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

李立[1],周奇[1],郑光[2],郭洪涛[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. 兰州大学信息学院,兰州 730000)

[摘要] **目的**:利用文本挖掘技术探索中成药、西药对慢性胃炎的治疗规律。**方法**:在中国生物医学文献数据库(CBM)中收集治疗慢性胃炎的相关文献,建立 Access 数据库,运用 SQL 对数据进行处理,结合人工降噪,分析中成药、西药对再障的治疗用药规律。**结果**:四逆散、理中丸等为中成药治疗慢性胃炎文献中出现的高频药物,阿莫西林、奥美拉唑、克拉霉素等为西药治疗慢性胃炎文献中出现的高频药物,在药物联合应用方面,中药四逆散、西药雪尼替丁和奥美拉唑出现频率较高。**结论**:利用文本挖掘技术可以从海量的中成药文献中发现治疗慢性胃炎的用药规律,协助规范中西医联合治疗方案。

[关键词] 慢性胃炎;文本挖掘;用药规律
[中图分类号] R287 [文献标识码] A [文章编号] 1005-9903(2011)24-0228-04

Based on Text Mining Techniques to Explore Medication Regularity of Chinese Patent Medicine and West Medicine Application for Chronic Gastritis

LI Li[1], *ZHOU Qi*[1], *ZHENG Guang*[2], *GUO Hong-tao*[1], *LV Ai-ping*[1*]
(1. *Institute of Basic Research In Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*; 2. *School of Information Science and Engineering Technology, Lanzhou University, Lanzhou* 730000, *China*)

[**Abstract**] **Objectine**: To analyze the regularity of Chinese patent medicine and west medicine application for chronic gastritis with text mining technique. **Method**: Collected large literature about treatment against chronic gastritis in CBM (Chinese Bio Medical Literature Database). Then, We transfer XML type data sets to the structured database of Microsoft® SQL® and visualize them into different graphs by Excel and Cytoscape software. **Result**: Sini San, Lizhong Wan were the commonly used Chinese patent medicine. The high frequencies of west medicine against chronic gastritis were amoxicillin, omeprazole and clarithromycin. Chinese patent medicine Sini San as well as west medicine omprazde and ranitidine of the combinated with othor drugs in treatment of the chronic gastrits. **Conelusion**: The text mining technology can find the medication regularity of Chinese patent medicine and west medicine application for chronic gastritis.

[**Key words**] chronic gastritis; text mining; using drugs regularity

慢性胃炎(chronic gastritis, CG)是指胃黏膜的慢性炎症性病变,可由多种不同的病因引起,组织学上表现为胃黏膜的炎性细胞浸润、改建,最终可导致黏膜固有腺体的萎缩,甚至消失。慢性胃炎发病率较高,有资料显示,在我国成年人80%以上具有不同程度的慢性胃炎[1]。2000 年在井冈山召开的慢性胃炎专题研讨会[2]提出治疗慢性胃炎的原则包括:消除或削弱胃黏膜攻击因子、增强胃黏膜防御能力、使用动力促进剂、中医药及抗抑郁和镇静药。

目前,关于中医或西医治疗慢性胃炎的文献报道已颇多,但因该病病因较为复杂,症状轻重不等,其治疗并无固定的模式。文本挖掘技术的应用可以帮助我们发现大量有用而未被发现的信息,已有文献开始尝试使用该技术探索中医

[收稿日期] 20110526(006)
[基金项目] 国家自然科学基金杰出青年项目(30825047);国家自然科学基金项目(30902003);国家科技部创新方法专项项目(2008IM020900)
[通讯作者] *吕爱平,博士生导师,研究员,从事疾病证候分类研究,E-mail: lap64067611@126.com

药治疗疾病的用药规律[3-4]。本文采用文本挖掘技术对医学文献数据库中大量的关于中成药及西医治疗慢性胃炎的文献进行分析,以求探索中成药、西药以及中成药和西药联合应用对慢性胃炎的治疗规律。

1 材料与方法

文本挖掘是从非结构化的文本数据中,抽取有意义的数据[5]。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[6-7]。

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)在主题检索下检索关键词"慢性胃炎",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6 506篇(检索日期2011年3月6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"。

1.2 文本数据处理 将收集来的数据,按照现在的先后顺序,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**1.1**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。提取出来的信息,主要是机标关键词(包括核心和非核心2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材料,然后导入SQL中进行下一步的挖掘分析。

1.3 数据一次清洗 根据**1.2**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。为了方便处理,将"序号"和"机标关键词"2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。

经过对原文献的分析,发现相同的关键词,在一篇文献的标题和摘要中,存在着重复出现的问题。对于文本挖掘来说,假设每一篇文献的贡献度是相同的,按照这个假设,对于一篇文献中,重复出现的关键词,只需要计算1次。据此,进行数据清洗工作。

1.4 数据挖掘以及分析 通过返查原文献,发现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。这种共同出现不是随机的,而是蕴含有一定的意义[8],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[5-6,8],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,构造了表1的算法来实现这一工作。经过表1算法的计算,得到名为DN_pairs的数据表。经过观察,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,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,构造了表2中的算法来实现。经过表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1次,并且都有1个对应的频数(Frequency)。

1.5 数据二次清洗 经过专业知识对表2(DN_pairs_frqcy)中的数据进行评估后发现,针对特定的疾病,表2中仍存在噪音问题。这些噪音,不再是关键词的简单重复,而是相对于专业只是来说的噪音问题。对此,针对特定的问题,对数据进行二次清洗。而这些噪音的产生,主要是自然语言的二义性和表达方式的多样性产生的。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。

表1 构建关键词对程序算法

```
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
```

表2 合并筛检关键词对程序算法

```
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
```

1.6 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,分别用Excel,Cytoscape 2.8等软件进行可视化处理,得到治疗慢性胃炎的中成药、西药及其联合用药的文献频数图。

2 数据挖掘结果的评价和分析

2.1 中成药治疗慢性胃炎的文献频数(图1) 根据数据挖掘结果发现治疗慢性胃炎的中成药共有66个,图1列出了文献频次在3次以上的17个药物,治法覆盖调和肝脾、温中驱寒、柔肝和胃、散瘀止血、消痞除满、健脾和胃等治疗原则。

从图表中可以看出,报道四逆散治疗慢性胃炎的文献数最多,达到 40 篇,其次是理中丸和胃康胶囊,反查原文发现,既有单独使用这些药物治疗慢性胃炎的报道,也有与其他药物联合应用的报道。

2.2 西药治疗慢性胃炎的文献频数(图 2) 根据数据挖掘结果发现治疗慢性胃炎的西药共有 103 个,图 2 列出了文献出现频次在 20 次以上的 18 个药物,包括抗生素、质子泵抑制剂、H_2 受体阻断剂、铋剂等。从图表中可以看出,使用阿莫西林治疗慢性胃炎的文章最多,达到 204 篇,其次是奥美拉唑和克拉霉素。查看文献原文,发现阿莫西林及克拉霉素的应用多与根除幽门螺杆菌(*Helicobacter pylori*, Hp)有关。出现频次最多的前 5 种药,有 4 种是抗生素,可见根除 Hp 是治疗慢性胃炎的重要手段。

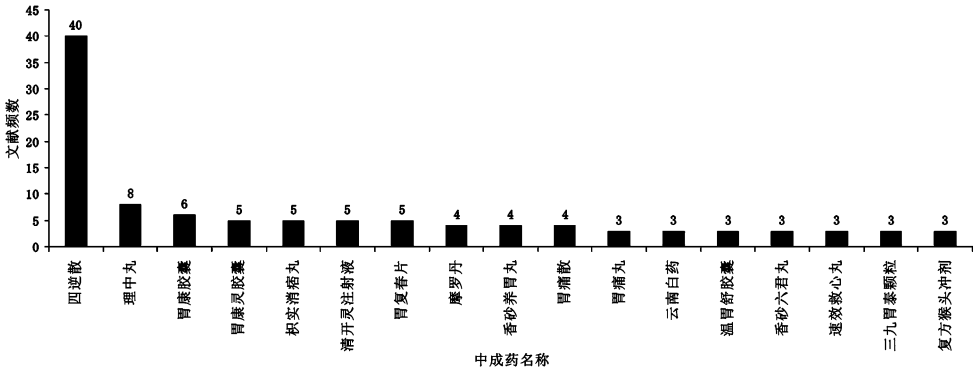

图 1 中成药治疗慢性胃炎的文献频数(频数≥3)

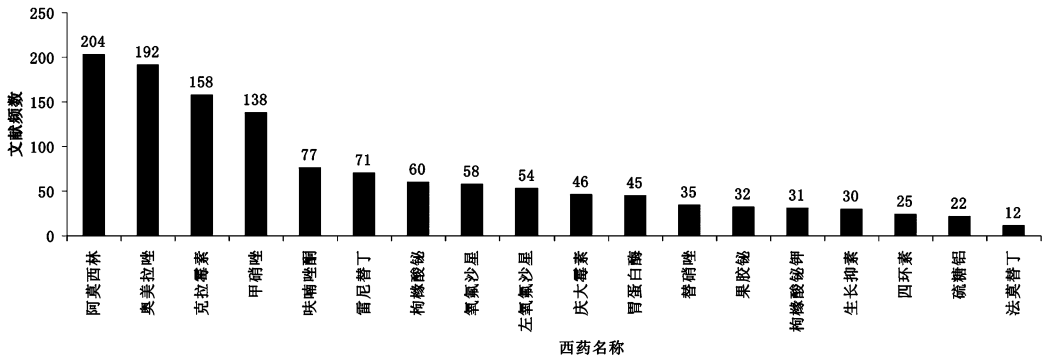

图 2 西药治疗慢性胃炎的文献频数(频数≥12)

2.3 中药与西药联合治疗慢性胃炎的文献频数(图 3) 根据数据挖掘结果发现中成药与西药联合应用治疗慢性胃炎的组合共 22 种,运用 Cytoscape 2.8 软件将数据结果转换成频数网络图(图 3),图中中成药与西药之间的连线代表 2 个药物联合应用。从图中可以看出,中药四逆散、西药雷尼替丁和奥美拉唑常与其他药物联合应用。

3 讨论

慢性胃炎在中医学中可以归属于“胃脘痛”、“痞满”、“吞酸”、“嘈杂”、“纳呆”等范畴。有学者在对 960 例 CG 证候分布规律研究发现,CG 中医证候的出现频率由大到小依次为肝胃不和证、肝郁脾虚证、脾胃虚弱证、脾胃湿热证、胃阴不足证[9]。四逆散出自于汉代张仲景的《伤寒论》,主要由柴胡、枳实、芍药、炙甘草组成,是调和肝脾,疏肝解郁的祖方,常用于治疗肝胃不和、肝郁气滞之胃痛、痞满等病。本研究结果显示,中成药治疗慢性胃炎的文献中,以四逆散为出现频次最高,这与目前该病主要证候类型肝胃不和证的治疗原则一致。理中丸亦来源于《伤寒论》,主要由人参、干姜、白术、甘草组成,具有温中祛寒,补气健脾的功效,适用于脾虚失运、脾胃虚寒等证。清开灵注射液具有清热解毒、化痰通络、醒神开窍的功效,有文献报道[10]清开灵注射液治疗脾胃湿热证、肝胃不和证、胃络瘀血证慢性胃炎的痊愈率、显效率高于胃乃安对照组($P<0.05$)。本研究通过文本挖掘技术得到的出现频次较高的中成药,与目前临床上的主要证候类型的治疗原则基本一致。

18 纪初西方学者就提出了慢性胃炎的概念,20 世纪 50 年代纤维胃镜的问世推进了人们对慢性胃炎的认识,近 20 年来 Hp 的发现,再把人们对慢性胃炎的认识推向一个新的高度。2000 年井冈山会议上提出的治疗慢性胃炎的方法有:根除 Hp、抑酸或抗酸治疗、胃黏膜保护剂治疗、胃运动障碍的治疗、助消化药物治疗等[2]。大量研究表明,Hp 感染是慢性浅表性胃炎的主要病因,根除 Hp 是防止复发的重要措施。本研究挖掘的西药中出现频数最高前 5 个药物,主要与根除

Hp 的规范的三联疗法或四联疗法有关,可见文本挖掘的结果与目前西医根除 Hp 的首要病因治疗一致。

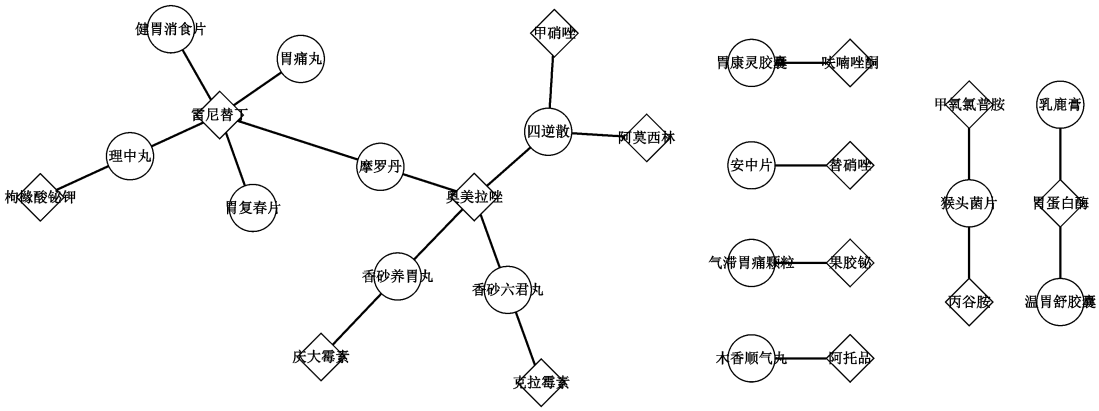

图 3　中药与西药联合治疗慢性胃炎的文献频数

通过对中成药与西药联合应用治疗慢性胃炎的文献进行挖掘,发现较多中西医联合应用的药对,比如四逆散与奥美拉唑、香砂六君丸与克拉霉素、胃康灵胶囊与呋喃唑酮等。目前,已有学者[11]对中医证候与 Hp 感染及胃部病理组织改变的关系进行了研究,但并未形成统一意见。查看原文可以发现,文献多报道中成药和西药临床联合应用时的协同作用,且联合应用的疗效优于单用,但从文献原文及本研究结果中的网络图中,未发现中西药物联用的原则和规律,亦未发现有关中西医药物联合应用不良事件的报道。可见,规范慢性胃炎中西药联合治疗的方案,已经成为当前中西医面临的共同课题。

文本挖掘(text mining)是数据挖掘的一个方向,它所挖掘的对象是非结构化或半结构化,即从数以百万计的文本数据中寻找潜在规律和趋势[12]。在生物学领域,由于生物学数据和生物医学文献数量的急骤增长,通过数据挖掘寻找规律和新知成了生物学研究的一个新热点和重要分支[13]。本研究从 6 506 篇文献中挖掘中西药治疗慢性胃炎的用药规律,基本能够反映临床西药、中成药及其联合用药的情况。我们有理由相信运用文本挖掘技术可以从海量的中医药文献中发现中医治疗规律,协助规范中西医联合治疗方案,从而促进中医临床研究和中医药事业的发展。

[参考文献]

[1] 吕农华.规范慢性胃炎的诊断与治疗[J].中华消化杂志,2005,25(2):65.

[2] 王崇文.慢性胃炎的分类、诊断及治疗现状[J].现代消化及介入诊疗,2003,8(3):164.

[3] 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.

[4] 吕毅斌,李立,王志飞,等.中药治疗类风湿性关节炎、痛风及骨性关节炎用药规律研究[J].世界科学技术——中医药现代化,2010,12(5):833.

[5] Jeffrey W Seifert. Data mining: An overview[M]. Nova Science Publishers, Inc, 2006: 201.

[6] Brigitte Mathiak, Silke Eckstein. Five steps to text mining in biomedical literature[C]. Pisa: the second european workshop on data mining and text mining for bioinformatics, held in conjunction with ECML/PKDD, 2004: 47.

[7] Zheng Guang, Jiang Miao, Xu Yusheng, et al. Discrete derivative algorithm of frequency analysis in data mining for commonly-existed biological networks[M]. Wuhan: The 2nd international symposium on computor network and multimedia technology(CNMT), 2010: 5.

[8] 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.

[9] 张声生.慢性胃炎中医证候学临床研究[D].北京:北京中医药大学,2005.

[10] 林壮民,伍俊杰.清开灵口服液治疗慢性胃炎的初步探讨[J].国际医药卫生导报,2006, 12(13):93.

[11] 朱飞叶,石灯汉,王丽,等.慢性胃炎中医证候研究进展[J].浙江中医药大学学报, 2008,32(5):692.

[12] 吕婷,姜友好.文本挖掘在生物医学领域中的应用及其系统工具[J].中华医学图书情报杂志,2010,19(4):56.

[13] Tari L, Anwar S, Liang S, et al. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism[J]. Bioinformatics, 2010, 26(18): 1547.

[责任编辑　邹晓翠]