基于引文分析的期刊论文文献需求与回溯保障的案例研究

——以上海 J 校高水平论文为例

郭承鹭 （上海交通大学图书馆）

摘要：本文通过引文分析法，分析高校图书馆用户对过往文献的文献需求，根据引用行为模式，对未来期刊需求做出预测。根据研究结果，论文的所属学科、论文的跨学科性或引用的跨学科性、论文被引文献的文献类型对引用间隔产生显著影响；而期刊的各项 JCR 指标不能作为是否需要回溯保障的预测变量；据此提出，相应的馆藏建设建议，平衡保障期刊的未来和当前需求。

关键词：引文分析 文献回溯 馆藏建设 文献保障率

1 引言

高校图书馆作为高校的文献资源保障机构，负有保障高校用户文献需求的使命和责任。对于高校图书馆来说，支持教学和科研一直是其职责所在 [1]。而国内的过往历史文献保障存在一定问题。这些问题可能来自三种情况：

一则，外文期刊历史采购存在一定缺失。新中国刚建立的 50 年代初期，由于国力有限和西方国家的封锁政策，当时的文献保障存在缺失；在十年动乱时期，外文期刊采购也一度中断；20 世纪末期又由于经费不足等问题，外文期刊采购不足。种种历史原因，导致我国的过刊保障存在一定缺失（郑建程，袁海波，2010）[2]。

二则，没有购买永久使用权。停止购买数据库时就不再保障相应期刊。三则，虽然购买了永久使用权，但是电子资源的长久保存和管理是难题和挑战（Walters，2013）[3]，这个问题是国内外学术图书馆普遍面对的。

对于历史原因导致的馆藏问题，在对国内高校图书馆进行文献调研时，肖珑等 [4]发现国内高校图书馆对早期人文社科文献资源的收藏率和保障较低；李峰，涂文波 [5]则认为国内高校图书馆的外文图书保障存在结构性缺失。

针对这种情况，国家机构或部分高校图书馆采取了回购措施。如国家科技图书文献中心（NSTL）建设回溯期刊库（郑建程，袁海波，2010），为高校图书馆用户获取过刊提供了一定保障。尽管如此，由于普遍存在的图书馆经费紧张，一个图书馆要如何平衡当前需求与过刊保障依然需要谨慎决策。

Lamothe [6]比较了图书馆的静态电子书合集和动态合集，其中前者指在一定时期内图书馆持续拥有的电子馆藏，后者指在一定时期内持有内容在变化的电子馆藏，发现静态合集中每种书的平均引用量持续下降，而后者在上涨。

但这并不意味着过于早期文献应该完全放弃。由于难以确定什么是早期文献中有价值的一部分——判断的困难来自即使用户也不知道以后的研究中需要用到什么过去的文献。从根

源上说，这与近年来纸本期刊大量被电子期刊取代、电子馆藏的建设指数级增长不无关系[7]。图书馆大量提供电子期刊访问而不是纸本期刊。但电子期刊通过租借模式提供获取而不是购买模式，成为可能也成为普遍趋势。但租借虽然可以降低短期成本，但会增加长期的可获取风险，获取的可持续性取决于支付的可持续性[8]。

当然这种情况并不限于期刊，图书、会议论文同样存在电子化趋势，这些出版物类型也在讨论之列，但由于电子期刊的使用量在近年来大幅增长[9]。期刊的被引分布也在本文专门分析之列。

已有的文献分布经验定律普赖斯定律表明，文献引用存在半衰期，这在电子化时代有了新特征。国内过刊保障研究更倾向于从国家整体层面或者文献资源的建设历史回顾进行，没有具体到微观层面，以单馆作为分析对象，而除了少量的国家统一保障之外，高校图书馆的文献资源建设预算和决算主要是独立进行；对用户而言，图书馆也应当关心本地用户的需求，因而，对单馆用户的需求进行考察和评估是必要的。

本文尝试基于引文分析的角度，研究用户对文献需求的时间分布规律及评估相应的图书馆保障情况。具体来说，需要回答如下研究问题：

1) 用户对不同年代的文献需求量如何，是否有偏好？
2) 用户需求的文献相应的保障率如何？
3) 用户文献需求的时间分布形成受什么因素影响？

为了回答上述问题，本文拟选取一个高校图书馆作为个案研究样本，采用引文分析法进行分析。

由于 J 校是国内规模较大的高校，有 5 万多的学生和全职教职工，发文量过多。本文只选取 J 校的高水平论文作为分析对象。高被引论文引用的参考文献通常比一般论文更多[10]；在情报学领域同样有此特征，高被引论文的引文相对更多，且更新颖（钟镇，2015；姜春林等，2015）。

2 数据收集与结果分析

2.1 数据收集方案

本文选取 2017 年在 ESI 网站分学科下载 J 校 2017 年的高被引论文（highly cited papers）和热被引论文（hot papers），其中前者指在过去十年内引用量在所在领域发文中排在前 1%的论文[11]，后者指在过去 2 年内引用量在所在领域被引量排在前 0.1%的论文[12]。2017 年 12 月日收集数据时， J 校有 21 个 ESI 学科有这两类论文，共计 792 篇高水平论文。

这些论文发表于 2007 年至 2017 年间。其中因统计时 2017 年未截止，故该年份的论文量较少。整体上 J 校高水平论文在统计时的分布随时间呈上升趋势。

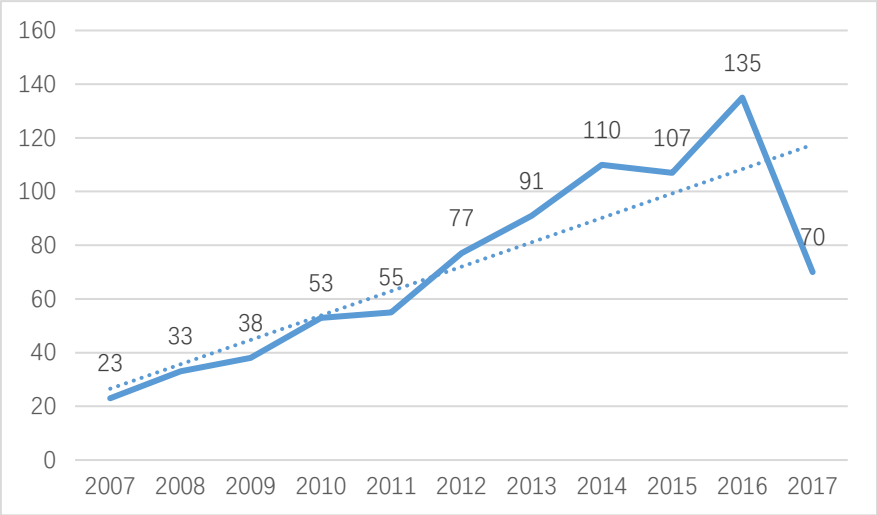

图1　高水平论文发表时间分布

在 Scopus 平台检索原文并导出相应的参考文献信息，共计 5.8 万条参考文献导出，基本情况见表 1。引用最早的参考文献发表于 1505 年，是一篇物理学论文的引用。从所有被引文献的发表时间平均值来看，经济学科为 1999.9，是所有学科中最早的。从被引文献总体来看，最近几年的引用占绝大多数。

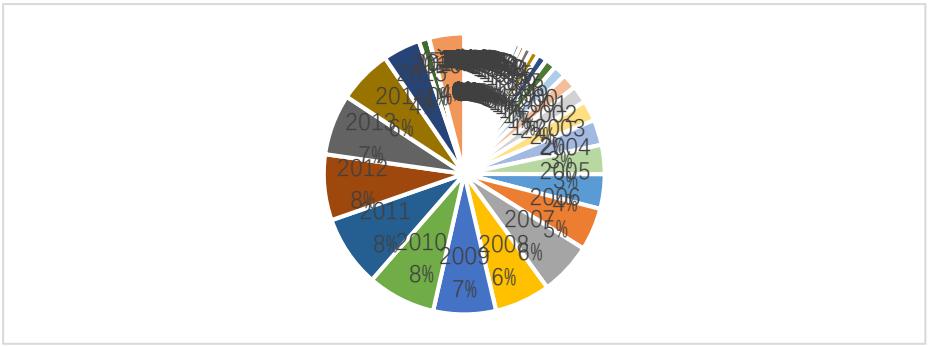

图2　分年代引用占比

这些引文来自期刊、会议论文、书等出版物，792 篇高水平论文共有 4.9 万篇被引文献有所在刊物的 ISSN 号，共来自 4379 种连续出版物。由于 Scopus 的连续出版物信息收录不全，来自非图书类连续出版物的引文量实际可能更多。这些连续出版物中，3 种期刊最早被引用和最晚被引用的载文之间跨度超过 100 年，分别是 *Lancet*，127 年； *Acta Mathematica*，107 年；*Proceedings of the London Mathematical Society*，102 年。

为了做更全面的分析，对引文收集了其他信息来做分析。包括：**1）是否保障**。根据 J 校图书馆的保障清单进行匹配，由于图书馆的保障清单未梳理完全，只梳理了部分期刊，其他未梳理期刊或出版物未进行赋值。**2）引文文献所在学科**，为了与论文所在学科保持一致性，

引文学科根据期刊的 ESI 学科属性表赋值，缺失值未做处理。

其余一些引文属性由 Scopus 导出信息或上述两个属性进行计算得到，不再详述。处理数据的软件为 EXCEL 和 SPSS。

表1　J 校高水平论文及参考文献概览

学科	论文量	总参考文献量	平均参考文献	原文发表年代范围	引文发表年代范围	引文发表时间均值
农学	10	624	62.4	2009~2017	1684~2016	2004.7
生物学与生物化学	34	1791	52.7	2007~2017	1936~2016	2006.1
化学	73	6088	83.4	2007~2017	1894~2017	2008.3
临床医学	163	6925	42.5	2007~2017	1825~2017	2006.8
计算机科学	39	1385	35.5	2008~2017	1948~2017	2006.4
经济学与管理学	7	511	73.0	2009~2016	1911~2015	1999.9
工程	123	8565	69.6	2007~2017	1811~2017	2005.3
环境/生态学	8	610	76.3	2009~2017	1974~2016	2007.6
地球科学	1	26	26.0	2009~2009	1971~2012	2003.7
免疫学	11	639	58.1	2007~2017	1968~2016	2006.9
材料科学	102	6879	67.4	2007~2017	1852~2017	2007.6
数学	21	689	32.8	2007~2017	1823~2016	2001.7
微生物学	8	425	53.1	2010~2015	1953~2015	2006.5
分子生物学与遗传学	33	1845	55.9	2008~2017	1908~2017	2007.1
神经科学与行为学	4	333	83.3	2008~2017	1943~2016	2007.2
药理学和毒理学	13	1078	82.9	2008~2017	1946~2016	2006.9
物理	110	7134	64.9	2007~2017	1505~2017	2004.3
植物与动物科学	14	1071	76.5	2008~2017	1937~2016	2004.3
精神病学/心理学	6	364	60.7	2014~2017	1960~2016	2005.2
一般社会科学	9	413	45.9	2007~2017	1967~2017	2006.6
空间科学	3	410	136.7	2015~2017	1936~2016	2009.6
所有学科	792	53022	66.9	1505~2017	2007~2017	2006.4

注：总参考文献量依据 Scopus 导出，不以实际导出的引文条目数为准。

2.2 被引文献时间分布特征

通过相关分析、独立样本 T 检验等方式检验引文时间分布上的影响因素。相关分析用于分析两个标度变量之间是否存在相关关系，独立样本 T 检验用于检验分类变量是否对标度变量有影响。

相关分析可以使用皮尔逊相关、肯德尔 tau-b 相关或斯皮尔曼相关系数。由于皮尔逊相关要求变量都为连续变量且符合正态分布，斯皮尔曼相关适用于有序变量或者不符合正态分布的等距变量，肯德尔 tau-b 相关对数据分布没有严格要求，可以用于有序变量的相关性检验 [13]。要检验的变量中有有序变量，或者等距变量的分布不清楚，在分析中统一使用了肯德尔 tau-b 相关分析。

其中部分引文信息不全或有误，这些数据在做分析时存在的问题及处理如下：

1）缺省值做分析时被排除。截至统计时间（2017 年底），Scopus 平台收录的参考文献有 14 亿条，遥遥领先于排在其后的竞争对手，但还是收录不全，部分文献的参考文献信息提取不全或无法识别全参考文献信息。本次研究中，只对 Scopus 能提取的正式出版物做分析。受限于 Scopus 收录的引文信息，部分引文无法识别或识别不全，因不清楚分布方式，无法通过推测增补相应信息，在做后续分析时相应信息不全的条目选择了直接排除个案的方式。

2）出错值做分析时被排除。论文发表时间只提取年份，被引文献的发表时间也只提取年份，但出现了少量前者晚于后者的情况。经检查，是如下原因导致：Scopus 引文识别提取出错，或者由于论文发表在不同平台上，Scopus 在计算时没有选取文献最先发表的文献，或者论文在正式发表之前就被引用了。共有 42 条引文出现了这种情况，这些数据被视为异常值，做与引文发表时间有关的分析时被排除在外。

2.2.1 影响分布的分类变量

只有两个取值的分类变量用独立样本 T 检验检验其对时间分布的影响；多个取值的采取单因素方差分析。

1）是否自引。分类变量中期刊自引与被引对引用间隔产生显著影响。属于期刊自引的引用，被引文献与原文文献时间跨度更长。在独立样本 T 检验中，二者在 0.01 的水平上存在显著差异。

2）是否保障。被保障期刊的最早被引年份比未保障期刊的最早被引年份更早。在独立样本 T 检验中，二者在 0.05 的水平上存在显著差异。

表2　　分类变量独立样本 T 检验

检验变量	分类变量	显著性（双尾）（上，假定方差均等；下，不假定方差均等）	引文保障	个案数	平均值	标准差
文献被引时间间隔	是否期刊自引	0.00	是	11920	39.52	247.096

		0.00	否	453	5.19	5.653
	是否同学科引用	0.008	是	14725	6.84	7.563
		0.007	否	13216	6.6	7.136
期刊被引平均时间间隔	期刊是否被保障	0.001	是	4123	4.816	8.207
		0.00	否	256	3.015	4.558

3）是否跨学科引用。被其他学科引用个数越多，平均引用时间距离越小。被其他几个学科引用与平均引用时间距离之间的皮尔逊相关系数为-0.206（这里似乎应该改用方差分析之类的），在 0.01 的水平上显著。

其中，再对期刊被几个其他学科引用及引用量与期刊被引平均时间距离做肯德尔 tau-b 相关分析，可以发现被其他多个学科引用的期刊，平均间隔较短。在总体的被其他学科引用量，被其他学科引用越多，平均距离越短，肯德尔 tau-b 系数为-0.436。而期刊被引跨度与被跨学科引用量和被几个其他学科引用正相关，可能是因为一本期刊的学术价值同时决定的，即更有学术价值的期刊被跨学科引用更多，在时间跨度上也更长。

表3　跨学科引用与引用间隔的相关分析

样本特征	被其他学科引用量	被几个其他学科引用
期刊被引平均时间距离	-.436**	-.340**
期刊被引跨度	.344**	.478**

注：1）期刊被引跨度指在样本中最晚被引用时间减去最早被引时间；2）被几个其他学科引用指期刊除了被本学科期刊引用外，还被其他几个学科引用；“**”表示在 0.01 的水平上显著，其中显著性均为双尾检验。

4）论文所在学科

对所在学科做单因素方差分析，除了地球科学因样本量较小，预测区间范围较大，难以检测出与其他学科的差异之外，其他各学科中，存在一定差性。生物学、医学及其较为临近的免疫学、临床科学等学科的引用时间距离较短。而经济学与管理学（后称经管）、精神病学/心理学、数学、物理等较长。其中，在事后检验中，经管与其他各学科均存在显著差异。

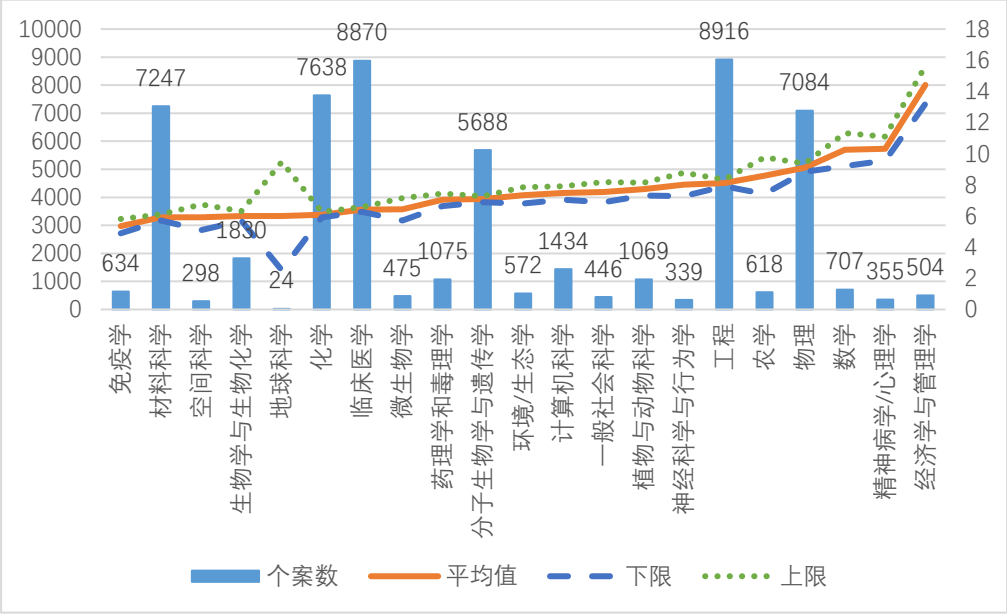

图3　论文所在学科平均引用时间距离

5）被引文献的文献类型。文献类型由 Scopus 定义，未做人工干预，Scopus 未赋值的作为缺失处理。主要取值有学术论文（article）、会议论文、短调查（short survey）等，书包括 Scopus 定义的书（book）和书的章节（book chapter）。其他类型的文献有勘误、预印本。

通过单因素方差分析检验得到文献类型显著影响引用时间间隔。其中其他这一类型因样本量较小，与其他类型无显著性差异；通信在样本中引用间隔较长，与其他类型均差异显著。除此之外，学术论文、会议论文，可以认为性质类似，它们互相之间无显著差异，而与其他类型有显著差异，引用间隔较长；综述与社论、注释一组较为类似，它们之间无显著差异；短调查居于这二组之间，时而与其他组有差异，时而没有。

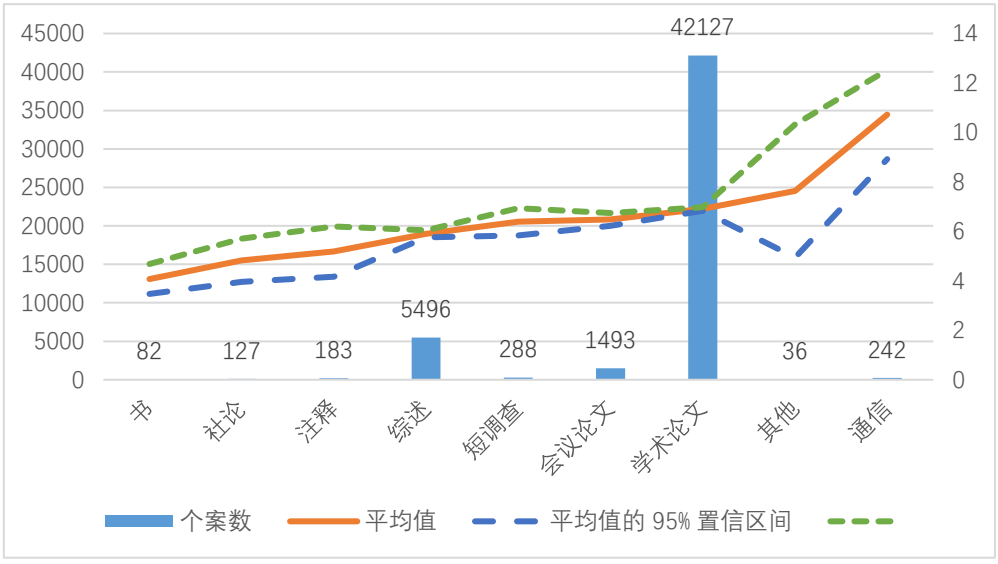

图4　文献类型平均引用时间距离

表4　单因素方差分析事后比较（假定方差不等）

塔姆黑尼	学术论文	综述	会议论文	社论	通信	短调查	注释	其他
学术论文		-1.007*	-0.426	-2.073*	3.816*	-0.522	-1.722*	0.732
综述	1.007*		.581*	-1.065	4.823*	0.485	-0.714	1.739
会议论文	0.426	-.581*		-1.646*	4.242*	-0.095	-1.295	1.158
社论	2.073*	1.065	1.646*		5.888*	1.551	0.351	2.804
通信	-3.816*	-4.823*	-4.242*	-5.888*		-4.338*	-5.537*	-3.084
短调查	0.522	-0.485	0.095	-1.551	4.338*		-1.2	1.253
注释	1.722*	0.714	1.295	-0.351	5.537*	1.2		2.453
其他	-0.732	-1.739	-1.158	-2.804	3.084	-1.253	-2.453	

注：“*”表示在 5%的水平上具有显著性。

2.2.2 时间分布与期刊指标的关系

主要通过相关分析，检验期刊指标是否能预测期刊的历史文献需求关系。其中期刊指标根据 JCR 导出的指标进行了相关分析检验。

期刊平均被引时间间距与绝大多数期刊指数都显著相关，但相关性不强，即相关系数的绝对值不高。换句话说，基于期刊本身的取值特征难以预测其时间分布。

表5　肯德尔 tau-b 相关分析（检验变量：期刊平均被引时间距离与期刊指数）

期刊指数	总被引量	半衰期（<=10.0）	半衰期（分等级）	影响因子	他引影响因子	5 年影响因子
相关系数	-.315**	.099**	-.033**	-.274**	-.266**	-.260**
相关系数	载文量	特征因子	论文影响力	论文占载文量比（%）	影响因子百分位	标准特征因子
显著性	-.280**	-.357**	-.178**	.036**	-.211**	-.357**

注：1）“**”表示在 0.01 的水平上显著，其中显著性均为双尾检验；2）半衰期均指期刊的被引半衰期；3）样本特征变量个案数都为 4379，期刊指数变量来自 JCR，缺少部分期刊的，除了 5 年影响因子、被速率、论文占载文比、半衰期（<=10.0）有不到 50 个个案的缺失外，其他都为 3471；4）JCR 导出的期刊属性中，超过 10 的半衰期只注明“>10”，没有注明具体值，采用了两种处理方式，一种放弃这部分样本，即表格中的半衰期（不包括大于 10.0 期刊）；另一种是分为 4 个等级，分别是小于等于 3.3，3.3~6.6（包含），6.6~10（包含），大于 10，分别赋值为 1、2、3、4，即表格中的半衰期（分等级）。

3 讨论与结论

根据前述的分析，文献引用的时间分布特征主要可以归纳为受如下因素影响：所在学科、内容跨领域、跨学科性，文献的可获取性，文献类型特征。

1）学科属性。经管类学科的引用间距较长，发展历史较长的学科如物理、数学、植物

学和生物学的引用间距也较长。

2）跨领域、跨学科的引用，引用时间距离较短。期刊他引越多，引用时间距离越小；期刊跨学科引用越多，引用时间距离越小。这可以推论，一篇文献如果在某一领域内，可能会追溯较久远的文献；反之，跨领域引用可能不会追溯太远或者跨领域的论文的半衰期较短。当然仅凭文献所在期刊的 ESI 学科属性或者发表在同一期刊上来推定文献的内容专业性并不精确，还需要对施引文献与被引文献做内容分析才能得到更精确的结论。

跨领域、跨学科引用可能本质上也是因为所在领域或学科较为新颖，尚未形成较为独立的学科知识体系，知识更新迭代较快。

3）引用时间距离受文献类型影响。更学术的论文平均引用时间距离更长，即半衰期更长，非正式学术载文的半衰期不确定。在学术论文、会议论文与短调查三种比较规范的科研论文中，学术论文体现了更久远的生命力，其次是会议论文，再次是短调查。

综述和社论的生命力较短。而书的平均引用时间距离较短，可能是因为书本身是对以往知识性成果的总结，在新颖性上本身就较为欠缺。而高水论文通常代表前沿领域（需要参考文献），从内容属性上说与书的距离就较远。而少量通信和其他类型的文献的时间距离较远，一方面，这些类型的文献样本量较少，95%置信水平上的上下限区间较大，不能完全肯定该结论是否可以推广；另一方面，人们引用书是为了查找期刊中没有的资料（Tripathi ，Jeevan，2013），引用通信等类型的文献可能也是为了补充某段零星资料，具有不确定性。

各类期刊指数对期刊被引用时的时间分布预测性并不强。这可能是因为一方面期刊指数表现良好的期刊，既是前沿的，当前的期刊能得到大量引用；过去的期刊也可能因较高的学术价值被引用更多。而表现不良好的期刊可能也是只有当下才有机会被引用。这导致从宏观数据上预测二者之间的差别没有可能。图书馆可以考虑从其他特征来做馆藏建设决策。

此外，本次数据的分析样本存在一定不足。一方面，引文分析只能考察用户在学术研究中需要的文献，而用户的教参需求、休闲需求并未被考虑进来。已有研究表明，人文社科类用户需要大量阅读图书。对浙江大学的学生用户调查也表明，本科生阅读休闲类电子书更多[16]。这种阅读需求难以直接通过引文分析调研出来。另一方面，J 校为理工科较强的综合类院校，分析样本主要是理工科文献。

另一方面，少量取值数量较少，不能做出预测。如高水平论文发文较少的学科，如空间科学。做出需要的回溯馆藏较难。需要增加样本量，或借鉴其他学校或该学科总体的引用距离，当然， 更直接的办法是走访相应专业的研究者。

4 馆藏建设建议

国外一些图书馆在馆藏建设时采取了根据用户意见的方法来确定预算优先级，再分配用于采购的资金额度[17]。根据前述讨论，可以根据引用行为特征并、用户特征建设和文献特征三者综合考虑建设馆藏，以期为用户提供更精准的保障服务。

1）高校图书馆建设馆藏时考虑学校的学科发展特征。在不同学科对回溯文献的需求不

同、是否跨学科引用影响过刊需求的情况下，高校图书馆可以根据本校学科发展特征进行决策，选择所需要的期刊。

J 校高水平论文存在大量跨学科引用。因而跨学科引用也可以作为馆藏建设的影响因素之一。跨学科需求的文献距离较短。跨学科引用较多的学科可以考虑较少的回溯馆藏建设预算。

预测未来用户对馆藏的时间分布规律，并用于馆藏建设决策，与领域发展速度、领域的知识更新迭代速度有关。

知识更新快的学科可以在建设回溯馆藏时设置较少的预算；更新慢的学科需要设置较多的预算。在研究样本中，免疫学、临床医学等学科即可以设置较少预算的回溯馆藏，而经管、物理、数学等自成体系更久的学科需要更多回溯馆藏保障。

2）结合文献类型考虑回溯库建设。建设回溯馆藏可适当在文科类学科上倾斜。J 校为理工科的学科，而文科对回溯馆藏的需求更为强烈。这让 J 校图书馆可以花费较少的预算在回溯馆藏建设上。

最后，建设回溯馆藏从用户需求的角度考虑，是要在用户对研究的前沿和历史之间做出平衡选择。一个基于引文分析的预测只能给出较为形而上学的预测，做更精确的预测需要对引文做内容分析或者对用户做调研，以期更深入了解他们的需求。

参考文献

[1] Corrall S, Kennan M A, Afzal W. Bibliometrics and research data management services: Emerging trends in library support for research[J]. Library trends, 2013, 61(3): 636-674.

[2] 郑建程,袁海波.NSTL 外文科技期刊回溯数据库的国家保障策略[J].图书情报工作,2010,54(13):10-13.

[3] Walters W H. E-books in academic libraries: Challenges for acquisition and collection management[J]. portal: Libraries and the Academy, 2013, 13(2): 187-211.

[4] 肖珑,张洪元,钟建法,武桂云,李浩凌,李峰.建国后高校文科外文文献的发展状况与未来保障研究[J].大学图书馆学报,2013,31(02):5-13+92.

[5] 李峰,涂文波.基于重点学科引文分析的我国高校人文社会科学外文文献保障率研究[J].图书情报工作,2013,57(02):64-69.

[6] Lamothe A R. Comparing usage between a Dynamic and a Static e-monograph Collection[J]. Collection Building, 2015, 34(1): 17-26.

[7] Bullis D, Smith L. Looking back, moving forward in the digital age: A review of the collection management and development literature, 2004–8[J]. 2011.

[8] Walters W H. Criteria for replacing print journals with online journal resources: The importance of sustainable access[J]. 2004.

[9] Tripathi M, Jeevan V K J. A selective review of research on e-resource usage in academic libraries[J]. Library review, 2013, 62(3): 134-156.

[10] 梁春慧,孙艳,万跃华.高被引论文的参考文献特征研究——以化学领域为例的实证分析[J].科技与出版,2014(07):119-122.

[11] 参考网页: http://ipscience-help.thomsonreuters.com/incitesLiveESI/ESIGroup/indicatorsGroup/citationThresholds/thresholdHighlyCited.html

[12] 参考网页: http://ipscience-help.thomsonreuters.com/incitesLiveESI/ESIGroup/indicatorsGroup/citationThresholds/thresholdHot.html

[13] 丁国盛, 李涛. SPSS 统计教程:从研究设计到数据分析[M]. 机械工业出版社, 2006.252-253.

[14] Jamali H R, Nicholas D, Rowlands I. Scholarly e-books: the views of 16,000 academics: Results from the JISC National E-Book Observatory[C]//Aslib proceedings. Emerald Group Publishing Limited, 2009, 61(1): 33-47.

[15] Staiger J. How e-books are used: A literature review of the e-book studies conducted from 2006 to 2011[J]. Reference & User Services Quarterly, 2012, 51(4): 355.

[16] Wang S, Bai X. University students awareness, usage and attitude towards e-books: Experience from China[J]. The Journal of Academic Librarianship, 2016, 42(3): 247-258.

[17] Oseghale O. Faculty opinion as collection evaluation method: A case study of Redeemer's University library[J]. Library Philosophy and Practice (e-journal), 2008: 221.