中国中药杂志
China Journal of Chinese Materia Medica

正则表达式在上市中药文献信息提取中的应用

王志飞[1,2]，谢雁鸣[1*]，王永炎[1]

（1. 中国中医科学院 中医临床基础医学研究所，北京 100700；
2. 中国中医科学院 博士后流动站，北京 100700）

［摘要］ 上市中药文献分析任务繁重，利用计算机自动提取目标信息之后再进行人工检查，可以简化工作、提高效率。该文分析了上市中药文献信息提取的难点，认为非术语信息的匹配和不同分组的同类信息的区分是影响常用信息提取技术在上市中药文献信息提取中应用的两大关键问题。该文分析了正则表达式的模式匹配对于解决上述两大问题的价值，并选取“病例数”（非术语信息）和“有效率”（不同分组同类信息区分）2 种信息，以此为例探讨了基于正则表达式的信息提取方法，初步实践了这一思路。

［关键词］ 中药上市后再评价；文献分析；正则表达式；信息提取

中成药上市前研究的局限性，主要表现在对研究对象、用药的严格限定上。因此中药上市后再评价，就要求在更广泛人群、更长时间及合并用药的情况下考察其安全性、有效性和经济学的特点。因此，上市后再评价对于研究的全面性有较高的要求。这种要求反映于文献分析，就要求广泛、全面地占有资料，因此导致了文献分析巨大的人力和时间耗费。

笔者于 2010 年 11 月检索鱼腥草注射液的中文文献，不包括灰色文献共获得 1 151 篇。如此巨量的文献，如果以传统的方法逐一阅读、甄别、摘录目标信息，需要耗费巨大的人力和时间。一种可行的方法是使用 OCR 工具将全文批量识别、校正，利用计算机自动提取信息之后进行人工检查。实践证明，完成千篇左右文献量品种的文献分析报告，人机结合的方法可将时间从 2 ~ 3 个月缩短为 1 ~ 2 周。

1 上市中药文献分析须提取的信息及方法学挑战

开展中药上市后再评价文献分析，不但要关注文献的题名、摘要、作者、作者单位、刊名、影响因子、出版时间、主题词、关键词等常规信息，更要关注所报道中药的类别、剂型、剂量、疗程、生产厂家、治疗病种、证候、症状、试验类型、是否对照组、试验对象的特征（如人群的年龄、性别等）、病例数、联合用药、不良反应、有效率等特定信息。提取上述特定信息会面临 2 个方面的问题：一是非术语信息的匹配问题，二是相似信息的区分问题。

常用的信息提取方法，如分词和词性标注，主要面向较为固定的术语，如剂型、生产厂家、治疗病种等，较难应用于非术语的信息提取；而中药上市后再评价文献分析的目标信息，往往都是非术语信息，如证候、症状等，其表达虽遵循一定的规律，但更多的是表现出随意性，会因为报告者个人的学术背景和语言习惯而不同。另外，中药上市后再评价关注的信息有相当一部分是量化的信息，如剂量、疗程等。对于这 2 类信息，基于简单匹配的信息提取方法基本上无能为力。

对于中药上市后临床再评价来说，提取信息最为关键的一点是区分同类信息。临床试验往往会采用对照设计，因此要抽取的目标信息往往在试验组和对照组同时存在，提取时须做出组间的区分。比如有效率，治疗药物的有效率与对照药物的有效率必须在信息提取时就区分开来。这种要求，只有通过对上下文语境的充分分析才可能做到。因此，中药上市后临床再评价的信息提取方法，一定是考虑了上下文匹配的方法，这一点，也是简单匹配的信息提取方法无法完成的。

2 正则表达式用于上市中药文献信息提取的优势

正则表达式（regular expression）是一种可以解决上述 2 个问题的潜在方法。正则表达式出现于 1950 年，最初用于描述神经元模型[1]；因为强大的模式表达功能，其应用范围快速扩展；目前已广泛应用于信息技术领域，是强大、便捷、高效的文本处理工具。

正则表达式功能的强大，在于其对“模式”的表达[2]。模式是一类事物的共有特征，是人们思考问题和描述事物时不可或缺的逻辑工具。中药上市后再评价文献中的目标信息可以被阅读者理解，关键在于这些目标信息处于各种各样的模式之中，模式表达了目标信息与其前后相关信息的关

［稿件编号］ 20110806036
［基金项目］ 国家“重大新药创制”科技重大专项（2009ZX09502-030）；中国中医科学院自主选题研究项目（Z0171）

［通信作者］ *谢雁鸣，研究员，博士生导师，研究方向为中医临床评价方法研究，Tel：（010）64014411-3302，E-mail：zhinanb2010@yahoo.com.cn
［作者简介］ 王志飞，助理研究员，主要从事中药上市后临床再评价研究，Tel：15001007822，E-mail：wzhftcm@163.com

系,它与目标信息本身的含义一起构成了阅读者理解的基础。

中药上市后再评价文献提取的 2 个难点,都可以通过模式的匹配来解决。对于正则表达式来说,数字也是一种模式;各种形式的数字也可以用各种表达式来匹配;同时,模式强调了目标信息与其上下文的关系,而这正是区分同类信息的关键。因此,正则表达式用于中药上市后再评价相关信息的提取,具有十分明显的优势。

3 正则表达式在上市中药文献信息提取中的应用

3.1 中药上市后再评价文献中病例数的提取

3.1.1 从题名中提取病例数信息 比较规范的论文,一般会在题名标示出病例数,如"某某药口服治疗心绞痛 50 例疗效观察"、"某某药治疗冠心病心绞痛 72 例临床研究"等。病例数信息出现于题名,其格式相对较简单,一般表现为"整数 + 例"的模式;同时,"整数 + 例"的模式出现于题名,其含义也是确定的,就是指试验组的病例数。因此,制定正则表达式如下(引号内字符串,不包括引号,下同):"[0 ~ 9] + 例",式中,"[0 ~ 9]"匹配 0 ~ 9 的任意数字;" + "表示这样的数字至少有 1 个(包括 1 个);"例"是这个表达式的指示词,说明 1 个或几个数字之后,必须有"例"字。

笔者全面收集了血栓心脉宁(包括片剂和胶囊)有效性和安全性文献共 76 篇,考察其题名设置,50.00% 的文献(38 篇)题名中含有病例数信息;应用上述正则表达式从题名中提取病例数,共获得 38 篇文献的病例数信息,成功率 100%。

3.1.2 从摘要中提取病例数信息 与从题名中提取不同,从摘要中提取病例数信息,要考虑所提取的数字是试验组的病例数,还是对照组的病例数,抑或是总的病例数。要做到这一点,制定正则表达式时,必须结合数字所在的上下文。

摘要中的病例数信息一般以下面几种形式出现:A,"将 106 例高血脂患者随机分为 2 组,治疗组 53 例,口服某药,对照组 53 例,口服某药";B,"采用某药治疗肺心病心力衰竭患者 45 例(治疗组),并与单纯西药治疗组 42 例(对照组)对照";C,"某药组(治疗组)40 例,某药组(对照组)38 例";D,"将 100 例冠心病患者随机分成 2 组,治疗组(50 例)口服某药,对照组口服某药";E,"将 40 例患者随机分为 2 组,每组 20 例";F,"治疗组和对照组各 80 例";G,"将 31 例高脂血症患者分为治疗组(n = 16)和安慰剂组(n = 15)"。

概括上述表述模式,分别制定如下正则表达式:①"(治疗|试验|试验)组[共] * [0-9] + 例" ;② "[0-9] + 例[\(\(]?(治疗|试验|试验)组[\)\)]";③ "[\(\(]?(治疗|试验|试验)组[\)\)][0-9] + 例";④ "[0-9] + 例. *? 分[为成作]. *? 每组[0-9] + 例";⑤ "治疗组[和及与]对照组各[0-9] + 例";⑥"(治疗|试验|试验)组\(n = [0-9] + \)"。式中," * "表示前面的内容存在 0 个或 0 个以上;" + "表示前面的内容至少有 1 个(包括 1 个);". *?"表示最小匹配。

有 2 种情况需要注意:①摘要不标明哪个整数是治疗组,哪个是对照组,而是用药名来代替,如"采用甲药治疗脑血栓形成 158 例,乙药治疗 158 例",这种情况下,可否用正则表达式提取病例数取决于是否能确定研究药物是试验组还是对照组;②有些文献只是简单的病例报道,未设定对照组,这种情况下正则表达式亦不适用,只能人工提取。

同样以 76 篇血栓心脉宁(包括片剂和胶囊)有效性和安全性文献为例,排除上述 36 篇题名中包含病例数的文献,剩余 36 篇文献中有摘要信息的文献共 25 篇;25 篇中有病例数信息的文献共 17 篇,占 68.00%;以上述正则表达式匹配,共获得 9 篇文献的病例数信息,占 52.94%。

本例以 6 个正则表达式匹配仅获得 9 篇文献的病例数,可见,针对摘要的病例数提取,与针对题名的提取相比,其效能降低许多。需要说明的是,本例试验文献数量太少,以至效能太低;如果文献数量较大,则其效能会大大增加。

3.2 中药上市后再评价文献中有效率的提取

有效率的信息一般出现在摘要中。有效率有十分突出的模式,它会以一个百分比的形式出现,但数字是多少并不固定,而且文献中出现的百分比不一定就是有效率。另外,与病例数类似,有效率有试验组和对照组或是其他组的不同,亦有总有效率和分项有效率的不同。

文献中对有效率的表达主要有以下几种形式:A,"结果:临床总有效率治疗组 96.67%,对照组 80.0%";B,"治疗组与对照组的总有效率分别为 95.8%,80.2%";C,"治疗组总有效率 92.5%,对照组总有效率 77.3%"。

根据上述表达形式概括正则表达式,主要有 2 个模式:①"(治疗|试验|试验)组[总] * 有效率. *? [0-9] + [. 0-9] * ";②"[总] * 有效率(治疗|试验|试验)组. *? [0-9] + [0-9] * "。获得匹配结果后,通过正则表达式 "[0-9] + [0-9] * ",将有效率(数字)提取出来。

需要注意的是,有些文章根据试验内容自己为治疗组命名,如"甲药 + 乙药组",这样表述的有效率,正则表达式无法制定普适的模式,无法提取,需要人工进行。

仍以血栓心脉宁(包括片剂和胶囊)为例,全面检索共获得有效性文献 73 篇,其中有 44 篇含有摘要信息,13 篇在摘要中含有有效率的信息。应用上述正则表达式匹配,共获得 8 篇文献的有效率信息,占 61.54%。

4 讨论

在中药上市后再评价文献分析的过程中,基于简化工作、提高效率,以快速形成文献分析报告的考虑,应用正则表达式辅助人工来提取目标信息,具有可行性。但是,论文文献采取自然语言表达思想,不同作者的表述变异很大,制定一个能包容所有表述的表达式,基本上不可能实现。实践中,可以应用定制的表达式来提取信息,检查其未能提取成功的信息,分析这些遗漏信息的模式,然后再修改表达式或制定新的表达式,而后再进行实践,从而形成循环,促进表达式在不断的实践中不断完善。因此,制定正则表达式只是第

一步,只有通过大规模的实践,才可能形成高效能的表达式。即便如此,面对自然语言的复杂语境,仍无法保证表达式可匹配全部目标信息,亦无法保证匹配到的信息一定完全正确。因此,正则表达式应用于中药上市后再评价文献信息提取,一定是采取人机结合的方式,这样才能保证信息的全面和准确。总之,正则表达式对中药上市后再评价文献信息的提取具有可观的应用价值,值得深入研究。

[参考文献]

[1] 邵瑛, 陆月明. 基于优化正则表达式的文本告警信息的提取和分析[J]. 微型电脑应用, 2010, 26(5):16.
[2] 王志飞, 李晓君, 郭霞珍, 等. 正则表达式在中医文献研究中的应用初探[J]. 中国中医药信息杂志, 2010, 17(3):98.

Application of regular expression in extracting key information from Chinese medicine literatures about re-evaluation of post-marketing surveillance

WANG Zhifei[1,2], XIE Yanming[1*], WANG yongyan[1]
(1. *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*;
2. *Post-doctoral Station of China Academy of Chinese Medical Sciences, Beijing* 100700, *China*)

[**Abstract**] Computerizing extracting information from Chinese medicine literature seems more convenient than hand searching, which could simplify searching process and improve the accuracy. However, many computerized auto-extracting methods are increasingly used, regular expression is so special that could be efficient for extracting useful information in research. This article focused on regular expression applying in extracting information from Chinese medicine literature. Two practical examples were reported in this article about regular expression to extract "case number (non-terminology)" and "efficacy rate (subgroups for related information identification)", which explored how to extract information in Chinese medicine literature by means of some special research method.

[**Key words**] re-evaluation; post-marketing surveillance Chinese medicine literature; regular expression; information extraction

doi:10.4268/cjcmm20112035

[责任编辑 马超一]