

3 大常用引文数据库检索中医药文献功能探究

周 晴 邵明坤 余恒先 蒯国斌

(上海市中医文献馆 上海 200020)

〔摘要〕 选取《中国生物医学期刊引文数据库》(CMCI)、《中国引文数据库》(CCD) 以及《中医药引文数据库》(TCMCD) 为研究对象, 从收录范围、检索方式、检索结果(查全率、查准率)、输出格式等各方面比较 3 大常用引文数据库在检索中医药文献方面的不同, 分析其优缺点和产生的原因, 提出修改建议。

〔关键词〕 引文数据库; 中医药; 检索功能

Discussion and Research on Retrieval Function of the Three Popular Citation Databases in Traditional Chinese Medicine Literature ZHOU Qing, SHAO Ming-shen, YU Heng-xian, KUAI Guo-bin, Shanghai Literature Institute of Traditional Chinese Medicine, Shanghai 200020, China

〔Abstract〕 The paper chooses Chinese Medical Citation Index, Chinese Citation Database and Traditional Chinese Medicine Citation Database as the research targets, compares the differences among the popular three citation databases from collected scopes, retrieval methods, retrieval results (recall ratio, precision ratio), results output format on retrieving traditional Chinese medicine literatures, analyzes the advantages and disadvantages and their causes and proposed suggestions for modifying.

〔Keywords〕 Citation database; Traditional Chinese medicine; Retrieval function

1 前言

引文数据库是基于文献之间的引用与被引用的关系, 由来源文献和被引文献 (Citation) 有序合理组合形成的二次文献^[1]。随着计算机技术的发展, 各家研制的引文数据库如雨后春笋发展迅猛, 各数据库在收录范围、检索效果等方面都有一定的差别。本文从引文检索功能方面对查询中医药论文时常用的中文引文数据库 (包括上海市中医文献馆自行开发的数据库) 进行比较研究, 包括收录范围、

检索方式、检索结果 (查全率、查准率)、检索结果输出格式等, 以使用户了解它们的优点和缺点, 从而更有效地选择和利用。

2 资料与方法

选取解放军医学图书馆研制的《中国生物医学期刊引文数据库》(CMCI)^[2], 清华同方研制的《中国引文数据库》(CCD)^[3], 上海市中医文献馆和重庆维普公司共同开发的《中医药引文数据库》(TCMCD)^[4] 进行分析。

1973 年美国情报学家 Lancaster 和 Fayen 列出了衡量信息检索系统效果的 6 项指标, 即收录范围、查全率、查准率、响应时间、用户负担和检索结果输出格式^[5]。因客观因素的限制, 本文仅选用收录

〔修回日期〕 2009-07-09

〔作者简介〕 周晴, 副主任医师, 长期从事中医药文献检索工作, 发表论文 7 篇。

范围、查全率和检索结果输出格式 3 个指标，并另外补充了检索字段、检索年限^[6]两项检索方式指标进行评价。

3 结果与分析

3.1 收录范围

表 1 3 大引文库收录范围			
项目	CMCI ^[7]	CCD	TCMCD
收录期刊种类	1 000 种医学期刊，其中中医类期刊 156 种	整刊 1 500 多种，部分刊 4 000 多种，其中中医类期刊 154 种	中医类期刊 130 种
收录期刊年限	1995 年至今	1912 年至今	1999 ~ 2006 年
数据量	270 万条来源数据，330 万条引文数据	390 万条来源数据，684 万条引文数据	34 万条来源数据，78 万条引文数据

3.2 检索方式

表 2 3 大引文库检索字段			
项目	CMCI	CCD	TCMCD
来源文献检索字段	题名、作者、单位、刊名 + 年份、检索词（关键词）	题名、作者、单位、刊名 + 年份、关键词、主题、中图分类号	题名、作者、单位、刊名 + 年份
引文检索字段	题名、作者、单位、刊名 + 年份	题名、作者、单位、关键词、摘要、中图分类号、被引频次数	题名、作者、单位、刊名 + 年份

从表 2 可知，CCD 的检索字段最多（14 个），CMCI 次之（9 个），TCMCD 最少（8 个）。来源文献和引文可检索字段的数量排序与之相同。CCD 很多字段是它特有的，如关键词、中图分类号、摘要、被引频次数。其他两个数据库 CMCI 和 TCMCD 也具备了最基本的检索入口，给用户提供了准确的选择。这 3 个数据库都没有提供全文检索字段，这是非专业检索用户最易用的，可以提高查全率，但相应降低了查准率。

3.3 检索结果（查全率）

3 个数据库共有检索字段为题名、作者、单位、刊名 + 年份，避免因各数据库收录范围不同以及录入篇名时存在差异，选择第 1 作者（同一单位）字段进行比较。为了消除数据库间收录年限不统一的因素，将来源文献年限定为 1999 ~ 2003 年（来源于期刊），分别在 3 个数据库中作者字段输入作者姓名进行检索。

进入统计的数据必须符合 4 个条件：引文作者为第一作者；引文单位是上海市中医文献馆；来源期刊年限为 1999 ~ 2003 年；来源文献不包括图书。

从表 1 可以看出，在收录期刊的种类方面，CCD 要优于另两个数据库，但中医类期刊与 CMCI 基本相同，TCMCD 收录期刊最少；在收录文献的年限和数据量方面，CCD 的收录年限最长，数据量最大，TCMCD 年限最短，数据量最少。

选择上海市中医文献馆已退休的 11 名职工作为被引作者。为了避免有非上海市中医文献馆的同名作者出现，笔者又通过《中国生物医学文献数据库》（CBMdisc）核对检出的被引文献，以保证引文作者确为上海市中医文献馆职工。检出结果后计算查全率。计算公式为：查全率 = （各个数据库检出的被引次数/被引总次数）× 100%，被引总次数为查重后各个数据库中检出的被引次数之和。

各数据库检出的被引次数及查全率有较大差异。如 CMCI 中“周家珩”的查全率是 0，CCD 的查全率为 100%。综合 11 名作者的被引结果进行比较，对查全率进行汇总，得出结论：CCD 的查全率最高（68%），TCMCD 次之（54%），CMCI 最低（39%）。在检索过程中，发现 CCD 数据库包含了《中国优秀硕士学位论文全文数据库》、《中国优秀博士学位论文全文数据库》和《中国重要会议论文全文数据库》，尤其是前两个全文数据库文章很多，但未在期刊上正式发表，所以是否进入检索结果存在着分歧^[8]。含有硕博论文的 CCD 检索结果与剔除硕博论文后的检索结果差异很大。

此时，除了 CCD 查全率下降外，CMCI 和 TCM-

CD 的查全率都有了很大的提高。TCMCD 的查全率从原来的 54% 一跃成为最高（70%），CCD 次之（59%），CMCI 查全率虽有提高，但仍为最低（50%），这可能与 CMCI 对刊种和文献的高度选择

性有关。《中国优秀博硕士学位论文全文数据库》是《中国引文数据库》（CCD）优于其他数据库查全率的关键。如果没有硕博论文库，CCD 的查全率将屈居第二，略微高于 CMCI。

表 3 3 大引文库漏检原因						
数据库	期刊收录不全		文献收录不全		参考文献著录错误	
	次数	比例（%）	次数	比例（%）	次数	比例（%）
CMCI	1	2	2	3	61	95
CCD	0	0	1	2	52	98
TCMCD	37	95	0	0	2	5

分析各数据库检出的 11 名作者被引次数不同的原因。从表 3 可知，第一，CMCI 和 CCD 的期刊收录总数是很齐全的，缺漏比例很低，而 TCMCD 在收录期刊时经过严格的筛选，收录期刊具有高度选择性，所以几乎所有的漏检都是由于期刊不全造成的。第二，在期刊收录齐全的基础上，CMCI 和 CCD 都有漏期的现象出现，而 TCMCD 在这方面比较严格，有其刊必有其期。第三，CMCI 和 CCD 参

考文献著录失误较多，是造成漏检的主要原因^[9]。有关参考文献著录错误的原因，可大致归纳为以下几种情况：单位名称不统一；个别数据著录时有作者名字，但单位名称为空；作者姓名写错；文章标题写错，尤其是标题中含有“的”字，经常有数据库任意添加或删除。

3.4 检索结果输出格式

表 4 3 大引文库检索结果输出格式						
项目	CMCI		CCD		TCMCD	
每页显示条目数	10		20		15	
是否提供全文	否		是（付费）		是（付费）	
是否需下载浏览器	否		否		否	
显示来源文献字段	题名、作者、单位、期刊、 年期页、资助、参考文献数		题名、作者、单位、期刊、 年期页、资助、关键词、摘 要、分类号		题名、作者、单位、期刊、年期页、 具体引用的参考文献	

从表 4 可以看出，CMCI 每页显示条目数最少，CCD 显示最多，且根据不同的来源数据库进行分列。3 者的全文浏览是有限制的，有的没有此项功能，有的需要付费。CCD 显示字段最多，能给用户提供较多的信息，方便用户判断相关文献，尤其是文摘对用户判断相关文献是很重要的。TCMCD 还列出了引证文献所有参考文献的具体信息。

4 讨论

4.1 本次研究的不足

一是因为客观因素的限制，本次研究未纳入《中国科学引文数据库》（CSCD），对各数据库的响

应时间、用户费用负担等指标没有进行比较。二是在检索功能方面，没有对各数据库是否能运用布尔逻辑符、是否有特殊功能和二次查询进行分析。三是在检索时虽然发现个别数据库在标引方面有些问题，如各个数据库篇名的标引有微小差别，可能会导致检索结果的差异，但并没有深入研究。四是对来源文献的年限进行限定可能会带有一些主观成分，会影响到分析的结果。

4.2 3 个数据库的不足

给用户提供多途径的查询是评价数据库的一个重要指标；方便用户检索的字段设置也同样重要。对非专业检索人员来讲，有许多检索字段，如表 2

所示的中图分类号、关键词等,用处并不大。而全文检索是最方便用户使用的,就像互联网上常用的谷歌(Google)、百度(Baidu)一样,但现在还没有一个数据库有这项功能,这说明这几个数据库的建设还不完善。比较检索结果后发现,这3个数据库的查全率都不高,说明各数据库在引文检索功能方面比较欠缺。

CMCI 有独特的“智能模糊扩展搜索功能”^[10],但查全率最低,说明参考文献著录方面失误较多。另外,CMCI 有复合检索项,但用第一作者和单位复合检索后,没有检索结果,程序出错,所以无法判断作者的单位一项,无法纠正重名作者,笔者只能通过其他的题录数据库逐篇加以确定。

CCD 期刊总数最全,文献最多,但主要依靠独有的硕博论文库,如果没有这两个库,查全率明显下降,它将搜索结果和全文数据库相关联,只显示在全文库中能找到的数据,所以造成大量的引证文献看不到。

TCMCD 虽然查全率最高,但是数据量太少,年限太短,还需要大量积累。而且,虽然有被引作者和被引单位的复合检索,但有些单位为空,无法100%查准,只能通过题录库对单位为空的文章加以确定。

5 结论

收录范围:在刊种方面,CCD 要明显优于其他库,但中医类期刊 CCD 和 CMCI 基本相同,TCMCD 最少;在年限和数据量方面,CCD 量最多,TCMCD 量少。检索方式:在检索字段方面,CCD 提供的最多,TCMCD 最少。检索结果:如果包含硕博论文库,CCD 的查全率最高(68%),如果不包含,TCMCD 的查全率最高(70%),两种情况下 CMCI

的查全率都是最低。结果输出格式:CCD 的显示和输出字段最多。

建议 CMCI 和 CCD 要加强对参考文献的纠错,对收录的期刊尽量收齐,不要缺期,保证数据库的查全率。CCD 要增强网络稳定性,因为在查询中,经常发生网页停滞无反应现象,重新启动浏览器后才能恢复正常使用。TCMCD 要扩大期刊的收录种数,延长年限,加快数据更新速度。

参考文献

- 1 赵蓉英,雷将,马瑞敏,等.我国五大数据库引文功能的比较研究[J].情报理论与实践,2008,9(4):589-593,605.
- 2 <http://156.156.156.111:1012/cdweb/leadpage.htm> [EB/OL]. [2009-6-1].
- 3 <http://ref.cnki.net/knsref/index.aspx> [EB/OL]. [2009-6-1].
- 4 <http://222.177.19.132:85/> [EB/OL]. [2009-6-1].
- 5 (美) F. W. Lancaster 著.陈光祚,王知津,王津生,译.情报检索系统——特性、试验与评价[M].北京:书目文献出版社,1984:5.
- 6 王飞,侯跃芳,崔雷.常用中文文献数据库引文检索功能的比较研究[J].中华医学图书情报杂志,2005,14(4):52-56.
- 7 楼海萍,杜娟.国内两大引文数据库的比较研究[J].中华医学图书情报杂志,2004,13(5):39-41.
- 8 史春扬,叶协杰.应用《中国引文数据库》分析高校发表论文被引状况[J].医学信息,2007,20(5):768-770.
- 9 董时军,张艳.医学期刊论文参考文献错误分析[J].预防医学情报杂志,2008,24(4):298-300.
- 10 王明洁.《中国生物医学期刊引文数据库》在医院的应用[J].医学信息学杂志,2007,28(3):250-251.