搜索引擎的学术应用对图书馆数据服务的启示

罗晓兰
（上海中医药大学图书馆，上海 201203）

摘要：开放数据是科研重要的数据来源，但在目前的科研数据开放共享中却被忽视。本文以科研中常用的搜索引擎数据为例，分析科研成果对开放数据的使用情况。从中国知网和万方数据库获取国内以谷歌和百度产品为研究数据来源的期刊论文（4 212篇）作为研究数据，通过人工标引和词频统计的方式，对国内科研论文中对搜索引擎产品数据使用情况进行统计，分析搜索引擎数据的使用特点、研究领域和发展趋势，为图书馆制定合理的科研开放数据服务政策、构建合适的科研数据服务模式提出建议。

关键词：学术研究；搜索引擎；开放数据；图书馆
中图分类号：G252.7
DOI：10.3772/j.issn.1673-2286.2017.08.007

1 研究背景

开放数据在政策、经济和社会发展条件的不断促进下，已具备深厚的数据积累。数据开放要求政府重视数据的搜集、公布、开放和运用[1]，促使政府自身或大众能较为便利地获取和再利用这些信息[2]。数据共享让大众生活更加便利和智能，也为科学研究提供海量数据资源。促进开放数据的利用有助于释放大数据的能量，以大数据为动力支持社会创新，以创新发展智能经济[3]。目前数据共享的研究主体主要是科研过程中产生的科学数据[4]，但免费公开获取的开放数据源常被忽视。

开放数据源涵盖在科学研究中利用的开放数据集、公共搜索和统计服务、开放平台和研究工具等，具有公开性、可得性、完整性、即时性等特点，主要由政府、科研机构、非政府组织、开放的商业平台、互联网应用程序等提供[5]。在各学科领域都存在具有行业特色的开放数据，这些数据被广泛使用到科研项目中。另外，还存在部分各学科通用数据，如搜索行为数据、在线社区文本、微博话题关注数据等。

在开放数据集中，来自于搜索引擎的数据是常用数据类型，贯穿科研工作的始终。搜索引擎提供的开放数据具有使用范围更广、适用性更强、可获得性更高等优势。尤其是在大数据科研理念下，基于用户搜索行为、在线交互行为、健康行为、学术行为、网络舆情等数据的数据量更大，更具有时效性。

作为重要的公开信息资源，搜索引擎的学术应用主要有两种形式：一是以搜索引擎产品提供的数据作为研究数据来源，二是以搜索引擎提供的平台或产品作为研究对象。其在科研中的应用优势主要包括数据可获得性高、数据覆盖范围广、实时性强、数据的认可度较高，还可通过实时、丰富的开放数据增加科研成果的创新程度。相关科研成果涉及舆情分析、反恐、人口统计和决策、金融投资、旅游服务、健康管理、教育教学等领域[6-7]。此外，搜索引擎也为科学研究提供许多高效实用的研究工具，如Google Earth、Google App Engine、百度云等。

在目前科研数据共享的趋势下，面对科研第四范式下科研人员对科学数据及其应用的需求变化，作为数据资源服务方的图书馆须思考如何有效提升开放数据资源建设的服务能力，开展科学数据服务[8-9]。但与出版商和期刊杂志社相比，图书馆并没有获取科研数据的先天优势条件，由此在科研过程中产生的科研数据共享进程也推进缓慢。

2 数据获取及分析

2.1 数据获取及预处理

在中国知网和万方期刊文献数据库中检索国内利用百度、谷歌提供的工具、数据、资源（如谷歌趋势、谷歌地图、百度指数）等进行学术研究的期刊论文。检索式为：篇名/题名 OR 文摘=百度 OR Baidu OR 谷歌 OR Google，搜索时间为2016年8月20—31日。通过题录信息进行初步筛选，保留有效题录，包括利用搜索引擎提供数据和工具进行研究、以搜索引擎产品为研究对象的论文，随后删除重复记录，形成统一格式的有效文本集。

对筛选过的文献题录进行人工标引，标引格式为“技术/产品/数据—年代”，如Google Earth—2016。产品和数据编码目录来源于文献筛选过程中对产品和数据类型的统计。

为保证标引质量，进行两组人工标引。人工标引的评分者信度系数为0.986（P<0.01，Sig双侧为0），说明两组标引者一致性很高，但仍存在二者标注不一致的情况。如对产品名称描述的不一致、对多种数据共同使用的标引缺失、标引中的错误等，随后根据论文内容进一步确认标引信息，形成一致的结果。

经过筛选后的期刊文献题录数据共4 212条，基于百度提供的数据进行研究的有1 121篇，基于谷歌数据的有3 091篇。从文献量看，2001—2013年，利用百度或谷歌提供的数据进行研究的成果处于上升态势。

2.2 搜索引擎在学术研究中应用的频次统计

搜索引擎积累了大量的搜索行为数据、语料资源、具有社交属性的用户自生成内容、搜索趋势和地理信息数据。搜索引擎不仅可为科研活动提供丰富的数据，还能将部分服务或产品二次开发嵌入新的科学研究，降低科研初始开发的时间成本和经济成本，为科研工作创造更多可能性。谷歌和百度在平台构建和开发工具方面为广大科研人员提供较大的扩展空间，如谷歌地球、谷歌地图和百度地图等为地理、地质、交通运输、航空航天、林业、畜牧业、农业、资源环境等学科提供基础资源数据，一般研究模式是在免费版的谷歌地球基础上二次开发，并与实际需求相结合进行个性化加工，实现功能扩展。

根据数据分析结果，基于谷歌的学术研究常用资源有谷歌地球、谷歌地图、谷歌搜索、谷歌数字图书馆、谷歌学术、谷歌安卓平台、谷歌云计算等；基于百度的学术研究常用资源有百度搜索、百度地图、百度指数、百度贴吧、百度文库、百度百科、百度知道等。

3 搜索引擎在学术研究中的应用分析及讨论

3.1 搜索引擎在学术研究中的应用统计

搜索引擎在学术研究中的应用主要有搜索工具和统计、地理信息系统、在线文档系统和百科类问答系统、社交互动平台、开发平台及其他专业类数据（见表1）。作为获取行为数据和查询文献信息的重要补充来源，搜索数据是研究中使用最多的数据来源，而地理信息系统在专业领域中使用最频繁。

（1）搜索应用。学术研究常用的搜索应用包括通用搜索、学术搜索、专业搜索、搜索指数和趋势统计，不同类型的搜索类产品在科研中的应用情况，如表2所示。

表 1 搜索引擎在学术研究中的应用统计

常用数据类别	常用数据细分	应用名称
搜索数据	通用搜索类、学术搜索类、专业搜索类	谷歌搜索、百度搜索、谷歌学术、百度学术、百度视频、百度拇指医生、百度图片
搜索趋势统计	搜索行为统计、搜索趋势分析、热点分析	百度指数、谷歌趋势、百度移动统计、百度舆情
地理信息系统	地理信息共享、地图信息查询	谷歌地球、谷歌地图、百度地图
在线文档系统和百科类问答系统	在线文档分享平台、问答系统、在线百科	百度文库、百度知道、百度百科、百度拇指医生
社交互动平台	在线论坛、基于兴趣的社区、知识社区	百度知道、百度贴吧
开发平台和工具	提供开发平台或技术支持、共享资源和协作学习	谷歌安卓平台、谷歌协作平台、百度众包平台、百度开放服务平台等
其他特色数据	大数据、多媒体资源、特色专业资源	百度大数据+、百度阅读、百度旅游、谷歌线上艺术博物馆等

表 2 搜索类产品在学术研究中的应用统计

产品类型	学术研究中的应用
通用搜索类	查新、名词释义、公开范围的搜索数据和搜索量、获取研究文本
学术搜索类	获取文献资源、获取文献引证评价数据
趋势统计类	特定名词或规定领域的搜索量、关注度、发展趋势
专业搜索类	针对特定研究领域进行数据搜集和研究，如新闻搜索、图片搜索、视频搜索、文本翻译、健康文本等

其中，在科研中常用的搜索趋势统计数据是搜索服务的衍生品，常用于经济学、行为学和信息科学研究，如百度指数和谷歌趋势。百度指数主要用于投资行为、旅游管理、流行病预测、产品关注度及变化趋势、票房预测、房地产、就业、舆情分析、空气质量、食品监管、城市发展研究、消费者信心指数、受众行为分析、出版发行等方面（见表3）。

表 3 百度指数在学术研究的应用统计

研究领域	成果数量/篇	主要研究内容
旅游管理	30	旅游景点关注度及时空特征分析、景点游客流量预测、搜索行为与景点关注度
经济学	23	投资行为、消费者信心指数、关注度与股票投资行为、关注度与收益
城市发展	13	城市网络联系、区域关注度差异
产品或搜索关注度	12	搜索趋势，产品、行为关注度和需求分析
医疗卫生	11	流行病监测或预测、突发公共卫生事件关注度及应对、医疗事件关注度、健康信息行为
房地产	5	搜索行为与房价、关注度与房价、成交量预测
编辑出版学	5	期刊影响力、核心期刊关注度

（2）地理信息系统。谷歌地球、谷歌地图、百度地图等是研究使用频率最高的专业数据来源，主要应用于地理、地质、测绘、交通运输、电力工程、地球物理学、电信、建筑工程、林业、水利水电、气象、资源环境以及教育培训等领域。研究模式主要有两种：一种是通过谷歌地球等软件的API及KML二次开发接口快速提取相关模型、数据和影像资料，实现工程设计的高度仿真，用于设计、施工、评估等多个阶段；另一种是通过前期测量获取数据，将设计方案通过谷歌地图或百度地图提供的二次开发接口进行加工呈现，使设计方案可视化并可通过移动客户端进行浏览和调用。

地理信息系统作为谷歌和百度在学术应用中数量最多的数据类型，存在严重“偏科”的特点。除测绘、地理信息、农业、林业、牧业、渔业等学科外，社会科学的很多创新研究也可借助地理信息系统开展，如将百度地图、谷歌地图与图书馆的读者服务、自助图书馆等结合起来，利用在线地图构建基于地理位置的用户服务和用户行为分析。

（3）在线文档系统和百科类问答系统。在线文档系统及百科类问答系统为科研提供大量持续增长的研究数据和文本资料，是文本分析研究的常用数据来源。在线文档系统是国内近年来发展较快但颇受争议的研究热点，主要用于文献获取，或作为课程教学资源数据库使用[7]，最受关注的是版权问题[10]。百度百科和百度知道是常用的百科类问答系统，其产生的大量文本资源是很有价值的科研数据来源，主要用于问答系统的运作模式和用户行为[11]、网络文本处理方法[12]、在线知识分享行为和模式[13-14]、网络信息行为及信息扩散研究[15]、在线问答系统的文本质量[16]、语义分析和文本挖掘[17-18]等研究。从研究成果发布时间看，这类系统研究属于比较新的研究领域，还有继续深入研究的价值。

（4）社交互动数据。使用社交互动平台数据的研究主要集中在图书馆学、情报学、计算机科学、新闻传播学和教育学。百度贴吧、百度知道等根据用户需求对行业和学科类别进行细分，形成数个聚集大量用户且具有社交互动性质的平台。在国内期刊论文中使用百度贴吧的情况更多，主要用于研究电视节目关注与传播[19]、网络群体管理[20]、网络语言传播[21]、在线互动行为[22-23]、网络互动文化[24]等。百度知道数据主要用于研究在线互动平台机制[25-26]、基于社交网络的信息

和知识传播[27-28]等。由于这些产品的社交属性不如微信、微博等社交媒体，因而并不是社交网络研究的主流数据来源，研究领域较狭窄，数据被挖掘的程度不高。智能移动终端普及以后，移动终端使传统论坛和互动问答平台的社交属性更强，便于继续追踪基于平台APP的移动使用行为数据，开展研究工作。

（5）开发平台和工具。搜索引擎提供开源的平台工具为科研和教学提供帮助，这些平台和工具大致可为两类，一是提供开发平台或技术支持，二是共享资源和协作学习。如谷歌协作平台是侧重于团队协作的网站编辑工具，可帮助企业创建企业内网、进行项目管理跟踪等，用户通过谷歌协作平台将所有类型的资源（包括文档、视频、图片、日历等）与好友、团队或网络分享[29]。基于此特性，谷歌协作平台被应用到“云计算辅助教学”实践，并取得较好成效[30-31]。谷歌云计算开发平台（Google App Engine）同样在应用软件开发、教学课程资源库建设、自主学习、协作学习、数字图书馆建设等领域中使用[32-33]。同样，百度推出的同类产品（百度云、百度众包平台、百度开放服务平台等）也在科研中得到应用[34-35]。

除以上列举的主要应用形式，在大数据决策的发展趋势下，搜索引擎提供的相关数据平台和服务（百度大数据+）、专业信息服务（百度健康、百度阅读、百度旅游）、在线特色服务（谷歌线上艺术博物馆）、在线特色小工具等都可成为可用的科研开放数据源，但目前鲜有成果出现。

3.2 搜索引擎学术应用总结

基于搜索引擎的开放数据价值得到学界普遍认可，但目前对开放数据的应用程度还不够深入，从目前国内科学研究对搜索引擎数据的使用中发现以下两个问题。

（1）数据应用层次太浅，只有部分数据被有效利用。如利用谷歌和百度地理信息系统的相关研究约1 700条，占总体研究数量的40%，而60%中像百度拇指医生、百度阅读、谷歌眼镜、谷歌线上虚拟博物馆等未被开发利用的数据就有可能存在科研的创新点。未来基于不同平台、不同场景的数据融合，将是科研数据利用的发展趋势。

（2）在学术研究中频繁使用的数据存在偏好，基于同类型数据的研究设计、方法和过程雷同，科研创新性不高，如何让有价值的数据在科研实践中发挥更大的作用，在研究模式、方法和切入点上值得学者进一步探索。

4 讨论及未来研究方向

4.1 搜索引擎的学术应用对图书馆数据服务的启示

从搜索引擎数据应用的案例分析可见，开放数据在各学科学术研究中应用广泛，是可利用的有效科研创新资源。但国内对开放数据的关注较少，成果的严重“偏科”也从侧面说明部分学科对开放数据的应用太少，开放数据的发现、抓取、整理、发布、利用和评估整个流程缺乏规范和指导。而资源获取、整合和推广等工作是图书馆的优势，因此可基于科研用户对开放数据的使用特点，在图书馆进行馆藏资源建设、特色数据库开发和服务，以及在资源整合过程中吸取有用经验，提供更人性化和个性化的资源服务。

（1）构建开放数据资源目录，做好开放数据资源发现和导航工作，帮助用户获取更多开放数据用于科学研究，提升科研创新性。现阶段科研数据共享工作提倡用户公开共享科研活动产生的数据，但不能忽视开放数据的应用。图书馆应抓住大数据分析和决策的发展趋势，更多应用公开、易获得的原始数据，为学术研究和管理决策提供资源支持。图书馆虽然不拥有科研成果数据的版权，但可根据用户学科背景和需求，对公开研究数据集进行搜集和加工，整理开放数据资源目录，为用户提供免费数据参考咨询和数据推荐服务。目前全球范围内有价值的开放数据集数量巨大，用户在研究中所使用的占比较少，大部分有价值的开放数据还未被利用。此外，在开放数据主题新颖性、研究模式的创新等方面有待加强。

（2）开展学科数据服务，做好开放数据获取的辅助工作。目前科研人员处于信息和数据海量增长的时代，图书馆可通过开展学科数据服务、嵌入式科研辅助服务等形式，帮助科研人员搜索可利用的开放数据、制定数据获取方案、寻求最佳获取途径；此外，还可提供存储空间和运行设备，评估数据价值，辅助数据分析工作，实现科研贡献和学术共享空间的职能。

图书馆的信息资源、用户行为数据、空间数据等也颇具研究价值，可开放给用户进行科学研究。如上海

图书馆利用整理的家谱数据开展开放数据应用开发竞赛，力求更充分地释放开放数据的价值，最大程度挖掘其背后的应用潜力，激发创新能力，这种双赢互动的形式值得推广。

（3）建立开放数据获取平台，整合学科资源、知识和数据，加强不同层次资源和数据的关联性。科研工作需要文献、知识、数据和工具等多种资源，但大部分资源分散在不同机构和平台，并未进行整合。图书馆可利用其在文献资源整合管理方面的经验和优势，将用户所需科研资源、知识、数据和工具进行整理和发布，利用关联技术建立不同类型资源的联系，构建开放数据获取平台，方便科研用户使用；还可提供数据使用情况分析报告和研究进展供科研人员参考，通过资源和服务整合提升科研成果创新。

（4）做好对科研人员的数据素养培训。从搜索引擎数据在研究中使用可以发现，学科、研究主题间差异明显，部分学科并没有利用可公开获取的庞大数据进行科研活动。除学科特点有所不同，科研数据素养是限制国内科研人员充分使用开放数据的制约条件之一，图书馆在进行数据资源服务过程中可向用户提供相关培训，包括对相关研究设计、数据抓取、工具使用、数据可视化等进行辅导，结合图书馆提供的数据资源服务，针对基于数据研究的发展趋势，组织数据分析培训课程，更好地辅助教学科研。

4.2 本文不足之处和未来研究方向

公开的科研数据将有望成为图书馆在数据服务阶段进行深度挖掘和整合的服务资源，在学术研究中还有极大的价值等待挖掘。本文在样本数据选取时只选取搜索引擎开放数据，不够全面。在下一步研究中应将国内学术研究中常用的搜索引擎服务、社交网络数据、政府机构和非营利机构提供数据等的利用情况纳入研究范围，如豆瓣、微博、微信和行业特色数据源集等。然后，与国外研究情况进行对比分析，获取科研用户使用行为特点和需求，为国内基于开放数据的科学研究事业提供宝贵经验，为图书馆构建开放数据资源服务目录和服务体系提供支持。

参考文献

[1] 张毅菁.从信息公开到数据开放的全球实践——兼对上海建设“政府数据服务网”的启示[J].情报杂志,2014(10):175-178,183.

[2] 陈美.美国开放政府数据的保障机制研究[J].情报杂志,2013(7):148-153.

[3] BERTOT J C,郑磊,徐慧娜,等.大数据与开放数据的政策框架:问题、政策与建议[J].电子政务,2014(1):6-14.

[4] 刘晶晶,马建华.论科研数据开放共享的三种途径[J].情报杂志,2015(10):146-150,96.

[5] 毕秋灵.数据新闻中的开放数据应用[J].湖北社会科学,2016(7):190-194.

[6] 杨滨.论云计算辅助教学(CCAI)中协作学习产生的设计机制——以Google sites下的协作学习为例[J].现代教育技术,2009(11):95-99.

[7] 王玉龙.基于百度文库的微课资源社区构建策略研究[J].中国远程教育,2015(2):73-78.

[8] 黄金霞,马雨萌.大数据时代开放信息资源的数据服务能力思考[J].数字图书馆论坛,2016(8):54-59.

[9] 陈建新.科学数据服务:图书馆服务的新领域[J].图书与情报,2013(4):93-95.

[10] 张丽波,马海群,周丽霞.避风港原则适用性研究及立法建议——由百度文库侵权案件说起[J].图书情报知识,2013(1):122-127.

[11] 常静,杨建梅,欧瑞秋.基于TAM的百度百科用户参与意向的影响因素研究[J].软科学,2010(12):34-37.

[12] 陆勇,章成志,侯汉清.基于百科资源的多策略中文同义词自动抽取研究[J].中国图书馆学报,2010(1):56-62.

[13] 夏火松,王瑞新.百度百科词条特性对知识共享意愿影响的实证研究[J].科学学研究,2010(12):1877-1883,1890.

[14] 黄令贺,朱庆华,沈超.差异与稳定:网络百科用户兴趣动态变化研究[J].图书情报知识,2016(2):101-113.

[15] 张洋,卢桥.中文社会化媒体信息老化的计量分析[J].情报杂志,2015(3):77-84.

[16] 孙晓宁,赵宇翔,朱庆华.基于SQA系统的社会化搜索答案质量评价指标构建[J].中国图书馆学报,2015(4):65-82.

[17] 许坤,冯岩松,赵东岩,等.面向知识库的中文自然语言问句的语义理解[J].北京大学学报(自然科学版),2014(1):85-92.

[18] 段利国,陈俊杰.综合句法结构及语义相似度的问题推荐技术[J].计算机科学,2012(1):203-206.

[19] 张倩,戴建华,闫萌萌.基于电视剧网络点播量分析的社会化媒体价值研究[J].现代传播(中国传媒大学学报),2013(11):59-62.

[20] 张郁文.浅析贴吧粉丝群体的管理——以“罗志祥吧”为例[J].新闻世界,2014(8):154-156.

[21] 贺洁.从大众传媒看“土豪”的传播与发展[J].青年记者,2014(24):86-87.

[22] 万力勇.网络百科用户协同创作的互动机制研究——以百度百科贴吧为例[J].情报杂志,2014(1):167-172.

[23] 王国华,刘菊,杨腾飞,等.网络空间中艾滋病的社会支持研究——以百度贴吧“HIV吧”为例[J].情报杂志,2015(11):105-110.

[24] 李可安.新媒体传播方式下的粉丝文化——以新浪微博和百度贴

吧为例[J].科技传播,2015(12):92-93.

[25] 赵丽红.互动式知识问答分享平台对虚拟参考咨询服务的启示[J].图书馆建设,2009(5):62-64.

[26] 霍建梅,李书宁.图书馆数字馆藏建设用户参与激励机制探究[J].图书情报工作,2015(2):5-10.

[27] 宁寒松.线上互动系统中“舆论领袖”的缺失及成因——以百度知道为例[J].新闻世界,2012(8):119-120.

[28] 王小立.百度“知道”知识传播对个人数字图书馆资源共享的启示——基于系统动力学方法[J].图书馆,2016(2):83-87.

[29] 百度百科.Google Sites[2016-11-7].http://baike.baidu.com/link?url=osu8ZSzbSC_yozf1NaziwxabhN79UBmlyhleJvg1OW3jlKneQVkshAvO1Gah1uU5mI5n61Nu3t0YMsUUkxSQ1dldtesmhwsRiawt-xQ3vqu.

[30] 杨滨.论云计算辅助教学(CCAI)中协作学习产生的设计机制——以Google sites下的协作学习为例[J].现代教育技术,2009(11):95-99.

[31] 徐瑞.Google协作平台在中小学教师教育技术培训中的应用探究[D].上海:华东师范大学,2010.

[32] 刘晓刚.基于开源云计算的远程教育系统的设计与实现[J].中国教育信息化,2011(9):40-43.

[33] 王佳隽,吕智慧,吴杰,等.云计算技术发展分析及其应用探讨[J].计算机工程与设计,2010(20):4404-4409.

[34] 陈霞,闵华清,宋恒杰.众包平台作弊用户自动识别[J].计算机工程,2016(8):139-145,152.

[35] 丁峰,梅晓亮,张丽.专业群教学资源信息化面向移动APP题库的设计及实现[J].信息系统工程,2016(5):148-149,152.

作者简介

罗晓兰，女，1985年生，博士研究生，讲师，研究方向：信息检索、健康信息行为、技术采纳与行为，E-mail：miaoqu11@126.com。

The Academic Applications of Search Engine and Its Inspiration to Library Data Services

LUO XiaoLan
(Shanghai University of TCM Library, Shanghai 201203, China)

Abstract: Open data is an important source of data for scientific research, but it is neglected in the scientific data sharing system. This study takes the search engine as an example to analyze the usage of open data in scientific research. 4 212 items got from CNKI and Wanfang to analyze the Google and Baidu data use behavior by the way of manual indexing and word frequency statistics, including their characteristics, frequency and tendency. Based on this, the author made recommendations to develop open data service policy and service mode for library.

Keywords: Academic Research; Search Engine; Open Data; Library

（收稿日期：2017-04-24）