Modern Management 现代管理, 2020, 10(6), 1134-1140
Published Online December 2020 in Hans. http://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2020.106137

Hans汉斯

关于开放存取OA类期刊的数据要素及趋势研究

王宁波，刘　峰*

无锡太湖学院，江苏省物联网应用技术重点建设实验室，江苏　无锡
Email: *lsttoy@163.com

收稿日期：2020年9月30日；录用日期：2020年12月24日；发布日期：2020年12月31日

摘　要

大数据时代的到来让人们可以通过技术手段对于海量数据进行挖掘和分析来获取和筛选信息。本研究以汉斯出版社2011~2019年之间发布全种类期刊文章作为研究对象，研究OA类期刊在学术期刊发文类别、研究视角变迁规律，并对以此挖掘文章的所属学科、关键词、引用等数据要素背后的知识。主要方法是通过时间序列分析方法，建立以时间、文章所属学科、关键词等文章重要元素之间的关系，来从通过大数据分析后的聚合数据得出OA类期刊逐渐被主流接受等结论。本研究通过对OA类期刊相关信息及发展趋势进行深度解析，推动了人们对于OA类期刊的理解。

关键词

大数据，开放存取OA期刊，数据挖掘，时间序列分析

Research on the Data Elements and Trends of Open Access OA Journals

Ningbo Wang, Feng Liu*

Key Construction Laboratory of Internet of Things Application Technology of Jiangsu Province, Wuxi Taihu University, Wuxi Jiangsu
Email: *lsttoy@163.com

Received: Sep. 30th, 2020; accepted: Dec. 24th, 2020; published: Dec. 31st, 2020

*通讯作者。

文章引用: 王宁波, 刘峰. 关于开放存取 OA 类期刊的数据要素及趋势研究[J]. 现代管理, 2020, 10(6): 1134-1140.
DOI: 10.12677/mm.2020.106137

Abstract

The advent of the era of big data allows people to mine and analyze massive amounts of data through technical means to obtain and filter information. This study took Hans Publishing House which published a full range of journal articles between 2011 and 2019 as the research object, studied the types of articles published by OA journals in academic journals, and the law of research perspectives, and explored the disciplines, keywords, and keywords of the articles and quoted the knowledge behind other data elements. The main method is to use time series analysis methods to establish the relationship between important elements of the article, such as time, the subject of the article, and keywords, to draw conclusions that OA journals are gradually being accepted by the mainstream from the aggregated data after big data analysis. This research promotes people's understanding of OA journals through in-depth analysis of relevant information and development trends of OA journals.

Keywords

Big Data, Open Access OA Journals, Data Mining, Time Series Analysis

1. 引言

21 世纪，是大数据的世纪。最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡，麦肯锡称：“数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素[1]。”随着时间的推移，数据无时无刻不在产生，而如此庞大的数据背后，人们迫切想要了解的宝贵数据可能却被埋没在幕布之下。正是因为这一需求，才推动知识挖掘技术的发展。如何从数据中提取各种有效信息，这一工作正在如火如荼地开展。

想要得出珍贵的数据就必须有好的数据源，本文选择的汉斯出版社，作为聚焦于国际开源中文期刊的出版发行机构，覆盖以下领域：数学物理、生命科学、化学材料、地球环境、医药卫生、工程技术、信息通讯、人文社科、经济管理等。目前汉斯出版社的所有期刊均被知网等数据库收录。其中有 23 本被美国《化学文摘 Chemical Abstracts》收录，30 本被 EBSCO 收录[2]。

本文从汉斯出版社发布的核心文章出发，探究近九年其发布各学科文章的数量、质量的变化，学者学术研究领域重心的转移方向，及文章关键词的演变趋势。在此基础上进一步分析大数据时代对生活的影响，并给出数据挖掘和有关政策性建议。

本文余下部分结构如下：第二部分是文献综述；第三部分将介绍研究过程的设计以及数据的整理；第四部分是建立时间序列以及图集来研究汉斯出版社发布文章的内在规律；第五部分是相关讨论和建议；第六部分做出结论。

2. 文献综述

当前，伴随着互联网信息技术的不断发展进步，大数据技术逐渐广泛应用于社会生活的各个领域，人们生活的各个方面都可以使用数据信息的方式来体现出来，让你们越发的离不开数据挖掘与分析技术

的发展[3]。“从大数据中发掘大洞察”等理念意味着对数据分析提出了新的、更高的要求，简言之，大数据时代就是数据分析的时代[4]。目前，数据挖掘技术正以前所未有的速度发展，并且扩大着用户群体，在未来越来越激烈的市场竞争中，拥有数据挖掘技术必将比别人获得更快速的反应，赢得更多的商业机会[5]。例如，早在 08 年等人曹毅，罗新星在《电子商务推荐系统关键技术研究》一文中通过数据挖掘和分析、机器学习等知识为恰当的用户在恰当的时间方便快捷地提供恰当的信息，并能根据用户的兴趣爱好自动地推荐给每个用户可能感兴趣且满意的商品[6]。在医药领域，数据分析的应用也变得更加多样。用例如可以运用在医药产品应用上，陈凤仪在 2010 年对番禺区部分医院 2007~2009 年口服降糖药物的应用做了研究，作者利用排序法和用药频度排序法，对番禺地区 2 家区级医院 2007~2009 年口服降糖药物应用数据进行统计挖掘，从而得出相关的现状分析及发展趋势，结论 a-糖苷酶抑制剂采购金额虽连续 3 年位居首位，但用药频度均小于双胍类及磺酰脲类，双胍类 DDDs 值连续 3 年位居首位，反映临床对该类药品的选择倾向较大[7]。此外，还有自然科研，分析了 100 多年来，《自然》发布的 8.8 万篇论文。研究发现，所有学科都出现了学科交叉性的增长，且没有放缓的迹象。随着研究人群、科研论文以及知识的增加，不同学科会变得愈来愈融合。研究机构以及资助单位应该意识到，学科交叉正在成为主流[8]。

综合现有研究结果，随着大数据时代的到来，移动互联网、物联网、交际网络、电子商务等新型信息技术的运用层出不穷。同时这些应用的推进使得各种数据正在迅速膨胀，不停地产生大数据。面向大数据市场的新技术、新服务、新业务等也在不断涌现。企业信息化建设以及对于大数据的利用将会成为企业提高核心竞争力的关键元素。各行各业的重要决定正在从业务驱动转向为大数据驱动，利用大数据来分析人们的喜好，消费趋势等，可以帮助商家、企业实时掌握市场变化并迅速做出决策反应。在医疗方面，通过大数据分析，可以不断提高问诊准确性以及测试药物有效性。在公共事业领域，大数据也开始发挥促进经济发展、维护社会稳定等方面的作用。以上研究主要集中在电子商务、医药等领域，但是对于 OA 类期刊分析研究案例较少，已有的也是对文章整体进行分析，往往忽略了文章中的重要数据要素。本文基于上述案例中的数据挖掘和分析技术，以汉斯出版社发布的文章为例，参考 5 万多篇论文，分析研究每篇文章的重要数据要素，得出结论并对公开存取 OA 类期刊的发展提出意见。

3. 研究设计与数据整理

(一) 研究方法

如图 1，本文的主要研究流程，首先是数据的爬取，接着对源数据进行处理，通过计数统计和词云分析来对数据进行提炼和分析，最后总结归纳得出结论。

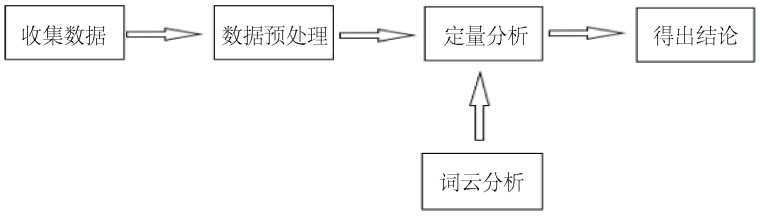

Figure 1. Research analysis flow chart
图 1. 研究分析流程图

(二) 数据收集和处理

本文选取 2011 年 4 月~2019 年 12 月汉斯出版社发布的各学科文章作为样本，使用 xpath 解析网页，在网页上找到所需数据的位置，然后利用爬虫技术进行爬取，最后将原始数据存入 csv 等待下一步操作。在数据获取时也出现一些问题，当 xpath 解析的路径不对时，这时要根据网页代码标签一层层找到数据

的位置，保证 xapth 路径准确无误。

下面对原始数据进行预处理，笔者使用的是比较常用且功能强大的 numpy 和 pandas，两者结合数据处理起来特别简单快捷。由于种种原因，爬取的数据可能会不够完整，或会出现缺失值，这是一种较为常见的现象。关键是如何处理？本研究爬取的数据多为中文，利用常规的填补方法是不可行的，所以是利用数据的权重不同，处理方法也不甚相同，若权重过高且缺失严重，即删除这一整条数据，若权重不高，即立一个 flag 进行标记注明此处有缺失值现象。

4. 数据分析及结果

(一) 文章学科分析

通过堆栈图(图 2)，对 2011~2019 年间汉斯出版社发布各学科文章的数量及其变化进行分析。总体来看，汉斯出版社每月发布文章数量在是不断增加，其中地球与环境、数学与物理这两门学科在同一时间发表的论文中占比增加明显。从月份上分析，一个有趣的现象是，汉斯出版社每隔两个月即达到一次当年发布文章数量的顶峰。以 2017 年作为一条分水岭，在此之前几乎每月发布文章数都不超过 300 篇，而 2017 年之后几乎每月都超过了 300 篇。

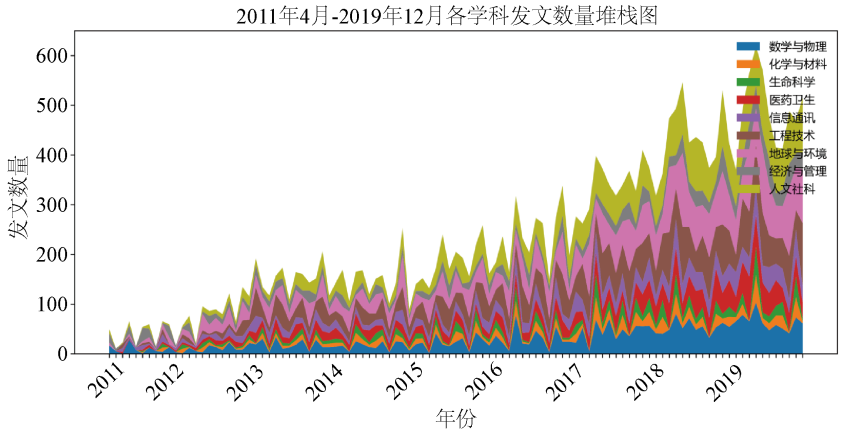

Figure 2. Stack chart of the number of articles published by Hans Publishing House in various disciplines from 2011 to 2019
图 2. 2011~2019 年汉斯出版社各学科发布文章数量堆栈图

(二) 文章引用分析

这些文章包含近 36 万条参考文献。为了准确识别论文所属学科，我们采用了汉斯出版社的分类信息，即使分类较为粗略，但通过这些庞大的数据依然可以得出一定结论。

结合图 2 和图 3 整体分析，地球与环境的引用占比占据所有学科的三分之一，当然这也和它发文数量成正比。相比之下数学与物理这门学科的文献研究更为有趣了。通过分析，可以得出数学与物理这门学科在发文数量上相比其他学科发文较多，而且每年都在增长，发文数量仅仅次于地球与环境，但文章引用在所有学科占比中却逐年下降。这里推测可能与数学与物理这门学科更加注重数学运算、推导公式有关，因此对于问题的探究与解决，相比其他学科没那么多。

(三) 词云分析

图 4 为汉斯出版社近九年的关键词生成的词云图，首先利用 jieba 库将所有关键词分解合并分好词，再用 TF-IDF 算法对关键词进行提取，TF-IDF (Term Frequency-Inverse Document Frequency)是一种统计方法，用以评估一个词语对于文件集或语料库中的一份文件的重要程度。其原理为：一个词语在一篇文章

中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章[9]。我们选取权重排名前 60 的关键词，以此生成了下面的词云图。

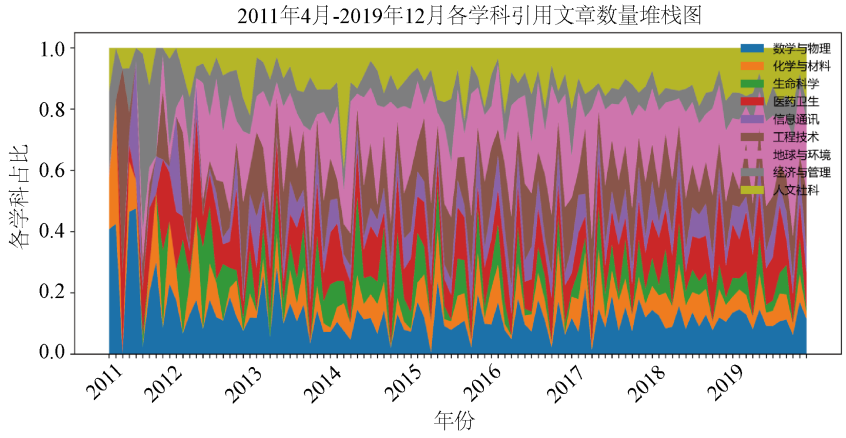

Figure 3. The flow chart of the proportions of articles published by Hans Publishing House from 2011 to 2019
图 3. 2011~2019 年汉斯出版社发布文章的各学科占比流动图

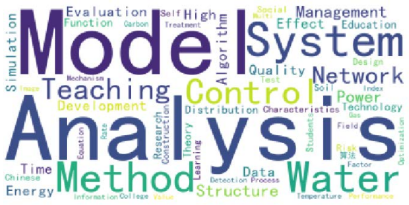

Figure 4. Word cloud diagram
图 4. 词云图

(四) 词频谱分析

根据 2011~2019 年发布文章中英文摘要关键词，进行整体词频输出得出如表 1。可以看出热门的词汇包括是“Analysis”分析、“Model”模型、“System”系统和“Method”方法。

Table 1. Ranking of main keywords of open access journals from 2011 to 2019
表 1. 2011~2019 年开放存取期刊的主要关键词排行

关键词	出现次数
Analysis	1527
Model	1480
System	1308
Method	1161
Water	775
Control	729
Network	659
Management	563
Teaching	546
Power	519

同时结合时间序列特性，根据关键词的权重得出了以下折线图(图 5)所示，可充分显示科研学者的主要研究方式方法的转变。

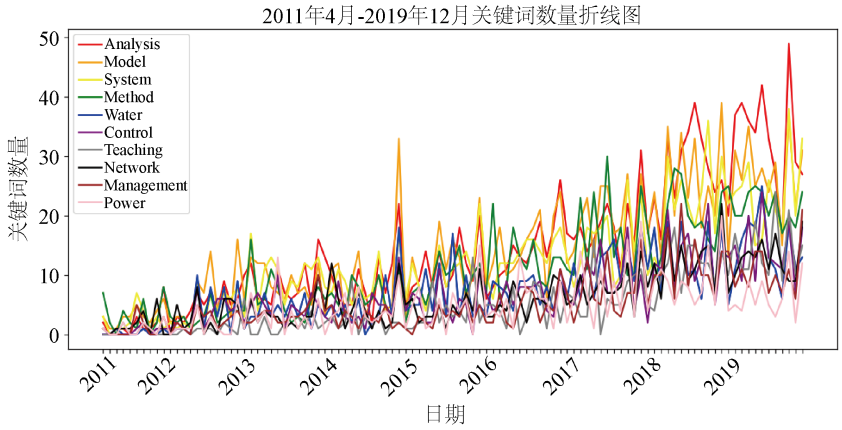

Figure 5. Line chart of the number of keywords with fusion weight
图 5. 融合权重的关键词数量折线图

2011 年学者主要关注在模型层面，随着计算机的迅速发展，模型的应用领域也在发生着革命性变化，模型已经向一切领域渗透，各行各业日益依赖于建模。同时在新设备、新技术的研制与开发中，对于质量、精度、速度、效益等指标的考查无一不需要在模型下运用相应的方法和借助计算机控制实现。根据词云图中 2019 年的数据也不难看出经过九年的发展，“模型”这一关键词依然排在前几名，可见其重要性。“分析”一词在 19 年大火起来也不奇怪，遇事离不开分析，在探索的路上分析清楚，也就离正确答案更近一步。

通过上述相关数据分析可以得出以下三点结论：

1、各学科文章的发布数量都在逐年上升，且各学科的发文数量逐渐走向均衡。

2、在各学科当中，地球与环境的引文占比居多，而数学与物理虽然发文数量占比较多，但引文数量却逐年下降。

3、对于开放存取的期刊，投稿学者的研究方法普遍倾向于通过建模进行分析研究，来取代了原先固定学科方法的研究模式。

根据以上三点结论，对公开存取 OA 类期刊的发展有如下建议：

1、开放存取 OA 类期刊的接受度越来越高，传统的期刊杂志社需要考虑自身发展的多样化，接受趋势并发展出自己的 OA 类期刊。

2、投稿 OA 类期刊的学者更加青睐于社科和应用类研究，而基础学科和理工科的研究相对较薄弱。因此开展了 OA 类业务的期刊杂志社需注意提高对社科和应用类研究论文的发文门槛，避免成为“水刊”，同时需要鼓励学者们对基础研究的投稿。

3、鼓励学者多样化研究方式方法，注重结果的完整性，鼓励实验数据的提交，打造优质开放存取期刊样板。

5. 总结

大数据时代下，最核心的就是对大数据进行研究分析。只有通过调查研究的数据，才能对于出现的问题提供有效准确的解决办法。本文通过研究发现汉斯出版社发布文章学科逐渐多元化，各学科文章数

量不断增加，近几年的文章研究热点在 Analysis 和 Model 等方面。因此得出结论：大数据技术为我们提供了全新的视角来分析和审视开放存取的期刊相关数据要素，以数据驱动的方式进行分析进而挖掘出相关知识，取代过去凭借经验和直觉的方式方法，后者往往带有人们的主观性，对于问题判断不够客观。

我们认为研究一篇论文包含哪些数据要素，可能有助于学科之间的比较，也有助于增加对论文影响力的正确性。作为数据分析的研究员，我们希望学科之间不再那么封闭。为分析学科之间的交叉性以及开放存取期刊的影响和驱动的关键因素，下一步工作则会围绕这些数据要素在学科之间相关性进行研究，挖掘交叉领域中文章自身的关注知识点是否有发生迁移，迁移过程中在哪些点聚变成新的知识点、知识面甚至知识领域，从而以更深层次的视角来研究这些数据要素对开放存取期刊的影响和驱动。

基金项目

本课题得到江苏省物联网应用技术重点建设实验室资助。

参考文献

[1] 官建文, 刘扬, 刘振兴. 大数据时代对于传媒业意味着什么?[J]. 新闻战线, 2013(2): 18-22.

[2] 汉斯出版社官网. https://www.hanspub.org/AboutUs/Index.aspx

[3] 曾怡. 大数据时代下数据分析理念的辨析[J]. 科技展望, 2016(34): 1-8.

[4] 李广建, 杨林. 大数据视角下的情报研究与情报研究技术[J]. 图书与情报, 2012(6): 1-8.

[5] 黄翔. 浅谈数据分析在电子商务中的应用[J]. 商情, 2010(13): 90-90.

[6] 曹毅, 罗新星. 电子商务推荐系统关键技术研究[J]. 湘南学院学报, 2008(5): 63-66.

[7] 陈凤仪. 2007-2009 年番禺区部分医院使用口服降糖药数据分析[J]. 广州医药, 2010, 41(2): 49-52.

[8] Gates, A.J., Ke, Q., Varol, O., *et al.* (2019) Nature's Reach: Narrow Work Has Broad Impact.

[9] 路永和, 李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作, 2013(3): 90-95.