社会科学进展
2022 年 12 月第 4 卷第 6 期

分类文献下载量数据统计特征分析系统设计与应用

魏立杰[1] 刘荫明[1] 孙江文[1] 杨久强[2] 林年添[2]

1．山东科技大学图书馆，青岛；
2．山东科技大学地球科学与工程学院，青岛

摘　要 | 透过分类文献下载量的大小及分布特征，可以了解读者的阅读行为及其与学科发展的关联性。本文根据分类文献下载量数据统计特点，有针对性地设计及研发数据统计特征分析系统，并将其应用于实际案例分析中。应用结果表明，多样性的分析方法提高了分类文献下载量数据统计分析的可靠性，有助于更好更全面地了解读者阅读规律和下载特点，为数字图书馆或资料库建设，为优化学科及学科的有效发展，提供科学量化管理依据[1]。

关键词 | 分类文献下载量；系统设计；下载量热力值；聚类分析；特征分析

信息技术的不断发展进一步改变了大学图书馆用户服务和寻求信息的行为，即从实体的资源分享拓展到了虚拟数字资源分享。显然，互联网的发展使图书呈现的形式发生了重大变化，尤其是数字化图书的出现改变了人们的阅读行为[2]。

基金项目：山东省软科学项目“高校国家知识产权信息服务中心建设路径与发展策略研究（2021RKY04065）”；山东省本科教学改革研究项目“新工科背景下行业高校地质类传统专业的升级改造与创新人才培养模式（M2020257）”；山东科技大学课程思政培育项目“地震勘探原理（KCSZ201904）”。
通讯作者：林年添，博士，山东科技大学地球科学与工程学院教授，研究方向：专业课教学理论及应用研究，E-mail：377237866@qq.com。
文章引用：魏立杰，刘荫明，孙江文，等．分类文献下载量数据统计特征分析系统设计与应用［J］．社会科学进展，2022，4（6）：467-475.
https://doi.org/10.35534/pss.0406041

了解线上分类文献（数字图书或文章等）下载量的大小及分布特征，有助于了解数字化图书或分类文献的利用情况，基于此，分析读者的阅读倾向以评估学科布局和发展的有效性。通过对图书馆数字化图书或分类文献的下载情况进行统计分析，对促进图书馆数字图书或资料库建设，对优化学科及学科的有效发展，提升读者服务工作水平起到参考借鉴作用[1]。

1 系统方案设计思路

该系统设计的核心思想是，基于分类文献下载量数据统计特点，进行有针对性的设计及研发，且能在目前主流的操作系统（如在 Windows 系统）中便捷安装运行，并能直接利用目前主流文档管理系统（如 OFFICE 操作系统中的 EXCEL 表格）的数据进行统计特征分析。系统或软件的功能不仅能静态分析分类文献下载量数据分布特征，还能动态分析分类文献下载量变化趋势，具有一定的预测功能。通过该系统的应用，了解数字化图书或分类文献的利用情况及分类文献学科的被关注程度。据此，通过读者的阅读倾向分析评估学科布局和发展的有效性。

基于上述设计总体思路，我们进行了系统软件的研发。该项研发是在 Windows 系统下的 Matlab 平台上开展的，该软件可在 Windows 系统下运行，运行时所需内存较小，可以在现存的几乎所有的计算机平台运行。

2 系统基本架构及模块功能

系统主要由“学科年度下载量静态分析模块（含‘学科年度下载量’与‘重点年年度下载量’）”“学科下载量动态分析趋势模块”“下载量热力值”及“下载量谱系分析模块（聚类分析）”五部分组成，具体如图 1 所示。本文案例以某大学（SKD）图书馆中国知网期刊全文库的文献下载量为数据基础。其中，大类学科 10 种，小类学科 168 种，大类学科分别为：A 为基础科学、B 工程科技Ⅰ辑、C 工程科技Ⅱ辑、D 农业科技、E 医药卫生科技、F 哲学与人文科学、G 社会科学Ⅰ辑、H 社会科学Ⅱ辑、I 信息科技、J 经济与管理科学[3]。

2.1 学科年度下载量静态分析模块

在软件主界面，如图 1 所示，单击“学科年度下载量”，在下拉菜单中，选择年份（本案例加载了 2011—2021 年不同分类文献的下载量，该模块可以根据需要，随时加载新增年份或更早期年份的下载量），如 2016 年，软件会自动绘制出该年度不同分类文献的下载量分布图。也可以选择多个年份，并叠加显示，如图 2 所示为 2011 年、2015 年、2019 年、2021 年不同分类文献的下载量统计图像。在此过程中，可以根据实际需要，重点显示某些学科的索引号，如在本案例中，共显示了 42 个小类学科的文献索引号，可以看到，B021 学科的下载量要明显高于其他学科的下载量。在实际操作过程中，可以根据需求，调整显示的文献索引号，以达到进行不同或相近学科之间分析比较的目的。此外，还可以通过调整下载量的取值范围，以显示文献下载量在某一借阅量范围之间的学科，便于进行学科下载量的统计分析。

图 1 软件主界面

Figure 1 Main interface of the software

https://doi.org/10.35534/pss.0406041

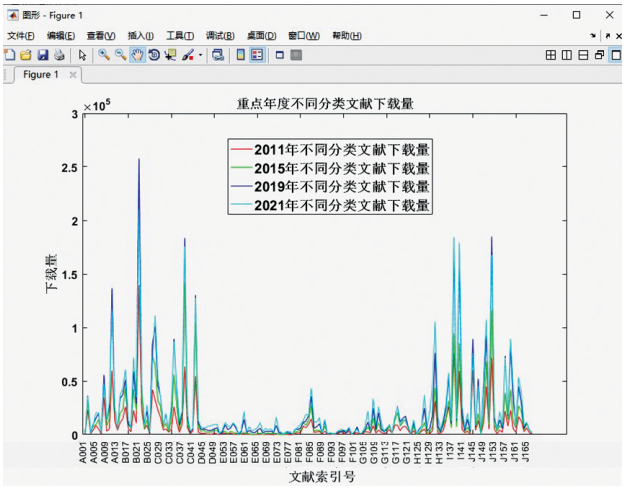

图 2　不同分类文献下载量对比

Figure 2　Comparison of downloads of different classified literature

2.2　学科下载量动态分析趋势模块

单击如图 1 所示的“重点学科下载量”模块，选择不同学科，并叠加显示。如图 3 所示为 A011、B021、C038、I138、I140、J152 学科 2011—2021 年度下载量变化趋势图。依此，不仅可直观观察到某学科的下载量的变化趋势，还可以同时观察不同学科下载量的变化趋势对比情况。可以看到，在此过程中，B021、C038、I138、I140、J152 五类学科的变化趋势中有交叉现象，从图中可以看到大体的趋势分布。如果想得到更为精细的分析数据（如在 2012 年时 C038、I138、I140 三类学科的下载量是否一致）比较困难，此时，可以通过调整下载量范围的大小进行局部放大，即通过缩小下载量的范围，突出不同学科下载量的差异，从而达到数据精细分析的目的。

2.3　下载量热力值图

“下载量热力值图”或称为“下载量变化趋势图”，用于展示多学科分类文献的年度下载量变化态势。可以是所有学科（如图 3 所示），也可以是部分学科（如相邻学科）。在此过程中，可以通过调整图书类别，以对不同学科下载量数据进行分析。如图 4 中间区域学科相较于上下区域学科的热力值（即变化趋势）不明显，

可以通过对图书类别进行调整，仅对上下区域学科的热力值进行比较，以获得近些年下载量有明显增加的学科数据特征。此外，还可以通过调整热力图的取值范围，以获得下载量大于（或小于）某一值的学科，便于对数据进一步分析。

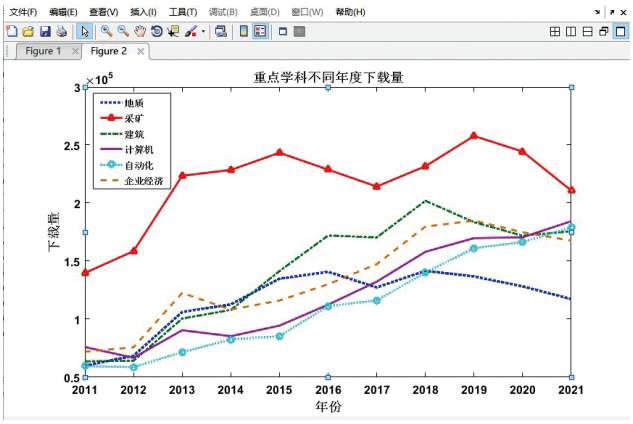

图 3　分类学科下载量变化趋势对比图

Figure 3　Comparison of change trend of downloading amount of classified disciplines

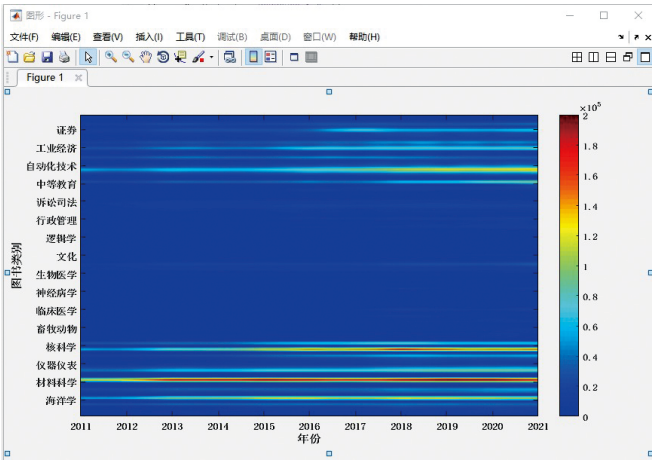

图 4　分类文献下载量变化趋势对比图（热力值平面图）

Figure 4　Trend chart of downloads of classified literature

2.4　下载量谱系分析模块

下载量谱系分析是通过聚类分析实现的。聚类分析是一种探索性分析方法，与判别分析不同，聚类分析事先并不知道分类的标准，甚至不知道应该分成几类，而

https://doi.org/10.35534/pss.0406041

是根据样本数据的特征自动进行分类。作为一种无监督学习方法，聚类分析被广泛地应用于数据的统计和分析。本系统采用系统聚类方法进行数据的分析。步骤如下：

（1）首先，输入图书的类别（本案例为图书大类），共 10 种。因此，在聚类分析开始时，每类图书自己划分为一类，共划分为十类，即每种图书和其类别是一一对应，从而将所有的研究数据进行划分成相应类别。

（2）利用几何数学方法计算任意两类之间的相似性（距离），以距离最近为准则，将最接近的两类重新归为一类。

（3）重复上述步骤（2）的过程，直至分类完成，获得最终的分类结果。

利用此原理，点击图 1 中“聚类分析”模块获得如图 5 所示的聚类树形图（下载量谱系图），其所展示的为 10 大类不同学科总下载量聚类分析系谱图。用于分析各大类学科总下载量的相关度。可以看到，在该图中，系统聚类分析自动对所有学科之间的关系进行调整，将数据特征相近的学科归位一类。并且，根据需求可以将学科分为不同的大类，如分为两类时，将 6、8、7、4、5、1、9 分为一类；2、10、3 分为一类。此外，不仅可以对大类学科进行聚类分析，也可以对小类学科进行分析，此时，仅需将大类学科数据替换为小类学科数据，无须进行其他修改，即可进行聚类分析过程。相较于监督学习算法来说，当修改为小类学科数据时，无须进行参数调整、模型训练等操作，因此聚类系数分析方法操作简单，是一种高效、快捷的数据分析方法。

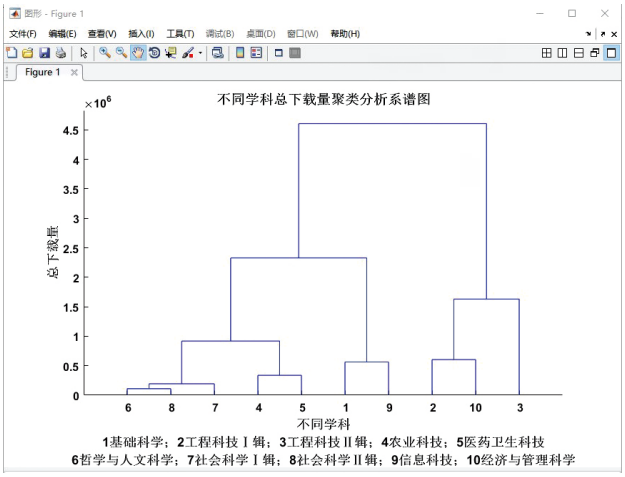

图 5　不同大类学科总下载量聚类分析谱系图

Figure 5　Cluster analysis pedigree of total downloads of different major disciplines

3 系统综合应用评价

本部分将“系统基本架构及模块功能”应用于实际案例，从“分类文献下载量分布特征分析”及“读者阅读习惯与学科发展态势分析”两个方面进行讨论，以评估所设计系统的有效性。

3.1 分类文献下载量分布特征分析

从图 2 和图 4 不难看出，本案例中（2011—2021 年）下载量最大和较大的区域主要在前区和后区，前者主要是 A、B、C 三个区，分别对应的是基础科学和工程科技（Ⅰ、Ⅱ），后者主要对应的是 I 与 J 两个区，分别对应信息科技及经济与管理科学。在前区中尤以 B 区中的矿业工程、C 区中的建筑及 A 区中的地质学最为显著。在后区中主要以 I 区中的计算机、自动化及 J 区中企业经济为突出。如图 5 所示，下载量谱系图也有良好的对应关系。谱系图中的 2（B 区工程科技Ⅰ）、10（I 区计算机、自动化）与 3（C 区工程科技（Ⅱ）），作为一类区，也是下载量主要贡献区。

3.2 读者阅读习惯与学科发展态势分析

图 2、图 4 及图 5 展示了分类文献下载量的基本分布特征。那么他们之间的变化趋势如何？图 3 是对如上所分析中下载量最大与较大区的一定年段内的变化趋势的对比。下载量最大的 B 区中的矿业工程呈现出下载量逐年增加到 2015 年出现拐点，2017 年后上升，到 2019 年又出现下降的拐点，A 区中的地质学的拐点出现在 2016 年，而 C 区中的建筑出现下降拐点是在 2018 年。J 区中企业经济的拐点出现在 2019 年，而 I 区中的计算机与自动化一路呈上升趋势。反映了不同学科不同时期其热度存在此起彼伏的现象。

3.3 学科发展综合分析

通过此系统对近十年分类文献下载量数据特征进行统计和分析，可以发现不同学科近些年的发展规律和特点，以及读者们的阅读倾向，从而有针对性地采取有效措施，提高文献的利用率，并合理地调整策略以更好地满足读者的阅读和科研需求。

该系统通过不同的技术方法较好地反映了一定时间范围内不同分析文献下

https://doi.org/10.35534/pss.0406041

载量的数据特征，通过不同文献下载量变化趋势，可以反映读者的阅读需求。在对实际数据进行分析时，可以以该系统为基础，通过对不同文献的下载量统计常态化，不断改进数据分析的技术方法，实时更新数据信息。通过对各种指标进行分析研究，以便准确把握不同学科最新的下载量趋势，获得读者的阅读倾向。针对各种数据变化特点，灵活地调整数据库的采购策略[4]。此外，不仅要考虑读者的阅读习惯，还应不断关注学校的学科建设情况和科研发展动向，及时提供相应的文献资源，保障学科建设和科研发展的顺利进行。为更好地满足教学科研的实际需求，图书及数据库的订购也可采取荐购的做法[5]，针对一些小类学科，对于多于这些针对性较强的学科文献，可将书目信息提供给相关专业的研究人员和读者，由其选择推荐购买的书目。

4　结语

根据分类文献下载量数据统计特点，所设计及研发的目的性明确、专业性更强的分类文献下载量数据统计特征分析系统，其所具有的定量化与多样性分析作用提高了分类文献下载量数据统计分析的科学性。通过对图书馆数字化图书或分类文献的下载情况进行统计分析，能更好地了解读者阅读规律和下载特点，对促进数字图书馆或资料库建设，对优化学科及学科的有效发展，提供了科学量化管理依据，为充分发挥图书馆的教育职能作用提供借鉴[1]。

参考文献

[1] 任丽丽．馆藏图书借阅量统计分析：以浙江警察学院图书馆为例[J]．电子世界，2013（4）：84–85.

[2] 魏立杰，林年添，丁仁伟，等．读者线下阅读行为变迁对大学隐性教育教学的启示[J]．科技视界，2021（5）：90–91.

[3] 杨晓萍．2012中文专业数据库检索：中国知网期刊全文[EB/OL]．[2022–11–20]．中国知网ppt课件，http://www.doc88.com/p%2D2344903511470.html.

[4] 周珊．高校图书馆电子图书使用评价与分析：以海南大学图书馆为例[J]．图书情报导刊，2016，1（10）：57–61，77.

［5］王荣宗，钟克理，隋晶晶，等．大学生图书借阅档案数据分析：以中国石油大学（华东）为例［J］．江苏科技信息，2022，39（8）：7–9.

Design and Application of Statistical Characteristic Analysis System for Downloads of Classified Literature

Wei Lijie[1] Liu Yinming[1] Sun Jiangwen[1]
Yang Jiuqiang[2] Lin Niantian[2]

1. Shandong University of science and Technology Library, Qingdao;
2. College of Earth Science and Engineering, Shandong University of Science and Technology, Qingdao

Abstract: Through the size and distribution characteristics of the downloads of classified literature, we can understand the readers' reading behavior and its relevance to the development of the discipline. In this paper, according to the statistical characteristics of the downloaded data of classified literature, we designed and developed a data statistical characteristics analysis system, and applied it to the actual case analysis. The application results show that the diversity analysis method improves the reliability of the statistical analysis of the download data of classified literature, helps to better and more comprehensively understand the readers' reading rules and download characteristics, provides scientific quantitative management basis for the construction of digital libraries or databases, and optimizes the effective development of disciplines and disciplines.

Key words: Downloads of classified literature; System design; Downloads heating value; Cluster analysis; Characteristics analysis

https://doi.org/10.35534/pss.0406041