

基于文本挖掘和医患互动论坛的深静脉血栓防治规律研究^{*}

李艳红

欧阳聪聪

(上海财经大学信息管理与工程学院 上海 200433)

(延边大学医学院 延吉 133002)

汪乐纯

欧敬民

(上海财经大学信息管理与工程学院
上海 200433)

(上海交通大学医学院附属新华医院普外科
上海 200092)

〔摘要〕 基于“好大夫在线”、“寻医问药”和“有问必答”的医患互动论坛，利用中文分词、数据清洗过滤、词频统计、语义网络图、可视化等文本挖掘技术或方法，提取和归纳深静脉血栓的症状、病因、常用药、并发症、非药物治疗、护理方法等知识，验证文本挖掘方法在探索疾病防治规律方面的可用性，以及医患互动问答网络数据源用于医学研究的可行性。

〔关键词〕 深静脉血栓；文本挖掘；防治规律；医患互动论坛

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.04.014

Research on the Law of Prevention of Deep Vein Thrombosis (DVT) Based on Text Mining and Doctor - Patient Interaction

Forum LI Yan - hong, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China; OUYANG Cong - cong, Medical College, Yanbian University, Yanji 133002, China; WANG Le - chun, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China; OU Jing - min, Department of General Surgery, Xinhua Hospital, Medical College of Shanghai Jiaotong University, Shanghai 200092, China

〔Abstract〕 Based on the doctor - patient interaction forum of "HaodaiFu Online", "Xunyiwenyao" and "Youwenbida", taking advantage of the text mining technologies and methods such as Chinese word segmentation, data cleaning and filtering, word frequency statistics, semantic network map and visualization, extract and summarize the symptoms, causes, common drugs, complications, non - drug therapy and nursing method of Deep Vein Thrombosis (DVT), the paper verifies the availability of text mining method in exploring the prevention law and method of diseases and the feasibility of application of doctor - patient interaction Q&A network data source in medical research.

〔Keywords〕 Deep Vein Thrombosis (DVT); Text mining; Prevention law; Doctor - patient interaction forum

〔收稿日期〕 2016 - 12 - 08

〔作者简介〕 李艳红，博士，副教授，发表论文 40 余篇；
通讯作者：欧敬民，博士，教授。

〔基金项目〕 国家自然科学基金资助项目（项目编号：81572673）。

1 引言

在生物医学领域，由于生物学数据和文献的急

剧增长,通过数据挖掘寻找规律和新知识成了研究热点和重要分支。文本挖掘属于数据挖掘的一个具体研究领域,其主要任务是对海量数据进行整合、分析,获得更具代表性、可信度高的结果,从文献中挖掘出来试验假设和试验建议,以便生物学家验证得到新的发现,从而提高人们对生物医学现象的认识。

深静脉血栓 (Deep Vein Thrombosis, DVT) 是指血液在深静脉腔内异常凝结,阻塞静脉管腔,导致静脉回流障碍。DVT 为血管外科常见疾病,常引起远端静脉高压、肢体肿胀、疼痛及浅静脉扩张等临床症状,甚至会发生血栓脱落,引起肺栓塞,造成猝死。20 世纪 60 年代,此病在国外开始引起关注,近些年才被国内重视。目前,中医药在 DVT 的治疗中取得了较好的疗效,已积累大量的文献资料,也有不少学者对中医方剂从不同的角度进行数据挖掘研究,取得了一定进展,证明其不仅能指导临床用药,也能为中医新药的临床和试验研究提供目标和思路。以文献数据为基础进行数据挖掘固然权威,但随着 Web2.0 时代的到来,互联网应用迅速普及,已成为公众获取信息、表达信息的重要平台,基于网络舆情进行数据挖掘正成为研究热点。因此,本文选择在线医疗网站的医患问答数据,利用文本挖掘技术,以 DVT 为研究对象进行研究,力求验证文本挖掘方法在探索疾病防治规律方面的可用性,以及医患互动问答网络数据源用于医学研究的可行性。

2 数据与方法

2.1 数据

以“好大夫在线”、“寻医问药”、“有问必答”3 大网站作为数据源,采集医生与患者对 DVT 疾病的互动交流信息,其维度包括患者描述、医生回答、用药情况、患者年龄、医生科室、研究方向,以前 3 个维度为主。结果好大夫在线获得数据 987 条 (2008.10 - 2016.04)、寻医问药网 7 790 条 (2012.03 - 2015.03)、有问必答 630 条 (2009.01 - 2015.12),清洗后共获得有效数据 6 871 条。

2.2 方法

2.2.1 概述 本研究主要采用文本挖掘方法。整个文本挖掘过程涉及文本采集、文本预处理、文本分析和知识获取 4 个环节。

2.2.2 文本采集 基于网络进行文本采集,通过网页的链接关系层层递进,最终在目的网页上抓取信息。本研究使用八爪鱼采集器作为数据采集工具获取上述维度数据。

2.2.3 文本预处理 (1) 文本分词。不同于英文,中文需要对文本进行分词处理,分词的准确性对于数据挖掘效果具有很大影响。为保证分词效果,分词过程中需要导入专业词典、自定义词典等;也可根据需要进行名字识别和词性标注。(2) 数据清洗。对分词后的文本去除标点符号、数字、日期、连词、介词、助词等停用词,必要时进行同义词替代或专有名词保留,避免一些冗余无用信息干扰后续文本挖掘结果,保障文本的有效性和准确性。本研究文本预处理过程以及涉及的方法和技术如下:(1) 下载停用词表(哈尔滨工业大学停用词表、四川大学机器智能实验室停用词库、百度停用词列表),编写代码将 3 个词表的内容合并。(2) 从搜狗词库中下载“药物不良反应术语”、“药品名称 商品名”、“药学词库”、“医学词汇大全”、“中外药品名称大全”、“中医中药大全”词典,使用深蓝词库转换 2.1 软件将其转为 TXT 格式。(3) 浏览数据对词库进行人工修订,加入一些较长的语句,如低分子肝素、月经时间延长等,补充其他药名,避免分词工具无法区分,完善词库。(4) 使用结巴分词进行分词并统计词频。

2.2.4 文本分析和知识获取 主要利用一些专业技术进行文本分析,提取有用信息,获取知识,进行可视化展示。此环节涉及的主要方法是词频统计、语义网络图等,词频统计主要统计特征词出现的频率,语义网络图能够帮助进一步了解关键词之间的联系。构建语义网络图步骤包括生成高频词表、生成行特征词表、生成共现矩阵词表和绘制语义网络图。

3 文本分析和知识获取

3.1 整体分析

3.1.1 词频 将“好大夫在线”、“寻医问药”上采集的数据分为患者描述、医生回答和用药情况 3 个文件，保存为 ROST CM6 接受的 TXT 格式，然后进行词频统计，结果见图 1 - 图 3。从词频统计结果来看，患者描述和医生回答涉及范围基本一致；但患者更为关注疾病带来的痛苦（肿胀、疼痛）、用药（肝素、华法林、口服、注射），医生更为关注

症状或并发症（栓塞、凝血、水肿、静脉曲张等）以及用药剂量。除去“静脉”、“血栓”这种无用词，从患者和医生的排序来看，“肿胀”、“疼痛”是排名第 1 的词汇，说明 DVT 的表现方式以肿胀为主，给患者带来非常大的痛苦；其次是华法林，说明此病使用最广泛的是这种药。将以上 3 类分词结果进行词性分类，得到症状的描述，如肿胀、疼痛、水肿、静脉曲张、凝血；药物的描述，如华法林、肝素；并发症的描述，如栓塞、综合征。这些可作为后继细化研究的出发点。

寻医问药_患者描述_词频			好大夫_患者描述_词频		
文件(F)	编辑(E)	格式(O)	文件(F)	编辑(E)	格式(O)
静脉	5164		疼痛	232	
血栓	4971		华法林	172	
肿胀	638		肝素	83	
华法林	470		口服	70	
疼痛	452		药物	57	
静脉曲张	264		输液	55	
药物	240		阿司匹林	51	
栓	143		分子	49	
中	130		静脉曲张	45	
肝	107		栓	38	
水	106		性	36	
口服	104		坏死	32	
输液	103		溃疡	30	
阿司匹林	90		凝血	36	
肠	80		方案	25	
！	74		水肿	24	
动脉	72		不	23	
分	63		发	23	
子	60		凝	22	
药	60		无	20	
方	54		动脉	20	
凝	50		中	15	
石	48		林	14	
无	39		剂	13	
力	34		压	13	
量	32		痛	12	
注射	32		血	12	
液					
病					

图 1 患者描述词频统计

寻医问药_医生回答_词频			好大夫_医生回答_词频		
文件(F)	编辑(E)	格式(O)	文件(F)	编辑(E)	格式(O)
栓塞	106		血栓	1263	
药物	89		静脉	1160	
肿胀	80		药物	341	
疼痛	76		华法林	249	
咳嗽	39		口服	227	
静脉	34		肿胀	166	
维	31		栓	138	
中	31		综合征	115	
静脉曲张	31		血	109	
水	30		剂	89	
高	25		里	63	
坏	24		弱	61	
性	21		弹性	59	
溃疡	21		颗粒	58	
胸痛	20		肝素	54	
结	20		子	48	
华	18		作用	44	
法	17		抗	42	
解	17		体	36	
作	15		规范	35	
用	14		场	33	
子	14		曲张	30	
栓	13		静	25	
酶	13		力	23	
动	13		该	22	
脉	12		因	20	
栓	12		林		
系			狼		
肝					

图 2 医生回答词频统计

用药情况_词频:	
文件(F)	编辑(E)
药物	164
华法林	150
血栓	64
静脉	56
阿司匹林	52
化	33
肝	27
素	19
酶	15
子	15
口服	14
液	10
输	10
不	9
凝	8
血	7
肿	6
胀	6
分	5
子	5
药	5
方	5
凝	5
石	5
无	5
力	4
量	4
注射	4
液	4
病	3

图 3 用药情况词频统计

3.1.2 语义网络图 基于词频统计可以得到出现频率较高的关键词，但关系不明确，如华法林与肝素是一起使用还是分开等，所以构建语义网络图来明确关键词之间的关系。由于篇幅关系，展示个别网络图，见图 4 和图 5。

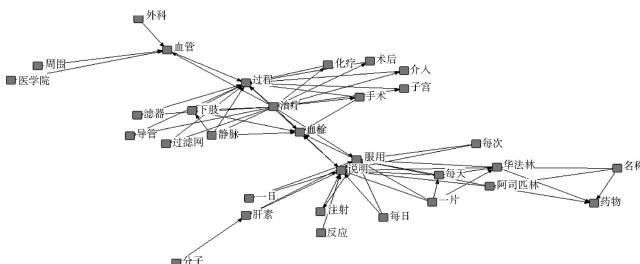


图 5 用药情况语义网络

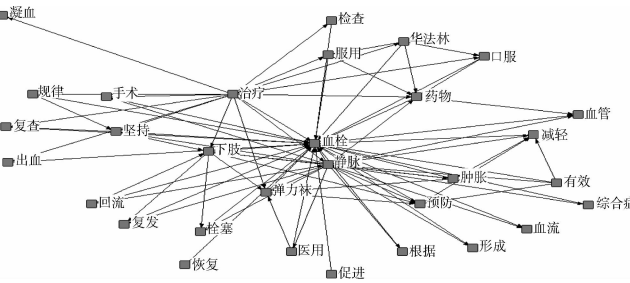


图 4 医生回答语义网络

基于语义网络图，结合相关医疗文献，可总结如下结果：(1) DVT 多发于术后，发病于下肢，最初表现为下肢腿部肿胀疼痛。根据“静脉曲张”、“蚯蚓”的词频，可见很多患者有静脉曲张的症状。

以“深静脉血栓”、“静脉曲张”作为主题在知网数据库搜索,可得到数篇静脉曲张诱发 DVT 的文献^[11-12],大多是在手术治疗静脉曲张时会有 DVT 的隐患。丁香园的一位血管外科博士明确指出 DVT 与静脉曲张存在误诊问题^[19]。(2) DVT 治疗使用的药物以华法林(口服)、肝素(注射)、阿司匹林为主,为保证血液顺利回流,会伴以弹力袜的治疗。(3) 由于血栓脱落的隐患,会造成肺栓塞出现,所以会放置滤网,以防血栓脱落堵塞。DVT 并不是只带来肿痛这一痛苦,还有潜在的猝死风险。(4) 验证了 DVT 的两大病因:血液高凝和血流缓慢。糖尿病可能会造成血液黏度增高,“妇产科”高频词说明怀孕是高凝的原因,“骨科”高频词说明术后的制动状态也是致病诱因。

3.2 细化分析

3.2.1 关于肝素的数据分析 用“肝素”查询筛选“好大夫在线”、“寻医问药”上采集的数据。将数据分为患者描述、医生回答和用药情况 3 类,进行分词,统计词频,见图 6-图 8,绘制语义网络图,见图 9。可知:(1) 肝素主要起到对凝血机制的治疗作用。通过对文献以及原始文本数据的回溯,认定肝素的作用是为了解决血液的高凝问题。(2) 发现了出血与肝素之间存在关系。肝素在缓解血液凝结的同时,可能会造成出血,需要格外注意;结合之前出现的关键词“剂量”,说明肝素的使用剂量是一大影响因素,用量过大可能会导致自发性出血,使用时需要进行血液学监护,以防发生不测。(3) 使用肝素治疗的同时,出现了华法林、弹力袜的关系,说明仅使用肝素对于一些患者来说是远远不够的,还需要口服华法林和辅助腿部的弹力袜。对患者来说,康复可能会是一个痛苦且漫长的过程,通过丁香园的药物报价查询,华法林、肝素价格均在 500 元以上,这对于患者是一个不小的负担。而血栓是一个坚持治疗的过程,不能停药,在原数据中出现了因停药病情加重的案例。

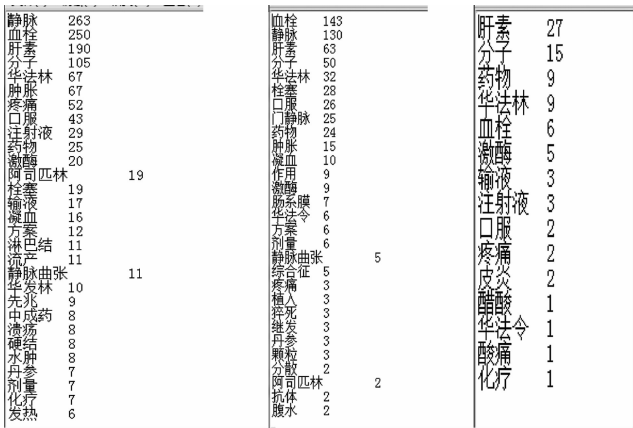


图 6 患者描述

图 7 医生回答

图 8 用药情况

词频统计

词频统计

词频统计

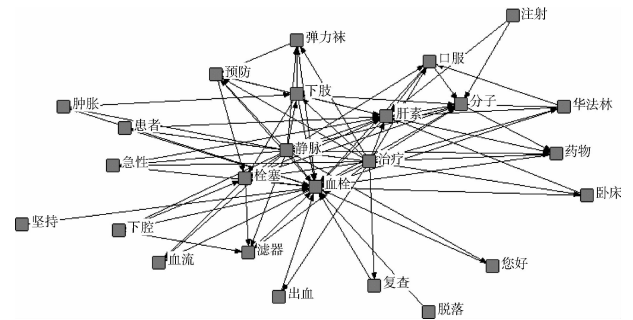


图 9 医生回答语义网络

3.2.2 关于高血压、糖尿病的数据分析 除去静脉曲张这种高频疾病,对比词频统计,发现高血压、糖尿病在词频中位于出现的疾病前 3 位。筛选出高血压、糖尿病的词语,人工浏览发现糖尿病患者多有 DVT 的并发症,而高血压与 DVT 关系不明确,有待于临床继续研究。

3.2.3 关于中药使用的数据分析 筛选出中药的数据,分为医生回答、患者描述两类。通过分词可以看到中药主要起到活血化瘀的功能,根据中药本身的药性,其只能作为辅助类药物,或者说对于病情较缓的患者起到改善的作用。具体经常使用哪几种中药,从“寻医问药”患者描述的词频统计中可知为丹参、当归、黄芪,以丹参应用最多。也有研究说明丹参对 DVT 的治疗是有效果的^[10]。

3.2.4 关于胸痛症状的数据分析 对胸痛的子集进行语义网络分析,发现胸痛与栓塞具有关联,胸痛是栓塞的表现,回溯原数据验证了这一点。急性肺栓塞缺乏特异性的临床症状和体征,所以患有 DVT 的患者发生胸痛症状时,应首先考虑肺栓塞发生的可能性。

3.3 其他信息提取

3.3.1 患病性别 对寻医问药网的数据单列出性别一列可知 DVT 的患者人群男性(56%)要比女性(44%)高。根据西班牙 Manuel Monreal 教授的研究,性别对 DVT 的患者人群的确有影响。女性 DVT 的病例较低,这可能意味着女性本身的体质、血液循环、生活习惯等比男性较优越,可以根据这一点为今后的医学探索提供思路。

3.3.2 患病年龄 将患者年龄按照每 10 年一个间隔,见图 10,观察发现 20 岁以后发病率逐渐升高,50~60 岁为高发期。根据我国人口平均寿命为 73 岁来看,70 岁以上患者人群减少不一定代表这个年龄层不会患病,而是因为整个社会这个年龄段的人较少而已,提示中老年人应将 DVT 作为预防腿部疾病的一大重点,需要经常活动。

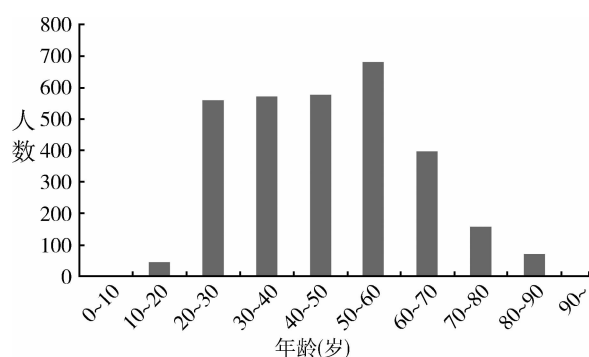


图 10 患者年龄统计

3.3.3 抗凝药物与饮食的关系 数据处理过程中,看到有医生对饮食问题加以嘱咐。根据常识来看,一般术后患者或重病患者都需要格外注意饮食问题。在丁香园搜索“深静脉血栓”、“饮食”、“华法林”等词语进行研究。华法林属于抗凝药物的一种,主要抑制维生素 K 依赖的凝血因子合成,从而防止血栓的形成和发展^[20]。食物中维生素 K

含量多种多样,使得华法林成为与食物关系最密切的药物之一。减弱华法林作用的食物主要是绿叶植物和水果、各种植物油等,增强华法林作用的食物包括银杏、大蒜、生姜、茴香等。

4 讨论

4.1 研究总结

文本挖掘技术应用于生物学领域,在国内外均取得了较好的应用效果^[9]。随着医疗信息文献数据库的建立和完善,目前很多国内学者青睐利用文本挖掘技术,基于文献数据研究某种中药成分的用药规律,如基于中医药文献研究玉屏风散^[1]、青风藤^[2]、复方丹参注射液^[8]的用药规律,或研究某种疾病的中医治疗用药规律,如慢性肾脏疾病^[3]、反流性食管炎^[4]、慢性阻塞性肺疾病^[5],以及中药的病证方药相应规律^[6-7]等。由于中医药在我国历史较长,学者基于文本挖掘技术对中医药的研究较多,西医方面的文献较少。在数据来源上,基本上都选用“中国生物医学文献数据库”,来自网络的数据源尚未使用,也未与临床数据相结合。在国外,基于网络和社会媒体数据,有相应的工作发布,如进行症状监测^[18]、疾病监测和生物事件的探测^[13]、提取药物不良反应信息^[14]、探索自我披露健康信息支持公共健康监测和医疗保健^[15]、监控滥用处方药^[16]、调查是否在线医疗社区的社交支持交换有利于患者的心理健康,如忧郁症^[17]等。根据国外的研究成果,已知网络数据在医学某方面应用的有效性,直接从网上抓取的数据更新及时,具有客观性、先进性以及临床价值。所以本文在研究方法和数据源选择方面进行了新的尝试。

4.2 研究的局限性

由于 DVT 近些年在我国才引起重视,医患信息的互通还存在一定不足,尚有很多问题,因此本研究也存在一定的局限性:(1) 在线医疗网站数据具有一定的时效性,也未必全面。通过对比医疗文献中的病因,静脉损伤这一病因就未在数据中发现,说明还需要更多的数据或更深层挖掘。(2) 对于发

病率不高的病种, 病例也是有限的, 对数据的搜集需要长期跟踪, 增加数据量, 提高分析结论的准确度和可信度。(3) 对数据仍需要细化, 进行某方面的细致分析, 如挖掘药物剂量与患者病情的关系、患此病的孕妇如何治疗不会对胎儿造成危害等问题。(4) 一些发现和疑点还需要进行临床的验证。

5 结语

鉴于“好大夫在线”、“寻医问药”、“有问必答”3 个网站的全国覆盖性, 从患者和医生言论可以反映临床诊断与治疗, 可以看出当前对 DVT 的整体认识和诊治水平。本研究主要验证了文本挖掘技术用于疾病防治规律获取的有效性, 以及医患互动问答网络数据源用于医学研究的可行性, 作为一个巨大的病患样本库, 能为医学研究提供有价值的信息。本文抛砖引玉, 希望更多学者投入于此, 用新的发现拓展临床诊断、用药的范畴和思路。

参考文献

- 1 郑康, 郑光, 罗丹, 等. 利用文本挖掘技术探索玉屏风散的用药规律 [J]. 中华中医药学刊, 2015, 33 (5): 1131 - 1133.
- 2 李雨彦, 郑光, 刘良. 文本挖掘探讨青风藤用药规律研究 [J]. 世界中医药, 2015, 10 (6): 823 - 827.
- 3 黄允瑜, 巴燕, 郑光, 等. 基于文本挖掘技术的慢性肾脏疾病中医治疗用药规律研究 [J]. 中国中医基础医学杂志, 2011, 17 (8): 912 - 914.
- 4 贺丹, 姜淼, 郑光, 等. 利用文本挖掘技术探索反流性食管炎疾病的症状、证候以及用药规律 [J]. 中国中医基础医学杂志, 2014, 20 (9): 1242 - 1244.
- 5 展俊平, 吴萍, 郭洪涛, 等. 基于文本挖掘技术的慢性阻塞性肺疾病临床用药规律分析 [J]. 中国中医药信息杂志, 2011, 18 (12): 27 - 30.
- 6 梁非, 展俊平, 李立, 等. 基于文本挖掘方法探索寒性热性中药的病证方药相应规律 [J]. 中国实验方剂学杂志, 2013, 19 (15): 333 - 337.
- 7 谭勇, 郭洪涛, 郑光, 等. 利用文本挖掘技术探索中医治疗疾病的用药规律 [J]. 世界科学技术 (中医药现代化), 2010, 12 (5): 823 - 827.

- 8 田景平, 郑光, 郭洪涛, 等. 利用文本挖掘技术探索复方丹参注射液应用规律 [J]. 中国实验方剂学杂志, 2013, 19 (4): 330 - 334.
- 9 翟兴, 唐燕, 林琳, 等. 文本挖掘在生物医学领域中的应用文献计量学分析 [J]. 现代生物医学进展, 2015, 15 (8): 1534 - 1539.
- 10 余盛飞. 丹参注射液预防普外科手术后患者下肢深静脉血栓形成的临床研究 [J]. 中外医疗, 2012, (2): 79, 81.
- 11 林启谋, 李勇, 周毅. 大隐静脉曲张术后深静脉血栓形成的预防 [J]. 当代医学, 2013, 19 (24): 49 - 50.
- 12 李大林, 张鲲, 颜京强, 等. 低分子肝素预防下肢静脉曲张术后深静脉血栓形成的效果 [J]. 中国普外基础与临床杂志, 2014, 21 (12): 1552 - 1554.
- 13 Corley CD, Cook DJ, Mikler AR, et al. Text and Structural Data Mining of Influenza Mentions in Web and Social Media [J]. Int J Environ Res Public Health, 2010, 7 (2): 596 - 615.
- 14 Nikfarjam A, Sarker A, O' Connor K, et al. Pharmacovigilance from Social Media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features [J]. J Am Med Inform Assoc, 2015, 22 (3): 671 - 681.
- 15 Ku Y, Chiu C, Zhang Y, et al. Text Mining Self - Disclosing Health Information for Public Health Service [J]. Journal of the Association for Information Science and Technology, 2014, 65 (5): 928 - 947.
- 16 Sarker A, O' Connor K, Ginn R, et al. Social Media Mining for Toxicovigilance: automatic monitoring of prescription medication abuse from Twitter [J]. Drug Saf, 2016, (39): 231 - 240.
- 17 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.
- 18 Gesualdo F, Stilo G, Gonfiantini MV, et al. Influenza - Like Illness Surveillance on Twitter through Automated Learning of Naïve Language [J]. PLoS ONE, 2013, 8 (12): e82489.
- 19 静脉曲张是怎么形成的? 要不要治? [EB/OL]. [2016 - 05 - 11]. <http://dx.y.com/column/2956>.
- 20 饮食对抗凝药物的影响究竟有多大? [EB/OL]. [2016 - 08 - 25]. <http://dx.y.com/column/3687>.