基于在线评论的政府数据开放平台用户增量需求研究

李冠 赵毅
（山东科技大学计算机科学与工程学院，青岛 266590）

摘要：深入研究政府数据开放平台用户需求特征，更有效地为用户提供数据服务，让开放数据创造更大的经济和社会价值具有重要意义。本文采集9个省（市、自治区）级政府数据开放平台的用户评论数据，首先运用LDA模型挖掘平台初建期和发展期用户需求主题，分析其需求热点；其次选取朴素贝叶斯算法研究用户需求主题的情感倾向；最后计算两个时期的平台用户需求主题相似度，揭示用户增量需求的动态演进路径。研究结果表明，“数据改进”“社会保障”“公共安全”等需求主题是用户持续关注的领域，其中“社会保障”主题的情感倾向概率为0.75，用户情感倾向积极，“数据改进”主题的情感倾向概率为0.22，用户情感倾向消极。本文研究平台用户需求主题、情感诉求以及需求变化，促进政府数据开放平台根据用户增量需求不断持续改进，为平台建设发展提供有力支持。

关键词：政府数据开放平台；用户增量需求；主题挖掘；情感倾向；文本相似度；主题演进
中图分类号：G203　DOI：10.3772/j.issn.1673-2286.2022.12.005
引文格式：李冠，赵毅. 基于在线评论的政府数据开放平台用户增量需求研究[J]. 数字图书馆论坛，2022（12）：37-46.

政府数据开放是数字化政府和社会发展到一定时期的产物[1]。政府向公众开放和共享数据资源，充分发挥了数据资源的价值，提高了政府综合管理服务能力。随着智慧城市、智慧乡村等城乡数字化建设，公众对数据资源的开放共享需求日益增长。因此，我国在推进数字化进程中正在加快政府数据开放平台建设。2015年8月，国务院颁布的《促进大数据发展行动纲要》指出，到2018年，中央层面构建统一的互联网政务数据服务平台[2]。根据复旦大学数字与移动治理实验室统计，截至2021年10月，我国已有193个政府数据开放平台，基本实现了公共数据的平台化服务[3]。为了更加精准、有效地为用户提供数据服务，让开放数据创造更大的经济和社会价值，深入研究平台用户的增量需求具有重要意义。

1 相关研究

近年来，国内外学术界非常关注政府数据开放平台在用户需求方面的研究。本文对相关文献进行梳理，总结归纳为以下三个方面的研究内容。①对用户需求影响指标的相关研究。Aguilera等[4]提出政府数据开放的关键主体是社会群众，用户既是数据的提供者也是数据的使用者，通过数据分析提出了用户的不同需求类型，为研究用户需求提供了一种路径。Zuiderwijk等[5]提出影响用户的需求指标，运用实证方法研究了用户对平台的接受度和使用意愿，并对用户的接纳度进行量化分析。Danneels等[6]使用3种认识论定义了3种不同类型的数据开放平台，指出尽管用户对政府开放数据寄予厚望，但是目前政府数据开放平台仍未达到公众预期。Susha等[7]使用解释性荟萃分析的方法，分析了用户需求的主要影响指标，结合具体实证研究了公众对政府数据开放平台的接受程度和推广意愿。莫祖英等[8]从用户视角出发，构建了政府数据开放质量的模糊评价模型，为我国政府数据开放质量的用户评价提供了评判标准。刘迪[9]结合文献分析法、问卷调查法和观察法等分析了平台对用户需求的满意情况，对用户的需求类型

做出分类，提出了通过定期评估数据开放成果健全政府数据开放平台功能。用户需求指标的相关研究为本文划分用户需求类型以及衡量用户满意度等提供了参考。②结合开放数据政策对平台用户需求进行分析。陈玲等[10]运用TF-IDF算法和LDA模型对我国数据开放政策进行了实证分析，揭示了我国政府开放数据政策的供给和需求存在不平衡的现象，未来要以满足用户需求为中心，做好数字服务型政府的工作。朱晓峰等[11]通过挖掘政府数据开放相关政策与实施成效间潜在因果关系，得出良好成效的政策前因条件组合，提出了政策支持是国家开放数据战略的重要组成部分。相关研究阐明了开放数据政策与用户需求的关联因果，为本文对用户需求的研究提供了新视角。③以某一平台为例分析用户需求特征。陆敬筠等[12]统计了上海市政府数据开放平台用户的访问次数，通过分析平台用户的行为特征构建用户画像，有助于平台了解用户需求，提升平台的服务水平。刘桂琴[13]以武汉市政府数据开放平台的用户评论为研究对象，提出从用户情感视角研究用户需求，发现有7个主题的情感趋向消极，2个主题的情感趋向积极，平台服务在一些领域没有较好地满足用户需求。杨洋[14]深度挖掘政府数据开放平台中的城市交通数据，从用户的兴趣点和兴趣面角度预测了开放数据在用地功能识别、职住人口分布和交通出行方面的需求，为交通规划拓展了数据获取渠道。

综上所述，学者们已经取得了较为丰富的研究成果，相关研究结论对政府数据开放平台的发展具有重要的参考价值。但经过分析发现，平台用户需求研究方面的样本多来源于某一省或市级平台，而以多个省或市级平台为研究样本，并将需求主题聚类和情感分析相结合的研究较少，该领域仍存在进一步改进空间。因此，本文以9个省（市、自治区）级政府数据开放平台为研究样本，采集平台用户在线评论数据，挖掘用户需求主题特征，选取朴素贝叶斯算法研究用户需求主题的情感倾向，通过对初建期和发展期的平台用户需求主题相似度计算研究平台用户需求主题演进路径，分析用户需求的动态变化，从而深入研究政府数据开放平台用户需求主题、情感诉求以及增量需求变化，为政府数据开放平台的发展提供借鉴。

2 样本选取与数据处理

2.1 样本选取

本文以复旦大学数字与移动治理实验室于2021年10月发布的《中国地方政府数据开放报告指标体系与省域标杆》中的18个省级政府数据开放平台为初始样本，以平台用户浏览量大于30万且用户评论大于500条为筛选依据，选取上海市、广东省、山东省、湖北省、江苏省、浙江省、贵州省、广西壮族自治区、内蒙古自治区（排名不分先后）9个省（市、自治区）的政府数据开放平台为研究样本。

2.2 数据采集和预处理

本文采用八爪鱼采集器爬取2017年1月1日—2022年6月30日9个省（市、自治区）政府数据开放平台的用户评论数据19 908条，数据字段包括用户名称、评论内容、评论时间，整理得到的初始平台用户评论数据如表1所示。

由于评论文本包含许多无效的干扰信息，主题挖掘的结果受数据质量影响较大，因此需要对文本数据进行预处理。首先，通过Jieba分词对已爬取的文本进

表1 初始平台用户评论数据表（部分）

序 号	用户名称	评论内容	评论时间
1	随风**	请问，这个接口该如何传参啊，使用python该怎么调用，能不能发到我的邮箱 lovih**@qq.com，麻烦了，感谢感谢	2018-08-14
2	y****	希望可以快捷查看每个数据集的评论数	2022-08-16
3	刘**	网站预留的两个联系电话两天都没打通，望回应山东省生态环境厅，“互动交流”→“您的建议”板块无法进行留言，总是出现验证码输入错误	2020-06-29
4	s******	建设工程扬尘与噪声在线监测管理系统日均值，接口服务地址怎么下载	2020-11-09
5	一只**	请更新一下出版物经营许可证数据表，闲鱼入驻需要经营许可证官网查询，谢谢	2021-04-17
6	H*****	整体目录更新有点慢，建议加快下更新周期	2022-08-16

行分词，去除无关的字符。其次，去除用户评论中的停用词。经过数据预处理后，最终得到17 871条用户评论数据，如表2所示。

表2 预处理后的平台用户评论数据表（部分）

序 号	处理结果
1	接口 传参 使用 调用 发 邮箱 麻烦
2	希望 快捷 查看 数据集 评论 数
3	网站 预留 两个 联系 电话 两天 打通 回应 山东省 生态 环境 厅 互动 交流 建议 版块 留言 出现 验证码 输入 错误
4	建设 工程 扬尘 噪声 在线 检测 管理 系统 日 均值 接口 服务 地址 下载
5	更新 出版物 经营 许可证 数据 表 闲鱼 入住 官网 查询
6	整体 目录 更新 慢 建议 加快 周期

2.3 政府数据开放平台的时期划分

复旦大学联合国家信息中心数字经济研究所共同发布的《2020中国地方政府数据开放报告》显示，全国有130多家政府数据开放平台，政务数据开放共享已成为政府公开信息治理和建设数字型政府的新标准，“开放数据，蔚然成林”的目标也基本实现[15]。2020年12月发布的《关于加快构建全国一体化大数据中心协同体系的指导意见》指出，我国已实现政府数据开放平台的初步建设目标，下一阶段将在各省数据开放平台的基础上形成分配优化、绿色集约的一体化格局[16]。鉴于此，本文以2020年为时间节点，将政府数据开放平台的建设分为初建期（2017年1月—2020年12月）和发展期（2021年1月—2022年6月）两个时期，进行用户需求主题研究。

3 基于LDA模型的平台用户需求主题聚类与热点分析

3.1 LDA主题模型

主题模型是一种对文本隐含主题进行建模的方法，通过将高维度的词的集合映射到低维度的主题空间，实现对目标数据的降维。在现有的研究中，根据适用对象的不同可将主题模型分为两种：第一种是适用于长文本处理的主题模型；第二种主要面向的是不超过10个词的短文本，隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）[17]是其中的一种经典模型，具有分类准确性高、可解释性强、应用广泛等特点。考虑到研究数据中用户评论的篇幅较短，运用LDA可以有效地识别语料库中潜藏的主题信息，因此，本文选择LDA主题模型实现对用户需求的主题聚类。

3.2 平台用户需求主题聚类

基于2.2中已清洗的用户评论数据构建词频矩阵，并建立LDA主题模型。利用LDA模型对数据进行训练可以得到不同主题的分类情况。训练参数设置为：选定主题范围1~20；设置超参数α=20/K，β=0.01，其中K为主题个数；同时设置训练次数为500次。通过调用LDA主题模型中的Log_Perplexity，分别计算平台初建期和发展期的困惑度数值，并依据其数值确定LDA最优主题数。

Azzopardi等[18]认为，困惑度数值越低，文档归属某一潜在主题的概率越高，表明模型的聚类效果越好，由此可以确定出最优主题数。因此本文得出，当平台初建期的困惑度数值为63时达到最小值，当前语料的最优主题数目为8；当平台发展期的困惑度数值为121时达到最小值，平台发展期的最优主题数为11。

结合LDA模型主题词的自动提取和现有研究的领域分布归类[19]以及人工判读，对初建期和发展期各主题下的高频词内容进行总结归类得到主题描述，分别如表3、表4所示。

3.3 基于LDA主题聚类的平台用户需求热点分析

本文依据上述平台用户需求LDA主题聚类，分别从平台初建期和发展期两个时期进一步分析政府数据开放平台的用户需求热点，为相关部门精准分析用户的需求特征以及平台的优化与发展提供参考。

表3 初建期的主题列表

主 题	主题描述	各主题下的部分高频词
主题1	数据质量	来源 全面 分类导航 数据集 格式 更新周期 增长 可视化 统计 机构团体 精准 开放授权 下载量
主题2	基本服务	养老保险 旅游 医疗卫生 社保卡 交通出行 核酸 农业 环境保护 教育科技 文化 休闲 监督 商贸流通 体育 挂失 机构 项目 申请公示
主题3	安全保障	覆盖 系统 监控 数据来源 隐私 平台 信息泄露 统计 政策 可靠性 稳定 无效身份 公开 追溯
主题4	公共服务	证书 考试 无差别 企业 地理空间 工具 App 系统 开放授权 研究 金融 财政分析报告 地理空间
主题5	数据获取	接口 元数据 大数据 权限 可机读 变更 数据搜索 覆盖 加工处理 涵盖 搜索栏 免费 准确性 完整数据集
主题6	分类导航	主题分类 部门检索 展示 搜索 导航栏 定位 图标引导 可视化 关键词 查找 热门数据 排序 标明
主题7	平台建设	推广 订阅 身份认证 引擎 升级 排序 平台界面 注册登录 权限 加载 更新 设施建设
主题8	互动交流	交流 反馈 内容意见 开放生态 数据需求 纠错 问卷调查 常见问题 研究 申请公示 咨询

表4 发展期的主题列表

主 题	主题描述	各主题下的部分高频词
主题1	民生服务	教育科技 证书 考试 商业 文化休闲 企业 技能考试咨询 农业 地理空间 机构 财税金融 用途 开放授权 研究 刷卡支付 工业
主题2	信息化建设	接口 数据集 可读性 信息 权限 更新动态 精准 定位 智慧城市 研发 一体化 政策 协同 集约化
主题3	功能优化	接口 元数据 增长 定期更新 订阅 注册 可机读 推广 收藏 检索 移动端 稳定 空间工具
主题4	创新平台	海量数据 预览 在线查看 高速 引擎 网页应用 创新 监督 卫星遥感 方案 升级 智慧
主题5	效能	参与度 覆盖面 适配 终端 开放程度 搜素 复杂度 区域 资金 加快 协同 大数据 检测
主题6	智慧服务	人工智能 动态 智慧 更新 区域链 链接 授权 大数据 算法 学习 系统 共享体系 提升 示范
主题7	数据应用	可视化 评审 智能化 地理空间 工具 功能描述 下载链接 App 稳定 精准 定位 需求 解读
主题8	公共安全	原始 隐私 平台 生产安全 信息泄露 统计 无效身份 公开 安全体系 认证 数据运行 指标
主题9	数据改进	格式 更新周期 全面 可视化 数据开发 研究 分类导航 准确 覆盖 开放授权 动态更新 下载量
主题10	评价反馈	改进 互动 反馈途径 公示 咨询 客服 问卷调查 讨论区 申请 下载动态 反馈机制 纠错 举报
主题11	社会保障	养老保险 医疗卫生 社保卡 城建住房 核酸 生活服务 旅游 市场监管 挂失 交通出行 机构 项目 权益保护

3.3.1 初建期 (2017年1月—2020年12月)

①数据质量主题下的高频词主要包括“全面”“数据集”“更新周期”“开放授权”等，可以看出这些词语都是围绕开放数据服务需求所展开的，通过这些词可以分析出，用户对数据的分类、数据的格式和数据可视化等方面的需求意愿比较强烈。

②基本服务主题下的高频词主要包括“养老保险”“卫生医疗”“社保卡”等，该主题所反映的高频词与群众的日常活动息息相关，这是平台所提供最基础、也是非常重要的服务，因此将其归纳为“基本服务”。

③安全保障主题下的高频词主要包括“隐私”“平台”“隐私泄露”等，不难看出，这些内容都是与开放数据的安全相关的。用户在使用政府所提供开放数据的同时，意识到了数据安全的重要性。数据隐私的保护、部分信息存在泄露、登录身份无效等是用户所关心和亟待解决的问题。

④公共服务主题下的高频词主要有“证书”“地理空间”“考试”“研究”“企业”等。这些词语都与政府职能部门所提供的公共服务关联性较强。用户在政府数据开放平台上查询与生活相关联的内容；查询一些资格等级证书或评定证书等；此外，用户或者企业做研究调查，相关的数据来源需要在数据开放平台下载。可见该主题是用户对数据开放平台所提供服务的描述，因此将其归纳为“公共服务”。

⑤数据获取主题下的高频词有“接口”“可机读”“权限”等，这些词语都是跟数据获取相关，由此可见，用户对数据的获取方式、处理加工等也是极为关注的。

⑥分类导航主题下的高频主要包括“主题分类”“部门检索”“可视化”等。在浩如烟海的数据中，如何快速、准确地检索到自己想要的数据是用户所关心

的问题之一。该主题主要是对数据资源分类的描述，因此将该主题归纳为“分类导航”。

⑦平台建设主题下的高频词有“推广”“订阅”“平台界面”等。政府数据开放平台为用户提供了最基础的功能服务，但平台处于起步期，基础建设存在一定的改进空间，用户对平台的界面设计、订阅等功能存在着优化需求。由此可以看出，该主题主要是用户对于平台建设需求的描述，因此将该主题归纳为“平台建设”。

⑧互动交流主题下的高频词主要有“反馈”“开放生态”“问卷调查”。政府数据开放平台设立了互动交流板块，用户可以向平台提出自己的数据需求，该板块还有需求列表、内容建议等功能，展示用户需求的热门数据以及平台建设过程中哪里存在的不足、改进方案等意见。

3.3.2 发展期 (2021年1月—2022年6月)

①民生服务是政府数据开放平台所提供的基础服务，查询资格等级考试、医疗卫生健康、气象服务、财税金融等功能与用户的日常息息相关，避免不了用户对该主题数据的大量需求。

②信息化建设主题下的高频词主要包括“信息”“协同”“集约化”。2020年12月国家发改委发布的《关于加快构建全国一体化大数据中心协同创新体系的指导意见》提出，进一步打破政府与部门间、政府与企业间的数据壁垒，大大提高数据资源的流通性。用户积极响应国家号召，对通信、计算机、数据库技术等现代信息化相关的数据需求增多，信息化建设逐渐成为热门主题。

③功能优化主题下的高频词主要有“定期更新”“订阅”等，所反映的主题词与平台功能迭代升级相关，因此将该主题归纳为“功能优化”。

④创新平台主题下的高频词有“海量数据”“预览”“在线查看”。政府数据开放平台不仅要在开放的数据量上做文章，还要对平台的前端展示上下功夫，通过对平台功能的迭代升级为用户提供独特服务。

⑤效能主题下的高频词主要包括“参与度”“覆盖面”“复杂度”，用户根据平台所提供的数据衡量平台的效能，可以看出，该主题主要是对平台效能的描述。

⑥智慧服务主题下的高频词有“区域链”“共享体系”。根据2018年发布的《智慧城市顶层设计指南》，国家给出了智慧城市顶层的总体构架方案。在2020年召开的全国两会上，政府报告将“新基建”首次纳入其中，并提出“城市大脑”是智慧城市建设突破的关键。智慧城市的发展日趋成熟，一方面带动着智慧数字政务的发展，要求政府数据开放的服务更加智慧化；另一方面用户对互联网、区块链、人工智能、大数据等技术浏览需求量越来越高。

⑦数据应用主题下的高频词主要有“可视化”“地理空间”“工具”等。所反映的主题词与数据的实际应用有关，因此将该主题归纳为“数据应用”。

⑧公共安全主题下的高频词主要有“隐私”“生产安全”“信息泄露”。不仅在生活中存在安全问题，用户在使用数据的同时也充斥着安全隐患，这些是用户所关心的问题。因此，将该主题归纳为“公共安全”。

⑨数据改进主题下的高频词有“格式”“更新周期”“数据开发”。随着近几年的发展，政府数据开放平台的建设日趋成熟，用户对平台的功能以及开放的数据提出了更高的要求，例如优化平台的网页链接、加快数据的更新周期等。

⑩评价反馈主题下的高频词有“客服”“反馈机制”“申请”等，政府数据开放平台的建设离不开对其评估和用户反馈，对平台数据的开放程度、覆盖面等效能进行有效评估并做出改进，认真接受用户提出的评价反馈。

⑪社会保障主题下的高频词主要包括“养老保险”“医疗卫生”“生活服务”等。用户在需求创新服务的同时，离不开基础的社会保障服务，该主题主要是用户对社会生活需求的描述，因此将该主题归纳为“社会保障”。

4 平台用户需求主题情感倾向分析

本文利用各个主题下的用户评论数据进行情感分析，得到用户评论中各个主题的情感倾向，可以从情感诉求层面了解平台用户满意度，有助于有关部门做出相应的改进措施。

4.1 用户评论数据预处理

用户评论通常由一句话或多句话组成，对于一些比较短的评论，用户通常只表达一个主题，但对于一些较长的评论，用户可能表达对多个主题的评价。为了提高评论主题识别的准确性，将评论中的长句切割成短句，再对评论短句进行主题识别[20]。

通常情况下，以标点符号分割的每个短句仅包含一个主题。因此，可以根据标点符号作为界限对评论进行分割。首先对每条评论设置一个id值，然后根据标点符号将评论切割成多个短句，切割后的评论短句与切割前的原评论id一致。本文将17 871条用户评论切割，并去除无关短句，最终得到28 476条评论短句，分句后的评论格式如表5所示。

表5 用户评论分句示例

id	新序列号	切割后的评论短句
132	326	希望开放更多有价值数据
132	327	加快更新速率
158	374	数据发布者联系方式希望可以更加详细
158	375	提交数据申请时网页有一点卡顿

4.2 构建训练集和测试集

本文基于机器学习对用户需求主题进行情感分析[21]，将4.1中已处理的28 476条评论短句作为语料数据集。筛选出用户正面情感倾向的评论，用标签“1”标记；筛选出用户负面情感倾向的评论，用标签“0”标记。根据相关研究结论，中立性的评论不纳入情感处理中。经过人工筛选和标记后，得到正面评论6 854条，负面评论7 356条。将人工标注好的样本按照训练集80%，测试集20%的比例随机采样。

4.3 分类器性能测试

在本次的文本二分类任务中，使用sklearn库中常见的5种文本分类器[22]，包括逻辑回归分类器、支持向量机分类器、K邻近分类器、决策树分类器以及朴素贝叶斯分类器等。本文以准确率、F1分数（F1-score）和召回率为模型有效性的评估标准，用5种文本分类器进行测试，最终得到各个分类器的性能测试结果。由表6可见，朴素贝叶斯在此次情感分析任务中表现最好，因此选用朴素贝叶斯分类器对各用户需求主题进行情感分析。

表6 分类器性能测试结果

分类器名称	准确率	F1	召回率
朴素贝叶斯	0.857 564	0.865 939	0.885 75
逻辑回归	0.781 195	0.863 024	0.835 461
决策树	0.835 415	0.849 906	0.868 212
SVM	0.834 77	0.773 146	0.823 796
KNN	0.715 439	0.762 954	0.854 58

4.4 用户需求主题情感倾向分析

首先，使用训练好的朴素贝叶斯分类器计算评论短句的情感倾向概率。朴素贝叶斯分类器计算出的情感倾向为正向概率，小于0.5的情感倾向为负面情感倾向，大于0.5的为正面情感倾向[23]。其计算结果如表7所示。

表7 各用户需求主题的情感倾向概率计算结果（部分）

id	序 号	用户评论内容	所属需求主题	情感倾向概率
257	158	济宁市住房公积金管理中心数据描述和实际数据不符	社会保障	0.363 621
2165	1362	首页“疫情防控”找不到自疫情发生以来的所有实时数据	民生服务	0.334 756
47	23	接口调用样例丰富	功能优化	0.945 753
4356	3137	客服的回复内容太过官方	评价反馈	0.039 285
854	492	提供数据的格式种类较多	数据改进	0.913 027
68	37	我申请接口服务一个月了竟然还是待审核的状态	评价反馈	0.018 294

然后，使用朴素贝叶斯分类器对各个需求主题的情感倾向概率进行可视化分析，其结果如图1所示，X轴表示需求主题，Y轴表示需求主题的情感倾向概率。

图1中“社会保障”“民生服务”“数据应用”“创新平台”需求主题的情感倾向偏向积极，而“数据改进”“评价反馈”“智慧服务”“功能优化”“效能”“信息化建设”和“公共安全”需求主题的情感倾向偏向消极。

基于上述分析结果，建议相关部门可以根据“数据改进”“信息化建设”等负面情感主题数据，研究和解决平台存在的问题，进一步满足用户在数据质量、平台服务等方面的需求。针对“社会保障”“民生服务”等正面情感主题数据，继续扩大该领域的数据开放量，提高公共数据的利用价值。

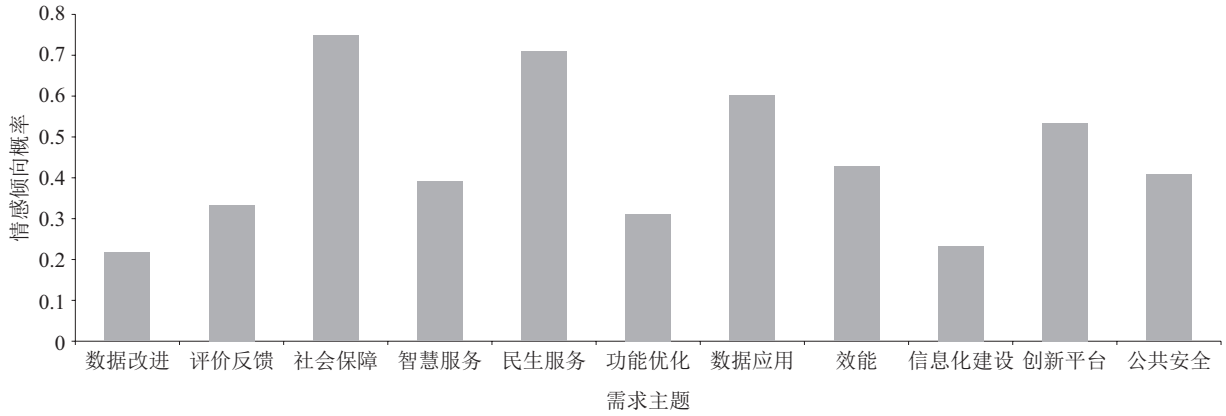

图1　用户需求主题情感倾向

5　基于相似度的平台用户增量需求主题演进分析与建议

在主题演进过程中，主题词出现频次、新增词汇、词间关系等方面的变化在一定程度上能够反映主题内容的变迁。通过对平台初建期和发展期的用户需求主题相似度计算，揭示用户的增量需求变化，为平台的建设提供借鉴。

5.1　用户需求主题相似度计算

相似度可以应用于不同的研究领域，研究领域不同其含义也不同。余弦相似度量法是计算文本相似性中最常用的算法之一[24]，用向量空间中的两个向量夹角间的余弦值作为度量两者间差别大小的尺度。若余弦值愈接近1，表示夹角的角度愈趋近0°，则说明两个向量越相似；若余弦值趋于0，且夹角趋近90°，则表明两个向量越不相似。

Kong等[25]从信息论的视角入手，阐明了主题间是否存在演化关系可以根据相似度大小确定。刘自强等[26]提出，将0.3确定为相似度阈值，假如相邻阶段的主题相似度数值大于0.3，那么可以确定主题间具有演进关系，并且属于同一条演进路径。

本文依据3.2中已得到的主题聚类，并计算两个主题关键词的余弦相似度，每个主题由一组关键词确定，通过主题聚类将语义特征相符的关键词聚类为同一主题，不同主题间的相似度不同。利用Python中Gensim库BOW模型和TF-IDF模型对平台初建期和发展期的用户需求主题进行相似度计算，计算得到相似度大于0.3的主题演进路径5条，包括：初建期主题1—发展期主题9，相似度为0.61；初建期主题2—发展期主题11，相似度为0.52；初建期主题3—发展期主题8，相似度为0.47；初建期主题4—发展期主题6，相似度为0.43；初建期主题1—发展期主题2，相似度为0.33。相似度小于0.3的主题演进路径有31条。

5.2　用户增量需求主题演进路径分析

对不同阶段用户需求主题相似度计算可以揭示其相似性和差异性，本文筛选出5条相似度大于0.3的主题演进路径，下面对其进行具体分析。

①初建期主题1—发展期主题9：这条路径对数据质量、数据改进主题较为关注，该演进路径体现了用户对开放数据需求在数据质量领域关注的延续性。随着大数据时代的发展，用户对数据的需求与日俱增，不仅表现在对数据集的宽度，而且对数据集的深度提出了更高的要求。

②初建期主题2—发展期主题11：主要关注基本服务、社会保障服务，这条演进路径体现了用户对社会保障服务领域的持续关注。社会民生相关主题的数据与用户的日常生活息息相关，公众的衣食住行都离不开社会保障，所以用户对社会保障服务领域相关的数据需求与日俱增。

③初建期主题3—发展期主题8：该路径主要关注安全保障、公共安全，这一主题演进体现了用户对安全保障的持续关注，表明用户有着安全防范和自我保护意识，相关部门在对数据开放利用时理应进行风险管控，定时定期对开放数据进行动态监管和风险评估，以防止数据泄露和滥用。

④初建期主题4—发展期主题6：该路径主要关注

公共服务、智慧服务，这条演进路径体现了用户对智慧服务的新需求。与传统服务不同，智慧服务基于平台与用户的隐性交互过程感知用户需求，可以更加准确地挖掘用户内在需求。

⑤初建期主题1—发展期主题2：数据质量和信息化建设是该路径下较为关注的主题，信息化建设是发展期新兴的需求主题之一，在大数据背景下，信息化建设不仅能够符合时代的发展趋势，还能提高政府数据开放平台服务水平。

5.3 政府数据开放平台发展建议

本文基于上述用户需求主题演进路径分析，为政府数据开放平台提出以下建议，以供参考。

（1）满足用户的数据改进需求。数据质量的高低决定了政府数据开放平台建设的好坏以及是否能够满足用户的使用需求，随着大数据时代的发展，用户对数据的需求与日俱增，不仅表现在对数据集的宽度，而且对数据集的深度提出了更高的要求[27]。政府相关部门在进行管理时，一方面，要及时更新数据目录清单；另一方面，还应注意备份和清理“过时”数据，减少系统内部存储压力，提高平台操作的使用效率。开放的数据集往往以相应的格式进行存储，格式种类越丰富，在一定程度上表示政府数据开放平台开放程度越高，但是对格式规划分类时，尽可能避免类似格式都被采用，例如“XLS”和“XLSX”、“DOC”和“DOCX”等，避免给存储系统带来不必要的压力。

（2）完善社会保障相关的数据供应。社会民生相关主题的数据与用户的日常生活息息相关，公众的衣食住行都离不开社会保障，因此，针对社会民生主题数据应当继续加大开放力度[28]。首先，提高数据集的覆盖面，相关部门制定总体战略，提升大数据资源的储备容量和运营水平；其次，减少二次数据等处理过的数据，开放更多的原始数据从多角度满足用户的不同需求，在提高数据量的基础上保证数据质量；最后，重点关注API接口的开放性，开放式API接口在提高数据价值方面发挥着重要作用。

（3）推进智慧服务的发展。与传统服务不同，智慧服务基于平台与用户的隐性交互过程来准确感知用户的需求。通过对用户信息与检索查阅行为的分析，可以充分挖掘数据信息的内在价值，为用户精准推送所需数据集。政府数据开放平台应借助大数据、神经网络与云计算技术构建强大的数据挖掘与分析系统[29]，提升数据分析效能，通过收集数据构建用户画像，在充分尊重和保护用户隐私的前提下构建用户知识体系，开展精准推送。

（4）重视用户需求反馈。用户所提供的反馈意见，一方面有利于用户之间互相交流，另一方面更有利于政府数据开放平台更好地满足用户的需求以及优化后续的建设。政府相关部门不仅要收集用户的需求反馈，更要对用户的反馈进行解答[30]，把切实解决用户的问题作为目标，尽量避免空话、套话，因为这样会使用户的体验感较差，从而对用户与开放数据平台之间的互动交流产生不利影响。

（5）加快数字化建设。为提高数据开放平台的数字化建设效果，在实际工作中需要完善信息管理模式，引进先进的技术平台，实现管辖数据资源的数字化发展[31]，可以在内部设立数字化信息平台，对已有的数据资源进行有效储存，再配合检索平台，多方位满足用户的检索需求。

6 结语

近年来，随着服务型数字政府的建设，政府数据开放平台的服务理念不断增强，深入研究用户需求，不断提高平台服务能力具有重要意义。本文通过LDA模型对平台用户评论主题聚类，挖掘出用户的增量需求热点，在此基础上选取朴素贝叶斯算法计算用户需求主题的情感倾向，发现社会保障主题是用户持续关注的领域且用户情感趋向积极，而用户对数据改进主题的情感趋向消极；最后通过文本相似度计算出用户需求的演进路径，结果表明平台不仅要满足社会保障等基础的主题服务，还要完善信息化建设、智慧服务等主题数据，以满足用户增量需求的发展。

本文从平台用户需求的整体架构出发研究并挖掘用户的需求主题，为当前处于建设发展期的政府数据开放平台更好满足用户需求提供有效借鉴。在后续的研究中，将通过研究平台用户的信息行为特征构建用户画像，希望有助于提升政府数据开放平台个性化、智慧化服务水平。

参考文献

[1] 肖冬梅，苏莹. 我国政府数据开放中的安全风险及其防范对策 [J].

现代情报，2022，42（6）：112-120，131.

[2] 陈伟. 《促进大数据发展行动纲要》解读 [J]. 中国信息化，2015（10）：11-14.

[3] 复旦大学数字与移动治理实验室. 中国地方政府数据开放报告——省域（2021年度）[R/OL]. [2022-01-20]. http://www.ifopendata.cn/static/report/%E4%B8%AD%E5%9B%BD%E5%9C%B0%E6%96%B9%E6%94%BF%E5%BA%9C%E6%95%B0%E6%8D%AE%E5%BC%80%E6%94%BE%E6%8A%A5%E5%91%8A%EF%BC%88%E7%9C%81%E5%9F%9F%EF%BC%89.pdf.

[4] AGUILERA U，LÓPEZ-DE-IPIÑA D，PÉREZ J. Collaboration-centred cities through urban apps based on open and user-generated data [J]. Sensors，2016，16（7）：193-204.

[5] ZUIDERWIJK A，JANSSEN M，DWIVEDI Y K. Acceptance and use predictors of open data technologies：Drawing upon the unified theory of acceptance and use of technology [J]. Government Information Quarterly，2015，32（4）：429-440.

[6] DANNEELS L，VIAENE S，VAN DEN BERGH J. Open data platforms：Discussing alternative knowledge epistemologies [J]. Government Information Quarterly，2017，34（3）：365-378.

[7] SUSHA I，ZUIDERWIJK A，JANSSEN M，et al. Benchmarks for evaluating the progress of open data adoption [J]. Social Science Computer Review，2015，33（5）：613-630.

[8] 莫祖英，邝苗苗. 基于用户视角的政府开放数据质量评价模型及实证研究 [J]. 大学图书情报学刊，2020，38（4）：84-89.

[9] 刘迪. 基于用户需求的地方政府数据开放平台优化策略研究 [D]. 哈尔滨：哈尔滨工业大学，2019.

[10] 陈玲，段尧清. 我国政府开放数据政策的实施现状和特点研究：基于政府公报文本的量化分析 [J]. 情报学报，2020，39（7）：698-709.

[11] 朱晓峰，葛锐，洪磊. 共生理论解构下政府数据开放政策导向与实施成效的组态分析 [J]. 图书情报工作，2022，66（14）：77-88.

[12] 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.

[13] 刘桂琴. 政府数据开放平台用户评论情感差异分析 [J]. 数字图书馆论坛，2019（2）：18-23.

[14] 杨洋. 面向城市交通需求分析的多源数据分析及应用研究 [D]. 南京：东南大学，2020.

[15] 复旦大学数字与移动治理实验室. 中国地方政府数据开放报告（2020下半年）[EB/OL]. [2022-12-12]. http://ifopendata.fudan.edu.cn/static/report/%E4%B8%AD%E5%9B%BD%E5%9C%B0%E6%96%B9%E6%94%BF%E5%BA%9C%E6%95%B0%E6%8D%AE%E5%BC%80%E6%94%BE%E6%8A%A5%E5%91%8A%EF%BC%882020%E4%B8%8B%E5%8D%8A%E5%B9%B4%EF%BC%89.pdf.

[16] 四部门印发加快构建全国一体化大数据中心协同创新体系指导意见 [J]. 中国信息化，2021（1）：24-27.

[17] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research，2003，3：993-1022.

[18] AZZOPARDI L，GIROLAMI M，VAN RISJBERGEN K，et al. Investigating the relationship between language model perplexity and lR precision-recall measures [C] //Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval，New York：ACM Press，2003：369-370.

[19] 陈美，何祺. 基于文本挖掘的我国省级政府开放数据平台比较研究 [J]. 图书情报工作，2022，66（7）：88-98.

[20] WITTEN IH，PAYNTER G W，FRANK E，et al. KEA：practical automatic keyphrase extraction [C] //Proceedings of the 4th ACM Conference on Digital Libraries.New York：ACM Press，1999：254-256.

[21] PANG B，LEE L，VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [J]. CoRR，2002，cs.CL/0205070.

[22] 陈祎荻，秦玉平. 基于机器学习的文本分类方法综述 [J]. 渤海大学学报（自然科学版），2010，31（2）：201-205.

[23] 周妹，常建华，陈思成，等. 一种基于朴素贝叶斯分类器的气溶胶类型识别模型 [J]. 光学学报，2022，42（18）：49-57.

[24] LIN D. An Information-theoretic Definition of Similarity [C] // Proceedings of the 15th International Conference on Machine Learning.1998.

[25] KONG Z J，Ll B. Method for integration and selection of innovation demands based on three-dimensional coordinates and cosine similarity measures [J]. Operations Research and Management Science，2018，27（11）：87-94.

[26] 刘自强，王效岳，白如江. 多维度视角下学科主题演化可视化分析方法研究——以我国图书情报领域大数据研究为例 [J]. 中国图书馆学报，2016，42（6）：67-84.

[27] 赵需要，郭义钊，姬祥飞，等. 政府开放数据生态链上数据要素价值分析及评估模型构建——基于“数据势能”的方法 [J]. 情报理论与实践，2022，45（12）：50-59.

[28] 张晓娟，莫富传，冯翠翠. 政府数据开放价值实现的机理：基于系统动力学的分析 [J]. 情报理论与实践，2022，45（5）：75-83.

[29] 王意，莫富传，张晓娟. 政府数据开放生态系统的理论、要素与模型探究 [J]. 情报理论与实践，2022，45（12）：42-49.

[30] 王法硕. 我国地方政府数据开放绩效的影响因素——基于定性比较分析的研究 [J]. 情报理论与实践，2019，42（8）：38-43.

[31] 侯晓丽，赵需要，樊振佳，等. 政府开放数据生态链的形成机理与培育策略 [J]. 情报理论与实践，2021，44（6）：7-17.

作者简介

李冠，女，1970年生，硕士，副教授，研究方向：智能信息处理、数字资源管理，E-mail：liguan0105@163.com。
赵毅，男，1998年生，硕士研究生，研究方向：图书馆与信息服务。

Research on User Incremental Demand of Government Data Open Platform Based on Online Comments

LI Guan ZHAO Yi
(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, P. R. China)

Abstract: It is of great significance to deeply study the characteristics of users' needs of government data open platform, provide users with data services more effectively, and make open data create greater economic and social value. This thesis collects user comment data of nine provincial (municipal, autonomous region) government data open platforms. First, LDA model is used to mine user demand themes in the initial construction period and development period of the platform to analyze their demand hotspots. Second, naive Bayesian algorithm is selected to study the emotional orientation of user demand themes. Finally, the similarity of platform user demand themes in the two periods is calculated to reveal the dynamic evolution path of user incremental demand. The research results show that "data improvement", "social security", "public security" and other demand topics are areas of continuous concern for users. Among them, the emotional probability of "social security" topics is 0.75, the emotional tendency of users is positive, the emotional probability of "data improvement" topics is 0.22, and the emotional tendency of users is negative. This thesis studies the theme, emotional appeal and demand change of the platform users, promotes the continuous improvement of the government data open platform according to the incremental needs of users, and provides strong support for the construction and development of the platform.

Keywords: Government Data Open Platform; User Incremental Demand; Topic Mining; Emotional Tendency; Text Similarity; Theme Evolution

（收稿日期：2022-11-21）

会议报道

第十三届全国知识组织与知识服务学术研讨会圆满闭幕

开放科学运动下，新的知识生态正在形成，知识组织与知识服务的模式、结构与方法亟需更新。2022年12月22日，中国科学技术信息研究所联合国家图书馆、国家科技图书文献中心、中国科学技术情报学会，以“开放视角下的知识系统重构与服务”为主题，召开“第十三届全国知识组织与知识服务学术研讨会”。会议以线上直播方式进行，由《数字图书馆论坛》编辑部协办。中国科学技术信息研究所信息资源中心主任潘云涛、国家图书馆中文采编部主任王洋担任主持。

中国科学技术信息研究所党委书记、所长赵志耘和国家图书馆馆长熊远明、国家科技图书文献中心主任许倞、中国科学技术情报学会理事长戴国强分别为会议致辞。中国科学技术信息研究所研究馆员常春、国家图书馆古籍馆副研究馆员赵大莹、武汉大学信息管理学院教授吴江、中国科学院文献情报中心研究馆员张智雄、中国人民大学信息资源管理学院教授贾君枝、中国科学技术信息研究所研究员张运良、福州大学经济与管理学院教授成全、安徽财经大学管理科学与工程学院教授魏瑞斌共8位专家结合各自的研究领域作了精彩报告，内容涉及知识组织的技术方案、知识分类的数据模型、知识工具的创新应用、知识挖掘的方法手段、知识组织系统的新发展趋势以及知识服务的实践案例等。

本届研讨会的召开有助于进一步推动知识组织与知识服务的升级优化，促进知识的共享、传播、利用和增值，对于加快实施创新驱动发展战略、提升自主创新能力具有重要意义。