DOI:CNKI:11 - 3495/R. 20110314. 0945. 016

·文献研究·

基于文本挖掘技术的肥胖和高脂血症处方规律研究

黄允瑜[1],陈慕芝[2],郑光[3],郭洪涛[3],姜森[3],赵宏艳[4*]
(1. 北京中医药大学东直门医院,北京 100700;
2. 新疆维吾尔自治区中医医院,乌鲁木齐 830000;
3. 中国中医科学院中医临床基础医学研究所,北京 100700;
4. 中国中医科学院中医基础理论研究所,北京 100700)

[摘要] **目的**:利用数据挖掘技术探索肥胖和高脂血症的中药用药规律。**方法**:在中国生物医学文献数据库中收集中医药治疗"肥胖"和"高脂血症"的文献,建立 Access 数据库,运用 SQL 对数据进行处理,Cytoscape 2.7 软件对数据分析结果进行可视化,分析中医药治疗肥胖和高脂血症的用药规律。**结果**:治疗肥胖的常用药物为丹参、山楂、泽泻、黄芪、大黄、茯苓、白术、半夏;其治疗的核心药物是黄芪、白术、大黄。治疗高脂血症的最常用药物为丹参、山楂、泽泻、黄芪、大黄、何首乌;其治疗的核心药物是丹参和山楂。**结论**:中医治疗肥胖和高脂血症的处方虽有相同之处,但是其治疗的核心却大不相同。文本挖掘分析技术在中医临床用药规律研究中具有良好的应用前景。

[关键词] 肥胖;高脂血症;文本挖掘;中药;配伍规律
[中图分类号] R242 [文献标识码] A [文章编号] 1005-9903(2011)09-0236-03

Medicinal Principle Exploration of Treating Obesity and Hyperlipidemia with Chinese Herbal Medicine Based on Text-mining Approach

HUANG Yun-yu[1], *CHEN Mu-zhi*[2], *ZHENG Guang*[3], *GUO Hong-tao*[3], *JIANG Miao*[3], *ZHAO Hong-yan*[4*]
(1. *Dongzhimen Hospital Affiliated to Bejing University of Chinese Medicine*, *Beijing* 100700, *China*;
2. *Traditional Chinese Medical of Xinjiang Uygur Autonomous Region*, *Wulumuqi* 830000, *China*;
3. *Institute of Basic Research in Clinical Medicine*, *China Academy of Chinese Medical Sciences*, *Beijing* 100700, *China*;4. *China Academy of Chinese Medical Sciences*, *Beijing* 100700, *China*)

[**Abstract**] **Objective**: To explore the medication principle of treating obesity and hyperlipidemia with Chinese herbal medicine by a novel text mining approach. **Method**: We proposed a data mining algorithm to retrieve simple and meaningful networks from large data sets of obesity and hyperlipidemia. We transfered XML type data sets to the structured database of Microsoft ® SQL ® 2 000 and visualized them into different graphs by

[收稿日期] 2010-12-30
[基金项目] 国家自然科学基金青年基金项目(30902003);中国中医科学院自主选题项目(Z0134);中国中医科学院自主选题项目(YZ-0701)
[第一作者] 黄允瑜,主治医师,从事中医肾病、内分泌、中医营养学临床工作,在职博士研究生,Tel:010-84013290,E-mail:huangyunyu1978@126.com
[通讯作者] *赵宏艳,医学博士,从事中医药免疫药理学研究、中药复方作用机制研究,Tel:010-64014411-2556,E-mail:zhaohongyan1997@163.com
[网络出版时间] 2011-03-14 09:45

software Cytoscape version 2.7 in order to detect the potential medication principles. **Result**: The most common used Chinese herbal medicine applied for treating obesity commonly proved to be Radix Salviae Miltiorrhizae, Fructus Crataegi Pinnatifidae, Rhizoma Alismatis, Radix Astragali Mongolici, Radix et Rhizoma Rhei Palmati, poria, Rhizoma Atractylodis Macrocephalae, Rhizoma Pinelliae; the core couplet medicine is Radix Astragali Mongolici, Rhizoma Atractylodis Macrocephalae, Radix et Rhizoma Rhei Palmati; the most common used herbal medicines applied for treating hyperlipidemia are Radix Salviae Miltiorrhizae, Fructus Crataegi Pinnatifidae, Rhizoma Alismatis, Radix Astragali Mongolici, Radix et Rhizoma Rhei Palmati, Radix Polygoni Multiflori; the core couplet medicine is Radix Salviae Miltiorrhizae and Fructus Crataegi Pinnatifidae. **Conclusion**: although the primary prescriptions for treating obesity and hyperlipidemia share some common herbs, the cores of the two formulae are significantly distintive, which indicates that the pathogenesis of the two diseases are different according to the theory in Chinese medicine. The novel text mining technology could be a promising approach in the medication principle detecting research of Chinese medicine.

[**Key words**] obesity; hyperlipidemia; text mining; Chinese herbal medicine; medication principle

文本挖掘技术是在数据挖掘的基础上针对文本开发的一种信息提取分析技术[1]。应用文本挖掘技术可以智能地从信息库中检索出符合用户需求的信息,还可以从文本数据中梳理、发现和提取其中隐含的知识并且形成用户可理解的信息知识。中医治疗疾病讲究的是理法方药,中药作为中医治病的物质基础,在文献中大量的被刊载。如果能对其进行全面的分析和整理,将有助于发现疾病治疗的核心规律,为进一步提高中医临床治疗效果、深入开展中医科研提供依据。

多项流行病学调查报告指出目前中国的超重和肥胖患病形势严峻,学龄儿童、成年人、老年人均呈现出增长速度加快的形势,应该尽早开展肥胖的防治与研究[2-4]。

中医药在肥胖和高脂血症的治疗中也发挥了应有的作用,基于现有大量相关文献,应用文本挖掘技术对中医治疗肥胖和高脂血症文献报道所使用的中药进行挖掘和分析,是全面整理总结其配伍经验的有益探索[5-6]。

1 材料与方法

1.1 文献选取方法 登录中国生物医学文献数据库(英文全称:Chinese bio medical literature database,简称 CBM)在主题检索下以“肥胖”、“高脂血症”为关键词进行检索。共得到文献34 611篇,其中肥胖15 788篇,高脂血症18 823篇,(检索日期:2010 年 6 月 4 日)。显示格式中选择“详细”和“显示全部”,以显示每篇文献的流水号、标题、摘要、主题词等信息。

1.2 文献处理方法 将收集来的相关文献数据,按照下载的先后顺序,分别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,对 **1.1** 中下载的非结构化的 TXT 文本数据进行信息提取,所提取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。

1.3 数据挖掘以及分析 根据 **1.2** 中生成的 Access 数据库,将基础数据导入 SQL 中,以“table_initial”为表名称,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,针对“序号”和“机标关键词”进行处理。

首先,从初始数据表(Table_Initial)中运用“关键词组合算法”,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,构造针对每一篇文献共同出现的关键词对。最后输出到“关键词对数据表”(DN_pairs)中。

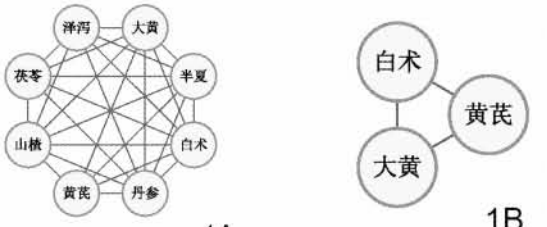
针对 DN_pairs 的数据表。通过构造“关键词对频数统计”的算法将其中相同的关键词对进行合并处理,只保留他们出现的频数。将结果输出到名为 DN_pairs_frqcy 的数据表中。表中所有的关键词对都只出现 1 次,且都有一个出现的频数(Frequency)。

1.4 数据的可视化 根据 **1.3** 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，分别得到治疗各个疾病的中药用药网络图，从中选出有代表性的两层进行分析讨论。

2 结果

2.1 肥胖常用中药及核心组成　从图 1A 中可以看出，治疗肥胖的常用药物为丹参、山楂、泽泻、黄芪、大黄、茯苓、白术、半夏；从图 1B 中可以看出，其治疗的核心药物是黄芪、白术、大黄。

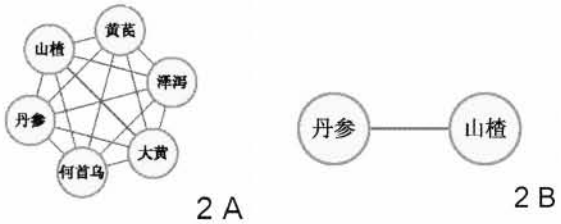


A. 第三层　　B. 最高层

图 1　治疗肥胖常用中药频数关系

图中圆圈内药物为检索文献中治疗肥胖所涉及到的频数较高的中药。

2.2 高脂血症常用中药及核心组成　从图 2A 中可以看出，治疗高脂血症的最常用药物为丹参、山楂、泽泻、黄芪、大黄、何首乌；从图 2B 中可以看出，其治疗的核心药物是丹参和山楂。



A. 第三层　　B. 最高层

图 2　治疗高脂血症常用中药频数关系

图中圆圈内药物为检索文献中
治疗高脂血症所涉及到的频数较高的中药

3 讨论

中医对肥胖的诊治有着悠久的历史，在《灵枢·卫气失常》篇中有肥胖的最早分型记录，将肥胖者分为“脂人”、“膏人”、“肉人”3 种类型。现代的中医学者也十分重视对肥胖的诊治。因为肥胖患者多合并有高脂血症，为了有效地区分两种疾病的中药治疗组方，提高治疗效果，应用文本挖掘技术进行潜在的组方分析是一种新的思路和方法。由结果可以看出，丹参、山楂、泽泻、黄芪、大黄在两种疾病的治疗中都高频率地出现，说明此两种疾病的治疗有其相同之处。而两者的核心治疗药物却大相径庭，肥胖以黄芪、白术、大黄为核心治疗药物，体现出脾胃在中医治疗肥胖中的重要性，补气健脾通腑为主要的治法。高脂血症以丹参和山楂为核心治疗药物，体现出饮食精气淤阻脉络是其病机的核心，正如《经脉别论》所云：食气入胃，散精于肝，淫气于筋。食气入胃，浊气归心，淫精于脉。如果精气不能顺利输注至脏腑、肌肉、皮毛，则淤滞在脉络，导致气血运行不畅而变证百出。因此化饮食，健脾胃，行结气，消瘀血也为其主要的治法。由此可见，肥胖和高脂血症的治疗有着本质区别，但在实际工作中两种疾病常合并出现，导致两种疾病中均高频率的出现同样的药物，因此应用文本挖掘技术寻找治疗疾病的核心组方是可行的也是十分有必要的。

本研究的目标仅仅集中于肥胖和高脂血症的用药规律研究，从这个研究中延伸至证候辨识规律、药证对应规律、以及治疗新药的研发研究，将是我们下一步的工作目标，也将是更有意义的工作。总之，运用现代数据挖掘、文本挖掘技术，在海量文献的基础上探索规律，从中得到新的启发或线索，从而获得新的知识，将为中医药基础研究、新药开发研究、临床实践提供有益参考与崭新思路[7]。

[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)[M]. montreal, AAAI Press, 1995: 112.

[2] 王文绢，王克安，李天麟，等. 中国成年人肥胖的流行特点研究：超重和肥胖的现患率调查[J]. 中华流行病学杂志，2001，22(2)：129.

[3] 陈捷，赵秀丽，武峰，等. 我国 14 省市中老年人肥胖超重流行现状及其与高血压患病率的关系[J]. 中华医学杂志，2005，85(40)：2830.

[4] 马军，吴双胜. 中国学龄儿童青少年超重肥胖流行趋势分析[J]. 中国学校卫生，2009，30(3)：195.

[5] 查青林，余俊英，余飞，等. 基于代谢相关 MeSH 词文本挖掘分析治疗咳嗽中药五味分类的生物学特征[J]. 中国中医基础医学杂志，2010，16(7)：616.

[6] 谭勇，郭洪涛，郑光，等. 利用文本挖掘技术探索中医药治疗疾病的用药规律[J]. 世界科学技术——中医药现代化，2010，12(5)：823.

[7] 姜淼，查青林，郭玉明，等. 基于中医药科学思维的生物医学创新研究思路与方法[J]. 中国中医基础医学杂志，2010，16(5)：354.

[责任编辑　何伟]

网络出版地址：http://www.cnki.net/kcms/detail/11.3495.R.20110314.0945.016.htm