光。

** 联系人: 郭洪涛, 在读博士, 主要研究方向: 中医药治疗自身免疫性疾病机制, Tel: 010-64014411-3401, E-mail: guoht2009@126.com; 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。

息。我们尝试利用现有的数据库平台,收集客观的数据,然后进行无关联规则的数据挖掘整理,再介入专业知识进行评估分析,希望从中得到有用的信息,并发现新的知识。本文利用数据挖掘技术,针对类风湿关节炎(Rheumatoid arthritis, RA)和糖尿病(Diabetes mellitus, DM)两个不同的疾病,从数据库中收集数据,然后进行数据分析处理,试图寻找一些“同证”的证据。

一、方法和材料

文本挖掘是数据挖掘的一个分支,它是从非结构化的文本数据中,抽取有意义的数据^[4]。具体说,文本挖掘应用到生物、医学上,可以分为文本数据收集、处理、结构化分析、可视化以及评价 5 个步骤^[5]。

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”。经过检索,出现款目词、主题词、命中文献数,合并检索主题词,共得到文献 13692 篇(检索日期:2010 年 6 月 4 日)。为了能看到每篇文献的流水号、标题、摘要、主题词等信息,在显示格式中选择“详细”和“显示全部”。应用同样方法在 CBM 数据库收集 DM 的数据,得到 109276 篇(检索日期:2010 年 6 月 4 日)。

2. 文本数据处理

将收集来的 RA 和 DM 的数据,分别按照下载的先后顺序,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“文本数据收集”中下载的非结构化的 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,以下简称 SQL)处理的格式。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)。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基础数据,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“文本数据处理”中生成的 Access 数据库,我们将基础数据导入 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。

我们认为,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^[6],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^[4-6],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 DN_pairs_frqcy,抽出不同频数的关键词对,根据中药间相关频次手工分类,用 Cytoscape 2.7 软件进行可视化处理,分别得到治疗 RA、DM 的中药用药,以及 RA 和 DM 交互时中药使用情况的网络图,从中选出有代表性的两层进行分析讨论。

5. 结果

(1)治疗 RA 中药网络图(见图 1~2)。

图中圆圈内药物为检索文献中治疗 RA 所涉及到的频数较高的中药,药物间连线代表两个药物应

* 网址 <http://sinomed.cintcm.ac.cn/index.jsp>.

用时的相互联系程度,连线越密,说明配伍的频次越高;连线中点上的数字代表两个药物出现的频数,数字越大,代表该药使用的频次越高。

图 1 中的频数范围在 7~10, 配伍频次较高的有黄芪、桂枝、附子、当归、鸡血藤、川芎、羌活等,说明在治疗 RA 时应用药物以补气温阳活血化瘀为主。图 2 为高频数图,频数在 11~45 之间,即黄芪、桂枝、当归、川芎、防己、地龙、全蝎、威灵仙、芍药、知母、乳香、没药、川乌、雷公藤、独活、羌活等为中药治疗类风湿性关节炎常用核心药物;并且乳香和没药、草乌和川乌、白芍和雷公藤、羌活和独活以对药的形式出现;尤其是在高频数的同时,黄芪、当归、桂枝还以高配伍的情况出现,说明补气活血温阳在 RA 治疗中可能是核心治疗法则。

(2) 治疗 DM 中药网络图(见图 3~4)。

图 3 中的频数范围在 20~87, 配伍频次较高的有黄芪、人参、山药、麦冬、丹参、当归、川芎、葛根等,说明在治疗 DM 时应用药物以补气养阴活血化瘀为主。图 4 为高频数图,频数在 102~258 之间,即黄芪、丹参、山药、当归、人参、红花、葛根等为中药治疗 DM 常用核心药物,并且黄芪为治疗 DM 的核心配伍药。

(3) RA 和 DM 交互中药网络图(见图 5~6)。

图 5 及图 6 是通过计算两病的源数据,使用相同的中药互相匹配,从而得到的数据表,再可视化后得到网络图。图 5 是频数在 5~9 范围,可以看到,两个病共同使用的中药以黄芪、当归、桂枝、知母、鸡血藤等为核心,同时滋阴清热的沙参、天花粉、麦冬、玄参也归为一类。图 6 频数在 10~17 范围,高频药物以黄芪、当归为核心,桂枝、川芎、地龙等次之;同时全蝎、蜈蚣,即是虫类药,又是对药,在两病的治疗中也很常用。

二、讨 论

“异病同治”是当前中医临床实践的重要法则之一,其核心内容是不同的疾病可能存在于一种证候类型之中,因而可以用同一种治疗方法,也就是说中医在治疗疾病时不是着眼于病的异同,而是着眼于“证”的异同,遵循“证同治亦同,证异治亦异”,“异病同治”与“异病同证”共同体现了中医辨证论治的特色^[7]。虽然这一法则为中医界所公认,但确切的客观的证据并不能很好呈现。中医临床能够认同,RA 和 DM 作为两种不同疾病,病程长,久病多虚多瘀,因此

两病的发病过程中伴有气虚血瘀证,这也即中医的“异病同证”,根据中医理论,“同证”即可“同治”。对于气虚血瘀证,临床就可以采用补气活血化瘀的方

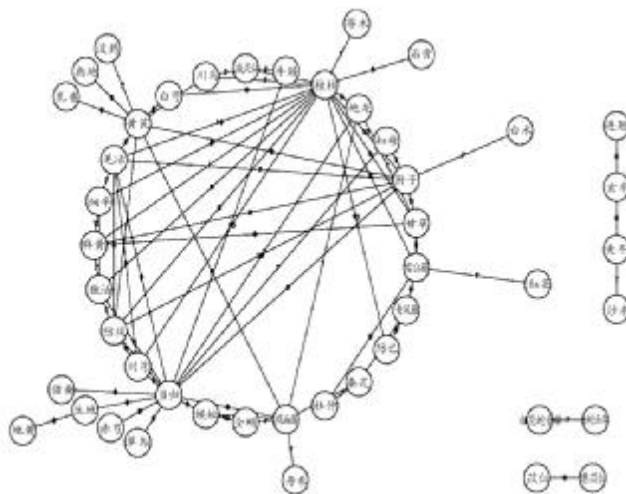


图 1 治疗 RA 中药网络图第 1 层

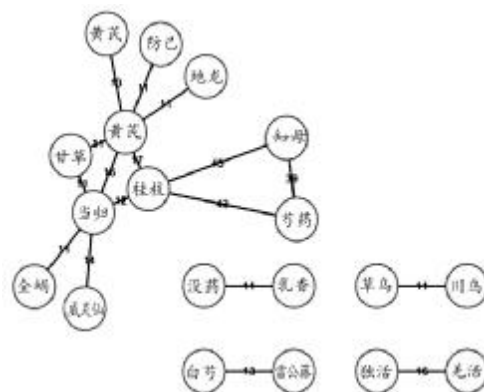


图 2 治疗 RA 中药网络图第 2 层

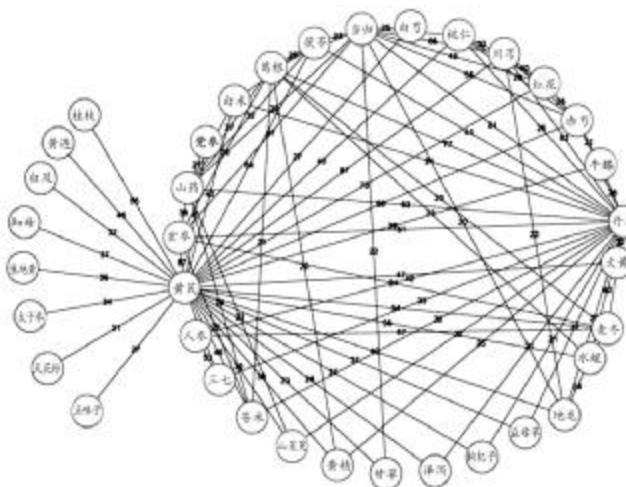


图 3 治疗 DM 中药网络图第 1 层

法进行治疗。

由以上网络图可以发现,中医治疗 RA 常用的药物以补气助阳(黄芪、桂枝)、活血化瘀(当归、川芎、鸡血藤)药为主,辅以祛风湿药止痛(草乌-川乌、乳香-没药、雷公藤-白芍、羌活-独活、防己、全蝎等),体现了治疗以活血化瘀、祛风湿止痛的治则。中药治疗 DM 常用的药物以补气药(黄芪、人参)、活血药(当归、丹参、川芎、红花)、养阴药(麦冬、山药、玄参)等药物为核心,治疗方法以补气活血化瘀、养阴清热为主。而两病中相同用药以黄芪、当归为核心用药,桂枝、川芎,地龙次之,说明补气活血化瘀是其共同的治法。

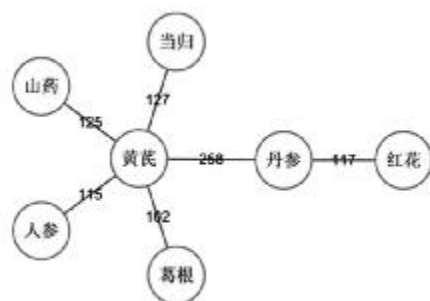


图4 治疗 DM 中药网络图第2层

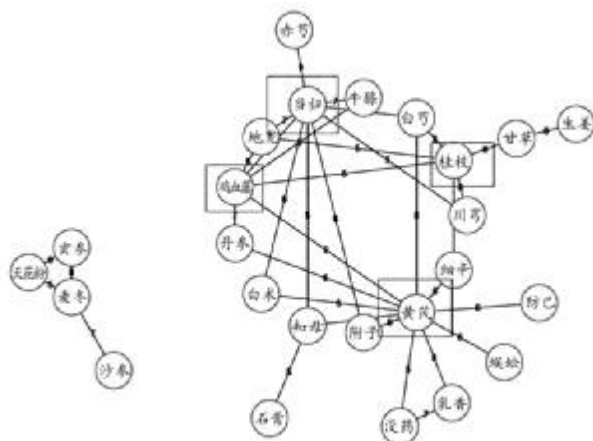


图5 RA 和 DM 交互中药网络图 1



图6 RA 和 DM 交互中药网络图 2

数据挖掘过程中发现,类风湿性关节炎、糖尿病虽为两种不同疾病,但其均使用了频次极高的补气及活血化瘀的药物,因此其证候特点中肯定存在相同点,即类风关和糖尿病在发病的过程中,都出现了气虚及血瘀的症状,针对这些症状,两种疾病都采用了活血化瘀的治疗方法。从而也反推出 RA 和 DM “异病同证”是客观存在的。故此,利用数据挖掘技术证明,两种疾病“同证”是有科学基础的。

三、展 望

我们通过对海量数据 (RA13692 篇, DM109276 篇) 的挖掘计算,呈现的结果与中医治疗两种疾病的传统认识基本一致,如前所述的核心用药、对药、常用配伍药,甚至常用的方剂等。并且还发现一些新的知识点,如一些新的可能的组方、对药、配伍等,而这些新发现可能为中医临床用药提供有益的参考。我们将会对更多的疾病展开类似研究,以期获得更多更全面的数据。我们还尝试对 PUBMED 的中医药数据进行挖掘,并且把中文数据和英文数据进行合并,已经发现了一些中药治疗某些疾病的通路,当然,这些通路还需要我们进一步的确认。综上,数据挖掘技术作为一个从海量数据中获取知识的有力工具,具有广阔的应用前景,其在中医药领域的应用将会对中医药学术水平的提高、现代化研究的进程、生存空间的扩展产生巨大的促进作用。

参考文献

- 1 Frawley WJ, Piatetsky SG and CJ. Knowledge Discovery in Databases: An Overview. Communications of the ACM, 1996, (39):1-34.
- 2 Grupe FH, Owrang MM. Database mining: Discovering new knowledge and competitive advantage. Information Systems Management, 1995, 4 (12):26-31.
- 3 Fayyad U, Shapiro GP, Smyth P. From data mining to knowledge discovery in database. AI magazine, 1996, 17(3):37-54.
- 4 Jeffrey W Seifert. Data mining: An overview. CRS Report RL31798, 2004.
- 5 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.
- 6 Andrea Campagna, Rasmus Pagh. Finding associations and computing similarity via biased pair sampling. 2009 Ninth IEEE International Conference on Data Mining, 2009:61-70.
- 7 吕爱平, 陈可冀. 疾病的证候分类研究思路. 中国中西医结合杂志, 2005, 25(9):843-844.

Exploring Commonly Existed Networks of Chinese Herbal Medicines Used in Rheumatoid Arthritis and Diabetes Mellitus through Data Mining

Guo Hongtao^{1,2}, Zheng Guang¹, Zhang Chi¹, Lv Cheng¹, Zha Qinglin¹, Jiang Miao¹, Lv Aiping¹

(1. Institute of Basic Research In Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China;

2. Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China)

Abstract: One important terminology in Traditional Chinese Medicine (TCM) is called treating different diseases with the same therapy. In TCM practice, some patients with Rheumatoid Arthritis (RA) and some other patients with Diabetes Mellitus (DM) can be treated with similar therapies. This suggests us that there should be networks of Chinese herbal medicines commonly existed in the literature on both RA and CHD. In this paper, we propose a data mining algorithm to retrieve simple and meaningful networks of Chinese herbal medicines used in treating both RA and DM. When the networks are built, we analyzed them in the theory of TCM, i.e., Chinese herbal medicines of Huang Qi (Radix Astragali), Gui Zhi (Ramulus Cinnamomi), Dang Gui (Radix Angelicae Sinensis), Chuan Xiong (Rhizoma Chuanxiong), Mai Dong (Tuber Ophiopogonis), Shan Yao (Radix Dioscoreae) and Xuan Shen (Radix Scrophulariae) are commonly used in treating both RA and DM. What's more, Huang Qi (Radix Astragali) and Dang Gui (Radix Angelicae Sinensis) are core association. They function as buqi (tonify qi), huoxue (blood-activating), and huayu (stasis-resolving).

Keywords: Treating different disease with the same therapy, Rheumatoid arthritis, Diabetes mellitus, Data mining, Visualization

(责任编辑:李沙沙,责任译审:吕爱平)