doi:10.3772/j.issn.2095-915x.2019.03.001

开放科学
（资源服务）
标识码
（OSID）

用户参与视角下微博用户兴趣差异识别

周海晨 李永明

南京农业大学信息科技学院 南京 210095

摘要：深入了解不同参与维度用户对微博服务的兴趣差异，对于各行各业更好地利用新媒体为用户提供个性化服务具有重要意义。以高校图书馆微博为例，采集世界一流大学建设高校图书馆微博的全样本数据，借助文本挖掘技术和可视化工具，分析不同参与维度下用户关注的微博语义特征。从微博文本中识别出反映各维度用户的兴趣特征词汇，并且成功获取不同参与维度用户的兴趣关注点：活跃用户关注于有奖竞赛、参考咨询、电子资源、交流分享，普通用户则更关注舆论热点、福利体验、活动投票、考研毕业。

关键词：用户参与；微博；文本挖掘；差异识别；高校图书馆

中图分类号：TP393 G250

Difference Identification of Micro-blog User's Interest from the Perspective of User Participation

ZHOU Haichen LI Yongming

College of Information science & Technology, Nanjing Agricultural University, Nanjing 210095, China

Abstract: Understanding of user's interest and preferences for microblog services has important value for the industry to make better use of new media to provide personalized services for users. Taking the library Weibo as an example, the full sample data of the microblog of university libraries in a world-class university were collected. Combining text mining techniques and visualization tools, the research analyzed differences in posts and comments of microblogs under different

---

基金项目：国家留学基金委公派出国留学项目（201806850077）；江苏省研究生科研与实践创新计划项目（KYCX18_0740）。

作者简介：周海晨（1993-），博士研究生，研究方向：社交媒体、自然语言处理应用，E-mail：haizenchow@gmail.com；李永明（1989-），博士研究生，研究方向：信息服务与用户。

participating dimensions. The study found that there are feature vocabularies in the library microblog and comment text reflecting the differences in the topics, and successfully identifying the focus topics of users with different participation dimensions. Active users pay attention to prize competition, reference service, electronic resources, social interaction, and ordinary users are more concerned about public opinion, welfare, voting event, postgraduate examination and graduation.

Keywords: User participation; micro-blog; text mining; difference identification; university library

引言

微博作为开放、便捷、高效的信息分享、传播及获取平台，吸引了数以亿计的个人参与其中。庞大的用户群体，致使众多企事业单位、政府、高校、媒体纷纷开通官方微博，利用这一基于用户关系的交互式信息交流平台为用户提供服务。微博数据中心发布的《2018 微博教育行业报告》[1] 指出：仅教育行业蓝 V账号数量便达到了 3.7 万，整体教育兴趣用户高达 1.48 亿。假设一个用户只关注一个教育行业蓝 V 账号，平均每个账号也能拥有 4000 名粉丝。面对如此庞大的兴趣用户群体，微博运营者该如何把握机会，制定出合理的粉丝管理策略，从而有效的将粉丝用户转化为忠诚用户，成为当下亟待探索的热点话题。围绕这一问题的思考，可以从微博服务方式、服务理念、服务内容等角度切入。但无论何种角度，均离不开用户为导向的发展理念，离不开对用户兴趣的精准把握与识别。

泛在网络环境下，用户导向的服务与发展理念，赋予了微博用户更丰富多元的内涵和新的定位。用户作为微博服务的利益相关者，参与微博服务已经成为各界共识。为了更好地了解用户对微博服务的需求和偏好，从用户参与视角研究微博，识别不同参与维度下（活跃、普通）用户的兴趣差异，对于各行各业利用微博为用户提供个性化服务，提升服务质量、增加关注用户数量、维持用户粘性以及用户满意度具有重要意义。

1 研究综述

新媒体已经成为各行各业开展线上服务的重要手段 [2]，微博作为一种重要的新媒体服务，不仅可以利用微博发布消息、共享资源，还可以与用户交流、互动 [3]。企业、政府等在利用微博提供服务的同时，可以更好地了解用户需求和偏好，有针对性地开展个性化服务。目前，国内外学术界都非常重视对微博（以 Twitter 和新浪微博为主）的应用 [4-5]，相关研究主要集中在微博的应用现状调查 [6]、发展策略 [7-8]、服务现状调研 [9-10]、影响力测评 [11]、营销推广 [12] 等几个方面，这些研究局限于微博应用的探讨。

微博作为基于 Web2.0 技术的新媒体，用户是其存在和发展的根基。用户参与也是 2.0 的要素之一 [13]。用户参与是拉近用户距离、了解用户需求的有效途径。有关用户参与微博的研究，存在基于用户参与（关注、浏览和互动）的微博社区构建策略 [14]、微博关注数据的关联

doi:10.3772/j.issn.2095-915x.2019.03.001

分析[15]、微博用户参与行为研究[16]等少量研究，更多地关注用户参与行为（浏览、点赞、转发、评论等），对不同参与维度下微博及其评论数据的主题挖掘研究甚少，同时，已有研究的样本量也不足[17]。

随着数据种类不断丰富、计算能力不断提升，出现研究者采用多种技术开展以微博用户兴趣的研究。如袁园探索了微博用户关注兴趣中的规律[18]，何黎基于用户信息与关系数据总结出核心用户的兴趣特征[19]，唐晓波构建出针对沉默用户的兴趣模型[20]等。尽管当下围绕用户兴趣的研究取得了一定进展，但研究者普遍未对用户的参与深入程度进行区分，或只分析单一的用户群体。缺乏对不同参与维度用户兴趣的对比分析。

综上所述，作者引入用户参与，对活跃用户与普通用户产生的微博文本进行对比分析，识别用户主题差异。本文选取“双一流”高校图书馆微博为实证对象，通过 Python 编写爬虫程序采集微博数据，使用 TextRank 算法抽取关键词，按照用户参与维度划分用户群体，分析不同参与维度下用户关注的正文主题差异与评论差异，为微博管理运营者更深入地了解用户需求和偏好，更好地借助新媒体为用户提供个性化服务提供参考和借鉴。

2 技术路线

为了更好地探究活跃用户与普通用户可能存在的兴趣差异，本文提出了基于用户参与视角下的微博用户兴趣差异研究的技术路线（见图 1），希望探索出一条集数据采集、文本挖掘、可视化分析为一体的微博主题差异识别方法。

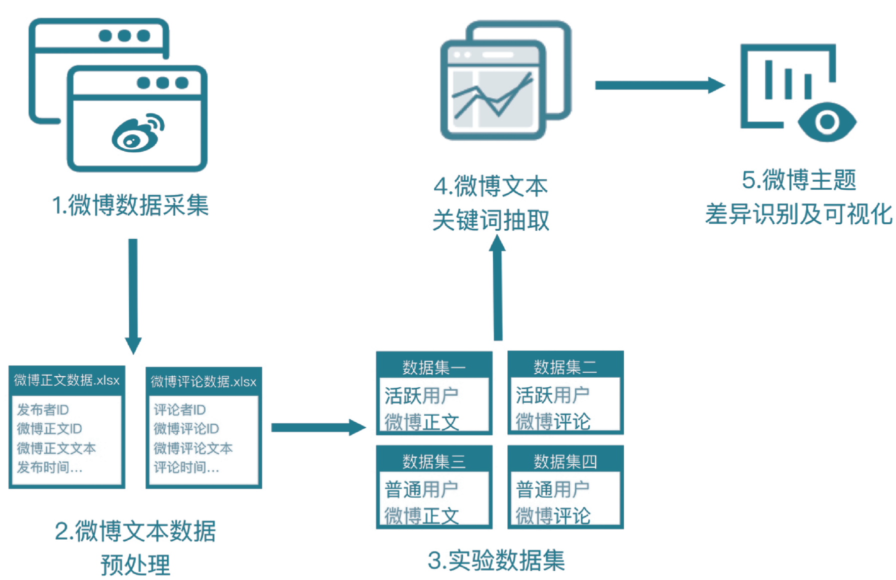

图 **1** 微博兴趣差异识别的技术路线

以下为用户参与视角下微博用户兴趣差异识别的具体步骤：

2.1 数据采集

2017 年 9 月 21 日，教育部、财政部、国家

发展改革委联合发布《关于公布世界一流大学和一流学科建设高校及建设学科名单的通知》，本文选取“双一流”中的世界一流大学建设的42所高校图书馆微博作为研究样本。通过网络调研发现，37所高校图书馆已开通了微博账号，中国海洋大学图书馆、华中科技大学图书馆、国防科技大学图书馆、云南大学图书馆和新疆大学图书馆这5所高校未开通。2017年9月30日，设计网络爬虫采集37所高校馆发布的所有微博及其评论数据，数据时间段为2011年8月至2017年9月，即37所高校馆从第一家开通微博到数据采集时间节点。

2.2 数据预处理

数据采集完成后，将获取到的数据导出为xlsx文件，并对数据进行去重、清洗、关联以及纠错。关联是指将微博正文数据与微博评论数据通过相同的字段进行关联，得到完整数据集。经过文本预处理，共得到微博正文数据58829条，微博评论数据71288条。其中，同一用户在不同高校馆微博下的评论分开统计，如用户“书骨精”在清华大学图书馆微博中评论7次，北京大学图书馆微博中评论6次，统计为两位用户。

2.3 实验数据集划分

（1）用户参与维度划分

有关用户参与的定义有很多种，本文借鉴A.M. Klem等[21]的研究成果，从参与行为视角研究用户参与。国内学者张宏涛[22]认为用户参与是行为特征，具体表现为时间、精力、智力、情感等方面的投入行为；马凌云[23]认为用户参与是用户对信息资源以及各项服务的关注、利用和分享的行为。综合已有成果，本文从行为特征方面划分用户参与维度。

结合二八定律，本研究假设20%的用户贡献了80%评论。对数据进行初步统计发现，当设定筛选条件为用户总评论数大于3条时，共得到5477条数据，约占微博评论用户总数的19%；剩余的用户总评论数小于等于2条，共23306条数据，约占81%。因此，本文根据用户评论数量将用户参与分为高度参与和低度参与两个维度，对应用户群体为活跃用户和普通用户，即高度参与用户（活跃用户）可界定为评论次数大于等于3的微博用户，低度参与用户（普通用户）界定为评论次数小于等于2的微博用户。也就是说，高度参与用户（活跃用户）约占总体用户数量的20%，低度参与用户（普通用户）约占80%。

（2）数据集划分

本文根据用户参与维度，将处理好的微博数据划分为四个集合，分别为：

数据集一，活跃用户微博正文文本；

数据集二，活跃用户微博评论文本；

数据集三，普通用户微博正文文本；

数据集四，普通用户微博评论文本。

2.4 文本关键词抽取

本文使用TextRank算法[24]计算出每个数据集中关键词的重要程度，生成关键词序列表。该算法的核心思想来自Google创始人Larry Page的网页排名算法PageRank[25]其将整个互联网视为一张有向图、网页为节点。TextRank则将拆分后的文本，即词汇，作为网络节点，

doi:10.3772/j.issn.2095-915x.2019.03.001

组成词汇网络图模型。该算法的主要优势体现在无需提前训练语料，因此较易应用于各种不同语言或领域。其主要步骤如图 2 所示。整体流程中，分词与去停用词的效果直接影响着 TextRnk 算法的计算结果。本研究使用了包括百度、搜狗、腾讯等企业提供的词典以及哈尔滨工业大学、四川大学等高校采用的停用词表，同时结合语料完善出未登录词，进一步提高分词效果。分词与去停用词后，需要对词语进行词性判断，保留名词或名词性短语作为关键词候选词。

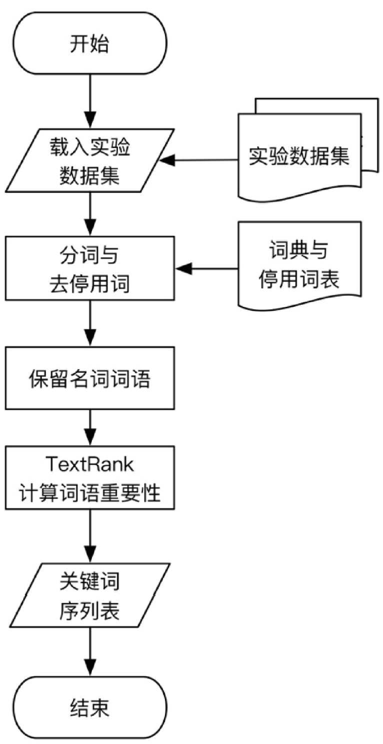

图 2 基于 **TextRank** 的关键词生成流程

2.5 差异识别与可视化

根据用户参与高低两个维度与微博正文、微博评论两类语料可得到四张表，分别为活跃用户微博正文关键词序列表、活跃用户微博评论关键词序列表、普通用户微博正文关键词序列表、普通用户微博评论关键词序列表。通过比较不同用户参与维度下微博正文与微博评论的关键词，分析关键词异同、排名位置变化，结合原始数据进行考察，可迅速识别出活跃用户与普通用户兴趣差异。

3 数据分析

3.1 微博兴趣差异识别

使用 TextRank 算法对于四个数据集分别进行计算，获得的重要性结果呈现为取值区间从 0 到 1 的实数。某一词汇数值越大则表明该词汇对于文本的重要性越高。根据关键词排名分别绘制出活跃用户与普通用户相应的微博正文与微博评论关键词对比图（由于版面有限只列出排名前 50 的关键词），左侧为活跃用户，右侧为普通用户。数字代表关键词在表中排名，连线两端为相同关键词。根据相同关键词在左右两边的排名相减，获得关键词排名差。当某关键词排名差较大时，表明该关键词所代表的兴趣在活跃与普通两类用户中的关注程度存在较大差异。反之，当排名差较小时，可以认为相应兴趣在两类用户中有着相似的关注度。在此基础上，通过分析可以归纳出文本反映出的重要兴趣差异。

（1）微博正文兴趣差异识别

表 1 为活跃微博正文关键词排名。同样地，可得到普通用户微博正文关键词排名表。基于微博正文关键词排名，可分别绘制出不同参与维度下用户关注的微博正文关键词对比图（见图 3）。

表 **1** 活跃用户微博正文关键词排名（部分）

排名	关键词	TextRank值	排名	关键词	TextRank值
1	网页	1.0	11	校区	0.219
2	链接	0.941	12	活动	0.219
3	同学	0.523	13	资源	0.208
4	读者	0.485	14	礼品	0.184
5	图书	0.316	15	讲座	0.182
6	开放	0.274	16	全文	0.179
7	时间	0.252	17	阅览室	0.178
8	通知	0.249	18	访问	0.171
9	数据库	0.235	19	大家	0.168
10	服务	0.225	20	分享	0.167

首先，图 3 列出 56 个关键词，其中 44 个词在两侧均出现，属于重叠词，12 个词只出现了一次，属于特征词，分别为：“安排”、“网址”、“抢答”、“电子”、“搜索”、“一楼”、“试用”、“毕业”、“学习”、“厦大”、“欢迎”、“二楼”。

其次，从排名差角度可知，排名差在 0 到 4 以内的关键词共有 23 个，约占重叠关键词的 52%。排名差在 5 或 5 以上的关键词共有 21 个，约占 48%。若将排名差处于 0 到 4 之间的关键词定义为低度差异关键词，大于等于 5 的定义为高度差异关键词，则能发现在用户微博正文关键词中，高度差异关键词与低度差异关键词各占一半，侧面反映出不同参与维度的用户有着较大的关注差异。

最后，从排名角度而言，排列靠前的关键词排名高度一致，绝大部分词语排名没有发生改变，只有相邻的“时间”与“通知”二词互换了位置。表明包括开闭馆时间、数据库相关资讯、读者服务等内容是所有用户重点关注的三大主题，也是图书馆微博提供的基础性服务内容。

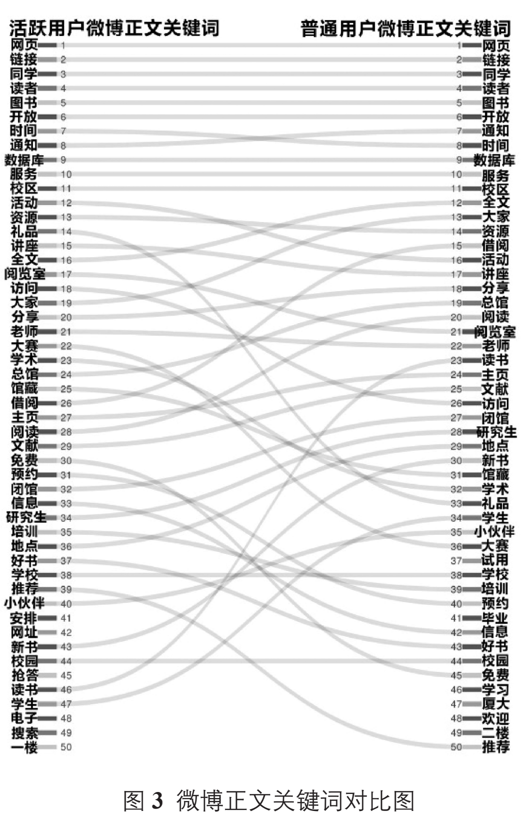

图 **3** 微博正文关键词对比图

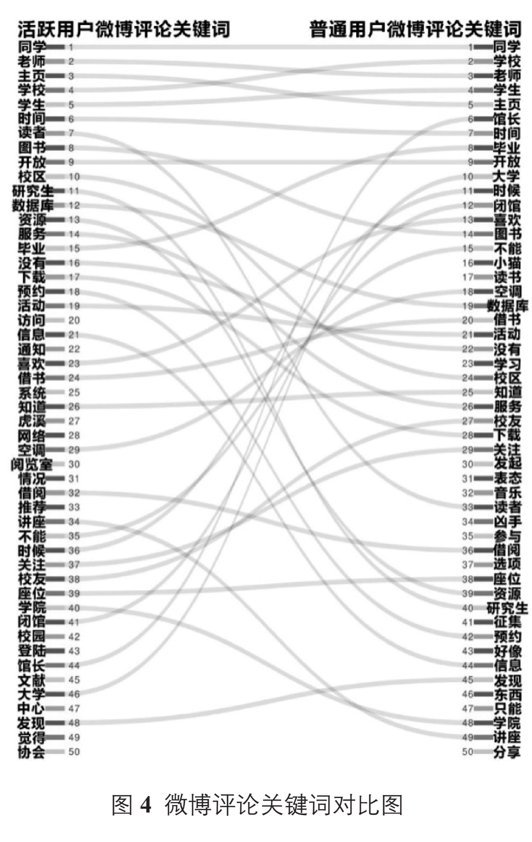

图 **4** 微博评论关键词对比图

doi:10.3772/j.issn.2095-915x.2019.03.001

（2）微博评论兴趣差异识别

根据不同用户参与维度下的微博评论关键词排名，绘制出活跃用户与普通用户微博评论关键词对比图（见图 4）。不同用户参与维度下微博评论数据共出现了 64 个关键词，比图 3 以微博正文为文本的关键词数多了 8 个。微博用户相较高校馆发布者而言群体更为庞大，语言表述更为丰富多样，因此活跃用户与普通用户在微博评论关键词上也会存在更多特征词，这些词汇也是深入研究不同层次用户参与过程中兴趣差异的重点词汇。

排名差在 0 到 4 以内的关键词共有 13 个，约占重叠词的 36%。排名差在 5 或 5 以上的关键词共有 23 个，约占 64%。对比图 3 中微博正文关键词可知，微博评论关键词在不同参与维度的用户中存在更为显著的兴趣差异。

3.2 微博用户兴趣分析

为直观的展示活跃用户与普通用户兴趣差异及强度，本研究基于前一阶段计算得到的重叠词排名差数据，绘制出微博用户参与主题偏好柱状图。其中图 5 横轴为微博正文中前 100 的重叠词，图 6 横轴为微博评论中前 100 的重叠词。每一个重叠词所对应的长度，代表了不同层次用户间关注差异情况。差值的绝对值越大，表明关注差异程度越大，负数代表活跃用户重视程度更高，正数代表普通用户重视程度更高。

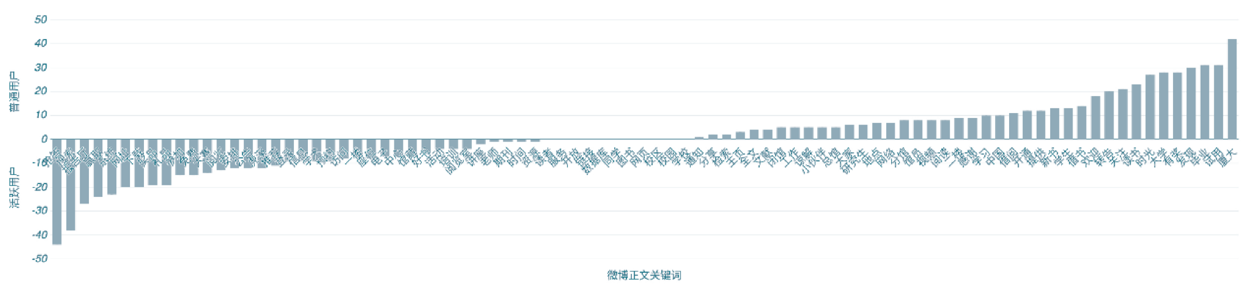

图 5 微博正文用户兴趣强度柱状图

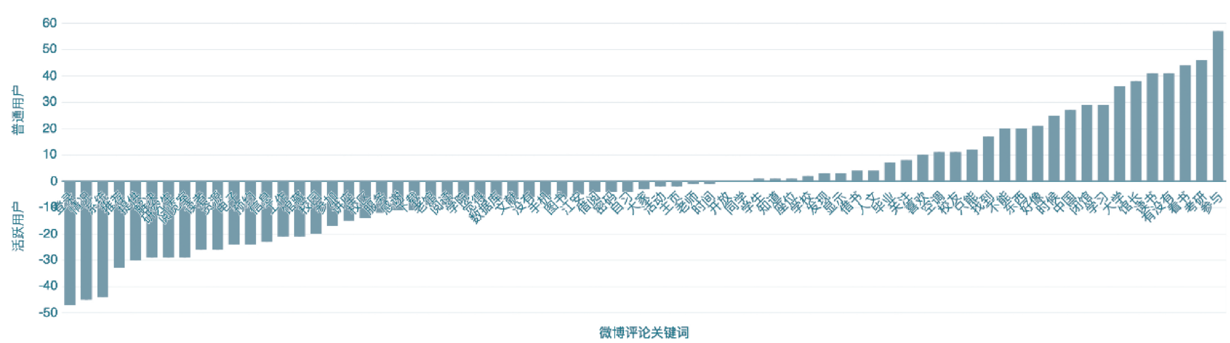

图 6 微博评论用户兴趣强度柱状图

总体而言，重叠词间差异呈递进态势，未观察到较为明显的断层现象，体现出活跃用户与普通用户间兴趣差异并非泾渭分明而是在拥有共同兴趣的基础上，有着不同程度的偏好。

（1）特征差异分析

还存在部分特征词如表 2 所示，特征词

为某类用户所特有，因此比重叠词具有更强的偏好指向性，能直接反应出特定用户的兴趣。可以看出，微博评论特征词数量远大于微博正文特征词，且从微博评论特征词中也更能体现出用户兴趣。如小猫、凶手、可惜、变态、天堂这 5 个仅出现在普通用户微博评论中的特征词，直接指向了清华大学流浪猫遇害事件。

表 2 活跃用户及普通用户特征词表（部分）

活跃用户	微博正文特征词	移动硬盘、挑战赛、咨询、晚安、提醒、领取
	微博评论特征词	访问、通知、虎溪、网络、中心、协会、咨询、一楼、电脑、学生会、网站、阅览、二楼、可能、申请、检索、小图、办公室、抱歉、校园网、登陆、提醒
普通用户	微博正文特征词	好友、毕业生、烫死、体验、书籍、福利
	微博评论特征词	小猫、发起、表态、音乐、凶手、选项、征集、分享、期待、厦大、记得、免费、可惜、投票、变态、图书馆、好书、世界、毕业生、天堂、书籍、高校、视频

综合图 5 图 6 以及表 2 进行分析，最终结果如表 3 所示。

（2）内容差异分析

通过表 3 可以看出活跃用户在图书馆开展的阅读推广、信息检索等主题活动中表现积极，主要原因是活跃用户自身拥有提升信息素养的愿望以及奖品的物质激励。对于参考咨询、数据库资源、校外访问、书籍与座位预约等图书馆核心资源与服务有着强烈的需求。同时更注重与高校馆及用户间的日常交流互动，如相互道晚安、转发名人名言等。

表 3 活跃用户与普通用户兴趣

用户类别	用户兴趣
活跃用户	有奖竞赛、参考咨询、电子资源、交流分享
普通用户	舆论热点、福利体验、活动投票、考研毕业

相较活跃用户，普通用户则更容易受到舆论热点事件的影响，如清华大学图书馆内一只深受喜爱的流浪猫遇害，清华大学图书馆微博第一时间发布了相关消息，在短时间内引起了众多在校师生前来评论哀悼，同时表达了对凶手的强烈谴责。对于图书馆开展的电子阅读设备体验申请、闭馆音乐投票等较为轻松愉悦、无需投入过多精力的主题活动更为容易接受。图书馆作为研究生考试备考的主要场所，考研相关主题的内容也是许多普通用户会重点关注的内容。此外，毕业季是众多师生离愁别绪之时，图书馆作为大学生活的重要场所之一许多普通用户会在此时与高校馆微博展开互动。

4 结语

通过邀请高校图书馆微博运营者对微博兴趣差异识别结果进行检验，与实际情况较为相符，证明了方法在实践中的切实可行。本研究从用户参与视角，基于文本挖掘中的重要算法 TextRank，对世界一流大学建设高校图书馆微博数据进行分析。通过用户评论数量区分出活跃用户与普通用户，对二者数据分别使用自动化、可视化分析方法，成功的识别出了高校馆

doi:10.3772/j.issn.2095-915x.2019.03.001

不同用户群体的兴趣偏好。本方法可以为舆情检测、社交媒体运营等专职人员提供高效、便捷的微博用户兴趣差异分析手段。

本研究的局限性在于用户参与维度的区分仅考虑评论数量，且基于图模型的 TextRank 算法抽取关键词的效果有限，增加了分析人员处理负担。在接下来的研究中，将会设计更为科学全面的用户区分方法；并适当考虑结合当下广泛采用的机器学习技术 [26]，提升关键词识别效果，以期得到更为准确的分析结果。

参考文献

[1] 微博数据中心 . 2018 微博教育行业报告 [EB/OL]. [2018-10-21]. http://data.weibo.com/report/reportDetail?id=414.

[2] 吴建中 . 新常态新指标新方向 (2012 中国图书馆年会主旨报告)[J]. 图书馆杂志 , 2012(12):2-6.

[3] 马爱华 . 论公共图书馆的新媒体服务 [J]. 图书情报工作 , 2014, 58(6):70-74.

[4] 张耀坤 , 邢文涛 . 基于微博内容分析的高校图书馆网络形象研究 —— 以厦门大学图书馆新浪微博为例 [J]. 图书馆研究 , 2016, 46(6):1-7.

[5] American Library Association. State of America's Libraries Report 2014[R]. [2018-05-22]. http://www.ala.org/news/sites/ala.org.news/files/content/2014-State-of-Americas-Libraries-Report.pdf

[6] 吴朝红 . 国内图书馆微博应用现状调查与发展对策研究 [J]. 图书馆工作与研究 , 2013, 1(3):44-47.

[7] 郑满生 . 3G 时代图书馆的微博应用现状及发展建议 [J]. 图书馆 , 2013(1):118-120.

[8] Shulman J, Yep J, Tomé D. Leveraging the power of a Twitter network for library promotion[J]. The Journal of Academic Librarianship, 2015, 41(2):178-185.

[9] 唐琼 , 袁媛 , 刘钊 . 我国高校图书馆微博服务现状调查研究 —— 以新浪认证用户为例 [J]. 大学图书馆学报 , 2013, 31(3):97-103.

[10] Loudon L, Hall H. From triviality to business tool: The case of Twitter in library and information services delivery[J]. Business Information Review, 2010, 27(4): 236-241.

[11] 王冰 . 高校图书馆微博影响力测评研究 [D]. 郑州 : 郑州大学 , 2015.

[12] 郭文玲 . 高校图书馆阅读推广专题微博之建设 —— 以 @ 大学生阅读分享平台为例 [J]. 图书情报工作 , 2015(20):23-28.

[13] Maness J M. Library 2.0 theory: Web 2.0 and its implications for libraries[EB/OL]. [2018-10-21]. http://www.webology.org/2006/v3n2/a25.html.

[14] 金玲娟 , 朱松挺 . 基于用户参与的图书馆微博社区构建策略研究 [J]. 新世纪图书馆 , 2015(4):61-64.

[15] 程秀峰 , 李重阳 , 陈莉玥 . 基于关联规则的高校图书馆微博关注趋势分析 [J]. 图书情报工作 , 2014, 58(8):73-78.

[16] 刘国敏 . 图书馆微博社区的用户参与行为研究 [J]. 图书馆论坛 , 2014, 34(1):57-61.

[17] 邢文明 , 杨玲 , 陈继丽 . 中美高校图书馆微博应用比较研究 —— 基于新浪微博和推特的分析 [J]. 图书情报工作 , 2017, 61(19):74-79.

[18] 袁园 , 孙霄凌 , 朱庆华 . 微博用户关注兴趣的社会网络分析 [J]. 数据分析与知识发现 , 2012, 28(2):68-75.

[19] 何黎 , 何跃 , 霍叶青 . 微博用户特征分析和核心用户挖掘 [J]. 情报理论与实践 , 2011, 34(11):121-125.

[20] 唐晓波 , 梁梦婕 . 融合结构与内容特征的微博沉默用户兴趣模型构建研究 [J]. 情报学报 , 2015, 34(11):1214-1224.

[21] Klem A M, Connell J P. Relationships matter: Linking teacher support to student engagement and achievement[J]. Journal of school health, 2004, 74(7):262-273.

[22] 张宏涛 . 用户参与图书馆服务策略研究 [J]. 图书馆工作与研究 , 2012(3): 48-50.

[23] 马凌云 , 康红 . Web2.0 环境下图书馆用户参与的分析与评价 [J]. 图书馆理论与实践 , 2010(4):15-17.

[24] Mihalcea R. TextRank: Bringing order into texts[J]. Emnlp, 2004:404-411.

[25] Page L. The PageRank Citation Ranking: Bringing Order to the Web[J]. Stanford Digital Libraries Working Paper, 1998, 9(1):1-14.

[26] 曾金 , 陆伟 , 陈海华 , 等 . 基于多模数据的微博用户兴趣识别研究 [J]. 情报科学 , 2018, 36(1):124-129.