·数据挖掘·

利用文本挖掘技术探索高血压病症状、证候以及用药规律

贺丹[1,2]，姜淼[2]，郑光[3]，张弛[2]，杨静[2]，赵宁[2]，沈姗姗[2]，边艳琴[2]，吕爱平[1,2*]
（1. 上海中医药大学 E 研究院，上海 201203；2. 中国中医科学院中医临床基础医学研究所，北京 100700；3. 兰州大学计算机信息学院，兰州 730000）

［摘要］ **目的**：利用文本挖掘技术探索高血压病中医证治方药规律。**方法**：在中国生物医学文献服务系统中收集治疗高血压的文献数据，采用基于敏感关键词频数统计的数据分层算法，挖掘高血压中医症状、证候以及用药规律。这些规律通过一维频次表及二维网络图进行展示。**结果**：高血压的最常见症状为头痛（2 650次），头晕（1 734次），证候以肝阳上亢为主，其次为阴虚阳亢和肝火亢盛，病变涉及肝、肾等脏腑。治法方面，为平肝潜阳，补益肝肾，并与活血化瘀法配合使用。汤药以天麻钩藤饮，钩藤饮以及半夏白术天麻汤为最常用。中药以天麻、钩藤应用最多。**结论**：文本挖掘能够比较客观地总结中医证治方药规律，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；高血压；证治方药
［中图分类号］ R287 ［文献标识码］ A ［文章编号］ 1005-9903(2014)19-0214-03
［doi］ 10.13422/j.cnki.syfjx.2014190214

Exploring Relationship Among Symptom, Pattern and Medication Regularity of Hypertension Based on Text Mining Technology

HE Dan[1,2], *JIANG Miao*[2], *ZHENG Guang*[3], *ZHANG Chi*[2], *YANG Jing*[2], *ZHAO Ning*[2], *SHEN Shan-shan*[2], *BIAN Yan-qin*[2], *LV Ai-ping*[1,2*]
(1. *E-research Institute of Shanghai University of Traditional Chinese Medicine, Shanghai* 201203, *China*; 2. *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*; 3. *School of Information Science and Engineering Technology, Lanzhou University, Lanzhou* 730000, *China*)

［**Abstract**］ **Objective**: This investigation aimed at applying data mining technology to explore the relationship among symptoms, patterns and medication regularity of hypertension. **Method**: All the relative literatures were collected from sinomed database, then subject to data slicing algorithm to mine out the relationship among symptoms, patterns and medication regularity of hypertension. **Result**: The most common symptoms of hypertension are headache (2 650), dizziness (1 734), and the patterns are manifested as liver Yang hyperactivity, Yin deficiency and Yang excess as well as upward liver fire, the most injured organ are liver and kidney. Treating methods would be to suppress the upward liver yang and strengthen liver and kidney which should be accompanied by activating Qi and Blood. The most common applied herbs are Gastrodiae Rhizoma and Uncariae Ramulus Cum Uncis. **Conclusion**: Text mining is a branch of data mining and a practicable technology which can be used in the field of traditional Chinese medicine (TCM). Through putting this technology in the research of TCM clinical practice, both doctors and patients could benefit from it.

［**Key words**］ text mining; data slicing algorithm; hypertension; medication regularity

［收稿日期］ 20131018(003)
［基金项目］ 国家中医药管理局中医行业专项课题（201207012）
［第一作者］ 贺丹，博士在读，讲师，从事疾病证候分类研究，Tel：010-64014411-3401，E-mail：fm873@126.com
［通讯作者］ *吕爱平，博士，研究员，从事疾病证候分类研究，Tel：010-64067611，E-mail：lap64067611@126.com

随着人口老龄化的到来,高血压的发病率呈现逐年增加的趋势,严重影响着人们的生活质量与健康水平,西医西药在血压控制上成果显著,但在改善患者各种临床症状方面,中医中药凸显了它的优势,传统中医学在很早就有高血压病相关的症状、证候描述和治疗方法的记载,现代中医在高血压的理论和临床研究方面也取得了很大的进展,通过临床辨证分型施治,使患者获得了较好的疗效[1],并且在降压、消除症状、改善生化指标等方面都积累了一定的临床经验[2]。然而,各种文献报道浩如烟海,如何通过收集现有文献,总结临床治疗的实证经验用于临床的再指导是当前科研工作者的使命。文本挖掘是数据挖掘技术的一个分支,其在中医药领域的应用已日渐成熟,通过检索相关文献,发现已有将文本挖掘技术应用于高血压疾病中成药及西药用药规律的分析[4],并且取得了一些研究成果,但应用文本挖掘技术对高血压疾病的各种临床要素之间的规律进行深度探讨尚属首例,本文将借助这一技术手段对高血压疾病的症状,证候和用药规律进行挖掘,以期得出一些有价值的结论供临床医生参考。

1 材料与方法

1.1 文本数据收集 登录中国生物医学文献数据库(Chinese BioMedical Literature Database, CBM, http://sinomed.cintcm.ac.cn/index.jsp),在"缺省"状态下,以"高血压病"为关键词进行检索,共得到文献175 011篇(检索日期:2013年9月24日),并选择"详细"和"显示全部"的显示格式,以获得每篇文献的流水号、标题、摘要、主题词等信息备用。

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。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。假设每一篇文献的贡献度是相同的,一篇文献中重复出现的关键词,只需要计算一次,据此构建算法进行数据清洗工作[3]。清洗完毕后的数据,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,进行可视化呈现。抽出不同频次的关键词对,用Cytoscape 2.8软件进行可视化处理,形成可视化网络图,然后结合专业知识进行解析,发现不合理的结果,即回溯原文献数据集,如果是噪音,仍按算法进行噪音清洗,直至噪音降到满意为止。最后的结果可视化成图,结合专业知识进行解析。

1.3 数据一次清洗 根据"文本数据处理"中生成的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。为便于处理,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)表示。为确保下载数据真实,需要对原文献进行回溯分析,相同的关键词存在着在一篇文献的标题和摘要重复出现的情况。在文本挖掘中,每一篇文献的贡献度是相同的,因此,对于一篇文献中重复出现的关键词,只需要计算一次。据此,需要对重复文献进行删除,即数据清洗[4]。

1.4 数据挖掘处理 通过返查原文献发现,在同一篇文献中出现的关键词,在关键词这一抽象层面上,部分反映整篇文章的信息。对某篇具体的文献来说,相关的关键词之间存在着"共同出现"这一基本事实。这种共同出现不是随机的,而是蕴含有一定的意义[5],尤其对于高频协同出现的关键词对,在一定的程度上反映了科研工作者的关注程度。更重要的是,针对目前的文本挖掘技术来说[6],这些协同出现的关键词也是很好的分析素材。基于上述分析,首先构造针对每一篇文献共同出现的关键词对,然后构造算法来实现这一工作。经过计算,得到名为DN_pairs的数据表。研究发现数据表DN_pairs中存在着大量相同的关键词对,这些重复的数据对于进一步分析来说,大部分属于噪音,对此,通过构造算法将相同的关键词对进行合并处理,只保留它们出现的频数,得到了名为DN_pairs_frequency的数据表,在这个数据表内,所有的关键词对,都只出现一次并且都有一个对应的频数(Frequency)。

1.5 数据二次清洗 利用专业知识对数据进行评估后会发现,针对特定的疾病,可能仍然存在噪音问题。这些噪音不再是关键词的简单重复,而是独立于专业知识以外的噪音问题。对此,针对特定问题,对数据进行二次清洗。这些噪音主要是由自然语言的二义性和表达方式的多样性所产生。对于这类问题,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。

2 文本挖掘结果的评价和分析

挖掘结果:通过整理挖掘结果,可以得出高血压病的治疗用药按频次高低依次为天麻,钩藤,半夏等,常见症状如图1,依次为头痛,头晕,呕吐,恶心和胸闷等。而通过二维频次构建的网络图可以看出头痛,头晕,呕吐,恶心和胸闷等症状之间的相互关系(图2的中间部分),通过回溯原文献发现,这些症状与高血压疾病过程中最常见的症状高度吻合。

图2为挖掘出的症状,证候以及治疗用药三者之间的网络关系图,图的中间部分为高血压病的常见证候:肝阳上亢,肝风内动,气虚血瘀,痰浊中阻等。与肝阳上亢相对应的常见症状为头晕,头痛,失眠,烦躁,而与肝风内动相吻合的症状则有抽搐,乏力,恶心,呕吐,同时挖掘出的治疗用药以及构成的方剂也与症状,证候高度契合,例如治疗肝风内动的镇肝熄风汤,主要成分有牛膝,龙骨,牡蛎,白芍,麦冬以及甘草;治疗气虚血瘀的补阳还五汤,其组成成分为黄芪,当归,赤芍,地龙,川芎,红花;还有化痰祛湿的半夏白术天麻汤,其组成为半夏,白术,天麻,茯苓,泽泻等。图2的底部用深色圆圈标注的就是构成特定方剂的药物组成,而方剂与治法之间的对应关系也通过深色箭头予以了标注。

3 讨论

文本挖掘技术应用于中医药已有多年历史,在将此技术

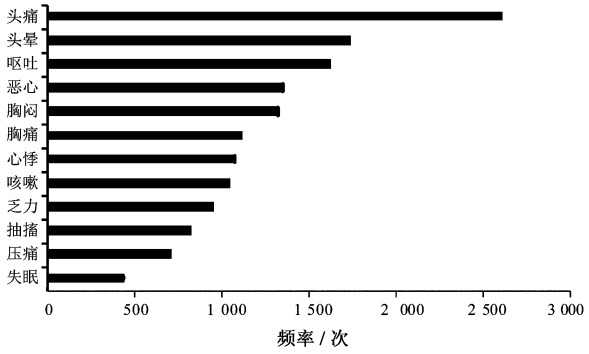

图 1　高血压病常见临床症状文献频率由高到低的排列

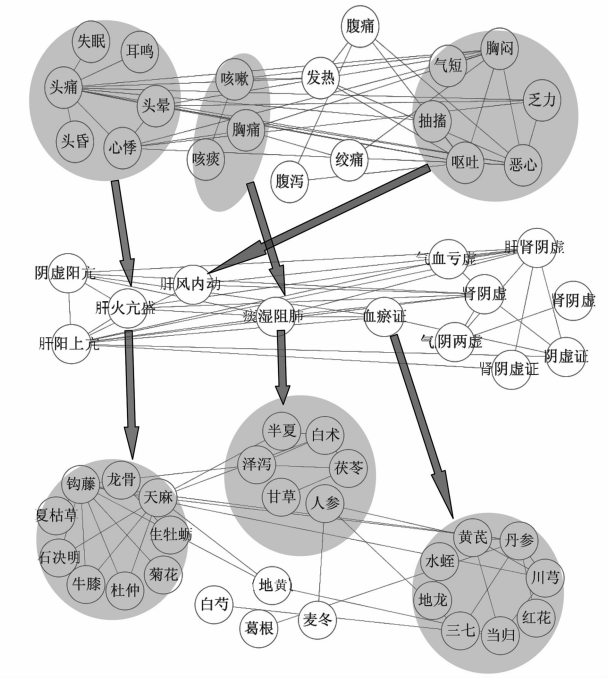

图 2　高血压病常见症状,证候以及用药关系

反复不断的与中医药理论结合的过程中,作者更深的体会到中医药知识与现代化科技结合的必要性,而且在经过论证之后发现,中医药现代化是可以实现的,只要不懈的探索,不断的尝试,充分利用现代科技知识,可以更深入的

挖掘中医药宝库的内涵。本文以高血压病为挖掘对象,通过对175 011篇文献进行症状,证候以及治疗方法的挖掘,作者发现其结果与临床实际高度吻合。

在挖掘过程中,笔者发现高血压病的合并病种类繁多,例如各类脑血管病,心血管疾病,糖尿病等,因而进一步了解高血压病在中医中药治疗下的规律,指导临床医生更准确的进行治疗,可谓意义重大。各项流行病学调查证实,我国面临着高血压的巨大负担。原发性高血压病(简称高血压病)因其常伴眩晕、头痛等症状,故中医学大多数文献将其归属于眩晕、头痛等病证范畴,并认为其病机多与肝有关,其中肝阳上亢、肝肾阴虚和阴虚阳亢是主要临床证型,治疗也从平肝潜阳或滋阴潜阳论治。有很多传统的方剂在高血压临床治疗上取得了较好的疗效,在现代医学中应用于各种症状类似的疾病,如血管神经性头痛、颈性眩晕、老年舞蹈病、脑血管意外等。在各相近方剂之间,对治疗高血压在功能主治方面既有相似之处,又有区别,临床应用各有侧重点[1]。本文利用文本挖掘技术发现中医药治疗高血压的过程中,其症状,证候以及用药之间存在一定的规律。通过挖掘作者获得一些具体的发现,例如,中药治疗高血压频次最高的依次是天麻,钩藤,半夏,而方剂使用频率最高的是天麻钩藤饮,钩藤饮,半夏白术天麻汤,常见证候则为肝风内动,肝火亢盛,并且这三者之间相互印证,高度吻合。王晓聪[7]对 31 例中青年高血压病病人,采用口服天麻钩藤饮加味煎剂治疗,连服 3 个月,随诊 3 个月,有效率为 90.3%。郭维琴[8]分 6 型论治高血压病,肝火上扰型治以平肝潜阳、清泻肝火,方用天麻钩藤饮加减;痰湿中阻型治以燥湿化痰、健脾和胃,方用半夏白术天麻汤加减;瘀血阻络型治以活血化瘀,方用通窍活血汤加减。王丽艳[9]分 5 型论治,肝阳上亢型宜平肝潜阳,方用龙胆泻肝汤等;阴虚阳亢型宜滋阴平肝,方用天麻钩藤饮、杞菊地黄汤等。由此可见,文本挖掘技术能很好的挖掘出中医药治疗与症状,证候之间的相互关系,并且与临床实际相一致,值得在未来的研究中继续推广运用。

[参考文献]

[1] 白宇明,魏国义,郝近大. 常用治疗高血压相近方剂对比分析[J]. 中国药房,2010,21(20):1885.

[2] 谢璞. 中医治疗高血压研究近况[J]. 中西医结合心脑血管病杂志,2008,6(5):566.

[3] 郭洪涛,郑光,赵静,等. 基于文本挖掘分析甲型 H1N1 流感的中医药治疗特色[J]. 世界科学技术——中医药现代化,2011,13(5):772.

[4] 王丽颖,郑光,郭洪涛,等. 基于文本挖掘技术的高血压病中成药与西药用药规律分析[J]. 中华中医药杂志,2013,28(1):60.

[5] Andrea Campagna, Rasmus Pagh. Finding associations and computing similarity via biased pair sampling [C]. Miami, Florida: 2009 Ninth IEEE International Conference on Data Mining, 2009:61.

[6] Wu X G, Huang Z D, Stamler J, et al. Changes in average blood pressure and incidence of high blood pressure 1983-1984 to 1987-1988 in four population's cohorts in the people's republic of China [J]. J Hypertension, 1996,14(11):1267.

[7] 王晓聪. 天麻钩藤饮加味治疗中青年高血压病疗效观察[J]. 中医药临床杂志,2004,16(2):109.

[8] 秦建国,王亚红,梁晋普,张莹. 郭维琴治疗高血压病经验[J]. 中医杂志,2007,48(7):586.

[9] 王丽艳. 高血压病辨证论治探要[J]. 中华中医药学刊,1999,18(5):32.

[责任编辑　邹晓翠]