基于引文统计分析的西文期刊馆藏发展策略

张建勇　刘筱敏
中国科学院文献情报中心　北京 100080
〔摘要〕 运用引文统计分析法研究我国科研人员对西文期刊的引用情况，并据此获得理想的西文期刊馆藏数量。提出需求满足
率的引文频次算法，论证阅读数和引文数及学科对引文的影响等因素，分析引用频次的时间分布状况。提出基于定量分析的馆
藏建设指导原则。
〔关键词〕 引文统计　馆藏建设　用户需求
〔分类号〕 G253

Development Strategy of Western Periodicals Collection Based on Citation Statistic Analysis

Zhang Jianyong　Liu Xiaomin
Library of Chinese Academy of Sciences, Beijing 100080
〔Abstract〕The paper analyzed the status of researchers use western language periodicals as reference in China by applying citation statistic
methods. As an analysis result, the reasonable amount of western language periodicals collection in library was gotten. Citation frequency analysis
algorithm that satisfies the readers' needs was offered in this paper. As the effect factors, the amount of citations and papers read, and the kinds
of scientific field, were demonstrated. The temporal distribution of the citation frequency was also analyzed. The strategy of how to develop
library collection was provided based on above quantitative analyses.

〔Keywords〕citation analysis　collection development　user requirement

1. 引言

图书馆馆藏发展始终是需要图书馆重点考虑的问题，国内外各个图书馆都制定了自己详细的馆藏发展政策和发展重点，在用户需求和资源利用环境发生改变的情况下，馆藏发展面临诸多的挑战。

经费预算是图书馆馆藏发展最具有影响的因素，经费的多少决定着图书馆馆藏的数量和质量。据 ARL（美国研究型图书馆协会）的最新数据统计，2003 年美国研究型图书馆用于购买期刊的费用相对于 1986 年而言虽然增长了 260%，但订阅的期刊总数只比 1986 年增加了 14%[1]。

用户阅读模式的变化和出版模式的变化也是影响图书馆馆藏模式的重要因素，用户更多地利用网络工作和学习，拥有与获取的争执在也逐渐向获取方式倾斜。多元化的出版模式成为主流。其中网络版期刊的品种在不断增加。

图书馆如何在一定的经费预算的情况下，制定合理的馆藏发展策略，购买足够的资源，

选择恰当的资源，找到经费预算和购置资源数量质量间的平衡点，成为满足用户需求，促进图书馆可持续发展的重要论题。

期刊的利用状况，最好的一个例证是发表论文中引用文献的数量。通过引用量的统计可以测评用户真正使用期刊的品种，并发现潜在资源。本文基于引文统计分析，选择科技期刊为研究对象，分析我国科研人员过去 5 年（2000-2004 年）在国内外科技期刊上发表的学术论文中引用过去 10 年（1995-2004 年）西文期刊的频度表，来分析我国科研人员实际的文献需求，分析阅读行为与引用行为间的关系，分析阅读与引用文献的学科差异等因素，由此探讨如何优化馆藏发展策略。

2. 引文统计与结果分析

2.1 引文统计结果

本次统计数据来源为中国科学引文数据库（简称 CSCD）与 Science Citation Index（简称 SCI）光盘版两个数据库，这两个数据库都是综合性的、收录自然科学和工程技术类期刊论文的引文数据库。CSCD 收录了我国出版的优秀期刊 1046 种[2]。SCI 收录全球出版的核心期刊 3700[3]余种，集中了我国以及世界优秀的科研论文，通过对文后参考文献的统计，我们可以发现科研人员使用期刊一些客观数据。数据统计范围为 2000 年—2004 年我国科研人员在国内外科技期刊上发表论文中引用 1995 年—2004 年的西文期刊，两个数据库中我国科研人员 5 年的发文的数量为近 60 万篇，引用过去 10 年的西文期刊论文数为 90 万篇，采用计算机模糊匹配和人工判断的方式对 90 万篇论文的母体文献即期刊刊名进行规范化处理，最后的数据收拢，得到了被引用的西文期刊引用频次数据，将数据集合按期刊被引用频次降序排列后形成引用频次和引用期刊的数据表。引用频次是文献计量的重要指标，常用于科研绩效评价，它证明了一种期刊在学术研究中的作用和影响。西文期刊的引用频次，反映了科研人员对该类资源的利用程度，在某种程度上利用程度的大小反映了我国科研人员实际的文献需求程度。

以某种期刊引用频次（反映需求量）占总所有期刊引用频次（反映总需求量）的比例作为某种期刊的需求满足率，为便于分析比较我国科研人员使用期刊的集中程度，我们将引用频次由大到小按序排列后，计算出每种期刊的需求满足率，并将需求满足率进行累计，分别截取累计需求满足率为 75%，80%，85%，90%，95%，98%相应期刊数据，观察和分析需求满足率对应的引用西文期刊品种数分布。

表 1.需求满足率与引用期刊数量关系表

满足率%	引用期刊数

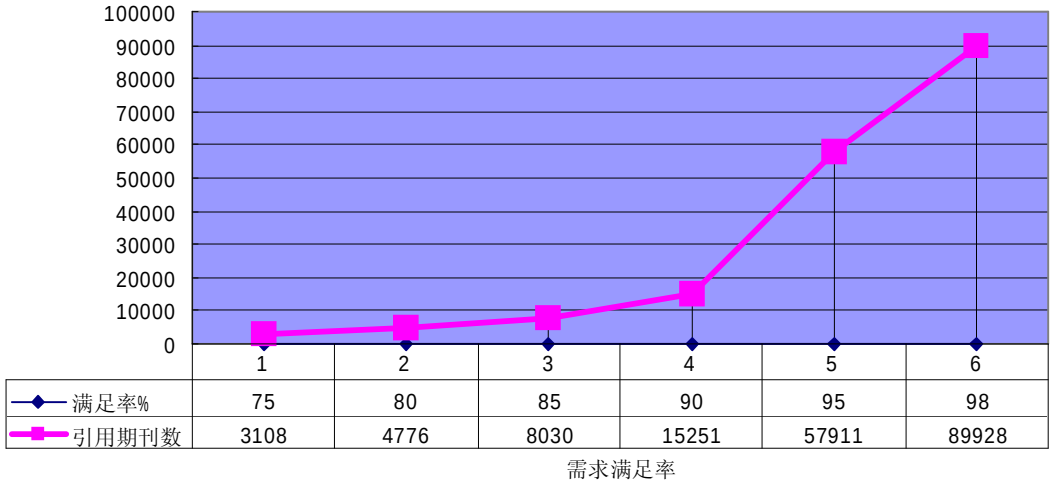
75	3108
80	4776
85	8030
90	15251
95	57911
98	89928

从统计表的结果看，需求满足率与引用西文期刊的数量曲线是符合布拉德福定律的，科研人员引用频次高的西文期刊呈高度汇聚的状态，累计满足率达 80% 时的期刊数量不足期刊总数的 5%，引用频次低的西文期刊呈离散状态，期刊种类高度分散。

通过引文数据可以得到西文期刊的引用情况，但无法判断作者引用的期刊是否还在正常出版，为给西文期刊的订购提供可靠的依据，我们利用 Ulrich’s 数据库，选取其核心数据库的期刊进行比较，这些期刊均是自然科学及其相关领域正在出版的期刊（现刊），共计 36318 种[4]。用这个期刊集合与引用期刊集合品种进行比较分析，发现引证期刊的需求满足率的前 75%中，有 3000 种期刊左右在 Ulrich’s 的核心期刊集合中。需求满足率中的期刊分布与 Ulrich’s 核心数据库中的期刊高度吻合。

需求满足率和引用关系图示图 1。

图 1. 需求满足率与引用期刊数关系图



2.2 阅读期刊与引用期刊间的相关关系分析

引文数据反映了期刊在科研工作中产生的影响，但引文分析也具有一定的局限性，引文行为与阅读行为存在一定的差异，引文分析在一定程度上缩小了期刊的使用品种以及期刊的

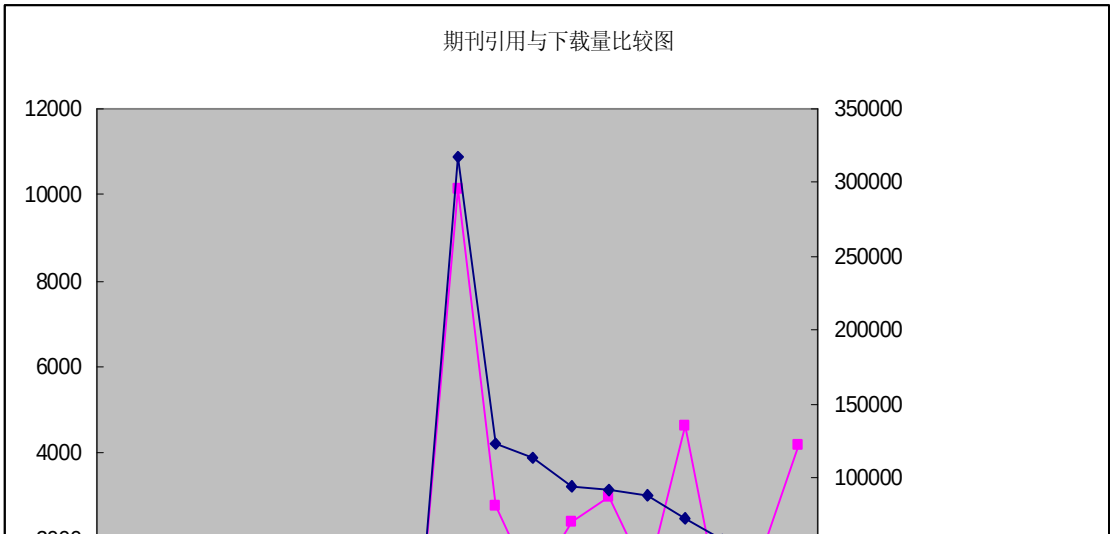
使用次数，一般情况下实际引用文献量要小于实际阅读文献量。为解决这个问题，我们通过在对中科院已订购的 ACS、SpringerLink 两个全文网络数据库中随机抽取 19 种期刊进行下载量统计，并比较这些期刊在中国科学引文数据库（CSCD）及中科院科研人员在 SCI 期刊上发表论文中的引用统计，这一点得到了证实。

表 2. SpringerLink、ACS 期刊引用量、下载量抽样统计表

journal	年均引用量	2005 年下载量
A1	1.33	480
A2	12.33	495
A3	139.33	1779
A4	46.33	2577
A5	60.67	3353
A6	49.67	585
A7	610.00	9657
A8	21.33	48
A9	33.33	645
B1	10148.00	317013
B2	2759.33	123412
B3	898.33	113253
B4	2385.00	94098
B5	2992.00	91820
B6	1134.00	88156
B7	4630.33	72022
B8	260.33	57570
B9	1666.00	53970
B10	4181.33	47597

年均引用量的产生依据是利用中国科学引文数据库（CSCD）和 SCI 数据库，对 2001-2004 年间中国科学院的作者在上述两个库的来源期刊上发表的论文中引用 1995-2004 年之间的抽样期刊论文频次除 10 年得到的平均数。2005 年下载量为 2005 年年度中国科学院用户使用网络数据库对抽样期刊的下载文献数量的统计。从统计数字看，实际引用文献数量远远小于实际阅读量，年平均阅读量为年平均引用量的 60 倍。

图 2.期刊引用与下载相关关系图



对 19 种抽样期刊的年均引用量和下载量比较，看两组数据的契合程度，见图 2，期刊引用与下载相关关系图，通过图 2 可以看出文献阅读行为与引用行为有比较强的相关关系，引用与下载量的曲线表示出当某种期刊的阅读量大时，其引用量也比较大，但也有反常情况出现，从期刊 B2-B10，可以看出引用行为与阅读行为之间的关系的非一致性，阅读量与引用频次之间关系并非线性比例关系，可能数据库开通时间和通过印本阅读的因素在影响该相关关系。引入相关函数

$$s_{x,y} = \frac{Cov(X,Y)}{s_x \times s_y}$$

计算引用行为与阅读行为之间的相关关系，两者的相关系数为 0.88，呈高度相关的状态。一般情况下实际阅读文献数量要大于实际引用文献数量，引用频次高的期刊，实际阅读频次也高，引用频次较低的期刊，阅读频次也低。根据文献计量学的布拉德福定律，针对某一学科领域用户完成一项科研任务的核心文献需求而言，阅读高引频的文章以及期刊即可满足其约 80%的核心需求；相反，若满足所有需求则需阅读大量的低引频文章和期刊。

因此，引入阅读离散度指标对引用品种数进行调整，高引频的期刊，阅读离散度相对较低，引用频次升高，其阅读离散度明显升高，当需求满足率达到 90%时，阅读离散度达到峰值，其后逐渐降低至 1（即以 90%为峰值点的正态分布）。

2.3 学科因素与引用之间的关系分析

另外，目前统计数据依赖 CSCD 和 SCI，这两个数据库在收录期刊的学科分布上具有不均衡性，化学、物理、生命科学等学科的期刊占有较高比例，同时每个学科的引用行为不同，会深化这种学科分布的不均衡，因此在构建外文期刊馆藏时，必须要考虑学科因素，纠正数据分析带来的偏差。表 3 统计了从 Ulrich's 系统核心集中筛选出的部分学科领域涉及到的期刊数，与上述领域在 CSCD 和 SCI 数据库有引文的期刊对比，引用期刊数大大低于核心期刊集，见表 3。

表 3．学科分布比较

DDC 学科分类	Ulrich's 核心集期刊数	引用期刊数
编程	98	26
程序	4	2
初等数学原理	54	27
处理模式	6	3
瓷器和相关技术	115	26
磁学	20	8
代数学	35	18
蛋白质	34	14
地理和历史	3765	131
地球科学	548	118
地质学	645	189
电技术、电子技术、超导	7	2
电学和电子学	63	17
动物学	694	190
动物遗传与进化	87	33
多媒体	14	3
发明和专利	60	1
纺织工业	140	8
分析化学	132	61
概率论与应用数学	167	86
工程材料	249	106
工程技术	520	118
古生物学;古动物学	117	52
光学	139	63
海洋工程	30	3
海洋生态学	6	3
海洋学	225	50
航空、航天工程	315	36
核工程	55	13
化学工程、化学工业	291	80

从表 3 可以看出，各学科引用行为是有差异的，在馆藏建设中要根据学科重点的不同对期刊收藏指标进行不同的加权，我们建立了学科调节因子的指标，学科调节因子为某学科购进期刊数量占用户引用期刊的种数的比例，如果该比例高则表示学科满足度比较好，如果该百分比较低，则表示学科满足度不理想，需要进行调整、增加新的期刊品种。在本研究中涉及到的学科调节因子，目的是来调整合适的馆藏数量指标。引用文献品种数越多，覆盖的研

究领域也越大，学科调节因子值则越小，当期刊数量达到 30000 种时，学科调节因子值降为 1。

2.4 调整后的数据结果

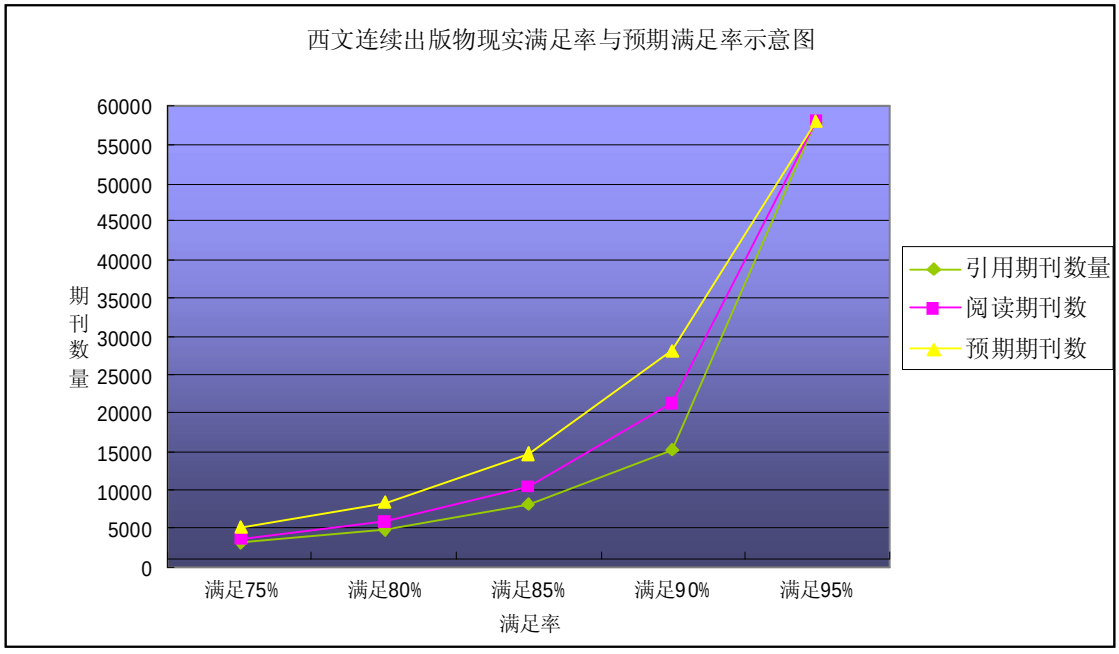
经过对引文反映的需求与引用期刊、阅读期刊及学科因素的分析，我们经合理调整，建立我国科研人员需求满足率与西文科技期刊数量的对应关系。我国西文科技期刊资源建设的目标至少应满足国内科研用户 90%的需求，西文期刊数量为 28183 种是一个比较理想的建设目标。见表 4。

表 4.需求满足率与引用期刊、阅读期刊和预期期刊数量关系表

满足率%	引用期刊数	阅读离散度	阅读期刊数	学科调节因子	预期期刊数
75	3108	1.10	3418	1.47	5025
80	4776	1.20	5731	1.45	8310
85	8030	1.30	10439	1.41	14718
90	15251	1.40	21351	1.32	28183
95	57911	1.00	57911	1.20	69493
98	89928	1.00	89928	1.00	89928

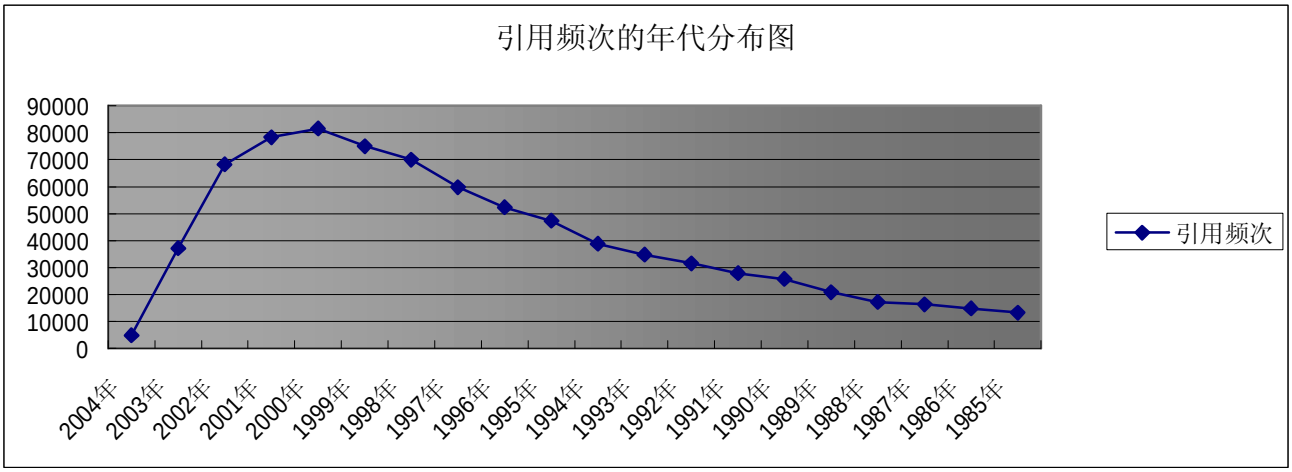
阅读离散度的值的高低是经验值，但更多是基于对网络期刊下载数量的观察，学科调节因子主要是针对总体情况，具体到某个学科时，要有不同的考虑。

图 3．西文期刊需求满足率与预期期刊品种关系图



2.5 引用频次在时间上分布

以往的馆藏发展多考虑的是现在和将来出版的资源，随着网络出版物的增多，在网上提供回溯档的出版商越来越多，到底这些回溯档的需求情况如何呢？从表中可以看出引用年代分布与期刊影响因子的经典计算方法有相同之处，引用高峰值集中在最近的 2 年。但随后几年的引用频次仍然保持较高的数量，呈缓慢下降的趋势。比较老的文献仍然有很高的引用数量，说明用户对回溯档的需求比较明显。



3. 西文期刊的馆藏发展策略

以需求驱动图书馆的资源建设是图书馆资源建设的出发点，使用期刊的引证指标作为西文期刊资源需求的客观数据，是了解用户需求比较实用的方法。通过对引文量的调整，包括阅读量的调节和学科调节因子的分析，以找到一个更接近事实的需求数据，从而构建一个相对比较实用的西文期刊采购策略。

3.1 总量满足原则

任何时候，图书馆拥有西文期刊的种数都是满足用户需求的一个关键的数字，从上面分析得出的数值，应该具有一定的指导意义，满足率为 75%时，预期期刊数为 5025 种，满足率为 80%时，预期期刊数为 8310 种，满足率为 85%时，预期期刊数为 14718 种，满足率为 90%时，预期期刊数为 28183 种，满足率为 95%时，预期期刊数为 69493 种。图书馆收藏期刊的种数要达到合适的总量，满足用户的合理的文献需要，从当前的实践看，满足率为 90%时，用户对图书馆的满意度会比较高。特别是从国家保障的角度看，90%的目标也是一个基本的目标。当然这个数字并不代表捆绑和拼凑的期刊种数，依然是经过选择的高引频的期刊。但总量是保持平衡的。

3.2 合作共享原则

从任何一个单馆的馆藏都无法满足用户 100%的需求，而合作共享则有可能满足用户100%的需求。小而全，大而全的馆藏模式导致的后果就是总量的不足，差别化才可能实现合作，也有共享的基础。基于学科特点选择单馆的期刊，参考所服务用户的引用期刊的频次等指标，制定馆藏采集政策。

3.3 拥有和获取相结合原则

网络出版模式和印本出版模式复合的时代，但从印本出发考虑问题肯定是多考虑拥有的方式，而网络出版模式，带给图书馆的多是方便的获取。无论是拥有还是获取，可提供给用户的资源就是能满足用户的图书馆馆藏。

3.4 全面保障原则

既要考虑期刊的总数，而且要考虑它的时间序列，回溯档同样具有需求，对回溯档的馆藏和现刊的采访给予同样的重视。

3.5 整合化建设原则

图书馆的馆藏建设要脱离过去那种单一的购买资源提供服务的方式，资源建设要从建设供应渠道出发，不仅是购买资源，也包括合作购买，交换，开放获取，原文传递等全方位的建设。过去用户到馆使用文献的情况，逐渐被资源开通到桌面的方式所替代。资源购买不足的情况下，采取合作购买和原文传递的服务模式解决用户的需求。这些服务模式的产生和发展，为图书馆馆藏发展提供了发展的广阔空间。

4. 结语

本文通过引文统计分析法来研究我国科研人员对西文期刊的引用情况，并据此获得理想的西文期刊馆藏数量。文章提出了需求满足率的引文频次算法，论证了阅读数和引文数间，学科对引文的影响，引用频次在时间上分布等因素。引文统计分析方法用于馆藏发展的研究是一个探索，需求满足率和预期期刊数的关系的分析，学科对引文频次的影响的分析，引文频次的时间分布状况分析仍需要深入研究。

参考文献

1. http://www.arl.org/stats/arlstat/graphs/2003/monser03.pdf
2. http://www.sciencechina.ac.cn
3. http://www.isinet.com
4. http://www.ulrichsweb.com/ulrichsweb/
5. Carol Hansen Montgomery，Measuring the Impact of an Electronic Journal Collection on Library Costs. D-Lib Magazine，October 2000
6. Diann Rusch-Feja，Evaluation of Usage and Acceptance of Electronic Journals. D-Lib Magazine，October 1999
7. Donald W. King ，After Migration to an Electronic Journal Collection. D-Lib Magazine，December 2002
8. Donald W. King，Measuring Total Reading of Journal Articles. D-Lib Magazine，October 2006
9. 邵晋蓉，高校图书馆重点期刊优化配置模型研究．情报杂志，2004（6）：96-97，100