

基于文本挖掘的互联网医疗平台用户画像模型构建^{*}

吕艳华 王康龙 钟小云 陈俊冶

(山西医科大学管理学院 太原 030600)

〔摘要〕 目的/意义 构建互联网问诊用户画像探究问诊主题,提升问诊服务质量,减少医患沟通障碍,以线上配合线下方式针对性治疗。方法/过程 利用 Python 爬虫获取好大夫在线医疗平台孤独症疾病问诊数据,运用隐含狄利克雷分布(latent Dirichlet allocation, LDA)与词频-逆文档频率(term frequency-inverse document frequency, TF-IDF)结合的模型划分数据,在降维聚类后实现用户群体分类。最后通过 logistic 回归模型计算输出不同用户群体特征集合,构建画像。结果/结论 用户问诊内容主要围绕 11 个主题展开,平台可通过主题内容的典型特征优化问诊填写模板,提高用户填写疾病描述准确性、问诊效率和患者满意度。

〔关键词〕 互联网医疗平台;主题模型;文本分析;用户画像

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.07.002

Construction of the User Portrait Model of Internet Medical Platform Based on Text Mining

LYU Yanhua, WANG Kanglong, ZHONG Xiaoyun, CHEN Junye

School of Management, Shanxi Medical University, Taiyuan 030600, China

〔Abstract〕 **Purpose/Significance** The internet consultation user portrait is constructed to explore the consultation topic, improve the consultation service quality, reduce the communication barriers between doctors and patients, and provide targeted treatment in an online and offline manner. **Method/Process** Python crawler is used to obtain the autism diagnosis data of a medical platform, and the combined model of LDA and TF-TFIDF is used to divide the data, and the user group classification is realized after dimensionality reduction clustering. Finally, the characteristic sets of different user groups are calculated and output by logistic regression model to construct the portrait. **Result/Conclusion** The consultation content of users mainly focuses on 11 topics. The platform can optimize the consultation filling template based on the typical characteristics of the subject content to improve the accuracy of the disease description, consultation efficiency and satisfaction of patients.

〔Keywords〕 internet medical platform; topic model; text analysis; user portrait

1 引言

〔修回日期〕 2023-12-03

〔作者简介〕 吕艳华,副教授,硕士生导师,发表论文 50 余篇。

〔基金项目〕 国家社会科学基金一般项目(项目编号:20BTK064)。

互联网医疗平台是指利用互联网和信息技术为用户提供医疗服务和健康管理在线平台。近年来随着信息技术不断发展和各类电子设备普及,我国

网民规模不断增长^[1]，互联网医疗行业日渐繁荣，涌现出丁香医生、好大夫在线、微医等具有影响力的互联网医疗平台。但是由于平台准入门槛较低，部分用户并不具备准确描述病情的能力，患者病情描述模糊在远程诊断中给医生工作带来困扰。为有效解决该问题，本文使用主题模型对患者描述分类，根据用户特点进行画像建模，提出平台改进策略，增强用户描述准确性，确保医生诊疗准确性和效率，优化患者就诊体验。

2 研究综述

2.1 互联网医疗平台相关研究

互联网医疗平台发展给患者、医院带来便利。探讨如何深入优化医疗平台体系、提升用户满意度已成为重点^[2]。在优化医疗平台体系方面，张宇等^[3]发现互联网医疗平台存在体制制约、商业模式不成熟、用户综合素养低等问题，并提出可能的解决方案；刘华^[4]则以互联网医疗平台用户和潜在用户为切入点，分析用户使用情况，根据不同用户特征向平台提出不同发展策略和建议。在提升用户满意度方面，李旭光等^[5]深度剖析用户知识互动与情感交互内容，揭示平台用户之间交互关系，进而优化用户互动引导模式；李月琳等^[6]利用平台患者问诊交互数据，分析影响交互效率的因素，帮助优化用户交互体验；尹相森等^[7]通过研究初次使用行为，构建患者初次使用互联网医疗平台的影响机制模型、可视化体验指数，推动平台不断完善。

2.2 医疗平台画像相关研究

“互联网+医疗”不断发展促进了用户画像技术在互联网医疗平台研究中的应用。画像技术在预测用户行为和发现用户潜在需求等方面具有重要作用，一直是管理决策的有效依据^[8-9]。互联网医疗平台画像的研究意义在于满足不同用户群体信息需求，提高用户忠诚度和留存率^[10-11]。滕春娥等^[12]通过综合标签、词云等方法构建模型提取患者需求，为医疗平台画像构建提供指导；翟姗姗等^[13]依据医疗平台健康信息推荐的不同场景，构建不同场景下的融合知识图

谱和病情画像信息推荐方案，辅助患者精准查找信息；盛姝等^[14]构建在线健康社区的用户画像并挖掘主题特征，解释不同角色群体用户行为差异及特征；高靖超等^[15]构建在线医疗社区健康焦虑画像模型探究用户情感及特征表现，为平台管理健康焦虑用户提供指导；刘忠宝等^[16]利用微博抑郁症患者数据构建画像为其提供合适的阅读材料，对抑郁症治疗起推动作用；江皓^[17]通过构建医生患者的双边画像建立医患匹配规则，尽可能避免医生选择的盲目性和随意性，实现精准的医生推荐服务。

2.3 相关研究评述

目前关于医疗平台画像的研究主要侧重于优化平台信息推荐精度和提升用户留存率，但很少有研究聚焦于用户问诊描述模糊带来的问题，缺少真正意义上以患者为信息主体的研究。此外，先前多是研究普通疾病，对于由家属替代患者就医的特殊案例，例如孤独症则鲜少提及。因此，本文聚焦于用户在互联网医疗平台填写问诊描述的质量，促进医患沟通，提高医生判断准确性和效率。借鉴前述学者研究成果，结合问诊描述特性，探索构建更全面、个性的用户画像，更好地理解患者需求，优化患者就诊体验，推动互联网医疗领域进一步发展。

3 平台画像研究设计：以孤独症疾病板块为例

3.1 构建流程设计

研究流程，见图 1，其中涉及两个重要步骤。一是用户分类，目前关于用户画像的研究多采用定性分析实现用户分类，即通过用户特征的相似性划分个体^[18-19]。本文根据疾病描述，确定用户关注主题内容，实现用户群体分类。采用多维标度法 (multidimensional scaling, MDS)，将原始高维数据映射到低维空间，使数据在低维空间中更易于被观察和理解^[20-21]。使用 K-means 聚类算法，在低维空间可视化数据聚类结构，以识别潜在群体类别。二是群体用户特征提取。采用 logistic 回归模型得到特征属性和主题内容间关系，确定最大概率群体属性集合，作为该用户群体的典型特征^[15,22-23]。

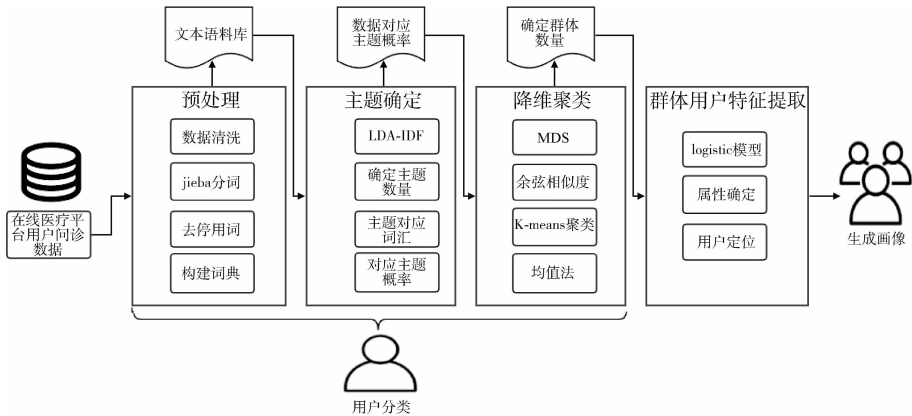


图1 互联网医疗平台画像构建流程

3.2 数据获取与处理

选取好大夫在线医疗平台问诊数据为数据源。孤独症是一种神经发育障碍，治疗主要侧重于早期干预，包括行为疗法、语言疗法、社交技能训练等，以及针对特定症状的药物治疗。然而，由于孤独症个体异质性强，早期筛查和诊治过程更为复杂。因此用户在互联网医疗平台问诊时对疾病描述的准确性尤为重要。通过爬虫程序抓取用户疾病描述、特征属性等问诊数据，总计8 314条（截至2023年7月20日），清洗后保留有效数据7 802条。基于患者年龄分为5个类别：1（0~3岁）、2（4~6岁）、3（9~12岁）、4（13~15岁）、5（15岁以上），性别分为两类：1（男）、2（女）。由于问诊用户均为患者家属，患者年龄以家属填写为准。收集到的年龄为7~8岁的患者数据量少且为无效数据，故不做处理。后续特征均采用上述变量表示。

3.3 主题内容获取

运用隐含狄利克雷分布（latent Dirichlet allocation, LDA）主题模型分析问诊数据的潜在主题数量。针对传统LDA模型中高频常用词与专有名词低频现象，采取LDA与词频-逆文档频率（term frequency-inverse document frequency, TF-IDF）相结合的方式改善词袋模型的词频问题^[24]；在分词过程中剔除“今天”“呢”“吗”等无意义词汇，将“利培酮”“肠道菌群”“粪移植”等词添入字典提升分词准确性。通过Python中SK-learn库构建模型，通常模

型分类最优主题数根据困惑度大小决定，但容易受噪声词汇干扰^[25]，因此采用主题一致性衡量主题数，考虑主题中词汇之间相关性，使主题质量评估更符合公众对主题的理解^[26]。将数据导入模型，运行得出结果，见图2，当主题数为11时，一致性分数最大，因此设定主题数为11，得出每个主题的主要内容和每条数据对应的主题概率，见表1。

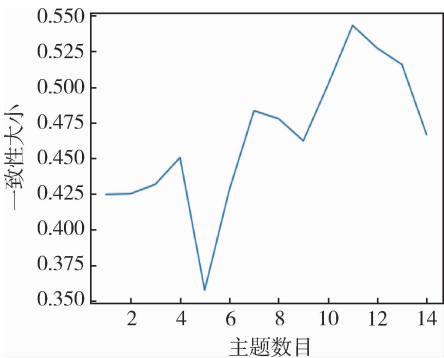


图2 主题一致性曲线

表1 主题对应内容

主题	对应内容
主题1	医学检查结果问题
主题2	疾病影响问题
主题3	中度孤独症问题
主题4	家庭影响问题
主题5	预后问题
主题6	粪菌移植影响问题
主题7	菌群失调及倾向问题
主题8	儿童情绪问题
主题9	日常生活状态问题
主题10	生活叛逆问题
主题11	异常举动问题

3.4 用户群体分类

根据主题提取结果，获取每条数据在 11 种主题中的概率，以 11 维概率向量表示。计算概率向量之间的距离，得到用户间相似度。计算向量相似度常用方法包括余弦相似度、欧氏距离、曼哈顿距离等。本文采用余弦相似度，通过计算两个向量之间的夹角余弦值度量其相似性^[27]：

$$\cos(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2}$$

(1)

依据主题间相似度降维，继而运用降维数据实施 K-means 聚类，使用手肘法和碎石图选定最优聚类数。当 K > 5 时，SSE 的下降幅度放缓，整图在此形成“手肘”状，见图 3，因此确定 K = 5 时为最优聚类效果。经 K-means 聚类后得出问诊数据分组，将结果进行可视化展示，见图 4。随后利用均值法计算每个簇的中心点和对应用户^[28-29]，将簇的中心用户确定为该类代表性用户，最终得出代表性用户和对应主题概率，见表 2。

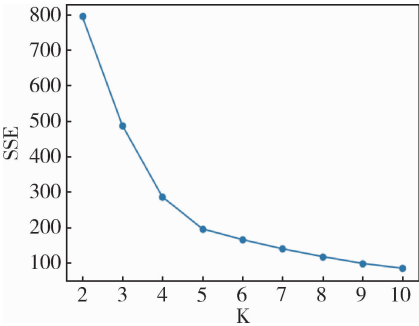


图3 误差平方和曲线

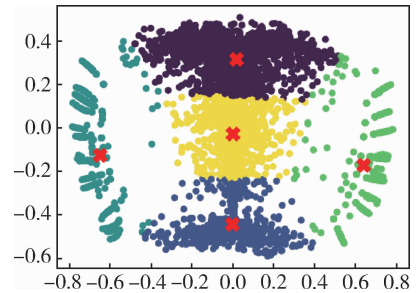


图4 K-means 聚类结果

表 2 簇团及中心点用户对应主题概率

主题	簇团 1	簇团 2	簇团 3	簇团 4	簇团 5
	用户 1677	用户 7608	用户 446	用户 2365	用户 1800
主题 1	0.039 354 57	0.546 461 12	0.029 138 01	0.039 287 22	0.292 025 37
主题 2	0.039 354 14	0.466 367 56	0.029 132 18	0.039 273 60	0.018 885 82
主题 3	0.039 354 14	0.028 241 72	0.029 132 18	0.039 274 27	0.251 930 70
主题 4	0.039 354 14	0.028 239 49	0.029 132 18	0.039 273 60	0.018 885 82
主题 5	0.039 354 24	0.028 239 44	0.029 132 18	0.039 304 88	0.018 886 48
主题 6	0.039 354 31	0.199 379 81	0.029 132 18	0.039 273 60	0.018 886 37
主题 7	0.039 354 27	0.028 239 43	0.029 132 18	0.607 205 36	0.018 886 02
主题 8	0.039 354 65	0.028 239 43	0.271 431 83	0.039 273 90	0.018 886 59
主题 9	0.606 457 22	0.028 240 54	0.029 137 35	0.039 285 26	0.304 954 67
主题 10	0.039 354 19	0.028 239 44	0.029 132 18	0.039 274 13	0.018 885 73
主题 11	0.039 354 14	0.028 240 15	0.029 132 18	0.039 274 17	0.018 886 43

3.5 画像特征提取

确定用户群体后，提取群体特征属性。之所以选取性别与年龄作为关键属性，是因为研究^[30]发现孤独症是大脑发育异常的结果，儿童大脑发育更容易受影响，且症状会伴随年龄增长而改变；此外有研究^[31]表明孤独症男性发病率明显高于女性，男孩

患孤独症概率大约是女孩的 4 倍，孤独症与年龄、性别有潜在联系。对群体属性变量进行 logistic 回归分析，将数据总量的 80% 共 6 241 条作为训练集，其余 20% 共 1 561 条作为测试集，经模型训练预测得出最大概率群体属性集合，见表 3。依据社会认同理论^[32-33]，这种组合最能反映用户群体关键特征。

分析新方法,推动医疗平台优化,为医疗实践和患者服务提供有力支持。

作者贡献: 吕艳华负责提出研究思路与方法,修改论文;王康龙负责撰写论文,处理数据,修改论文;钟小云负责处理数据,修改论文;陈俊冶负责格式校对,修改论文。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 中国互联网络信息中心. 第 52 次中国互联网络发展状况统计报告 [EB/OL]. [2023 - 08 - 29]. https://www.cac.gov.cn/2023-08/29/c_1694965940144802.htm#/.
- 2 陈曦. 互联网医疗研究现状及未来展望 [J]. 人民论坛·学术前沿, 2017 (24): 40 - 47, 95.
- 3 张宇, 张梦梦, 程繁银. 网络医疗咨询的开展现状及对策研究 [J]. 重庆医学, 2017, 46 (9): 1286 - 1287.
- 4 刘华. 互联网医疗发展现状及前景调查分析——以北京市为例 [J]. 调研世界, 2019 (3): 21 - 25.
- 5 李旭光, 李珊珊, 刘一凡, 等. 综合型社交平台上的在线医疗健康社区中知识互动和情感交互的关系研究 [J]. 情报理论与实践, 2021, 44 (8): 103 - 111.
- 6 李月琳, 张建伟, 张姍. 螺旋式与直线式: 在线健康医疗平台用户与医生交互模式研究 [J]. 情报学报, 2021, 40 (1): 88 - 100.
- 7 尹相森, 李俊儒, 张云秋, 等. “互联网 + 医疗”环境下用户行为影响机制研究——基于在线医疗平台患者用户初次使用行为的分析 [J]. 价格理论与实践, 2022 (6): 109 - 112, 194.
- 8 高广尚. 用户画像构建方法研究综述 [J]. 数据分析与知识发现, 2019, 3 (3): 25 - 35.
- 9 徐芳, 应洁茹. 国内外用户画像研究综述 [J]. 图书馆学研究, 2020 (12): 7 - 16.
- 10 李星. 在线健康社区用户画像模型构建 [D]. 绵阳: 西南科技大学, 2020.
- 11 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.
- 12 滕春娥, 何春雨. 在线医疗社区用户画像构建与应用 [J]. 图书情报工作, 2021, 65 (12): 147 - 154.
- 13 翟姗姗, 胡畔, 潘英增, 等. 融合知识图谱与用户病情画像的在线医疗社区场景化信息推荐研究 [J]. 情报科学, 2021, 39 (5): 97 - 105.
- 14 盛姝, 黄奇, 郑姝雅, 等. 在线健康社区中用户画像及主题特征分布下信息需求研究——以医享网结肠癌圈数据为例 [J]. 情报学报, 2021, 40 (3): 308 - 320.

- 15 高靖超, 彭丽徽, 张艳丰, 等. 在线医疗社区健康焦虑用户画像模型构建及实证研究 [J]. 图书情报工作, 2023, 67 (16): 124 - 134.
- 16 刘忠宝, 赵文娟. 基于患者画像的抑郁症阅读材料推荐模型研究 [J]. 大学图书馆学报, 2021, 39 (1): 85 - 93.
- 17 江皓. 基于医生患者双边画像的在线健康社区医患匹配研究 [D]. 北京: 北京外国语大学, 2023.
- 18 杨柳. 基于在线评论语义分析的肿瘤疾病患者情感画像研究 [D]. 太原: 山西大学, 2021.
- 19 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.
- 20 汪建旭. 基于主题模型和用户相似度的推荐算法研究 [D]. 广州: 广州大学, 2022.
- 21 陈富国. 多维标度法的理论与方法 [J]. 心理科学通讯, 1990 (4): 38 - 42.
- 22 袁绮蕊. 基于 K - MEANS 的在线健康社区用户画像模型构建 [J]. 科技情报研究, 2021, 3 (4): 95 - 106.
- 23 王晓芳, 贾宗维. 基于 K - means 的 MOOC 学习行为分析及用户画像研究 [J]. 中国教育信息化, 2019 (1): 43 - 46.
- 24 黄杰晟. 基于 TF - IDF 与 LDA 的杜甫全集情感和主题分析 [J]. 现代计算机, 2021 (10): 77 - 82.
- 25 于本海, 卢畅. 在线健康社区信息主题特征及其潜在价值研究——基于 LDA 模型对“百度痛风病吧”案例的分析 [J]. 价格理论与实践, 2022 (3): 195 - 198.
- 26 董文慧, 熊回香, 杜瑾, 等. 基于学者画像的科研合作者推荐研究 [J]. 数据分析与知识发现, 2022, 6 (10): 20 - 34.
- 27 冯莎. 豆瓣电影评论文本的情感分析研究——基于 2017 年电影《乘风破浪》爬虫数据 [J]. 中国统计, 2017 (7): 30 - 33.
- 28 胡畔. 基于用户画像的慢性病在线健康社区场景化信息推荐研究 [D]. 武汉: 华中师范大学, 2021.
- 29 黄晓斌, 张明鑫. 在线健康社区青少年群体用户健康信息需求研究 [J]. 中华医学图书情报杂志, 2020, 29 (5): 37 - 47.
- 30 赵梅菊, 肖非, 邓猛. 自闭症儿童适应行为发展特点的实证研究 [J]. 教育学术月刊, 2015 (8): 75 - 81.
- 31 徐光兴. 自闭症的性别差异及其与认知神经功能障碍的关系 [J]. 心理科学, 2007 (2): 425 - 427.
- 32 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.
- 33 张莹瑞, 佐斌. 社会认同理论及其发展 [J]. 心理科学进展, 2006 (3): 475 - 480.
- 34 王松, 王卫霞. 互联网开放式创新社区用户画像构建及用户创新需求识别 [J]. 科技管理研究, 2023, 43 (7): 158 - 170.