

# 基于 Usage Metrics 的中文学术论文用户平台偏好和兴趣偏好比较研究<sup>\*</sup>

陈必坤 周慧娴 钟周燕 王曰芬

**摘 要** 研究中文学术论文用户使用模式有助于分析使用规律,为应用决策提供依据。本文以八个学科被 CSSCI 或 CSCD 收录的 61 本开放获取期刊发表于 2014—2015 年的学术论文为样本,以期刊官网和信息集成平台上的使用数据为来源,采用 Usage Metrics 的方法,从用户平台偏好和用户兴趣偏好两方面比较中文学术论文的用户使用模式。研究发现:①期刊官网与信息集成平台的用户平台偏好存在差异。一是从学科视角看,用户整体上倾向于使用期刊官网而不是信息集成平台获取所需论文,而社会科学的信息集成平台篇均下载次数均大于自然科学。二是从期刊视角看,社会科学用户比自然科学用户更倾向于使用信息集成平台获取所需论文;与信息集成平台相比,期刊官网的用户下载数据呈现更为明显的“两级分化”现象;期刊官网下载次数与信息集成平台下载次数整体上的相关性较低。②期刊官网与信息集成平台的用户兴趣偏好存在差异。每本期刊官网和信息集成平台下载次数前 20% 的学术论文的 Jaccard 相似系数较低;在所列举的图书馆、情报与文献学中,期刊官网与信息集成平台的用户关注主题不同。图 7。表 5。参考文献 37。

**关键词** 学术文献 使用数据 Usage Metrics 学术期刊 学术交流 使用模式

**分类号** G250.2

## Exploring the User Platform Preference and User Interest Preference of Chinese Scholarly Articles: A Comparison Based on Usage Metrics

CHEN Bikun, ZHOU Huixian, ZHONG Zhouyan & WANG Yuefen

### ABSTRACT

With the rapid progress of networking and digitization, interactive data of users on various platforms have been recorded in real time. It has promoted the rise and development of usage metrics that provided new ideas and methods for users' academic communication behavior research and decision-making applications. Also, various Chinese academic literature databases have been increasingly utilized by users who are in need for high-quality information service, which can be promoted by usage metrics on user usage patterns.

In order to comparatively investigate the usage patterns of journal official sites and pay-for-access platform and then to support decision-making, this study selected academic papers, published in 2014–2015 of 61 Chinese open access journals indexed by CSSCI and CSCD in eight disciplines, “Library, Information and Archival Science”, “Management Science”, “Economics”, “Pedagogy”, “Computer Science”, “Earth

<sup>\*</sup> 本文系国家社会科学基金重大项目“面向知识创新服务的数据科学理论与方法研究”(16ZDA224)的研究成果之一。(This article is an outcome of the key project “Theory and Method Research of Data Science Towards Knowledge Innovation Service”(NO.16ZDA224) supported by National Social Science Foundation of China.)

通信作者:陈必坤,chenbikun@njust.edu.cn, ORCID: 0000-0002-7252-3906 (Correspondence should be addressed to CHEN Bikun, Email: chenbikun@njust.edu.cn, ORCID:0000-0002-7252-3906)

Science”, “Math” and “Biology”, and explored Chinese usage patterns from user platform preferences’ aspect and user interest preferences’ aspect by analyzing usage data of journal official sites and pay-for-access platforms. Operatively, comparisons of user platform preferences were implemented by descriptive statistics and correlation analysis and comparisons of user interest preferences were explored by Jaccard similarity coefficient and co-word analysis.

Firstly, it proved that there were differences in user platform preferences between journal official sites and pay-for-access platforms. From the perspective of discipline, users tended to use journal official sites rather than pay-for-access platforms to obtain papers, and the paper downloads of social science were higher than natural science in pay-for-access platforms. From the perspective of journals, social science users were more inclined to use pay-for-access platforms than natural science users to obtain papers. Compared with pay-for-access platforms, the paper downloads of journal official sites showed a more obvious Matthew effect. The paper downloads of journal official sites showed a lower correlation with pay-for-access platforms overall. Secondly, it proved that there were differences in user interest preferences between journal official sites and pay-for-access platforms. The top 20% most downloaded papers in journal official sites and pay-for-access platforms showed a lower Jaccard similarity. In the example of library, information and literature science, the users of journal official sites and pay-for-access platforms paid attention to different topics respectively.

In a word, this study tried to comparatively investigate the usage patterns and its influencing factors by analyzing the usage data of Chinese journal official sites and pay-for-access platforms. However, the factors that affect the usage patterns were very complicated. How to probe into the mechanism of influencing factors and then verify them was still an important problem to be solved in further study. 7 figs. 5 tabs. 37 refs.

## KEY WORDS

Academic literature. Usage data. Usage Metrics. Academic journals. Academic communication.  
Usage pattern.

## 0 引言

随着网络化、数字化的发展以及相关技术的进步,科研人员越来越倾向于通过网络获取数字学术文献从事科学研究,这使得科研用户的学术文献使用数据(即学术文献的 HTML 格式浏览数据与 PDF 等格式的下载数据<sup>[1]</sup>)得以被记录。在此背景下,越来越多的研究者对学术文献使用数据进行采集、整理和挖掘分析,以发现与用户使用行为相关的特点、规律,或者通过学术文献使用数据进行相关性研究,由此产生了被国际学术界称为 Usage Metrics 的研究热点。

在科学文献交流与利用的研究中,采集与分析用户的订购、阅览、检索、引用等数据并进

行相关探索一直是业界关注的主要内容。早在纸质文献出版和使用时代,就有学者通过追踪和收集纸质出版物的使用数据开展研究,比如图书馆通过分析馆内订阅的纸质出版物使用数据进行资源建设与评价<sup>[2-3]</sup>和用户研究<sup>[4]</sup>。随着各种学术交流平台对用户交互信息记录功能的不断完善,使用数据被视作一种独立的数据类型<sup>[5-6]</sup>。笔者经过调查发现,目前图书馆学情报学领域的学者主要从五个方面进行 Usage Metrics 研究:①通过使用数据研究用户行为模式,如科学家工作时间<sup>[7-8]</sup>、用户使用偏好<sup>[9-13]</sup>以及用户时序使用模式<sup>[14-15]</sup>等;②通过使用数据研究文献老化规律,一般从历时或共时两方面进行分析,如 Moed 和 Halevi<sup>[16-17]</sup>先后从期刊和国家等视角研究使用数据的历时和共时老化规

律, Gorraiz 等<sup>[18]</sup>从学科视角进行了相关研究; ③运用使用数据探测学科领域的研究趋势<sup>[19-20]</sup>; ④将使用数据作为评价期刊、作者、机构或国家影响力的指标, 具体包括作为单个指标进行评价<sup>[21-23]</sup>或与 Altmetrics 指标结合进行评价<sup>[24-25]</sup>两种形式; ⑤探索使用数据与其他数据的相关性, 主要包括使用数据与引用数据的相关性<sup>[26-31]</sup>、不同平台使用数据的相关性<sup>[9, 14]</sup>、使用数据与作者数量<sup>[32]</sup>或基金资助数据<sup>[33]</sup>的相关性等。

综上所述, 不同学科领域的学者已从多个方面研究学术文献使用数据, 取得了较为丰硕的成果。但上述研究的数据来源主要是英文使用数据(如 Web of Science、Elsevier 和 Springer-Nature 等), 而中文使用数据方面的研究仅限于 CNKI 平台, 数据类型单一且样本数量有限。目前越来越多的中文期刊官网开始提供学术论文的 HTML 浏览数据或 PDF 下载数据, 给中文 Usage Metrics 的研究提供了新契机。那么, 通过中文学术论文交流与利用平台, 尤其是对期刊官网平台和信息集成平台(比如 CNKI)的使用数据进行用户平台偏好与用户兴趣偏好比较分析, 可以得出哪些用户使用模式的规律, 并进而提供哪些应用决策的依据, 是本文研究的出发点。

本文选取“图书馆、情报与文献学”“管理学”“经济学”“教育学”“计算机科学技术”“地球科学”“数学”和“生物学”八个学科被 CSSCI 或 CSCD 收录的 61 本开放获取期刊发表于 2014—2015 年的学术论文为样本, 通过网络爬虫等方式获取其在期刊官网与 CNKI 上的浏览、下载与被引数据, 尝试比较研究不同平台中用户平台选择偏好与内容兴趣偏好, 进而发现用户使用模式的规律及其影响因素, 为新技术背景下中文 Usage Metrics 的理论研究与应用提供参考。

## 1 数据与方法

### 1.1 数据与处理

笔者自 2014 年 01 月起开始追踪调研

CSSCI 和 CSCD 中文期刊(均含扩展版)的使用数据, 选取当时即开放获取的期刊作为追踪对象。最终依据《中文社会科学引文索引(CSSCI)来源期刊和收录集刊(2017—2018)目录》和《中国科学引文数据库(CSCD)来源期刊(2017—2018)目录》, 选取“图书馆、情报与文献学”“管理学”“经济学”“教育学”“计算机科学技术”“地球科学”“数学”和“生物学”八个学科(前四个属于社会科学, 后四个属于自然科学)的 61 本学术期刊为研究样本。具体选择条件如下。①样本期刊官网所有论文可开放获取且被 CNKI 收录, 保证用户能够在期刊官网和 CNKI 浏览并下载学术论文。②样本期刊论文发表时间跨度为 2014 年 1 月至 2015 年 12 月, 以保证文献浏览、下载和引用数据积累到稳定状态(通常是文章发表后的 2—3 年<sup>[34]</sup>)。③样本期刊官网的 2014—2015 年论文浏览或下载数据必须完整并且为即时数据, 即从论文正式出版后, 期刊官网即可浏览或下载摘要或全文; 若是在 2014 年 2 月之后启用新网站或 2014—2015 年部分论文浏览或下载数据缺失, 则不纳入抽样范围。

数据处理步骤如下: 首先, 从 CNKI 获取所有论文的下载次数、被引次数及其他元数据(如论文标题、作者、关键词等); 其次, 通过网络爬虫从期刊官网采集所有论文的浏览次数、下载次数及其他元数据(如论文标题和作者等); 再次, 通过“论文标题”and“作者”字段或 DOI 字段匹配将 CNKI 与期刊官网数据融合; 最后, 删除公告、新闻等, 仅保留学术论文, 样本数据见表 1。“图书馆、情报与文献学”“管理学”“经济学”和“教育学”的样本数据获取与预处理时间为 2018 年 3 月 14 日至 3 月 21 日, “计算机科学技术”“地球科学”“数学”和“生物学”的样本数据获取与预处理时间为 2018 年 4 月 1 日至 4 月 2 日。此外, CNKI 使用数据是学术论文的 PDF 与 CAJ 格式下载次数总和, 期刊官网使用数据是学术论文的 HTML 浏览次数(摘要浏览或全文浏览)和 PDF 下载次数。由于 CNKI 不提供 HTML 浏览次数, 无法与期刊官网浏览次数进行比较, 故本文仅关注两个平台的学术论文下载次数。

表 1 样本数据

| 学科         | 论文数量  | 期刊列表                                                                                                                           |
|------------|-------|--------------------------------------------------------------------------------------------------------------------------------|
| 图书馆、情报与文献学 | 2 189 | 中国图书馆学报、情报资料工作、图书情报知识、情报杂志、现代情报                                                                                                |
| 管理学        | 3 237 | 管理学报、中国管理科学、系统工程理论与实践、科学决策、科技进步与对策                                                                                             |
| 经济学        | 1 378 | 经济学家、经济理论与经济管理、政治经济学评论、南方经济、贵州财经大学学报、经济与管理研究                                                                                   |
| 教育学        | 1 152 | 高校教育管理、研究生教育研究、华东师范大学学报(教育科学版)、远程教育杂志、学位与研究生教育                                                                                 |
| 计算机科学技术    | 9 173 | 计算机工程、计算机应用、计算机科学、计算机应用研究、计算机工程与应用、计算机科学与探索、计算机集成制造系统、小型微型计算机系统                                                                |
| 地球科学       | 4 466 | 地学前缘、地质科学、地球科学、地球科学进展、地球物理学报、地球物理学进展、天然气地球科学、石油地球物理勘探、大地测量与地球动力学、矿物岩石地球化学通报、吉林大学学报(地球科学版)                                      |
| 数学         | 1 418 | 数学进展、数学学报、应用数学学报、数学物理学报、系统科学与数学、应用数学和力学                                                                                        |
| 生物学        | 6 055 | 生态学杂志、生物多样性、微生物学报、水生生物学报、生物技术通报、生物工程学报、微生物学通报、医用生物力学、中国生物防治学报、中国生物制品学杂志、生态与农村环境学报、食品与生物技术学报、应用与环境生物学报、中国生物医学工程学报、中国生物化学与分子生物学报 |

## 1.2 研究范围与方法

确定研究范围和选择研究方法的关键,一是要弄清所分析的数据是如何产生以及如何表现出来的,二是要明辨通过数据可以做哪些研究以及如何研究。

首先,从来源与表现看,学术论文用户使用数据在学术交流的过程中产生,是交流要素在一定交流模式下相互关联、相互作用的表征。借助于经典的科学文献信息正式交流模式——兰开斯特模式<sup>[35]95-97</sup>,结合网络环境下信息交流的情景,梳理用户、期刊出版机构、信息集成商与信息中心(主要指非营利性的图书馆等机构)等交流要素在科学交流中的角色与相互作用关系,可将产生使用数据的学术交流过程与模式描述为:用户主要通过投稿等方式与期刊出版机构建立论文生产关系;期刊出版机构通过出版、转让纸质或电子版论文与用户、信息中心和

信息集成商建立出版发行关系;信息集成商通过购买多个期刊出版机构的电子版论文使用权构建学术数据库,与期刊出版机构建立关系;信息中心通过购买纸质期刊或学术数据库与期刊出版机构和信息集成商建立关系;用户是通过期刊出版机构(纸质版或者期刊官网)、信息集成商或信息中心等多种渠道使用学术论文(如图1所示)。本文重点关注用户通过学术期刊官网或信息集成商平台(用户直接访问或者通过信息中心间接访问)进行学术交流过程中的用户使用数据,基于 Usage Metrics 进行研究。

其次,结合已有研究,基于不同平台的数据,本文认为利用 Usage Metrics 可以比较研究用户使用偏好,从而揭示用户使用模式的特点与规律。用户使用偏好主要包括平台选择偏好(渠道)和研究兴趣偏好(信息资源),其中平台选择偏好会影响研究兴趣偏好(渠道选择对信息

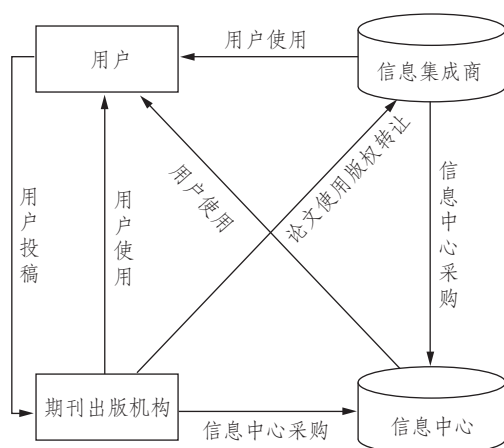


图1 产生使用数据的学术交流过程与模式图

资源的影响)。此外,已有研究表明:使用数据次数一定程度上反映了用户对论文的关注与兴趣程度(如论文主题、作者和机构等),使用次数越高,关注与兴趣越大<sup>[12, 17, 31]</sup>。因此,本文通过使用数据研究用户平台偏好和用户兴趣偏好,进而达到发现用户使用模式及其影响因素的目的。具体研究问题如下:学术期刊官网与信息集成平台存在哪些用户平台偏好差异?学术期刊官网与信息集成平台存在哪些用户兴趣偏好差异?其中影响用户使用模式的因素包括哪些?

由上,本文通过 Usage Metrics 研究和比较用户平台选择偏好与内容兴趣偏好,以探索用户使用模式的规律及其影响因素。本文选择的研究方法与设计的操作环节如下:首先,采用数量统计和斯皮尔曼系数计算方法,对 CNKI 使用数据和期刊官网使用数据分别进行描述性统计和相关性分析,以揭示用户平台偏好特征及其影响因素;然后,分别采用 Jaccard 系数计算和共词分析方法,从指标和内容两个不同测度进行比较分析,以研究用户兴趣偏好特征及其影响因素。其中,描述性统计、相关性分析和用户兴趣分析(指标视角)通过 R 语言<sup>①</sup>实现,用户兴趣分析(内容视角)通过软件 VOSviewer<sup>[36]</sup>实现(见图2)。

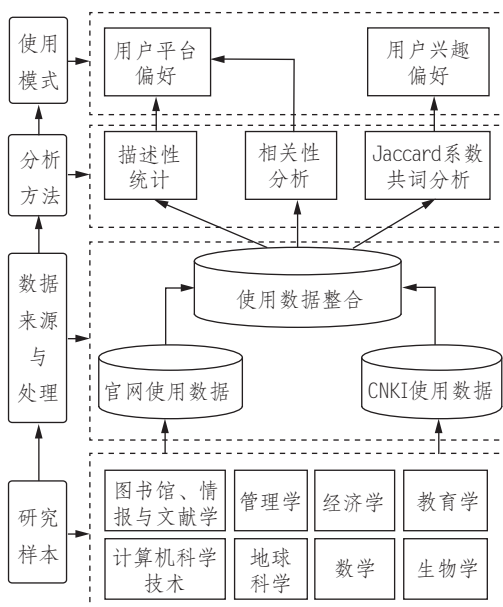


图2 研究的技术路线图

## 2 研究内容与结果

### 2.1 用户平台偏好比较分析

根据前面所获得的数据,与用户平台偏好有关的使用数据主要包括:学科属性、平台类型、期刊名称、下载次数、被引次数。下面分别对上述数据进行统计分析。

#### (1) 基本统计分析

一是从学科视角来看(见表2):除管理学外,其他学科的官网篇均下载次数均大于 CNKI 篇均下载次数,表明用户整体上倾向于使用开放获取期刊官网而不是付费的信息集成平台获取所需论文。对不同学科的官网篇均下载次数和 CNKI 篇均下载次数进行汇总,首先可知地球科学的篇均下载次数最高,其次是图书馆、情报与文献学、经济学、教育学、生物学和管理学,最后是计算机科学技术学和数学。此外,在 CNKI 篇均下载次数上,社会科学均大于自然科学。

① <https://www.r-project.org>

表 2 八个学科使用数据的基本统计

| 学科         | 类型      | 中值    | 均值    | 均值<br>汇总 | 学科      | 类型      | 中值  | 均值      | 均值<br>汇总 |
|------------|---------|-------|-------|----------|---------|---------|-----|---------|----------|
| 图书馆、情报与文献学 | 官网下载    | 366   | 894.9 | 1 389    | 计算机科学技术 | 官网下载    | 369 | 517.3   | 663      |
|            | CNKI 下载 | 270   | 494.4 |          |         | CNKI 下载 | 95  | 146.1   |          |
| 管理学        | 官网下载    | 334   | 434.1 | 928      | 地球科学    | 官网下载    | 475 | 1 978.3 | 2 221    |
|            | CNKI 下载 | 362   | 493.9 |          |         | CNKI 下载 | 175 | 242.9   |          |
| 经济学        | 官网下载    | 287   | 761.4 | 1 339    | 数学      | 官网下载    | 289 | 359.6   | 434      |
|            | CNKI 下载 | 370.5 | 577.2 |          |         | CNKI 下载 | 52  | 75.2    |          |
| 教育学        | 官网下载    | 371   | 586.1 | 1 139    | 生物学     | 官网下载    | 494 | 900.2   | 1 134    |
|            | CNKI 下载 | 312.5 | 553.2 |          |         | CNKI 下载 | 153 | 233.5   |          |

二是从单个期刊视角来看,其统计结果与学科视角存在差别,1/3 的 CSSCI 收录的期刊与 37/40 的 CSCD 收录的期刊官网篇均下载次数大于 CNKI 篇均下载次数,这表明:在社会科学中,大多数期刊的用户更倾向于使用付费的 CNKI 获取所需论文,而在自然科学中,大多数期刊的用户更倾向于使用官网获取所需论文。

#### (2) 偏离度分析

表 2 显示,各个学科的篇均下载次数均高于中值,表明数据的分布存在集聚效应。为了进一步揭示数据分布的集聚效应,本文应用偏离度(Skewness)方法展开分析,首先将不同学科学术论文的官网下载次数、CNKI 下载次数和被引次数分别降序排列,然后计算三种数据的累积百分比,最后绘制数据累积百分比与论文数量累积百分比函数图,如图 3 和图 4 所示。结合图 3 与图 4 可知,官网下载次数、CNKI 下载次数以及 CNKI 被引次数均存在集聚效应,其中 CNKI 下载次数的集聚效应最弱,官网下载次数和 CNKI 被引次数的集聚效应程度相当。

从学科视角来看,各个学科前 25% 的学术论文贡献了超过 50% 的下载次数和 60% 左右的被引次数。具体来讲,在 CNKI 平台中,图书馆、情报与文献学和教育学的下载次数集聚效应最强,其次是经济学、计算机科学技术、生物学、管

理学、地球科学和数学。教育学和数学的被引次数集聚效应最强,其次是计算机科学技术和图书馆、情报与文献学,最后是经济学、生物学、地球科学和管理学;在官网平台中,图书馆、情报与文献学和地球科学的下载次数集聚效应最强,其次是教育学、经济学、管理学和生物学,最后是计算机科学技术与数学。结合基本统计数据,可以发现,均值汇总数最高的专业性较强的地球科学的下载次数集聚效应在官网中较强,在 CNKI 平台上较低,其被引集聚也较低;在均值汇总数最低的基础性较强的数学的下载次数集聚效应在两个平台中都处于最弱,而被引却较强;均值汇总数处于中列的应用性较强的教育学的下载次数集聚效应在两个平台中都处于前列,被引也最强。

笔者还对图 3 和图 4 中八个学科的官网下载次数、CNKI 下载次数“尾部数据”(80%—100%)进行分析,发现官网下载次数普遍低于 CNKI 下载次数,表明“尾部数据”在信息集成平台的被关注度要高于期刊官网。从下载次数“头部数据”和“尾部数据”的整体分布情况来看,在学术论文交流与传播以及期刊影响力方面,信息集成平台具有吸引更宽泛用户并可能使每一篇学术论文有更多被利用机会的独特作用,而期刊官网呈现出的更为明显的“两级分化”

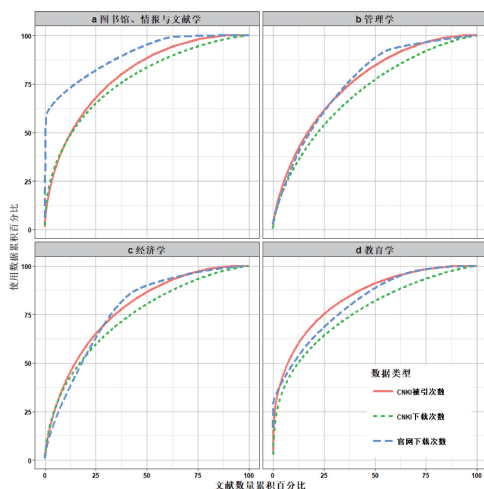


图3 社会科学不同类型数据的偏斜度分布

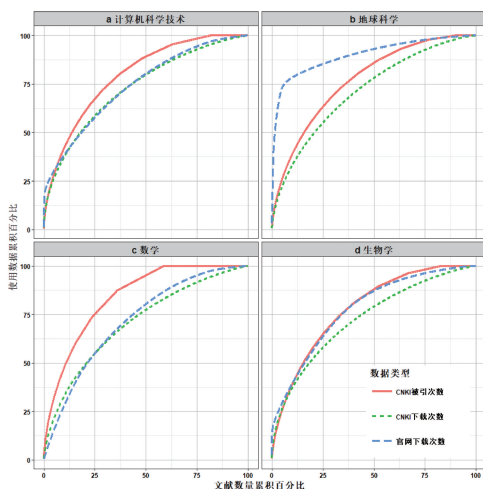


图4 自然科学不同类型数据的偏斜度分布

现象使其专业性的作用更强。进而证实,用户对于资源集聚平台与资源专业平台在使用选择上存在偏好差异。

### (3) 不同平台下载次数与被引次数分析

图5是不同学科期刊官网和CNKI下载次数与被引次数的散点图。图5a(期刊官网)与图5b(CNKI)是社会科学四个学科的散点图,红色圆点代表管理学、绿色三角形代表教育学、蓝色正方形代表经济学、紫色加号代表图书馆、情报与文献学;图5c(期刊官网)和图5d(CNKI)是自然科学四个学科的散点图,红色圆点代表地球科学、绿色三角形代表计算机科学技术、蓝色正方形代表生物学、紫色加号代表数学;图中的直线是不同学科的线性回归拟合线;图中的标签代表不同学科的线性回归系数与显著性检验 $p$ 值。为了清晰揭示下载次数与被引次数之间的关系,对期刊官网和CNKI下载次数分别取自然对数,并且仅保留数值小于125的CNKI被引次数。

由图5可知,社会科学与自然科学的整体分布情况类似,图5a和5c的学科分层明显而图5b和5d的学科分层模糊,并且图5a和5c的线性回归拟合线未呈现正相关关系,而图5b和5d呈现正相关关系(若CNKI下载次数未取自然对数,线性回归系数会更高,但是若官网下载次数

未取自然对数,线性回归系数依旧很低)。该现象与Web of Science数据平台存在相似之处。赵星<sup>[31]</sup>以Web of Science数据为例,证实在物理学和图情学中引文次数与使用次数的正相关性相对较强,计算机科学技术和经济学的相关系数强度不高,但显著性检验均已通过。具体来看,在期刊官网平台,社会科学和自然科学的线性回归系数均较低,且数学的回归系数为负数。在CNKI平台中,社会科学(图5b)的线性回归系数均大于自然科学(图5d),且社会科学中的四个子学科回归系数相近;自然科学中地球科学、计算机科学技术与生物学的回归系数相近,而数学较低。

上述研究表明,下载次数与被引次数在期刊官网平台上不存在相关关系,而在CNKI平台上具有相关关系。结合CNKI平台所提供的功能,发现:CNKI的默认检索结果页面同时提供论文下载次数与被引次数,用户点击下载或者被引链接就很容易看到“双高”的期刊论文,即论文下载次数或被引次数越高,被人关注较多且影响力越大,高关注度和高影响力的论文更能吸引用户下载阅读;而期刊官网不提供上述检索服务,用户只能根据需要自行判断下载。因此,上述研究结果也表明用户的下载和引用

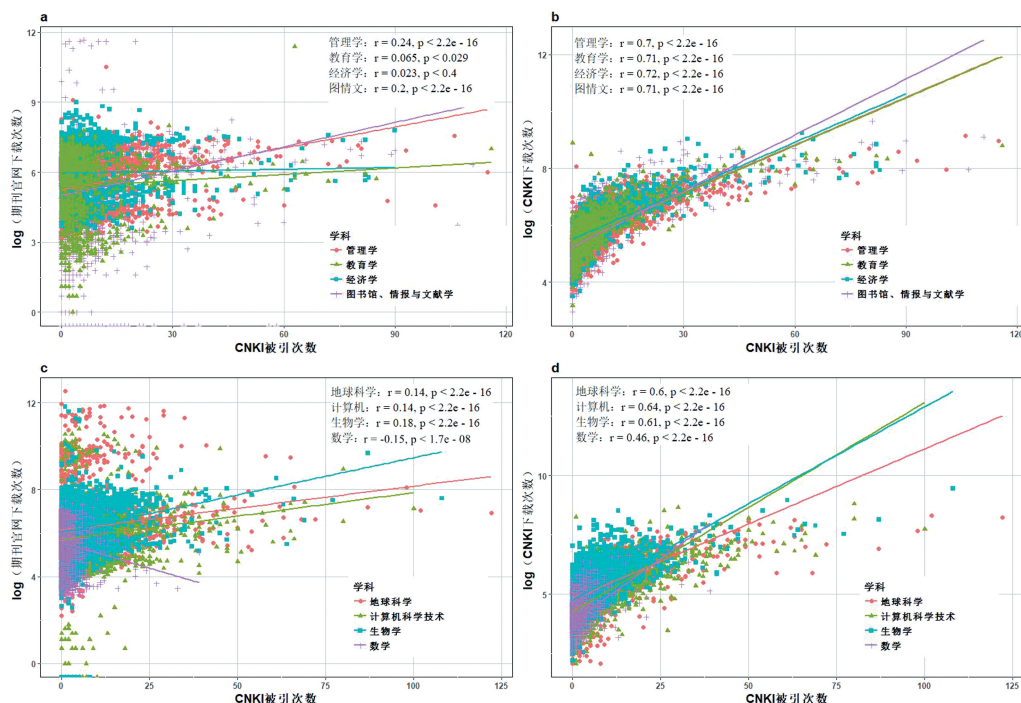


图5 不同学科的 CNKI 和期刊官网下载次数与 CNKI 被引次数散点图

模式受到平台所提供功能的影响。

#### (4) 不同平台下载次数的相关性分析

前文从基本统计和偏离度等角度分析了期刊官网与 CNKI 平台的学术论文下载次数,为了进一步分析二者的关系,本文计算了各个学科及期刊的斯皮尔曼系数,如表 3 和表 4 所示。由表 3 可知,在学科上,仅有图书馆、情报与文献学的官网下载次数与 CNKI 下载次数呈现中度相

关关系,其他学科均呈现不相关甚至负相关关系。由表 4 可知,在期刊上,仅有《中国生物制品学杂志》的官网下载次数与 CNKI 下载次数呈现很强的相关关系,另有 14 本期刊(见表 4 中加粗字体)的官网下载次数与 CNKI 下载次数呈现中度相关关系。整体而言,官网下载次数与 CNKI 下载次数的相关性较低,表明不同平台的用户使用偏好存在较大差异。

表 3 不同学科 CNKI 下载次数与官网下载次数的斯皮尔曼系数

| 学科         | 相关性      | 学科      | 相关性       |
|------------|----------|---------|-----------|
| 图书馆、情报与文献学 | 0.406*** | 计算机科学技术 | 0.087***  |
| 管理学        | 0.324*** | 地球科学    | -0.013    |
| 经济学        | 0.035    | 数学      | -0.219*** |
| 教育学        | 0.065*   | 生物学     | 0.316***  |

(\*  $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\*  $p < 0.001$ )

表 4 不同期刊 CNKI 下载次数与官网下载次数的斯皮尔曼系数

| 期刊              | 相关性                        | 期刊            | 相关性                        |
|-----------------|----------------------------|---------------|----------------------------|
| 中国图书馆学报         | 0.127                      | 地球物理学进展       | 0.317 <sup>***</sup>       |
| 情报资料工作          | <b>0.479<sup>***</sup></b> | 地学前缘          | 0.228 <sup>**</sup>        |
| 图书情报知识          | 0.291 <sup>***</sup>       | 地质科学          | -0.249 <sup>**</sup>       |
| 情报杂志            | 0.363 <sup>***</sup>       | 天然气地球科学       | 0.367 <sup>***</sup>       |
| 现代情报            | 0.143 <sup>***</sup>       | 地球科学          | 0.187 <sup>***</sup>       |
| 管理学报            | 0.249 <sup>***</sup>       | 大地测量与地球动力学    | 0.073                      |
| 系统工程理论与实践       | 0.24 <sup>***</sup>        | 吉林大学学报(地球科学版) | 0.336 <sup>***</sup>       |
| 中国管理科学          | 0.272 <sup>***</sup>       | 矿物岩石地球化学通报    | -0.247 <sup>***</sup>      |
| 科学决策            | 0.211 <sup>*</sup>         | 石油地球物理勘探      | <b>0.408<sup>***</sup></b> |
| 科技进步与对策         | 0.215 <sup>***</sup>       | 数学进展          | 0.076                      |
| 经济学家            | <b>0.406<sup>***</sup></b> | 数学学报          | 0.281 <sup>***</sup>       |
| 经济理论与经济管理       | -0.043                     | 系统科学与数学       | -0.075                     |
| 南方经济            | 0.156 <sup>*</sup>         | 应用数学和力学       | -0.019                     |
| 贵州财经大学学报        | -0.009                     | 应用数学学报        | 0.114                      |
| 经济与管理研究         | -0.017                     | 数学物理学报        | 0.123 <sup>*</sup>         |
| 政治经济学评论         | 0.3 <sup>***</sup>         | 生态学杂志         | <b>0.462<sup>***</sup></b> |
| 高校教育管理          | 0.311 <sup>***</sup>       | 生态与农村环境学报     | 0.242 <sup>***</sup>       |
| 研究生教育研究         | 0.06                       | 生物技术通报        | <b>0.613<sup>***</sup></b> |
| 华东师范大学学报(教育科学版) | 0.267 <sup>**</sup>        | 水生生物学报        | <b>0.414<sup>***</sup></b> |
| 远程教育杂志          | 0.146                      | 中国生物化学与分子生物学报 | <b>0.526<sup>***</sup></b> |
| 学位与研究生教育        | -0.018                     | 生物多样性         | <b>0.418<sup>***</sup></b> |
| 计算机工程           | 0.297 <sup>***</sup>       | 中国生物制品学杂志     | <b>0.951<sup>***</sup></b> |
| 计算机集成制造系统       | 0.302 <sup>***</sup>       | 中国生物医学工程学报    | <b>0.415<sup>***</sup></b> |
| 计算机科学与探索        | <b>0.559<sup>***</sup></b> | 中国生物防治学报      | 0.394 <sup>***</sup>       |
| 计算机应用           | 0.287 <sup>***</sup>       | 生物工程学报        | <b>0.502<sup>***</sup></b> |
| 小型微型计算机系统       | 0.333 <sup>***</sup>       | 食品与生物技术学报     | 0.069                      |
| 计算机工程与应用        | 0.38 <sup>***</sup>        | 微生物学报         | <b>0.489<sup>***</sup></b> |
| 计算机科学           | 0.084 <sup>**</sup>        | 微生物学通报        | <b>0.544<sup>***</sup></b> |
| 计算机应用研究         | 0.377 <sup>***</sup>       | 医用生物力学        | 0.091                      |
| 地球科学进展          | 0.341 <sup>***</sup>       | 应用与环境生物学报     | 0.339 <sup>***</sup>       |
| 地球物理学报          | <b>0.465<sup>***</sup></b> | —             | —                          |

(\* p&lt;0.05; \*\* p&lt;0.01; \*\*\* p&lt;0.001)

## 2.2 用户兴趣偏好比较分析

通过上述描述性统计分析与相关性分析获知,不同学科及期刊的用户平台偏好存在较大差异,那么不同学科及期刊的用户兴趣偏好到底如何?下面从指标测度与内容分析两个不同层面进行比较。

### (1) 基于指标测度的比较

在已有研究中,研究者大多选取某一学科或期刊在特定时间内(比如一年或一个月)下载次数前 10<sup>[31]</sup>、前 20<sup>[20]</sup>或前 20%<sup>[9]</sup>的论文作为用户兴趣分析的样本。另外,在图书馆、情报与文献学案例研究中,前 100、前 1 000、前 5%、前 10% 或前 20%的高频次实体(比如作者、机构或关键词等)通常被作为研究样本。本文根据帕累托提出的“二八定律”,分别选取每个期刊 CNKI

下载次数与官网下载次数前 20%的学术论文作为研究样本,运用 Jaccard 相似系数判断不同期刊 CNKI 下载次数与官网下载次数前 20%的学术论文的相似度,即用二者的交集与并集的比值从整体上来测度不同平台用户兴趣的相似程度。

由表 5 可知:15/61 的期刊 Jaccard 相似系数大于等于 0.3,即有 15 本期刊的 CNKI 下载次数与官网下载次数前 20%的学术论文拥有约 50%及以上的交集;9/61 的期刊 Jaccard 相似系数大于等于 0.25 且小于 0.3,即有 9 本期刊的 CNKI 下载次数与官网下载次数前 20%的学术论文拥有 40%—50%的交集;其余期刊的 Jaccard 相似系数小于 0.25,交集少于 40%。总体而言,期刊官网与 CNKI 的用户兴趣偏好存在较为明显的差异。

表 5 期刊 CNKI 下载次数与官网下载次数前 20%学术论文的 Jaccard 相似系数

| 期刊        | Jaccard | 期刊            | Jaccard |
|-----------|---------|---------------|---------|
| 中国图书馆学报   | 0.14    | 地球物理学进展       | 0.21    |
| 情报资料工作    | 0.24    | 地学前缘          | 0.21    |
| 图书情报知识    | 0.23    | 地质科学          | 0.10    |
| 情报杂志      | 0.25    | 天然气地球科学       | 0.24    |
| 现代情报      | 0.19    | 地球科学          | 0.22    |
| 管理学报      | 0.19    | 大地测量与地球动力学    | 0.14    |
| 系统工程理论与实践 | 0.21    | 吉林大学学报(地球科学版) | 0.18    |
| 中国管理科学    | 0.23    | 矿物岩石地球化学通报    | 0.04    |
| 科学决策      | 0.26    | 石油地球物理勘探      | 0.32    |
| 科技进步与对策   | 0.20    | 数学进展          | 0.19    |
| 经济学家      | 0.36    | 数学学报          | 0.23    |
| 经济理论与经济管理 | 0.05    | 系统科学与数学       | 0.09    |
| 南方经济      | 0.18    | 应用数学和力学       | 0.11    |
| 贵州财经大学学报  | 0.12    | 应用数学学报        | 0.17    |
| 经济与管理研究   | 0.09    | 数学物理学报        | 0.20    |
| 政治经济学评论   | 0.26    | 生态学杂志         | 0.30    |

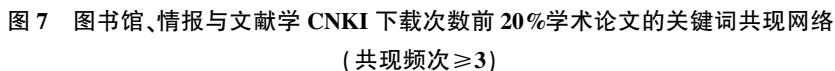
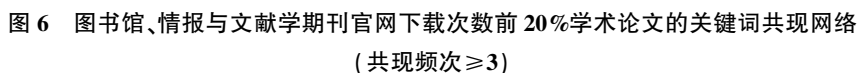
续表

| 期刊              | Jaccard | 期刊            | Jaccard |
|-----------------|---------|---------------|---------|
| 高校教育管理          | 0.30    | 生态与农村环境学报     | 0.24    |
| 研究生教育研究         | 0.13    | 生物技术通报        | 0.44    |
| 华东师范大学学报(教育科学版) | 0.20    | 水生生物学报        | 0.28    |
| 远程教育杂志          | 0.25    | 中国生物化学与分子生物学报 | 0.43    |
| 学位与研究生教育        | 0.11    | 生物多样性         | 0.31    |
| 计算机工程           | 0.23    | 中国生物制品学杂志     | 0.97    |
| 计算机集成制造系统       | 0.25    | 中国生物医学工程学报    | 0.31    |
| 计算机科学与探索        | 0.44    | 中国生物防治学报      | 0.25    |
| 计算机应用           | 0.22    | 生物工程学报        | 0.38    |
| 小型微型计算机系统       | 0.27    | 食品与生物技术学报     | 0.17    |
| 计算机工程与应用        | 0.31    | 微生物学报         | 0.38    |
| 计算机科学           | 0.15    | 微生物学通报        | 0.38    |
| 计算机应用研究         | 0.30    | 医用生物力学        | 0.09    |
| 地球科学进展          | 0.24    | 应用与环境生物学报     | 0.28    |
| 地球物理学报          | 0.24    | —             | —       |

## (2) 基于内容分析的比较

前文从指标层面证实了期刊官网与 CNKI 的用户兴趣偏好存在较为明显的差异,本部分利用共词分析方法从内容层面比较不同平台的用户兴趣偏好。限于篇幅及作者的专业背景,本文仅以“图书馆、情报与文献学”为例进行分析。首先,分别对每本期刊官网下载次数和 CNKI 下载次数前 20% 学术论文的关键词进行预处理(比如去除停用词,合并同义词等);然后,将处理后的结果输入到 VOSviewer 中进行共词分析,结果如图 6 和图 7 所示。图 6 和图 7 中的节点代表关键词,节点大小代表度数,节点之间的连线粗细代表共现关系强弱,不同颜色代表不同聚类。为了使图表更清晰,图 6 舍去 1 个离散节点,仅保留 89 个节点;图 7 无离散节点,共计 124 个相连节点。此外,为了更好地进行比较,图 6 与图 7 的节点大小、字体大小、连线粗细均采用相同参数。

由图 6 和图 7 可知:图 6 节点较少且连线较稀疏,共聚为 5 类(网络舆情相关、图书馆服务相关、科学计量相关、科学知识图谱相关、知识管理与知识服务相关);图 7 节点较多且连线密集,共聚为 5 类(图书馆相关、大数据相关、科学知识图谱相关、社会化媒体相关、网络舆情相关)。具体来看,图 6 中节点大小差异较小且主题分布较为均衡,而图 7 中节点大小差异较大且主题分布较为集中,这表明:期刊官网用户关注的研究主题比较平衡,而 CNKI 用户关注的研究主题比较集中,比如 CNKI 中的用户对“大数据”和“网络舆情”等社会热门主题的关注度要高于期刊官网中的用户。这也说明在热点研究主题采集上,用户更倾向于使用信息集成平台获取更多的相关研究成果。期刊官网锁定的用户主要是该领域的研究者,而 CNKI 用户群体则是多个领域的研究者。



### 3 结论与讨论

本文以八个学科的 61 本中文学术期刊为例,借助数理统计和共词分析等方法,通过对期刊官网和 CNKI 上的使用数据进行用户平台偏好和用户兴趣偏好比较分析,进而发现了中文学术论文用户使用模式的规律及其影响因素,得出如下结论。

(1) 在用户平台偏好比较上,学术期刊官网与信息集成平台的选择和使用存在着较大差异。①在平台的选择上,从学科视角来看,用户整体上倾向于选择开放获取期刊官网而不是付费信息集成平台来获取所需论文。此外,社会科学的 CNKI 篇均下载次数均大于自然科学的 CNKI 篇均下载次数。从期刊视角来看,在社会科学中,大多数用户更倾向于选择付费的 CNKI 获取所需论文,而在自然科学中,大多数用户更倾向于使用官网获取所需论文。②在平台的使用上,从集聚效应看,学术期刊官网与信息集成平台使用数据均存在集聚效应,与信息集成平台相比,学术期刊官网的用户下载数据呈现更为明显的“两级分化”现象;从相关性效应看,无论是学科视角还是单个期刊视角,官网下载次数与信息集成平台下载次数整体上的相关性较低。

(2) 在用户兴趣偏好比较上,学术期刊官网与信息集成平台的指标测度和内容分析都存着偏好差异。①基于指标测度,不同期刊官网下载次数与信息集成平台下载次数前 20% 的学术论文的 Jaccard 相似系数较低;②基于共词分析,

期刊官网与信息集成平台用户关注的研究主题内容不同。其中,期刊官网用户关注的研究主题比较平衡,而信息集成平台用户关注的研究主题相对集中。

(3) 通过比较发现,学术期刊官网与信息集成平台各自具有独特的特点。通过数据分析,笔者认为造成不同平台用户使用模式不同的影响因素包括以下方面。①学术论文交流与利用的生态:由于现有的“开放获取期刊官网资源有限,信息集成平台集中了绝大部分的学术期刊资源”的格局<sup>[37]</sup>,使得学术资源的多寡影响着用户的渠道选择;②学术论文交流与利用的付费方式:“机构付费、机构个体用户免费”是当前我国信息集成平台资源的主流使用方式,高校研究所大量用户能够直接免费使用信息集成平台的产品,也影响了用户的使用模式;③学术论文交流与利用平台的功能设置:学术期刊的自身质量、网站建设与网站访问设置,以及信息集成平台的检索结果排序机制等会影响用户的使用;④学术论文交流与利用的主体类别与需求:用户群体规模差异和任务目标差异也会影响用户使用资源,信息集成平台用户群体规模要远大于各个期刊官网,期刊官网用户多为该领域的研究者,而信息集成平台用户群体则是多个领域的研究者。此外,当用户需要检索特定主题的研究论文时,一般会优先选择信息集成平台作为浏览或获取的途径,信息集成平台的资源集聚优势能够保证用户所需论文的全面系统性,而期刊官网更多的是吸引本学科或相关学科领域的研究者定期浏览、下载学术论文,以把握本学科的整体研究主题以及研究动态与趋势。

### 参考文献

- [1] 王贤文,毛文莉,王治. 基于论文下载数据的科研新趋势实时探测与追踪[J]. 科学学与科学技术管理, 2014, 35(4):3-9. (Wang Xianwen, Mao Wenli, Wang Zhi. Detecting and tracing emerging research trends real-timely using Web data[J]. Science of Science and Management of Science & Technology, 2014, 35(4):3-9.)
- [2] Coombs K A. Lessons learned from analyzing library database usage data[J]. Library Hi Tech, 2005, 23(4): 598-609.

- [ 3 ] Kraemer A. Ensuring consistent usage statistics, part 2: working with use data for electronic journals[J]. The Serials Librarian, 2006, 50(1/2): 163-172.
- [ 4 ] Pesch O. Ensuring consistent usage statistics, part 1: Project COUNTER[J]. The Serials Librarian, 2006, 50(1/2): 147-161.
- [ 5 ] Glänzel W, Gorraiz J. Usage metrics versus Altimetrics: confusing terminology? [J] Scientometrics, 2015, 102(3): 2161-2164.
- [ 6 ] 王贤文.科学计量大数据及其应用[M]. 科学出版社, 2016: 30.(Wang Xianwen. Scientometrics big data and its applications[M]. Science Press, 2016: 30.)
- [ 7 ] Wang X, Xu S, Peng L, et al. Exploring scientists' working timetable: do scientists often work overtime?[J]. Journal of Informetrics, 2012, 6(4): 655-660.
- [ 8 ] Wang X, Peng L, Zhang C, et al. Exploring scientists' working timetable: a global survey[J]. Journal of Informetrics, 2013, 7(3): 665-675.
- [ 9 ] Chen B. Usage pattern comparison of the same scholarly articles between Web of Science (WoS) and Springer[J]. Scientometrics, 2018, 115(1): 519-537.
- [ 10 ] Davis P M, Solla L R. An IP-level analysis of usage statistics for electronic journals in chemistry: making inferences about user-behavior[J]. Journal of the American Society for Information Science and Technology, 2003, 54(11): 1062-1068.
- [ 11 ] Davis P M, Price J S. eJournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER[J]. Journal of the American Society for Information Science and Technology, 2006, 57(9): 1243-1248.
- [ 12 ] Wang X, Fang Z, Sun X. Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count[J]. Scientometrics, 2016, 109(2): 917-926.
- [ 13 ] Wang X, Xu S, Fang Z. Tracing digital footprints to academic articles: an investigation of PeerJ publication referral data[EB/OL]. [2018-10-28].<http://cn.arxiv.org/abs/1601.05271>.
- [ 14 ] Chen B, Zhong Z, Zhan C. Usage pattern analysis of academic articles from two Chinese journals[C]// ISSI. Proceedings of the 16th International Conference on Scientometrics and Informetrics. Wuhan, 2017: 366-375.
- [ 15 ] Khan M S, Younas M. Analyzing readers behavior in downloading articles from IEEE digital library: a study of two selected journals in the field of education[J]. Scientometrics, 2017, 110(3): 1523-1537.
- [ 16 ] Moed H F. Statistical relationships between downloads and citations at the level of individual documents within a single journal[J]. Journal of the American Society for Information Science and Technology, 2005, 56(10): 1088-1097.
- [ 17 ] Moed H F, Halevi G. On full text download and citation distributions in scientific-scholarly journals[J]. Journal of the Association for Information Science and Technology, 2016, 67(2): 412-431.
- [ 18 ] Gorraiz J, Gumpenberger C, Schlögl C. Usage versus citation behaviours in four subject areas[J]. Scientometrics, 2014, 101(2): 1077-1095.
- [ 19 ] Bollen J, Luce R, Vermulapalli S S, et al. Usage analysis for the identification of research trends in digital libraries[EB/OL]. [2018-10-28].<http://www.dlib.org/dlib/may03/bollen/05bollen.html>.
- [ 20 ] Wang X, Wang Z, Xu S. Tracing scientist's research trends realtively[J]. Scientometrics, 2013, 95(2): 717-729.
- [ 21 ] Wan J, Hua P, Rousseau R, et al. The journal download immediacy index (DII): experiences using a Chinese full-text database[J]. Scientometrics, 82(3): 555-566.
- [ 22 ] De Sordi J O, Conejero M A, Meireles M. Bibliometric indicators in the context of regional repositories: proposing

- the D-index[J]. *Scientometrics*, 2016, 107(1): 235-258.
- [23] Chi P S, Glänzel W. Comparison of citation and usage indicators in research assessment in scientific disciplines and journals[J]. *Scientometrics*, 2018, 116(1): 537-554.
- [24] Bollen J, Sompel H V D, Smith J A, et al. Toward alternative metrics of journal impact: a comparison of download and citation data[J]. *Information Processing & Management*, 2005, 41(6): 1419-1440.
- [25] Kurtz M J, Henneken E A. Measuring metrics-a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics[J]. *Journal of the Association for Information Science and Technology*, 2017, 68(3): 695-708.
- [26] O'Leary D E. The relationship between citations and number of downloads in decision support systems[J]. *Decision Support Systems*, 2008, 45(4): 972-980.
- [27] Kurtz M J, Bollen J. Usage bibliometrics[J]. *Annual Review of Information Science and Technology*, 2010, 44(1): 3-64.
- [28] Subotic S, Mukherjee B. Short and amusing: the relationship between title characteristics, downloads, and citations in psychology articles[J]. *Journal of Information Science*, 2014, 40(1): 115-124.
- [29] Schloegl C, Gorraiz J, Gumpenberger C, et al. Comparison of downloads, citations, and readership data for two information systems journals[J]. *Scientometrics*, 2014, 101(2): 1113-1128.
- [30] 曹艺, 王曰芬, 丁洁. 面向学术影响力评价的科技文献引用与下载的相关性研究[J]. *图书情报工作*, 2012, 56(8): 56-64. (Cao Yi, Wang Yuefen, Ding Jie. Relativity between citation and download indicators of scientific papers for academic influence evaluation[J]. *Library and Information Service*, 2012, 56(8): 56-64.)
- [31] 赵星. 学术文献用量级数据 Usage 的测度特性研究[J]. *中国图书馆学报*, 2017, 43(3): 44-57. (Zhao Xing. Exploring the features of usage data for academic literatures[J]. *Journal of Library Science in China*, 2017, 43(3): 44-57.)
- [32] Chi P S, Glänzel W. An empirical investigation of the associations among usage, scientific collaboration and citation impact[J]. *Scientometrics*, 2017, 112(1): 403-412.
- [33] Zhao S X, Lou W, Tan A M, et al. Do funded papers attract more usage?[J]. *Scientometrics*, 2018, 115(1): 153-168.
- [34] Lippi G, Favalaro E J. Article downloads and citations: is there any relationship?[J]. *Clinica Chimica Acta*, 2013, 415: 195.
- [35] 岳剑波. 信息管理基础[M]. 北京: 清华大学出版社, 1999: 95-97. (Yue Jianbo. *Information management foundations*[M]. Beijing: Tsinghua University Press, 1999: 95-97.)
- [36] Eck N J, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping[J]. *Scientometrics*, 2010, 84(2): 523-538.
- [37] 朱剑. 学术新媒体: 缘何难以脱颖而出——兼及学术传播领域媒体融合发展[J]. *北京交通大学学报(社会科学版)*, 2015, 14(4): 7-17. (Zhu Jian. When will academic new media excel on media integration and development in academic communications[J]. *Journal of Beijing Jiaotong University (Social Sciences Edition)*, 2015, 14(4): 7-17.)

陈必坤 南京理工大学经济管理学院信息管理系讲师。江苏 南京 210094。

周慧娴 南京理工大学经济管理学院信息管理系硕士研究生。江苏 南京 210094。

钟周燕 南京理工大学经济管理学院信息管理系硕士研究生。江苏 南京 210094。

王曰芬 南京理工大学经济管理学院信息管理系教授。江苏 南京 210094。

(收稿日期: 2018-09-17; 修回日期: 2018-10-12)