

# 一站式检索工具在系统评价中的应用及效果评价



苏石<sup>1</sup>, 翟中会<sup>2</sup>, 刘华<sup>3</sup>

1. 西安交通大学第一附属医院 (西安 710061)

2. 西安交通大学图书馆 (西安 710061)

3. 西安医学院 (西安 710021)

**【摘要】** 生物医学科研人员进行系统评价时,检索结果的敏感度和精确度直接影响系统评价最终结果,而不同的检索方法和检索工具可能产生不同的检索结果。本文首先评价了布尔逻辑检索、等级检索、主题检索和全文检索方法对系统评价检索结果的影响,接着介绍了一站式检索工具(Google Scholar 和资源发现系统)在系统评价中的应用,并评价其检索效果。

**【关键词】** 系统评价;检索;敏感度;精确度;一站式检索

## Application and effect of one-stop search tool in system review

SU Shi<sup>1</sup>, ZHAI Zhonghui<sup>2</sup>, LIU Hua<sup>3</sup>

1. The First Affiliated Hospital of Xi'an Jiao Tong University, Xi'an, 710061, P.R.China

2. Library of Xi'an Jiaotong University, Xi'an, 710061, P.R.China

3. Xi'an Medical University, Xi'an, 710021, P.R.China

Corresponding author: ZHAI Zhonghui, Email: zhaizh@mail.lib.xjtu.edu.cn

**【Abstract】** The sensitivity and accuracy of the retrieval results directly affect the results in the systematic review of biomedical researchers. However, different retrieval methods and retrieval tools lead to different retrieval results. In this paper, we first evaluate the effects of Boolean logic retrieval, hierarchical retrieval, topic retrieval and full-text retrieval for the results of systematic reviews. Then we introduce the application of one-stop search tool (Google Scholar and resource discovery system) in systematic review, and evaluate the effect of one-stop search tool.

**【Key words】** Systematic review; Retrieval; Sensitivity; Accuracy; One-stop search

生物医学文献的系统评价是基于原始临床医学研究产生二次证据的过程,是对已出版的文献进行系统总结、评价和分析<sup>[1]</sup>。常用生物医学文献数据库包括:CNKI、CBM、CMCC、VIP、WanFang Data、MEDLINE、EMbase、SCI、OVID、PQDT、The Cochrane Library 等<sup>[2]</sup>。目前 PsycINFO 数据库已有超过三百万条记录, MEDLINE 和 EMbase 也已超过两千万条。它们都支持布尔逻辑检索、通配符、词根扩展等检索方式。但类似的数据库的检索词间存在差别,这不仅增加了研究者的工作负担,同时对检索结果的敏感度和精确度也产生很大的影响。系统评价研究人员为了保证查全率,就必须检

索大量数据库,这无疑增加了系统评价团队的工作量,如能采用一站式检索工具 Google Scholar 或资源发现系统代替众多的数据库,将减轻系统评价时文献检索的工作量。本文采用实例,评价了布尔逻辑检索、等级检索、主题检索和全文检索方法对系统评价检索结果的影响,并与一站式检索工具结果比较,确定一站式检索工具是否能够满足系统评价要求的查全率和查准率。

## 1 现状

国内外学者做了大量关于系统评价检索策略优化的研究,主要集中在根据不同的检索目的构建相应的过滤器,其中 narrow 过滤器可提高查准率, broad 过滤器可提高查全率<sup>[3-6]</sup>。但系统评价检索策略还存在一些其他问题:①需从不同的数据库平

台检索文献,但不同数据库检索策略和收录范围均不同,这无疑给研究人员增加了工作负担,另外由于系统评价团队对一些检索平台不熟悉,往往导致大量的漏检、误检等。所以除了采用数据库检索外,还需要采用引文跟踪、跟踪登记的临床试验、浏览专题杂志、阅读相关会议文献等手段保证查全率;②检索结果的再现困难,系统评价要求检索策略能够再现相同的检索记录集合,但很多系统评价的检索式不能满足这个要求。造成这种原因主要包括:检索策略拼写错误、不恰当的逻辑组配、截词错误、不相关的 MeSH 词、错误的 MeSH 词、不合理的冗余以及同一检索策略不适用于所有数据库。另外,美国国立医学图书馆每年对医学主题词表(MeSH)的更新也可能使检索表达式不能重新执行或者不能获取相同的检索结果。

## 2 常用检索范式

系统评价检索策略主要采用布尔逻辑检索和等级检索 2 个检索范式。布尔逻辑检索已有 40 多年历史,但布尔逻辑检索不适用于最近几年出现的检索系统,如 Google Scholar 和资源发现系统。而等级检索近年来才出现,这种方式比布尔逻辑检索更适用于 Google Scholar 和资源发现服务等系统。

### 2.1 布尔逻辑检索

采用逻辑运算符连接术语、短语、子表达式等组成系统评价检索式,布尔检索式非常直观,用户能够从检索式中预计检索结果包括或不包括哪些内容。尽管制定布尔逻辑检索式非常复杂,但系统评价一直在使用该检索式,主要是由于布尔逻辑检索的再现性,如果数据库记录没有发生变化,运行同样的检索式能够返回相同的检索结果。另外一个原因是布尔逻辑表达式的灵活性,能够表达非常复杂的概念,如“(clin \$ adj25 trial \$).ti”标识标题中包含 clin 和 trial 词根并且能相邻 25 个字符。布尔逻辑检索能够明确知道包含和排除哪些文档,这有利于在出现新的论文后对系统评价进行更新,布尔逻辑检索还可以定义为一个查询模板,这些模板可以用于相应的子查询。

但布尔逻辑检索也存在一些不利因素,一方面很难控制检索结果数量,如错用一个“AND”将会丢失大量的检索结果,错用一个“OR”将导致返回大量的检索结果。另一方面由于检索结果没有按相关性排序,判断哪些结果纳入系统评价比较困难。再者,一些检索式过于复杂,如 The Cochrane Library 中的系统评价中有的检索式长达 100 多行,

容易出现错误。且较长的检索式很难重复获取到相同的检索结果,如由于主题词表的变化引起主题词扩展(EXP)意义发生改变或数据库加入新的元数据字段引起逻辑运算符“OR”的语义发生变化。

### 2.2 等级检索

等级检索依据词频、查询上下文、链接结构等技术对检索结果进行排序,用户能够通过等级判断哪些文献更符合系统评价纳入标准,而不需考虑整个检索结果的大小<sup>[7]</sup>。等级检索更适合迭代查询,在检索式改变后,能够很明确看到是否改善了检索结果,这个优点完全符合系统评价检索的要求,这是布尔逻辑检索进行迭代查询时无法做到的。

但值得注意的是,等级检索缺乏再现性,由于数据集合的改变,术语统计数据也发生了变化,因此等级排序结果也可能发生改变。自动扩展查询可能明显改变了查询的原来意义,另外由于检索机制经常调整,非常小的语法解析变化可能导致排序结果发生很大的变化。另外,排序文献中究竟排在前面多少条目是相关文献,这个数值也很难界定。

目前还没有充分的证据支持将等级检索式作为系统评价检索方式,现有系统评价还是采用布尔逻辑检索,但等级检索可以作为布尔逻辑检索的补充,如系统评价团队根据文献标题和摘要选择相关文献的时候,可以根据等级排序方法迅速找出相关文献,从而减少评价人员的工作量。另外布尔逻辑检索时需要注意 MeSH 词扩展仅仅支持最新版本的 MeSH 词表,如果 MeSH 词表发生了变化,重新运行检索结果将不能得到相同的结果集。OVID 提供的 MEDLINE 不同历史版本能够弱化这种影响。此外,由于词根检索对用户的不透明性,应使用手动截词进行更有效的控制。

## 3 常用检索方式

### 3.1 主题检索

主题检索是要求关键词出现在标题、关键词或摘要中,检索步骤为思考研究主题、分解主题概念、列举主题概念的同义词、建立逻辑组配检索式(A1 OR A2 OR A3) AND (B1 OR B2 OR B3) AND (C1 OR C2 OR C3)。这种检索方式主要适用于摘要或索引数据库(MEDLINE、SCI 等),特点是精确检索,即输入什么关键词就会检索到与该关键词相关的文献,这种情况下,大部分检索返回结果较少,有时甚至检索结果为零。一方面由于检索词仅仅和元数据匹配,如检索“youth”,而文章标题或摘要中仅有“teenager”时,将不能检索到这篇

文献。另一方面检索系统只能精确检索,没有自动词根覆盖单词的所有变体,如英式和美式拼法,因此在检索时必须包含单词的所有变体、单复数等形式。复杂的布尔逻辑检索确保了主题检索的准确性和命中率。

### 3.2 全文检索

全文检索是指关键词在全文中匹配而不仅仅是元数据,随着大量全文数据库的出现,检索环境也发生了很大的变化,数据库提供自动词根搜索并自动增加相关同义词,如检索“teenagers”,全文中可能出现“youth”、“adolescence”等,不需要用“teenagers”的同义词检索。即使采用最基本的检索如 A AND B AND C 也能够保证查全率。相反,选择不恰当的同义词或过度使用通配符反而会使检索结果变得更糟糕。

恰当的关键词是主题检索的关键,如采用“OR”连接药物名称列表、同一个概念列表等,但问题是许多研究人员对关键词列表没有充分的理解,经常过度延伸使用关键词。如检索式(A1 OR A2 OR A3) AND (B1 OR B2 OR B3) AND (C1 OR C2 OR C3),可能使用 A3、B3、C3 三个词并不确切,但用户仅仅是为了避免万一遗漏的情况,这时由于没有区分学科,相同的关键词可能出现在不同学科而检索意义完全相反。全文检索虽然保证了查全率,但查准率较低。

## 4 Google Scholar 检索工具

Google Scholar 从 2004 年诞生到现在,已经被广大科研人员接受,据 2010 年 OCLC 的统计,84% 的科研人员使用 Google Scholar 检索文献<sup>[8]</sup>。但许多信息检索专家提出了 Google Scholar 的大量缺点,如检索式有 256 个字符限制,最多显示 1 000 个检索结果,每个页面最多 20 个结果,不能成批导出结果,仅仅能通过 ZOTERO 导出每页 20 个结果,不能和其他专业文献选择过程整合等。此外,Google Scholar 没有截词操作符,自动词根机制仅用于常见词根,但检索时并不可靠,如“child”不能自动匹配“childhood”和“children”,逻辑操作符嵌套不能超过一层等缺点;甚至会出现由于检索词顺序不同而引起检索结果不同,如“esophagus OR oesophagus”和“oesophagus OR esophagus”由于关键词顺序不同,检索结果就相差了 1 万多条。和专业数据库如 PubMed 相比,Google Scholar 采用自由词匹配,而 PubMed 优先采用规范的主题词匹配,更容易控制查全率和查准

率。而由于网络原因,国内访问 Google Scholar 需要通过 VPN 访问。

Google Scholar 不是传统意义的文献数据库,他采用自动化爬行互联网的学术文档,然后基于全文建立索引<sup>[9]</sup>。索引数据可能来源于出版社的网页,也可能来自一篇全文,Google Scholar 没有自己的文献数据库,而是直接链接到文献的 web 页面。这种索引机制不能保证一个时间点检索到的内容和另外一个时间点检索的内容完全相同。但也正是由于这个原因,Google Scholar 可能代替 PubMed、EMbase 等数据库的联合检索,用户也不需掌握不同的数据库检索方法。

### 4.1 Google Scholar 检索方法

**4.1.1 数据来源** 从 The Cochrane Library 检索到最新的 10 篇 Cochrane Database Systematic Reviews 文献(2016 年 12 月 1 日),然后从这 10 篇论文中提取其纳入研究和检索表达式。因这些研究的检索表达式一般都采用了截词、逻辑嵌套且字符大大超过了 256 个字符限制,所以要对这些检索式进行修订才能适应于 Google Scholar 检索。修订原则为:检索表达式限制在 256 个字符内,尽可能用简单的逻辑组配连接术语、短语和子表达式,尽可能保证检索结果的敏感性。其中 1 篇系统评价(Antidepressants versus placebo for depression in primary care)的检索式转换如下,框 1 为系统评价中的检索表达式,框 2 为修订后适合 Google Scholar 检索的表达式:

为了检查系统评价中纳入研究是否能在 Google Scholar 中检索到,对修订后的检索式进一步缩小范围,如给检索表达式加入作者或标题的一部分进行确切匹配,框 3 为加入了作者的检索表达式。

**4.1.2 不同检索结果的评价** 采用 Google Scholar 中精确匹配的检索结果 R (如采用框 3 的确切匹配检索式)和 10 篇论文中提取到的纳入研究 Ds 之比表示检索的敏感性(R/Dselect);采用检索结果 R 和 Google Scholar 中检索到所有结果 Dall 之比(R/Dall)表示检索的准确性。通过这两个比值判断 Google Scholar 检索的敏感性和准确性。

10 篇最新系统评价共纳入了 340 个原始研究。我们的检索结果显示,Google Scholar 的每篇系统评价搜索结果在 4 100 ~ 61 000 篇之间,共计 271 430 篇。R/Dall 在 74% ~ 100% 范围内,其中 6 篇系统评价的 R/Dall 值为 100%,R/Dall 整体值为 91%,说明 Google Scholar 检索能够满足系统评级检索准确性。R/Dselect 最小值为 0.10%,最大值为 17.8%,R/Dselect 整体值为 4.7%,说明检索敏感性

### 框 1 系统评价中的检索表达式

Free-text=(antidepress\* OR “monoaminic oxidase inhibitORS” OR “selective scrotonin reuptake inhibtiors” OR “tricyclic drugs” OR acetylcarnitine OR alaproclate OR amersergide OR amiflamine OR amineptine OR amitriptyline OR amoxapine OR befloxtone OR benactyzine OR brofaromine OR bupropion OR butriptyline OR caroxazone OR chorpoixiten OR cilosamine OR cimoxatone OR citalopram OR clomipramine OR clorgyline OR clorimipramine OR clovoxamine OR deanol OR demexiptiline OR deprenyl OR desipramine OR dibenzipin OR diclofensine OR dothiepin OR doxepin OR duloxetine OR escitalopram OR etoperidone OR femoxetine OR fluotraccn OR fluoxetine OR fluparoxan OR fluvoxamine OR idazoxan OR imipramine OR iprindol\* OR iproniazied OR isocarboxazid OR litoxtin\* OR lofepramin\* OR maprotilin\* OR medifoxamin\* OR melitracen OR metapramin\* OR mianscrin OR milnacipran OR minaprin\* OR mirtazapine\* OR moclobemid\* OR nefazodon\* OR nialamid\* OR nomifensin\* OR nortriptyline\* OR noxiptilin\* OR opipramol OR oxaflozan\* OR oxaprotilin\* OR pargylin\* OR paroxetine\* OR phenelzin\* OR piribedil OR pirlindol\* OR pivagabin\* OR prosulprid\* OR protriptylin\* OR quinupramin\* OR reboxetin\* OR rolipram OR sertraline\* OR setiptilin\* OR trimipramin\* OR venlafaxin\* OR viloxazin\* OR viqualin\* OR zimeldin\*) and free-text=placebo\* and keyword=(depress\* OR dysthymi\* OR “adjustment disorder\*” OR “mood disorder\* OR “affective disorder” OR “affective symptomts”) and free-text=(“general practice” OR “primary care” OR “community mental health” OR “family practice” OR “health maintenance organization” OR HMO OR home OR “university clinic” OR private OR ambulatory)

### 框 2 修订后适合 Google Scholar 检索的表达式

Placebo (“anxiety disorder” OR “obsessive-compulsive” OR “post-traumatic”) (randomized OR randomized OR random OR controlled OR “double-blind” OR “clinical trial” OR prospective) (child OR children OR adolescent OR pediatric)

较低。但由于研究方法设计没有考虑到 Google Scholar 对相关文献的排序,所以本实验得到的 R/Dselect 低于实际值。系统评价检索式字符长度 (Lr) 为 141 到 1 800 字符之间,中位数为 810 个字

### 框 3 加入作者的 Google Scholar 检索表达式

Author: blaski depression treatment placebo (antidepressant OR “monoamine oxidase inhibitors” OR “selective serotonin reuptake inhibitors” OR “tricyclic drugs”) (“general practice” OR “primary care”) (randomized OR randomized OR random OR trial)

符,修订后适合 Google Scholar 检索的表达式长度 (Lg) 为 93 ~ 220 个字符之间,中位数为 180 个字符,系统评价检索式字符长度和修订后适合 Google Scholar 检索的表达式长度 (Lr/Lg) 比值范围为 0.09 ~ 0.37,平均值为 0.2。

## 5 资源发现系统检索工具

近年来各高校图书馆为了提高资源利用率和用户体验,纷纷引进了资源发现系统,国内引进的发现系统主要为 Primo、EDS、Summon、Worldcat Local 四种系统。资源发现系统和 Google Scholar 类似,都为用户提供了一站式检索各种筒仓数据的功能。但资源发现系统在某些方面优于 Google Scholar,如没有 256 个字符限制,提供了类似于学术数据库的高级搜索、方面搜索等,检索功能更接近传统的学术数据库。本节采用了以上类似方法评价资源发现系统是否能单独作为系统评价的检索工具,方法学部分同上。

由于不同单位采用的发现系统不同,且使用相同的发现系统的不同单位,其资源配置也不同,导致收录范围和最终检索结果都不同,故本研究仅评价了西安交通大学图书馆的 Summon 资源发现系统。

检索同样的 10 篇系统评价纳入的 340 个研究,发现系统检索到的结果在 2 200 ~ 7 800 篇之间,共计 4 700 篇。R/Dall 在 75% ~ 100% 范围内,其中 4 篇系统评价的 R/Dall 值为 100%, R/Dall 整体值为 76%,资源发现系统检索能够满足系统评级的检索敏感性。R/Dselect 最小值为 0.20%,最大值为 31%, R/Dselect 整体值为 12%,该值虽然高于 Google Scholar,但和系统评价要求的检索精确性还有很大的差距。系统评价检索式字符长度 (Lr) 为 141 到 1 800 字符之间,中位数为 810 个字符,修订后适合资源发现系统检索的表达式长度 (Lg) 为 93 ~ 310 个字符之间,中位数为 237 个字符,系统评价检索式字符长度和修订后适合资源发现系统检索的表达式长度 (Lr/ Lg) 比值范围为 0.13 ~ 0.48,平均值为 0.30。

## 6 小结

本文分析了系统评价检索的敏感性和准确性,在进行系统评价时,应将布尔逻辑检索和等级检索结合起来,采用布尔逻辑检索获取相关文献,利用等级检索筛查相关文献。建议进行字段和元数据搜索,而不需要进行全文检索。因一站式检索系统 Google Scholar 和资源发现系统能够满足检索敏感性,但精确度较低,所以进行系统评价时,仍需进行传统数据库的检索,而将一站式检索系统 Google Scholar 和资源发现系统检索作为补充。

### 参考文献

- 1 陈匡阳, 屈丽娜, 胡芳, 等. 动物实验系统评价/Meta分析检索策略报告情况调查. 中国循证医学杂志, 2016, 16(3): 348-353.
- 2 葛龙, 安妮, 曾巧铃, 等. 我国干预类系统评价/Meta分析文献检索新挑战. 中华医学图书情报杂志, 2013, 22(5): 2-8.
- 3 Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. *Stud Health Technol Inform*, 2004, 107(Pt 1): 311-316.
- 4 Wilczynski NL, McKibbin KA, Walter SD, *et al*. MEDLINE clinical queries are robust when searching in recent publishing years. *J Am Med Inform Assoc*, 2013, 20(2): 363-368.
- 5 Wilczynski NL, Haynes RB. Optimal search filters for detecting quality improvement studies in Medline. *Qual Saf Health Care*, 2010, 19(6): e31.
- 6 Lokker C, Haynes RB, Wilczynski NL, *et al*. Retrieval of diagnostic and treatment studies for clinical use through PubMed and PubMed 's clinical queries filters. *J Am Med Inform Assoc*, 2011, 18(5): 652-659.
- 7 Srikantaiah KC, Srikanth PL, Tejaswi V, *et al*. Ranking search engine result pages based on trustworthiness of websites. *Int J Comput Sci*, 2012, 9(4): 318.
- 8 翟中会. 提升图书馆资源可用性策略研究. *图书馆理论与实践*, 2015, (3): 89-92.
- 9 薛俊峰, 张梅, 牛晓芳. Google Scholar与PubMed搜索比较. *图书情报工作*, 2008, 52(7): 133-135.

收稿日期: 2017-09-15 修回日期: 2018-03-21

本文编辑: 樊斯斯