



数据驱动的在线健康社区 用户信息行为研究

A Data-Driven Study of
User Information Behavior
in Online Health Communities

赵月华 等 著



上海社会科学院出版社
SHANGHAI ACADEMY OF SOCIAL SCIENCES PRESS



数据驱动的在线健康社区 用户信息行为研究

A Data-Driven Study of
User Information Behavior
in Online Health Communities

赵月华 等著



上海社会科学院出版社
SHANGHAI ACADEMY OF SOCIAL SCIENCES PRESS

图书在版编目(CIP)数据

数据驱动的在线健康社区用户信息行为研究 / 赵月华等著 .— 上海 : 上海社会科学院出版社, 2024
ISBN 978 - 7 - 5520 - 4305 - 1

I. ①数… II. ①赵… III. ①智能技术—应用—社区卫生服务—研究 IV. ①R197.1 - 39

中国图家版本馆 CIP 数据核字(2024)第 026848 号

数据驱动的在线健康社区用户信息行为研究

著 者: 赵月华、苏新宁、许 鑫

责任编辑: 张 晶

封面设计: 霍 覃

出版发行: 上海社会科学院出版社

上海顺昌路 622 号 邮编 200025

电话总机 021 - 63315947 销售热线 021 - 53063735

<https://cbs.sass.org.cn> E-mail: sassp@sassp.cn

排 版: 南京展望文化发展有限公司

印 刷: 上海龙腾印务有限公司

开 本: 710 毫米×1010 毫米 1/16

印 张: 20.5

字 数: 364 千

版 次: 2024 年 3 月第 1 版 2024 年 3 月第 1 次印刷

ISBN 978 - 7 - 5520 - 4305 - 1/R · 073

定价: 98.00 元

版权所有 翻印必究

国家社科基金后期资助项目 出版说明

后期资助项目是国家社科基金设立的一类重要项目,旨在鼓励广大社科研究者潜心治学,支持基础研究多出优秀成果。它是经过严格评审,从接近完成的科研成果中遴选立项的。为扩大后期资助项目的影响,更好地推动学术发展,促进成果转化,全国哲学社会科学工作办公室按照“统一设计、统一标识、统一版式、形成系列”的总体要求,组织出版国家社科基金后期资助项目成果。

全国哲学社会科学工作办公室

目 录

第 1 章	绪论	1
1.1	研究背景	1
1.2	研究意义	3
1.3	研究思路	4
1.4	研究方案	6
1.5	研究方法	7
1.6	研究数据	9
1.7	研究内容	13
第 2 章	在线健康社区分析框架	15
2.1	信息维度的在线健康社区分析	16
2.2	用户维度的在线健康社区分析	34
2.3	社区维度的在线健康社区分析	59
2.4	本章总结	74
第 3 章	基于特征的角色识别及用户行为模式探测	78
3.1	在线健康社区用户分类研究	78
3.2	研究方法	81
3.3	用户分布	88
3.4	用户角色识别	92
3.5	用户角色识别及行为模式分析	98
3.6	本章总结	109
第 4 章	基于信息交互的意见领袖识别及群组探测	110
4.1	在线健康社区信息交互行为分析及群组探测	110
4.2	本章研究方法	115

4.3	信息交互行为统计分析	122
4.4	信息交互网络分析	128
4.5	意见领袖识别及特征分析	139
4.6	用户群组分析	152
4.7	本章总结	163
第5章	基于信息交互内容的主题识别及演化探测	166
5.1	基于用户生成内容的主题分析	166
5.2	本章研究方法	177
5.3	基于信息交互内容的主题及特征词分布	184
5.4	基于信息交互内容的主题演化分析	199
5.5	基于交互内容的用户贡献度分析	207
5.6	基于交互内容的用户行为模式分析	214
5.7	本章总结	219
第6章	基于信息交互的社会情感支持识别及用户类型探测	222
6.1	在线社交平台社会及情感支持	222
6.2	本章研究方法	229
6.3	在线健康社区社会情感支持分布	235
6.4	基于用户类型的用户行为模式分析	243
6.5	本章总结	254
第7章	基于用户角色和主题识别的用户行为探测	256
7.1	本章研究方法	256
7.2	用户行为模式分析	265
7.3	本章总结	273
第8章	后记	276
8.1	在线健康社区分析框架构建	276
8.2	基于特征的角色识别及用户行为模式探测	277
8.3	基于信息交互的意见领袖识别及群组探测	278
8.4	基于信息交互内容的主题识别及演化探测	278
8.5	基于信息交互的社会情感支持识别及用户类型探测	279
	参考文献	281

图 目 录

图 1-1	研究思路	5
图 1-2	在线健康社区架构	6
图 1-3	自闭症吧主页示例	12
图 1-4	主要研究内容	14
图 2-1	在线健康信息内容研究框架	21
图 2-2	在线健康信息传播研究框架	28
图 2-3	在线健康社区信息评价研究框架	33
图 2-4	在线健康社区患者视角研究框架	49
图 2-5	在线健康社区医生视角研究框架	55
图 2-6	在线健康社区医患关系视角研究框架	59
图 2-7	在线健康社区模式与价值研究框架	63
图 2-8	在线健康社区风险与价格研究框架	66
图 2-9	在线健康社区应用现状与发展研究框架	69
图 2-10	在线健康社区平台评价研究框架	73
图 2-11	在线健康社区研究内容框架	76
图 3-1	研究基本流程	81
图 3-2	角色识别模型的平均性能	86
图 3-3	用户等级分布	89
图 3-4	用户性别分布	89
图 3-5	用户活跃时长分布图	91
图 3-6	用户最近活动时间分布图	92
图 3-7	用户发帖情况	100
图 3-8	用户发帖回复情况	102
图 3-9	用户发帖占比变化	105
图 3-10	用户活跃情况	108
图 4-1	本章研究流程图	121

图 4-2	发帖量分布图	123
图 4-3	回帖量分布图	124
图 4-4	主帖回复数量分布	125
图 4-5	多级回复数量分布	126
图 4-6	整体网络结构图-UCINET	128
图 4-7	发帖和回帖数量分布	132
图 4-8	参与互动的用户数量分布	133
图 4-9	交互网络整体结构图	134
图 4-10	网络密度变化	136
图 4-11	网络聚类系数变化	136
图 4-12	网络平均距离变化	137
图 4-13	基于角色的整体交互网络	138
图 4-14	基于角色的整体中心交互网络	138
图 4-15	各类用户内部交互网络	139
图 4-16	意见领袖交互网络整体结构图	151
图 4-17	用户群组聚类	153
图 4-18	“子群 15”内部的骨干交互网络	154
图 4-19	“子群 7”内部的骨干交互网络	155
图 4-20	“子群 16”内部的骨干交互网络	155
图 4-21	用户群组子结构变化	156
图 4-22	2017 年 1 月—2017 年 6 月的凝聚子群	157
图 4-23	2017 年 1 月—2017 年 6 月的凝聚子群	157
图 4-24	2017 年 7 月—2018 年 2 月的狭长子群演化	158
图 4-25	2018 年 3 月—2018 年 11 月的扇形讨论小组演化	158
图 4-26	2018 年 3 月—2018 年 11 月的凝聚子群	159
图 4-27	2018 年 12 月—2019 年 5 月的凝聚子群	159
图 4-28	2018 年 12 月—2019 年 5 月的凝聚子群形态	159
图 4-29	规模排在前 3 位的子群用户的重合度变化	161
图 4-30	不同规模的凝聚子群涉及用户人数	162
图 5-1	知网中关于“用户生成内容”主题的文献数量趋势图	166
图 5-2	LDA 主题模型的主题生成结构图	168
图 5-3	本章研究流程图	177
图 5-4	主题数——一致性曲线	180
图 5-5	时间切片相关设置	182

图 5-6	管状图示例	182
图 5-7	矩阵转换示意图(id 表示用户, tp 表示主题)	184
图 5-8	主帖主题聚类可视化效果	194
图 5-9	回复帖主题聚类可视化效果	195
图 5-10	大主题下的小主题数量分布	199
图 5-11	主帖的 50 个主题数量分布	199
图 5-12	主帖的大主题数量分布	200
图 5-13	回复帖的 50 个主题数量分布	200
图 5-14	回复帖的大主题数量占比分布	200
图 5-15	主帖主题获得回复量分布	202
图 5-16	主帖主题与回复帖主题对应情况	203
图 5-17	主帖大主题与回复帖大主题对应情况	203
图 5-18	回复帖大主题月演化情况	205
图 5-19	回复帖的大主题各时期讨论数量趋势	206
图 5-20	主帖的主题用户最大贡献度	207
图 5-21	回复帖的主题用户最大贡献度	212
图 5-22	核心用户可视化结果	215
图 5-23	Netdraw 用户聚类图	218
图 6-1	困惑度—主题数量曲线	230
图 6-2	SEE - k 值	232
图 6-3	聚类效果图	235
图 6-4	机器学习结果分布	236
图 6-5	各类别共现弦图	236
图 6-6	不同类型情感值对比	238
图 6-7	5 种社会支持类型在 3 种帖子中的分布	240
图 6-8	3 种帖子类型的平均情感值对比	241
图 6-9	主帖与回复帖不同类型帖子相应回复的类型分布图	242
图 6-10	用户发帖篇数平均值及分布情况	244
图 6-11	用户收到回帖篇数平均值及分布情况	245
图 6-12	用户活跃时长及活跃天数平均值及分布情况	246
图 6-13	主题分布情况	249
图 6-14	用户情感值平均值、标准差及分布情况	250
图 6-15	用户交互网络	252
图 6-16	各社群相对类别占比分布	253

图 7-1	本章研究流程	257
图 7-2	t-SNE 聚类结果	263
图 7-3	pyLDAvis 结果	264
图 7-4	不同角色用户发布主帖情况	266
图 7-5	不同角色用户发布回复帖情况	267
图 7-6	主帖和回复帖分布	267
图 7-7	不同角色用户发布主帖的回复量情况	268
图 7-8	不同角色积分等级值	269
图 7-9	用户活跃时间	270
图 7-10	不同角色主题分布	272

表 目 录

表 3-1	训练集标注结果	82
表 3-2	二分类混淆矩阵	84
表 3-3	角色识别最优结果	87
表 3-4	用户信息特征	87
表 3-5	用户发帖情况分布	90
表 3-6	用户回帖情况分布	90
表 3-7	用户活跃时间分布	91
表 3-8	“自闭症患者及亲友”用户类型细分	93
表 3-9	“专业人士”用户类型细分	94
表 3-10	“第三方”用户类型细分	96
表 3-11	“其他无关人员”用户类型细分	96
表 3-12	训练集标记情况	97
表 3-13	训练集用户角色分布情况	98
表 3-14	用户角色识别结果	98
表 4-1	参与交互的用户量及发帖和回帖数量	122
表 4-2	发帖量前 20 名的用户	123
表 4-3	回帖量前 20 名的用户	125
表 4-4	被回复总量前 20 名的用户	126
表 4-5	平均每帖被回复数量前 20 名的用户	127
表 4-6	节点中心度前 10 名的用户	130
表 4-7	接近中心度前 10 名的用户	131
表 4-8	中间中心度前 10 名的用户	131
表 4-9	交互网络指标	135
表 4-10	节点中心度前 10 名的用户	140
表 4-11	节点中心度前 10 名的用户特征	140
表 4-12	节点中心度前 10 名的用户在自闭症吧内发帖	141

表 4-13	节点中心度前 10 名的用户在自闭症吧外发帖	142
表 4-14	接近中心度前 10 名的用户	145
表 4-15	接近中心度前 10 名的用户特征	145
表 4-16	接近中心度前 10 名的用户在自闭症吧内发帖	146
表 4-17	接近中心度前 10 名的用户在自闭症吧外发帖	148
表 4-18	中间中心度前 10 名的用户	149
表 4-19	中间中心度前 10 名的用户特征	149
表 4-20	调解用户与引导用户关系对照	150
表 4-21	中间中心度前 10 名的用户在自闭症吧内发帖	150
表 4-22	意见领袖交互网络指标	151
表 4-23	凝聚子群聚类结果	152
表 4-24	用户重合度	160
表 4-25	大规模子群涉及的用户	162
表 4-26	小规模子群涉及的用户	163
表 5-1	主题分析方法概览	169
表 5-2	数据字段结构说明	178
表 5-3	主帖的 50 个主题及各主题的前 30 个特征词分布	184
表 5-4	回复帖的 50 个主题及各主题的前 30 个特征词分布	189
表 5-5	主帖的 50 个主题的最大贡献用户及内容关键词	208
表 5-6	回复帖的 50 个主题的最大贡献用户及内容关键词	212
表 5-7	回复帖的最大贡献用户总结	214
表 5-8	10 位核心用户的相关信息	215
表 5-9	不同角色的用户主题多样性	217
表 5-10	聚类中的用户角色和共同主题分布	218
表 6-1	社会支持类型及网络结构分析相关研究	228
表 6-2	网络社会支持行为性别差异相关研究	228
表 6-3	社会支持类型示例	229
表 6-4	算法评价	231
表 6-5	预测结果	232
表 6-6	初始聚类中心	233
表 6-7	聚类结果	233
表 6-8	分类结果共现分析示例	237
表 6-9	非参数检验结果及成对比较	239
表 6-10	次级社群网络数据	253

表 7-1	数据字典	258
表 7-2	训练集标注结果	259
表 7-3	模型平均性能	261
表 7-4	所有用户角色分布	262
表 7-5	主题分布	263
表 7-6	用户发帖和回帖情况	265
表 7-7	不同角色用户发布主帖数和回帖数	266
表 7-8	不同角色等级积分情况	269
表 7-9	用户活跃时长情况	270
表 7-10	不同角色活跃情况	271
表 7-11	主题分布	271

第1章 绪 论

习近平总书记在党的十九大报告中指出,“人民健康是民族昌盛和国家富强的重要标志。要完善国民健康政策,为人民群众提供全方位全周期健康服务”;《“健康中国 2030”规划纲要》指出,当前是推进健康中国建设的重要战略机遇期,并将积极发展基于互联网的健康服务视为重要的战略目标之一。可见,全民健康问题已经成为国家和地方政府最为重视的问题之一。“互联网+”医疗和“健康中国 2030”等国家战略的先后出台,更是为网络医疗健康服务的快速发展提供了契机。随着健康 2.0 时代的到来,用户对网络医疗健康服务的需求日益增长;与此同时,随着越来越多的用户通过网络渠道获取健康信息以及健康信息服务,近年来健康信息学,尤其是消费者健康信息领域,迎来了蓬勃发展。本书以在线健康社区为研究对象,聚焦在线健康社区(Online Health Community, OHC)中的用户行为及信息交互,探讨、分析在线健康社区中的用户角色、用户交互行为、主题演化、社会情感支持等议题。

1.1 研究背景

根据美国国家医学图书馆(The United States National Library of Medicine, NLM)的定义,健康信息是指一般健康、药物和补品、特殊人群、遗传学、环境健康与毒理学、临床试验和生物医学等方面的文献与信息。^① 美国医学信息学协会(American Medical Informatics Association, AMIA)进一步将“健康信息用户”定义为搜寻有关保健、疾病的预防和治疗、各种健康状况管理及慢性病等信息的人。^② 帕特里克(Patrick)等人指出,“用户健康信息”是让

① 张进,赵月华,谭莹.国外社交媒体用户健康信息搜寻研究:进展与启示[J].文献与数据学报,2019,1(01):108—117.

② Suess S. Consumer Health Information[J]. *Journal of Hospital Librarianship*, Routledge, 2001, 1(4): 41—52.

个人能够理解其健康状况,并为自己和家人做出健康决定的信息,包括个人和社区健康的促进和加强、私人护理、分享(专家—患者)决策、患者教育、患者信息和康复、健康教育、对医疗保健系统的使用、对保险或医疗提供者的选择。^①

随着互联网用户群体的壮大,在线搜寻和获取健康信息的用户日益增多。美国健康信息用户的比例在2001年为15.9%,2007年大幅上升至31.1%,2010年达到32.6%。^②2015年的一项调查显示,苏格兰有68.4%的患者曾利用过在线健康信息。^③皮尤研究中心(Pew Research Center)发现,对特定疾病和医疗问题的关注主导了美国人的在线查询,16个主要的健康话题涵盖了各种具体疾病、健康饮食和医疗保险。涂(Tu)总结了用户搜寻健康信息的三类主要信息源:互联网、出版物(图书、杂志、报纸)和其他人(朋友、亲戚)。其中,在互联网上搜寻健康信息的比例从2001—2010年持续增长,互联网已经成为用户搜寻健康信息的主要来源。由此可见,网络健康信息在用户的自我健康管理中扮演着越来越重要的角色。

由此,越来越多的在线健康社区应运而生。在线健康社区是不断发展的在线社交媒体平台与不断增强的公民医疗健康意识共同促生的产物。由于传统医疗资源的获取成本较高且分布不均衡,除了传统的面对面医患沟通以外,更多的患者会通过浏览和参与各类在线健康社区的讨论来获取健康信息。对于在线健康社区而言,社区内的信息交互是影响用户信息采纳、需求满足和服务体验的重要因素,并且有可能对用户的健康状况产生深远影响。

与传统的以医生为中心的医疗健康服务模式相比,在线健康社区为用户提供了一个就健康医疗相关问题进行信息交流、经验分享、问答咨询及社会支持的开放式网络平台。^④在线健康社区由于其开放性以及交互性的特点,对于普通公众、患者,尤其是慢性疾病患者而言,能够提供一个有效的获取信息以及与其他用户或患者建立联系的平台。此外,除了信息的交互,通

① Niederdeppe J, Hornik R C, Kelly B J, Frosch D L, Romantan A, Stevens R S, Barg F K, Weiner J L, Schwartz J S. Examining the Dimensions of Cancer-related Information Seeking and Scanning Behavior[J]. *Health Communication*, 2007, 22(2): 153-167.

② Tu Ha T. Surprising Decline in Consumers Seeking Health Information[J]. *Tracking report*, 2011(26): 1-6.

③ Moreland J, French T L, Cumming G P. The Prevalence of Online Health Information Seeking Among Patients in Scotland: A Cross-Sectional Exploratory Study[J]. *JMIR Research Protocols*, 2015, 4(3): e85.

④ 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(09):134—142.

过与其他用户的交流,在线健康社区能够有效地为用户提供社会及情感支持,这对于患者的健康管理同样有着重要意义。

国务院《“十三五”卫生与健康规划》指出:“培育健康医疗大数据应用新业态”,“促进云计算、大数据、物联网、移动互联网、虚拟现实等信息技术与健康服务的深度融合,提升健康信息服务能力”。在“互联网+”医疗和“健康中国 2030”以及《“十三五”卫生与健康规划》等国家战略相继出台的大背景下,互联网健康医疗迎来了蓬勃发展的契机。大量互联网健康医疗服务和平台的涌现,也将为进一步推动全民健康的实现,助力健康中国战略的实施。

1.2 研究意义

近年来国内外对在线健康社区的研究已经得到了越来越多的关注。对相关文献进行梳理,可以发现国内关于在线医疗社区的研究起始于 2008 年,2015 年开始进入快速发展阶段。研究对象主要包括:在线医疗平台、在线患者论坛、在线问答平台上的与医疗健康相关的主题板块等。国外相关研究重点关注社交媒体上的医疗健康相关信息群组、专门的医疗健康在线讨论社区、针对特定疾病或健康问题的患者社区等。通过对在线健康社区中用户行为以及信息交互的挖掘,不仅能够真实地反映用户的健康信息需求,并且能够进一步了解用户在进行健康信息交互过程中的行为模式特征。然而,目前针对在线健康社区的研究往往集中于单一的视角,缺乏从多维视角将用户行为与信息交互进行融合分析的研究。因此,如何从非结构化的、用户生成的内容中挖掘用户的信息需求,如何通过交互内容和交互行为识别用户角色以及探析其行为模式,进而揭示在线健康社区对用户的信息支持以及社会及情感支持,并形成在线健康社区分析框架,就是本书研究内容的意义和价值。

1.2.1 理论意义

从学科发展角度来看,消费者健康信息学(Consumer Health Informatics)领域作为图书情报学、医学信息学、公共管理等学科的交叉领域,近年来进入了蓬勃发展的阶段。在国外,消费者健康信息学已经逐步成为较为成熟的学科,国内的相关研究数量也在迅速增长。因此,本书提出的在线健康社区分析框架及应用有利于丰富和发展消费者健康信息学领域的研究方法体

系,并为相关学科领域的研究提供研究方法层面的借鉴。

从用户行为的研究角度来看,以往的相关研究多通过用户实验来分析用户行为模式,而本书的研究则通过在线健康社区中用户生成的真实数据,能够以真实情境中的用户行为数据为基础,挖掘用户行为模式并探测用户角色以及用户需求。有关在线健康社区中不同角色用户的信息行为的挖掘和分析,对更加深入的用户健康信息行为研究具有借鉴意义。

1.2.2 实践及应用意义

对于用户而言,本书的研究能够更全面地挖掘用户的健康信息需求。通过对在线健康社区中用户交互信息的挖掘,能够发现用户的隐性信息需求,从而有利于网络知识发现和网络健康信息组织,更好地为用户提供全面的健康信息。此外,通过对用户角色的自动识别,能够发现不同角色用户的细分的信息需求,实现更加精准的健康信息推送。

对于网络健康信息服务平台而言,本书研究中的用户角色识别及信息行为特征挖掘能够应用于在线健康社区的平台建设实践中。通过更加系统地了解用户的信息需求和行为模式特征,以及对用户身份角色和用户社区角色的识别,能够为平台建设、机制设计等提供科学参考,制订更加合理的激励机制以增强社区内的用户参与度,从而推动平台的发展并提升用户对社区平台的黏性。

对于国家健康事业而言,对在线健康社区的深入挖掘有利于充分发挥出网络健康信息资源对于提高公众健康信息素养和健康水平的作用,从而在一定程度上缓解医疗资源分布不均衡所导致的诸多社会问题。此外,对网络健康信息资源的内容挖掘能够为相关医疗和网络信息监管部门的制度建设提供思路,引导在线健康医疗服务、电子健康和智慧医疗相关实践探索进一步走向规范和成熟。

1.3 研究思路

在线健康社区是以交互性、社区化、共享性、知识性等为特点的网络时代的产物,其核心是由信息、用户和社区三个要素共同组成的。因此,如图 1-1 所示,本书各章节聚焦于三个要素之间的相互作用,并通过多种数据分析方法,从多维度对在线健康社区中的用户角色、主题分析、用户交互行为、社会与情感支持进行研究。