

多维度社交网络舆情 用户群体聚类分析方法研究

王晰巍 贾若男 韦雅楠 张 柳

[摘 要] [目的]通过舆情用户群体聚类为舆情监管部门和社交网络服务提供商定位用户群体特征、实施有针对性的管控措施提供新的视角和思路。[方法]以群体理论为基础,从用户的影响力特征、情感特征和行为特征出发进行聚类,通过采集新浪微博平台用户数据,利用 Canopy、K-Means 算法进行聚类,最终通过 Neo4j 和 Weka 进行可视化呈现。[结果]聚类结果表明,同一舆情事件的用户群体在情感、影响力和行为等方面存在差异,不同舆情事件的用户群体在上述方面也会存在相同点。[局限]两事件均为高校舆情事件,并且仅以新浪微博平台作为数据来源。[结论]根据聚类结果可针对相同舆情事件和不同舆情事件中的各个用户群体提出对应的管控策略。

[关键词] 多维度;社交网络;舆情;用户群体;用户聚类

[作者简介] 王晰巍,吉林大学管理学院,吉林大学大数据管理研究中心,吉林大学网络空间治理研究中心;贾若男(通讯作者)(ORCID:0000-0002-4262-7982),吉林大学管理学院,E-mail:2943442131qq.com;韦雅楠,吉林大学管理学院;张柳,吉林大学管理学院(长春 130022)。

[原文出处]《数据分析与知识发现》(京),2021.6.25~35

[基金项目] 本文系吉林大学国家发展与安全(生物安全)专项研究课题(项目编号:2020JDGFAZ003)和吉林大学研究生创新基金资助项目(项目编号:101832020CX057)的研究成果之一。

1 引言

据《2020-2021 年中国移动社交行业研究报告》数据显示,中国社交行业用户规模突破9亿人,其中微博是目前市场主流应用^[1]。微博作为中国最具影响力的社交网络平台,垂直覆盖领域达到60个^[2]。随着互联网普及率和网民数量的迅速增长,以微博为代表的社交网络平台已经逐渐成为突发事件舆情产生、发酵、演变和传播的核心阵地。社交网络舆情事件的明显特征是具有用户的群体性,当持某一立场的网民群体达到一定规模时,群体的意见倾向表达不仅能够感染和同化群体外部的网民,甚至可能会影响舆情事件的演进发展态势。因此,对于社交网络舆情中用户群体的聚类特征进行分析,是社交网络舆情监管和控制的重要切入点。

目前,国外在此领域的研究主要关注社交网络舆情中的信息传播、观点挖掘或情感分析等方面,并多利用数学建模、超网络、复杂网络动力学模型和仿真模型等方法研究舆情事件的传播和演变。Garcia 等^[3]通过分析恐怖袭击发生后6万多名 Twitter 用户的数据,对社交网络中集体消极情绪反应的影响进行分析,为事后社交网络中社区危机管理提供了借鉴;Qiu 等^[4]指出随着 Twitter、Facebook 和微博等社交媒体高度普及,用户集群已成为越来越重要的研究主题;Liu 等^[5]将用户信息和网络全局信息共同纳入分析范围,提出一种用于社交网络中用户社群监测的划分和聚集算法(DA);You 等^[6]通过仿真实验,提出“中央节点识别-标签传播-社群组合”的三段式社群发现算法,为社交网络中的用户社群研究提供了新

的思路。目前国内对于社交网络舆情用户群体的分析和研究主要体现在用户画像构建、情感分析、社群和用户行为等方面。林燕霞等^[7]基于社会认同理论构建了微博群体用户画像,并依据不同群体用户的典型特征提出了相应的网络舆情治理建议;何高奇等^[8]提出了一个基于SIRS传染病模型改进的群体情绪感染模型E-SIRS,动态模拟了突发舆情事件下群体情感的变化趋势;张海涛等^[9]基于复杂网络理论,通过社群发现算法识别话题社群,对热点事件中的话题社群及网民情感波动进行研究;孙越恒等^[10]以复杂网络理论为基础,融合在线社交群体的多维特征构建了社交群体行为模式分析框架,对比分析了事件驱动下的社交群体行为规律。从国内外研究成果来看,用户集群的研究受到重视,目前从主题、社群监测、标签、群体情绪和社群发现等角度对用户群体进行分析的研究较为常见,但是,从多维度对用户群体特征进行聚类和多角度分析的研究较为少见,并且利用可视化图谱对用户特征进行展示和分析的研究也有限。

因此,基于目前的突发事件社交网络舆情发展背景和相关研究现状,本文拟解决以下问题。

(1)如何基于群体理论构建社交网络舆情事件中的用户群体聚类图谱模型?

(2)如何量化社交网络舆情事件中用户群体聚类图谱特征信息?

(3)多话题网络舆情事件的群体聚类图谱有何异同?

通过对相关文献的梳理,以群体理论作为理论基础,从情绪结构、权力结构和关系结构三个维度构建突发事件社交网络舆情用户群体聚类图谱,并对聚类后的群体情感结构、权力结构和关系结构进行可视化分析。本文力求为学界的社交网络舆情可视化图谱构建与分析提供新的研究视角和方法,同时为产业界建立健全社交网络舆情监管平台提供实践上的借鉴和参考。

2 理论基础

2.1 群体理论

群体理论(Group Theory)指的是社会心理学中探索和解释群体过程或群体动态及其相关问题的一般

理论观点^[11]。影响力较大的群体理论是法国社会心理学家古斯塔夫·勒庞提出的,主要从心理学角度考察群体问题,通过对组成群体的个体特征进行分析和研究,达到解释群体特征的目的。这些群体特征包括两个方面,即群体的组成和群体的结构。其中,群体的结构包括情感结构、权力结构(领导与被领导)和群体关系结构等^[12]。目前,群体理论在高校学生管理、个人社会化和大众传播等方面已经有了广泛的应用^[13]。

2.2 社交网络用户的聚类

随着Twitter、Instagram、微博、知乎等社交网络应用的大量兴起,社交网络用户的数量显著增加。社交网络用户聚类分析近年来已经逐渐成为计算机和图书情报领域研究的热点问题^[14]。目前,社交网络用户聚类研究主要从用户的行为数据和发布的内容数据出发,构建新的模型或方法对用户进行聚类分析。Liang等^[15]提出动态的用户集群主题模型(UCT),使用用户生成内容推断用户当前的分布,并通过实验验证了所提出的用户聚类模型与最新标准相比的有效性;Qiu等^[16]构建动态社交网络主题模型(DSM),通过在短文本中为用户创建具有动态特征和社交连接性的主题实现用户聚类,实验结果表明在带有社交网络短文本的用户聚类中,该聚类模型优于其他方法;Hu等^[16]利用用户标签、位置以及时间数据,提出一种基于非负矩阵三因子分解的协同聚类方法,并通过真实数据集验证了该方法对比其他聚类方法的有效性;Koc等^[17]将Twitter数据按照用户、用户讨论的问题和用户在讨论中使用的术语进行解释,并在这三个维度上采用协同聚类形成多个用户集群,通过实际数据集和合成数据集测试了算法以说明算法的适用性。

2.3 社交网络舆情用户群体聚类图谱构建的提出

在本文的研究中,将社交网络舆情用户界定为通过社交平台参与事件讨论、转发和点赞等过程的用户。由于社交网络和舆情事件传播的特殊性,面对同一舆情事件,不同用户的情感态度、影响范围、关系网络等都是不同的,通过对用户群体的有效聚类可以将行为、立场和情感等性质相近的用户归为一类群体,差异较大的用户个体划到不同的群体类别。通过图谱可视化工具可以更为直观地展示和分析出不同

用户群体的聚类特征,进而制定个性化、具体化和有针对性的舆情处理和调控策略,并对用户的舆情参与和传播行为进行预测,这对社交网络舆情事件的管控及引导具有重要的参考价值和实践意义。

3 社交网络舆情用户群体聚类分析方法

3.1 社交网络舆情用户聚类特征

社交网络舆情用户的群体聚类特征体现在用户情感特征、权力特征和关系特征三个方面。

社交网络舆情用户情感特征是指舆情事件中用户的情感类别,反映了群体的情绪结构。情感分类是指基于用户文本进行挖掘和分析,识别出其中的情感倾向性,并对情感进行分类,如褒义、贬义等^[18]。在舆情事件发生时,数量规模较大群体的情感倾向不仅会影响群内用户个体,还会对群外用户个体产生影响。因此,舆情用户的情感类别是聚类特征的一个重要维度,对于群体中情感倾向分布情况的及时把握和预测,有利于对每个用户的情感倾向进行有针对性的引导。

社交网络舆情用户权力特征是指舆情用户节点影响力的大小。由于社交网络的开放性,在网络舆情事件传播和发展的过程中,普通用户节点对关键用户节点或传播影响力大的用户节点具有一定的依赖性,会形成以关键节点或传播影响力大的用户节点为中心的群簇。关键用户节点和传播影响力大的用户节点对普通用户节点具有影响和引导作用,类似于群体内部的“领导”与“被领导”的关系,这也是群体理论中群体权利结构的体现。

社交网络舆情用户关系特征是指用户的点赞、评论和转发等动态数据信息,体现了社交网络舆情用户的关系网络,类似于群体中的人际关系结构。对于用户关系数据信息的分析,不仅能够揭示舆情事件传播过程中各个用户节点之间的关系网络,而且能够帮助舆情管理者更好地理解舆情信息的传播路径和方向,对舆情监控和舆情热点监测具有重要的意义。

3.2 分析方法

使用火车头脚本编辑爬虫软件实现数据采集。获取的数据通过 Excel、Access 等软件进行数据的预处理,包括筛选、去重、删除空白及无效文本、取消超链接等。本文基于群体理论、情感分析和聚类分析

等算法,构建了社交网络舆情用户群体聚类分析方法,具体方法过程如图1所示。

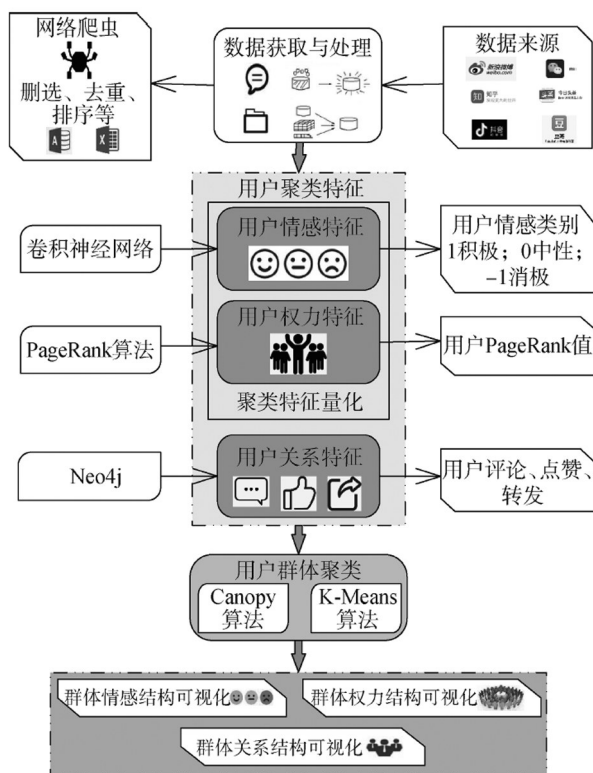


图1 社交网络舆情用户群体聚类分析方法

(1)群体情感特征

卷积神经网络(Convolutional Neural Networks, CNN)是一种包含卷积计算、深度结构的前馈神经网络,是一种典型的深度学习模型^[19]。近年来,越来越多的学者倾向于通过卷积神经网络解决文本情感分类问题^[20-21]。不同学者在进行情感分类研究时使用的研究标准是不同的,较为常见的是三级(正、负和中性)分类^[22-23]。因此,本文通过卷积神经网络模型进行三级分类,并在群体聚类过程中,用-1代表消极情感,0代表中性情感,1代表积极情感。

(2)用户权力特征

社交网络舆情用户影响力的测量方法多采用PageRank、中心性(Centrality)和入度(In-degree)等算法实现影响力大小的量化^[24]。PageRank算法是一种适合大规模网络全局排序的计算方法,在计算过程中考虑到了直接相邻和间接相邻节点的影响,能够反映节点的全局属性^[25],在影响力计算中得到广泛的应用^[26-27]。因此,本文采用PageRank算法计算用

户的影响力。

(3) 用户关系特征

在用户关系特征的构建和分析过程中,本文使用 Neo4j 图数据库作为工具。Neo4j 图数据库作为一种高性能的图形数据库,对比传统的数据库应用,可以更加高效地解决大量结构化较低、关系较复杂的数据信息,为数据的可视化和综合分析提供了较大的便利,在社交网络中有着广泛的应用^[21,28-29]。

(4) 用户群体聚类分析

Canopy 算法是一种快速近似的聚类算法,只要遍历一次数据即可得到聚类结果,但缺乏高精度^[30]; K-Means 算法是目前应用最广泛的、基于划分的聚类算法^[31]。K-Means 算法需要事先确定类簇的个数 K 和 K 个初始中心点,因而对于 K 值的确定直接影响最终的聚类结果和误差。针对 Canopy 算法和 K-Means 算法的不同特性,本文将 Canopy 算法和 K-Means 算法结合使用,将 Canopy 算法作为 K-Means 聚类的前奏,通过 Canopy 算法首先对数据进行“粗”聚类,为 K-Means 算法提供初始的 K 值,然后使用 K-Means 算法继续调节 K 值直至达到合适标准。同时,运用图谱工具对群体情感结构、权力结构和关系结构进行可视化分析。

4 数据采集和聚类分析

4.1 数据采集和预处理

随着活跃用户(为网站带来价值的用户)、头部用户(月阅读量超过 10 万的用户)的不断增长和积累,新浪微博已经成为众多社交网络平台中在线活跃人数最多和影响力最为广泛的社交媒体^[32],因而在社交网络的样本选择上,选择新浪微博作为数据采集的平台。同时,近年来校园安全和学术不端逐渐成为教育部和各高校院所的重点舆情管控对象,也引

起了社会各界的广泛关注。这两类事件虽同属于校园舆情,但事实本质不同,所引发的社会关注程度和网民情感表达存在差异,因而本文选取这两类高校舆情事件进行对比研究,以期为不同类型的舆情事件分析提供更有针对性的管控建议。

“北京交通大学实验室爆炸”和“南京大学 404 教授”事件在事发后较短时间内迅速登上各大话题热搜榜单,在社会各界引起了广泛而激烈的舆论关注,跻身高校十大热点事件^[33]。因此,在话题的选择上,本文选取新浪微博校园安全舆情热点话题“#北交大实验室爆炸#”和学术不端舆情热点话题“#南京大学 404 教授#”作为数据源进行数据的采集。数据字段主要包括用户昵称、ID 号码、点赞、转发及评论的内容、时间和工具端等。时间范围从 2018 年 10 月 24 日到 2019 年 11 月 2 日,共采集到原始数据 72 758 条,经处理后最终得到有效数据 70 005 条。

数据聚类分析和可视化分别采用新西兰怀卡托大学(The University of Waikato)开发的开源机器学习以及数据挖掘软件 Weka(Waikato Environment for Knowledge Analysis)^[34]和 NoSQL 图形数据库 Neo4j 实现。

4.2 聚类图谱及数据结果

(1) 聚类特征量化

用户情感特征的量化通过 Python 构建情感分类模型实现。主要经过输入层输入文本矩阵、卷积层提取特征、池化层最大池化、全连接层 Softmax 函数进行最终输出等过程实现情感分类。为了防止过度拟合,全连接层进行正则化输出优化,采用 Dropout 优化策略,Dropout 取值为 0.5。最终得到用户的情感分类及置信度如表 1 所示。

用户权力特征的量化采用 PageRank 算法实现。首先利用用户关系特征数据构建舆情事件传播网

表 1 用户情感分类结果及置信度(部分)

用户昵称	文本内容	情感分类	置信度
回头万里人常叹	真是讽刺,再次失望。	消极	0.87
diamondli99	……可悲,中国的博士,绝对弱势群体……	消极	0.84
小兜里满满的幸福	@中国大学生在线@中国教育在线考研频道@央视新闻	中性	0.94
秋桐小宅女	看调查结果。	中性	0.98
真无羽	逝者安息,实验都具有风险性,感谢他们为科学做出的贡献,安息。	积极	0.82
三七二十一个酥	……珍惜身边的每个人吧!……愿逝者安息,实验室安全警钟长鸣……	积极	0.88
放风筝的灰原哀	……学术诚可贵,生命价更高,一定要注意实验安全……	积极	0.70
叫我杏仁	岂止是难过,含辛茹苦培养出的人才,真的活不了了。	消极	0.70
CMLY、F	切记水火无情。	中性	0.73

络,节点代表用户,边代表用户间的评论和转发关系,然后将用户关系数据导入 Neo4j,利用 Cypher 语言调用 PageRank 算法,计算得到用户的 PageRank 值,如表 2 所示。

表 2 用户 PageRank 值(部分)

用户昵称	PageRank 值	用户昵称	PageRank 值
北京消防	1736.80	澎湃新闻	897.18
江宁公安在线	52.31	南京大学	719.89
中国消防	2.24	陈迪 Winston	343.35
北京交通大学	997.05	头条新闻	317.21
懒懒的周小姐	0.15	小姐姐爱学习	0.15
慎独明智	0.15	一只阿迟儿	0.15
KaiHugo	0.15	北欧 DJ	0.15

(2)群体聚类

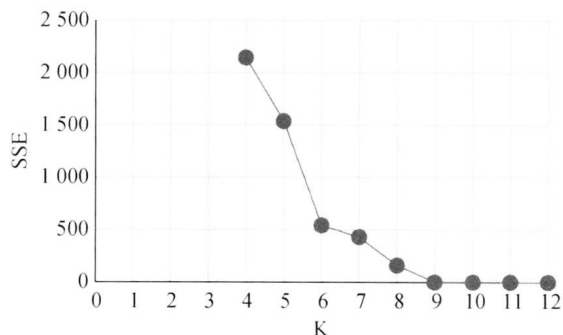
借助数据挖掘软件 Weka 实现聚类过程,将用户的情感特征数值、基于用户关系特征利用 PageRank 算法得到的影响力特征数值导入 Weka 中,调用聚类算法进行聚类过程。误差平方和(Sum of the Squared Errors, SSE)是所有样本的聚类误差,可以用来判断 K-Means 聚类效果的好坏^[35]。通过 Canopy 先确定初始的 K 值,再根据 K 值与 SSE 的变化曲线确定最终的聚类结果。两个事件的 SSE 变化趋势如图 2 所示,可

知当 K=9 时 SSE 值的变化最先趋于平缓状态,故最后确定的聚类系数 K 值为 9。

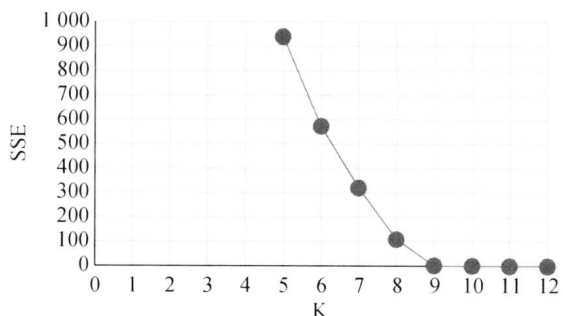
最后的聚类结果将节点划分为 9 个类簇,每个类簇所包含的节点数和占总节点数的比例如表 3 所示。可视化图谱如图 3 所示,横坐标为类簇编号(Cluster 0 ~ 8),纵坐标为用户编号(0 ~ N, N 为该事件中用户总数),以不同颜色表示不同类簇。

表 3 事件用户群体聚类结果

类簇	聚类结果	
	“北交大”事件	“南大”事件
0	9491(24%)	1675(6%)
1	2158(5%)	1485(5%)
2	14974(38%)	2729(9%)
3	1685(4%)	15144(50%)
4	3250(8%)	3360(11%)
5	5434(14%)	1446(5%)
6	458(1%)	2540(8%)
7	1457(4%)	1210(4%)
8	901(2%)	610(2%)

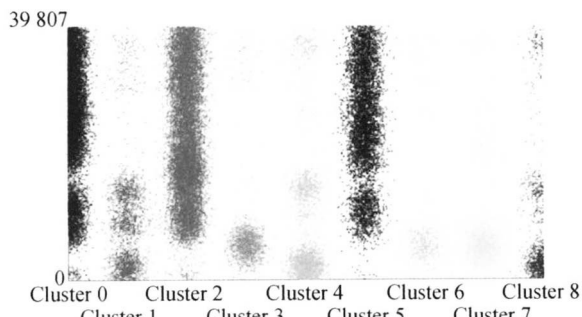


(a) “北交大”舆情事件

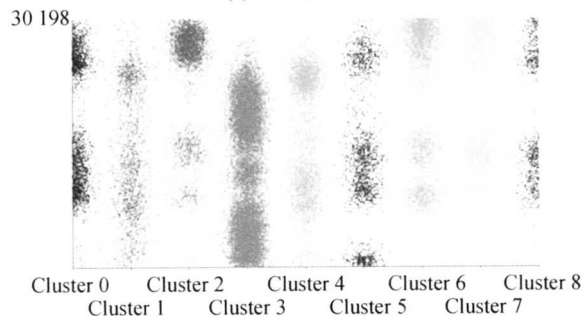


(b) “南大”舆情事件

图 2 舆情事件 SSE 变化趋势图



(a) “北交大”舆情事件



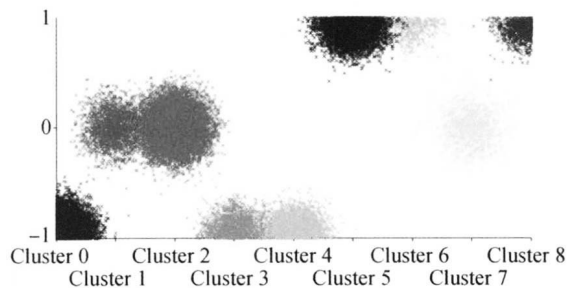
(b) “南大”舆情事件

图 3 舆情事件聚类结果

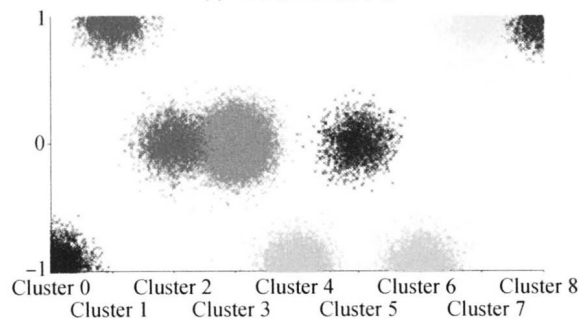
4.3 用户群体聚类图谱特征

(1) 用户群体情感特征

两事件用户情感结构如图4所示,其中纵坐标表示用户情感类别(-1表示消极情感,0表示中性情感,1表示积极情感),横坐标表示类簇编号(Cluster0~Cluster8),不同类簇以不同颜色表示。两类事件用户群体的情感总体上均呈现消极情感占比大于积极情感占比的情况。“北交大”事件中消极情感占比明显大于“南大”事件,原因与两事件的性质有关。“南大”事件属于学术不端事件,其中人为因素更加明显,网民意见绝大部分倾向于对事件当事人的声讨和谴责;“北交大”事件属于突发校园安全事件,网民发声的重要部分是表达对事故受害者的同情。“北交大”事件无辜受害者和“南大”舆情事件在用户的情感上形成鲜明的对比,广大网民对“北交大”事件无辜受害者形成同情的共意,在情绪的表达上也更为正面。



(a) “北交大”舆情事件

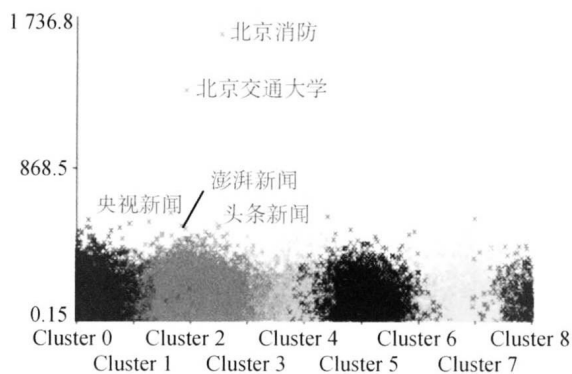


(b) “南大”舆情事件

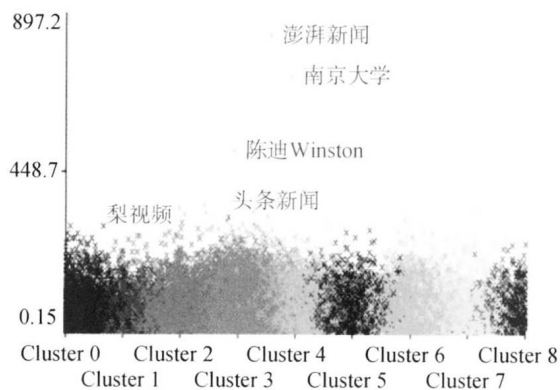
图4 各类簇情感分布

(2) 用户群体权力特征

两事件用户权力结构如图5所示,其中纵坐标表示用户的PageRank值,横坐标表示类簇编号(Cluster0~8),不同类簇以不同颜色表示。两类事件中用户群体的影响力呈现相同的分布,即绝大部分节点的



(a) “北交大”舆情事件



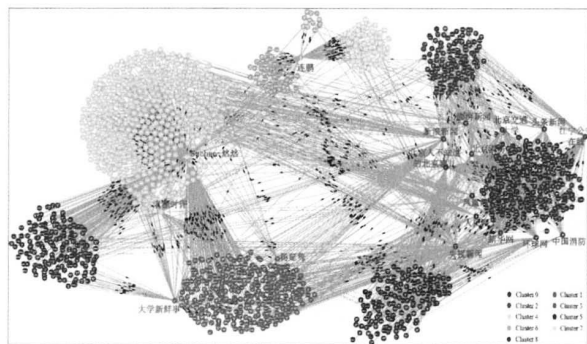
(b) “南大”舆情事件

图5 各类簇PageRank值分布

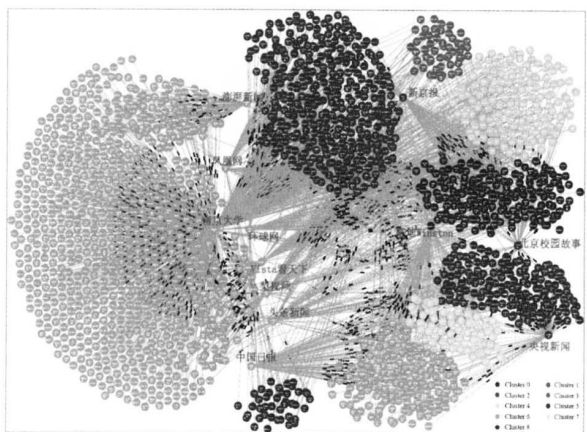
影响力是相同的并处在一个较低的水平,只存在有限个用户节点的影响力水平相对较高的情况。另外,两类事件中影响力较大的关键节点均出现聚集现象,呈现圈群化,如“北交大”事件中的Cluster2和“南大”事件中的Cluster3。并且在这些高影响力节点聚集的群落中,节点类型呈现多样化,除了高校官微还有政府官微、主流媒体和意见领袖节点出现。同时,这类群体内均有影响力最高的节点出现,如“北京消防”和“澎湃新闻”,这也与该群簇内影响力节点聚集有关,最高影响力的节点信息辐射范围广,能够吸引更多同类型的用户节点。

(3) 用户群体关系特征

由下页图6可知,每个类簇中均存在关系结构,且呈现“跨类簇”的用户节点。例如,“北交大”事件中的Cluster2和“南大”事件中的Cluster3,这样的群簇应该是舆情管理中重点需要关注的用户群体,因为他们不仅仅与本群体内部的用户节点间有联系,同时与其他类簇中的用户节点间也有直接关系。有部分



(a) “北交大”舆情事件



(b) “南大”舆情事件

图6 用户群体聚类关系

群体处于被影响的状态,如图6(a)中的Cluster 5~Cluster 8,以及图6(b)中的Cluster 4、Cluster 6、Cluster 7、Cluster 8。这类群体虽然与其他群体之间存在交叉关系,但是群体中没有影响力较大的关键节点,是在舆情事件中容易被其他群体影响和引导的用户群体。校园官微、政府官微和主流媒体大多数聚集在同一类簇中,如图6(a)中的Cluster 2和图6(b)中的Cluster 3;而意见领袖节点由于立场和观点较为个性化,受影响的用户属性不同,群簇分布较为分散,如图6(a)中意见领袖节点分散分布在Cluster 1和Cluster 4中,图6(b)中意见领袖节点分散分布在Cluster 0和Cluster 2中。

5 讨论分析

5.1 群体情感分析

公众舆情事件中,用户的情感与舆情事件的类型、媒体报道和群体归属等之间均存在相关性^[36]。高校舆情事件按照性质可以划分成多种类型,在这两类舆情事件中,消极情感均占上风,但由于事发原

因不同,公众的情感表达也不尽相同。有研究表明,对于人为因素较重、信息内容不透明、处理不利的事件,公众负面情感的传播和演变更加显著^[37],负面信息传播速度要比正面信息传播速度快^[38]。因此,在舆情的管控和引导过程中,应该重点关注呈现负面情感趋向的用户群体,同时充分利用积极情感群体的优势,引导整体情感朝着正能量、积极向上的方向发展。

在高校负面社交网络舆情事件中,公众多希望通过自身或群体情感的表达唤起某种行动,如了解更多与事件相关的信息、推动相关责任方对事件的进一步处理等。因此,在舆情事件中公众情感的管控上,可以从公众情感的二元化疏导上展开。一是通过高校官微和其他媒体的叙事报道调动公众的同情感,引发共情效应,将个体性意见逐步转化成同质的公众性意见,并通过拥有情感优势的群体进一步影响公众情绪,放大正面情感信息流,减少负面情绪的占比;二是高校在官方媒体的运营和建设过程中,应该转变自身定位,给予公众情感一定的回应和反馈,在日常的互动交流中传达人文情怀,拉近与公众之间的情感距离,从源头上削减负面情感,在事件爆发后使情感积极化和中性化,扭转舆情事件中消极情绪占上风的局势,同时防止事件的二次发酵和衍生舆情的出现。

5.2 群体权利分析

由于两类舆情事件的性质不同,舆情事件中关键节点的存在也呈现动态性、不固定性和事件依赖性,线上社群的发展不依赖于相同的意见领导者,涉及不同领导者之间的“接力”^[39]。同时,只有少数关键节点在两个事件中同时具备较高的影响力,有研究表明这些少数关键节点在向他人传播思想方面有着极强的说服力^[40]。对于那些高影响力节点聚集的群落,应该采取重点关注、重点引导的策略,鼓励节点形成意见圈群,扩大同质意见圈群的辐射范围。

舆情监管部门不仅要用动态的、发展的眼光看待社交网络中的关键节点及所在群体,同时还应探索不同事件中此类群体出现的规律,以便为后续同类事件的预防和处理提供参考。对于容易被其他群簇影响的用户群体,则主要利用关键节点进行引导,

扩大同质意见圈群的范围。此外,虽然两个事件中的关键节点是不同的,但高校官微均作为高影响力节点出现。因此,对于高校,平时应该注重校园媒体的建设。一是通过高校官方媒体平台宣扬正确、主流的价值导向,引导学生积极参与学校、社会间的良性互动,宣传和弘扬社会正气,提高学生的社会责任感、自我管理能力和信息辨别能力,在舆情事件爆发时降低学生被不良势力煽动、不理智、发表不当言论的可能性;二是通过高校媒体平台的宣传,提升高校软实力、树立良好的公众形象,增强高校在公众社交网络空间发表言论的公信力;三是校园突发事件发生后,应及时通过高校官方媒体平台发布事件相关信息,对舆情事件的动态、走向进行实时的参与和引导,掌握舆论的主动权。

5.3 群体关系分析

两类舆情事件中,均有“跨类簇”用户节点的出现,这类节点可以将信息从一个群簇传递到另一群簇,从而导致更高的传播性能和更广的传播范围^[41]。这类“跨群簇”节点聚集的群落应该是事件传播网络的核心,舆情的管控和引导离不开此类用户群体的贡献。另外,关键节点大部分聚集在同一类簇中,通过彼此之间相互关注、评论、点赞和转发,具有相同意见的用户组成意见社群,这与Li等^[42]的研究相一致。这些节点通过彼此的影响关系形成的“同质”意见圈群,对其他普通用户节点的影响范围会像“滚雪球”一样持续不断地扩大,因而此类群体也应该是有关部门重点关注的对象。

舆情管理部门和社交网络服务平台提供商应该重点关注“跨类簇”用户群体,从“治理”和“合作”两个方面展开。对于那些为了热度而歪曲事实、蓄意煽动的群体,可以联合社交网络运营商对其进行“限流”操作,阻断其信息传播途径,避免波及更大范围的社交网络用户,必要时可以采取法治手段,落实追责机制。而对于尊重客观事实、正向宣传引导的群体,可以采取联动策略,透过其在群落间的交叉联系多群体合作网络,并联合社交网络运营商增加“曝光度”,发挥该类型群体的群内和群外影响效应,推动“同质”意见网络群体的形成,起到传递正能量、同化消极情感倾向、过滤负面信息的作用。除了群体,对

于“跨类簇”节点自身,可以从“律己”和“律他”两个方面展开,“律己”是指要提高自身社会责任感,理性对待“热度”和“曝光度”;“律他”是指必要时可以采取删帖和拉黑等手段净化自己所辖的网络空间。

6 结语

理论层面,本文在梳理社交网络舆情和用户聚类相关研究的基础上,将群体理论与社交网络用户的聚类研究相结合,构建了社交网络舆情用户群体聚类图谱,为社交网络情境下舆情用户聚类特征研究提供了新的理论视角和可视化分析模型。实践层面,本文结合两种具有代表性的不同类型高校社交网络舆情事件进行实证对比分析,从用户群体情感结构、权力结构和关系结构的可视化直观地展现了网络舆情用户的群体聚类特征,对政府有关部门和社交网络平台服务提供商了解社交舆情事件中的用户群体聚类特征,进行舆情引导具有重要的现实意义。

本文选取的社交网络舆情事件虽然话题不同,但是均属于高校舆情的范畴,在以后的研究中将进行自然灾害事件、社会安全事件等多类型社交网络事件的不同类型话题的对比分析。另外,对于用户间的不同关系,如“点赞”“评论”和“转发”等在聚类中所占权重的界定和区分,也是未来深入研究的重要方向。

作者贡献声明:

王晰巍:提出研究思路和框架,论文修改;
贾若男:数据收集与分析,撰写和修改论文;
韦雅楠:英文文献收集与整理,文章校对;
张柳:数据预处理,英文文献收集。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储,E-mail:2943442131@qq.com。

[1]王晰巍,贾若男.北交大全部数据.xlsx.微博采集数据。

[2]王晰巍,贾若男.南大全部数据.xlsx.微博采集数据。

参考文献:

- [1]艾媒咨询.2020-2021年中国移动社交行业研究报告[EB/OL].[2021-05-18].https://www.sohu.com/a/442605711_533924.
- [2]新浪微博数据中心.2018微博用户发展报告[EB/OL].[2019-11-01].<http://www.199it.com/archives/847890.html>.
- [3]Garcia D, Rime B. Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack[J]. Psychological Science, 2019, 30(4): 617-628.
- [4]Qiu Z C, Shen H. User Clustering in a Dynamic Social Network Topic Model for Short Text Streams[J]. Information Sciences, 2017, 414: 102-116.
- [5]Liu Z Y, Ma Y H. A Divide and Agglomerate Algorithm for Community Detection in Social Networks[J]. Information Sciences, 2019, 482: 321-333.
- [6]You X M, Ma Y H, Liu Z Y. A Three-stage Algorithm on Community Detection in Social Networks[J]. Knowledge-Based Systems, 2020, 187: 104822.
- [7]林燕霞,谢湘生.基于社会认同理论的微博群体用户画像[J].情报理论与实践,2018,41(3):142-148.
- [8]何高奇,边晓晖,孙菲,等.基于传染病机制的突发事件下群体情绪感染模型[J].华东理工大学学报(自然科学版),2018,44(6):909-917,949.
- [9]张海涛,刘雅姝,张泉慧,等.基于模块度的话题发现及网民情感波动研究——以新浪微博“中美贸易摩擦”话题为例[J].图书情报工作,2019,63(4):5-14.
- [10]孙越恒,刘晓彤,王文俊.事件驱动的在线社交群体演化行为预测[J].情报杂志,2019,38(6):110-117.
- [11]顾明远.教育大辞典(增订合编本)[M].上海:上海教育出版社,1998.
- [12]古斯塔夫·勒庞.乌合之众:大众心理研究[M].冯克利译.北京:中央编译出版社,2005.
- [13]王泉,陈云奔.旁观者对校园欺凌影响及其纠偏策略——基于群体理论的分析[J].学习与探索,2019(3):44-48.
- [14]张海涛,唐诗曼,魏明珠,等.多维度属性加权分析的微博用户聚类研究[J].图书情报工作,2018,62(24):124-133.
- [15]Liang S S, Ren Z C, Zhao Y K, et al. Inferring Dynamic User Interests in Streams of Short Texts for User Clustering[J]. ACM Transactions on Information Systems, 2017, 36(1): 10.
- [16]Hu L, Xing Y H, Gong Y L, et al. Nonnegative Matrix Tri-factorization with User Similarity for Clustering in Point-of Interest[J]. Neurocomputing, 2019, 363: 58-65.
- [17]Koc S S, Ozer M, Toroslu I H, et al. Triadic Co-clustering

of Users, Issues and Sentiments in Political Tweets[J]. Expert Systems with Applications, 2018, 100: 79-94

- [18]王晰巍,张柳,文晴,等.基于贝叶斯模型的移动环境下网络舆情用户情感演化研究——以新浪微博“里约奥运会中国女排夺冠”话题为例[J].情报学报,2018,37(12):1241-1248.
- [19]毕殿杰,魏苏林,赵涛,等.基于卷积神经网络的在线评论情感分析模型[J].河北科技师范学院学报,2019,33(2):41-47.
- [20]张柳,王晰巍,黄博,等.基于字词向量的多尺度卷积神经网络微博评论的情感分类模型及实验研究[J].图书情报工作,2019,63(18):99-108.
- [21]李慧,柴亚青.基于卷积神经网络的细粒度情感分析方法[J].数据分析与知识发现,2019,3(1):95-103.
- [22]王晰巍,邢云菲,韦雅楠,等.大数据驱动的社交网络舆情用户情感主题分类模型构建研究——以“移民”主题为例[J].信息资源管理学报,2020,10(1):29-38,48.
- [23]Zhang S X, Wei Z L, Wang Y, et al. Sentiment Analysis of Chinese Micro-blog Text Based on Extended Sentiment Dictionary[J]. Future Generation Computer Systems, 2018, 81: 395-403.
- [24]林青,李立煊,杨腾飞.社交网络用户影响力量化模型研究——以新浪微博为例[J].情报杂志,2018,37(8):203-207.
- [25]陈思菁,李纲,毛进,等.突发事件信息传播网络中的关键节点动态识别研究[J].情报学报,2019,38(2):178-190.
- [26]王珂,刘东苏.基于PageRank的动态网络核心节点检测及演化分析[J].情报学报,2018,37(7):703-711.
- [27]陈晓威,史昱天.社会网络中关键节点的识别——基于符号网络的PageRank算法改进[J].数据分析与知识发现,2017,1(8):68-75.
- [28]张风军.基于Neo4j图数据库的社交网络数据的研究与应用[D].长沙:湖南大学,2016.
- [29]Holzschueher F, Peinl R. Performance of Graph Query Languages: Comparison of Cypher, Gremlin and Native Access in Neo4j[C]//Proceedings of the Joint EDBT/ICDT 2013 Workshops. ACM, 2013: 195-204.
- [30]Owen S, Anil R, Dunning T,等.Mahout 实战[M].王斌,韩冀中,万吉,译.北京:人民邮电出版社,2014:134-138.
- [31]张琳,牟向伟.基于Canopy+K-means的中文文本聚类算法[J].图书馆论坛,2018,38(6):113-119.
- [32]Alexa. Traffic Detail(sina.com.cn)[EB/OL].[2019-11-08].<http://www.alexa.com/siteinfo/www.sina.com.cn>.
- [33]新媒田宇.校园舆情·2018年高校十大热点事件[EB/OL].

[2019-11-08].<https://baijiahao.baidu.com/s?id=1623436522696496935&wfr=spider&for=pc>.

[34]Cho S W, Cha M S, Sohn K A. Topic Category Analysis on Twitter via Cross-media Strategy[J]. *Multimedia Tools & Applications*, 2016, 75(20): 12879-12899.

[35]Nainggolan R, Perangin-Angin R, Simarmata E, et al. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error(SSE) Optimized by Using the Elbow Method[J]. *Journal of Physics Conference Series*, 2019, 1361: 012015.

[36]Fernandez-Gavilanes M, Juncal-Martinez J, Garcia-Mendez S, et al. Differentiating Users by Language and Location Estimation in Sentiment Analysis of Informal Text During Major Public Events[J]. *Expert Systems with Applications*, 2019, 117: 15-28.

[37]Zhang W, Wang M, Zhu Y C. Does Government Information Release Really Matter in Regulating Contagion-Evolution of Negative Emotion During Public Emergencies? From the Perspective of Cognitive Big Data Analytics[J]. *International Journal of In-*

formation Management, 2020, 50: 498-514.

[38]Chen S J, Mao J, Li G, et al. Uncovering Sentiment and Retweet Patterns of Disaster-related Tweets from a Spatiotemporal Perspective-A Case Study of Hurricane Harvey[J]. *Telematics and Informatics*, 2020, 47: 101326.

[39]Lee J Y H, Yang C S, Hsu C, et al. A Longitudinal Study of Leader Influence in Sustaining an Online Community[J]. *Information & Management*, 2019, 56(2): 306-316.

[40]Ahajjam S, Haddad M E, Badir H. A New Scalable Leader-community Detection Approach for Community Detection in Social Networks[J]. *Social Networks*, 2018, 54: 41-49.

[41]Zhang L F, Su C, Jin Y F, et al. Cross-network Dissemination Model of Public Opinion in Coupled Networks[J]. *Information Sciences*, 2018, 451: 240-252.

[42]Li C L, Bai J P, Zhang L, et al. Opinion Community Detection and Opinion Leader Detection Based on Text Information and Network Topology in Cloud Environment[J]. *Information Sciences*, 2019, 504: 61-83.

Clustering User Groups of Public Opinion Events from Multi-dimensional Social Network

Wang Xiwei Jia Ruonan Wei Ya'nan Zhang Liu

Abstract: [Objective] User groups are the main units to disseminate public opinion. This study identifies the characteristics of user groups through clustering techniques, which could help social network companies provide better services. [Methods] With the help of Group Theory, we clustered users based on their influence, sentiments, and behaviors. First, we collected user data from the Sina Weibo. Then, we utilized Canopy and K-Means algorithms to cluster users. Finally, we visualized our findings with Neo4j and Weka. [Results] User groups of the same public opinion event were different in emotion, influence, and behaviors, while user groups from different public opinion events shared common characteristics. [Limitations] Both public opinion events in this study happened at Chinese universities, and we only collected data from Sina Weibo. [Conclusions] Based on the clustering results, we could propose effective administration strategies for each user group in the same or different public opinion events.

Key words: Multi-dimensional; Social network; Public opinion; User group; User clustering