基于文本挖掘的在线健康社区用药咨询研究 *

杨　平　肖遗规
（湖南工业大学商学院，株洲　412007）

摘　要：[**目的 / 意义**] 在线健康社区是人们获取健康信息的重要途径，研究在线健康社区用户用药咨询需求，有助于在线健康社区药品服务优化及可持续发展。[**方法 / 过程**] 以 39 健康网为例，首先，利用 Python 编码爬取肠胃用药类药品的 59 048 条用药咨询评论，并进行预处理；其次，使用 TF-IDF、TextRank、LDA 主题模型等文本挖掘方法对实验数据进行主题关键词挖掘，并进行关键词共现网络分析；最后，综合分析在线健康社区用户用药咨询需求的主题特征，并提出优化建议。[**结果 / 结论**] 研究结果发现在线健康社区用户主要关注药品治疗效果、服用方式、不良反应、药品区别以及孕妇等特殊人群在用药时的注意事项等。本研究一方面为药品厂商调整优化药品说明书内容提供了理论依据，另一方面为在线健康社区优化药品说明书内容布局以及建立或完善药品科普服务指引了方向。

关键词：在线健康社区　用药咨询　文本挖掘
分类号：G203　C93-03　R194
DOI：10.31193/SSAP.J.ISSN.2096-6695.2024.01.07

0　引　言

随着互联网信息技术的不断发展，人们获取健康信息和知识的方式不再局限于传统的医患面对面交流，互联网成为人们获取健康信息的便捷途径[1-2]。据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》显示[3]，截至 2023 年 6 月，我国互联网医疗用户规模达 3.64 亿人，占总体网民规模的 33.8%。“互联网 + 医疗”模式的不断发展，吸引了大量用户通过在线健康社区（Online Health Communities，简称 OHCs）获取医疗健康相关服务，常见的在线健康社区有“39 健康网”“寻医问药网”“好大夫在线”等，这些健康社区可为用户提供药品服务、预约挂号、健康科普、疾病治疗经验分享等服务与信息，极大地方便了健康社区用户[4]。

用药咨询是在线健康社区中重要的药品服务，当用户对药品使用存在疑惑，且通过查阅药品

* 本文系湖南省自然科学基金青年基金资助项目“服务型领导对组织绩效影响机制研究——组织公民行为与团队断裂带视角”（项目编号：2021JJ40183）的研究成果之一。
[作者简介] 杨平（ORCID：0009-0004-8440-843X），男，硕士生，研究方向为健康大数据文本分析、用户健康信息行为、健康知识管理等，Email：3109605327@qq.com；肖遗规（ORCID：xiaoyigui@hufe.edu.cn），男，讲师，硕士生导师，博士，研究方向为人力资源管理、战略管理等，Email：754828090@qq.com。

说明书未得到解决时，便会向药师咨询，这些咨询内容能够直接反映用户的实际用药需求，挖掘这些需求有助于发现用户在用药过程中高频关注的内容，进而为在线健康社区优化药品相关服务提供可行的参考依据，有助于在线健康社区的可持续发展。本文以 39 健康网中肠胃用药类药品的用户用药咨询为例，通过词频统计分析、文本主题特征分析、关键词共现网络分析等，发现用户用药咨询需求主题特征，进而为在线健康社区药品服务优化提供参考建议。

1　相关研究

1.1　健康信息与服务研究

在线健康社区是以互联网信息技术为依托，以医疗健康为主题的社交平台，平台中入驻了大量的一般公众、患者及其看护人员、医生以及医疗健康服务机构等各类用户，他们以在线互动替代传统的面对面交流，通过提供健康服务、共享医疗健康信息、传播疾病治疗经验等方式优化医疗资源的时空局限［5–6］。

在线健康社区可为用户提供各类健康信息与服务，且信息与服务密不可分，两者相辅相成。从信息的发布者视角来看，在线健康社区主要提供两类信息：一类是平台发布的信息，包括医院、医生、药品以及疾病等信息，可为用户提供医院与医生推荐、电子预约挂号、药品购买和疾病查询等服务，国内外学者基于在线健康社区所提供的信息，对在线健康社区的医院与医生推荐［7–9］、疾病诊断［10］等服务进行了大量研究。另外一类是医生和患者等用户发布的信息，一些学者基于医生发布的健康科普文章，研究如何为社区用户提供个性化的健康科普推荐服务［11］，以及各类在线问诊和咨询服务；还有一些学者根据患者发布的评论信息，如患者向医生咨询疾病和药品、患者之间的经验分享以及情感交流等，分析用户的健康信息需求［12–13］、健康信息行为［14–15］、情感特征［16］等。

1.2　健康信息挖掘技术研究

在线健康社区中的各种医疗健康信息，是由社交行为产生的交互信息，大多数以文本形式呈现，具有数据量大、结构复杂等特点。如何挖掘此类信息的主题内容、探索用户的健康信息需求和行为，一直是学术界关注的热点。

此前，对于信息主题内容的挖掘主要采用内容分析、人工统计标注的方法，如金碧漪等为了解消费者对于糖尿病信息的需求，采集雅虎问答社区中糖尿病相关的提问记录，通过人工编码、文本处理、多维尺度分析、中心词聚类等方法，发现热点主题是日常疾病管理、疾病确诊和治疗［17］；施亦龙等采用内容分析法对中美两个最大的在线社区百度知道和雅虎问答上采集的自闭症问答数据进行分析，发现美国用户对于疾病的基础知识掌握比中国用户好，提问内容更加详细、多样，对疾病的探索性提问更加积极［18］。但传统的内容分析法和人工统计标注需要耗费大量的人力和时间成本，随着文本挖掘技术的不断成熟，越来越多的学者将文本聚类算法、LDA 主题模型等自动识别技术应用到在线健康社区的信息挖掘研究中，如唐晓波等对在线健康社区高血压问答文本进行聚类分析，发现用户最关心疾病的治疗、并发症和生活方式等［19］；张丽等基于 LDA 主题模型、情感分析等，对医药电商在线评论进行文本分析，挖掘出疫情背景下消费者对网购医

药商品的需求重点与痛点[20]。

1.3 用药咨询研究

用药咨询是指药师应用药学专业知识和临床技能，对患者、患者家属等咨询者提供药物治疗和合理用药的药学服务[21]。药品作为特殊的商品，其安全性直接影响着使用者的身体健康，而保证安全的原则之一就是合理规范使用，但如今大多数患者医学素养不高，对于药品的使用经常存有疑问，因此用药咨询就显得非常重要。

当前，用药咨询的方式主要分为线下途径和线上途径。传统的线下用药咨询通常是咨询者去药店或医院等，药师对其进行面对面的指导，这种形式的指导能够使医生获得足够的信任，但通常也需要花费咨询者大量的时间。随着“互联网 + 医疗”的快速发展，用药咨询服务逐渐扩展到在线平台，咨询者可以直接通过互联网向药师咨询用药相关问题，为咨询者提供了极大便利，特别是受到新冠肺炎疫情的影响，人们更青睐于足不出户的就医方式，因此在线用药咨询就有了巨大优势，大量用户纷纷尝试。

在线健康社区作为“互联网 + 医疗”时代的产物，同样具有在线用药咨询功能，咨询者遇到任何药品相关问题都可以直接通过社区向药师提问。目前，国内外有关用药咨询服务的研究多聚焦于线下门诊[22-23]，但仍有一些学者对线上用药咨询服务进行了探索，具体涉及到用药咨询服务质量管理与评价、用药咨询服务模式等方面，如梅昕等分析了儿科药师通过“问药师”平台提供儿童用药咨询服务的实践效果，为后续出台药师参与临床合理用药相关规范及行业法规提供了参考[24]。

综上所述，国内外有关在线健康社区健康信息的研究较为成熟，研究成果丰硕，但聚焦于药品信息和药品服务的研究较少，而用药咨询作为在线健康社区重要的药品服务，对在线健康社区的发展有着重要影响。用户作为用药咨询的主体，其药品咨询内容能够直接反映用药需求，对于在线健康社区服务的提升有着重要价值，加之利用文本挖掘技术在医疗健康领域有较为成熟的研究基础，因此，本研究以在线健康社区用户用药咨询评论为研究对象，利用 TF-IDF、TextRank、LDA 主题模型等多种文本挖掘方法，并进行关键词共现网络分析，以发现用户高频关注的用药需求信息。

2 研究方案

本文研究方案主要包括数据采集及预处理、文本挖掘、实验结果分析、结果讨论等，具体研究框架如图 1 所示。

首先，从在线健康社区采集用药咨询模块用户评论等数据，并对采集的数据进行预处理；其次，利用 TF-IDF、TextRank、LDA 主题模型等方法，挖掘用药咨询评论主题关键词；再次，对用药咨询评论进行主题特征分析，主要包括词频统计分析、文本主题分析、关键词共现网络分析等；最后，在上述分析的基础上，对结果进行综合分析与讨论。下面对数据采集及预处理、文本挖掘部分进行具体论述。

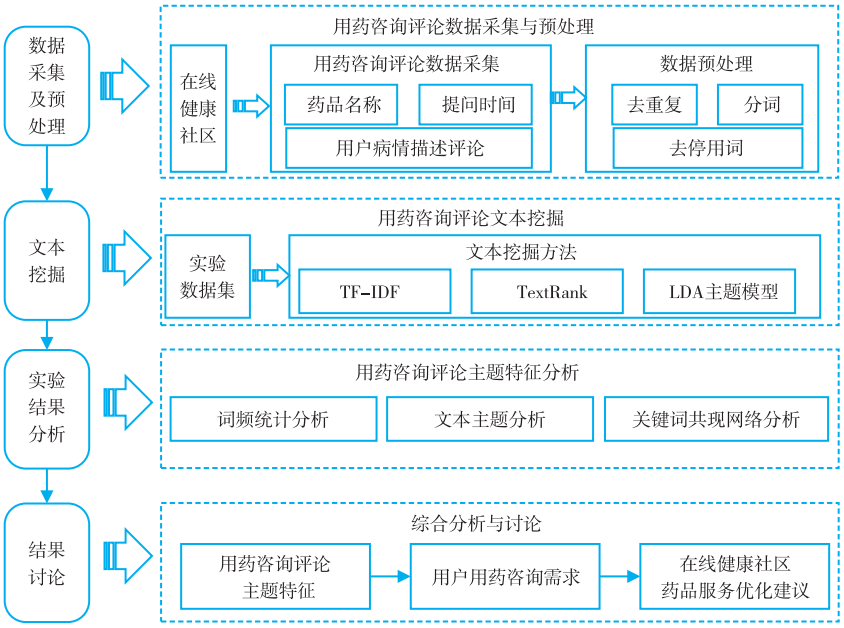

图 1　总体研究框架图

2.1　数据采集及预处理

首先，选取具有代表性药品数据库的在线健康社区；其次，使用 Python 编写网络爬虫程序，采集用户用药咨询相关的信息；再次，对获取到的文本数据进行清洗、去重复、分词和去停用词等操作；最后，得到实验数据集。

2.2　文本挖掘

本研究采用 TF-IDF、TextRank、LDA 主题模型等当前主流的关键词挖掘方法，对实验数据集进行文本挖掘，下面分别对这三种方法进行简要介绍。

2.2.1　TF-IDF 算法

TF-IDF 算法是一种基于词频统计的加权技术，优点是简单快速，计算效率高，可用于表示特征项在整个语料库中的重要性，其基本思想是特征项的重要程度与在文档中出现的频率成正比，与在语料库中出现的频率成反比[25]。TF-IDF 算法由两部分组成，即 TF 算法与 IDF 算法，其中 TF 算法的基本思想是一个特征词在一个文档中出现的次越数多，则这个特征词越能表达这个文档，而 IDF 算法的基本思想是一个特征词在越少的文档中出现，则对文档的区分能力越强。TF-IDF 算法计算公式[26]如（1）所示。

$$\begin{aligned} TF-IDF &= \text{词频}（TF）\times \text{逆文档频率}（IDF） \\ &= \frac{\text{某个特征词在文档中出现的次数}}{\text{文档的总词数}} \times log \frac{\text{语料库的文档总数}}{\text{包含该特征词的文档数}+1} \end{aligned} \tag{1}$$

2.2.2　TextRank 算法

TextRank 算法是一种基于图模型的排序算法，考虑了词频与词语间的关系，其基本思想来

源于谷歌的 PageRank 算法，能够脱离语料库的基础，将文本转化为图结构，并使用迭代计算的方式计算每个节点的权重值，权重值越大，则表示单词或短语越重要[27]。其计算公式如（2）所示。

$$T_{TextRank}(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}T_{TextRank}(V_j) \quad (2)$$

其中，V_i、V_j 为词语节点，d 为阻尼系数，一般取值为 0.85，In（V_i）为指向词语节点 V_i 的词语节点集合；Out（V_j）为词语节点 V_j 指向的词语节点集合；w_{ji}、w_{jk} 为词语节点 V_j 到 V_i、V_j 到 V_k 的边权重[28]。

2.2.3 LDA 主题模型

LDA 模型是一种基于语义的方法，考虑了上下文语义关系。在利用 LDA 模型进行主题挖掘前，需确定提取的最优主题数目，当前研究中常借助困惑度或一致性得分等方法来确定最优主题数目[29]。本研究使用困惑度计算用药咨询文本的困惑度值，并绘制困惑度曲线图，从而确定所需提取的最优主题数。困惑度由 Blei D. M. 等[30]于 2003 年提出，主要用于评估语言模型的优劣程度，虽然较小的困惑度得分意味着模型对文本有较好的预测作用，但是同时也要考虑困惑度曲线是否出现拐点，以此来综合评判合适的主题数量，具体的计算公式（3）如下：

$$Perplexity(D)=\exp\left(-\frac{\sum_{i=1}^{M}\log p(w_{di})}{\sum_{i=1}^{M}N_{di}}\right) \quad (3)$$

式中，D 是目标数据中的文本文档，D_i 表示第 i 个文本文档，N_{di} 表示文档所有词项的总数，p（w_{di}）表示文档集中各文档的产生概率。

LDA 是由 Blei D. M. 等针对早期的 Latent Semantic Analysis(LSA) 和 Probabilistic Latent Semantic Analysis(PLSA) 两种主题挖掘模型的缺陷所提出的一种无监督机器学习技术[30]，该模型是基于“文档 – 主题 – 词项”三层贝叶斯概率模型，利用概率统计的思想对文档进行建模，可自动识别发现大规模文档集或语料库中隐藏的主题信息，对于处理海量文本信息十分有效，且可以提高文本分类的精度[31]。LDA 主题模型核心表达式如公式（4）所示。

$$P(w|d)=P(w|t)\times P(t|d) \quad (4)$$

其中，w 代表文档中的一个词项，d 代表文档集中的一篇文档，t 代表一个主题。本研究使用 LDA 主题模型对用户用药咨询文本进行主题挖掘时，首先，需确定 3 个参数，即 α、β、K，其中 α =50/K、β取 0.01，K 为最优主题提取数目，由困惑度曲线求解确定。

3　在线健康社区用药咨询分析

3.1　数据获取

39 健康网是国内领先且具有代表性的在线健康社区，拥有国内最大的各级医院、医生、药品及个人医疗资料数据库，因此本文选取 39 健康网作为本实验的数据来源。39 健康网药品通功能模块下具有众多类型的药品，例如镇热解痛类、肠胃用药类、呼吸系统类等药物。根据阿里健康研究院联合中康科技发布的《2022 线上用药趋势白皮书》显示，随着用户群体和用户健康需求的快速增长，OTC 市场诞生了包括肠胃健康等多个新趋势赛道，另外根据药融云发布的《药融云 2022 年度医药电商白皮书》，在线销售渠道中，消化系统类用药占比最大，因此本研究选取肠胃类药品作为研究对象，具有更好的代表性。

通过 Python 编写网络爬虫程序，采集 39 健康网药品通功能模块下，肠胃类药品中有关助消化、炎症、胃脘疼痛三类药品的用户用药咨询评论数据，采集时间为 2023 年 10 月 12 日，采集数据项包括药品名称、用户病情描述、提问时间等，最终获得 2996 个药品下的 59 048 个用户评论文本。

3.2　数据预处理

数据预处理主要包括数据清洗、去重复、分词以及去停用词等操作，由于爬取的数据较为杂乱，存在大量不相关的内容，例如网站链接等，会影响最终的计算结果，因此，进行数据清洗非常必要。另外 39 健康网许多同功能药品的用药咨询内容完全一样，例如，清开灵颗粒（白云山）与清开灵颗粒（远达）两个药品用药功效无明显差异，二者药品说明书存在区别，但用药咨询内容却完全相同，因此需进行去重复操作。通过预处理，最终获得 28 250 个用户评论文本数据，作为本实验的数据集。

然后对实验数据集进行分词和去停用词操作。本研究借助 Python 中的 jieba 工具进行分词操作，考虑到药品的特殊性，jieba 库对药品专业名词的划分存在不足，因此将采集的数据集中药品名称抽取出来作为 jieba 新增词库，最终经过处理得到包括丹七片等 2298 个药品名称，这样使得分词划分更加科学；去停用词操作，本研究综合使用了中文停用词表与哈工大、四川大学、百度等停用词表，以及自定义停用词。

3.3　文本主题特征分析

本研究对经过预处理得到的数据，借助 Python 编写程序得到用药咨询评论文本的词云图，以及 TF–IDF、TextRank、LDA 主题模型等挖掘的主题关键词，最后进行关键词共现网络分析等。具体结果如图 2、图 3、图 4 与表 1 所示。

3.3.1　词云图分析

词云图为基于词频统计分析的可视化结果。从词云图可以看到，高频词主要有“作用”“功效”“副作用”“治疗”“服用”“多久”“效果”“区别”等，因此在线健康社区用户用药咨询评论主要涉及药品的作用效果、服用方式、副作用等不良反应以及药品区别等方面。特别地，对于药品作用、效果、功效等有关药品有效性的问题描述最多。

图 2　用户用药咨询关键词词云图

3.3.2　文本主题关键词分析

本研究计算用药咨询评论文本排名前 10 的主题关键词及各自的 TF-IDF 值与 TextRank 值，利用 LDA 模型进行主题提取前，需计算困惑度以确定最优的主题个数，经过计算并绘制困惑度曲线图，如图 3 所示，发现主题数为 5 时，困惑度最低，并且为拐点，因此最优主题数选择 5 较为合适。然后利用 gensim 工具计算评论文本的 LDA 分布，并设置参数 α =50/K、β =0.01，最优主题数 K 为 5，迭代次数 passes=100，并提取每个主题下的前 10 个关键词输出结果。最终得到用药咨询评论文本主题关键词结果如表 1 所示。

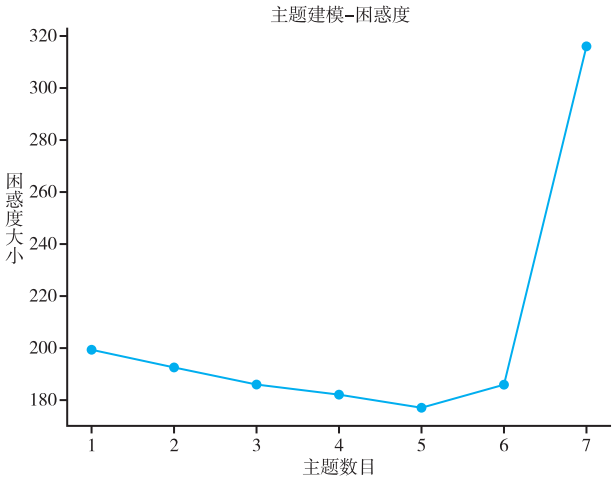

图 3　主题困惑度曲线

表 1　用药咨询主题关键词提取结果

方法	评论主题
TF-IDF	“功效”：0.29，“副作用”：0.20，“服用”：0.17，“作用”：0.17，“治疗”：0.11，“多久”：0.10，“效果”：0.07，“区别”：0.07，“拉肚子”：0.05，“健胃消食片”：0.04
TextRank	“功效”：1.0，“作用”：0.99，“服用”：0.59，“副作用”：0.54，“治疗”：0.53，“效果”：0.33，“区别”：0.25，“拉肚子”：0.14，“孕妇”：0.10，“胃炎”：0.10

续表

方法	评论主题
LDA	Topic1：0.047*“功效” + 0.027*“服用” + 0.020*“治疗” + 0.016*“作用” + 0.013*“逍遥丸” + 0.011*“效果” + 0.011*“区别” + 0.008*“有用吗” + 0.008*“阿莫西林” + 0.008*“多久”
	Topic2：0.046*“作用” + 0.040*“功效” + 0.031*“副作用” + 0.026*“多久” + 0.019*“治疗” + 0.018*“健胃消食片” + 0.015*“服用” + 0.014*“效果” + 0.013*“区别” + 0.009*“注射用”
	Topic3：0.035*“作用” + 0.026*“服用” + 0.025*“治疗” + 0.025*“功效” + 0.019*“副作用” + 0.016*“区别” + 0.012*“藿香正气水” + 0.009*“肠溶片” + 0.009*“哺乳期” + 0.009*“氧氟沙星”
	Topic4：0.035*“作用” + 0.022*“副作用” + 0.017*“功效” + 0.016*“服用” + 0.015*“区别” + 0.014*“多久” + 0.014*“香砂养胃丸” + 0.012*“胃痛” + 0.012*“治疗” + 0.011*“维生素”
	Topic5：0.057*“功效” + 0.054*“作用” + 0.038*“副作用” + 0.029*“服用” + 0.023*“效果” + 0.015*“治疗” + 0.014*“多少钱” + 0.013*“胃炎” + 0.011*“多久” + 0.011*“注射液”

由表 1 可见，首先，TF–IDF 和 TextRank 提取的关键词，以及 LDA 提取的各个主题的关键词，存在非常高的相似性，均包含“功效”“治疗”“作用”“效果”“副作用”“区别”等词，因此药品的作用效果、是否有副作用以及药品之间的区别等是咨询者高频关注的内容。其次，LDA 主题模型提取的五个主题之间存在较高的相似性，通过观察很难看出每个主题所代表的内容，但是观察每个主题之间的差异，发现药品的不同是主题呈现差异的关键因素，从侧面也反映出，无论是何种药品，用户用药咨询的内容通常涉及到药品的作用功效、服用方式、药品区别以及药品价格等方面。最后，综合表 1 结果，可将用户用药咨询内容归纳为以下几个方面：药品治疗效果方面，“功效”“作用”“治疗”“多久”“效果”等词高频出现，反映了用户对药品使用效果的关注；药品种类方面，“健胃消食片”“逍遥丸”“阿莫西林”“藿香正气水”“氧氟沙星”“香砂养胃丸”等药品是患有肠胃疾病用户常咨询的药物品种；疾病类型方面，“拉肚子”“胃炎”“胃痛”等词反映了咨询者常伴有拉肚子、胃疼等疾病症状；而“服用”“副作用”“区别”等词高频出现，则分别反映了用户对于药品服用方式、药物不良反应、药品区别等内容的关注。

3.3.3 关键词共现网络分析

通过词云图分析和文本主题关键词分析，可以发现用户用药咨询的关键内容，但无法发现不同关键词之间的关联强度，本研究使用关键词共现网络以进一步发现主题关键词之间的内在联系。关键词共现网络分析是一种计算文本中多个关键词同时出现的频次，以此判断它们之间的相似性关系并进行分析的研究方法[32]，其特点是以高频关键词为节点，以节点两两之间的共现关系为基础，将词与词之间的关系数值化处理，再通过图形化的方式揭示词与词之间的结构关系。使用 Python 将用药咨询文本进行分词，抓取关键词，并生成共词矩阵；再筛选出现频次前 25 的关键词进行共现分析，并利用 Ucinet 软件的 NetDraw 可视化功能绘制关键词共现网络，结果如图 4 所示。图中节点表示高频关键词，节点越大表示出现的频次越高，节点与节点之间的连线表示两个关键词的共现关系，连线越粗，表示关键词两两共现的次数越多，联系越密切。

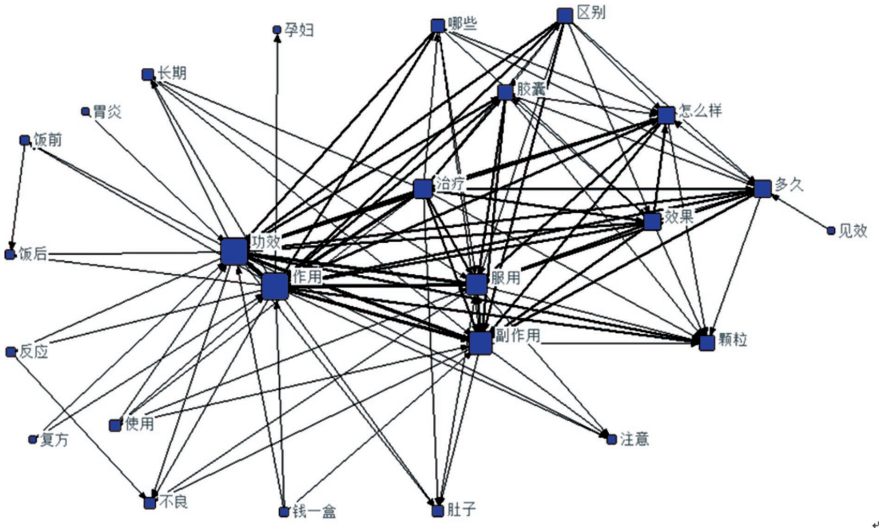

图 4　用药咨询关键词共现网络

图 4 中，“功效”“作用”“副作用”“服用”“治疗”“效果”等词是整个网络的重要节点，是咨询者在用药咨询过程中核心咨询的内容。根据关键词共现网络，咨询内容主要体现在以下几个方面：药品有效性方面，“服用”与“功效”“作用”“效果”等词联系密切，说明咨询者非常关注药品服用后的效果；药品安全性方面，“不良”与“反应”、“服用”与“副作用”等词组联系密切，反映了咨询者较为关注药品服用后是否存在不良反应以及副作用等安全性问题；服用方式方面，“饭前”“饭后”“功效”“作用”等词联系紧密，可见咨询者对于药品服用方式较为关注；适用人群方面，儿童和孕妇属于药品使用的特殊人群，在药物使用方面存在诸多注意事项，图 4 中“孕妇”一词的出现，反映了孕妇这类人群在使用药物时经常会咨询药师，以便安全用药。

3.4　实验结果讨论与建议

综合以上实验分析结果，发现用户用药咨询内容主要涉及药品治疗效果、服用方式、不良反应、药品区别以及孕妇等特殊人群在用药时的注意事项等内容。

本文结合实验结果与在线健康社区实际运营方式，对在线健康社区优化药品服务提出一些建议。首先，用户在购买或者使用药品前，通常会优先查阅药品说明书，而用药咨询只是平台为用户提供药品相关咨询的补充服务，但本文实验发现，虽然药品治疗效果、服用方式、药物不良反应等核心内容，药品说明书均会对此作出说明，但用户仍然会去问药师，侧面表明药品说明书仍有优化空间，因此在线健康社区及医药企业可对药品说明书中药物作用效果、服用方式、药物不良反应等用户高度关心的内容进行补充，尽可能对用户实际用药中的问题进行解答，进而缓解药师解答咨询的压力；其次，在线健康社区可对药品说明书布局进行优化，药品说明书内容多，且较为专业化，可对有关药物作用效果、服用方式、注意事项等内容进行突出显示，使得用户一目了然；最后，在线健康社区可使用本实验的研究方法挖掘用户关注的问题，并据此建立或完善药品科普宣传服务。

4 结束语

本研究采集 39 健康网药品通功能模块下肠胃类药品中有关助消化、炎症、胃脘疼痛等三种药品的用户用药咨询评论数据，通过词频统计分析、文本主题特征分析、关键词共现网络分析，以发现用户在药物使用过程中最为关心的问题。研究结果表明，在线健康社区用户对于药品作用效果、服用方式、不良反应、药品区别以及孕妇等特殊人群在用药时的注意事项等方面的内容较为关注。研究结果一方面为药品厂商调整优化药品说明书提供了理论依据，另一方面为在线健康社区优化药品说明书内容布局以及建立或完善药品科普内容指引了方向。

【参考文献】

［1］Cao B, Huang W, Chao N, et al. Patient activeness during online medical consultation in China: Multilevel analysis［J］. Journal of Medical Internet Research, 2022, 24(5): e35557.

［2］Fu Y, Tang T, Long J, et al. Factors associated with using the Internet for medical information based on the doctor-patient trust model: a cross-sectional study［J］. BMC Health Services Research, 2021, 21(1): 1268.

［3］第52次《中国互联网络发展状况统计报告》［EB/OL］.（2023-8-24）［2023-10-07］. https://www.cnnic.cn/n4/2023/0828/c88-10829.html.

［4］杨平，肖遗规，钟军. 基于SOR模型的在线健康社区用户药品购买意愿研究［J］. 绵阳师范学院学报，2024，43（1）：47-57.

［5］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.

［6］Qiao W, Yan Z, Wang X. Join or not: the impact of physicians' group joining behavior on their online demand and reputation in online health communities［J］. Information Processing & Management, 2021, 58(5): 102634.

［7］王若佳，王继民. 用户认知视角下在线问诊平台医生推荐研究［J］. 图书情报工作，2023，67（10）：128-138.

［8］黄锦泉，张雨欣，张楚，等. 基于在线问诊文本信息的线下就诊医院推荐研究［J］. 情报探索，2023（9）：88-93.

［9］Yuan H, Deng W. Doctor recommendation on healthcare consultation platforms: an integrated framework of knowledge graph and deep learning［J］. Internet Research, 2022, 32(2): 454-476.

［10］Kumar P M, Lokesh S, Varatharajan R, et al. Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier［J］. Future Generation Computer Systems, 2018, 86(1): 527-534.

［11］周欢，张培颖. 融合LDA和TF-IWF的健康科普文章混合推荐方法研究［J］. 图书馆研究，2022，52（3）：26-35.

［12］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）：195-198，206.

［13］吴菊华，王煜，黎明，等. 基于加权知识网络的在线健康社区用户知识发现［J］. 数据分析与知识发现，2019，3（2）：108-117.

［14］林萍，吕健超. 基于Stacking集成学习的在线健康社区问答信息采纳识别研究［J］. 情报科学，2023，41（2）：135-142.

［15］董伟，李建红，陶金虎. 在线健康社区活跃用户识别及其交互类型分析［J］. 文献与数据学报，

2020，2（1）：89–101.

［16］吕健超，林萍. 用户健康信息素养与问答文本情感特征对在线健康社区问答采纳影响分析［J］. 智能计算机与应用，2023，13（1）：5–11.

［17］金碧漪，许鑫. 社会化问答社区中糖尿病健康信息的需求分析［J］. 中华医学图书情报杂志，2014，23（12）：37–42.

［18］施亦龙，许鑫. 中美在线问答社区中的自闭症信息分析［J］. 中华医学图书情报杂志，2015，24（4）：5–8，31.

［19］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.

［20］张丽，张祯. 基于文本挖掘的新冠肺炎疫情下医药在线消费者的需求研究［J/OL］. 运筹与管理，1–8［2024–02–20］. http://kns.cnki.net/kcms/detail/34.1133.G3.20230901.0849.002.html.

［21］王海莲，闫素英，甄健存，等. 用药咨询标准制订与解析［J］. 医药导报，2022，41（10）：1439–1441.

［22］Mulder M B, Doga B, Borgsteede S D, et al. Evaluation of medication–related problems in liver transplant recipients with and without an outpatient medication consultation by a clinical pharmacist: a cohort study［J］. International Journal of Clinical Pharmacy, 2022, 44(5): 1114–1122.

［23］邓梓辛，徐传新. 一则利用药师服务患者流程进行门诊用药咨询实例［J］. 中国医院药学杂志，2021，41（20）：2142–2145.

［24］梅昕，冀连梅，彭小丹，等. 儿科药师基于“问药师”平台开展互联网用药咨询服务的实践［J］. 中国药房，2023，34（12）：1520–1523.

［25］景丽，何婷婷. 基于改进TF–IDF和ABLCNN的中文文本分类模型［J］. 计算机科学，2021，48（S2）：170–175，190.

［26］周欢，刘嘉，张培颖，等. 复杂网络视角下在线健康社区评论有用性研究［J］. 情报科学，2022，40（9）：88–97.

［27］杨冬菊，胡成富. 基于改进TextRank的科技文本关键词抽取方法［J/OL］. 计算机应用，1–9［2024–02–20］. http：//kns.cnki.net/kcms/detail/51.1307.tp.20230825.1605.006.html.

［28］杨延娇，赵国涛，袁振强，等. 融合语义特征的TextRank关键词抽取方法［J］. 计算机工程，2021，47（10）：82–88.

［29］冉从敬，李旺. 基于LDA的企业竞争对手识别模型构建——以蔚来汽车有限公司为例［J］. 情报理论与实践，2023，46（8）：88–95.

［30］Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation［J］. Journal of Machine Learning Research, 2003, 3(4–5): 993–1022.

［31］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.

［32］Kim C, Na Y. Consumer reviews analysis on cycling pants in online shopping malls using text mining［J］. Fashion and Textiles, 2021, 8(1): 38.

杨平，肖遗规 . 基于文本挖掘的在线健康社区用药咨询研究［J］. 文献与数据学报，2024，6（1）：080–091.

A Study of Online Health Community Medication Counseling Based on Text Mining

Yang Ping　Xiao Yigui

(College of Business , Hunan University of Technology, Zhuzhou 412007, China)

Abstract: ［**Purpose/Significance**］Online health community has become an important way for people to obtain health information, and studying the demand of medication consultation of online health community users contributes to the optimization and sustainable development of online health community drug service. ［**Method/Process**］Taking 39Health.com as an example, firstly, 59,048 medication consultation comments of gastrointestinal medicines are crawled using Python coding and pre-processed; secondly, the experimental data are subjected to theme keyword mining using text mining methods such as TF-IDF, TextRank, and LDA topic model, and keyword co-occurrence network analysis is carried out; lastly, a comprehensive analysis of the online health community users' medication consulting needs of the theme characteristics, and put forward optimization suggestions. ［**Result/Conclusion**］The results of this study show that online health community users are mainly concerned about the therapeutic effect of drugs, the way of taking drugs, adverse reactions, the difference between drugs, and the precautions to be taken during pregnant women and other special groups using drugs. This study provides a theoretical basis for drug manufacturers to adjust and optimize the content of drug manuals, and on the other hand, it provides a direction for online health communities to optimize the content layout of drug manuals and to establish or improve drug popularization services.

Keywords: Online health communities; Medication counseling; Text mining

（本文责编：王秀玲）