安欣宇,陈育新,张　晗. 网络健康社区用户的新冠肺炎信息需求[J]. 中华医学图书情报杂志,2021,30(2):53-58.

DOI:10.3969/j.issn.1671-3982.2021.02.009

·信息组织与信息服务·

网络健康社区用户的新冠肺炎信息需求

安欣宇[1],陈育新[1,2],张　晗[1]

[摘要]搜集网络健康社区用户中关于新冠肺炎的提问记录,利用 LDA 模型对提问主题进行建模生成“主题-词”分布概率矩阵和分析主题一致性。结果显示网络健康社区用户对新冠肺炎的临床症状、检查诊断、用药治疗、疫区防控等方面关注较多。对网络健康社区用户关注的新冠肺炎热点主题进行分析发现,用户对新冠肺炎的信息需求与国家相继颁布的诊疗方案及防控指南内容相一致,说明对用户关于新冠肺炎提问的主题分析能够准确反映网络健康社区用户对新冠肺炎的信息需求。

[关键词]新冠肺炎;LDA 模型;网络健康社区;信息需求

[中图分类号]R193;TP391.1　　[文献标志码]A　　[文章编号]1671-3982(2021)02-0053-06

Need of online health community users for information on COVID-19 infection pneumonia

AN Xin-yu[1], CHEN Yu-xin[1,2], ZHANG Han[1]

(1. China Medical University Medical Informatics School, Shenyang 110122, Liaoning Province, China; 2. Southwest University Computer and Information Science School, Chongqing 400000, China)

Corresponding author: ZHANG Han

[**Abstract**]The questions about COVID-19 infection pneumonia put by online health community users were recorded. The question topic model was established according to the LDA model to generate the distribution probability of "topics-words" and to analyze the coherence of topics, which showed that more attentions were paid to the symptoms and signs, examination and diagnosis, medications, prevention and control of COVID-19 infection pneumonia by online health community users. Analysis of the hot topics on COVID-19 infection pneumonia showed that the need of online health community users for information on COVID-19 infection pneumonia was consistent with the contents in the treatment plan and prevention and control guidelines issued by our country, which indicates that topic analysis of questions about COVID-19 infection pneumonia put by online health community users can accurately show the need of online health community users for information on COVID-19 infection pneumonia.

[**Key words**]COVID-19 infection pneumonia; LDA model; Online health community; Information need

新型冠状病毒肺炎(COVID-19)(以下简称“新冠肺炎”)是由新型冠状病毒引起的乙类传染病,世界卫生组织将新冠肺炎疫情列为“国际关注的突发公共卫生事件(Public Health Emergency of International Concern, PHEIC)”[1]。疫情期间使用网络平台查找健康专业知识的人数急剧增加[2],网络健康社区有效弥补了线下医疗服务的空间限制,逐渐成为公众交流疾病信息的重要沟通渠道[3]。在网络健康社区,用户根据自身信息需求进行提问交流,实

[基金项目]辽宁省社会科学规划基金项目“基于机器学习算法的药物新适应症多库融合精准预测模型研究”(L19BTQ006)成果之一

[作者单位]1. 中国医科大学医学信息学院,辽宁 沈阳 110122;2. 西南大学计算机与信息科学学院,重庆 400000

[作者简介]安欣宇(1998-),女,硕士研究生,研究方向为医学文本挖掘与知识管理。

[通讯作者]张　晗(1974-),女,博士,教授,研究方向为医学文本挖掘与知识管理、数据科研与信息服务。E-mail:zhanghan@cmu.edu.cn

现信息共享。

信息需求是公众为解决自身的某种未知而产生的信息渴望[4],公众为消除自身对某种信息的“不确定感”进行相关信息搜索[5]。因此研究网络用户的信息需求可以了解用户在一定时间内的信息渴望,从而为其提供有效的资源帮助。有学者以“妈妈网”的“备孕”板块为例,通过主题和情感分析研究备孕女性的信息需求,总结出了 21 个主题及三大类情绪,其研究结果显示,在备孕阶段,女性缺少相应的健康信息支持与心理抚慰[6]。有学者通过数据编码等方法对 Yahoo! Answers 和 Diabetic Connect 论坛中与糖尿病相关的记录进行主题分析,将用户的糖尿病信息需求分为 8 个主题,发现 2 个论坛的“疾病管理”“治疗”“疾病预防”等主题较相近,“诊断检查”“社会生活”等主题存在差异[7]。有学者对“39 健康网”的高血压问答记录进行主题聚类的结果显示,治疗类问答最多(占 36.28%),但其信息需求呈下降趋势,说明常见慢性病高血压的治疗方式已经逐渐被大众所熟悉和掌握[8]。

综上所述,网络健康社区中的用户问答记录可以在一定程度上反映该阶段用户的信息,为信息工作者的资源组织工作提供有效帮助。本文利用隐含狄利克雷分布(latent dirichlet allocation, LDA)模型对“快速问医生”网络健康社区中有关新冠肺炎的问答记录进行主题分析,了解用户对新冠肺炎的信息需求的主题特征,旨在为网络健康社区和信息机构准确提供公众健康信息需求,为政府制定相关决策提供参考,进而为公众提供针对性的健康信息服务。

1 数据与方法

1.1 数据来源

网络健康社区“快速问医生”(www. 120ask. com)是一个求医问药的健康网站,其用户日均访问 1 600 万次以上(独立 IP),实名认证医生注册量超过 110 万人[9]。本文选取“快速问医生”的“肺炎”板块(板块路径:有问必答→内科→呼吸内科→肺炎)中用户关于新冠肺炎的提问作为数据来源,利用后羿采集器爬取 2020 年 4-6 月的用户提问记录(采集时间为 2020 年 7 月 1 日),通过人工筛选出新冠肺炎相关问题 2 676 条,去重后得到的 2 672 条提问记录作为数据集。

1.2 研究方法

1.2.1 数据预处理

在对新冠肺炎用户提问记录进行主题分析之前,首先对数据集进行分词、去停用词、同义词转换等预处理工作。

分词:应用 Python 的 jieba 包对数据集进行分词处理。为弥补 jiaba 包对专业词汇及新型词汇的分词处理缺陷,本文构建了含有 883 个词汇的自定义词典。

去停用词:对于分词过程中出现的量词、代词及“帮帮我”“谢谢”等与研究内容无关的动词、名词等词频过高且没有实际意义的词汇,本文将其加入停用表,以减少信息干扰。

同义词转换:由于用户的提问语言不规范,本文将“测量体温”“量了体温”“量体温”等词统一替换为“测体温”,以提高模型主题的准确性,最终得到含有 474 个词汇的同义词表。

1.2.2 LDA 主题分析

LDA 主题模型由 David M. Blei 等人于 2003 年提出[10],之后 Thomas L. Griffths 对模型进行了改进[11]。LDA 模型是基于“文档-主题-词”的 3 层贝叶斯概率生成的,用于推测文档的主题分布。本文使用 Python 的 gensim 库[12]和 mallet 包[13]对新冠肺炎相关问题集进行 LDA 建模和主题提取。在 LDA 模型中每个词在不同主题中出现的概率不同,利用概率分布对每篇文档的主题及每个主题中的词汇进行计算。LDA 主题模型的核心概率公式如下:

$$P(\text{词}|\text{文档}) = P(\text{词}|\text{主题}) \times P(\text{主题}|\text{文档}) \quad \text{公式(1)}$$

枚举所有主题,得到所有的 P(词|文档),根据所有概率结果选择与文档中的词对应的主题[14]。LDA 模型最终生成“文档-主题”概率分布矩阵、“主题-词”概率分布矩阵,用于进一步分析。

本文利用 Coherence Model 模块计算主题数取不同值时模型的主题一致性(coherence)[15]。其计算公式如下:

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log \frac{D(v_i, v_j) + 1}{D(v_j)} \quad \text{公式(2)}$$

公式(2)中,$D(v_j)$ 为词汇 v_j 的文档频率,$D(v_i,$

v_j)为词汇 v_i 和 v_j 的文档共现频率。

主题一致性由主题中的高频词主导,以实现语意连贯性,其值越大说明主题连贯性越好。本文测试了主题数为 1 到 15 时的主题一致性值,同时结合各主题中词的概率分布,确定最终主题数为 9。

2　结果与讨论

2.1　新冠肺炎词汇统计

通过统计核心词汇可以分析该领域的研究热点及发展趋势。很多学者利用词频分析法进行肿瘤治疗主题分析[16]、医院管理制度热点趋势分析[17]、肺康复研究热点聚类分析[18]等。为了解用户提问的热点词,本文利用 Python 中的 wordcloud 库绘制高频词词云,用以反映现阶段新冠肺炎用户的关注热点。本文对 2 672 条新冠肺炎用户提问进行分词、去除停用词、规范同义词等预处理后,最终得到 5 540个词汇,其中“新冠肺炎”“发烧”“咳嗽”“吃饭”“新冠肺炎病毒”“病毒感染”“喉咙”“体温”等词汇词频较高。将前 70 个高频词绘制成词云(图 1)。从词云中大致可以看出,用户信息需求主要集中在新冠肺炎的临床症状、传播途径、检查诊断 3 方面。其中,“发烧”“测体温”“疼痛”等词汇反映出用户对自身症状是否与新冠肺炎症状相吻合存在疑惑,对新冠肺炎具体临床症状掌握信息不全面;“病毒感染”“戴口罩”“接触”等词汇反映出用户对新冠肺炎传播途径的信息需求。深入了解疾病传播途径可有效保障自身安全,遏制疾病传播。从“核酸检测”“医院”“CT”等词汇可以看出新冠肺炎检查诊断同样是用户关心的内容。用户有效掌握新冠肺炎初步诊断方法,遵循“早发现、早报告、早隔离、早诊断、早治疗”原则,可避免疫情大范围传播。



图 1　新冠肺炎高频词词云

2.2　新冠肺炎主题分布统计

词频分析法虽然能在一定程度上反映当前阶段新冠肺炎用户的信息需求,但其仅以词频的形式展现,故不能对用户的信息需求进行有效分类,无法看出各词汇之间的内部联系。因此本文利用 LDA 模型对“快速问医生”网络平台中 2 672 条新冠肺炎相关问题进行主题提取,结合主题一致性值及“主题–词”概率分布矩阵中的具体内容将最优主题数定为 9。设定每个主题下的关键词提取数为 20,主题相关词汇按词分布概率递减顺序排列(表 1)。由表 1 可知,主题 1、主题 2 是新冠肺炎症状主题,主题 4、主题 8 是新冠肺炎检查诊断主题,主题 3 是新冠肺炎疫区主题,主题 5 是用药相关主题,主题 6、主题 7、主题 9 是新冠肺炎传染途径及防护措施相关主题。

表 1 新冠肺炎用户提问热点主题分布

主题编号	提问数量/条	比例/%	主题相关词汇
1	506	0.19	咳嗽、喉咙、痰、不适、流鼻涕、痒、喘气、感冒、鼻塞、打喷嚏、鼻涕、鼻、吹气、嗓子痒、肺炎、空调、喉咙痒、有点痛、厉害、上火
2	401	0.15	发烧、新冠肺炎、乏力、胸部、拉肚子、喘不上气、身体、干咳、腹泻、全身酸痛、烧、睡、接触史、病例、全身、浑身、闷热、着凉、慌、深呼吸
3	319	0.12	新冠肺炎、核酸检测、接触、新冠肺炎病毒、武汉、害怕、隔离、胸闷、病患、阴性、火车、疫区、消毒、恶心、影响、北京市、阳性、湖北省、人员、病毒携带者
4	281	0.10	医院、肺炎、检查、CT、肺部、治疗、血常规、症状、原因、排除、抽血化验、难受、住院、查询、好转、宝宝、缓解、打吊瓶、白细胞、门诊
5	227	0.08	吃饭、疼痛、头疼、头晕、药、感冒、喉咙疼、吃药、吐、退烧、感冒药、退烧药、躺、发炎、不疼、咽喉部、莲花清瘟、消炎药、阿莫西林、医院
6	337	0.13	病毒感染、新冠肺炎病毒、戴口罩、传染、口罩、冠状病毒、病毒、快递、讲话、戴、超市、孩子、酒精、喷、洗澡、距离、电梯、就医、骑车、消毒
7	159	0.06	新冠肺炎、担心、喝、买、害怕、疫情、新冠肺炎病毒、餐具、肚子、朋友、吃饭、坐、导致、宾馆、小区、餐、网上、家人、上班、经期
8	230	0.09	体温、症状、新冠肺炎、测体温、睡觉、测、头、出门、温度、出汗、学校、起床、家、热、冷、晕、发冷、测温仪、无力、干咳
9	212	0.08	病毒感染、碰、回家、手、眼睛、洗手、洗、手机、衣服、脸部、身边、病、口沫、不适、睡觉、擦、爸妈、摸、厕所、水

2.3 新冠肺炎信息需求分析

结合表 1 中用户关于新冠肺炎的提问热点主题,可以将新冠肺炎信息需求归纳为以下 5 类。

一是症状信息需求。新冠肺炎与普通感冒极为相似,患者会出现发烧、干咳、乏力等症状,但二者临床症状存在差异。新冠肺炎相较于普通肺炎会出现腹泻、呕吐、呼吸困难等症状,如主题 1 中“流鼻涕”“嗓子痒”等词所描述的症状为普通感冒症状,主题 2 中“拉肚子”“喘不上气”等词说明此主题侧重描述新冠肺炎相关症状。用户无法有效区分新冠肺炎与普通感冒,因此国家发布了新冠肺炎的临床特点相关文件[19],详细介绍了新冠肺炎临床表现、观察病例临床表现,其中患者以发热、乏力、干咳为主要表现,伴有呼吸道疾病,用户可根据自身症状进行初步鉴别。

二是诊断信息需求。由于发热为新冠肺炎最主要的临床症状,因此体温测量是最方便快捷的初步诊断和筛查手段。主题 8 中汇集了与体温测量密切相关的词汇。若患者对自身症状无法自行诊断,可以去医院做进一步的实验室检查,包括鼻咽拭子、CT 影像、血液检测等加以证实,参见主题 4 中的诊断主题词汇。用户及医护人员可根据《新型冠状病毒肺炎诊疗方案(试行第七版)》中的“病理改变”一节[20]对比患者胸部影像、抗体检测结果等进行诊断,以保证检测结果的准确性。

三是疫区信息需求。新冠肺炎期间的另一个关注热点是与来自疫区及感染人员的接触情况。由于新冠肺炎的传染性强,群集性疫情爆发使疫情防控形势严峻,落实常态化防控要求、提高病毒检测能力、警惕外来输入病例等工作刻不容缓。如主题 3 中的“武汉”“北京”“湖北省”“隔离”“火车”等词以及情感类词汇“害怕”,表明公众对外来人员,尤其是对来自疫区人员进行病毒检测与隔离工作的重视程度。疫情的及时通报可保障群众的知情权[21],有效管理可抑制疾病的扩散[22],使正常生产生活秩序得以稳定。

四是治疗用药信息需求。用户出现发热、干咳等症状时,优先选择服用退烧药、感冒药。如主题 5 中的“莲花清瘟”“阿莫西林”等词汇表示的药物以求缓解症状。若症状未好转,患者应及时前往医院进行诊治。为保证患者用药正确,国家颁布的《新型冠状病毒肺炎诊疗方案(试行第七版)》“治疗”一

节[20]中明确给出中西医用药及仪器治疗的方法,患者可在患病初期服用莲花清瘟等药物进行初步治疗。

五是传染防护信息需求。准确了解新冠肺炎的传播途径及防护措施有助于遏制疾病的传播。用户对新冠肺炎传播途径及防护措施相关信息存在疑惑,如主题 6 中“戴口罩”“酒精”,主题 7 中“喝”“吃饭”,主题9 中“洗手”“口沫”等词反映了用户对防护的信息需求。针对用户关注的防护问题,国家颁布了相关防控指南,引导群众采取正确的个人防护措施[23-26],同时规范公共场所的疾病防控工作[27-28],以求在有效控制疾病传播的同时,维护社会稳定。

3　结论

本文收集网络健康社区用户的新冠肺炎提问记录,采用 LDA 模型提取用户关注的热点问题,并对各主题进行深入分析。结果显示,疫情期间网络用户对新冠肺炎的临床表现、诊断、治疗和预防等方面较为关注。针对以上热点问题,国家相继颁布了相关诊疗方案及防控指南,给予了正确的解释和引导,有效遏制了疫情的发展,使民众能够精准防疫,有序复产、复工。

突发公共卫生事件具有明显的时间节点特征,采用时序分析方法研究网络用户对公共卫生事件关注热点的时间变化趋势,可为信息服务者提供针对性的借鉴参考。

4　结语

本文针对健康社区用户提问进行主题分析,有效揭示大众的医疗信息需求,发现用户信息需求特征,为健康信息服务机构开展高质量的服务、为政府制定更加符合群众健康需求的决策提供参考。由于“快速问医生”网站仅保留 3 000 条用户提问,本文收集的数据也仅覆盖了“快速问医生”网站 2020 年 4-6 月的帖子,因此本文的研究存在不足之处。今后将通过增加网站收集数量、延长数据收集时间、扩充研究样本量等方式对用户的健康提问进行深入分析,进一步了解不同人群、不同时段健康信息需求的主题特征变化,从而为用户提供精准的信息服务。

【参考文献】

[1] WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV)[EB/OL]. (2020-01-30)[2020-06-31]. https://www.who.int/zh/director-general/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov).

[2] 中国互联网络信息中心. CNNIC 发布第 45 次《中国互联网络发展状况统计报告》[EB/OL]. (2020-04-28)[2020-07-11]. http://www.cnnic.cn/gywm/xwzx/rdxw/20172017_7057/202004/t20200427_70973.htm.

[3] 郝　军,戚淼杰, 韩优莉. 在线医疗社区的发展现状及启示:以北京市为例[J]. 卫生经济研究,2020,37(11):18-22.

[4] Case D O. Looking for information: A Survey of Research on Information Seeking, Needs and Behavior[M]. England: Emerald Group Publishing,2012:97.

[5] Belkin N J, Oddy R N, Brooks H M. ASK for information retrieval: Part I. Background and theory[J]. Journal of Documentation, 1982,38(2):61-71.

[6] 刘　冰,历　鑫,张赫钊,等. 网络健康社区中身份转换期女性信息需求主题特征及情感因素研究:以“妈妈网”中“备孕板块”为例[J]. 情报理论与实践,2019,42(5):87-92.

[7] 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.

[8] 唐晓波,李　津. 在线健康社区信息需求主题分析[J]. 数字图书馆论坛,2019(2):12-17.

[9] “快速问医生”的优势[EB/OL]. [2020-07-29]. https://www.120ask.com/company/marketing.html.

[10] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003(3):993-1022.

[11] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences USA,2004,101(1):5228-5235.

[12] Gensim[EB/OL]. [2020-07-17]. https://radimrehurek.com/gensim/.

[13] Mallet Homepage[EB/OL]. [2020-05-24]. http://mallet.cs.umass.edu/.

[14] 杨　磊,王子润,侯贵生. 基于 Q-LDA 主题模型的网络健康社区主题挖掘研究[J]. 数据分析与知识发现,2019,3(11):52-59.

[15] Roberts M, Stewart B, Tingley D. stm: R Package for Structural Topic Models[J]. Working Paper,2014,57(1):445-460.

[16] 李　佳. 1998-2007 年我国肿瘤治疗研究的主题分析[J]. 医学信息,2009,22(1):57-61.

[17] 张　年,余昌胤,袁成菊,等. 我国现代医院管理制度的研究热点及趋势分析[J]. 中国医院管理,2019,39(2):62-64.

[18] 彭司淼,李乐之,欧尽南. 近 10 年肺康复研究热点共词聚类分析[J]. 护理学杂志,2020,35(11):91-94.

[19] 国家卫生健康委办公厅,国家中医药管理局办公室. 关于印发新型冠状病毒感染的肺炎诊疗方案(试行第五版)的通知

[EB/OL]. (2020-02-04)[2020-06-31]. http://www.gov.cn/zhengce/zhengceku/2020-02/05/content_5474791.htm.

[20] 国家卫生健康委办公厅,国家中医药管理局办公室. 关于印发新型冠状病毒肺炎诊疗方案(试行第七版)的通知[EB/OL]. (2020-03-03)[2020-07-23]. http://www.gov.cn/zhengce/zhengceku/2020-03/04/content_5486705.htm.

[21] 医政医管局. 关于黑龙江省新冠肺炎聚集性疫情有关情况的通报[EB/OL]. (2020-05-01)[2020-07-23]. http://www.nhc.gov.cn/yzygj/s7659/202005/9115d2334dfb4b2badcd9a5482f4a9b8.shtml.

[22] 疾病预防控制局. 关于做好离京人员新冠肺炎健康管理服务工作的通知[EB/OL]. (2020-06-19)[2020-07-23]. http://www.nhc.gov.cn/jkj/s7916/202006/7166534cf9b8434ea7213b93acd8fe48.shtml.

[23] 疾病预防控制局. 关于印发公众科学戴口罩指引的通知[EB/OL]. (2020-03-18)[2020-07-25]. http://www.nhc.gov.cn/jkj/s3577/202003/0a472cc09e744144883db6a74fe6e76 0.shtml.

[24] 疾病预防控制局. 关于印发外卖配送和快递从业人员新冠肺炎疫情健康防护指南的通知[EB/OL]. (2020-07-10)[2020-07-25]. http://www.nhc.gov.cn/jkj/s7934td/20200 7/c95f49f34989404398c13c7a0897edc9.shtml.

[25] 疾病预防控制局. 一图读懂:新型冠状病毒肺炎防控方案(第五版)[EB/OL]. (2020-02-22)[2020-07-23]. http://www.nhc.gov.cn/jkj/s3578/202002/dc2f3b1200144ae1a50ffa4e0df62da8.shtml.

[26] 综合监督局. 国家卫生健康委办公厅关于印发消毒剂使用指南的通知[EB/OL]. (2020-02-19)[2020-07-25]. http://www.nhc.gov.cn/zhjcj/s9141/202002/b9891e8c86d141a08ec45c6a18e21dc2.shtml.

[27] 疾病预防控制局. 关于印发新型冠状病毒肺炎流行期间商场和超市卫生防护指南的通知[EB/OL]. (2020-02-15)[2020-07-25]. http://www.nhc.gov.cn/jkj/s3577/202002/6a13deef74604f39a16390679d98283c.shtml.

[28] 疾病预防控制局. 关于严格预防通过交通工具传播新型冠状病毒感染的肺炎的通知[EB/OL]. (2020-01-24)[2020-07-27]. http://www.nhc.gov.cn/jkj/s3577/202001/e5e8c983baba4c1589512e6c99fdaa4e.shtml.

[收稿日期:2020-11-11]

[本文编辑:孙伟娟]