·数据挖掘·

基于文本挖掘探索心律失常药物治疗规律

刘学文[1],蔡峰[2],郑光[1,3],姜淼[1],吕爱平[2*]
(1. 首都医科大学潞河教学医院中医科,北京 101100;
2. 中国中医科学院中医临床基础医学研究所,北京 100700;
3. 兰州大学应用数学与统计学院,兰州 730000)

[摘要] **目的**:利用文本挖掘技术探索心律失常治疗规律,以期为临床经验总结提供参考。**方法**:利用数据挖掘技术,结合可视化,首先在 CBM 数据库中收集治疗心律失常文献数据,采用基于敏感关键词频数统计的数据分层算法,结合文献回溯、人工阅读降噪,挖掘心律失常的中西药治疗规律。这些规律通过一维频次及二维的网络图进行显示。**结果**:心律失常常用西药为肾上腺素、尿激酶、地高辛等,常用中成药为稳心颗粒、复方丹参滴丸等,常用中药为地黄、丹参、人参、天冬、炙甘草等,其中肾上腺素、稳心颗粒、桂枝与其他药物关联最密切,中西药联用时常用中成药为稳心颗粒。**结论**:文本挖掘技术能够比较客观地总结心律失常药物治疗规律,为临床经验总结提供有益的探索与参考。

[关键词] 文本挖掘; 数据分层算法; 心律失常; 证候; 药物治疗
[中图分类号] R287 [文献标识码] A [文章编号] 1005-9903(2013)17-0350-05
[doi] 10.11653/syfj2013170350

Explore Medication Principle of Treating Arrhythmia Based on Text Mining

LIU Xue-wen[1], *CAI Feng*[2], *ZHENG Guang*[1,3], *JIANG Miao*[1], *LV Ai-ping*[2*]
(1. *The Luhe Teaching Hospital of the Capital Medical University*, *Beijing* 101100, *China*;
2. *Institute of Basic Research in Clinical Medicine China Academy of Chinese Medical Sciences*, *Beijing* 100700, *China*;
3. *Applied Mathematics and Statistics Academy*, *Lanzhou University*, *Lanzhou* 730000, *China*)

[**Abstract**] **Objective**: To explore the medication principle of treating arrhythmia based on text mining and to provide a reference for clinical experience summary **Method**: We used data mining technology combined with visualization. Based on the literature data of treatment of arrhythmia collected in the CBM database, we explored the medication principle of treating arrhythmia, both traditional Chinese medicine (TCM) and Western medicine. The means we used include frequency statistical data based on sensitive keywords hierarchical algorithm, retro-read and manually noise reduction. These laws displayed by the frequency of one-dimensional and two-dimensional network diagram. **Result**: Common Western medicine for treating arrhythmia are Epinephrine, urokinase, digoxin, etc, common Chinese medicine are Wenxin grain, compound Danshen Dripping pill, etc, common TCM are Rehmanniae, Salvia, Ginseng, Aspartame, Licorice, etc. Among these, Epinephrine, Wenxin grain, Cassia twig are most closely related to other medicine. Commonly used Chinese medicine is Wenxin

[收稿日期] 20130628(020)
[基金项目] 国家科技部创新方法学专项项目(2008IM020900);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)
[第一作者] 刘学文,博士后,主治医师,从事中医老年病防治研究,Tel:010-69543901,E-mail:tjtcmliu@126.com
[通讯作者] *吕爱平,博士,研究员,从事基于中医证候分类的中医药临床评价方法学研究、中医药防治自身免疫性疾病的机制研究,Tel:010-64067611,E-mail: lap64067611@126.com

Grain when Chinese medicine combined with Western medicine. **Conclusion**: Text mining technology can be used for summary of medication principle of treating arrhythmia, thus could provide useful exploration and reference for clinical experience summary.

[**Key words**] text mining; data stratification algorithm; arrhythmia; syndrome; drug treatment

心律失常(cardiac arrhythmia)是指心脏冲动的频率、节律、起源部位、传导速度或激动次序的异常。药物治疗主要包括抗心律失常药物治疗和抗凝药物治疗等,抗快速心律失常药物分为4大类,缓慢心律失常应用增强自律性和加速传导的药物[1]。中医药治疗心律失常的效果也被相关临床研究证实[2-5],机制研究涉及离子通道、离子泵、电生理、能量代谢、细胞因子等方面[6-10],相关文献大量存在于现有数据库中。本研究借助不断成熟的数据挖掘技术[11-13],结合可视化、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分析中西药物治疗心律失常的规律。

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(英文全称:Chinese BioMedical Literature Database,简称CBM,网址http://sinomed.cintcm.ac.cn/index.jsp),在"缺省"状态下,以"心律失常"为关键词进行检索,共得到文献68 773篇(检索日期:2012年4月13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 将收集数据按照下载先后顺序,整合到一个平面文件(后缀TXT)中,以ANSI编码格式保存,并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"文本数据收集"中下载的非结构化TXT文本数据进行信息提取,并以格式化形式存储在大型关系型数据库(Microsoft SQL Server,以下简称SQL)中。其中,提取的信息主要为"机标关键词"。

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。

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。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。这种共同出现不是随机的,而是蕴含有一定的意义[12],尤其对于高频协同出现的关键词对,在一定的程度上,这些词对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[13],这些协同出现的关键词,也是很好的分析素材。

基于上面的分析,第一步,就是构造针对每一篇文献共同出现的关键词对。就此,构造了图1的算法,来实现这一工作。经过图1算法的计算,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,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,构造了图2中的算法来实现。经过图2中算法的处理,得到了名为DN_pairs_frqcy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)。

```
USE Table_Initial
FOR each PMID
  k =Number_of_DescriptorName(PMID)
  j = 1
  FOR DescriptorNames(i) (i =, 1, 2, ..., k)
    DO while j ≤ k
      DescriptorNames_Pair=DescriptorNames(i)+
          DescriptorNames(j)
      j = j + 1
      OUTPUT DescriptorName_Pair INTO
          table DN_pairs
    ENDDO
    j = 1
  ENDFOR
ENDFOR
```

图1 构建关键词对程序算法

1.5 数据二次清洗 经过专业知识对图2(DN_pairs_frqcy)中的数据进行评估后发现,针对特定的疾病,图2中仍存在噪音问题。这些噪音,不再是关键词的简单重复,而是相对于专业来说的噪音问题。

```
USE table DN_pairs
k =max_line_number
DO while k ≥ 1
GO top
  FOR DescriptorName_Pair(1) //The 1st pairs in CHD_RA
    COUNT its Frequency
  EndFor
  OUTPUT DescriptorName_Pair, Frequency INTO table
         DN_pairs_Frqncy
  DELETE all DescriptorName_Pair(1) from table
         DN_pairs
  k =max_line_number
ENDDO
```

图 2　合并筛检关键词对程序算法

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

根据“数据二次清洗”中得到的数据表 DN_pairs_frqcy,按照频数由高到低的顺序对关键词进行排序,确定其相互关联程度,并在此基础上,对某一频数以上(即大于等于关系)的数据进行切片处理。对于数据的可视化工作,在 Cytoscape 2.8 软件中进行,经过处理,得到治疗心律失常的中药、中成药、西药、证候以及中西药联合应用等结果图。然后,对分析结果进行原文献溯源性阅读,进一步评价分析挖掘结果、降噪,现择其部分结果进行呈现分析。

3　结果

3.1　西药文本挖掘结果

3.1.1　西药文本挖掘一维频数结果　文本挖掘显示,心律失常相关西药共 201 种,使用前 10 名由高到低依次为:肾上腺素、尿激酶、地高辛、阿替洛尔、门冬氨酸、盐酸胺碘酮、胰岛素、阿司匹林、卡托普利、辛伐他汀(见图 3)。

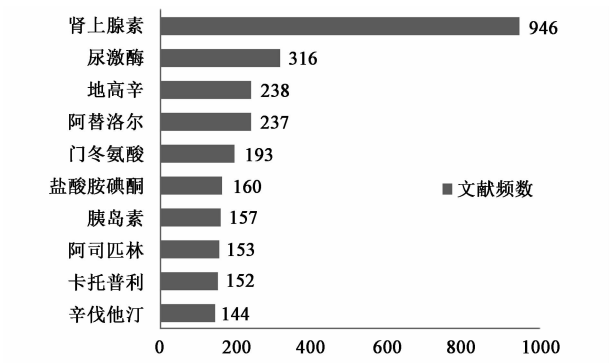

图 3　心律失常西药文献前 10 名频数

3.1.2　西药文本挖掘二维网络结果　文本挖掘显示,心律失常相关西药组合共 372 种,选取大于等于 6 的文献频数构建网络图。如图 4 所示,网络中药物的显示度和节点的大小正相关。肾上腺素、卡托普利、维生素 C、辛伐他汀、阿司匹林、地高辛、门冬氨酸、尿激酶、维生素 A、阿替洛尔、辅酶 Q10、螺内酯、盐酸胺碘酮、硝酸异山梨酯等等具有较高显示度。

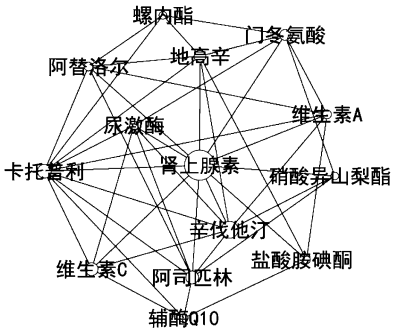

图 4　心律失常西药网络(频数≥6)

3.2　中成药文本挖掘结果

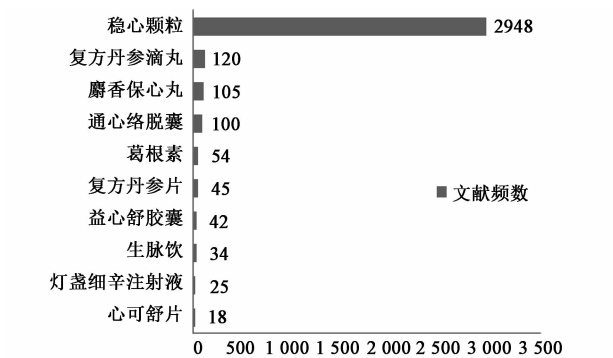

图 5　心律失常中成药文献前 10 名频数

3.2.1　中成药文本挖掘一维频数结果　文本挖掘显示,心律失常相关中成药共 47 种,使用前 10 名由高到低依次为:稳心颗粒、复方丹参滴丸、麝香保心丸、通心络胶囊、葛根素、复方丹参片、益心舒胶囊、生脉饮、灯盏细辛注射液、心可舒片。

3.2.2　中成药文本挖掘二维网络结果　文本挖掘显示,心律失常相关中成药组合共 47 种,选取≥3 的文献频数构建网络图。如图 6 所示,网络中药物的显示度和节点的大小正相关。稳心颗粒、通心络胶囊、地奥心血康胶囊、丹参片、葛根素、复方丹参滴丸等具有较高显示度。

3.3　西药、中成药联用文本挖掘结果　文本挖掘显示,心律失常相关西药、中成药联用组合共 54 种,选取≥1 的文献频数构建网络图,如图 7 所示,联用时常用中成药为稳心颗粒,常用西药为肾上腺素、门冬氨酸、阿替洛尔、酒石酸美托洛尔、地高辛、阿司匹林。

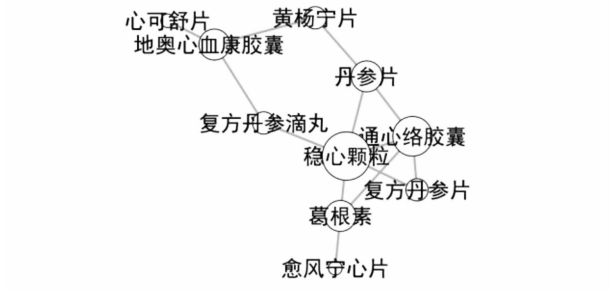

图 6　心律失常中成药网络(频数≥3)

图 7　心律失常西药、中成药联用网络

3.4　中药文本挖掘结果

3.4.1　中药文本挖掘一维频数结果　文本挖掘显示,心律失常相关中药共 223 种,使用前 10 名由高到低依次为:地黄、丹参、人参、天冬、炙甘草、麦冬、黄芪、附子、儿茶、黄连(见图 8)。

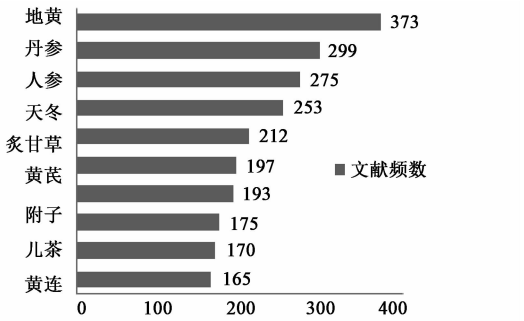

图 8　心律失常中药文献前 10 名频数

3.4.2　中药文本挖掘二维网络结果　文本挖掘显示,心律失常相关中药组合共1 851种,选取大于等于 20 的文献频数构建网络图。如图 9 所示,网络中药物的显示度和节点的大小正相关。桂枝、地黄、麦冬、附子、川芎、黄芪、丹参、人参、炙甘草、红花、当归、三七、五味子、党参、葛根、苦参、细辛等等具有较高显示度。

4　讨论

文本挖掘(text mining)技术是以统计数理分析、计算语言学为理论基础,服务于医药、生物、文献研究等学科的新兴的交叉学科[14]。应用于中医药领域,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。

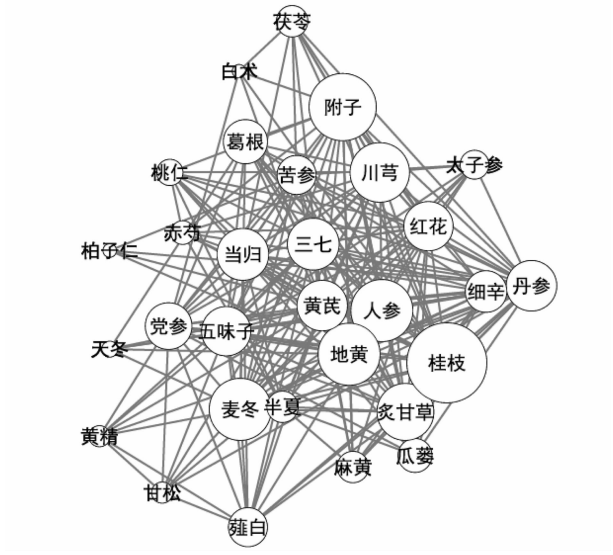

图 9　心律失常中药网络(频数≥20)

根据中医理论或专业知识,计算机挖掘中医药文献库与生物医学信息库的非关联知识发为中西医结合研究提供新的思路和途径,并且结果更加客观,可重复性强[15]。

本文应用文本挖掘快捷、全面系统的总结了心律失常临床实践中的中西用药、中医证候等,为临床医生提供了参考。如稳心颗粒为中成药单用和连用中最常见的药物,其功效为益气养阴,活血化瘀,定悸复脉,不仅具有大量临床证据[16-17],基础实验亦证明其确切机制[18-19]。近年来研究结果显示,血管紧张素转换酶抑制剂(ACEI)、血管紧张素受体阻滞剂(ARB)、他汀类药物等心律失常上游药物也有抗心律失常作用[20],本研究结果与此有相符之处。中药文本挖掘二维网络结果表明,经常联用的中药为丹参、黄芪、麦冬、人参、红花、麻黄、附子、细辛,体现中医药针对心律失常采用益气养阴,温补心阳,活血化瘀通脉的治法与功效[21-22],与证候文本挖掘结果亦相符。

但毕竟计算机计算的结果会存在实际偏差,需要结合回溯原文,借助专业知识进行分析判别,这对于图例的理解十分重要,例如:在中药的使用方面,一维频次较高的中药反而在二维网络中并不突出,结合文献资料可以判定二维网络中较高显示度的中药更符合实际。因此在创新知识点的发现上,仍然依赖分析者的专业判断,而尽量提高计算机自动化分析能力,降低噪音,逐步减少人工工作量是对文本挖掘技术提出的下一步要求。

总之,文本挖掘技术结合可视化可以快捷、全面系统地总结心律失常临床实践中的用药情况与规

律,为临床医生提供参考,而结合人工回溯原文献,借助专业知识进行分析评判,可以提高挖掘的深度和精度,使挖掘的结果更客观,从而为知识更新、新的临床指南升级提供证据。

[参考文献]

[1] 张澍. 中国心律失常学科进展[J]. 中华医学信息导报,2012,27(4):8.

[2] 魏潇,娄彬,徐重白. 中医药治疗心律失常研究进展[J]. 中国中医急症,2010,19(12):2110.

[3] 阳永扬,徐彤彤. 解郁定悸汤治疗中青年妇女心律失常[J]. 中国实验方剂学杂志,2012,21(18):300.

[4] 雷智锋,刘影. 传统方剂在心律失常治疗中的应用及研究进展[J]. 世界中西医结合杂志,2011,6(6):533.

[5] 张会超,韩丽华,王振涛,等. 律复康胶囊对心肌梗死后患者窦性心律震荡的影响[J]. 中国实验方剂学杂志,2011,17(19):273.

[6] 许洁睿,何燕. 中药复方抗心律失常的实验研究进展[J]. 河北中医,2011,33(11):1739.

[7] 陈钰,刘晓秋,郭丽丽,等. 中药在心肌细胞钙信号转导通路研究进展[J]. 中国实验方剂学杂志,2012,18(18):319.

[8] 施慧,龙子江,王靓. 中医药治疗心律失常临床及实验的研究进展[J]. 安徽医药,2009,13(1):95.

[9] 莫薇. 中医药治疗心律失常的新进展[J]. 中国中医药咨讯,2010,15(8):250.

[10] 杨杰,龙子江,穆磊,等. 葛根芩连汤抗大鼠心肌缺血再灌注致心律失常作用及其机制研究[J]. 中国实验方剂学杂志,2013,19(1):284.

[11] 郭洪涛,郑光,张弛,等. 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[J]. 世界科学技术-中医药现代化,2010,12(5):818.

[12] Guang Zheng, Miao Jiang, Yusheng Xu, et al. Discrete derivative algorithm of frequency analysis in data mining for commonly-existed biological networks[J]. CNMT, 2010,35(4):5.

[13] Guang Zheng, Miao Jiang, Xiaojuan He, et al. Discrete Derivative: A data slicing algorithm for exploration of sharing biological networks between rheumatoid arthritis and coronary heart disease[J]. BioData Mining, 2011, 4:18.

[14] 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.

[15] S. Li, Z. Q. Zhang, L. J. Wu, et al. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network[J]. IET Syst Biol, 2007,1(1):51.

[16] 王瑛,付强,刘海霞,等. 稳心颗粒治疗老年缺血性心律失常的临床研究[J]. 中华心血管病杂志,2003,31(12):890.

[17] 吉俭,邱健强,黄艳平,等. 稳心颗粒对冠心病患者心肌缺血及心律失常疗效观察[J]. 中国实用内科杂志,2002,22(11):704.

[18] 李小威,黄从新,范新荣,等. 稳心颗粒对人超极化激活环核苷酸门控阳离子通道 2 电生理特性的影响[J]. 中国心脏起搏与心电生理杂志,2011,25(3):249.

[19] 孙小霞,周筠,兰燕平,等. 稳心颗粒对家兔 3 层心室肌细胞瞬时外向钾电流的影响[J]. 中华医学信息导报,2011,32(5):561.

[20] 张凤祥,曹克将. 心律失常上游药物的治疗研究进展[J]. 中华心律失常学杂志,2011,15(1):68.

[21] 中华中医药学会. 中医内科常见病诊疗指南(西医疾病部分)室性早搏[J]. 中国中医药现代远程教育,2011,9(18):142.

[22] 宋丹,丁碧云. 心律失常的中医药治疗进展[J]. 中医药临床杂志,2012,24(1):76.

[责任编辑 邹晓翠]