医患问答社区热点主题分析研究*

王 煜 魏 理 姜顺军

（广东工业大学管理学院 广州 510520） （广州医科大学附属第一医院 广州 510120）

〔摘要〕 介绍在线健康社区主题分析、主题识别方法的相关研究情况，以寻医问药网的糖尿病社区帖子为主要研究对象，通过文本挖掘和文本聚类方法，利用矢量空间模型和 K－Means 聚类模型进行主题分析，识别关键特征词，为在线健康平台提供信息服务建议，提升患者体验和促进健康管理。

〔关键词〕 医患问答社区；主题分析；疾病；文本挖掘；文本聚类

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.11.001

Study and Analysis of Hot Spot Topics of Q&A Community for Doctors and Patients

WANG Yu, *School of Management*, *Guangdong University of Technology*, *Guangzhou 510520*, *China*; *WEI Li*, *JIANG Shunjun*, *The First Affiliated Hospital of Guangzhou Medical University*, *Guangzhou 510120*, *China*

〔**Abstract**〕 The paper introduces study situation related to topic analysis and topic recognition methods of the Online Health Community (OHC). It takes posts in the diabetes community of the XYWY.com as the major study object, implements topic analysis and recognizes key characteristic words by making use of text mining, text clustering, Vector Space Model (VSM) and K－Means clustering model, as well as provides suggestions on information service for online health platform to improve patient experience and health management.

〔**Keywords**〕 Q&A community for doctors and patients; thematic analysis; disease; text mining; text clustering

1 引言

我国相关医疗健康政策的推动和互联网医疗的发展使在线健康的理念深入人心，公众对互联网健康服务的参与度也不断提高。2016 年 10 月政府文件《“健康中国 2030”规划纲要》提出重点发展基于互联网的健康服务，促进个性化健康管理服务发展；2017 年 2 月《“十三五”全国人口健康信息化发展规划》提出构建“互联网＋健康医疗”服务新模式。在我国约 56% 的 60 岁以上老年人具有通过互联网咨询慢病知识的经历。这些数据充分说明越来越多的病患由于自身医疗知识有限，更趋向于通过在线健康社区（Online Health Community，OHC）获取相关疾病的诊疗知识。患者既是信息提供者又是消费者[1]，从被动的信息接收者转变成主动的信息生产者。这种转变极大地激发患者在线参与社交活动的积极性，生成大量的在线健康数据，与此同时社区用户（包括患者、医生和第 3 方服务机构等）可以获取更好的疾病管理知识[2－4]。因此分析在线健康社区的数据资源已成为大数据时代的重要

〔收稿日期〕 2018－06－05

〔作者简介〕 王煜，硕士研究生；通讯作者：魏理，主任药师，发表论文 16 篇。

〔基金项目〕 广州市哲学社科资助项目“信息网络技术驱动广州制造业转型”（项目编号：502170111）；广州市科技计划资助项目“基于临床大数据的呼吸系统急危重症风险预警模型及药物治疗评价系统研究”（项目编号：201803010063）。

研究问题，充分利用在线健康社区的数据价值能够有效地提高疾病的诊疗效率。

Moorhead S A 等在综述中指出社交媒体在健康交流方面的 7 大主要用途，其中包括加强与他人的互动，促进分享和获取健康信息[3]。而较早的研究指出用户参与在线健康社区主要是交流医疗健康知识和获得情感支持，如寻求和分享个人健康保健和疾病诊疗经验、对健康主题提出个人观点、抒发个人情感和寻求他人支持等[4]。有数据表明互联网用户中 80% 在网上查看关于特定疾病或治疗等健康话题的信息，其中 34% 的人浏览网站或博客的在线健康评论或经验，24% 的人咨询特定药物或治疗[5]。然而大多在线健康社区并没有清晰的主题组织结构，信息往往杂乱无序。本文选取寻医问药网糖尿病社区的数据，基于医患问答社区对在线健康社区的热点主题进行挖掘、分析，旨在促进在线健康平台提供高效服务，使用户获得切实有用的医疗健康知识。

2 相关研究

2.1 概述

虚拟社区指由具有共同兴趣爱好或目的的用户组成的基于互联网技术的在线集合体，通过信息生产及消费来完成在线协作、知识分享及在线交易等活动[6-7]。在线社区源自于虚拟社区，在线健康社区是虚拟社区在主题与活动上的细化，分为 3 类：包括 PatientsLikeMe、好大夫在线等在内的专业性医疗网站；包括天涯社区健康版块在内的综合性社区网站的医疗频道或子版块；包括病症交流为主的微信群等在内的即时聊天群组[8]。OHC 属于在线网络聚集社区，是用户通过互联网对健康相关问题进行知识无偿分享、专家一对一咨询和成员自由交流等活动的在线社区，它将用户聚集在同一个网络平台，用户有更好的匿名性和相关性，相互之间更能施加有利影响，给患者提供更多的情感鼓励和精神支持。

2.2 热点主题分析[9]

2.2.1 不同 OHC 类型的主题分析 在线社区的类型不同，其健康主题的特征分布也不尽相同。一些学者对综合类网站从主题词分类角度对其相关主题进行分类，进而对热点关键词进行抽取和分析，旨在提高互联网医疗的服务水平。Bowler L 等以 Yahoo！Answers 中有关饮食失调的问题为例，通过分析词频、词性和情感等方式总结出相关主题[10]。吕英杰将在线健康社区 Medhelp 的肺癌、乳腺癌和糖尿病 3 种典型疾病作为研究对象，使用文本聚类的方法分析发现 7 个健康热点主题[11]。金碧漪等以问答社区网站 Yahoo！Answers 和糖尿病社区 Diabetic Connect 为数据来源，借助数据编码和文本处理的方法得出在线健康社区有关糖尿病的信息主题分布特征[12]。Lu Y 等同样以在线社区 MedHelp 上的肺癌、糖尿病和乳腺癌论坛消息作为实验数据，使用文本挖掘技术识别利益相关者[13]。可见国内外对专业的医患问答社区的热点主题研究还较少，对于健康社区热点主题的挖掘一般集中在综合类的网站和情感分享型的社区，相比综合类的在线健康社区，专业的医患问答社区的专注度无疑更高，直接体现患者最关注的问题。本文将专业医疗问答社区寻医问药网作为研究对象，通过对其中糖尿病社区的文本知识进行分析和整理，获取患者普遍关心的知识。

2.2.2 不同疾病的 OHC 主题分析 消费者信息需求因健康问题不同而各不相同，特定疾病的话题代表用户的信息需求[14]。关于较难攻克的疾病癌症，Park H 调查网上社区的相关信息，发现大部分帖子与医疗专题有关，包括治疗、诊断、症状等在内的 9 个子主题[15]。对于宫颈癌，Westbrook L 等在一个在线问答论坛上发现病因、预防、症状、诊断、预后和治疗以及缓解和生命结束等 8 个主题[16]。对于常见的慢病糖尿病，Zhang J 等确定包括营养、诊断和检查、体征和症状、教育和信息资源等在内的 12 个主题[17]。关于慢病高发症状高血压，邓胜利等利用文本挖掘软件对百度知道的相关提问进行分词和标准化处理，证明用户最为关注高血压日常疾病管理、疾病确诊和治疗[18]。对于肥胖，Liang B 等从肥胖支持组中的手术、药物和自我支持线索中得出包括商业减肥计划、体重减轻、健康和医疗问题、社交焦虑在内的 11 个主题[19]。针对不同病症的研究层出不穷，

但作为慢病中最常见的一种代谢性疾病，针对糖尿病的主题研究仍是热门，尤其是医患问答社区中的糖尿病研究，其对患者的慢病管理意义重大。

2.3 主题识别方法

2.3.1 统计调查方式 早期大部分关于医疗信息资源使用情况以及疾病关注程度的研究是通过问卷调查、访谈和实验研究等方式进行的[20-21]。Basch EM 等针对门诊和癌症中心患者和同伴的调查问卷中发现通过印刷品和互联网寻求的主题相似，印刷产品是癌症患者最常见的信息来源[22]。Buis LR 对来自 8 个在线社区的 3 717 个职位进行定量内容分析，重点关注具有高和低 5 年相对生存率的癌症，发现高存活率社区包含情感支持内容比例更高，低生存率社区包含信息支持内容的比例更高[23]。以上这些基于调查问卷和访谈的方法虽以网络社区中的真实帖子文本为研究对象，但依靠大量的人工注释效率低下，难以应对庞大用户群体产生的海量数据分析，经常受到样本数量限制、问卷和访谈对象不精准等因素的影响，很难在整体上客观、全面、真实地反映在线健康社区中的热点主题。因此基于文本挖掘的关键词自动抽取技术得以广泛应用。

2.3.2 基于文本挖掘的主题识别方法 在线社区的主题识别方法主要有 4 种：基于主题模型[24]、基于关键词统计[25]、基于文本聚类[26]和基于网络图的方法[27]。主题模型是语义挖掘的核心，用来发现文档中的抽象主题，是处理非结构化数据的常用方法，其主要功能是从文本数据中提取潜在的主题信息，主要思想[28]为：文档是若干主题的混合分布；每个主题又是词语的概率分布。Blei DM 等提出的隐含狄利克雷分配（Latent Dinchlet Allocation，LDA）模型是目前最主流的主题模型[29]，它是在 Thomas Hofmann 提出的概率性潜在语义索引（Probabilistic Latent Semantic Indexing，PLSI）[30]基础上加入贝叶斯框架改进的，更具一般化特征。基于关键词统计的方法，统计文中候选词出现的频次和位置，对这两类特征值加权求和并排序，选取权值最高的 N 个词作为关键词。如 TFIDF 算法[31]用于评估某个词对某个文档集或语料库中的某份文档的重要程度。郭红梅等[27]认为基于网络图的方法是依据图中结点和边的属性识别图中核心的术语或关联子团，包括中心度方法、紧密关联子图查找和图聚类 3 种。

很多研究将以文本挖掘技术为代表的智能化处理手段应用于在线健康社区的文本处理中，从中有效获取有价值的信息，重点关注用户的角色行为、知识共享、忠诚度分析和情感分析等方面[32-34]。有学者针对不同领域对主题分析方法进行研究。如 Do N 等采用基于本体的特定领域建模方法，提出 Keyphrase 和匹配签名算法[35]。HaCohen-Kerner Y 等则不限定领域使用随机森林法提出自动提取关键短语以供检索、分类和聚类的一般方法[36]。在基于文本聚类的主题识别研究中，基于划分的 K-Means 算法[37-39]应用最为广泛。本文创新性地采用 K-Means 算法对医患问答社区的糖尿病帖子进行主题分析研究，旨在探索特定在线社区的主题分布规律，提供慢病管理和信息服务等方面建议。

3 研究方案

3.1 概述

本文以寻医问药网（http://www.xywy.com/）的糖尿病帖子为研究对象，运用文本挖掘的研究方法，通过文本分词得出特征集，以特征词的权重将文本表示为矢量空间模型（Vector Space Model，VSM）[40]，通过 K-Means 算法对文档集进行聚类操作，得出两个结果簇，对每个结果簇的特征词进行分析，基于频繁值对聚类结果进行表示，最后提取主题，形成医患问答社区的热点主题，其主题识别框架，见图 1。

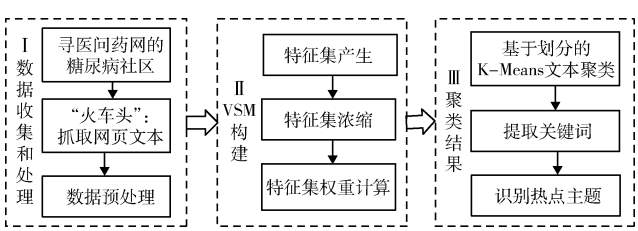

图 1 医患问答社区的热点主题识别框架

3.2 数据收集和处理

3.2.1 数据收集 本文通过火车头数据采集器获取寻医问药网糖尿病社区 2016 年 6 – 12 月末的帖子数据，共采集到 3 662 条包括标题、问题和回答在内的患者提问数据。

3.2.2 数据预处理 将文本数据表示成 VSM 之前，要对数据进行预处理，包括句子切分和停用词消除。(1) 句子切分。句子是包含完整语义的最小单位，而每个文本信息通常由多个句子组成，需要将文档切分成若干个句子。由于对精度要求不是很高，本文使用换行号、句号、问号和感叹号来识别句子是否结束，示例的切分结果，见表 1。(2) 停用词消除。旨在减少文档中出现过于频繁而对检索没有区分意义的词，如冠词、介词和连词等，以压缩存储空间，可采用停用词表和词频统计的处理方式，本文采用前者，去除“可以、吗、4、7、许多、有、用、看、这个、你、的、比如”等停用词，得到“停用词清除表”。

表 1 示例的句子切分

原帖标题	原帖问题	原帖回答	句子切分后文本
尿蛋白高一个加号可以服用盐酸二甲双胍吗?	47 女，糖尿病 7 年，前阵子腿压迫神经吃了许多治腿疼得药，导致糖没控制好，尿蛋白有个加号，空腹血糖 11，可以用盐酸二甲双胍吗？感觉吃完有点冒虚汗，心脏不太舒服，看说明说肾病患者不宜服用这个药品。补充提问：症状：视力下降，心慌，乏力	指导意见：有肾病不可以服用是指肾功能不全的情况下不可以服用。目前你有蛋白尿，需要加用减少蛋白尿的药物，比如金水宝。还需要积极地控制好血糖，减少其对肾脏的损伤。注意如果服用，二甲双胍有明显的不适，可以换成拜糖平等其他降糖药	(1) 尿蛋白高一个加号可以服用盐酸二甲双胍吗 (2) 47 女，糖尿病 7 年，前阵子腿压迫神经吃了许多治腿疼得药，导致糖没控制好，尿蛋白有个加号，空腹血糖 11，可以用盐酸二甲双胍吗 (3) 感觉吃完有点冒虚汗，心脏不太舒服，看说明说肾病患者不宜服用这个药品 (4) 补充提问：症状：视力下降，心慌，乏力 (5) 指导意见：有肾病不可以服用是指肾功能不全的情况下不可以服用 (6) 目前你有蛋白尿，需要加用减少蛋白尿的药物，比如金水宝 (7) 还需要积极地控制好血糖，减少其对肾脏的损伤 (8) 注意如果服用，二甲双胍有明显的不适，可以换成拜糖平等其他降糖药

3.3 研究过程及结果

3.3.1 VSM 构建 经过数据收集和预处理后，需要通过构建 VSM 将已获取的文本数据转化成结构化数据，其构建过程包括特征集的产生、缩减和权重计算。(1) 特征集的产生。通过张华平博士开发的 NLPIR 汉语分词系统能准确统计所分词语的频数，且可通过较大的样本数据提高分词精确度，得到的分词和其统计结果，见表 2。(2) 特征集的缩减。通过信息增益算法对3 662个文档特征集进行缩减，选取排名前 3 的特征词，大大降低文本 VSM 的维度。示例的 3 个代表特征词为服用、肾病和蛋白尿，见表 3。(3) 特征词的权重计算。为将文档表示为空间向量，计算帖子特征词的权重。示例帖子可以用三维空间向量 (0.96, 0.74, 1.24) 来表示，见表 4。特征词“服用”的词频和信息增益比“肾病”、“蛋白尿”高，但对于整个文档集而言，特征词“蛋白尿”的权重反而最高，更具代表性。

表 2 示例的分词统计结果

分词	词性	词频	分词	词性	词频	分词	词性	词频
服用	v	4	导致	v	1	指导意见	n_ newword	1
不	d	4	糖	n	1	一个	m	1
尿	n	2	没	d	1	是	v	1
蛋白	n	2	个	q	1	指	v	1
控制	v	2	加	b	1	肾	n	1
吃	v	2	号	n	1	加号	n	1
好	a	2	空腹血糖	n_ newword	1	全	a	1
盐酸	n	2	感觉	n	1	女	b	1
二	m	2	完	v	1	情况	n	1
甲	m	2	有点	d	1	下	f	1
双	q	2	冒	v	1	目前	t	1
胍	n	2	虚汗	n	1	糖尿病	n	1
腿	n	2	心脏	n	1	年	q	1
肾病	n	2	太	d	1	加	v	1
可以	v	2	舒服	a	1	用	p	1
减少	v	2	说明	v	1	功能	n	1
的	u	2	说	v	1	药物	n	1
蛋白尿	n	2	高	a	1	金水宝	n	1
需要	V	2	患者	n	1	还	d	1
前	f	1	不宜	v	1	积极	a	1
阵子	q	1	药品	n	1	地	u	1
压迫	v	1	补充	v	1	血糖	n	1
神经	n	1	提问	v	1	其	r	1
了	u	1	症状	n	1	对	p	1
治	v	1	视力	n	1	肾脏	n	1
疼	v	1	下降	v	1	损伤	v	1
得	u	1	心慌	a	1	–	–	–
药	n	1	乏力	a	1	–	–	–

表 3 示例的信息增益

特征词	I（Tk）	I（T）	I（T，Tk）	Gain（T，Tk）	词频
服用	0.041	2.1	1.67	0.43	4
肾病	0.027	2.1	1.81	0.29	2
蛋白尿	0.025	2.1	1.84	0.26	2

表 4 示例的特征词权重

特征词	ni	N	TFi，j	IDEi	Wi，j
服用	2 098	3 662	4	0.24	0.96
肾病	1 550	3 662	2	0.37	0.74
蛋白尿	871	3 662	2	0.62	1.24

3.3.2 K－Means 聚类 采用 SPSS Modeler 软件和 NLPIR 汉语分词系统分别进行文本聚类和词频统计，得到聚类结果和热点主题。（1）聚类结果。通过软件得到聚类结果，见图 2。最大和最小聚类大小分别为 2 207 条和 1 455 条，比例各占 60.3% 和 39.7%。（2）热点主题提取。选用基于频繁值的表示方法，通过分词系统对两个结果簇分析，选取最终聚类结果的关键词，见表 5。分析第 1 类和第 2 类主题可知“低血糖，感染，昏迷，口渴，恶心，疲劳，尿频，饥饿，视力下降”等词语为糖尿病的症状描述，“肾脏疾病，肥胖，神经病变，病足”等词语为糖尿病的并发症描述。因此将热点主题定义为“症状与并发症”和“治疗”。

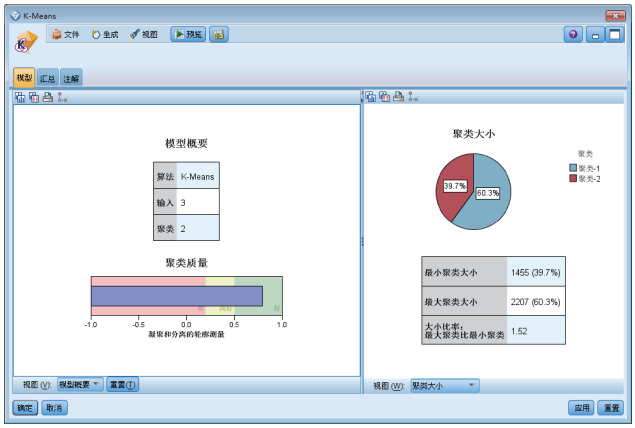

图 2 聚类结果

表 5 关键词提取结果

聚类 ID	关键词	帖子数
1	低血糖，感染，昏迷，肾脏疾病，肥胖，神经病变，病足，口渴，恶心，疲劳，尿频，饥饿，视力下降	1 455
2	胰岛素，注射，治疗，截肢，拜糖平，运动，降糖药，饮食，住院，内分泌科	2 207

4 结语

本文采用文本挖掘的方法提取医患问答社区的热点主题，为在线健康社区提供健康知识搜索方面的服务建议，促进其快速发展。同时为医疗健康知识发现、OHC 的主题分析以及知识图谱构建作出一定的理论贡献，促进健康知识管理及相关理论的发展和延伸。此外识别热点主题利于患者诊疗决策的积极参与以及日常的健康管理。研究取得了一定的成果，但仍有不足之处：其一，本文采用矢量空间模型对文本内容进行结构化转变，采用文本聚类的方法对文本数据进行分类以识别主题，未来更适合于医患问答社区热点主题的提取方法有待研究。其二，本文通过主题的发帖量判断主题热度，进行健康热点主题的识别研究，未来还需要进一步研究其他相关影响因素。

参考文献

1 Fichman R G, Kohli R, Krishnan R. The Role of Information Systems in Healthcare: current research and future trends [J]. Information Systems Research, 2011, 22 (3): 419 – 428.

2 Yan L, Tan Y. Feeling Blue? Go Online: an empirical study of social support among patients [J]. Information Systems Research, 2014, 25 (4): 690 – 709.

3 Moorhead S A, Hazlett D E, Harrison L, et al. A New Dimension of Health Care: systematic review of the uses, benefits, and limitations of social media for health communication [J]. Journal of Medical Internet Research, 2013, 15 (4): e85.

4 Finn J. An Exploration of Helping Processes in an Online Self – help Group Focusing on Issues of Disability [J]. Health and Social Work, 1999, 24 (3): 220 – 231.

5 Fox S. The Social Life of Health Information, 2011 [EB / OL]. [2017 – 06 – 20]. http://www.Pew internet. org/2011/05/12/the – social – life – of health information 2011/.

6 Tsai H T, Bagozzi R P. Contribution Behavior in Virtual Communities: cogniyive, emotional, and social influences [J]. Mis Quarterly, 2014, 38 (1): 143 – 164.

7 周军杰，左美云. 线上线下互动、群体分化与知识共享的关系研究——基于虚拟社区的实证分析 [J]. 中国管理科学，2012，20 (6)：185 – 192.

8 周军杰. 用户在线参与的行为类型——基于在线健康社区的质性分析 [J]. 管理案例研究与评论，2016，9 (2)：173 – 184.

9 Gooden R J, Winefield H R. Breast and Prostrate Cancer Online Discussion Boards. A Thematic Analysis of Gender Differences and Similarities [J]. Journal of Health Psychology, 2007, 12 (1): 103 – 114.

10 Bowler L, Oh J S, He D, et al. Eating Disorder Questions in Yahoo! Answers: information, conversation, or reflection? [J]. Proceedings of the Association for Information Science and Technology, 2012, 49 (1): 1 – 11.

11 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.

12 金碧漪，许鑫. 在线健康社区中的主题特征研究 [J]. 图书情报工作，2015，59 (12)：100 – 105.

13 Lu Y, Wu Y, Liu J, et al. Understanding Health Care Social Media Use From Different Stakeholder Perspectives: a content analysis of an online health community [J]. Journal of Medical Internet Research, 2017, 19 (4): e109.

14 Z Zhao Y, Zhang J. Consumer Health Information Seeking in Social Media: a literature review [J]. Health Information &

Libraries Journal, 2017, 34 (4): 268 - 283.

15 Park H, Park M S. Cancer Information - seeking Behaviors and Information Needs Among Korean Americans in the Online Community [J]. Journal of Community Health, 2014, 39 (2): 213 - 220.

16 Westbrook L, Zhang Y. Questioning Strangers About Critical Medical Decisions:" what happens if you have sex between the HPV shots?" [J]. Information Research, 2015, 20 (2): 1 - 12.

17 Zhang J, Zhao Y, Dimitroff A. A Study on Health Care Consumers' Diabetes Term Usage Across Identified Categories [J]. Aslib Journal of Information Management, 2014, 66 (4): 443 - 463.

18 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.

19 Liang B, Scammon D L. E - word - of - mouth on Health Social Networking Sites: an opportunity for tailored health communication [J]. Journal of Consumer Behaviour, 2011, 10 (6): 322 - 331.

20 Klemm P, Nolan M T. Internet Cancer Support Groups: legal and ethical issues for nurse researchers [J]. Oncology Nursing Forum, 1998, 25 (4): 673 - 676.

21 Schultz P N, Stava C, Beck M L, et al. Internet Message Board Use by Patients with Cancer and Their Families [J]. Clinical Journal of Oncology Nursing, 2002, 7 (7): 663 - 667.

22 Basch E M, Thaler H T, Shi W, et al. Use of Information Resources by Patients with Cancer and Their Companions [J]. Cancer, 2004, 100 (11): 2476 - 2483.

23 Buis L R, Whitten P. Comparison of Social Support Content within Online Communities for Gigh - and Low - survival - rate Cancers [J]. Comput Inform Nurs, 2011, 29 (8): 461 - 467.

24 Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences, 2004, 101 (suppl 1): 5228 - 5235.

25 Qin P, Xu W, Guo J. A Novel Negative Sampling Based on TFIDF for Learning Word Representation [J]. Neurocomputing, 2016, 177 (C): 257 - 265.

26 He X, Ding C H Q, Zha H, et al. Automatic Topic Identification Using Webpage Clustering [C]. San Jose: Proceedings IEEE International Conference on Data Mining, 2001: 195 - 202.

27 郭红梅, 张智雄. 基于图挖掘的文本主题识别方法研究综述 [J]. 中国图书馆学报, 2015, 41 (6): 97 - 108.

28 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述 [J]. 软件学报, 2017, 28 (9): 2431 - 2449.

29 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3 (1): 993 - 1022.

30 Hofmann T. Probabilistic Latent Semantic Indexing [C]. Berkeley: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999: 50 - 57.

31 Salton G, Buckley C. Term - weighting Approaches in Automatic Text Retrieval [J]. Information Processing & Management, 1988, 24 (5): 513 - 523.

32 Bekhuis T, Kreinacke M, Spallek H, et al. Using Natural Language Processing to Enable in - depth Analysis of Clinical Messages Posted to an Internet Mailing List: a feasibility study [J]. Journal of Medical Internet Research, 2011, 13 (4): e98.

33 刘璇, 汪林威, 李嘉, 等. 在线健康社区中用户回帖行为影响机理研究 [J]. 管理科学, 2017, 30 (1): 62 - 72.

34 张克永, 李贺. 网络健康社区知识共享的影响因素研究 [J]. 图书情报工作, 2017, 61 (5): 109 - 116.

35 Do N, Ho L V. Domain - specific Keyphrase Extraction and Near - duplicate Article Detection Based on Ontology [C]. Can Tho: Proceedings of International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for the Future (RIVF), 2015: 123 - 126.

36 HaCohen - Kerner Y, Vrochidis S, Liparas D, et al. Keyphrase Extraction Using Textual and Visual Features [C]. Dublin: Proceedings of the 25th International Conference on Computational Lingnistics, 2014: 121 - 123.

37 张琳, 牟向伟. 基于 Canopy + K - Meanss 的中文文本聚类算法 [EB/OL]. [2018 - 01 - 07]. http://kns.cnki.net/kcms/detail/44.1306.G2.20171206.0358.008.html.

38 刘江华. 一种基于 kmeans 聚类算法和 LDA 主题模型的文本检索方法及有效性验证 [J]. 情报科学, 2017, 35 (2): 16 - 21, 26.

39 刘欣, 佘贤栋, 唐永旺, 等. 基于特征词向量的短文本聚类算法 [J]. 数据采集与处理, 2017, 32 (5): 1052 - 1060.

40 Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18 (11): 613 - 620.