

网络健康社区信息需求特征测度 ——基于时间和主题视角的实证分析^{*}

李重阳, 翟姗姗, 郑路

(华中师范大学信息管理学院, 武汉 430079)

摘要: 以在线问答社区中癌症信息为例, 对其信息需求从时间和主题两个角度进行综合测度。对不同时间的数据进行分组, 采用LDA主题挖掘法来确定相关编码规则, 探求在线问答社区中用户信息需求特征, 发现用户对癌症信息的需求主要集中在若干个主题, 并且对各主题的关注度随时间呈现出一定的变化趋势。进一步探究在线问答社区健康信息服务存在的问题, 为促使各网络健康社区信息服务水平的提高提供参考和指导。

关键词: 在线问答社区; 信息需求; 特征测度; 癌症

中图分类号: G203; R730

DOI: 10.3772/j.issn.1673-2286.2016.9.006

1 引言

随着“互联网+”时代的到来, 互联网与各领域的融合发展已经成为不可阻挡的潮流。2015年7月, 我国政府正式公布《国务院关于积极推进“互联网+”行动的指导意见》(以下简称《意见》)^[1], 针对医疗健康领域, 《意见》提出要推广在线医疗卫生新模式、促进智慧健康养老产业发展, 充分利用互联网、大数据等手段建立医疗网络信息平台。互联网时代, 网络越来越成为人们获取健康信息的重要途径。根据2013年皮尤研究中心发布的调查报告显示, 35%的美国成年人在通过网络为自己或他人查询健康信息^[2]。相对于专业医学人士, 普通用户由于自身医疗知识有限, 更趋向通过健康社区获取健康保健知识、疾病治疗经验等。因此, 对网络健康社区信息进行研究, 发现用户信息需求, 可以更好地促进网络健康社区的运转和信息服务。用户对哪些主题的网络健康信息感兴趣, 用户的信息需求随时间的变化呈现出怎样的趋势, 这就需要对网络健康社区信息需求特征进行测度。本文以在线问答社区中癌症信息为例, 借助Latent Dirichlet

Allocation (LDA) 获取其主题范围, 并据此建立基于内容主题的编码规则, 通过对不同时间段的信息进行主题分布探测、编码分类、关键词提取和对比分析, 实现对用户信息需求特征的测度, 探索和预测网络健康社区信息需求的特征和变化趋势, 帮助其更好地组织信息资源, 增强用户获取健康信息的效率。

2 相关研究

2.1 网络健康信息研究现状

消费者健康信息 (Consumer Health Information) 指与大众、病患及其家属的有关健康和医学信息^[3]。网络健康信息在用户健康服务方面发挥巨大作用, 其研究主要集中在网络健康信息需求、网络健康信息搜寻行为和网络健康信息质量评估等方面。

在网络健康信息需求方面, 主要侧重对不同类型用户的健康信息需求内容、特点的研究。如刘亚君等将用户群体分为一般用户和特殊用户来探讨国外不同类型用户健康信息需求特点^[4]; Valero-Aguilera等发现

^{*} 本研究得到教育部人文社会科学研究青年基金项目“基于用户兴趣挖掘的电子政务门户知识整合研究”(编号: 13YJC870029)资助。

乳腺癌和泌尿系统癌症患者都希望了解不接受治疗的后果方面的信息^[5]。在网络健康信息搜寻行为方面,主要集中在搜寻行为的影响因素研究。施亦龙等提出健康信息搜寻行为受性别、种族、学历、收入等因素影响^[6];王猛将影响健康信息搜寻行为的因素分为人口学特征、社会支持、网络支持和信息需求的满足等五方面^[7];Yi等^[8]和Mesch等^[9]发现文化差异对使用互联网访问医疗信息的动机影响。网络健康信息种类繁多,来源广泛,因而其质量评估也是重要的研究方向。国内目前还没有成熟完整的医疗健康信息评价系统,而在20世纪90年代,国外就有商业机构、非营利性组织、图书馆和个人等对网络健康信息质量问题进行研究^[10],并且存在很多评价系统,如DISCERN系统等^[11];在评价机制方面,评价者可以是用户、第三方认证平台或者学科信息门户;在评价维度方面,主要涉及信息内容、网站结构和网站服务等方面^[12]。

2.2 网络健康社区信息需求研究进展

目前,互联网医疗健康信息服务平台数量巨大、种类繁多,其中既有面向专业人士的医疗健康网站,也有面向普通大众的门户网站。随着Web 2.0技术的发展,网络健康社区凭借其注重交互的特性,迅速成为广大用户搜寻相关信息的重要平台。目前与健康主题相关的内容主要集中在以专家问答、健康博客和病友论坛等文本信息为交流互动方式的网络健康社区平台上,其参与者包括普通民众、病患人群、病患家属、专业医师等各种类型的用户^[13]。对网络健康社区信息需求的研究有助于了解用户需求,提升网络健康社区服务水平。例如,金碧漪等以Yahoo!Answers和Diabetic Connect为数据来源,借助统计分析发现网络健康社区中有关糖尿病的信息主题分布特征,并对在线健康信息服务提出建议^[14];Sun等以Yahoo!Answers中有关饮食失调的问题为例,通过词频、词性、情感分析等,总结出相关主题^[15];Arden等选取400条涉及英国国家卫生医疗质量标准署发布的关于孕中体质量控制指导文件的论坛帖子为研究对象,发现孕中体质量受到很多因素的影响^[16];Valero-Aguilera等发现83%患乳腺癌的女性对治愈率、存活率以及疾病对外形的影响等方面的信息感兴趣^[5]。

网络健康信息数量巨大,网络健康社区形式多样,关于用户信息需求特征的测度研究也较多,但在测度对象、测度视角和测度方法上仍有研究空间。在

测度对象方面,多是以专门的医疗网站、医学论坛、医疗社区、专家问答平台等为对象,这类网络健康社区涵盖主题多、信息内容质量高、数据访问量大,是多数研究的数据来源;在大众化在线问答社区方面,以Yahoo!Answers为研究对象的居多,针对国内在线问答社区,探测用户健康信息需求的则很少;在测度视角上,多针对某一具体领域,而以时间为视角探讨信息需求主题变化的研究较少;在测度方法上,多是基于统计分析的方法(如人工标注和词频统计等)总结用户信息需求特征。近年来,主题模型LDA被应用于科技文献主题发现^[17]、微博主题及用户推荐^[18]、话题追踪^[19]等领域,也被应用于医学领域,如利用LDA和LSA(Latent Semantic Analysis)为医学文本和影像建立分析模型^[20];根据医学主题词在文献中的重要程度,采用LDA对不同时间窗的主题进行抽取,计算主题间的关联度等^[21],但是将LDA应用于在线问答社区健康信息主题测度的研究则较少。

3 网络健康社区信息需求特征测度研究方案

基于上述内容,本文通过对测度对象、测度视角和测度方法的改进进行研究,并绘制基于时间和主题视角的特征测度流程示意图(见图1)。

(1) 测度对象:以中国最大的在线问答社区“百度知道”为对象。在线问答社区基于User Generated Content模式建立,是一种供用户提问和回答的知识分享网站^[22],其中的信息已经成为普通大众日常生活、工作的重要参考依据。在线问答社区健康信息是网络社区健康信息的重要组成部分。与专业医学网站、医学论坛等相比,在线问答社区有用户来源范围广、层次多的特点,以此为研究对象更能了解普通大众的信息需求主题特征;问题针对性强,问答多是一对一或一对多,有助于了解用户个性化信息需求;涵盖领域多,涉及生活、教育、学习等众多领域,可全方位、多角度地了解用户信息需求。

(2) 测度视角:从时间和主题两个角度进行。将数据按照不同时间段进行分组,并采用LDA方法确定主题编码规则。

(3) 测度方法:结合LDA主题识别和人工标注等手段确定编码规则,通过分析不同时间段信息需求的主题分布来探索其变化趋势。

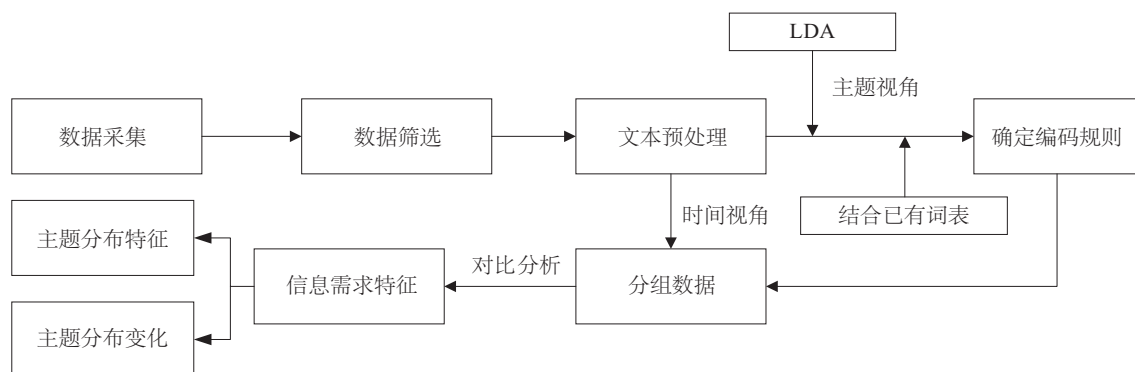


图1 基于时间和主题视角的特征测度流程示意图

3.1 数据采集与筛选

百度知道是全球最大的中文互动式知识问答分享平台。2015年9月,百度知道宣布在其平台上解决的问题量已突破4亿个,成为广大网民的生活伙伴^[23]。其涉及问题包括经济金融、法律法规、科学教育等15类,其中和医疗健康信息有关的3类(心理分析、健康生活、医疗卫生)。与本文研究主题最相关的是医疗卫生类,包括皮肤科、儿科、外科、内科等。根据WHO发布的《全球癌症报告2014》,全球癌症病例将呈现迅猛增长态势,由2012年的1 400万人,预计至2035年将达到2 400万人^[24],由此可见,癌症已经成为威胁人们健康生活的主要病症之一。

为保证数据有较好的比对性,本文选取医疗卫生类中的问答记录作为数据来源。在百度知道“医疗卫生”栏目,以“癌症”为关键词进行检索,利用八爪鱼数据采集工具对检索的问答记录结果进行抓取,并辅以人工判读剔除无效检索结果,截至2015年12月31日,共获取1 500条问答记录。

3.2 文本预处理

在线问答社区的问答记录来自网络用户,由于用户知识水平和专业背景各异,以及使用词汇的极大不确定性,因此数据质量参差不齐,需要对其进行文本预处理。本文利用NLPIR汉语分词系统对数据进行预处理,保留与主题分析相关的高频词。

3.3 编码规则

采用数据编码方法可将众多数据划分到对应具体

类目中,从而使数据有序化,因此常被用于社会化问答平台的问题分析^[25]。传统网络健康社区的数据编码规则多采用人工方法确定,研究者结合已有主题词表,根据采集数据的实际情况不断调整编码规则,最终形成主题划分策略。这种方法简单实用,但人为影响因素大,易出现分类不客观的情况。基于此,本文采用LDA和人工结合的方式确定主题编码规则。

LDA是由Blei等提出的主题发现模型^[26],其在概率潜在分析模型(Probabilistic Latent Semantic Analysis, PLSA)的基础上增加Dirichlet先验分布^[27]。LDA是一个三层贝叶斯概率模型,包含文档、主题和词项三层结构,可用于识别大规模文档集或语料库中潜在主题信息。为更好地解释LDA基本原理,本文对一些符号定义说明(见表1)。

表1 符号定义说明

参 数	定 义
D	文档集合
a	θ 的狄利克雷分布先验参数
β	φ 的狄利克雷分布先验参数
θ	文档—主题分布
φ	主题—词分布
d	集合中单个文档
z	集合中主题
M	文档数
m	θ_m 表示第 m 个文档—主题分布
K	主题数
k	φ_k 表示第 k 个主题—词分布
V	每个主题包含词数
N	文档包含词数
n	W_n 表示文档中第 n 个单词
W	文档中的词

将每条问答记录作为一个单独文档,对三层贝叶斯概率模型而言,一篇文档的生成过程包括(1)选择 N , $N \sim \text{Dir}(\beta)$, N 表示文档长度;(2)抽取 θ_m , $\theta_m \sim \text{Dir}(a)$, θ_m 是文档在主题上的分布;(3) For $n = 1$ to N ; (4)选择 $z_{mn} \sim \text{Multinomial}(\theta_m)$; (5)抽取单词 $w_{mn} \sim \text{Multinomial}(\phi_{z_{mn}})$ 。

借助LDA模型,要估计的参数有两类: $\bar{\theta}_1, \dots, \bar{\theta}_M$ 和 $\bar{\phi}_1, \dots, \bar{\phi}_K$ 。

假设数据集 $\bar{z}$ 中的第 i 个词对应的Topic为 z_i , 其

中 $i = (m, n)$ 是一个二维下标,对应于第 m 篇文档中的第 n 个词,用 $\neg i$ 表示去除下标为 i 的词。根据Gibbs Sampling算法要求,假设已经观测到的词 $w_i = t$, 得到文档—主题分布和主题—词分布。

$$\hat{\theta}_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^K (n_{m,\neg i}^{(k)} + \alpha_k)} \quad (1)$$

$$\hat{\phi}_k = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^V (n_{k,\neg i}^{(t)} + \beta_t)} \quad (2)$$

最终得到若干主题及主题下对应的关键词,整体主题挖掘效果见图2。

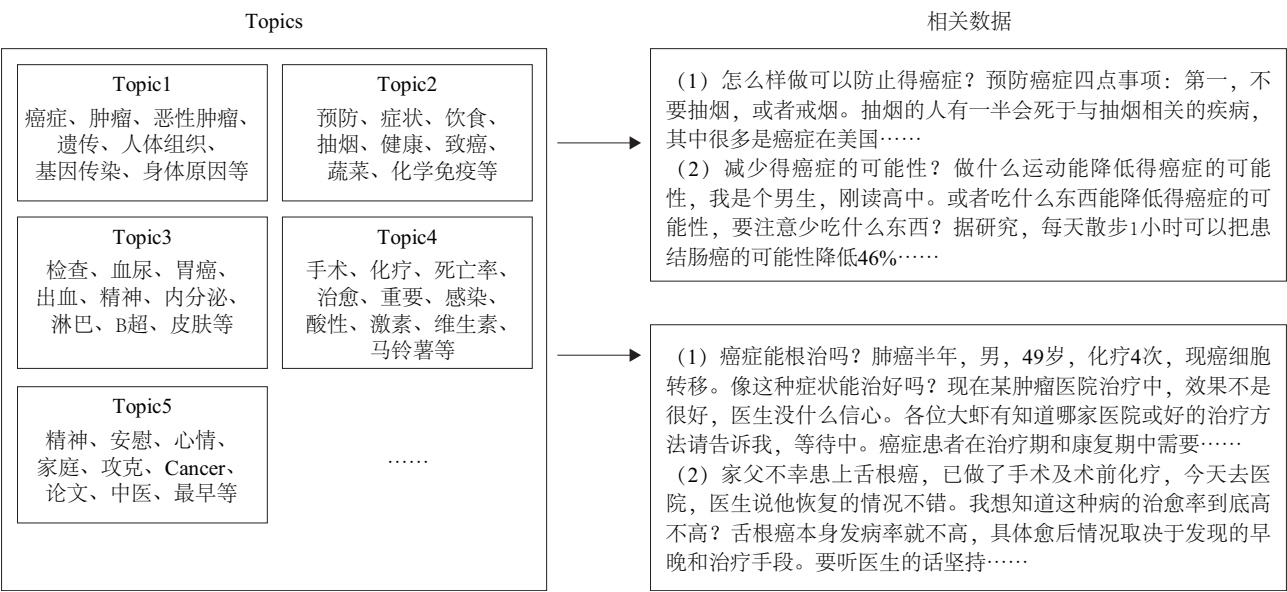


图 2 主题挖掘效果图

利用LDA模型对文档集合进行训练得到相关参数,包括Topic以及每个Topic具体对应的主题词和概率。对于某一篇文章,假设其主题—词分布不变,借助已有模型,可以探测其文档—主题分布,并把其中概率最大的Topic作为该文档的主题,以此作为文档归类的

依据。因为在线问答社区中的信息涉及医学领域知识,单纯依靠LDA分类不能保证完全准确,因此需结合其他信息,如PubMed、ACS (American Cancer Society) 和中国抗癌协会等相关文献,并最终形成癌症信息主题划分策略,见表2。

表 2 癌症主题分类策略

主 题	子类目	具体内容
T1基础病理知识	T1.1定义、病因、种类	癌症病因、关于癌症的基础知识、基本原理等
	T1.2 症状	患癌的表现或某种癌症病症的症状
T2疾病预防	-	询问如何预防癌症,包括健康的生活习惯等
T3诊断、检查	-	直接描述症状或通过指标询问是否得癌症或询问身体状况; 询问癌症的遗传问题,描述家庭癌症病史,询问自己患癌的可能性; 是否得病的检查以及得病例行检查等

续表

主 题	子类目	具体内容
T4治疗	T4.1药物、手术	治疗癌症的特效药、偏方; 手术治愈的概率等
	T4.2 注意事项	癌症防止复发、转移; 癌症病人饮食注意问题、食疗、锻炼等
T5其他	T5.1救助措施	保险、癌症协会、救助保障等
	T5.2情感生活	病人个人情绪、心理, 以及对家庭的影响等
	T5.3教育研究	提到作业、教育中的一些癌症问题、研究等

从表2可见, 用户的信息需求主要集中在基础病理知识、预防、诊断、治疗及其他等方面。这些信息需求主题与癌症疾病的特点密不可分, 由于癌症逐渐成为常见病、多发病, 死亡率极高, 发病初期不易被察觉, 因而用户对病因、症状等基础病理知识方面的信息需求尤为迫切, 也占据较大比重。癌症发病与环境有较大关系, 随着健康意识的提高, 用户对疾病预防方面的信息需求逐渐增多。此外, 不同类型人群信息需求的内容不同, 未患癌用户较多关注疾病预防、诊断检查方面信息; 已患病用户和其家属, 多关注药物手术治疗以及日常生活注意事项等信息; 部分用户关注社会救助, 如商业保险、抗癌协会等信息。癌症对患者心理产生巨大影响, 使其出现焦虑、恐惧、抑郁等不良情绪, 患者及其家属十分关注心理健康辅导、情感安抚方面的信息。对教育研究方面信息有需求的用户, 基本不是病患或其家属, 多属于学生或教育研究者。

4 网络健康社区信息需求特征测度结果分析

本文根据实际采集情况将数据按照时间维度分为3组。首先, 对数据进行整体描述; 其次, 分别对3组数据进行文本预处理和主题编码; 最后, 对3组数据信息需求内容进行特征测度, 探测用户信息需求主题分布和变化, 并预测其发展趋势。

4.1 数据整体情况描述

为保证不同分组有足够数据量, 也为比较不同时间段信息需求的主题变化, 根据不同时间的数据采集量和采集总量, 将数据分为2005—2009年、2010—2012年、2013—2015年, 数据量总体呈逐年上升趋势

(2015年采集量较少), 这和发展实际情况相符。每个时间段为一组, 保证每组的问答记录, 具体情况见表3。

表3 数据整体情况描述

组 别	时间/年	问答记录数/条	问答记录总数/条
第1组	2005	19	477
	2006	72	
	2007	101	
	2008	136	
	2009	149	
第2组	2010	133	514
	2011	154	
	2012	227	
第3组	2013	245	509
	2014	239	
	2015	25	

首先, 对3组数据分别进行文本预处理, 将每组所有记录合并为一个文档, 经过引入停用词表、去除无意义词、分词等操作, 探测其中关键词, 具体结果见表4。

因为数据采集以用户提问中包含“癌症”为依据, 因此“癌症”“肿瘤”等关键词频次较高。权重主要由词频TF、倒转文档频率TDF及关键词出现位置3个指标综合决定。

4.2 信息需求主题分布特征

从表5可见, 不同时间段用户信息需求内容主题分析包括两方面。一是主题分类分析, 即根据编码规则对数据进行分类处理, 形成各自类目数据, 每组都

表 4 分组数据高频关键词部分情况

第1组		第2组		第3组	
关键词	权 重	关键词	权 重	关键词	权 重
癌症	5.497 3	癌症	5.382 4	癌症	5.744 0
肿瘤	4.807 2	治疗	4.557 5	治疗	4.923 1
治疗	4.529 9	肿瘤	4.460 1	细胞	4.843 4
细胞	4.516 7	患者	4.376 2	肿瘤	4.823 2
食物	4.425 9	细胞	4.212 7	癌细胞	4.733 3
.....					
检查	4.078 0	预防	3.913 1	疾病	4.214 3
遗传	4.054 4	恶性肿瘤	3.857 0	恶性肿瘤	4.166 7
疾病	3.920 4	体质	3.706 3	化疗	3.998 3
抑制	3.862 3	食品	3.696 4	生长	3.987 4
抗癌	3.846 3	手术	3.637 0	健康	3.983 0

对应若干子类目文档。二是关键词提取, 首先, 对各组文档进行处理, 将每个目录下所有词出现总频次和所有词数目的商作为选取关键词的阈值, 保留高于阈值的词; 其次, 借助LDA训练模型, 得到相应主题对应的高频词; 最终, 确定各类目关键词, 并进行主题分析。

在实际操作中, 不能将LDA方法作为文档归类的唯一手段, 因为部分问题可能涉及若干类目, 如出现某一个问答记录属于两个主题情况, 针对此类问题, 要以实际解答的内容为依据, 将其归入较符合的主题类目。

此外, 同一关键词也会被归入不同的类目, 例如食物类关键词, 当用户询问“哪些食物可以预防癌症”时, 则应将其归入疾病预防类目; 当用户询问“癌症患者是否可以食用某类食物”时, 应将其归入治疗类目。此外, 由于某些关于症状的问答记录无法判断用户的真实意图是简单咨询癌症的基础知识, 还是通过询问判断是否得病, 不能简单将这类问题归入“基础病理知识”类, 也不能归入“诊断、检查”类, 因此将“症状”作为一个单独子类目。

表 5 主题编码结果

主 题	子类目	问答记录数/次			关键词数/个			比 例/%		
		第1组	第2组	第3组	第1组	第2组	第3组	第1组	第2组	第3组
T1 基础病理知识	T1.1 定义、病因、种类	131	125	95	46	40	37	27.46	24.31	18.66
	T1.2 症状	38	28	18	10	15	11	7.97	5.45	3.54
T2 疾病预防	-	78	94	103	13	16	28	16.35	18.29	20.24
T3 诊断、检查	-	70	82	82	21	30	31	14.65	15.95	16.11
T4 治疗	T4.1 药物、手术	52	70	72	20	23	26	10.90	13.62	14.15
	T4.2 注意事项	40	44	50	11	9	13	8.39	8.56	9.82
T5 其他	T5.1 救助措施	5	9	12	4	3	8	1.05	1.75	2.36
	T5.2 情感生活	26	21	24	9	8	11	5.45	4.09	4.71
	T5.3 教育研究	37	41	53	14	12	17	7.78	7.98	10.41
总计		477	514	509	148	156	182	100	100	100

4.3 信息需求主题分布变化

4.3.1 用户信息需求主题整体分布特征

对不同分组数据进行编码和主题分析后,发现3组数据主题分布基本集中在基础病理知识、疾病预防、诊断检查、治疗和其他(社会保障、情感生活和教育研究)方面,并呈现较一致的分布状态。其中基础病理知识、疾病预防以及诊断、治疗这三方面是用户较为关心的,相关问答记录数也较多,占有较大的比重。以2010—2012年数据为例,可以看出各主题分布的大致情况,见图3。

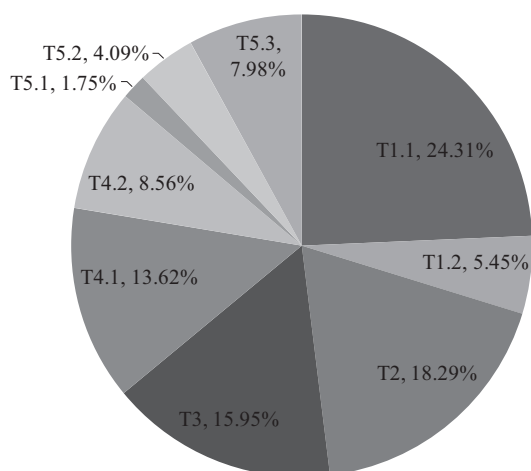


图3 2010—2012年各主题分布情况

国外关于在线问答社区癌症信息需求的研究多以Yahoo!Answers为研究对象^[28]。在线问答社区中用户对癌症的信息需求可分为医学类和非医学类,医学类用户较为关注治疗方面信息,非医学类,关于医院或医生的建议是用户经常讨论的话题;在具体类型癌症中,乳腺癌是用户关注最多的病症,其次是宫颈癌和肝癌^[29];此外,其他学者发现患者和医疗保健人员更关注预防和诊断类信息^[30]。由此可见,国内外用户在网络健康社区信息需求主题存在细微差别。

4.3.2 信息需求主题分布变化趋势

用户对健康信息的需求也呈现出一些变化,见图4。

(1) 用户对基础病理知识的需求虽然仍占比较大,但总体呈现下降趋势。这与癌症疾病自身特点紧密

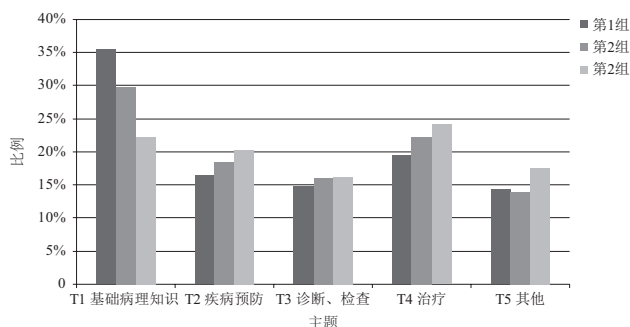


图4 3组数据主题分布情况

相关,用户较为关心诸如“癌症和肿瘤的关系”“癌症病因”“致死机理”等基础性问题。随着健康意识和知识的增强,人们越来越关注癌症信息,对其了解也越来越深入,因而对基础病理知识方面的信息需求呈现下降趋势。

(2) 关于癌症预防和治疗方面的信息需求呈现上升趋势。特别是癌症预防中,经过分析具体的文本信息,发现关于健康生活习惯及抗癌食品等讨论居多。从有关癌症症状问题中已逐渐可以判断用户意图是单纯知识普及还是进行自我诊断,说明用户对待癌症的态度有所转变。

(3) 关于用户对其他方面的信息需求,如癌症的社会救助、癌症对患者以及家庭、社会的影响等,在2013年前没有明显变化,从2013年之后则得到较多关注。

4.3.3 在线问答社区存在问题

在线问答社区健康信息存在一些问题。(1) 无法确定用户的真实意图。比如“胃癌的症状是长时间的反复胃痛吗?”和“我这几个月一直胃痛,是胃癌吗?”两个问题,后者可以判断出用户在进行简单自我诊断,而前者无法判断用户是简单询问还是主动自我诊断,在实际归类时可能会出现偏差。(2) 信息质量参差不齐。由于用户自身知识背景和素质等存在较大差异,因此提问、回答的质量差别较大。在实际操作中,相互参考答案现象比较严重,甚至出现答非所问的情况,给编码工作和文本处理带来困难,同时也影响在线问答社区的健康信息服务水平。(3) 类目不够细化。根据专业医学主题词表和相关网站等资料,癌症种类涉及几十种,年龄、性别、地区等均是癌症相关影响因素。但在线问答社区对相关信息的类目划分不够具体,不利于用户个性化信

息的获取。

5 总结

分析在线问答社区中健康信息需求主题分布特征和变化趋势,探讨其中的现象、原因和存在问题,可以对在线问答社区健康信息服务提供三点启示。

(1) 对在线问答社区健康平台服务者而言,要明确用户感兴趣的**健康主题**以及各主题的**差异程度**。用户对不同种类、不同主题的健康信息需求不同,例如,糖尿病是慢性疾病,其治疗多依靠病人在日常生活的自我调理,因而用户较为关注日常管理^[31];而癌症的形成多与遗传或者环境因素相关,良好的生活习惯会降低患癌的可能性,因此用户对于预防方面的信息有较大需求。

(2) 细化类目,提供更加个性化的信息服务。在线问答社区能够给广大用户提供健康指导,但其信息处在内容量大且杂乱无章的状态。在线问答社区应提高专业水准,针对不同领域的健康信息进行主题挖掘分析,对用户创造的信息进行管理,为其创建科学、专业、更加细化的类目体系,引导用户发现自身真正需要的信息,并以主题为依据进行分类导航,不同主题侧重程度根据用户关注程度进行区分,以此设置导航节点和更加人性化的服务界面,从而提高在线问答社区健康信息的服务质量。

(3) 增强用户对健康信息的管理。除在线问答社区管理人员对信息加强管理外,还应允许用户添加社会化标签,引导用户对自身健康信息进行归类。社会化标签作为用户直接产生的数据,可以直接反映用户需求及其变化,为在线问答社区更好的信息服务提供依据。

当然,本文存在诸多不足之处,接下来进行更深入的研究。

(1) 为保证数据有较好的对比性,使3组数据量相同,同时导致各组间界限不明显,因而用户需求变化虽呈现一定趋势,但不甚明显。接下来在测度体系中要扩大数据量,应包含各类型网络健康社区和信息,还要细化研究粒度,保证其科学性。在分类操作中,采用LDA进行数据训练提高科学性,但要结合人工判断,思考如何结合在线问答社区信息特点对LDA模型进行改造,从而提高模型效率和准确度。

(2) 在研究中发现,单纯根据用户问答记录不能完全判断用户的意图和情感状态,后续研究应该综合各种因素,探究用户真实需求和情感倾向,包括对不同

群体的需求特征以及需求的影响因素等方面的研究。

(3) 在线问答社区中的信息多是用户自发组织形成,其质量参差不齐,因而健康信息质量评估也是重要研究方面。

参考文献

- [1] 国务院发布医疗健康领域“互联网+”行动指导意见(附全文)[EB/OL]. (2015-07-06)[2016-01-12].<http://news.hc3i.cn/art/201507/33839.htm>.
- [2] FOX S,DUGGAN M.Health Online 2013[EB/OL].[2016-01-12].<http://www.pewinternet.org/2013/01/15/health-online-2013/>.
- [3] MLA,CAPHIS.The librarian's role in the provision of consumer health information and patient education[EB/OL].[2016-01-20].<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC299415/pdf/mlab00375-0088.pdf>.
- [4] 刘亚君,兰小筠.国外网络健康信息用户研究进展及启示[J].中华医学图书情报杂志,2011(7):38-41.
- [5] VALERO-AGUILERA B,BERMÚDEZ-TAMAYO C,JIMÉNEZ-PERNETT J, et al.Information needs and internet use in urological and breast cancer patients[J].Supportive Care in Cancer Official Journal of the Multinational Association of Supportive Care in Cancer, 2014,22(2):545-552.
- [6] 施亦龙,许鑫.在线健康信息搜寻研究进展及其启示[J].图书情报工作,2013(24):123-131.
- [7] 王猛.美国网络健康信息用户获取行为的影响因素研究[J].信息资源管理学报,2013(3):47-58.
- [8] YI Y J,STVILIA B,MON L. Cultural influences on seeking quality health information: an exploratory study of the Korean community[J]. Library & Information Science Research, 2012,34(1):45-51.
- [9] MESCH G,MANO R,TSAMIR J.Minority status and health information search: a test of the social diversification hypothesis[J]. Social Science & Medicine, 2012,75(5):854-858.
- [10] 魏萌萌,马敬东,夏晨曦.国内外网络健康信息质量评估工作研究综述[J].中国卫生事业管理,2012(7):551-553.
- [11] Discern.Quality criteria for online consumer health information on treatment choices[EB/OL].[2016-01-30].<http://www.discern.org.uk/hoti.php>.
- [12] 孙丽,曹锦丹.国外网络健康信息质量评价系统的应用现状及启示[J].医学与社会,2011(7):15-17.
- [13] 吕英杰.网络健康社区中的文本挖掘方法研究[D].上海:上海交通大学,2013.
- [14] 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(12):100-105.

- [15] SUN O J, HE D, WEI J, et al. Linguistic characteristics of eating disorder questions on Yahoo! Answers content, style, and emotion[J]. Proceedings of the American Society for Information Science & Technology, 2013, 50(1):1-10.
- [16] ARDEN M A, DUXBURY A M, SOLTANI H. Responses to gestational weight management guidance: a thematic analysis of comments made by women in online parenting forums[J]. Bmc Pregnancy & Childbirth, 2014, 14(1):1-12.
- [17] 王平. 基于层次概率主题模型的科技文献主题发现及演化[J]. 图书馆情报工作, 2014, 58(22):70-77.
- [18] 邸亮, 杜永萍. LDA模型在微博用户推荐中的应用[J]. 计算机工程, 2014, 40(5):1-6, 11.
- [19] 张晓艳, 王挺, 梁晓波. LDA模型在话题追踪中的应用[J]. 计算机学, 2011(S1):136-139, 152.
- [20] 李博. 基于LDA和LSA的医学文本和影像分析模型及应用研究[D]. 长春: 吉林大学, 2012.
- [21] 安新颖. 基于LDA的医学新主题监测模型研究[C]//中华医学会第十八次全国医学信息学术会议. 中华医学会第十八次全国医学信息学术会议论文集. 成都: 中华医学会, 2012.
- [22] SHAH C, OH J S, OH S. Exploring characteristics and effects of user participation in online social Q & A sites[J/OL]. First Monday, 2008, 13(9)[2016-01-13]. https://www.researchgate.net/publication/220167918_Exploring_Characteristics_and_Effects_of_User_Participation_in_Online_Social_QA_Sites. DOI:10.5210/fm.v13i9.2182 • Source: DBLP.
- [23] 中国新闻网. 百度知道已解决问题量破4亿 解决13亿人困惑[EB/OL]. (2015-09-21)[2016-01-30]. <http://www.chinanews.com/it/2015/09-21/7536004.shtml>.
- [24] 世界卫生组织(WHO)发表《全球癌症报告2014》[EB/OL]. [2016-01-30]. <http://web2.fimmu.com/yxy/new/news/show/?id=476>.
- [25] ZHAGN J, ZHAO Y. A user term visualization analysis based on a social question and answer log[J]. Information Processing & Management An International Journal, 2013, 49(5):1019-1048.
- [26] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.
- [27] HOFMANN T. Probabilistic latent semantic analysis[C]//Proceedings of the 15th conference on uncertainty in artificial intelligence. San Francisco: Morgan Kaufmann Publishers, 1999:289-296.
- [28] SANGHEE O, Zhang Y, Min S P. Cancer information seeking in Social Q & A: identifying health-related topics in cancer questions on Yahoo!Answers [J/OL]. Informing Research, 2016 [2016-08-30]. [http://www.researchgate.net/publication/305465646_Cancer_Information_Seeking_in_Social_QA_I](http://www.researchgate.net/publication/305465646_Cancer_Information_Seeking_in_Social_QA_Identifying_Health-Related_Topics_in_Cancer_Questions_on_YahooAnswers)dentifying_Health-Related_Topics_in_Cancer_Questions_on_YahooAnswers.
- [29] PARK H, MIN S P. Cancer information-seeking behaviors and information needs among Korean Americans in the online community[J]. Journal of Community Health, 2014, 39(2):213-220.
- [30] CHO J, NOH H I, HA M H, et al. What kind of cancer information do Internet users need?[J]. Supportive Care in Cancer, 2011, 19(9):1465-1469.
- [31] 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, (12):37-42.

作者简介

李重阳, 女, 1991年生, 情报学硕士。

翟姗姗, 女, 1986年生, 副教授, E-mail: Zhais@mail.ccnu.edu.cn。

郑路, 男, 1990年生, 情报学博士。

Measurement of Information Demand Characteristics in Online Health Community: an Empirical Analysis Based on Time and Theme Perspective

LI ChongYang, ZHAI ShanShan, ZHENG Lu

(School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract: Taking cancer as an example, this paper measures the features of information demand in online Q&A platform from the perspectives of time and theme. It gets the features of information demand by the methods of dividing the data into different groups and coding based on LDA. The results show that the need for cancer information is mainly focused on some topics and changing over time. This paper also analyses the problems of online Q&A platform. All these finds are benefit to improve online health information service.

Keywords: Online Q & A Platform; Information Demand; Characteristics Measurement; Cancer

(收稿日期: 2016-08-31)