·业务研究·

多维属性融合视角下的在线健康社区关键用户识别研究

张　军,李新旺,李　鹏

(山东理工大学 管理学院,山东 淄博 255012)

摘　要:【目的/意义】在线健康社区已成为公众获取医疗信息和服务的重要形式。识别在线健康社区关键用户及其特征,为提升健康社区服务质量和效率提供理论依据。【方法/过程】基于信息行为学理论构建了包括交互行为属性、信息质量属性、情感倾向属性的多维分析框架,利用AttriRank算法和网络抗毁性评估方法识别在线健康社区关键用户。【结果/结论】在胆系癌症疾病QQ群中识别出15个关键用户。他们不仅具有高活跃性和高互惠度的交互行为特征,还具备多样性水平高且结构均衡的信息质量特征,且多数持有正向情绪倾向。"行为+内容+情绪"的分析框架和考虑属性的用户排序算法能准确识别在线健康社区关键用户,为在线健康社区的持续运营供了科学的决策支持。【创新/局限】构建多维属性分析框架进行在线健康社区关键用户识别,丰富了在线健康社区关键用户识别的理论体系。

关键词:在线健康社区;关键用户;属性融合;AttriRank算法;网络抗毁性评估

中图分类号:G252.0　　**DOI:**10.13833/j.issn.1007-7634.2022.03.011

1 引　言

随着"互联网+医疗"及"健康中国2030"战略导向的发展,公民健康意识与信息素养正不断提升,促进了一系列在线健康社区的形成与快速发展,如甜蜜家园、百度高血压吧和以医疗健康为主题的QQ群、微信群等[1]。中国互联网信息中心发布截至2020年12月,在线医疗用户规模达2.15亿,占网民整体的21.7%,新冠肺炎疫情期间部分第三方互联网服务平台咨询量同比增长了20多倍[2]。在线健康社区已逐渐成为了患者获取健康信息的重要渠道。病患及家属借助不同类型的在线健康社区平台与病友、医生、志愿者等进行交互,搜寻医生医院、治疗方案等信息。他们共同构成了在线健康社区,并在交互过程中不断学习和积累经验,实现了健康信息搜寻、交互、获取方式的进化。研究结果显示,与知乎问答社区、Linux在线交流社区及Wiki编辑者社区等在线社区一样[3],在线健康社区中也存在一类特殊用户[4],对其存在和发展起到了重要的作用。他们通过积极提供健康知识、开展情感交流,带动新用户、潜伏用户积极参与在线活动[5],是社区存在和发展的重要驱动力,被称为在线健康社区中的关键用户。由此,构建多维属性融合的在线健康社区关键用户识别方法并分析他们的行为特征,对提高在线健康社区的管理水平和服务质量有重要的参考价值。

2 相关研究现状述评

在线社区上的关键用户可能是舆情传播中的"意见领袖",也可能是某个特殊领域的自媒体。他们往往有大量的粉丝或关注者,由其提供的信息会有较大的概率被转发和评论[6],进而能影响更多的普通大众。所以已有研究成果主要考虑了关键用户在社会网络中的结构位置进行识别,如:在线社会网络中节点的度数、介数、特征向量中心性等[7-9],其优势是评价方法比较客观,但是评价指标比较单一。而在线健康社区(Online Health Community)是一类特殊的知识型在线社区,医生、患者及其家属都可以在不受时间和空间条件的约束下,用发文、回复、点赞、转发等形式完成健康信息或者专业医疗知识交互[10]。因此,在线健康社区上的关键用户不仅在社会网络中位置重要,还能提供医疗领域知识并促进用户之间的交流[11]。所以从信息行为学视角出发,识别在线健康社区中的关键用户需要综合考虑用户的交互行为、内容和主题、甚至情感等多个维度[12-13]。

在线健康社区用户行为研究涉及了用户搜索行为、分享行为和交互行为等方面。早期的用户行为研究通过调研和统计分析方法,解释了用户的健康信息行为动机,认为利他主义、享受、自我价值等满足感会对用户的医疗知识共享行为产生显著正向影响[14-16]。随着在线健康社区应用的普及,

收稿日期:2020-05-11

基金项目:国家自然科学基金项目"社交媒体上观点、信息和行为的耦合动力学机制研究"(G71801145);山东省社科规划基金项目"新旧动能视域下在线科普信息传播机制及效果评价研究"(20CGLJ22)。

作者简介:张军(1978-),女,山东淄博人,博士,教授,主要从事复杂网络和舆情动力学研究;李新旺(1996-),男,山东德州人,硕士研究生,主要从事信息管理与智能科学研究;李鹏(1977-),男,山东泰安人,硕士,副教授,主要从事管理科学理论与方法研究,通讯作者:Lp_sdut@163.com。

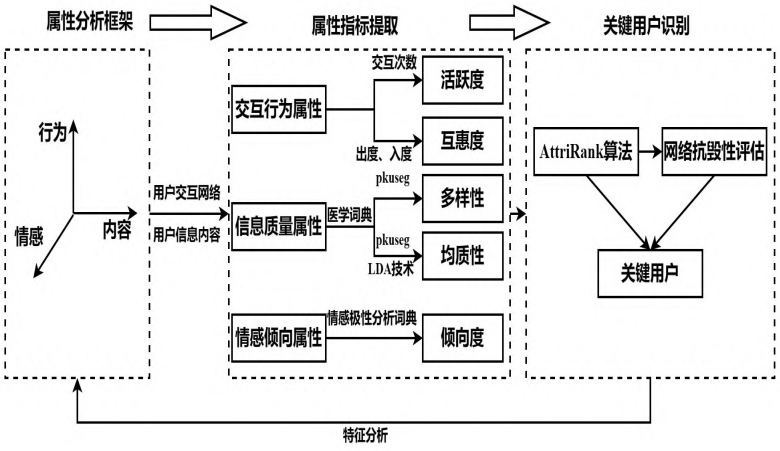

图1 关键用户识别和特征分析研究框架图

Figure 1 The research framework of key users´ identification and feature analysis

研究热点迁移到用户的交互行为。刘璇等[5]研究了在线健康社区用户的回帖行为,发现活跃度比较高的用户在后续交互中获得回复的概率比较大。这说明了高活跃度用户是在线健康社区交互关系产生的动力[17]。此外,从社会互惠理论视角来看,在线健康社区的用户在利他心理的驱动下,通过交互行为能够实现信息价值资源的交互,比如:疾病治疗的专业建议和情感支持。用户所持有的互惠交互模式也是维持在线社区运营的重要社会性规范之一[4]。

健康信息内容也是在线健康社区研究的一个重要分支。学者们主要采用文本挖掘、机器学习等方法对健康信息内容或主题分布进行识别、分类、统计等分析[18-21],并关注到较差的信息质量[22]不仅会影响用户参与社区交互行为的发生,还会直接导致社区用户流失。由于在线健康社区是一种典型的UGC社区,用户是提供知识和信息的主要来源,故而用户所持有的信息质量高低就决定了在线健康社区专业领域知识服务的水平。一般而言,评价一个用户信息质量可以从信息内容多样性和信息结构均衡性两个方面出发。如:范昊等[23]研究发现在Yahoo Answers社区中,有些关键节点虽然度比较大,但其贡献的内容比较单调,使得社区交互结构比较松散;翟羽佳等[24]研究则表明,百度戒烟吧由于中高活跃用户贡献的信息不均衡,导致论坛话语权偏移,阻碍了新用户的参与。因此,信息质量是识别在线健康社区中关键用户不可或缺的维度。

此外,在对用户的信息需求进行研究时发现,在线健康社区上的患者及家属不仅需要专业知识的支持,情感支持和陪伴[25]等也十分重要。Beaudoin等[26]通过研究在线健康社区中的情感支持对癌症病人情绪的影响,发现在线健康社区确实能够减少癌症病人的负向情感。同时,患者之间的交互行为可以将用户的负向情感逐渐转变为正向情感,表明在线健康社区可以给患者及家属提供必要的情感支持[27]。因此,对于在线健康社区来说,用户情感属性也是关键用户识别的重要维度。

综上所述,目前对在线健康社区的关键用户识别研究还较少,需要从“行为+内容+情绪”等多维度综合考虑开展研究。鉴于此,本文从信息行为学角度出发,构建了交互行为属性、信息质量属性和情感倾向属性指标;然后基于AttriRank算法和网络抗毁性评估方法进行关键用户的识别和检验,并分析了关键用户的基本特征。

3 研究设计

3.1 关键用户识别和分析框架

本文设计了在线健康社区关键用户识别和分析框架,如图1所示。根据信息行为学理论,提出“行为+内容+情感”的多维度分析框架,构建用户交互行为属性、信息质量属性、情感倾向属性。其次,根据在线健康社区用户交互行为数据,构建用户交互网络并计算用户交互行为属性;根据在线健康社区用户的信息内容数据,结合百度名医百科文本内容及医学词典生成用户交互关键词,用pkuseg切词工具和LDA主题技术识别并抽取在线健康社区的主题,并用基于词典的情感极性分析方法分析用户情感属性。然后,基于AttriRank算法对用户进行评分和排序,通过网络抗毁性评估和检验社区内的关键用户效用,并分析关键用户的属性特征。

3.2 用户属性

在线健康社区中的用户可以通过提问、回答、评论等在网络上进行交互。以用户作为节点,用户之间的交互关系作为连边,两者构成了一个复杂的交互网络,可以记为$G=(V,E)$。其中V是交互网络中的节点集合,且V中的每个节点i均具有交互行为属性、信息质量属性和情感倾向属性。E是用户交互关系构成的有向边集合,$\forall_e < i,j > \in E$,表示节点i提及(@)、转发或者评论节点j的信息,即节点i主动与j进行交互。在线健康社区上用户之间的交互内容构成

了关键词集合$KW=\{key_1,key_2...key_n\}$,在此基础上可以获得相应的交互主题集合,表示为$\Gamma=\{T_1,T_2...T_S\}$。其中$T_l$表示第l个交互主题由f个关键词构成,记为$T_l=\{key_1^l,key_2^l...key_f^l\}$。

3.2.1 交互行为属性

在线健康社区用户的交互行为属性包括交互的活跃度和互惠度。

活跃度是指用户在社区中活跃程度,通过用户i的交互次数与社区中所有用户的交互次数之和的比值来进行度量,活跃度计算方式如公式(1)所示。其中,w_i是用户i的加权度,代表交互网络中节点i的交互次数;$\sum_{i=1}^{|V|}w_i$是交互网络中的总加权度,表示了社区中所有用户的交互次数之和。

$$a_i=w_i/\sum_{i=1}^{|V|}w_i \tag{1}$$

互惠度是指用户参与社区的互动模式,通过用户i的主动交互次数和被动交互次数比值进行计算,互惠度r计算方式如公式(2)所示。其中,k_i^{out}是节点i的出度,代表用户主动交互次数;k_i^{in}是节点i的入度,代表用户被动交互次数。$r>1$表示用户主动交互次数占优;反之,用户被动交互次数占优。

$$r_i=k_i^{out}/k_i^{in} \tag{2}$$

3.2.2 信息质量属性

在线健康社区用户的信息质量属性包括信息多样性与信息均质性。

信息多样性是指用户在交互过程中提及信息类型的丰富程度。由每个用户使用的关键词数量q_i与社区中关键词总数n的$\log_2^n$之比进行度量,计算方式如公式(3)所示。

$$d_i=q_i/\log_2^n \tag{3}$$

信息均质性表示用户在交互过程中信息结构的均衡程度[28]。通过用户交互过程中提及的主题信息熵H_i与最大信息熵的比值来度量,计算方式如公式(5)所示。其中,信息熵的计算方式如公式(4)所示,$P(T_{ij})$表示用户i在第l个主题上的概率,S_i代表用户i的信息主题个数。信息均质性u的取值范围为[0-1],当u越接近1时,信息均质性越高;反之,信息均质性越低。

$$H_i=-\sum P\left(T_{ij}\right)*\log_2^{P\left(T_{ij}\right)} \tag{4}$$

$$u_i=H_i/\log_2^{S_i} \tag{5}$$

3.2.3 情感倾向属性

在线健康社区用户的情感倾向度是指用户在交互过程中情感的倾向性,通过正向情感频数与负向情感频数之比进行度量,计算方式如公式(6)所示。其中,pe_i表示用户交互过程中正向情感出现的次数,ne_i表示用户交互过程中负向情感出现的次数。当情感倾向度m大于1时,表明该用户在信息交互过程中倾向于表达正向情感;反之,用户倾向于表达负向情感。

$$m_i=pe_i/ne_i \tag{6}$$

综合上述定义,用户属性指标如表1所示。

表1 用户属性指标及计算公式

Table 1 User´s attributes and calculation formula

用户属性维度	属性	计算公式
交互行为属性	活跃度	$a_i=w_i/\sum_{i=1}^{\|V\|}w_i$
	互惠度	$r_i=k_i^{out}/k_i^{in}$
信息质量属性	多样性	$d_i=q_i/\log_2^n$
	均质性	$u_i=H_i/\log_2^{S_i}$
情感倾向属性	倾向度	$m_i=pe_i/ne_i$

3.3 关键用户识别方法

3.3.1 AttriRank算法

AttriRank是同时考虑网络结构和网络中节点属性的无监督节点重要性排序方法[29]。该算法以PageRank算法为基础,同时考虑节点的多维度属性,对网络中的每个节点进行打分。设第t次迭代过程中网络G中的任意节点i的属性向量为$x_i(t)$,对应得分为$\pi_i(t)\geq 0$,则网络中节点集合的得分构成了向量$\pi(t)=(\pi_0(t),\pi_1(t),...)\in R^N$,$\sum_{i\in V}\pi_i(t)=1$是一个离散的马尔科夫空间。因此,本文将基本迭代规则设置为:具有相似属性的两个节点,其分类或回归结果应该是相似的。所以,在第t次迭代过程中节点得分的计算步骤如下:

步骤1:按照公式(7)计算节点集V中任意两个节点i和j之间的属性相似性$S_{ij}(t)$。

$$S_{ij}(t)\equiv e^{-\gamma\left\|x_i(t)-x_j(t)\right\|_2^2} \tag{7}$$

其中,$x_i(t)$,$x_j(t)$是两个节点的属性向量,参数γ控制属性间距离的影响程度。

步骤2:按照公式(8)计算出节点i与节点j属性的相互影响力$Q_{ij}(t)$。

$$Q_{ij}(t)\equiv S_{ij}(t)/\sum_{k\in V}S_{kj}(t) \tag{8}$$

其中,$\sum_{k\in V}S_{kj}$代表节点集V中所有节点与j节点的相似性之和。

步骤3:根据PageRank算法的计算规则,如公式(9)计算网络中节点i与节点j结构的相互影响力$P_{ij}(t)$。

$$P_{ij}(t)\equiv\begin{cases}1/k_j^{out},<j,i>\in E\\1/|V|,k_j^{out}=0\\0,otherwise\end{cases} \tag{9}$$

其中,k_j^{out}代表用户j的出度,$|V|$代表网络中节点的总个数。

步骤4:按照公式(10)更新用户集合的得分$\pi(t)$。

$$\pi(t)=(1-d)Q+dP \tag{10}$$

其中,阻尼系数$d\in(0,1)$,矩阵Q代表节点之间的属性影响力,矩阵P代表节点之间的结构影响力。

步骤5:计算误差$\varepsilon=|\pi(t-1)-\pi(t)|$,当$\varepsilon$足够小停止迭代,输出节点得分。

表2 胆系癌症QQ群社区数据样例

Table 2 Gallbladder cancer QQ group data

form	to	content
u1	u2	@u2早上只吃了点肠内。中午想吃点稀饭行吗?
u2	null	稀饭里也给她加点
u3	u1	炖鲫鱼汤喝吧,可以缓解白细胞低 @u1
u4	u3	@u3我家没有吃靶向药,医生有建议阿帕替尼,但群里反应不太好。
u5	u6	@u6病人压力更大
u6	null	医生查房还是建议阿帕替尼,效果不错,说现在报销,价格降下来了~~
u7	null	我们家每天二盒脱脂牛奶,2一3袋蛋白粉
u8	null	如果饮食跟不上,癌症这个病就是耗元气。
u9	u6	@u6阿帕替尼胆囊好像报不了,不是适应症
u6	u9	我家是肝门胆管@u9

3.3.2 网络抗毁性评估

网络抗毁性评估方法是复杂网络研究领域中节点重要度评估的有效模型。它的基本研究假设是关键节点对网络的基本结构形态有重要的影响。因此,可以统计特定节点移除后网络的效率、鲁棒性等指标的变化来评价节点的作用,并判断节点是否为关键节点。本文所采用的网络效率计算公式如下:

$$\eta = 1/N(N-1)\sum_{i\neq j\neq C} 1/d_{ij} \tag{11}$$

其中,η的取值范围为[0-1],两节点i和j之间的效率为$1/d_{ij}$,$\eta=1$表示网络连通性最好,$\eta=0$则表示网络是由孤立的节点组成的。

3.3.3 关键用户识别过程

根据AttriRank算法得到每个用户得分,并计算出高分用户移除后的网络效率变化的程度,进而进行关键用户识别。具体过程如下:

步骤1:根据AttriRank算法计算出每个用户的得分,并按得分进行降序排列。

步骤2:从得分最高的用户开始逐步移除对应节点,并计算移除节点后的网络效率。

步骤3:观察移除节点后网络效率的变化情况,根据网络崩溃的阈值判断关键用户个数。

4 实例及分析

4.1 数据集简介

本文选择了肝胆外科中胆系癌症QQ群社区为研究对象,它是国内较大、人数较多、交流较为活跃自发病友群,共包含2780个用户,涉及30个省市(台湾、西藏、香港和澳门除外)。本文收集了社区内2020.07.01-2020.12.31共6个月的聊天文本,共获取197649条有效数据,原数据样例如表2所示。

由于胆系癌症发病率低且恶性程度高,患者和家属对此病症知识了解途径比较少,所以在QQ群社区中,用户间交互比较频繁。一般而言,用户可以根据社区中的上下文随意加入到聊天列表中,如表2的第2行数据所示。另外一种方式是用户通过@提及某个特殊用户,进行提问或者回答等交互。由此,将QQ群社区中用户作为交互网络的节点,根据第二种交互方式生成有向边$<u_1,u_2>$添加到交互网络。最终该健康社区的交互网络由2780个节点和11000条有向边组成,过滤掉孤立节点后得到社区最大连通分量图,如图2所示。

图2 胆系癌症QQ群社区交互网络的最大连通分量

Figure 2 Maximum connected component of interaction network of gallstone cancer QQ group

图2中节点的颜色是根据节点所属不同的模块进行相应的配色,节点的大小与该节点度的大小相对应。交互网络结构统计结果为:节点平均度为3.957,即每个用户会与3至4个人发生交互;模块度为0.173,说明用户交互中形成了明显的社区结构;平均路径长度为2.698,聚类系数为0.17,两者表明该交互中存在较短的信息传输路径,可以使健康信息在社区中快速传播。

此外,根据在线健康社区用户的聊天内容数据,结合百度名医百科文本内容及医学词典生成用户交互关键词集合KW,包含关键词89516个;用pkuseg切词工具与LDA主题技术结合,识别并抽取在线健康社区的主题集合Γ,共包含6个

表3 胆系癌症QQ群社区主题及主题词

Table 3 Topics and key words of gallbladder cancer QQ group

主题	主题名称	主题词
T0	医生医院	医院,手术,医生,时间,检查,复查,推荐,价格…
T1	疾病检查	问题,报告,办法,病理,中药,影像,门诊,单位…
T2	疾病症状	难受,感染,发烧,血小板,睡觉,舒服,穿刺,胆红素…
T3	药物情感	治疗,谢谢,肿瘤,便宜,药物,学习,帮忙,黄疸…
T4	术后治疗	医生,化疗,方案,手术,基因检测,副作用,转移,术后 …
T5	术前治疗	影像报告,化疗,引流,支架,靶向药,放疗,手术,靶向…

主题,其中主题的高频词详见表3。

从表3中可以看出,胆系癌症QQ群社区中用户主要关注的是医生医院、疾病检查、疾病症状、药物情感、术后治疗、术前治疗共六类信息主题,且每个主题之间具有紧密的联系。

4.2 关键用户识别结果

首先,根据用户属性定义从数据集中获取用户交互行为属性、信息质量属性和情感倾向属性值。以节点741用户的属性计算为例,计算过程如下:

(1)交互行为属性

① 活跃度。节点741的交互次数为7340,社区总的交互次数为197650,故其活跃度为:

$$a_i = w_i \Big/ \sum_{i=1}^{|V|} w_i = 7340/197650 = 0.037$$

② 互惠度。节点741的出度为316,入度为250,故其互惠度为:

$$r_i = k_i^{out}/k_i^{in} = 316/250 = 1.264$$

(2)信息质量属性

① 信息多样性。节点741的交互过程中使用关键词10238个,社区所有用户交互过程中使用的关键词为89516个,故其多样性为:

$$d_i = q_i/\log_2^n = 10238/\log_2^{89516} = 622.38$$

② 信息均质性。节点741在交互过程中涉及了6个主题,且每个主题下的交互次数分别为[1987,1352,785,1057,1082,1077],总交互次数为7340。其均质性计算过程分3步:

Step1:计算节点741在各个主题上的交互概率。其中,第0个主题上的交互概率计算过程如下。

$$P(T_{i,0}) = T_{i,0}/w_i = 1987/7340 = 0.271$$

以此类推,得到节点741在其他主题上的交互概率分别为:$P(T_{i,1}) = 0.184$),$P(T_{i,2}) = 0.107$,$P(T_{i,3}) = 0.144$,$P(T_{i,4}) = 0.147$,$P(T_{i,5}) = 0.147$。

Step2:利用Step1中得到节点741在每个主题上的概率$P(T_{i,l})$,计算节点741的信息熵。

$$H_i = -\sum P\left(T_{i,j}\right) * \log_2^{P\left(T_{i,j}\right)}$$

$$= (-0.271 * \log_2^{0.271}) + (-0.184 * \log_2^{0.184}) + (-0.107 * \log_2^{0.107}) + (-0.144 * \log_2^{0.144}) + (-0.147 * \log_2^{0.147}) + (-0.147 * \log_2^{0.147})$$

$$= 2.521$$

Step3:计算节点741的信息均质性。

$$u_i = H_i/\log_2^{S_i} = 2.521/\log_2^6 = 0.975$$

(3)情感倾向属性

节点741的正向情感频数为3173,负向情感频数为3501,故其情感倾向度为:

$$m_i = pe_i/ne_i = 3173/3501 = 0.906$$

在获取每个节点的属性值后,通过AttriRank算法计算用户得分。其中,将阻尼系数d设置为0.85,$\varepsilon = 1E-10$。最后根据3.3中关键用户识别过程,将网络崩溃的临界值设为网络初始性能的50%[30-31],识别关键用户。图3中显示了逐步移除高分节点后,网络效率变化趋势。显然,当移除第Top15节点时网络达到崩溃临界阈值。因此将得分排名Top15的用户作为这个在线健康社区的关键用户,占整个社区的0.5%。他们的属性值及得分情况如表4所示。AttriRank与PageRank排名的区别在于第3~4、12~15名的次序,与Betweeness Centrality(中介中心性)排名和Eigenvector Centrality(特征向量中心性)排名比较分别有10个和13个用户存在次序差异。产生这些差异的主要原因是AttriRank算法同时考虑了用户的属性,也即用户属性有差异时,在不同的算法中排名就存在差异。

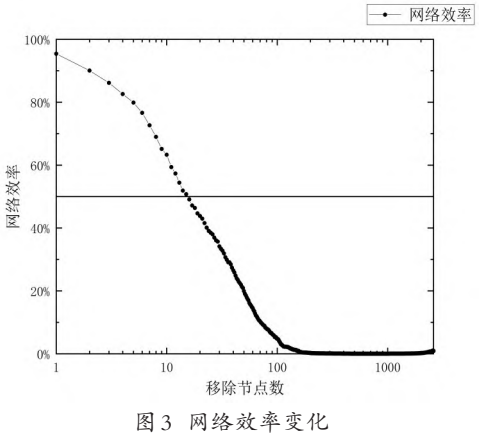

图3 网络效率变化

Figure 3 The change of network efficiency

4.3 关键用户特征分析

在新信息技术的驱动下,各类在线健康社区中都聚集了大量的用户。一般而言,用户在线的交互方式和结果与现实

表4 胆系癌症QQ群社区关键用户的属性值及得分(取3位小数)

Table 4 Attribute values and scores of key users' in gall cancer QQ group (take 3 decimal places)

node-id	活跃度 a_i	互惠度 r_i	多样性 d_i	均质性 u_i	倾向度 m_i	score	Attri Rank	Page Rank	Betweeness	Eigen vector
741	0.037	1.264	622.38	0.975	0.906	0.014	1	1	1	2
1278	0.050	1.290	1029.6	0.990	1.128	0.013	2	2	2	1
1166	0.059	1.178	750.10	0.980	2.518	0.011	3	4	4	5
676	0.041	1.350	637.27	0.990	0.938	0.011	4	3	3	3
222	0.063	1.218	641.65	0.962	1.171	0.010	5	5	5	4
1145	0.020	1.261	315.69	0.983	0.909	0.009	6	6	7	6
904	0.014	1.538	327.48	0.993	1.292	0.008	7	7	6	7
910	0.022	1.556	320.85	0.963	2.043	0.007	8	8	8	10
816	0.055	1.676	579.64	0.964	1.594	0.007	9	9	9	8
539	0.015	1.104	151.13	0.947	0.667	0.007	10	10	12	19
1244	0.022	1.521	380.25	0.982	2.153	0.006	11	11	10	9
140	0.003	1.008	118.91	0.962	0.776	0.006	12	13	13	24
1099	0.020	1.496	308.27	0.975	1.412	0.006	13	12	11	11
425	0.009	1.604	199.03	0.973	1.864	0.005	14	15	14	22
1307	0.018	0.793	169.06	0.914	0.645	0.005	15	14	16	14

社会有很大区别,且关键用户的交互行为能对知识扩散和信息传播有重要的影响。因此,本文分析和比较了QQ群社区中关键用户和普通用户的属性值差异,并总结了关键用户的特征,具体详述如下。

4.3.1 关键用户交互行为特征

图4(a-b)展示了关键用户和普通用户的活跃度和互惠度分布。其中X轴标识用户活跃度或互惠度的值,垂直虚线 $\bar{a}=0.00036$、$\bar{r}=0.00036$ 分别标记了社区中用户活跃度和互惠度的均值,以 $a/\bar{a}$、$r/\bar{r}$ 的值作为Y轴。图中均采用了双对数坐标系,清晰地显示了在本文研究的在线健康社区中用户活跃度和互惠度值差异比较大。

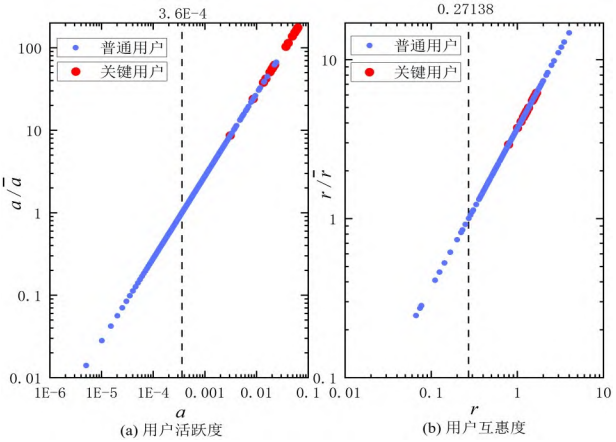

图4 用户活跃度、互惠度分布

Figure 4 The distribution of user's activity and reciprocity

图4(a)显示,有92.8%用户的活跃度值低于 $\bar{a}$,7.2%的用户活跃度值高于 $\bar{a}$。故总体而言,关键用户的活跃度均处于较高的水平,其中活跃度的最小值是0.0031,是均值的8.6倍,活跃度的最大值同时也是社区中全体用户的最大值,是0.06299,是均值的175.0倍。图4(b)显示,r 值大于1的用户只有8.2%。其中,关键用户互惠度最小值是0.79279<1,是均值的2.9倍;最大值是1.67586,是均值的6.2倍。又因为互惠度 r 值大于1表示的是用户进行主动交互,所以说明本文研究的QQ群中多数用户交互目的是获取信息和知识。特别的,关键用户的互惠度值并不是社区中最高的。这说明关键用户在社区中并不是有问必答,而是有选择性的为其他用户提供信息。

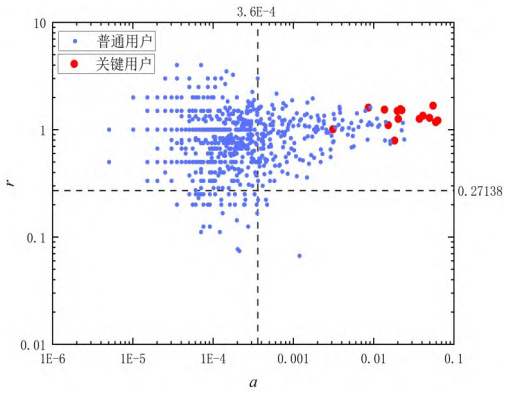

图5 用户活跃度与互惠度关系

Figure 5 The relationship between users' activity and reciprocity

图5中则展示了关键用户和普通用户的活跃度值与互惠度值的对应关系。在双对数坐标系下,发现用户的交互行为呈现出两类模式。第一类表现为活跃度值较小,互惠度值比较分散。这些用户多数是患者和家属,他们不会一直在线活跃,只有当其病程出现进展、病症发生变化时,才会在QQ群中发言、求助。第二类表现为用户的活跃度值偏大,而互惠度值接近于1。关键用户的活跃度和互惠度的对应关系

就是后者。这说明关键用户在QQ群社区中的交互行为有两种,一是利用碎片化时间浏览群里用户的聊天内容,主动通过@UID的形式为特定用户提供信息和知识,二是当群里用户通过@UID形式搜寻医疗健康知识时,对其进行回复。关键用户的这两种不同的交互行为总数占到了社区这两种交互行为总数的25.9%,进一步说明了关键用户是社区用户交互行为可以持续的主要动力。

4.3.2 关键用户信息质量特征

图6(a-b)展示了关键用户和普通用户的信息多样性和均质性差别,其中X轴标识用户信息多样性和均质性的值,垂直虚线 $\bar{d}$ = 6.27414、$\bar{u}$ = 0.30134分别标记了社区中用户信息多样性和均质性均值,以 $d/\bar{d}$、$u/\bar{u}$ 的值作为Y轴,同样采用了双对数坐标系。

图6(a)显示了在本文研究的在线健康社区中89.7%用户的多样性值低于 $\bar{d}$,10.3%的用户多样性值高于 $\bar{d}$。其中,关键用户的信息多样性均处于较高的水平:多样性最小值是118.9068,是均值的19.0倍,多样性的最大值是1029.614,是均值的164.1倍。说明关键用户信息多样性水平较高,能在交互过程中提供大量的信息和丰富的内容。图6(b)显示了在本文研究的在线健康社区中60.4%用户的信息均质性值低于 $\bar{u}$,39.6%的用户信息均质性值高于 $\bar{u}$。其中,关键用户的信息均质性值均处于较高的水平:最小值是0.9142,是均值的3.0倍,信息均质性的最大值是0.99267,是均值的3.3倍;且关键用户的信息均质性值均趋近于1,同时说明了关键用户在交互过程中信息结构的均衡程度高。

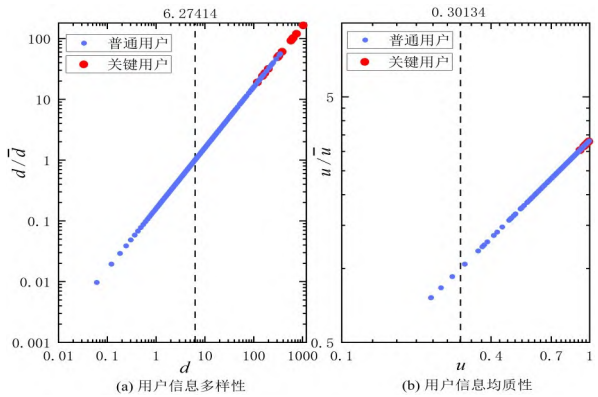

图6 用户信息多样性、均质性分布

Figure 6 The distribution of information diversity and homogeneity

图7中则展示了关键用户和普通用户的信息多样性值与信息均质性值的对应关系。在双对数坐标系下,多样性值较小的用户均质性值比较分散,多样性值偏大的用户均质性接近1。也就是对于普通用户而言,受到自身知识背景、病程、病症等因素影响,他们聊天内容涉及的信息量和主题范围有较大的局限性。而关键用户则是要给不同病程和病症的患者和家属提供健康医疗知识,所以其信息多样性和均质性值均名列前茅。

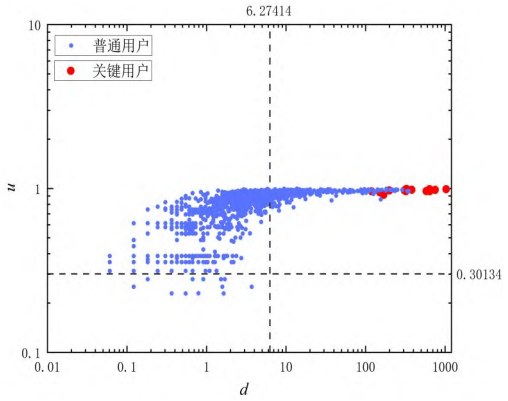

图7 用户信息多样性与均质性关系

Figure 7 The relationship between diversity and homogeneity

4.3.3 关键用户情感倾向特征

图8展示了关键用户和普通用户的情感倾向度差别,其中X轴标识情感倾向度m的值,垂直虚线 $\bar{m}$ = 0.4258标记了社区中用户情感倾向度的均值,以 $m/\bar{m}$ 的值作为Y轴,采用了双对数坐标系。

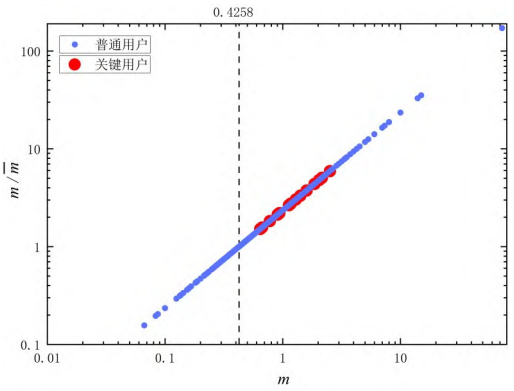

图8 用户情感倾向度分布

Figure 8 The distribution of users´ emotional disposition

图8显示本文研究的在线健康社区中,有70.1%用户的情感倾向度值低于 $\bar{m}$,29.9%的用户情感倾向度值高于 $\bar{m}$。其中,关键用户的情感倾向度均处于较高的水平:情感倾向度的最小值是0.64476,是均值的1.5142倍,情感倾向度的最大值是2.51826,是均值的5.91418倍。此外,关键用户情感倾向度主要集中在1附近,表明关键用户在交互过程中倾向于表达正向情感,且没有明显的极端性,这对于维持良好的社区氛围起到了重要的作用。

5 结 语

本文研究在线健康社区关键用户识别和特征分析问题。首先为关键用户的识别构建出多维融合属性指标体系,从交互行为属性、信息质量属性和情感倾向属性三个维度,构建了用户的活跃度、互惠度,信息的多样性、均质性,情感倾向

度等5个分析指标。其次是构建关键用户识别方法,诸如PageRank等经典的关键用户识别算法不适用于在线健康社区,应结合用户多维属性开展在线健康社区关键用户识别和特征分析。因此本文在AttriRank算法基础上获取用户评价结果和排名,通过网络抗毁性评估方法识别在线健康社区关键用户。

本文以胆系癌症的患者和家属自发形成的QQ群社区为研究对象开展了实证验证,在胆系癌症QQ群社区中识别出15个关键用户。研究结果表明:第一,在线健康社区中的关键用户比普通用户有较高的交互意愿,他们的作用不仅体现在与普通用户发生交互,而且还是社区中健康信息和医疗知识的主要传播者;第二,在线健康社区中的关键用户不仅比普通用户拥有更多信息,而且信息结构更均衡,是社区中主要的知识和经验来源及载体;第三,在线健康社区中的关键用户极少有极端情绪表现,能给普通用户提供适度的正向情感支持。

在未来的工作中,可结合时间感知框架设计动态环境下的关键用户识别算法,进一步提高健康信息服务水平,为促进社区发展提供更多的合理建议。

参考文献

1 赵栋祥.国内在线健康社区研究现状综述[J].图书情报工作,2018,62(9):134-142.

2 中国互联网络信息中心.第47次中国互联网络发展状况统计报告[EB/OL].[2021-02-03].http://www.cac.gov.cn/2021-02/03/c_1613923423079314.htm.

3 柯阳,隋杰.基于用户特征属性的微博话题关键用户挖掘[J].计算机应用研究,2019,36(6):1614-1617,1622.

4 吴江,周露莎.在线医疗社区中知识共享网络及知识互动行为研究[J].情报科学,2017,35(3):144-151.

5 刘璇,汪林威,李嘉,张朋柱.在线健康社区中用户回帖行为影响机理研究[J].管理科学,2017,30(1):62-72.

6 席海涛,聂文博,李闰臣,田慧敏,陈立.在线健康社区用户交互的研究现状与进展[J].情报科学,2021,39(4):186-193.

7 FREEMAN L C.A set of measures of centrality based on betweenness [J].Sociometry,1977,40(1):35- 41.

8 FREEMAN L C.Centrality in social networks conceptual classification [J].Social Networks,1978,1(3):215-239.

9 BONACICH P.Factoring and weighting approaches to status scores and clique identification[J]. Journal of Mathematical Sociology,1972,2(1):113-120.

10 Young C. Community Management that Works: How to Build and Sustain a Thriving Online Health Community[J]. Journal of Medical Internet Research,2013,15(6):e119.

11 阮光册,夏磊.高质量用户生成内容主题分布特征研究[J].图书馆杂志,2018,37(4):95-101.

12 王闯,王亚民.基于K核分解的网络知识社区关键用户挖掘研究[J].情报理论与实践,2019,42(6):149-153.

13 王英杰.在线知识社区用户协同价值共创情境构建研究[J].情报科学,2021,39(4):30-36.

14 Yan Z,Wang T,Chen Y,et al.Knowledge Sharing in Online Health Communities: A Social Exchange Theory Perspective [J].Information & Management,2016,53(5):643-653.

15 Zhao J, Wang T, Fan X C. Patient Value Co-Creation in Online Health Communities Social Identity Effects on Customer Knowledge Contributions and Membership Continuance Intentions in Online Health Communities[J].Journal of Service Management,2015,26(1):72-96.

16 Oh S. The Characteristics and Motivations of Health Answerers for Sharing Information,Knowledge,and Experiences in Online Environments [J].Journal of the American Society for Information Science & Technology,2012,63(3):543-557.

17 彭昱欣,邓朝华,吴江.基于社会资本与动机理论的在线健康社区医学专业用户知识共享行为分析[J].数据分析与知识发现,2019,3(4):63-70.

18 Coulson N S.Sharing, supporting and sobriety: A qualitative analysis of messages posted to alcohol-related online discussion forums in the United Kingdom[J].Journal of Substance Use, 2014, 19(1-2):176-180.

19 Attard A, Coulson N S. A thematic analysis of patient communication in Parkinson's disease online support group discussion forums[J]. Computers in Human Behavior, 2012, 28 (2):500-506.

20 Zhang J, Zhao Y M.A user term visualization analysis based on a social question and answer log[J]. Information Processing & Management,2013,49(5):1019-1048.

21 金碧漪,许鑫.网络健康社区中的主题特征研究[J].图书情报工作,2015,59(12):100-105.

22 Yi M Y, Yoon J J, Davis J M, et al. Untangling the antecedents of initial trust in Web - based health information:The roles of argument quality, source expertise, and user perceptions of information quality and risk[J].Decision Support System, 2013, 55(1):284-295.

23 范昊,张玉晨,吴川徽.网络健康社区中健康信息传播网络及主题特征研究[J].情报科学,2021,39(1):4-12,34.

24 翟羽佳,张鑫,王芳.在线健康社区中的用户参与行为——以"百度戒烟吧"为例[J].图书情报工作,2017,61(7):75-82.

25 LeBesco K.Book Review:Online Social Support:The Interplay of Social Networks and Computer-Mediated Communication [J]. Journal of Language and Social Psychology, 2008,27(3):312-314.

26 Beaudoin C E, Tao C C. Modeling the Impact of Online Cancer Resources on Supporters of Cancer Patients[J].New Media & Society,2008,10(2):321-344.

27 Biyani P, Caragea C, Mitra P, et al. Co-Training over Domain-Independent and Domain-Dependent Features for Sentiment Analysis of an Online Cancer Support Community[C]// Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,2013.

28 Pielou B C. Ecological diversity[M]. New York:John Wiley & Sons,1975.

29 Chin-Chi Hsu, Yi-An Lai, Wen-Hao Chen, Ming-Han Feng, Shou-De Lin. Unsupervised Ranking using Graph Structures and Node Attributes[C]. Proceedings of the 10th ACM International Conference on Web Search and Data Mining,2017:771-779.

30 张晋.城市轨道交通线网结构特性研究[D].北京:北京交通大学,2014.

31 张军.网络舆情时变演化机制及应对策略研究[M].北京:中国社会科学出版社,2020.

(责任编辑:赵红颖)

Key User Identification of Online Health Community Based on Multi-Dimensional Attribute Fusion

ZHANG Jun, LI Xin-wang, LI Peng

(School of Management, Shandong University of Technology, Zibo 255012*, China)*

Abstract:【Purpose/significance】Online health community has become an important form for the public to obtain medical information and services. This paper explores the identification of key users and their characteristics in online health community, with a view to provide theoretical basis for improving the quality and efficiency of health community service.【Method/process】Based on the information behavior theory, a multi-dimensional analysis framework was constructed, including interactive behavior attributes, information quality attributes and emotional tendency attributes. AttriRank algorithm and network invulnerability evaluation method were used to identify key users of online health community.【Result/conclusion】The results showed that 15 key users were identified in the QQ group of biliary cancer diseases and they not only have the characteristics of high activity and reciprocity, but also have the characteristics of high diversity and balanced information quality, and most of them have positive emotional tendencies. The analysis framework of "behavior + content + emotion" and the user ranking algorithm considering attributes can accurately identify the key users of the online health community, which provides scientific decision support for the sustainable operation of online health community.【Innovation/limitation】Build a multi-dimensional attribute analysis framework for online health community key user identification, which enriches the theoretical system of online health community key user identification.

Keywords: online health community; key users; attribute fusion; AttriRank algorithm; network invulnerability evaluation