情报科学
第37卷 第8期 2019年8月

基于用户信息行为的微信健康信息关注度研究

商丽丽，王　涛

（中国人民大学 信息学院，北京 100872）

摘　要：【目的/意义】旨在自动分析用户对微信健康信息的关注度。【方法/过程】构建了微信健康信息关注度分析模型，以丁香医生、丁香家庭健康和脉脉养生公众号推送的健康信息作为数据来源，识别微信健康信息的类别和主题分布，基于对用户信息行为的分析来评估用户对健康信息的关注度。【结果/结论】微信公众平台推送的健康信息主要有12类；通过分析用户对微信健康信息关注度发现，健康风险、饮食、药物、身体活动和癌症主题受关注程度较高；用户对各主题健康信息的关注程度与微信公众平台的信息供应分布并不一致。本研究提出的用户关注度自动分析模型具备可以移植性，是对传统关注度研究方法的有效补充。

关键词：信息行为；健康信息；群体关注度；社交媒体

中图分类号：G203；G206　　**DOI：**10.13833/j.issn.1007-7634.2019.08.022

Research on Collective Attention to Health Information on WeChat through Analyzing Information Behaviors of Users

SHANG Li-li, WANG Tao

(School of Information, Renmin University of China, Beijing 100872, *China)*

Abstract:【Purpose/significance】This paper aims to analyze users' attention to WeChat health information automatically. 【Method/process】In this paper, the automated analysis model of collective attention to health information on WeChat was developed. We collected the health information pushed by the public accounts of "dingxiangyisheng", "baojiandaifu" and "mmaijiu" as experimental data. The topic distribution of WeChat health information was mined and identified, and the collective attention to health information was evaluated through the analysis of users' information behaviors. 【Result/conclusion】WeChat public platform mainly pushes 12 types of health information. By analyzing users' attention to WeChat health information, it is found that health risks, diet, drugs, physical activities, cancer are highly concerned. Users' attention to health information on each subject is not consistent with the release distribution of WeChat public platform. The automated analysis model of collective attention proposed in this study is portable and it is an effective supplement to the traditional research methods of user concern.

Keywords: information behavior; health information; collective attention; social media

1 引　言

随着移动互联网的快速发展，健康信息的获取方式呈现出多样化和便捷化。用户除了根据自身需求主动搜索健康信息外，还可以通过订阅或关注的方式自动接收社交网络平台推送的健康信息，其中最典型的应用是微信公众号。作为微信的重要功能，微信公众号允许关注者接收、阅读和分享各种主题的健康信息。截至2017年底，微信用户数已达9.02亿，微信公众号数量已超过1000万，其中活跃账号350万，公众号已成为用户在微信平台上使用的主要功能之一[1]。微信公众号将健康信息以文本、图片和音视频等组织形式传递至用户端，同时微信具备人际传播、群体传播、大众传播及混合传播等相互结合所形成的多样化传播方式[2]，深刻改变了健康信息在网络环境中的传播和接收的方式，延伸了健康媒体的传播触角。现阶段，微信用户已普遍使用公

收稿日期：2019-05-05

基金项目：中国博士后科学基金第60批面上项目"数据驱动的医养结合为老服务模式研究"（2016M601201）

作者简介：商丽丽（1986-），女，山东滨州人，在读博士研究生，主要从事用户信息行为与文本挖掘研究；通讯作者：王　涛.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

众号获取健康信息[3]，具有专业认证资质的权威健康微信公众号得到了较广泛的认同[4]，而对于用户尤其患有慢性病的用户而言，他们对不同主题的健康信息的关注和需求程度存在差异[5]。

近年来，相关学者陆续对微信健康信息展开研究。侯筱蓉等[6]探索了用户对微信健康信息的感知（对健康信息真实性的主观判断）和效用（对主观判断为真的健康信息作出的反应），证实了微信是健康信息传播的有效平台，而多数调查对象缺乏有效识别健康信息真伪的能力；张克永等[7]以信息传播理论为指导，运用访谈法和层次分析法，确立了健康微信公众平台信息质量评价指标体系；金燕等[8]应用扎根理论的方法，通过开放式访谈研究健康类微信公众号关注度的影响因素，发现公众号质量、用户和环境是影响健康类微信公众号关注度的主要因素。而针对健康关注度的研究，戴龙等[9]以入户问卷调查的方式，采用分层多阶段整群随机抽样的方法分析厦门市居民对健康知识关注情况，发现居民最关注的健康知识依次为饮食卫生、慢性病、传染病和心理疾病；单婵娟等[10]通过对社区卫生服务中心问卷调查的方式，采用分层随机抽样的方法分析上海市社区人群对慢性肾脏病（CKD）的关注度及相关影响因素，结果表明社区人群对CKD的关注主要集中在治疗手段、预期寿命和身体症状等几个方面。

基于上述文献回顾发现，对于微信健康信息研究主要围绕微信健康信息真实性、公众号质量评价以及公众号的关注度，缺乏对微信健康信息关注度的研究。而现有关于健康关注度的研究主要采用问卷调查的方式，存在样本量小、主观性强的问题。而针对微信公众平台上不同主题特征的健康信息，自动化地分析用户群体对健康信息的关注度是有必要的。因此，本研究构建了微信健康信息关注度分析模型，以微信健康公众平台健康信息作为数据来源，挖掘微信健康信息的类别，自动识别健康信息的主题分布，基于对用户信息行为的分析来评估用户对健康信息的关注度。了解用户对健康信息的关注情况，有助于微信公众平台的实践者发现用户对健康内容的偏好，继而提供满足用户需求的健康信息服务，同时为微信平台的运营和发展提供启示。

2 相关研究

2.1 社会网络群体关注度

群体关注度（collective attention）是信息时代决策和观点传播的核心问题。社会网络用户群体关注度是社会网络中某些用户群体的共同兴趣偏好，或者特定时段内这些用户所关注的对象及其关注度变化的情况[11-12]。以往的研究从不同方面考察了影响群体关注度的因素，Frese等[13]认为媒介的组织结构、信息投放位置和内容叙述方式都影响着公共领域的群体关注度；Moussaid等[14]认为群体关信息注度的高低取决于信息发布的频率，再现了信息关注度与所涌现的信息数量的对数正态分布关系；阳德青等[12]构建模型以预测新生成的信息或新生事件的群体关注度演化趋势，证实了信息发布时间会影响群体关注度，而关注度随着时间呈现幂律下降趋势。本文认为，在社会网络中内容生成者针对某个主题发布信息，目标用户的浏览、阅读、评论、点赞、分享等信息行为在不同程度上体现了群体对信息的关注度，关注度的高低取决于信息本身的质量和信息主题是否满足用户需求。

2.2 用户行为分析

信息时代网络上产生了大量关于人类活动的数据，基于网络的用户行为分析是指运用多学科知识研究和分析网络用户的构成、特点及其在网络应用过程中行为活动上所表现出来的规律[15]。用户在社交网络平台上的活动记录也呈现出了规律性，例如，用户的转发、评价和点赞行为与新浪微博发布数量的关系近似满足幂律分布[16]；男性更愿意参与新浪微博热点事件，带有图片的内容或权威人士发布的信息更容易被转发和评论[17]；点赞行为侧面体现了Facebook用户的个性差异[18]。解析这些规律也可以识别出目前提供的信息服务可能存在的问题，并为进一步修正或重新制定服务策略提供依据。例如，齐超等[19]从微博用户、用户关系和发布内容三个方面提取特征，结合逻辑回归的方法构建了用户转发行为模型，可以预测转发用户的规模和信息传播的深度，该模型能够在热点发现和舆情监控中发挥作用。

2.3 基于用户信息行为分析的健康信息关注度

作为一种特定类型的用户行为，用户信息行为是指主体为了满足某一特定的信息需求，表现出的信息获取、查询、交流、传播、吸收、加工和利用的一系列过程[20]。用户在社交网络平台上的信息行为反映了自身的态度和意愿，可以通过分析这些行为数据来评估用户对信息的关注度，衡量信息传播的效果。本研究主要关注阅读和点赞两类信息行为：阅读性信息行为和点赞性信息行为。在本文的研究情境中，阅读性信息行为在用户点击健康类文章标题时产生，除了因突发事件导致的暴增式点击，用户的日常阅读行为主要体现了他们对特定健康主题的倾向性[21]。点赞性信息行为在用户阅读健康信息并主动确认后产生，与在朋友圈中由自我展示、主观规范等驱动的点赞举动不同，微信公众平台的点赞行为主要反映了用户对健康内容的认可、偏好和持续关注[22]。因此，用户的阅读行为和点赞行为共同反映了他们对特定健康信息的关注度。

健康信息涵盖了多个主题，用户群体对不同主题健康信息的关注程度也存在差异。以往的研究对特定疾病信息进行了人工编码分类，例如，金碧漪等[5]在文献资料的基础上对糖尿病健康信息进行类别界定，明确了糖尿病的病因及病理知识、诊断和检查、治疗、疾病管理、并发症、社会生活、疾病预防、教育和研究八个主题信息。有关信息需求测度的研究表明，在线问答社区的用户对癌症信息的需求主要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、

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

情感生活和教育研究）方面[23]。不同于以往文献偏重对特定领域的健康信息进行人工主题分类，本研究主要面向综合性健康信息，并且参考了世界卫生组织的健康信息主题界定标准[24]将用户关注内容与健康主题进行自动匹配。

3 研究方法

为了评估用户群体对微信健康信息的关注度，本文构建了微信健康信息关注度分析模型，主要包括数据收集和清理、文本分词和过滤、健康信息主题分析以及关注度评分四个部分，分析模型如图1所示。具体过程是：对从微信公众平台采集到的健康信息清洗以形成标准的实验数据集，针对数据集中的文本内容进行分词、过滤，对分词数据进行文本聚类和主题自动匹配来确定健康信息的主题分布，通过分析用户的阅读和点赞行为对健康信息关注度评分。

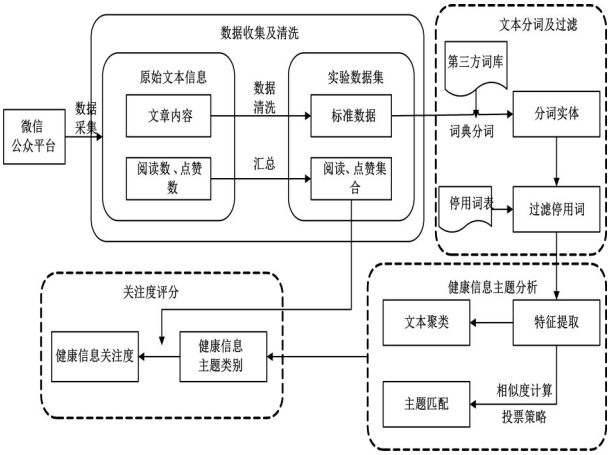

图1 微信健康信息关注度分析模型

3.1 数据收集及清洗

本研究开发了爬虫工具自动采集微信公众平台的健康信息。对用户群体健康信息关注度进行评估，需要选取经过专业资质认证的、影响力较广泛的公众号发布的健康信息作为样本。微信健康类公众号数量较多，活跃情况也各不相同，我们以清博指数[25]上2018年8月排名前三位的健康公众号丁香医生、丁香家庭健康和脉脉养生为研究对象。清博指数是清博大数据下的媒体排名指数，该指数从“整体传播力”、“篇均传播力”、“头条传播力”和“峰值传播力”四个维度来评估微信公众号的影响力。同时，以上三个公众号通过医疗健康组织进行运营，由掌握健康知识和健康教育技能的专业人员制作与维护大众健康相关的内容，赢得了较好的口碑。因此，以上三个公众号发布的健康信息作为样本是具有代表性的。

数据清洗的目的是提高数据质量，主要操作包括字段分割、验证及改正、数据标准化等[26]。我们以标签定位的方式分割字段，提取三个公众号的历史文章标题和正文信息，移除重复以及非健康相关的内容后格式化为成标准数据，并将点赞数和阅读数等结构化数据存入数据库形成实验数据集。经过数据清洗之后，我们将丁香医生、丁香家庭健康和脉脉养生自创建至2018年9月2日的5756篇文章（丁香医生2585篇，丁香家庭健康1968篇，脉脉养生1203篇）以及对应点赞数、阅读数纳入研究数据集。

3.2 文本分词及过滤

为实现对数据集中的文本信息进行自动分析，需对清洗后的文章内容进行分词。中文分词方法主要有词典分词、语义分词以及理解分词等。其中，词典分词是一种统计分词方法，可达到自动从语料库中学习词汇和词的结构并匹配切分的目的[27]。针对微信健康公众号内容专业性较强、词汇相对集中的特点，我们选取词典分词方法。同时，为了消除词典分词因词典不完善而影响分词准确性的影响，本研究在采用中文系统分词框架jieba[28]（基于词典分词的开源工具）进行分词的基础上，引入搜狗细胞词库[29]，使用传统医学词汇、疾病预防工作等专业词汇扩充词典，提高分词准确性。此外，为避免停用词对健康主题分析产生干扰，我们还使用停用词表对停用词进行过滤，最终形成有效分词数据。

3.3 健康信息主题分析

对健康信息主题的分析包括文本聚类和主题匹配两个阶段。为了获得微信健康信息的类别，我们首先对以上分词数据进行聚类。在文本聚类阶段，我们采用TF-IDF[30]方法对文本内容进行特征向量化，统计文本中各词的出现频率，获取健康类文章的特征关键词；之后对提取出的特征关键词进行聚类。由于微信健康信息类别数目是不确定的，因此我们选择K-Means[31]聚类方法。同时，为了能快速收敛以达到较好的聚类效果，本实验依据肘部法则[32]计算最优K值。

为了识别聚类后的健康信息所属的主题，需要进行类别-主题匹配。在主题匹配阶段，我们采用Skip-Gram模型[33]计算特征关键词与世界卫生组织界定的健康主题[24]的相似度，采用投票策略（所有关键词对主题词进行投票，得票最多的主题词即作为该类健康信息的主题）获得特定类别健康信息的主题。

微信健康信息类别与健康主题匹配过程如下：

（1）分词数据作为语料库训练Skip-Gram模型；

（2）针对第i类健康信息，计算该类健康信息的特征关键词（keyword）与健康主题词（subject term）的相似度，形成第i类健康信息与健康主题词的相似度矩阵 $\begin{pmatrix} sim_i(k_1,st_1) & \cdots & sim_i(k_1,st_n) \\ \vdots & \ddots & \vdots \\ sim_i(k_m,st_1) & \cdots & sim_i(k_m,st_n) \end{pmatrix}$；

（3）针对第i类健康信息中第j个关键词，扫描矩阵第j行，找相似度最大值 $Max(sim_i(k_j,st_t))$，并为与该关键词相似度最大的主题词投票；

（4）重复步骤（3），直到计算出第i类健康信息中所有m个关键词的投票；

（5）找到该类健康信息得票最高的主题 $subject_s$（若出现

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表1 健康信息聚类结果

类别	关键词
第1类	运动 动作 按摩 肌肉 身体 穴位 分钟 膝盖 气血 锻炼 经络疼痛 人体 关节 阳气
第2类	孩子 宝宝 妈妈 课程 家长 父母 直播 老师 儿童 家庭 用药教育 护理 咳嗽 时间
第3类	糖尿病 血糖 控制 胰岛素 患者 运动 饮食 主食 治疗 糖尿病食物 低血糖 餐后 并发症 监测
第4类	癌症 血管 吸烟 戒烟 风险 肿瘤 血栓 肺癌 疾病 研究 增加预防 致癌 发现 患者
第5类	皮肤 产品 皱纹 干燥 脸上 治疗 成分 效果 清洁 表面 油脂洗澡 维生素 作用 毛巾
第6类	食物 营养 维生素 鸡蛋 蔬菜 含量 水果 脂肪 蛋白质 含有 饮食 牛奶 食品 胆固醇 膳食
第7类	睡眠 熬夜 失眠 睡觉 睡前 质量 时间 小时 白天 晚上 身体枕头 影响 工作 规律
第8类	治疗 患者 检查 药物 疾病 医院 症状 甲状腺 手术 服用 疼痛 引起 导致 感染 发现
第9类	朋友 身体 生活 心里 减肥 老人 风险 头发 网络 疾病 危害健康 社会 影响 卫生
第10类	牙齿 刷牙 口腔 牙龈 出血 清洁 口臭 导致 细菌 治疗 敏感医院 疾病 重要 修复
第11类	痛风 尿酸 嘌呤 高尿酸 发作 关节 肾脏 患者 控制 食物 治疗 药物 运动 朋友 尿液
第12类	高血压 血压 降压 控制 患者 降压药 测量 饮食 药物 血管 用药 吃药 中风 升高 运动

表2 第12类健康信息与主题词相似度

	癌症	艾滋病	白内障	…	流感	高血压	健康风险	口腔卫生
高血压	0.349	0.424	0.404	…	0.208	1.000	0.319	0.392
血压	0.238	0.259	0.362	…	0.353	0.625	0.234	0.453
降压	0.243	0.270	0.390	…	0.298	0.513	0.278	0.341
控制	0.315	0.319	0.359	…	0.384	0.445	0.246	0.475
患者	0.268	0.389	0.407	…	0.386	0.439	0.237	0.336
降压药	0.273	0.324	0.390	…	0.346	0.542	0.271	0.401
…	…	…	…	…	…	…	…	…
饮食	0.294	0.280	0.302	…	0.241	0.411	0.246	0.473
运动	0.271	0.248	0.243	…	0.260	0.300	0.225	0.425

最高票数并列情况，由3位领域专家再进行一轮投票），输出健康信息类别–主题对 $<category_i, subject_s>$；

（6）重复步骤（2）（3）（4）（5），直到识别出所有类别的 $<category, subject>$ 对。

3.4 关注度评分

阅读性信息行为和点赞性信息行为共同反映了用户对微信健康信息关注度情况，因此我们通过分析用户的阅读和点赞记录对不同主题健康信息的关注度进行评分。为了保证统计数据的可靠性，我们选取的文章的发布日期与实验日期间隔大于3个月。已有的研究表明，网络文章的生命周期符合对数正态分布，即发布期传播较慢，随着转发的积累进入到爆发期，而在热度过后开始衰减[34]。一般情况下，文章发布后6小时内获得的传播和关注最多，24小时以后迅速衰减[35]。因此，我们有理由假定数据集中的阅读和点赞量已基本保持稳定。此外，在公众号创建初期由于缺乏品牌效应可能会导致阅读和点赞量极小，而某些突发事件又会引发阅读或点赞量飙升，这两种情况都可能会造成分析偏差。因此，我们采用截尾均值方法[36]分别计算用户对每一类健康信息的阅读和点赞均值，见公式（1）。然后，计算用户对每一类健康信息的阅读和点赞量，见公式（2）。最后，以用户对每一类健康主题的阅读和点赞量在所有主题阅读和点赞量中的占比评价关注度，见公式（3）。

$$\overline{X_\alpha} = \frac{X_{(n\alpha+1)} + X_{(n\alpha+2)} + \ldots + X_{(n-n\alpha)}}{n - n\alpha} \quad (1)$$

$$\begin{aligned} RI(i) &= \overline{R_\alpha} * (n-k) \\ LI(i) &= \overline{L_\alpha} * (n-k) \end{aligned} \quad (2)$$

$$A(i) = \frac{RI(i)}{\sum_{j=1}^{m} RI_i(j)} + \frac{LI(i)}{\sum_{j=1}^{m} LI_i(j)} \quad (3)$$

其中，n表示每类主题文章数，α表示截尾系数，α=k/n，k表示极值文章数。$X_{(1)}, X_{(2)}, \cdots X_{(n)}$ 表示分别将阅读数、点赞数按升序排列后的顺序序列。$RI(i)$ 为用户对第i类主题的阅读量，$LI(i)$ 为用户对第i类主题的点赞量，$\overline{R_\alpha}$ 以及 $\overline{L_\alpha}$ 分别表示阅读和点赞量的截尾均值。$A(i)$ 为用户对第i类健康信息的关注度，$\sum_{j=1}^{m} RI(j)$ 为用户对全部主题的阅读量，$\sum_{j=1}^{m} LI(j)$ 为用户对全部主题的点赞量。

4 实验结果分析

4.1 健康信息主题分布

通过聚类分析表明，微信公众平台发布的健康信息主要面向12个类别。此外，对提取出的特征词按权重排序，并选择前15个词作为关键词并转化为向量文档。健康信息聚类结果如表1所示。

对主题识别的结果分析以第12类健康信息为例。在第12类健康信息的15个关键词中，有7个关键词与主题词高血压的相似度最大，5个关键词与主题词药物相似度最大，3

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

个关键词与主题词口腔卫生相似度最大。因此,通过投票策略,我们可以确定该类健康信息所对应的主题是高血压。第12类健康信息关键词与主题词的相似度如表2所示。

对所有类别的微信健康信息进行主题匹配的结果如表3所示。微信公众平台发布的健康信息的主题包括身体活动、儿童发育、糖尿病、癌症、护理、饮食、睡眠、药物、健康风险、口腔卫生、痛风和高血压。按健康主题对健康类文章进行汇总,进一步分析微信公众平台推送文章的主题分布情况,如图2所示。其中,健康风险类文章的数量最多,饮食类次之,睡眠、痛风和口腔卫生主题的文章相对较少。

表3 微信健康信息主题匹配结果

类别	健康主题	文章数量(篇)
第1类	身体活动	624
第2类	儿童发育	267
第3类	糖尿病	185
第4类	癌症	355
第5类	护理	196
第6类	饮食	1007
第7类	睡眠	106
第8类	药物	806
第9类	健康风险	1851
第10类	口腔卫生	92
第11类	痛风	100
第12类	高血压	167

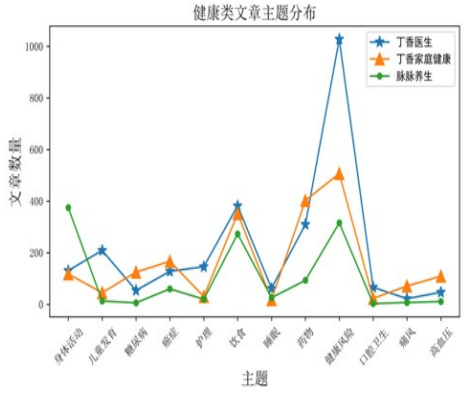

图2 健康类文章主题分布图

4.2 健康信息关注度评分

我们通过分析健康类文章的点赞和阅读量的分布情况来设置截尾系数去除离群数据,阅读和点赞分布如图3所示。结果表明,文章阅读量整天偏高,点赞量基本保持在一万以内。突发热门事件产生了离群点,离群点占比1%,截尾系数设置为1%。

对识别出的12类健康主题的信息关注度评分,评分结果如表4所示。用户对微信健康信息各主题的关注度分布如图4所示。根据评分结果可知,健康风险、饮食、药物、身体活动和癌症主题受关注程度较高。尤其是对健康风险因素的具有特别高的关注,表明微信用户愿意了解致使个人患病或受伤害的属性、特征或风险知识,继而采取预防性健康行为;对饮食和身体活动主题的青睐说明用户已经意识到平衡的饮食和有规律的身体活动是身体健康的基础;而用户对药物信息的获取则会影响他们对药物治疗方案的依从性。以上发现都在一定程度上说明了用户的健康理念发生了变化,从生物医学领域以治疗为主转变为社会方法上的预防为主。

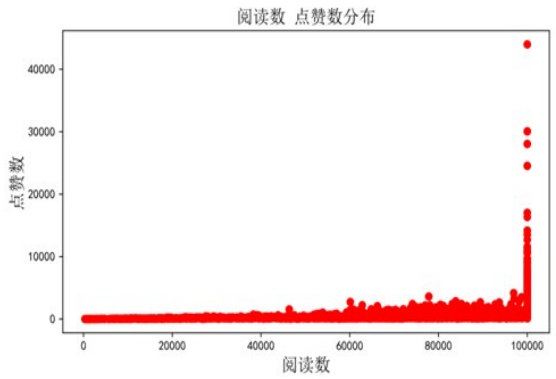

图3 文章阅读和点赞分布

表4 微信健康信息关注度评分

主题	评分	主题	评分
身体活动	0.169	睡眠	0.049
儿童发育	0.064	药物	0.232
糖尿病	0.052	健康风险	0.701
癌症	0.141	口腔卫生	0.028
护理	0.073	痛风	0.026
饮食	0.410	高血压	0.049

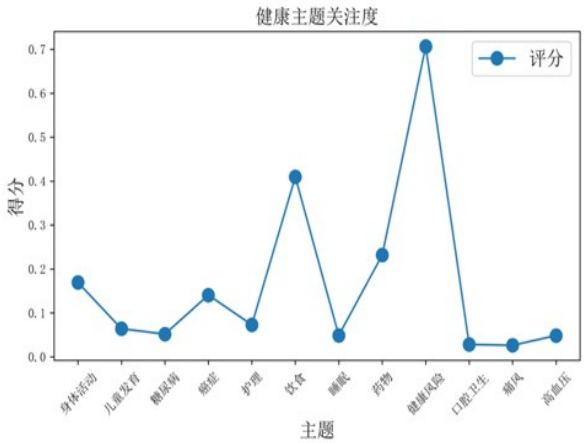

图4 微信健康信息关注度

本研究进一步评估了微信健康信息主题分布和用户对健康信息关注度分布的差异。由于微信健康类文章数目和关注度评分在不同的数值域内,我们对关注度得分值和健康文章数量进行处理,在对齐变量尺度的同时不会改变数据的性质和相关关系。健康信息主题分布和关注度分布趋势比较如图5所示,分析发现两者的分布趋势不完全一致。结果表明,用户对健康风险和饮食主题的关注程度高于微信公众

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

平台对此类健康信息的供应度；微信公众平台对痛风、高血压、糖尿病、护理、儿童发育主题的信息供应超出了用户关注度；而身体活动、癌症、药物主题的需求和供应有比较好的匹配。

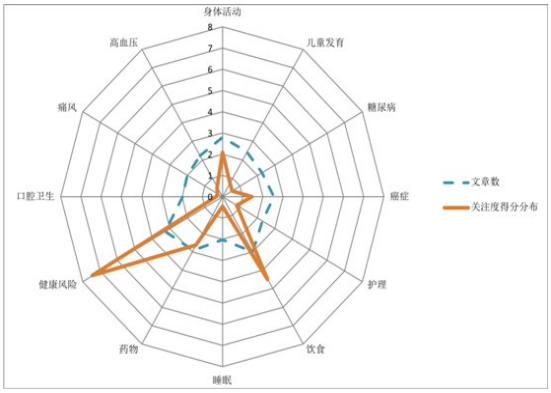

图5 健康信息主题分布与关注分布趋势比较

5 总结和展望

本文构建了微信健康信息关注度分析模型，以丁香医生、丁香家庭健康和脉脉养生公众号推送的健康信息作为数据来源，挖掘微信健康信息的类别，自动识别健康信息的主题分布，并基于对用户信息行为的分析来评估用户对健康信息的关注度。研究发现：第一，微信公众平台推送的健康信息主要有12类，对应的主题为身体活动、儿童发育、糖尿病、癌症、护理、饮食、睡眠、药物、健康风险、口腔卫生、痛风和高血压；第二，用户对健康风险、饮食、药物、身体活动和癌症主题的健康信息关注程度较高；第三，微信公众平台对各主题健康信息的供应量与其关注情况并不一致。其中，用户对健康风险和饮食主题的关注程度高于微信公众平台对此类健康信息的供应度，微信公众平台对痛风、高血压、糖尿病、护理、儿童发育主题的信息供应超出了用户关注度，而身体活动、癌症、药物主题的需求和供应有比较好的匹配。

本研究基于用户信息行为对微信健康信息关注度进行分析，可以为微信公众平台的实践者提供如下启示。第一，了解公众健康状况，有效响应用户对健康信息的需求。目前微信健康信息主要集中在12个主题上，实践者可以进一步了解公众的健康状况，满足用户对更全面健康信息的需求。第二，明确用户的健康观，合理安排预防性和治疗性健康信息的供应。用户对健康风险和饮食信息尤为关注，而平台对疾病信息的供应超过了群体的关注强度，这侧面反映了用户对预防性行为和生活方式的重视。因此，实践者可以根据实际状况合理调整对预防性和治疗性健康信息的发布。第三，推送优质的健康内容，引导用户形成科学、理性的健康理念。微信公众平台应当适当传播健康教育信息和基础的健康知识，同时确保健康信息的论证质量和可信性，促使用户尤其是缺乏健康素养的用户建立科学、理性的健康理念。

本研究还存在局限性。首先，我们只选取了三个公众号的健康类文章作为样本，未来会收集更多的健康信息，构建更全面的语料来提高主题识别的精度。其次，原始数据中的图片信息未纳入到研究当中，未来的研究会把图片解析和文本分析结合起来，探索更有价值的结论。

参考文献

1 中国信息通信研究院. 2017年微信经济社会影响力研究[EB/OL].http://www.199it.com/archives/720219.html,2018-05-04.
2 束芳彬. 微信的传播机制研究[D]. 武汉:华中师范大学, 2015.
3 石文惠,王静雷,李园等. 利用微信开展健康传播的探索[J]. 中国健康教育,2018,(3):326-333.
4 吕亚兰,黄 成,周 虎. 微信平台用户健康信息行为及其影响因素研究[J]. 医学信息学杂志,2018, 39(3): 77-80.
5 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.
6 侯筱蓉,付 扬,陈 娟. 基于微信平台的健康信息用户感知和效用研究[J]. 现代情报,2016, 36(10):89-93.
7 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,(11):143-148.
8 金 燕,张启源. 基于扎根理论的健康类微信公众号关注度影响因素研究[J]. 图书馆理论与实践,2018,(4):54-58.
9 戴 龙,田 丁,骆瑾瑜,等. 厦门市居民健康信息关注程度与获取途径分析[J]. 中国健康教育, 2013,29(4):348-351.
10 单婵娟,龙俊睿,邬碧波,等. 上海市社区人群对慢性肾脏病的关注度及其影响因素[J]. 第二军医大学学报, 2018, (1):37-43.
11 Wu F, Huberman B A. Novelty and Collective Attention [J]. Proc Natl Acad Sci USA, 2007, 104(45):17599-17601.
12 阳德青,肖仰华,汪 卫. 基于统计模型的社会网络群体关注度的分析与预测[J]. 计算机研究与发展, 2010, 47(z1): 378-384.
13 Frese M. Creating Collective Attention in the Public Domain: Human Interest Narratives and the Rescue of Floyd Collins[J]. Social Forces, 2002, 81(1):57-85.
14 Moussaid M, Helbing D, Theraulaz G. An individual-based model of collective attention[J]. arXiv preprint,2017, arXiv:0909.2757, 2009.
15 董富强. 网络用户行为分析研究及其应用[D]. 西安:西安电子科技大学, 2005.
16 Guan W,GaoH,Yang M, et al. Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events[J]. Physica A: Statistical Mechanics and its Applications, 2014, 395(2): 340-351.
17 解 军,邢进生. 基于KNN算法的新浪微博用户行为分析及预测[J]. 山西师范大学学报(自然科学版), 2016,(2):

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

38-45.

18 Ghavami S M, Asadpour M, Hatami J, et al. Facebook user´s like behavior can reveal personality[C]//2015 7th Conference on Information and Knowledge Technology (IKT). IEEE, 2015: 1-3.

19 齐　超,陈鸿昶,于　岩. 基于行为分析的微博信息传播效果[J]. 计算机应用,2014, 34(8):2404-2408.

20 胡昌平. 现代信息管理机制研究[M]. 武汉:武汉大学出版社,2004:124-132.

21 Ratkiewicz J, Flammini A, Menczer F. Traffic in social media I: paths through information networks[C]//2010 IEEE Second International Conference on Social Computing. IEEE, 2010: 452-458.

22 Gan C. Understanding WeChat users' liking behavior: An empirical study in China[J]. Computers in human behavior, 2017,68(1): 30-39.

23 李重阳,翟姗姗,郑　路. 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[J]. 数字图书馆论坛, 2016,(9):34-42.

24 世界卫生组织. 健康主题[EB/OL].https://www.who.int/topics/zh/,2018-12-04.

25 清博指数. 微信传播指数WCI[EB/OL].http://www.gsdata.cn/site/usage/,2018-09-01.

26 郭志懋,周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11):2076-2082.

27 张卫丰,张迎周, 周国强. 中文分词技术综述[C]// 全国web信息系统及其应用学术会议、全国语义web与本体论学术研讨会暨全国电子政务技术与应用学术研讨会. 2008:9-11.

28 Python Software Foundation. Jieba 0.39 [EB/OL].https://pypi.org/project/jieba/,2018-12-04.

29 搜狗. 细胞词库[EB/OL].https://pinyin.sogou.com/dict/, 2018-12-05.

30 Wu H C, Luk R W P, Wong K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. Acm Transactions on Information Systems, 2008, 26(3):55-59.

31 Pena J M, Lozano J A, Larranaga P. An empirical comparison of four initialization methods for the k-means algorithm [J]. Pattern recognition letters, 1999, 20(10): 1027-1040.

32 Ketchen D J, Shook C L. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique[J]. Strategic Management Journal, 1996, 17(6):441-458.

33 熊富林,邓怡豪,唐晓晟. Word2vec的核心架构及其应用[J]. 南京师范大学学报(工程技术版), 2015,(1):43-48.

34 Mac Kay D. An example inference task: clustering[M]. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press, 2003: 284-292.

35 Cha M, Kwak H, Rodriguez P, et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system[C]//Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007: 1-14.

36 Marazzi A, Ruffieux C. The truncated mean of an asymmetric distribution[J]. Computational Statistics & Data Analysis, 1999, 32(1):79-100.

（责任编辑:张连峰）

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net