利用文本挖掘技术分析慢性盆腔炎的用药规律

王妍[1,2], 郑光[1,3], 郭洪涛[1,4], 张弛[1], 姜森[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. 北京市中医学校, 北京 101101;
3. 兰州大学信息学院, 兰州 730000; 4. 上海中医药大学, 上海 201203)

[摘要] **目的**: 利用文本挖掘技术探索西药、中成药对慢性盆腔炎(chronic pelvic inflammation, CPI)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 CPI 的相关文献, 运用大型数据库(SQL)对数据进行处理, 结合人工降噪, 分析西药、中成药对 CPI 的治疗用药规律。**结果**: 甲硝唑、替硝唑、氧氟沙星、庆大霉素、地塞米松等西药依次为治疗 CPI 文献中出现的高频药物。妇科千金片、康妇消炎栓、桂枝茯苓丸、千金散等依次为治疗 CPI 文献中出现的高频药物。同时也发现西药、中成药联合治疗 CPI 的一些规律。**结论**: 利用文本挖掘的方法, 从文献报告频数方面呈现了西药、中成药治疗 CPI 的用药规律, 尤其是西药、中成药联合应用很值得进一步的研究。

[关键词] 慢性盆腔炎; 文本挖掘; 西药; 中成药
[中图分类号] R287 [文献标识码] A [文章编号] 1005-9903(2012)09-0286-04

Exploring the Association Rules of Chinese Patent Medicine and West Medicine for Chronic Pelvic Inflammation with Text Mining Technique

WANG Yan[1,2], *ZHENG Guang*[1,3], *GUO Hong-tao*[1,4], *ZHANG Chi*[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. *BeiJing School of Traditional Chinese Medicine (TCM), Beijing* 101101, *China*;
3. *School of Information Science and Engineering Technology, Lanzhou University, Lanzhou* 730000, *China*;
4. *Shanghai University of Traditional Chinese Medicine, Shanghai* 201203, *China*)

[**Abstract**] **Objective**: To mine the association rules of Chinese patent medicine and west medicine used for chronic pelvic inflammation. **Method**: Transfered XML type data sets to the structured database of Microsoft SQL Server and visualized them into different graphs by software Excel and Cytoscape. **Result**: The results suggested that metronidazole, tinidazole, ofloxacin, gentamicin, and dexamethasone were the commonly used west medicine. Gynecologic Qianjin tablets, Kangfu Antiphlogistic Suppository, Guizhi Fuling bolus, and Qianjin powder are of the core Chinese patent medicine frequently used for chronic pelvic inflammation. **Conclusion**: The use of text mining methods, From the frequency reported in the literature has shown the Western medicine, Chinese medicine treatment of CPI's drug laws, especially in Western medicine-Chinese medicine combin were worth further study.

[**Key words**] chronic pelvic inflammation; west medicine; chinese patent medicine; text mining

盆腔炎(pelvic inflammation, PI)是盆腔内生殖器官、盆腔周围结缔组织以及盆腔腹膜等发生炎性病变的总称, 分为急性与慢性两种, 而以慢性盆腔炎(chronic pelvic inflammation, CPI)多见, 是妇科的常

[收稿日期] 20111121(785)
[第一作者] 王妍, 博士研究生, 讲师, 主要从事中医药疗效评价和药效评价方法研究, Tel: 010-81530412, E-mail: 18901309816@126.com
[通讯作者] *吕爱平, 博士, 教授, 从事中医证候分类科学研究, Tel: 010-64067611, E-mail: lap64067611@126.com

见病、多发病,占妇科门诊 1/3[1]。慢性盆腔炎临床表现以长期反复发作的下腹部或腰骶部疼痛,白带增多,月经失调和痛经为主,部分患者可因本病并发输卵管阻塞性不孕以及异位妊娠[2]。中医药治疗慢性盆腔炎多标本兼顾,西医治疗慢性盆腔炎多采用抗生素治疗。中西药联用,优势互补、扬长避短,是目前治疗慢性盆腔炎的重要措施。本文利用文本挖掘技术分析中成药及西药治疗慢性盆腔炎的用药规律,为临床用药和药物研究提供依据。

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[3-5]。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]。

1.1 文本数据收集 首先,登录中国生物医学文献数据库(Chinese BioMedical Literature Database, CBM,网址 http://sinomed.cintcm.ac.cn/index.jsp)在主题检索下检索关键词"慢性盆腔炎"。经过检索,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4 416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 **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"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。

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,只需要计算一次。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[6]。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[3-5,7-12],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,都只出现一次,并且都有一个对应的频数(Frequency)。

表 1 关键词对构造算法

```
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
```

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,而是相对于专业知识来说的噪音问题。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 2 关键词对频数算法

```
USE table DN_pairs
k =max_linc_number
DO while k ≥ 1
GO top
  FOR DescriptorName_Pair(1)
    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（Shannon P 2003），得到治疗 CPI 的西药、中成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 治疗 CPI 西药文献频数图 数据挖掘到的西药有 76 种，为了方便展示，依据频数高低排序，如图 1 所示，选取大于等于 6 的文献频数，所涉及的西药包括以下几类。抗微生物类：甲硝唑、替硝唑、氧氟沙星、庆大霉素、左氧氟沙星、克林霉素、红霉素、环丙沙星、卡那霉素、阿奇霉素、头孢曲松钠、头孢拉定、阿莫西林；肾上腺皮质激素类：地塞米松；局部麻醉药：利多卡因；酚类有机物：苯酚；蛋白水解酶类：糜蛋白酶。抗菌药的种类最多，应用最广泛，其中甲硝唑、替硝唑频率最高，说明临床工作者在 CPI 治疗中对关注较多，其应用较广。

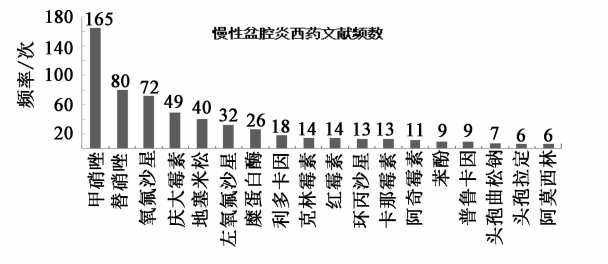

图 1 慢性盆腔炎药物文献使用频数图（频数≥6）

2.2 治疗 CPI 中成药文献频数图 数据挖掘到的中成药有 57 中，如图 2 所示，选取大于等于 6 的文献频数，所涉及频数较高的中成药包括妇科千金片、康妇消炎栓、桂枝茯苓丸、千金散、鱼腥草注射液、妇乐冲剂、丹参注射液等。其中妇科千金片、康妇消炎栓、桂枝茯苓丸使用频数最高，返查原文献，发现其应用有很大一部分单独应用中药治疗。

2.3 CPI 西药、中成药联合使用文献频数图（图 3）

数据挖掘到得西药和中成药联合应用的组合有 72 种，如图 3 所示，选取大于等于 2 的文献频数，图中连线代表药物两两之间的联系，连线数量代表两

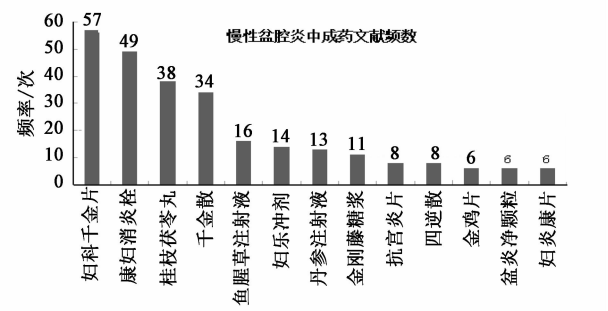

图 2 慢性盆腔炎中药文献使用频数图（频数≥6）

种药物联合使用在文献出现的篇数。从图中可以看出，中成药应用都是单独出现，两两之间没有发生交联，而其联合的西药可以是一种或多种。其中，妇科千金片联合应用药物的种类最广泛，可以分别联合氧氟沙星、阿托品、克林霉素、左旋咪唑、甲硝唑。

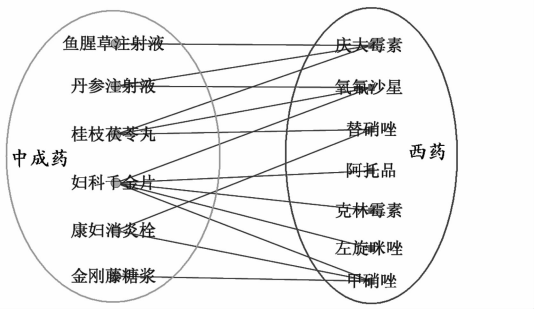

图 3 慢性盆腔炎西药、中成药联合使用文献频数图（频数≥2）

3 讨论

随着生物信息学、系统生物学等研究的兴起，科学家们越来越清楚地认识到，只有以信息系统视角研究同样是系统的、复杂的中医药理论体系，在合理整合和充分利用各种数据资源的基础上，进行科学分析、特征提取和规律探索，才能逐步揭示其本质和内涵[13]，通过数据挖掘寻找规律和新知成了生物学和医学研究的热点[14]。本研究从4 416篇文献中挖掘中西药治疗 CPI 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。

治疗 CPI 的西药频数排在前五位的依次是甲硝唑、替硝唑、氧氟沙星、庆大霉素、地塞米松，主要是抗感染和对症治疗，与治疗慢性盆腔炎的诊断标准和治疗指南[15]基本一致。

中成药的应用中，妇科千金片、康妇消炎栓、桂枝茯苓丸的应用频数最高。妇科千金片由千斤拔、单面针、金樱根、穿心莲、功劳木、党参、当归、鸡血藤组成。具有清热除湿，补益气血的作用。用于带下病，湿热下注，气血不足症。盆腔炎、子宫内膜炎、宫颈炎见上述证候者。有研究显示妇科千金片有明显的抑菌作用[16]。康妇消炎栓由苦参、败酱草、紫花地丁、穿心

莲、蒲公英、猪胆粉、紫草、芦荟组成,具有清热解毒,利湿散结,杀虫止痒的作用,用于湿热,湿毒所致的腰痛,小腹痛,带下病,阴痒,阴蚀。桂枝茯苓丸为经典的古代方药,出自《金匮要略》妇人妊娠病脉证并治方。现代研究表明其具有抗炎、增加外周血流量、抑制前列腺素 E 合成、抗纤维蛋白溶酶、增强巨噬细胞吞噬能力、刺激黄体酮分泌等作用[17]。

中成药和西药的联用中,妇科千金片和抗菌药联合应用明显较多。妇科千金片的临床报告近年较多,居于中成药的首位,返查文献发现,临床联合应用时,文献大多报道可以起到协同作用,合用的疗效优于单用,依据大量文献中西医联合用药疗效比较确切,具体中成药和西药联合使用的合理性,如何联用,其联合作用的机制是怎样的,有无潜在的风险,以及基础实验研究等方面的报道罕见,大部分临床医生根据临床经验进行中西药联合治疗。我们呼吁应从基础药理实验开始,不断探索中西药间的相互作用及配伍规律,以便合理使用率,逐步达到联合用药确实"安全、有效、经济、合理"。

本次研究基于文本挖掘技术,对现有数据库文献进行定向文本挖掘,结果基本呈现了慢性盆腔炎中成药、西药及中西药联合用药规律的临床研究现状,是一种新的经验总结方法。它可以快捷、客观、全面系统的总结慢性盆腔炎药物应用规律,为临床医生临床用药提供客观参考依据。

[参考文献]

[1] 朱兰,赵煜,刘军,等.慢性盆腔炎国内外的研究现状[J].西南国防医药,2007,17(3):376.

[2] 董建春,王波主译.女性生殖道感染性疾病.4 版[M].济南:山东科学技术出版社,2004:377.

[3] Seifert J W. Data mining: an overview[C]. CRS Report, 2004, RL31798.

[4] Cohen A M, Hersh W R. A survey of current work in biomedical text mining[J]. Brief Bioinform, 2005, 6(1):57.

[5] Bellazzi R, Diomidous M, Sarkar I N, et al. Data analysis and data mining: current issues in biomedical informatics[J]. Methods Inf Med, 2011, 50(6):536.

[6] Guang Zheng, Hongtao Guo, Aiping Lu, et al. Two dimensions data slicing algorithm, a New Approach in Mining Rules of Literature in Traditional Chinese Medicine[J]. CCIS Springer-Verlag Berlin Heidelberg, 2011, 237:161.

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

[8] Michaelw Berry M B. Lecture Notes in Data Mining[C]. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, 2006.

[9] H Shatkay, R Feldman, 'Mining the BiomedicalLiterature in the Genomic Era: An Overview'[J]. Computational Biology, 2003, 10(6):821.

[10] Sam Zaremba, M R S, Thomas Hampton. etc. Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens[J]. BMC Bioinformatics, 2009, 10:177.

[11] Nathan Harmston, W F, Michael P H. Stumpf. What the papers say: text mining for genomics and systems biology[J]. Human Genomics, 2010, 5(1): 17.

[12] Guang Zheng M J, Aiping Lu, et al. Discrete derivative: a data slicing algorithm for exploration of sharing biological networks between rheumatoid arthritis and coronary heart disease[C]. BioData Mining, 2011, 4(18):1.

[13] 李梢,张学工,季梁,等.复杂性疾病生物信息学研究的策略与方法[J].世界华人消化杂志,2003,11(10):1465.

[14] 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.

[15] Ault K A, Faro S. Pelvic inflammatory disease, Current diagnostic criteria and treatment guidelines[J]. Postgrad Med, 1993, 93(2):85.

[16] 管仲莹,向绍杰,孟莉,等.妇科千金片抑菌作用的实验研究[J].实用中医内科杂志,2010,24(6):29.

[17] 曹惠云.桂枝茯苓丸在日本的研究与应用[J].国外医学:中医中药分册,2003,25(2):78.

[责任编辑 古云侠]