项目结题验收单

专家验收表（主持人所在单位组织 3-5 名专家对项目进行验收、自评。）

项目名称	基于借阅时长和H指数分析借阅数据的实证研究		
主持人	姚震	职务/职称	馆员
所在单位	（加盖单位公章）		
专家意见	该项目研究深入统计分析了我馆流通数据，针对流通时长结合中图法一、二级类目做了多方面的统计分析，利用统计特征设计的借阅量化比之前的研究在合理性、可用性等方面有所提高。结合H指数的概念，计算出各大类的核心图书，对比量化前后的数据差别，总结出识别读者需求的方法。该研究方法创新方法，在流通数据分析利用方面做出了较好的应用效果，如图书下架、制定采购书单计划等工作中都具有良好的数据说服力和指导性。 鉴于H指数具有时间窗口等方面的限制，可继续坚持做更深层的数据分析工作。 该研究中的数据清洗、数据统计和分析的算法均通过编程实现，具有较强的实用性。数据与算法具有可重复性，有较强的科学性保证。 综合考虑，建议予以结题。 （如需要可增加页数）		
专家签字			
职务/职称			

项目编号：2021040
注:项目编号请查看立项通知，也可缺省

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：基于借阅时长和 H 指数分析借阅数据的实证研究

项目关键词：借阅时长、H指数、数据分析

项目单位（盖章）：山东农业大学图书馆

通信地址：山东省泰安市岱宗大街 61 号(271018)

项目主持人：姚震

联系电话：15552555715

电子邮件：yz@sdau.edu.cn

提交日期：2022 年 4 月 14 日

基于借阅时长和H指数分析借阅数据的实证研究

关键词：借阅时长、H指数、数据分析

1 研究背景、目的及意义

高校图书馆分析图书流通数据是保障图书馆的核心业务有效开展、提升服务水平的基础工作之一。流通数据的分析结果可用于图书推荐、调整采购规划、优化典藏分布策略等方面的工作中。近年来学术界对流通数据的分析成果丰硕，主要方法一般基于文献计量学，或结合数据挖掘、大数据的算法模型做交叉融合研究。如熊拥军等研究者总结了图书馆统计分析技术要点、资源利用大数据组织方式、设计了大数据技术框架并利用Elastic Search和Java技术实现了数据分析功能[1]。对借阅行为的分析同样丰富，如易明等利用人类动力学分析了借阅数据，统计了不同学历读者的借阅时间间隔特征，得到从群体到个体层面的借阅行为规律[2]。胡乃菲等则基于借阅数据做了大学生阅读需求的分析，得出大部分学生的阅读动机以专业图书满足应试需求为主、以文学小说满足消遣为辅的结论；统计了读者的平均借阅时间，得出专业图书和考试类材料的借阅时间较长的结论；并相应提出了图书馆服务策略[3]。各项研究均提出了相应的服务策略[4-6]，推动了图书馆的服务提升和转型。

2 研究内容及方法（思路、方法、具体内容）

本次研究的原始数据集来源是山东农业大学图书馆提供的图书馆业务数据集中的流通数据，包含2014级至2020级40516名学生（含本科和研究生）的853476条流通记录。为了突破图书管理系统以借阅次数为统计源的局限，实现对数据进行多种方式的组合试验等方面的考虑，在数据预处理阶段便依据借出时间和还回时间计算了借阅图书的持有时间，并清洗了索书号中分类字段外的字符（索书号生成规则是分类号加著者号），再按照中图法第五版的一级类目和二级类目截取相应位的字符串，作为后续组合统计的参考字段。

设持有时间的单位为天，借出未还的流通记录（还书时间字段为空）共4911

条，这些数据可归属为随机缺失类型，删除或设为 0 并不会导致有偏估计，因此可规定借阅时长定义为 0；另外因持有时间小于 14.4 分钟的因在计算过程中只保留两位小数而导致计算结果为 0 的借阅记录有 10919 条，此两种流通记录共 15830 条。考虑到借阅过程中产生的超期、临时延期等因素会造成借阅时间过长，且假期时间会导致统计有偏于无假期的借阅记录，因此在计算图书持有时间完成后，设置新字段存储减去临时延期（如疫情严重时期封校闭馆时段、默认还书时间处于寒暑假期中等情况）和超期后的图书持有时间，使借阅图书的持有时间都近似于在校期间的情况。经此步骤处理后统计新的图书持有时长字段，得到超过 120 天的记录有 31331 条（占总数的 3.67%），超 150 天的有 9267 条记录，按照概率论中小概率事件的定义——大量重复试验中出现的频率在 1%或 5%以下的事件，可将借阅时间超过 120 天的借阅视为小概率事件（超长借期的发生则以学生忘记归还为主要原因，少数则因图书丢失），因此统计时可剔除持有时间超过 120 天的记录。统计(0,120]区间内有续借操作的图书持有时间的平均值为 81.47，标准差为 25.52；无续借的图书持有时间的平均数是 32.7，标准差是 24.56，进一步按中图法分类后可得各大类图书借阅持有时间的统计特征值，详见表 1。

表 1. 图书借阅持有时间的统计特征值（单位：天）

分类号	平均数 v	中位数 m	标准差 sd	分类号	平均数 v	中位数 m	标准差 sd
A	29.80	26.96	22.80	N	31.35	27.29	24.43
B	33.16	30.04	24.91	O	35.32	33.5	25.15
C	34.01	30.96	25.72	P	36.43	33.71	26.48
D	29.89	25.67	24.94	Q	35.86	33.88	27.07
E	29.28	24.96	24.35	R	31.22	28.08	24.91
F	31.47	27.79	25.10	S	32.57	29	25.75
G	31.83	28.12	25.40	T	36.51	35.04	25.18
H	36.63	35.79	24.23	U	31.14	27.71	24.53
I	28.04	22.83	23.51	V	29.05	25.04	23.72
J	33.56	31.04	24.87	W	30.81	25.67	25.95
K	30.10	26.12	23.88	Z	28.05	23	23.47

为了更直观的查看图书持有时间的分布特点，以 0 为起始值、1 为区间单位，统计图书持有时间为 0、以及落在区间(0,1]、(1,2]……(119,120]的各大类借阅记录的条数，汇总后得图 1。可见各大类图书都有较多的短期借阅记录，[0,30]区间内流通量大的几类图书的散点分布基本相似，且是有较大波动的曲线；流通量小的各类图书的散点分布曲线则比较平缓；(30,80]区间内分布点组成的曲线基本相似，波动

点和幅度基本相同；接近 120 区间时基本重合，长期借阅的发生概率基本一致；其他借阅量小的分类，则分布基本平缓且相似，这几方面都意味着读者有相似的借阅行为习惯。短期借阅行为数量庞大，其内在原因可能包括图书种数多、内容参差不齐、读者喜好差异性大等方面。各大类图书在接近一个借期 60 天（设置各类图书借期相同）的区间段有更高的记录条数，说明读者对借阅有效期的关注度是比较高的。

读者借阅文献是对文献价值的肯定，经前文所述的数据探索后发现持有时间是可量化的信息。图书持有时间的影响是多方面的，例如文献长度、个人阅读速度、阅读时间安排等，这些都是随机性、差异性强的影响因素。总结流通工作中的经验，可归纳概括出超短期借阅行为代表着读者判断该书无法满足求知所求，或不符合阅读习惯、个人偏好等，与此对应的是超长期的借阅，也代表读者所借图书没有占据读者关注，未被充分利用，因此这两种借阅行为的量化值应较小。而高效利用的图书，读者会续借、临近借期满时归还，甚至归还后重新借阅。由此归纳设计了新的量化公式：设图书借阅量化值为$p\ (0 \le p \le 1)$，设t为该借阅行为的图书持有时间，t'为该借阅行为的超期时间，令有续借的借阅行为的量化值 $\mathrm{p}=\begin{cases}1 & ,\ t'=0\\ \frac{t-t'}{t} & ,\ t'<t\\ 0 & ,t'\ge t\end{cases}$，令没有续借的借阅行的量化值$p=\begin{cases}\frac{t-t'}{m_i} & ,\ t<m_i\\ 1 & ,\ m_i\le t<v_i+\frac{v_i+sd_i}{x}\\ \frac{120-(t-t')}{120-(v_i+sd_i)} & ,v_i+sd_i\le t<120\\ 0 & ,t\ge 120\end{cases}$，其中 v_i、m_i、sd_i分别是各一级类目的平均数、中位数和标准差。定参 120 是借期（60 天）和续借期（续借时间点后延 60 天）的最大值之和，也是可以区别小概率事件的阈值。该量化方程是先增后减且连续的。

借阅量化的值平均数为 0.58，中位数为 0.89，标准差是 0.36。按量化结果结合分类进行区间计次统计得表 2。可见 68982 条记录（占总借阅量的 7.5%）被量化为 0，其中 65066 条是超短时间借阅，3916 条是超长时间借阅。

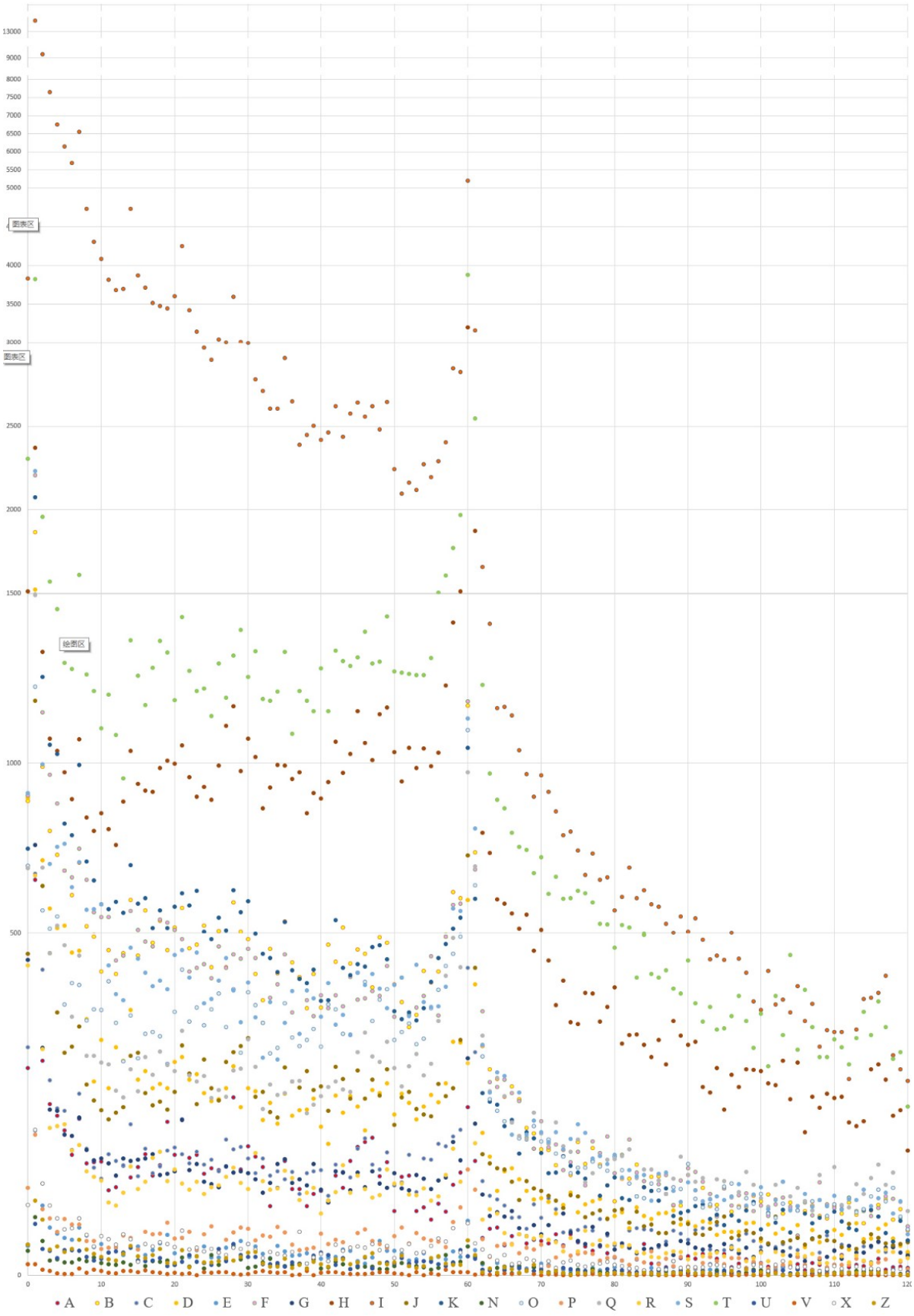

图 1. 分类区间计数散点图

表 2. 量化值分类区间计次表

	0	(0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]	总和
A 类	594	889	592	579	435	440	576	665	740	1047	5944	12501
B 类	1733	3187	2084	1907	1626	1939	1918	2046	2195	2538	22854	44027
C 类	672	1261	717	645	581	706	642	746	800	942	8644	16356
D 类	1524	2014	1518	1134	1006	978	1048	1077	1151	1518	13887	26855
E 类	135	265	178	155	142	107	144	124	174	200	1451	3075
F 类	1804	3463	2209	2151	1774	1694	1654	2064	2051	2423	22938	44225
G 类	747	1092	627	647	591	592	576	540	764	765	7525	14466
H 类	2783	4581	3894	3052	3712	4054	4118	4497	4832	5474	50936	91933
I 类	7501	22184	15232	15630	11939	10094	10654	13758	12677	21633	123363	264665
J 类	984	2063	1096	972	847	1145	961	1249	1318	1547	13147	25329
K 类	1592	3232	2706	2353	1747	1912	1890	2075	2344	2709	22726	45286
N 类	75	116	69	66	63	50	62	77	74	103	832	1587
O 类	1354	2111	1435	1613	1350	1412	1581	1624	1897	2105	19586	36068
P 类	228	352	275	211	164	253	261	249	277	341	2998	5609
Q 类	1527	2430	1582	1326	1084	1307	1336	1435	1419	1733	18992	34171
R 类	778	933	588	546	447	425	496	612	584	653	6600	12662
S 类	2034	3184	2106	2000	1654	1567	1618	1795	2116	2264	22417	42755
T 类	4454	6808	5415	4420	4947	4916	5976	5392	6650	6874	67754	123606
U 类	77	121	89	98	90	104	81	79	114	87	1017	1957
V 类	24	19	10	18	18	6	17	17	13	19	168	329
X 类	163	343	248	192	166	177	117	176	167	237	2081	4067
Z 类	84	178	91	111	81	62	83	91	106	162	898	1947
总和	30867	60826	42761	39826	34464	33940	35809	40388	42463	55374	436758	853476

将原 H 指数概念中的引用次数改为借阅次数则可得图书借阅的 H 指数概念：当且仅当某类（或某出版社或著者）有 H 种图书的借阅次数都不小于 H 时，该类图书的成就分值就是 H[9-10]。基于此概念可计算各一级类目的 H 指数（汇总数据如图 2 所示），并遴选出各类的核心图书。文学（I）类的 H 指数显著高于其他各类，H 指数排前 5 的其它一级类目是 H、K、T、Q 类。可见量化计算对文学类图书的 H 指数影响最大，指数值降幅达 34.5%。

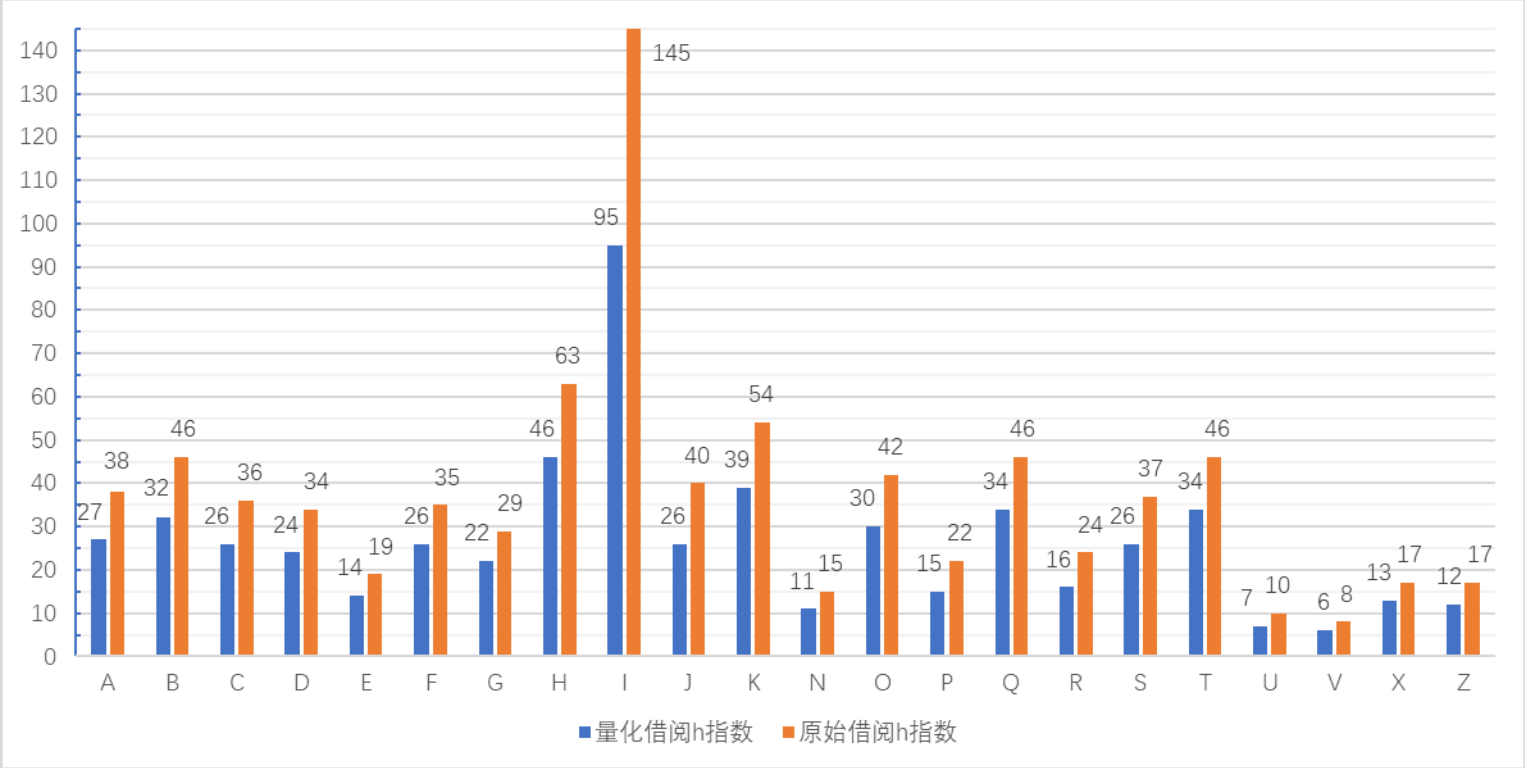

图 2. 一级类目的 H 指数对比柱状图

3 结论与建议

选取借阅量大的 F、H、I、K、S、T 大类中借阅量化值累计大于该类 H 指数的核心图书，做进一步分析，将此 6 大类的核心图书以量化降幅比（借阅次数减量化值再除以借阅次数）进行排序，截取各大类前 3 与后 3 的试验结果汇总在表 3 中。可见文学类图书中当代小说（I247）占文学类总借阅量的 25.56%，日本文学（I313）和英国文学（I561）占比排在其次；而从量化对借阅次数的影响方面来看，下降幅度最大的是日本文学、当代小说、现代散文（I266），下降幅度最小的则是哥伦比亚文学（I775），该类图书经详细查看图书信息后发现均是马尔克斯的小说，其次是古代至近代小说（I242）。此数据变化是符合图书特征和读者阅读习惯的，I242 和 I775 类涉及的图书均是长篇且是名著，所以内容质量高、所需阅读时间也较长，而其他小说或散文则因篇幅小、阅读喜好等不同原因则会存在更多的低量化值情况。语言（H）类的核心图书均是英、法、日语的词汇和考试相关图书，法语类图书受量化影响最大，分析其中可能性较大的原因是我校没有法语专业，图书需求集中在外语专业的第二外语选修和读者自学外语的情况中，此类需求与英语等级或考

研等考试的需求形成明显对比，同样情况也在 TP311 等类的图书中。工业技术（T）类的核心图书中，量化影响最大的 TS976 是“家庭生活”方面的图书，内容无法满足读者真正需求而导致超短期借阅量较大从而降低了其量化结果；其他专业图书的借阅则基本与我校专业设置相对应，且量化影响到的比例也相对不高。农学（S）类的核心图书中农学史（S-09）受量化影响最大，而作物遗传育种（S330）、观赏园艺（S68）、耕作学与有机农业（S34）等专业书受较小影响，说明科普或基础知识等图书并不能满足专业需求，读者在挑选图书时由于分类法等专业知识的缺少，导致对图书内容的判断存在误差，在阅读后再判断满足不了需求，进而产生超短期借阅。其他未说明的个大类图书也可推论出相同的分析结果。各类专业图书的受量化的影响较小，说明读者对专业图书的需求较为迫切，且可推测读者在选择专业图书时也面临困扰，难以精准选择可满足需求的图书。

表 3. 核心图书分类号借阅量化与借阅次数对比表

分类号	量化值总和	借次总和	下降比例	分类名称
I313	2288.67	4264	46.33%	日本文学
I247	3654.52	6712	45.55%	当代小说
I266	96.67	170	43.14%	当代散文
I217	97.7	151	35.30%	民国作品集
I242	218.02	330	33.93%	古代至近代小说
I775	303.63	456	33.41%	哥伦比亚文学
H32	57.79	112	48.40%	法语
H319	3067.66	4907	37.48%	英语教学
H315	47.45	73	35.00%	英语写作、修辞
H314	135.65	200	32.18%	英语语法
H36	144.5	209	30.86%	日语
H313	605.09	857	29.39%	英语词汇
TS976	136.44	237	42.43%	生活知识、家政服务
TV13	50.31	79	36.32%	水力学
TP312	340.8	501	35.05%	计算机编程语言
TU986	462.49	679	31.89%	园林规划与建设
TU984	58.47	85	31.21%	城市规划
TP311	35.8	52	31.15%	程序设计、软件工程
TB126	45.87	65	29.43%	工程流体力学
TP391	115.23	155	25.66%	信息处理
TH-39	74.39	99	24.86%	机电一体化
TH164	35.81	43	16.72%	计算机辅助机械制造
S-09	28.41	68	58.22%	农业史
S330	46.85	67	30.07%	作物遗传育种

S68	65.7	90	27.00%	观赏园艺
S34	40.01	54	25.91%	耕作学与有机农业
F821	30.14	59	48.92%	世界货币
F069	30.51	56	45.52%	经济学其他分支学科
F49	36.13	58	37.71%	信息产业经济
F0	599.39	866	30.79%	经济学
F224	32.27	45	28.29%	经济数学方法
F830	30.19	41	26.37%	金融、银行理论
K248	1301.47	2311	43.68%	明史
K835	94.5	161	41.30%	南亚人物传记
K827	226.41	348	34.94%	社会政治人物传
K204	178.45	275	35.11%	古代史籍
K10	193.27	284	31.95%	世界通史
K02	42.25	60	29.58%	社会发展理论

因此，图书馆可在以下三个方面提升服务水平：（1）应在专业图书文献资源建设方面更加细心，遴选采购图书不能只关注一级类号，而应认真考察所属分类，加强图书的专业性；（2）针对复本数量变化难以及时满足如公共文化热点等因素导致的读者对某一种资源的需求激增的现象，图书馆应设计更灵活、响应更快的复本采购工作方案；（3）在阅读推广活动中要更清晰明了的讲解图书馆藏书体系，并根据活动对象的专业特征做通识共需类和专业类图书的精准推广；（4）图书下架等工作更多注意超短期借阅图书，总结图书特点。

4、结论

基于借阅时长统计特征的量化方法将借阅时长、续借、超期等信息所能体现的图书流通的意义做了数值化，使原本模糊的统计意义变得清晰，所得数据更有参考价值。现常见的图书管理系统一般只对典藏图书做“零借阅”统计，作为图书下架等工作的参考指导，而本文提出的方法对有借阅记录的图书做利用情况分析，找出单原始借阅数据中隐藏的更为精确的读者需求。利用 H 指数的概念对流通数据进行分析可遴选出核心图书，使对读者需求的分析更为精细。借助该方法找到更符合读者需求的结果，对图书馆的馆藏资源建设优化、布局方案调整、策划更有针对性的图书推荐、根据不同专业调整读者培训中各专业图书相关知识内容等工作具有实际指导意义，有利于提升图书馆的服务质量和效果。在流通数据的预处理过程中应用该量化方法，然后结合数据挖掘或大数据分析模型做聚类、预测等试验，对比结果异同、求证该量化方法是否提升原始数据产出的准确性以及确定影响程度的大小，可作为下一步研究继续开展。

4 项目成果（发表的文章、开发的软件、取得的实践效果等）

设计编写了针对汇文导出的流通数据的数据清洗程序、量化公式的计算程序，该成果在第三届“慧源共享”全国高校开放数据创新研究大赛主赛道山东赛区获得“三等奖”。研究过程已总结为论文，目前尚在外审阶段。另外，研究结果在图书选购、图书推荐和图书下架等工作中作为参考数据，提高了工作效率和质量。

5 参考文献

[1]周志峰.h 指数应用于图书馆借阅数据分析的探索[J].图书馆建设,2009(11):82-84+89.
[2]净玲娣,王立宏,王芹,李雅.图书流通数据分析的 H 类指数遴选及应用——以西北农林科技大学图书馆为例[J].图书情报工作,2020,(09):65-72.
[3]赵雨薇.基于数据挖掘感知读者需求的高校图书馆差异化服务研究[J].图书馆工作与研究,2018(07):68-73.
[4]易明,张展豪,李怡.大学生图书借阅行为的人类动力学分析[J].图书馆学研究,2019(22):83-93.
[5]左平熙.大数据时代高校图书馆智慧服务的逻辑与路径[J].图书馆工作与研究,2021(05):48-54.
[6]江山.智慧图书馆要素研究及建设思考[J].图书馆工作与研究,2022(02):58-63.
[7]王红,袁小舒,原小玲,黄建国. 高校图书馆读者借阅趋势线性回归建模预测探析[J]. 图书情报工作,2020,64(3):59-70.
[8]初景利,赵艳.图书馆从资源能力到服务能力的转型变革[J].图书情报工作,2019,63(01):11-17.
[9]李洋,温亮明.我国高校图书馆科学数据开发现状调研与分析——以一流大学建设高校图书馆为例[J].图书馆工作与研究,2021(12):5-15.
[10]景民昌,于迎辉.基于借阅时间评分的协同图书推荐模型与应用[J].图书情报工作,2012,56(03):117-120.
[11]山东农业大学图书馆，″山东农业大学图书馆业务数据集″,http://hdl.handle.net/20.500.12291/10726 V1 [Version]