

# 艾滋病健康信息需求研究\*

刘艳丽 袁文娟

(杭州医学院图书馆 杭州 310053)

〔摘要〕 以百度知道中关于艾滋病的提问为数据基础,利用 Python 语言对提问内容进行词频和共词网络分析,揭示艾滋病健康信息需求 5 大主题,分析艾滋病健康信息用户特征,提出艾滋病健康信息传播建议,为精准推送健康信息提供参考。

〔关键词〕 文本挖掘; 艾滋病; 健康信息需求

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.07.010

**Study on AIDS Health Information Needs** LIU Yanli, YUAN Wenjuan, Library of Hangzhou Medical School, Hangzhou 310053, China

〔Abstract〕 Taking the questions about AIDS in Baidu Zhidao as data basis, the word frequency and co-word network of the question content are analyzed by using Python language. The paper reveals the five major topics of AIDS health information needs, analyzes the characteristics of AIDS health information users, and proposes the countermeasures for the spread of AIDS health information, so as to provide references for the accurate push of health information.

〔Keywords〕 text mining; AIDS; health information needs

## 1 引言

艾滋病是由艾滋病病毒即人类免疫缺陷病毒(HIV)引起的一种病死率极高的恶性传染病,目前没有疫苗可以预防,也没有治愈该疾病的有效药物和方法,严重危害公众健康<sup>[1]</sup>。加强艾滋病防治工作是实施健康中国战略的重要任务。国家卫生健康委联合多部门印发《遏制艾滋病传播实施方案

(2019-2022 年)》(以下简称《方案》),第 1 项措施即实施预防艾滋病宣传教育工程,要求居民、流动人口、老年人及易感染艾滋病危险行为人群艾滋病防治知识知晓率达 90% 以上;各地积极探索利用大数据信息、人工智能技术判断艾滋病防治重点人群和对象,通过互联网精准推送防治信息<sup>[2]</sup>。开展健康信息传播对于艾滋病防治工作至关重要,而分析用户对相关健康信息的需求特征能更精准地为其提供所需健康信息。

## 2 研究现状

### 2.1 国外

网络因其获取信息的便捷性,已经成为公众获取健康信息的重要途径。艾滋病患者由于社会歧视与排斥,也更偏向于通过网络获取相关健康信息。

〔收稿日期〕 2019-12-30

〔作者简介〕 刘艳丽,馆员,发表论文 7 篇,参编著作 1 部。

〔基金项目〕 杭州市哲学社会科学规划课题资助项目“‘双一流’战略下浙江省科研机构国际学术论文影响力评价研究”(项目编号:Z18JC072)。

通过网络用户不仅可以浏览信息,还可主动提问,有研究发现网络问答社区是用户获取健康信息的重要途径<sup>[3]</sup>。目前国内关于艾滋病健康教育的需求研究以面向特定人群的问卷调查为主,由于问题预先设定,不能全面反映公众真实需求。对社会问答平台中关于健康话题的提问进行研究,能有效反映用户对相关健康信息的需求。国外关于社会问答平台健康信息需求与行为的研究已逐步兴起<sup>[4]</sup>, Park MS 等采用内容分析与语义网络分析的方法分析网络用户关于乳腺癌话题,以研究用户信息搜索行为<sup>[5]</sup>; Ji Z 等通过分析雅虎问答和百度知道中关于癌症的问题,对中美两国用户健康信息需求进行分析<sup>[6]</sup>; Alasmari A 等对比分析雅虎问答和 WebMD 两种社会化问答平台的健康信息需求<sup>[7]</sup>。

## 2.2 国内

国内有少数学者开始尝试研究社会问答社区中用户健康信息需求,如邓胜利等利用 ROST CM 文本挖掘软件分析百度知道中有关高血压的提问<sup>[8]</sup>;赵安琪等采用内容分析法对知乎中关于抑郁症的提问进行研究<sup>[9]</sup>;曹树金等采用社会网络分析法揭示社会化问答网站用户糖尿病相关信息需求<sup>[10]</sup>;Zhou X 等对中国某健康网站中关于糖尿病的提问进行分析<sup>[11]</sup>。但国内关于艾滋病健康信息需求的研究鲜有报道,王国华等通过分析百度贴吧“HIV 吧”版块内容,发现其中寻求社会支持最普遍的是信息类,寻求建议和参考,需求信息远多于供应信息<sup>[12]</sup>,但对用户信息需求内容未做深入研究。本文对百度知道中关于艾滋病的提问进行文本挖掘,分析用户对艾滋病健康信息的需求特点,为精准推送相关健康信息提供科学依据。

## 3 资料与方法

### 3.1 数据来源

主要以百度知道为数据基础。百度知道是基于百度搜索的互动式知识问答分享平台,用户有针对性地提出问题,通过积分奖励机制发动其他用户来解决该问题。同时这些问题的答案会作为搜索结

果,提供给其他有类似疑问的用户,达到知识分享效果<sup>[13]</sup>。本文主要研究非医学专业人员对艾滋病相关信息的需求,在获取相关数据时未对艾滋病专业术语“获得性免疫缺陷综合征”进行检索。经过初步检索发现用户除以“艾滋病”一词进行提问外,还常用其英文缩写“AIDS”,“HIV”也较多地用于提问中,因此本研究以“艾滋病 + AIDS + HIV”为关键词(+表示逻辑或的关系),通过八爪鱼采集器获取百度知道相关提问内容,包含发布时间、回答者、回答个数、点赞数、标题、问题详情、浏览次数、最佳答案等信息。数据获取时间为 2019 年 12 月 11 日。利用 Excel 对采集到的问题内容进行清洗,筛选掉重复数据、没有提问内容或提问内容与艾滋病无关数据,最终研究样本共有 1 579 条提问数据。

### 3.2 方法

利用 Python 对采集的问题样本进行分词及词频分析,采用 Gephi 进行语义网络分析。分词是进行文本分析的基础,通过 Python 中的 jieba 分词对提问内容进行分词处理。在实际分词过程中,根据初步分词结果将提问中涉及但未被正确切分的词语添加为自定义词语并重新分词。对分词结果进行词频统计,使用 Python 中的 Wordcloud 库制作关键词云图,以揭示用户对艾滋病的主要关注点。采用 Gephi 绘制关键词共现网络图,以分析用户对艾滋病健康信息需求的主题分布。由于艾滋病及相关概念有多种表达方式,为更准确地分析语义网络,对这些同义词或不规范的表达进行统一处理,将概念统一为“艾滋病”和“HIV”,将“艾滋病病人”、“艾滋病人”统一为“艾滋病患者”,“艾滋病宣传日”统一为“艾滋病日”。

## 4 结果

### 4.1 高频关键词

对研究样本分词后进行词频统计,共得到 2 167 个词,筛除无实际意义的词如“请问”、“到底”、“谢谢”等,去掉征求艾滋病的作文或英文翻译等

方面的词如“作文”、“英文”、“一篇”等。取筛选后频率最高的前 100 个词, 绘制词云图, 见图 1。根据词云图的生成原理, 词语出现频率越高, 其显示规格越突出、离图中心越近。图 1 可以直观显示出用户对艾滋病健康信息需求的关键词: 艾滋病、艾滋病病毒 (hiv)、感染、患者、传染、检测、预防、症状、传播、检查等, 频率最高的 20 个关键词及词频, 见表 1。分析词频时 HIV 自动转换为 hiv, 词云图中以 hiv 形式出现。



图1 艾滋病相关问题关键词

表1 前20位高频关键词及其词频

| 序号 | 关键词 | 词频    | 序号 | 关键词  | 词频 |
|----|-----|-------|----|------|----|
| 1  | 艾滋病 | 1 292 | 11 | 不会   | 52 |
| 2  | hiv | 284   | 12 | 检查   | 52 |
| 3  | 感染  | 101   | 13 | 艾滋病日 | 49 |
| 4  | 患者  | 100   | 14 | 传播   | 45 |
| 5  | 传染  | 69    | 15 | 中国   | 45 |
| 6  | 检测  | 64    | 16 | 病毒   | 41 |
| 7  | 预防  | 57    | 17 | 哪里   | 40 |
| 8  | 世界  | 55    | 18 | 意思   | 35 |
| 9  | 症状  | 55    | 19 | 问题   | 34 |
| 10 | 哪些  | 54    | 20 | 如何   | 32 |

## 4.2 共词网络

为了解各关键词之间的关联,揭示信息需求主

题分布,根据高频关键词共现关系构造共现网络,见图2。节点代表关键词,节点越大其中介中心性越高,表明该关键词更容易与其他关键词共现;连边代表关键词之间的共现关系,连边越粗,关键词之间的共现频率越高,关系越密切;根据聚类系数对节点颜色进行区分,相同颜色代表节点间关联度较强。由于本研究的问题样本都围绕艾滋病,故在构建共现网络时将“艾滋病”这个节点去掉,以更清晰地体现其他节点之间的关联度。由图2可见节点主要分为5种颜色,分别代表5种不同主题。

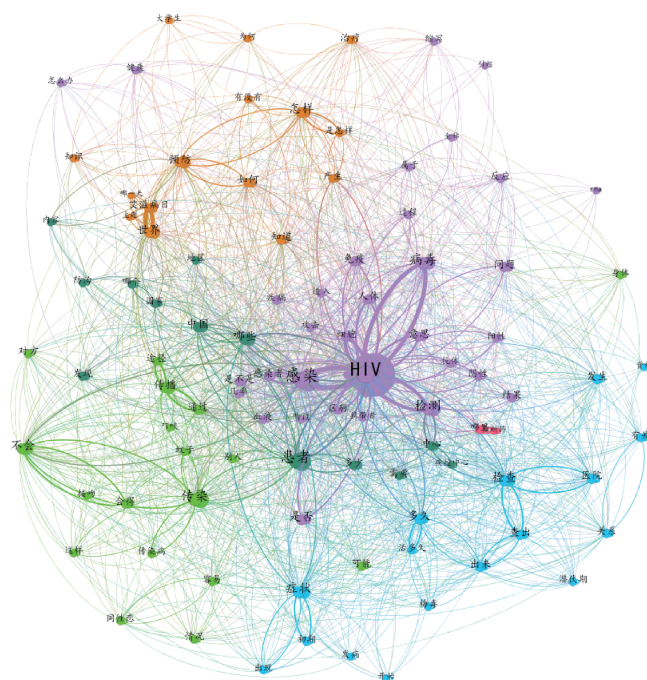


图2 关键词共现网络

## 5 讨论

### 5.1 主题分布

结合表1和图2分析高频词及词间关系,整理得出用户对艾滋病健康信息的需求主要集中在5个方面,见表2。此外还有少数与艾滋病密切相关的问题,如艾滋病起源、危害,HIV与艾滋病的区别,艾滋病患者能否生育等。

表 2 艾滋病健康信息需求主题及关键词

| 信息需求主题 | 主要内容                   | 关键词（前 5 位）        |
|--------|------------------------|-------------------|
| 艾滋病病毒  | 艾滋病病毒的感染与检测            | HIV、感 染、病 毒、检测、细胞 |
| 流行趋势   | 艾滋病患者数量、高发地区等          | 患者、哪些、中国、多少、国家    |
| 传播与传染  | 艾滋病的传播途径、与患者接触后被传染的可能性 | 传染、传播、患者、不会、通过    |
| 临床诊断   | 艾滋病的症状、医院检查、潜伏期、生存期等   | 症状、哪些、检查、多久、医院    |
| 预防知识   | 艾滋病的预防措施、世界艾滋病日等       | 预防、怎样、如何、世界、艾滋病日  |

5.2 用户特征

本研究中艾滋病健康信息需求来自不同人群。关注艾滋病健康信息的用户除感染者或高危人群外还有其他人员，这部分人群主要关注艾滋病流行趋势和预防知识以及传播途径和致病机理等，表明艾滋病具有比较广泛的社会影响。艾滋病病毒感染者或疑似患者更多地关注艾滋病传染与临床诊断，包括艾滋病病毒检测等，对治疗等信息关注较少。根据 HIV 传播途径，容易引起 HIV 传播的行为称为高危行为<sup>[14]</sup>。其中大部分用户未确诊，部分人与患者或病毒携带者接触过，部分人有过艾滋病高危行为，部分人已经出现某些症状，其提问主要是咨询是否患病或检测方面的问题；部分人已确诊为艾滋病，则询问生存期或者具体办法。怀疑感染艾滋病时大多以“不会感染/传染艾滋病了吧”这种方式提问，还有部分用户使用“救命”、“好怕”等词。可以看出这类用户流露出对艾滋病的恐慌与担忧。“治疗”一词出现频率为 20 次，大部分是针对艾滋病治疗药物和方法的提问，还有少数用户对艾滋病能否治好进行提问。此外有一部分提问围绕艾滋病患者的日常生活、职业、生育等方面。关于艾滋病的阻断药相关问题有 24 个，以询问阻断药购买地点为主。这一类用户是有过艾滋病高危行为的用户。发生艾滋病高危行为后，在 72 小时内服用艾滋病阻断药可以阻拦艾滋病病毒<sup>[15]</sup>。

5.3 相关信息传播建议

艾滋病作为一种危害性极大的传染病，受到社会广泛关注，健康信息需求的用户群体非常广泛。关于艾滋病的健康宣传教育不仅要针对艾滋病感染者、疑似患者以及高危人群，也要面向广大群众，而不同人群对艾滋病健康信息需求有所差别，要根据不同人群特点开发针对性和可接受性强的宣传材料<sup>[2]</sup>。本文基于文本挖掘分析各类人群信息需求特征，为不同人群提供健康信息。针对公众，在全面普及艾滋病危害、传播、症状等基本知识的基础上，应着重宣传艾滋病预防知识。艾滋病疑似患者潜藏于普通大众中，由于社会歧视，这部分人群不敢轻易暴露，其信息需求较难发现。因此在为公众提供健康信息时应考虑这部分人群需求，提供艾滋病检测方面的信息，如检测机构、方法等。本研究还发现由于大众对艾滋病的恐惧，与艾滋病患者有接触史者通常会担心感染艾滋病而发帖求助，因此应提供相关信息以消除这类人群的恐惧心理。针对高危人群除进行防护措施的科普外，还应加强关于艾滋病阻断药的宣传。由于阻断药的采购和发放受到国家严格管控，很多人不知道应用阻断药或如何购买，可能错失防治艾滋病的关键时期。艾滋病患者信息需求特征比较明显，主要关注生存期与治疗，其次关注日常生活、职业与生育方面。对于艾滋病患者，国家采取最大的隐私保护及救治措施，患者接受治疗的同时能接受相关健康教育。但仍有部分患者由于某些原因而未能接受健康教育，对相关健康信息特别是日常生活、职业与生育等方面的信息需求迫切但因害怕泄露隐私而不敢公开求助。因此关于艾滋病生存期与治疗、日常生活、职业、生育等方面的健康信息应面向公众传播。

6 结语

本文基于 Python 语言对百度知道中关于艾滋病的提问进行文本挖掘，通过词频和共词网络分析揭示用户对艾滋病健康信息需求的主题分布，在此基础上进一步分析用户特征，提出艾滋病健康信息传

播对策建议,为精准推送信息提供参考。无论是艾滋病患者、高危人群或是普通大众都存在对艾滋病健康信息的需求,信息传播内容应全面充分,传播对象应覆盖所有人群,且健康信息内容根据人群的不同而有所侧重。本研究还存在一定局限性,研究样本只是针对百度知道中关于艾滋病的提问进行研究,较为有限。此外艾滋病患者情感信息需求展现不够充分,只揭示艾滋病疑似患者以及少数确诊患者恐慌与无助的心理。而情感支持对于艾滋病患者非常重要,应挖掘艾滋病患者情感需求。另外还应调查用户对艾滋病健康信息的需求是否得到满足,为更好地开展艾滋病健康信息传播提供依据。

## 参考文献

- 1 国家卫生健康委员会. 2018 年我国卫生健康事业发展统计公报 [EB/OL]. [2019 - 12 - 15]. [http://www.nhc.gov.cn/guihuaxxs/s10748/201905/9b8d52727cf346049de8acce25ffcb0.shtml?tdsourcetag=s\\_pcqq\\_aiomsg](http://www.nhc.gov.cn/guihuaxxs/s10748/201905/9b8d52727cf346049de8acce25ffcb0.shtml?tdsourcetag=s_pcqq_aiomsg).
- 2 国家卫生健康委员会. 关于印发遏制艾滋病传播实施方案(2019—2022 年)的通知 [EB/OL]. [2019 - 12 - 15]. <http://www.nhc.gov.cn/jkj/s7925/201910/adc374d0613144b2b7bb5d6c58a60223.shtml>.
- 3 赵钊,孙伟新. 网络问答社区中健康信息使用情况调查研究 [J]. 医学信息学杂志, 2016, 37 (5): 14 - 19.
- 4 赵海平,邓胜利. 基于社会化问答平台的用户健康信息行为研究综述 [J]. 信息资源管理学报, 2016, 6 (4): 19 - 27.
- 5 Park M S, Park H. Topical Network of Breast Cancer Information in a Korean American Online Community: a semantic network analysis [J]. Information Research, 2016, 21 (4): 729 - 754.

- 6 Ji Z, Zhang Y, Xu J, et al. Comparing Cancer Information Needs for Consumers in the US and China [J]. Stud Health Technol Inform, 2017 (245): 126 - 130.
- 7 Alasmari A, Zhou L. What Health Information Are Consumers Seeking? A Comparison between Two Types of Online Q&A Sites [C]. Orlando: 11th International Conference on Social Computing and Social Media, 2019.
- 8 邓胜利,刘瑾. 基于文本挖掘的问答社区健康信息行为研究——以“百度知道”为例 [J]. 信息资源管理学报, 2016, 6 (3): 25 - 33.
- 9 赵安琪,赵海平,路培鑫. 基于社会化问答社区的抑郁症健康信息需求研究 [J]. 中华医学图书情报杂志, 2018, 27 (9): 38 - 45.
- 10 曹树金,闫欣阳. 社会化问答网站用户健康信息需求的演变研究——以糖尿病为例 [J]. 现代情报, 2019, 39 (6): 3 - 15.
- 11 Zhou X, Ni Y, Xie G, et al. Analysis of the Health Information Needs of Diabetics in China [J]. Stud Health Technol Inform, 2019 (264): 487 - 491.
- 12 王国华,刘菊,杨腾飞,等. 网络空间中艾滋病的社会支持研究——以百度贴吧“HIV 吧”为例 [J]. 情报杂志, 2015, 34 (11): 105 - 110.
- 13 百度. 百度知道 [EB/OL]. [2019 - 12 - 09]. <https://baike.baidu.com/item/%E7%99%BE%E5%BA%A6%E7%9F%A5%E9%81%93/92140?fr=aladdin>.
- 14 百度. 百度百科: 高危性行为 [EB/OL]. [2019 - 12 - 24]. <https://baike.baidu.com/item/%E9%AB%98%E5%8D%B1%E6%80%A7%E8%A1%8C%E4%B8%BA/5524694?fr=aladdin>.
- 15 金剑. 艾滋病高危后 72 小时内可服药阻断 [J]. 江苏卫生保健, 2019 (1): 26.

## 敬告作者

《医学信息学杂志》网站现已开通,投稿作者请登录期刊网站: <http://www.yxxxx.ac.cn>, 在线注册并投稿。

《医学信息学杂志》编辑部