

上海市公共数据开放平台用户画像构建与分析

陆敬筠 吕海艳

(南京工业大学经济与管理学院, 南京 211816)

摘要: 为帮助政府数据开放平台更精确地了解用户访问行为从而精准满足用户需求, 更有针对性地开放数据, 提升服务水平, 本文以上海市公共数据开放平台用户访问数据为研究对象, 构建上海市公共数据开放平台用户画像, 分析用户行为特征。基于用户画像构建与行为分析, 提出上海市公共数据开放平台的优化策略。该研究也为其他地方政府数据开放平台的用户画像研究提供参考。

关键词: 政府数据; 开放平台; 用户画像; 访问行为; 精准需求

中图分类号: D63; G203 **DOI:** 10.3772/j.issn.1673-2286.2021.10.008

引文格式: 陆敬筠, 吕海艳. 上海市公共数据开放平台用户画像构建与分析[J]. 数字图书馆论坛, 2021 (10): 54-59.

大数据时代的到来使各类数据应用的领域越发广泛, 如经济、政治、社会、医疗、教育等, 引发了社会公众对数据开放的强烈需求。政府部门作为社会数据资源的主要掌握者, 将不涉及国家安全、商业机密和个人隐私的政府数据开放出来, 可以满足社会各领域对政府数据的使用需求, 这不仅能提高政务透明度与公众满意度, 还能通过数据利用者的挖掘、再利用充分发挥数据价值从而服务于社会公众, 为社会进步注入新的活力。

我国政府数据开放的研究起步较晚, 政府数据开放平台还存在不少亟待完善的问题, 如数据质量不高、数据覆盖的领域不完善, 对用户需求的满足还存在不足。国内学者也从多个角度对政府数据开放平台的建设进行了探讨, 但已有研究大多从指标确定^[1-3]、模型构建^[4-5]、实证分析^[6-7]等角度出发, 探讨我国政府数据开放平台的发展现状并根据分析结果提出相应的优化策略, 以提高我国政府数据开放水平。从用户视角出发的研究大多从用户质量感知^[8-9]、用户体验^[10-11]与用户服务评价^[12]的视角进行相关探讨, 基本没有从用户访问行为入手分析政府数据开放平台的用户需求。

用户是政府数据开放平台的实际体验者和数据使用者, 用户在访问政府数据开放平台的过程中会发生检索、查询、申请等行为, 通过对用户行为进行分析, 能够发现用户更加关注的数据领域、用户对其他数据的需求等, 从而让政府数据开放平台的数据开放更有针对性。由于上海市公共数据开放平台发展较早且较为成熟, 用户访问行为数据量大且较容易获取, 所以对该平台的用户访问数据进行挖掘和分析具有一定的可行性与代表性。

本文基于上海市公共数据开放平台的用户访问数据构建用户画像。用户画像一般被定义为一组描述用户属性的结构化数据集^[13], 即信息标签化, 从不同维度的用户属性中获取有价值的用户信息, 提取用户标签, 进行用户特征表示^[14-15]。用户画像在挖掘用户真实需求^[16]、预测用户兴趣^[17]方面具有极大优势, 可以为决策者提供决策建议。通过构建上海市公共数据开放平台用户画像, 可以直观反映用户对政府数据开放平台的真实需求, 并以此对政府数据开放平台的优化提出建议。

1 研究方法

1.1 数据来源

本研究使用网络爬虫获取用户访问上海市公共数据开放平台^[18]的行为数据。上海市公共数据开放平台主要分为九大板块,分别是首页、数据资源、数据图谱、地图服务、典型应用、开放生态、互动社区、安全沙箱以及行业应用。数据资源板块下分为不同的数据领域,包括经济建设、民生服务、城市建设、资源环境、其他等13类,用户在该板块下可以进行数据检索、浏览和下载,同时还可以对数据集进行评分;典型应用板块下有用户注册的典型应用(如“长三角供应链平台”等)以及每个典型应用的浏览量;用户检索的高频关键词一般会以热图形式呈现在网站右上角,用户对数据的使用申请公示一般公布在网页的右侧边栏。爬虫获取的数据主要分为三类:第一类为检索行为数据,即用户对不同检索关键词的检索词频;第二类为操作行为数据,主要包括用户在数据资源板块下对属于不同数据领域数据集的浏览量及下载量、对数据集的综合评分和典型应用板块下用户注册典型应用情况以及典型应用浏览量;第三类为选择行为数据,即网站右侧边栏公布的用户使用申请公示等。

1.2 用户画像构建步骤

本研究分4个步骤构建用户行为画像。①数据处理阶段主要包括数据获取、数据预处理以及数据集构建3个子步骤,其目的是将原始数据中有价值的数据分离出来并将数据转换成合适的数据结构,方便后续的挖掘分析。②数据-标签映射阶段主要包括数据挖掘及信息标签化两个步骤,首先利用文本分析、描述性统计等方法对经过处理的数据集进行深入挖掘,再将挖掘结果进一步标签化。③标签体系构建阶段形成3个维度,即检索行为偏好、操作行为偏好、选择行为偏好。④用户画像构建阶段包括利用构建的标签体系进行用户画像建模和结果分析。

2 用户标签分析

经过数据处理和数据-标签映射后的标签化处理,可以获得用户访问上海市公共数据开放平台的标

签,包括用户的检索行为偏好、操作行为偏好和选择行为偏好。

2.1 用户检索行为偏好分析

用户检索行为偏好的直观体现是用户在数据资源板块下检索关键词的词频,上海市公共数据开放平台网站数据资源页面右上角有根据用户检索词频绘制的用户热搜关键词云图(见图1)。

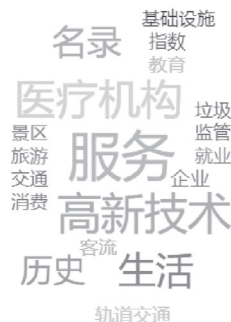


图1 上海市公共数据开放平台用户热搜关键词云图

用户检索频率较高的分别是服务、高新技术、医疗机构等,其中服务的检索词频为3 167次、高新技术的检索词频为2 263次、医疗机构的检索词频为2 153次,这反映了用户对上海市公共数据开放平台的数据需求集中在民生服务、经济建设、卫生健康等领域。

2.2 用户操作行为偏好分析

2.2.1 数据集浏览量及下载量分析

对上海市公共数据开放平台各种类数据集的浏览、下载情况进行统计分析(见图2),可以发现用户对政府数据开放平台资源的领域偏好。

用户对不同数据领域的数据集浏览情况存在显著差异,经济建设领域数据的浏览量远高于其他数据领域,高达1 336 136次。浏览量相对较高的是城市建设、道路交通、资源环境等数据领域。下载量和浏览量总体呈正相关的关系,浏览量越多,则下载量越多。

对各数据领域下数据集的相关字段,如“标题”“关键字”等进行深入分析,可以发现,经济建设领域中,企业、经营情况、纳税、进出口额、公共预算等是用户重点浏览与下载的内容;城市建设领域中,建筑面积、公积

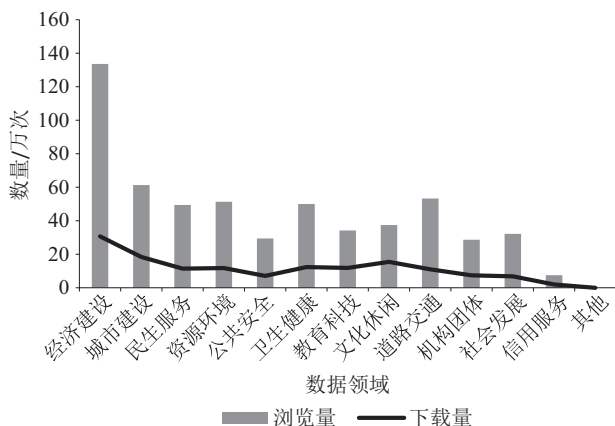


图2 上海市公共数据开放平台数据浏览及下载情况

金、绿化、建筑垃圾等更受用户关注；道路交通领域中，用户更多浏览和下载与城市道路信息、航班信息、天气预警信息、驾培学员信息相关的数据集；资源环境领域中，浏览量与下载量排名靠前的数据集主要是供水、用电、能源消耗、排污、噪声检测等。

2.2.2 典型应用注册情况分析

对用户注册典型应用的行为进行分析，可以得知用户使用数据做了什么，用户更倾向于使用哪个领域的数据进行应用开发以挖掘其潜在价值。

用户在上海市公共数据开放平台注册的典型应用共有50个，比如“长三角绿色供应链平台”“智行浦东”等，根据上海市公共数据开放平台典型应用注册情况（见图3）可知，其中大多属于服务型应用，即通过应用开发服务群众，方便群众在各个方面生活。利用经济建设、民生服务、卫生健康、道路交通领域的数据注册的典型应用数量占总注册应用数量的80%，而经济建设领域的应用数量占比高达44%，说明经济建设领域的的数据既是用户需求的重点，也是用户挖掘的重点。

2.2.3 典型应用浏览量分析

典型应用的使用效果可以通过用户对典型应用的浏览量进行体现。

根据上海市公共数据开放平台典型应用浏览量（见图4）可知，经济建设领域的典型应用浏览量最多，其次是民生服务、卫生健康。显然，这与典型应用的注册数量呈正相关关系。由于注册典型应用的用户通常

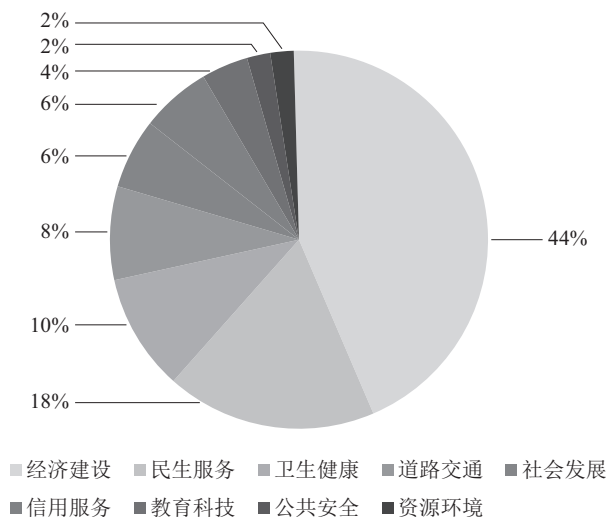


图3 上海市公共数据开放平台典型应用注册情况

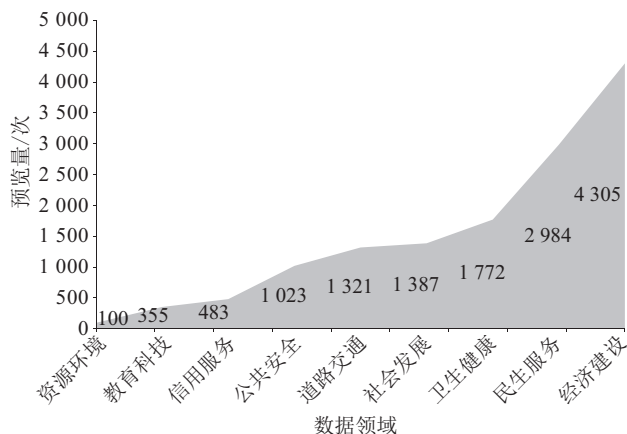


图4 上海市公共数据开放平台典型应用浏览量

为企业用户，预览典型应用的用户通常为普通用户，由此可知，企业用户的需求与普通用户的需求在某些方面是一致的。

2.2.4 数据集评分分析

获取用户对上海市公共数据开放平台数据集的综合评分情况，对不同数据领域的数据集根据综合评分进行排序，汇总得到表1。高达3 750个数据集的用户评分为0分，数据集为0分的原因可能有：①用户对数据集的质量不满意，数据不能满足用户的基本需要；②用户没有对数据集进行评分，参与度不高，这也是政府数据开放平台需要提升的重点，即提高用户使用的参与度。另外，综合评分为5分的数据集数量为714个，其中经济建设领域数据集203个，资源环境领域数据集179个，公共安全领域数据集72个，3个领域的数据集

的数据集需求量更高,且很多典型应用是基于经济建设领域的数据集进行开发的,所以经济建设领域的数据集要求有更高的数据质量。

(2) 政府数据开放平台上的数据评分出现较多0分的异常情况,这说明用户对很多数据的质量不满意或者没有参与到数据评分中,所以用户需要政府采取措施提高数据质量并定期更新数据评分,以更加全面真实地向用户显示开放数据的质量。

(3) 用户对诸如停车场、交通状况、垃圾处理情况等与切身生活密切相关的数据更加关注,也更希望政府及时开放相关数据。

4 政府数据开放平台的优化策略

用户画像技术在政府数据开放平台访问领域的应用能够为平台实现更精准的数据服务。根据用户访问行为构建用户画像,可使政府数据开放平台掌握用户访问的行为数据,对深入研究用户行为,了解用户需求,优化网站服务,实现数据更有针对性的开放,提高数据开放的质量,具有一定的实践意义。针对上海市公共数据开放平台用户访问数据构建的用户画像,本文提出以下政府数据开放平台优化策略。

(1) 加强经济建设领域的数据开放力度。由于经济建设与每个人息息相关且关系到每个人的利益,所以不管是企业用户还是普通用户,对经济建设领域的数据需求都是最大的。加大经济建设领域的数据开放力度可以更好地满足用户的数据需求。加大经济建设领域的数据开放力度包括:增加经济建设领域数据集开放量;提高经济建设领域数据集质量;提升经济建设领域数据集更新频度等。

(2) 定期评估政府数据开放效果。满足用户对政府开放数据的需求和提高政府数据开放的水平需要定期评估。由于政府开放的数据集存在大部分评分为0分的数据集,这部分数据集除了数据质量不高的原因外,还有很大可能是因为用户没有参与评分,用户的反馈没有及时反映到政府数据开放平台,这在一定程度上造成政府对用户的数据需求了解不全面、不真实。为了解决这个问题,政府可以定期向主要的平台用户展开调研,让用户参与到数据的评分以及对平台的评价中,以保证数据开放平台的持续有效运行。

(3) 提高数据开放的覆盖性。用户产生数据申请的行为是由于政府数据开放平台没有公布相关的数据

集,即目前政府数据开放的数据覆盖性具有局限性,用户的数据需求未能得到全面满足。针对用户申请行为,平台应加快审批速度,尽快公开符合开放要求的相关数据,并根据用户的申请,开放其他相关的数据,以更好地满足用户需求。

本研究使用用户画像方法对上海市公共数据开放平台的用户访问行为进行分析,反映了上海市公共数据开放平台的用户需求,不仅有利于政府数据开放平台更好地满足用户需求,也为其他地方政府数据开放平台的用户画像研究提供了参考。

参考文献

- [1] 邓崧,姚承慧. 基于DEA模型的省级政府数据开放效率评价研究[J]. 云南行政学院学报, 2020, 22(4): 2, 116-125.
- [2] 段忠贤,吴艳秋. 政府数据开放度评价指标体系构建与实证测度[J]. 统计与决策, 2019, 35(22): 23-27.
- [3] 林明燕,张廷君. 地方政府数据开放平台绩效评估指标体系实证研究[J]. 图书馆理论与实践, 2019(12): 46-54.
- [4] 迪莉娅. 政府数据开放成熟度模型研究[J]. 现代情报, 2019, 39(1): 103-110.
- [5] 赵树宽,孙彦明,张福俊,等. 基于跨界融合的政府数据开放共享模型研究[J]. 图书情报工作, 2018, 62(12): 21-29.
- [6] 曹慧琴,张廷君. 城市政府数据开放平台发展的影响因素及提升建议[J]. 城市问题, 2020(12): 100-109.
- [7] 刘畅,刘璇. 天津政府数据开放: 现状、问题与政策建议[J]. 图书馆学研究, 2018(11): 40-47.
- [8] 翟军,李晓彤,苗珍珍,等. 我国开放政府数据“脏数据”问题研究及应对——地方政府数据平台数据质量调查与分析[J]. 图书馆, 2019(1): 42-51.
- [9] 顾嘉琪,袁莉. 基于公众需求的政府数据开放服务质量提升研究[J]. 情报杂志, 2020, 39(6): 196-202.
- [10] 郑磊,韩笑,朱晓婷. 地方政府数据开放平台研究: 功能与体验[J]. 电子政务, 2019(9): 12-22.
- [11] 朱红灿,胡新,李顺利. 基于Kano模型的政府数据开放平台用户体验要素分类研究[J]. 现代情报, 2018, 38(12): 13-21.
- [12] 林平,何思奇,段尧清. 数据与用户视角下政府开放数据服务水平评价研究[J]. 图书情报工作, 2020, 64(2): 23-29.
- [13] ATOTE B S, SAINI T S, BEDEKAR M, et al. Inferring emotional state of a user by user profiling [C] //Proceedings of the 2016 2nd international conference on contemporary computing and informatics. Noida: IEEE, 2016: 530-535.

- [14] 金小月. 个性化档案数据服务——用户画像[J]. 卷宗, 2018 (15): 35.
- [15] 韩梅花, 赵景秀. 基于“用户画像”的阅读疗法模式研究——以抑郁症为例[J]. 大学图书馆学报, 2017, 35 (6): 105-110.
- [16] 朱会华, 肖海清, 梁红烂. 基于用户画像的智能荐购模式研究[J]. 图书馆学研究, 2020 (7): 34-38.
- [17] 赵璟博, 朱欣娟. 基于行为数据采集加权的用户兴趣预测[J]. 国外电子测量技术, 2021, 40 (3): 123-129.
- [18] 上海市公共数据开放平台[EB/OL]. [2021-09-10]. <https://data.sh.gov.cn/index.html>.

作者简介

陆敬筠, 女, 1971年生, 博士, 教授, 研究方向: 电子政务及数据分析, E-mail: 2269495034@qq.com。

吕海艳, 女, 1996年生, 硕士研究生, 研究方向: 电子政务及数据分析与处理。

Construction and Analysis of User Portrait on Shanghai Public Data Open Platform

LU JingYun LV HaiYan

(Department of Economics and Management, Nanjing Tech University, Nanjing 211816, China)

Abstract: In order to help the open government data platform understand the user access behavior more accurately, so as to accurately meet the needs of the users, open the data more pertinently, and improve the level of service, this paper takes the data accessed by the users of Shanghai Public Data Open Platform as the research object, constructs the user profile of Shanghai Public Data Open Platform, and analyzes the characteristics of user behavior. Based on user profile construction and behavior analysis, this paper puts forward the optimization strategies of Shanghai Public Data Open Platform. This study also provides a reference for the user profile research of other local government data open platforms.

Keywords: Government Data; Open Platform; User Portrait; Access Behavior; Precise Demand

(收稿日期: 2021-09-12)