·数据挖掘·

基于文本挖掘方法探索冠心病药-证对应规律研究

杨静[1],田景平[1,2],谭勇[1],蔡峰[1],姜淼[1],吕爱平[1*]
(1. 中国中医科学院中医临床基础医学研究所,北京 100700;
2. 广州中医药大学研究生院,广州 510006)

[摘要] **目的**:采用文本挖掘方法,探索中医药治疗冠心病药证对应规律。**方法**:在中国生物医学文献(CBM)数据库中收集有关冠心病文献数据,采用基于敏感关键词频数统计的数据分层算法,挖掘冠心病证候、中药及它们之间的规律,并利用 Cytoscape 2.8 软件进行可视化展示。**结果**:冠心病证候以气虚血瘀为最常见证候,其次为心血瘀阻、气阴两虚。丹参、黄芪、葛根、人参、红花、川芎、三七、麦冬为治疗冠心病的核心药物。定向挖掘结果显示治疗冠心病气虚血瘀证的药物与冠心病常用药物基本一致。**结论**:文本挖掘技术能够比较客观地总结疾病药证对应规律,为临床应用提供有益的探索和参考。

[关键词] 文本挖掘; 数据分层算法; 冠心病; 中医证候
[中图分类号] R285.5 [文献标识码] A [文章编号] 1005-9903(2013)21-0319-04
[doi] 10.11653/syfj2013210319

Exploring Association Rules of Syndrome and Herbal Medicine within Framework of Traditional Chinese Medicine on Coronary Heart Disease Through Text Mining Approach

YANG Jing[1], *TIAN Jing-ping*[1,2], *TAN Yong*[1], *CAI Feng*[1], *JIANG Miao*[1], *LV Ai-ping*[1*]
(1. *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*; 2. *Guangzhou University of Traditional Chinese Medicine, Guangzhou* 510006, *China*)

[**Abstract**] **Objective**: Explore the principle of pattern-herbs of Chinese medicine on coronary heart disease (CHD) by text mining analysis. **Method**: Download the literatures about CHD from Chinese Bio-Medical Literature Database. Mine the rules of traditional Chinese medicine (TCM) pattern and Chinese herbal medicines by data slicing algorithm. The results are showed visually with Cytoscape 2.8 software. **Result**: The pattern of qi deficiency and blood stasis is the main TCM symptom. Heart blood stasis syndrome and deficiency of qi and yin are the common pattern. The center TCM herbs of CHD are Radix Salviae Miltiorrhizae, Radix Astragali, Radix Puerariae, Radix Ginseng, Flos Carthami, Rhizoma Chuanxiong, Radix Notoginseng, Radix Ophiopogonis. As is shown in directional mining, the center herbs for damp-heat pattern are similar with those treatments for CHD. **Conclusion**: Text mining approach provides an important method in exploring the rules of pattern-herbs of Chinese medicine.

[**Key words**] text mining; data slicing algorithm; coronary heart disease; syndrome of traditional Chinese Medicine

---

[收稿日期] 20121223(020)
[基金项目] 国家自然科学基金杰出青年项目(30825047);国家自然科学基金青年基金项目(30902003);中国中医科学院第六批自主选题项目(Z0218)
[第一作者] 杨静,博士,助理研究员,从事中西医结合临床基础研究,E-mail:yangjingdr@163.com
[通讯作者] *吕爱平,研究员,博士研究生导师,从事病证关联研究,Tel:010-64014411-3301

冠状动脉粥样硬化性心脏病(coronary atherosclerotis heart disease)简称冠状动脉性心脏病或冠心病(coronary heart disease, CHD),指由于冠状动脉粥样硬化使管腔狭窄或阻塞导致心肌缺血、缺氧而引起的心脏病,为动脉粥样硬化导致器官病变的最常见类型。冠心病作为一种由环境因素和遗传因素共同作用导致的复杂疾病,已成为世界范围内死亡和致残的首要原因[1]。中医药防治冠心病疗效好,副作用小,本文借助不断成熟的文本挖掘技术[2],结合原文献回溯,人工阅读分析等方法,对现有中文文献进行挖掘,探索冠心病药证对应规律。

1 材料与方法

1.1 文本数据收集 在中国生物医学文献数据(Chinese Bio Medical Literature Database, CBM, http://sinomed.cintcm.ac.cn/index.jsp)中以"缺省[智能]"状态下检索"冠心病",共得到文献66 896篇(检索日期2012 年 9 月 23 日),依次下载所有文献并保存。

1.2 文本数据处理 将收集来的数据,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。由 2 人同时背对背人工阅读下载文献,初步找出噪音性文献,不属于冠心病中医证候分类及其中药治疗的文献均视为噪音性文献,核对后统一进行筛除,对检索下载的文献进行初步筛选。然后,利用专有的文本提取工具(软件著作权,软著登字第 0261882 号,登记号 2010SR073409),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、便于大型关系型数据库(Microsoft SQL Server,简称 SQL)处理的格式,然后导入 SQL 中进行下一步的挖掘分析。

以"Table_Initial"为表名称,针对"序号"和"机标关键词"进行处理。为方便处理,将以上 2 个字段分别于用 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。首先从初始数据表(Table_Initial)中运用"关键词组合算法",对同一片文献中出现的关键词进行配对。该关键词组合算法的核心,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,据此进行数据的一次清洗工作,将构造出来的关键词对,输出到"关键词对数表"(DN_pairs)中,供分析使用。

经过关键词组合算法的构造,得到名为 DN_pairs 的数据表。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,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,通过构造"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)。

经过对 DN_pairs_frqcy 数据表中的数据评估发现,针对特定的疾病,表中仍存噪音问题。这些噪音不再是关键词的简单重复,而是相于专业只是来说的噪音问题,对此针对特定问题,对数据进行二次清洗。这些噪音的产生,主要是自然语言的义性和表达方式的多样性产生的。这些问题只能逐个分析,建立规则,然后根据规则进行数据的二次清洗。清洗完毕后的数据,最终既可以提取挖掘对象的一维频次,也可以得到挖掘对象的二维关系。最终从 DN_pairs_frqcy 数据表中抽出不同频次的关键词对,根据药物间相关频次手工分类,用 Cytoscape 2.8 软件进行可视化处理。最后的结果可视化成图,结合专业知识进行解析。

2 冠心病证候及中药文本挖掘结果

2.1 证候文本挖掘结果 一维频数结果:文本挖掘共提取到 173 个证型,排名前 10 的证候由高到低依次为(括号内为文献频数,下同):气虚血瘀(537)、血瘀(471)、心血瘀阻(370)、气阴两虚(318)、痰湿阻肺(300)、气滞血瘀(216)、心脉痹阻(171)、心气虚(129)、心阳虚(107)、肾阴虚(97)。由于构建词表及挖掘词的包含关系,相同证候可能会重复出现。回溯原文献数据集血瘀多为"气虚血瘀"、"心血瘀阻"等词,因此应视为噪音。

二维网络结果:进一步构建冠心病证候两两之间网络关系图(图 1,2)。图中圆圈内为证型名称,连线代表证候两两之间的联系。证候的连线愈多,代表该证型与疾病的关联程度越高;圆圈越大,代表该证型在文献中出现的频次越高。冠心病相关证候共同出现的组合共1 684种,选取排名前 100 位的证候联合(文献频数≥2)的网络图用于相对全面展示冠心病证候组合;选取排名前 8 位的证候联合(文献频数≥51)的网络图用于展示最主要的证候组合进行分析。图 2 提示气虚血瘀、心血瘀阻、气阴两虚、痰湿阻肺、心气虚等证候与其他证候同时出现频率较高。提高频次分层后(图 3),气虚血瘀、气阴两虚、气滞血瘀、心血瘀阻、心脉痹阻为冠心病最常见的两两相关的证候。

从证候一维及二维结果来看,气虚血瘀为其主要证候,其次为心血瘀阻、气阴两虚、痰浊阻肺、气滞

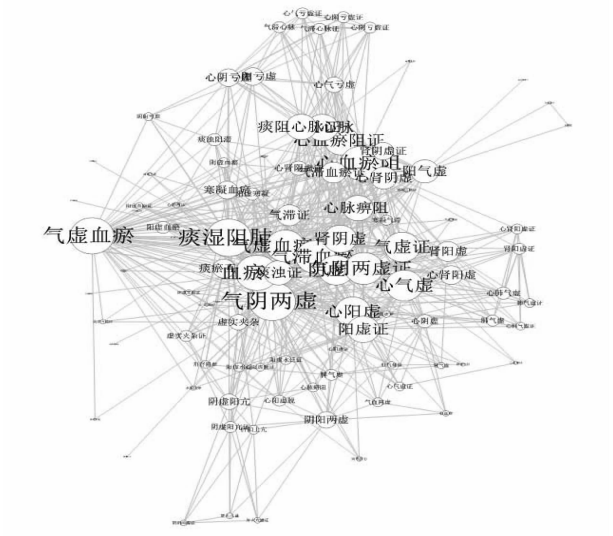

图 1　冠心病中医证候网络（频数≥2）

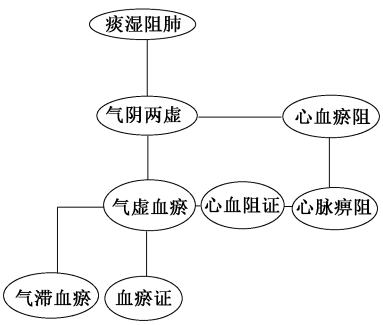

图 2　冠心病中医证候网络（频数≥51）

血瘀。

2.2　中药文本挖掘结果　一维频数结果：文本挖掘共提取到相关中药名称 237 个，排名前 10 的中药频次由高到低依次为：丹参（2 455）、黄芪（610）、葛根（579）、人参（560）、红花（426）、川芎（375）、三七（375）、麦冬（309）、麝香（302）、白及（240），见图 4。回溯原文献数据集发现：白及多为“C-反应蛋白及…”、“低密度脂蛋白及…”、“高密度脂蛋白及…”等词，因此为噪音，应予以剔除。此次挖掘得到的排名前 10 位的冠心病常用中药，与作者曾于 1 年前发表的文章中检索得到的排序前 10 名的中药，丹参、黄芪、人参、葛根、红花、麦冬、川芎、三七、麝香、生地黄，药味基本保持相同，排序有所变化，文献数量有所增加[3]。

二维网络结果：进一步构建冠心病中药两两之间网络关系图（图 3）。图中圆圈内为中药名称，连线代表中药两两之间的联系。中药的连线愈多，代表该中药与疾病的关联程度越高；圆圈越大，代表该中药在文献中出现的频次越高。冠心病相关中药组合共2 504种，选取排名前 100 位的中药联合（文献频数≥3）的网络图用于相对全面展示冠心病中药组合分类，选取选取排名前 10 位的药物联合（文献频数≥104）的网络图用于展示最主要的中药组合进行分析。图 5 提示黄芪、人参、丹参、当归、川芎、三七、瓜蒌、红花、桃仁、桂枝、白芍、葛根、半夏、麦冬、茯苓、党参、水蛭、檀香等中药与其他中药同时出现频率较高。当文献频数提高到 104 时，提示有 10 种药物经常联用，分别为：丹参、黄芪、红花、葛根、人参、三七、川芎、麦冬、瓜蒌、薤白（图 4）。

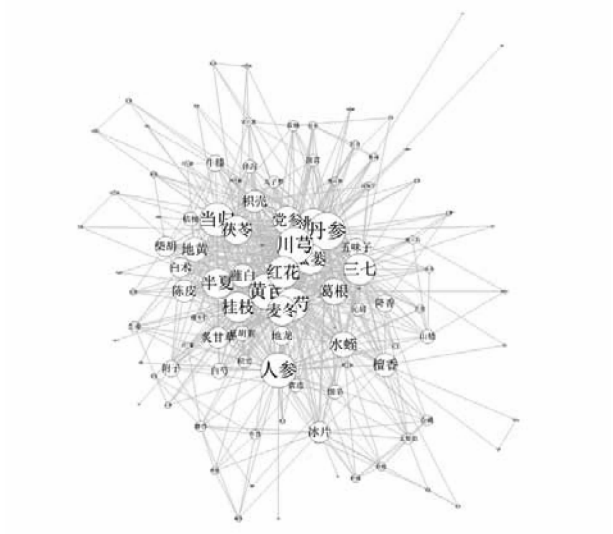

图 3　冠心病中药网络（频数≥3）

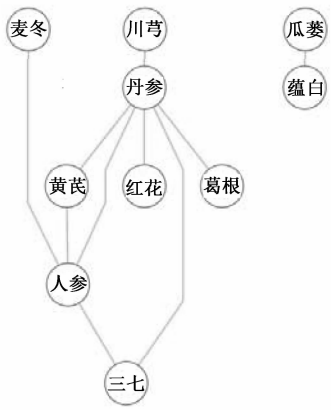

图 4　治疗冠心病中常用中药联用网络（频数≥104）

从中药的一维及二维的结果来看，丹参、黄芪、葛根、人参、红花、川芎、三七、麦冬为治疗冠心病的核心药物，且与其他药物联用的频数也最高。

2.3　定向文本挖掘结果　为进一步验证在文本挖掘技术下证候与中药的关系，本研究对气虚血瘀证对应的中药进行定向挖掘，发现治疗冠心病气虚血瘀证常用中药为：黄芪、党参、当归、丹参、黄精、人参、红花、川芎、赤芍、蒲黄、地龙。此项文本挖掘结果提示，治疗冠心病气虚血瘀证的中药与治疗冠心病的常用中药基本一致（图 5）。

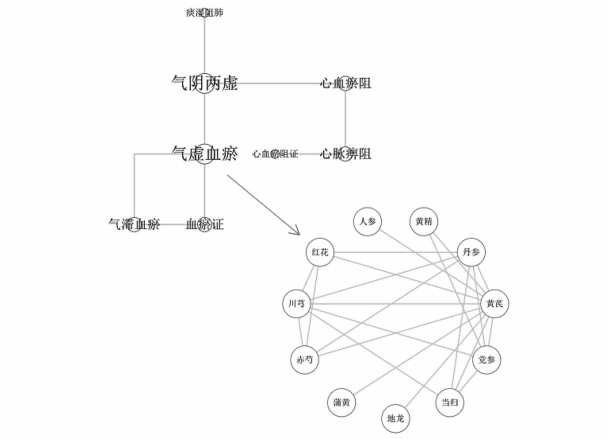

图 5　冠心病气虚血瘀证及相应中药的定向文本挖掘结果

3　讨论

文本挖掘技术是以计算语言学、统计数理分析为理论基础，服务于生物、医药、文献研究等学科的新兴的交叉学科[4]。文本挖掘能从海量的中医药文献中发现知识以促进中医临床研究和中药复方研发等多个方面。根据中医理论和专业知识，利用数据挖掘技术对中医药文献库与生物医学信息进行处理，为中西医结合研究提供新的思路和途径[5]，并使结果更加客观，可重复性强[6]。

冠心病属中医学“胸痹”、“心痛”、“心悸”等范畴。有学者研究近 10 年国内冠心病不同中医证型分布情况及不同地域的冠心病中医证型分布特征发现，冠心病实证多见于痰阻心脉、心血瘀阻，虚证多见于气阴两虚，虚实夹杂等，其中以气虚血瘀多见[7]。本研究利用文本挖掘方法从 66 896 篇“冠心病”文献中共提取到 173 个证候，其中以气虚血瘀型文献频次最高；其次为心血瘀阻、气阴两虚、痰湿阻肺、气滞血瘀；另外，心脉痹阻、心气虚、心阳虚、肾阴虚也较为常见，与以上学者的证型调查结果在某些方面有较高的相似性；传统中医学认为，冠心病病位在心，涉及到肺、肾；病理因素分为虚、实两大类，虚证有气虚、阳虚、阴虚，其中以气虚为主；实证有血瘀、气滞、痰浊，以血瘀因素为主，这也与本研究文本挖掘所得到的证型特征基本保持一致。

中药共提取到 237 个，以丹参文献频次最高，其次为黄芪、葛根、人参、红花；频次排列前 10 名的还有川芎、三七、麦冬、麝香。其中丹参、红花、川芎为活血化瘀之品，黄芪、人参具有补气之功，与以气虚血瘀为主要证候的药物挖掘结果相呼应。针对气虚血瘀证的定向挖掘结果显示治疗冠心病气虚血瘀证的常用中药基本属于治疗冠心病的常用中药之列。

以丹参、黄芪为例发现，丹参、黄芪既是文本挖掘得到的治疗冠心病的常用中药，又是定向挖掘得到的常用中药，在临床使用中中药丹参是治疗冠心病的活血化瘀之品，黄芪是补气升阳之药；现代药理研究显示，丹参酮ⅡA 具有抗动脉粥样硬化、缩小心梗面积、降低心耗氧量、保护心肌、抗心律失常等作用[8]，丹参素有清除氧自由基、调节钙平衡、保护线粒体和调节炎性细胞因子等功效，对心血管系统具体明细的保护作用[9]；黄芪的主要有效成分之一黄芪多糖具体有增强细胞免疫功能、抗肿瘤、免疫调节、降糖、抗氧化延缓衰老等功效[10]，两者的现代药理作用与各自的中药治疗功效相呼应，并且丹参、黄芪的现代药用制剂如丹参注射液、黄芪注射液在临床中治疗冠心病时广泛应用[11]，也从一个侧面证实了活血补气的药物在冠心病治疗中起着较重要作用。

[参考文献]

[1] Lopez A D, Mathers C D, Ezzati M, et al. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data [J]. Lancet, 2006, 367(9524): 1747.

[2] ZHENG G, JIANG M, HE X J, et al. Discrete derivative: a data slicing algorithm for exploration of sharing biological networks between rheumatoid arthritis and coronary heart disease [J]. Bio Data Min, 2011, 4: 18.

[3] 杨静，谭勇，郭洪涛，等. 基于文本挖掘技术的冠心病临床用药规律分析[J]. 中西医结合心脑血管病杂志，2011，9(11)：1281.

[4] 薛为民，陆玉昌. 文本挖掘技术研究[J]. 北京联合大学学报：自然科学版，2005，19(4)：59.

[5] 李立，周琦，郑光，等. 基于文本挖掘技术分析中成药、西药对慢性胃炎的治疗规律[J]. 中国实验方剂学杂志，2011，17(24)：228.

[6] LI S, ZHANG Z Q, WU L J, et al. Understanding ZHENG in traditional Chinese medicine in the contex of neuro-endocrine-immune network [J]. IET Syst Biol, 2007, 1(1): 51.

[7] 毕颖斐，毛静远，张伯礼，等. 基于文献的冠心病中医证型地域性分布特征研究[J]. 中医杂志，2012，53(3)：228.

[8] 李博，朱平先，吴清华. 丹参酮心血管药理作用及临床应用进展[J]. 江西医学院学报，2009，49(6)：126.

[9] 张宁，苏瑾，金磊，等. 丹参素心血管作用机制的研究概况[J]. 药学实践杂志，2009，27(6)：404.

[10] 何文涓，袁志坚，何晓升. 黄芪多糖的药理作用研究进展[J]. 中国生化药物杂志，2012，33(5)：692.

[11] 杨眉. 丹参、黄芪注射液的药理作用及临床应用[J]. 浙江中医药大学学报，2007，31(3)：379.

[责任编辑　邹晓翠]