

基于文本内容分析的互联网医疗平台信息质量评价研究^{*}

王 君¹ 姚 唐¹ 王亚妮²

(¹ 北京航空航天大学经济管理学院 北京 100191 ² 首都医科大学燕京医学院 北京 101300)

〔摘要〕 目的/意义 借助文本内容分析法对互联网医疗平台用户发布的信息质量进行评价研究。方法/过程 通过 Python 爬虫获取国内典型互联网医疗平台的文本内容，进行自然语言处理和聚类分析，提取特征指标设计逻辑回归模型并进行灰色关联度修正，从而构建信息质量评价指标体系。结果/结论 互联网医疗平台可重点从信息丰富度、信息发布者特征及用户交流互动方面加强信息的发布与呈现，引导用户发布高质量信息并快速识别有价值信息，促进互联网医疗平台的可持续发展。

〔关键词〕 互联网医疗；信息质量评价；文本内容分析；逻辑回归；灰色关联度修正

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.11.002

Study on Information Quality Evaluation of Internet Medical Platforms Based on Text Content Analysis
WANG Jun¹, YAO Tang¹, WANG Yani²

¹School of Economics and Management, Beihang University, Beijing 100191, China; ²Yanjing Medical College, Capital Medical University, Beijing 101300, China

〔Abstract〕 **Purpose/Significance** To evaluate and study the quality of information published by users of internet medical platforms using text content analysis method. **Method/Process** The text content of typical internet medical platforms in China is obtained by Python crawler, natural language processing (NLP) and clustering analysis are performed. Feature indexes are extracted and logistic regression model and grey relational degree correction are designed to construct an information quality evaluation index system. **Result/Conclusion** The information release and presentation of internet medical platforms can be strengthened from the aspects of information richness, characteristics of information publishers and user interaction, and users can be guided to release high-quality information and quickly identify valuable information, so as to promote the sustainable development of internet medical platforms.

〔Keywords〕 internet medical treatment; information quality evaluation; text content analysis; logistic regression; grey relational degree correction

1 引言

互联网医疗作为我国健康科技创新的重要组成部分，是互联网技术在医疗行业的落地应用。互联网医疗平台是以医疗健康信息交流为主的媒介，患者与医生可以通过其进行健康咨询、经验分享和交流互动等活动，对日常健康管理和疾病控制有着重要影响^[1-2]。目前国内比较知名的互联网医疗平台有“丁香园”“春雨医生”“好大夫在线”等。平台信息

〔修回日期〕 2023-08-02

〔作者简介〕 王君，教授，博士生导师，发表论文 80 余篇；通信作者：王亚妮，博士，讲师。

〔基金项目〕 首都医科大学自然培育项目（项目编号：PYZ22037）；北京市教委社科一般项目（项目编号：SM202310025001）。

质量是影响互联网医疗可持续发展的关键因素。然而互联网医疗平台在提供便捷资源的同时，面临着医疗数据庞大、信息质量参差不齐等问题^[3]，影响患者信息浏览和健康决策。提高互联网医疗平台信息质量，帮助患者更好地利用平台管理健康并预防疾病，使互联网医疗服务更好地满足患者需求，对互联网医疗平台的可持续发展具有重要意义。

目前学术界有关信息质量评价的研究已有一定基础^[4-5]。在互联网医疗平台方面，主要围绕平台信息服务质量^[6-7]、用户使用意愿^[8]及用户满意度^[9]进行研究。互联网医疗信息质量评价主要借助问卷调查^[10]、专家访谈^[11]等定性分析方法，采用文本内容分析法对信息质量进行评价的研究还比较欠缺。因此，本文基于国内典型互联网医疗平台数据，对用户发布的文本内容信息进行自然语言处理和聚类分析，提取信息质量评价的特征指标建立逻辑回归模型并进行灰色关联度修正，构建互联网医疗平台的信息质量评价指标体系，以期帮助互联网医疗平台构建合适的信息搜索规则、提高服务质量和效率，促进互联网医疗资源最大化利用。

2 资料与方法

2.1 数据来源

本研究聚焦“丁香园”论坛肿瘤医学板块用户

主页的信息及其发布的帖子信息。于 2021 年 9—12 月基于 Python 的 selenium 库编写程序爬取肿瘤医学板块的 29 300 篇帖子及其评论，通过对帖子进行清洗过滤，获得有效帖子 22 057 条，涉及发帖用户 10 725 名。爬取的信息主要有文本型和数值型两类，帖子信息包括标题、内容、发表时间、标签、评论、浏览量和点赞数、收藏数等；用户信息包括昵称、职业、等级、积分、既往发帖链接、粉丝数、作品总浏览量、帖子被收藏总次数等。

2.2 数据处理思路

首先对文本内容清洗和处理，借助北京大学语言计算与机器学习组提供的 pkuseg 多领域分词库对文本信息进行分词和词性标注；借助哈尔滨工业大学自然语言处理实验室创建的停用词表、百度停用词表和基于词频统计的人工选择停用词方法对数据进行停用词去除；然后根据词频 - 逆向文件频率 (term frequency - inverse document frequency, TF - IDF) 生成文档向量并基于 k - means 聚类算法进行文本聚类，结合聚类后的关键词抽取确定文本类别；最后抽取信息质量评价指标，利用 Python 开源库 statsmodels 提供的统计分析方法实现逻辑回归模型并进行灰色关联度修正，构建信息质量评价指标体系，见图 1。

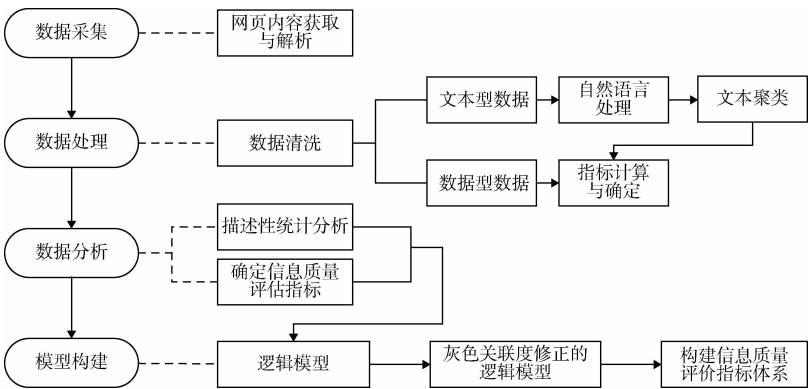


图 1 数据处理过程

2.3 数据预处理

pkuseg 分词工具致力于为不同领域数据提供个

性化的预训练模型。用户可以根据分词文本领域特点，自由选择不同模型^[11]。分词后，进行停用词去除。首先利用停用词表去除常规无用字词，接着采

用词频统计方法人工选择进一步去除停用词，最后基于词汇词性标注去除介词、副词、语气词、叹词、拟声词等对于文本分析无用的字词。

2.4 聚类分析

聚类分析可以帮助分析文本信息的大致分类属性。采用向量空间模型对每一个文档都用向量 d_j 表示：

$$d_j = (w_{1,j}, w_{1,j}, \cdots, w_{i,j}) \tag{1}$$

文档向量的每个维度对应一个词组。通过 TF-IDF 方法计算文档向量各维度的取值，选取权重较大的前 10 000 个词组作为 TF-IDF 各维度对应词组，构建针对整个文档集的 TF-IDF 矩阵。并采用基于 Python 的 numpy 库的 linalg.svd 方法对 TF-IDF 矩阵进行奇异值分解，在尽可能保留文档信息的情况下选取 $r=300$ ，得到原矩阵近似表示。TF-IDF 矩阵近似表示如下：

$$A_r = U_r \times W_r \times V_r^T \tag{2}$$

其中， U_r 矩阵是取 U 矩阵的前 r 列所得矩阵， W_r 矩阵是取对角阵 W 前 r 个对角元素所得对角阵， V_r 矩阵是选取 V 矩阵前 r 行所得矩阵。

利用构建的文档近似矩阵，采取基于 sklearn 库的聚类方法实现 k-means 聚类^[12]。在随机选取初始聚类中心的前提下，不同 k 值的聚类算法运行结果对应的轮廓系数不同。 k 取值为 6 时对应的轮廓系数最大，因此 k-means 聚类 k 值设置为 6。采用随机选取初始聚类中心并设定 k 值为 6 的条件将文档集聚成 6 类。对各类别文档进行词频统计，并根据高频词汇将各类别进行归纳。

2.5 指标提取及说明

借鉴医学信息和信息管理领域信息质量评价的相关研究，结合互联网医疗平台特征，选取一级指标和对应的二级指标，见表 1。

表 1 互联网医疗平台信息质量评价指标

一级指标	二级指标	说明
目标值	帖子的标签（target value, TV）	“丁香园”论坛对帖子标注的标签，分为优质和普通标签
帖子形式	完整性（completeness, CT）	帖子包含的数据种类，包括文字信息、图片信息、数值信息、链接
	丰富度（richness, R）	帖子包含的信息量和图片数，图文并茂的帖子被认为丰富帖
帖子内容	信息量（amount of information, AI）	帖子传递的有用信息量，即帖子经过分词、停用词等处理后所包含的字数之和
	可读性（readability, RE）	读者理解帖子文本内容所需能力
	主题相关度（topic similarity, TS）	帖子文本向量与类别中心的距离，采用余弦距离
	有效性（effectiveness, E）	帖子被其他用户收藏的数量，收藏量越高表明帖子越有效
用户浏览行为	受欢迎度（popularity, PO）	帖子被其他用户浏览的数量，浏览量越高表明帖子越受到用户欢迎
	一致性（consistency, CS）	帖子点赞数和收藏数，点赞数高且收藏数高则该帖子满足一致性
	评论相关度（comment relevance, CR）	帖子评论文本向量与帖子文本向量之间的距离，以相关度较高的评论占帖子总数的比例作为评论相关度
	互动性（interactivity, I）	其他用户对帖子的评论数量，即用户参与讨论的意愿
发布者属性	发布者权威（authorship, AS）	帖子发布者拥有的粉丝数、发布者在平台的等级与积分

2.6 模型构建

因变量为离散变量的计量模型称为离散被解释变量模型。在离散被解释变量模型中，逻辑回归模型因其具有简单、可并行化、可解释性强等特点而被广泛使用^[13]。逻辑回归模型思想是使个体为某一类别

的概率最大化，采用最大似然法进行参数估计。
灰色关联度分析是灰色系统理论中非常活跃的分支，其基本思想是根据各序列的相似程度来判断其之间联系是否紧密^[14]。灰色关联度分析可以从多角度对物品质量进行评价，具有操作性强、效果好等优点。基于互联网医疗信息质量评价指标，采用

逻辑回归模型和灰色关联度修正对影响信息质量评价的影响因素进行分析：

$$\ln P = \beta X^T \tag{3}$$

其中 P 为使帖子为高质量帖子的概率， β 为系数向量， X 为指标向量。

3 结果

3.1 数据处理结果

3.1.1 词云图展示 对文本进行分词和停用词处理后绘制词云图，见图 2。词云图主要是对文本数据进行视觉表示，通过不同的字体大小和颜色展示每个词的重要性，便于读者迅速直观地了解词的重要程度和文本内容主旨。“丁香园”肿瘤医学板块的帖子及评论出现频次最高的关键词主要涉及肿瘤、资讯、发现、情况、免疫、治疗等，这些关键词都与用户的健康咨询、知识分享、病例共享等内容和服务密切相关。



图2 基于文本内容的词云图构建

3.1.2 聚类分析结果 共得到 6 个聚类结果。第 1 类高频词包含肿瘤、研究、细胞、治疗、免疫、患者、癌症、基因、临床等，称为医学研究；第 2 类高频词包括下载、指南、链接、翻译、临床、肿瘤等，称为知识分享；第 3 类高频词包括治疗、患者、肿瘤、化疗、药物、手术、转移、方案、检查等，称为治疗方案；第 4 类高频词有患者、癌症、治疗、食物、作用、饮食、化疗、营养等，称为患者养护；第 5 类高频词有肿瘤、治疗、临床、手术、医师、化疗、内科、患者、解剖等，称为医患交流；第 6 类包含许多无医学含义词汇，归为杂项。对聚类后各类别分布情况进行分析，治疗方案

类帖子数量最多，医患交流类、知识分享类和杂项类帖子占比很小，见图 3。

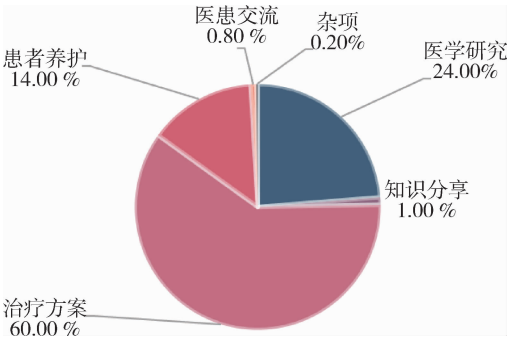


图3 聚类后的各类别分布

3.2 信息质量评价模型结果

在实证模型中，以帖子标签为被解释变量，提取到的二级指标为自变量，进行逻辑回归，见表 2。其中系数指各自变量对被解释变量的影响系数。对模型的整体检验 $\log - likelihood$ 值为 $-4\ 032.2$ ，说明自变量组合对被解释变量的影响具有统计学意义。在 $\alpha = 0.05$ 的显著性水平下，除可读性 (RE)、主题相关度 (TS) 及发布者粉丝数 (AS_1) 3 项指标外，其他指标对信息质量评价的影响均具有统计学意义 ($P < 0.05$)。

表2 信息质量评价模型（一）

因变量	系数	标准误	z	$P > z $
完整性 (CT)	0.173 2	0.087	2.001	0.045
信息量 (AI)	-0.000 3	3.72×10^{-5}	-7.173	0.000
丰富度 (R)	23.823 9	2.813	8.468	0.000
可读性 (RE)	-0.000 5	0.001	-0.880	0.379
主题相关度 (TS)	0.149 0	0.210	0.710	0.478
有效性 (E)	0.008 9	0.001	8.980	0.000
受欢迎度 (PO)	2.709×10^{-5}	9.16×10^{-6}	2.957	0.003
一致性 (CS)	1.679 7	0.064	26.268	0.000
评论相关度 (CR)	1.022 5	0.074	13.744	0.000
互动性 (I)	0.022 0	0.007	3.188	0.001
发布者粉丝数 (AS_1)	0.000 2	0.000	1.463	0.143
发布者等级 (AS_2)	0.108 7	0.019	5.825	0.000
发布者积分 (AS_3)	-0.002 7	0.000	-12.356	0.000
常数项	-3.115 6	0.278	-11.209	0.000

3.3 灰色关联度修正结果

对模型进行灰色关联度修正，见图 4。灰色关联度值大于 0.9 表示指标间具有较强的相关性，基于逻辑回归的结果剔除完整性（CT）、可读性（RE）、信息量（AI）、主题相关度（TS）及发布者粉丝数（AS_1）变量，对数据进行第 2 次逻辑回归，见表 3。结果显示所有变量均具有统计学意义

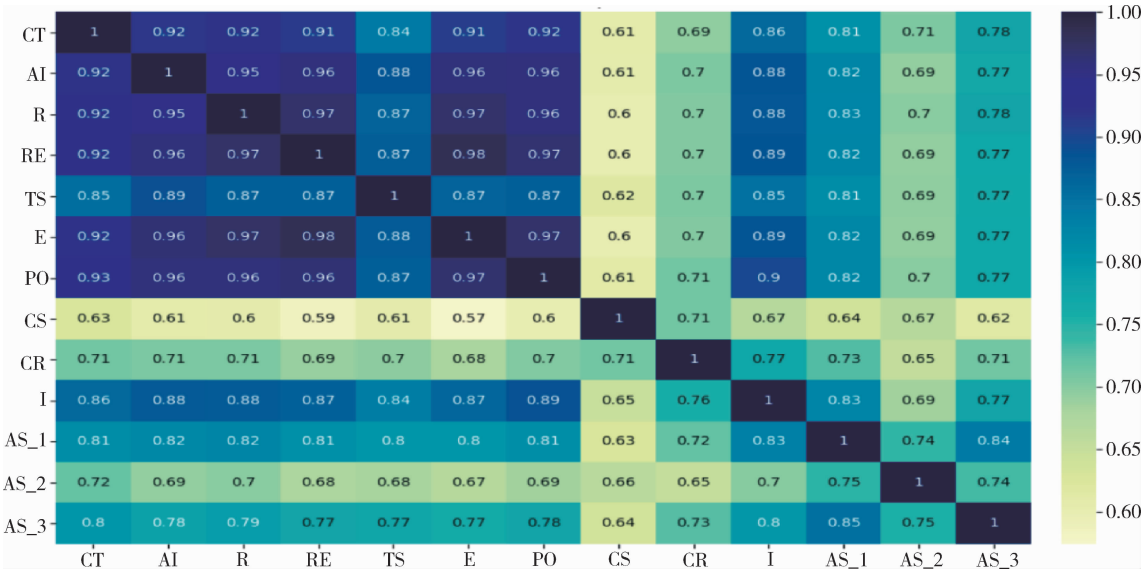


图 4 灰色关联度分析结果

表 3 信息质量评价模型（二）

因变量	系数	标准误	z	P> z
丰富度（R）	28.207 3	2.727	10.344	0.000
有效性（E）	0.007 1	0.001	8.041	0.000
受欢迎度（PO）	2.69 × 10 ⁻⁵	8.83 × 10 ⁻⁶	3.045	0.002
一致性（CS）	1.627 9	0.063	25.902	0.000
评论相关度（CR）	0.997 9	0.074	13.505	0.000
互动性（I）	0.025 6	0.007	3.831	0.013

($P < 0.05$)，修正前后回归模型的拟合系数 R^2 分别为 0.284 6 和 0.278 5，均方误差相近，表明两次回归结果对样本的拟合效果相近。灰色关联度修正之后的逻辑回归模型如下：

$$\ln P = 28.207\,3 \times R + 0.007\,1 \times E + 0.000\,026\,9 \times PO + 1.627\,9 \times CS + 0.997\,9 \times CR + 0.025\,6 \times I + 0.121\,8 \times AS_2 - 0.002\,6 \times AS_3 - 2.791\,4 \quad (4)$$

续表 3

因变量	系数	标准误	z	P> z
发布者等级（AS_2）	0.121 8	0.017	7.155	0.000
发布者积分（AS_3）	-0.002 6	0.000	-12.136	0.000
常数项	-2.791 4	0.077	-36.242	0.000

3.4 互联网医疗平台信息质量评价指标体系

基于逻辑回归和灰色关联度修正的模型分析后的信息质量评价指标体系，见图 5。

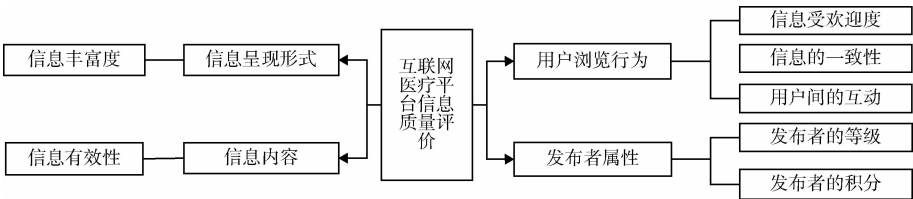


图 5 互联网医疗平台信息质量评价指标体系

4 讨论

4.1 互联网医疗平台在发布信息时应注重信息呈现的丰富度

互联网医疗平台中的信息特征在其质量评价中具有至关重要的作用,且主要体现在信息的丰富度方面。信息形式是用户浏览时的直观印象,对于尝试在平台上寻求信息支持的患者来说非常重要。互联网医疗平台应当采取措施鼓励用户发布信息量大、信息丰富多样的帖子,吸引更多用户参与平台互动,信息呈现的丰富度直接决定信息被接纳的难易程度^[15]。面对种类繁多的信息,互联网医疗平台发展不应局限于单一的文字信息形式,还可以发布如短视频、图片等有趣的信息吸引用户,让用户更愿意投入平台。

4.2 用户之间的交流互动有助于提升互联网医疗平台信息质量

互联网医疗平台上用户发布的信息被其他用户浏览并产生具体的互动如点赞、收藏和评论等行为时,此信息得到的正面评价会显著提升。用户之间的交流互动一方面使用户发布的信息被更多信息搜寻者知晓;另一方面,用户之间的交流互动也可以引导其他用户更精准、专业地贡献信息,平台用户作为医疗信息接收者的同时,也可作为信息提供者来分享知识,促进互联网医疗服务的发展^[16]。用户的积极参与和交流互动有助于实现互联网医疗平台信息价值。

4.3 信息发布者的权威性是互联网医疗平台信息质量评价的重要因素

发布者的权威性是指发布者对该领域知识的掌握及专业程度,身份权威发布者的帖子更具有说服力,对于其他用户来说更有价值^[17]。互联网医疗平台应注意采取适当措施控制用户积分的发放,因为当用户获得较高积分时,反而不利于其提供高质量信息。在信息过载时代,只有在源头保证信息质量,即确保信息发布者的权威性,平台中发布的信

息才会有参考和借鉴意义。平台可以让浏览信息的用户对发布信息进行打分,再结合发布者的特征如年龄、受教育程度、患病时长等加权得到该发布者的总体权威性得分。

5 结语

本研究以国内典型互联网医疗平台为研究对象,借助 Python 网络爬虫获取数据,采用自然语言处理和文本内容分析对用户发布的帖子进行聚类分析和特征提取,并设计了基于逻辑回归模型和灰色关联度修正实验的互联网医疗平台信息质量评价指标体系,有助于用户发布高质量信息并快速识别有价值的信息,也将有助于互联网医疗平台构建合适的信息搜索规则,以及提高信息服务质量和效率。

参考文献

- 1 MENG F, ZHANG X, LIU L, et al. Converting readers to patients? From free to paid knowledge – sharing in online health communities [J]. Information processing & management, 2021, 58 (3): 102490.
- 2 WU H, DENG Z, WANG B, et al. How online health community participation affects physicians' performance in hospitals: empirical evidence from China [J]. Information & management, 2021, 58 (6): 103443.
- 3 LI Y, FAN Z, YUAN X, et al. Recognizing fake information through a developed feature scheme: a user study of health misinformation on social media in China [J]. Information processing & management, 2022 (1): 59.
- 4 张克永,李贺.健康微信公众平台信息质量评价指标体系研究[J].情报科学,2017,35(11):143-149.
- 5 赵玉遂,许燕,吴青青,等.应用德尔菲法构建网络健康信息质量评价指标体系[J].预防医学,2018,30(2):121-124.
- 6 蒋知义,李巧,邢思佳,等.在线健康社区信息服务质量评价指标体系构建及实证研究[J].情报探索,2021,(4):29-36.
- 7 WU B. Patient continued use of online health care communities: web mining of patient – doctor communication [J]. Journal of medical internet research, 2018, 20 (4): e126.

(下转第 29 页)

5 结语

本文以医疗业务协同为切入点，以转诊业务为研究载体，构建集团双向转诊云平台，旨在推进集团医联体转诊业务的一体化、便捷化、规范化，确保医联体内部信息资源共享、管理服务协同、服务能力双赢。平台依托卫生服务专网，融合“云会诊”保障患者诊疗过程的连续性；实现检查、检验、治疗、处方医嘱协同互通，提升患者转诊全流程便捷性；规范线上审批流程，增强双向转诊过程规范性。基于区域协同的双向转诊模式为患者服务提供便利，在提升转诊效率的同时加强医联体医疗机构间的联系。未来将进一步拓展集团内医联体成员单位的双向转诊应用，形成统一、规范、便捷的服务模式。此外，平台与医联体医院间入院预处置的深度融合也是集团双向转诊云平台下一步拓展的研究方向。

参考文献

(上接第 12 页)

8

马骋宇,王启桢.在线医疗服务平台医生采纳行为及影响因素研究[J].中国卫生政策研究,2018,11(6):68-73.

9

YIN S, CAI X, WANG Z, et al. Impact of gamification elements on user satisfaction in health and fitness applications: a comprehensive approach based on the Kano model [J]. Computers in human behavior, 2022, 128 (3): 107106.

10

秦琴,柯青,谢雨杉,等.所见,所感与所知:用户的注意力,主观感受和在线健康信息质量评价关系探究[J].情报学报,2022,41(2):176-187.

11

张琪,范永胜.基于改进 T5 PEGASUS 模型的新闻文本摘要生成研究[J].电子科技,2022,11(4):1-7.

12

CHEN Z, WANG Y, ZHANG S, et al. Differentially private user-based collaborative filtering recommendation based on k-means clustering [J]. Expert systems with applications, 2021, 168 (15): 114366.

13

刘瑞平,苏思奇.含定性解释变量的 Logistic 回归模型的实例研究[J].北京信息科技大学学报(自然科学版),2022,37(4):88-93.

14

刘思峰.灰色系统理论创立,发展大事记(1982-2021)[J].南京航空航天大学学报(社会科学版),2022,24(4):39-40.

15

LIU S, LIAO H, PRATT J. Impact of media richness and flow on e-learning technology acceptance [J]. Computers & education, 2009, 52 (3): 599-607.

16

WANG J, YAO T, WANG Y. Patient engagement as contributors in online health communities: the mediation of peer involvement and moderation of community status [J]. Behavioral sciences, 2023, 13 (2): 152.

17

刘迎春,谢年春,李佳.虚拟学习社区中基于用户行为的知识贡献者信誉评价研究[J].现代情报,2020,40(3):117-125.

1

吴华丽,李婧,张红.基于分级诊疗的双向转诊服务模式设计与实现[J].中国数字医学,2018,13(6):16-18.

2

李佳林,申思哲,龙达.双向转诊模式及运行机制的探讨[J].中国中医药现代远程教育,2022,20(22):186-187.

3

印宏模.新时代紧密型医联体思想政治工作引领实践探索[J].中国卫生产业,2020,17(6):68-71.

4

董文勇.公医诊疗合作与我国复合转诊法制构建——健康治理现代化的训诫[J].河北法学,2020,38(10):115-137.

5

肖书琴.县域医共体建设中的利益相关者冲突研究——以广东省湛江市徐闻县为例[D].广州:华南理工大学,2020.

6

郝凤成,庄同义,王芳.新型医联体背景下的智慧医疗建设模式与路径研究[J].中国卫生产业,2020,17(7):163-165.

7

朱月兰,秦平,桑丘,等.区域卫生健康信息平台数据治理及应用[J].中国数字医学,2023,18(1):12-17.