基于 Web 数据挖掘的失眠症人群特征分析*

王林峰 晏峻峰 刘欢庆

（湖南中医药大学管理与信息工程学院 长沙 410208） （南华大学附属第一医院 衡阳 421001）

〔摘要〕 对 Web 数据挖掘中的一些常用方法进行介绍，包括网络爬虫技术、中文分词、关键词提取算法等，通过网络爬虫技术获取在线医疗网站中与失眠相关的数据，对数据进行清洗和分类处理，基于规则对文本数据进行分词、关键词提取，分析失眠患者的性别、年龄分布情况以及症状、病因等特征。

〔关键词〕 Web 数据挖掘；分词；失眠；关键词提取

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.08.015

Analysis on the Characteristics of Insomnia Groups Based on Web Data Mining *WANG Lin－feng, YAN Jun－feng, College of Management and Information Engineering, Hunan University of Traditional Chinese Medicine, Changsha 410208, China; LIU Huan－qing, The First Affliated Hospital of Nanhua University, Hengyang 421001, China*

〔**Abstract**〕 The paper introduces some common methods (including web crawler technology, Chinese words segmentation and keyword extraction algorithm) of Web data mining, acquires the data related to insomnia in the online medical website through the web crawler technology, classifies and processes the data, carries out words segmentation and keywords extraction of the text data based on the rules, and analyzes the gender and age distribution situations, symptoms, causes of disease and other characteristics of patients with insomnia.

〔**Keywords**〕 Web data mining; Text segmentation; Insomnia; Keywords extraction

1 引言

互联网技术的快速发展，民众信息素养的提升和国家政策的支持，使得在线医疗服务平台得到快速推广，越来越多的患者选择在网上寻求医生帮助，甚至是一对一的诊疗。中国互联网络信息中心发布的《第 37 次中国互联网络发展状况统计报告》显示，2015 年中国互联网络医疗用户数量为 1.52 亿人。在线医疗服务提供一个新的方式，让患者可以不用实地见到医生，就可以咨询病情，能够帮助患者快捷地了解自身的健康信息，获得及时诊疗，很好地解决了医疗资源分配不均导致的看病难问题。随着时间的推移，互联网医疗积累大量的诊疗数据并不断更新，但这些诊疗数据大多为非结构化数据，如何有效地利用这些数据是现阶段面临的主要问题。Web 数据挖掘技术可以自动化获取互联网上的非结构化数据并进行信息挖掘和知识提取。因

〔修回日期〕 2017－07－05

〔作者简介〕 王林峰，硕士研究生；通讯作者：晏峻峰，博士生导师。

〔基金项目〕 湖南省高校创新平台开放基金（项目编号：13K076）；国家重点学科中医诊断学开放基金（项目编号：2013ZYZD08）；湖南省 2011 数字中医药协同创新中心建设项目。

此，本文通过网络爬虫技术，从当前主流在线医疗服务平台中抽取与失眠主题相关数据，对数据进行预处理，利用自然语言处理中的分词、关键词提取技术分析失眠人群特征。

2 方法介绍

2.1 爬虫策略

2.1.1 工作原理 爬虫也称网络蜘蛛或者网络机器人，是一种按照一定的规则自动抓取万维网资源的程序或者脚本，也是搜索引擎的核心部件[1-2]。爬虫程序首先初始化一个 URL 列表集合，再从集合中按顺序抽取出 URL 并读取网页的内容，通过 HTML 标签抽取特定的内容，将该网页上的 URL 加入 URL 列表集合，重复上述过程，直到满足系统的一定条件停止抓取。

2.1.2 网页爬取策略 本研究利用 Python 标准函数库 urllib 与第 3 方函数库 BeautifoulSoup，结合多线程实现数据的并发采集。由于在实际数据采集过程中需要不定期采集更新的列表页面，而列表页的更新频率是动态变化的，因此为避免重复爬取已经获取的列表页，本研究设计连续双记录判别法，待爬取的页面结构，见图 1。

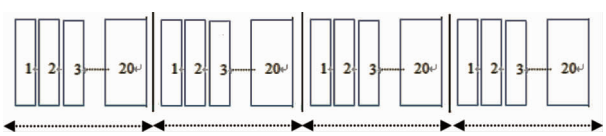

图 1 待爬取页面结构

假设列表页 L 共有 200 页，每一页有 20 个标题链接 T，可把该种结构看成队列，随着时间的推移，旧的标题链接页从右侧移出，新的标题链接从左侧进入。假设最近一次爬取的最后一个标题链接为 L_iT_j，那么 L_iT_j 存在 3 种情况：（1）L_iT_j 为列表页 1 的标题链接 1，即 $i=1$，$j=1$，意味着标题链接没有更新。（2）L_iT_j 为列表页 i 的标题链接 j，$1\leqslant i\leqslant 200$，$1\leqslant j\leqslant 20$，且 $i=1$ 时，$j!=1$，即标题链接 L_iT_j 处在中间或者最后的位置。（3）L_iT_j 已经不在队列中，说明页面信息已全部更新，不存在重复爬取的情况。在爬取列表页 1 时，将标题 1 和标题 2 存入数据表中，同时加上日期标注，再下一次爬取每个列表页时，取出比当前日期小的最后两条记录并逐页判断是否同时包含标题 1 和标题 2；一旦包含则停止爬虫程序。如果重复记录在该页的最后，理论上该算法最大遗漏记录数为 18 条（假设一个列表页有 20 条标题链接），但能够有效避免爬取重复数据。

2.2 中文分词

词是语言成分中的最小单元[3]。由于在中文句子中词与词之间没有分隔符，因此进行中文文本挖掘时，第 1 步便要对中文自然语言文本进行分割，变成合理的词语序列，分词的效率直接决定了文本挖掘结果的准确度。本研究使用的分词工具包正是采用基于词典与统计相结合的方式，算法步骤如下：（1）对句子进行正向扫描，在词典中逐词查找是否存在，将存在的词搭配情况构成有向无环图（DAG）。如对于句子“有力保障”，“有”、“有力”在词典中存在，“有力保”不存在，“力”、“力保”在词典中存在，“力保障”不存在，用 DAG 图表示为｛0：[0，1]，1：[1，2]｝。（2）利用动态规划算法寻找 DAG 中的最大概率路径，找出基于词频的最大概率切分组合。如字串 Z =“有力保障”可以拆分成 W_1 =“有力/保/障”，W_2 =“有力/保障”，最大概率分词就是要求得所有拆分组合中概率最大的组合，即 Max（P（W_1 | Z），P（W_2 | Z）），根据贝叶斯公式 P（W | Z）=P（Z | W）P（W）/P（Z），P（Z | W）表示出现词串的条件下字串的概率，很明显该值为 1，P（Z）为字串的概率，其对于各个拆分组合来说都一样，因此可以忽略，因此最大概率算法只要求出 P（W），即词串的概率。根据一元语法[4]，词的概率相互独立，有 P（W）=P（w_1，w_2.. w_n）=P（w_1）P（w_2）.. P（w_n）（公式 1），即拆分词组合的概率就是各个词出现的概率乘积，词的概率可以通过词出现的次数除以词典中总次数得到。由于词的概率是一个很小的小数，如果字串过长，拆成的词数也会越多，那么对于一组拆分组合来说，每个词的概率相乘以后乘积有可能接近于 0，因此通过数学公式变换，对

公式 1 两边取负对数，得到公式 2，即 $-\log P(W) = -\log P(w_1) - \log p(w_2) \cdots - \log p(w_n)$（公式 2），通过公式 2 可知当词概率越大，结果越小，因此求最大概率路径转换成求最短路径。根据第 1 步构造的有向无环图，可通过求最短路径进行求解，例如对于字串 Z = “有力保障”，可得到的有向无环图，见图 2（字母 a - g 为词的概率）。由图 2 可知，路径 b - >f 最短，因此为概率最大的切分组合，即“有力保障”的分词结果为“有力/保障”。（3）对于未登录词，采用隐马尔可夫模型进行分词。但该模型对专业词汇分词的效果不尽人意，因此本研究通过自定义词典保证部分未登录词的完整性，部分自定义词，见表 1。

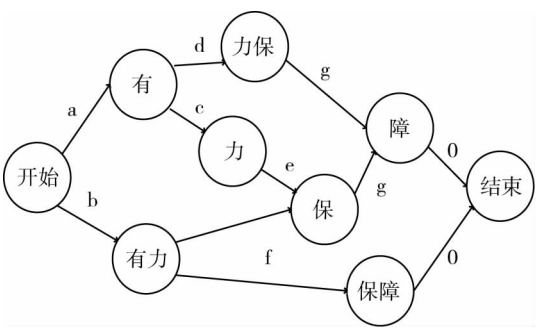

图 2 有向无环图

表 1 自定义词典

划分	自定义词
时间类	这几天，这两天，这些天，这阵子，几个月，半个月
表现形式类	难入睡，没睡意，难睡着，就醒了，醒的早，容易醒
病因类	气血不足，气血亏虚，气血虚，气血两虚，气血同虚，植物神经紊乱，植物神经功能紊乱
症状类	没力气，记忆力差，没记性

此外，由于中文语言表达的随意性和灵活性，仅仅通过自定义词典来保证分词的准确性不够，需要结合规则。例如从中医的角度分析失眠的病因，多与脏腑相关，故有很多词如“心肾不交”、“肝火扰心”、“肝肾阴虚”等在词库中不存在，通过自定义词典定义会很繁琐，且难以保证所有词都概括进来，通过构建如下规则 R = {s = “心 | 肝 | 脾 | 肺 | 肾 | 胃 *” or “ * 心 | 肝 | 脾 | 肾”，3 ≤ length（s）≤4}，如“心肾不交”在引入规则之前的分词结果为心肾/不交，引入规则后，由于首字匹配“心”，因此将其后缀词重新组成一个词，且长度为 4 满足条件，即“心肾不交”。

2.3 关键词提取

关键词提取是自然语言处理技术的基础与核心，在文档检索、自动摘要、文本分类与聚类、自动问答等方面有着广泛的应用[5]。TF - IDF 是一种基于统计的方法，通过计算词的权重来评估其在一段文本中的重要程度。TF 表示词在该文档中出现的频率，计算公式为 $TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k i,j}}$，式中：$n_{i,j}$为词在该文档中出现的次数，$\sum_k n_{k,j}$为该文档中所有词的频数之和。IDF 的算法原理是：如果在所有文档集中包含词 w 的文档越少，则 IDF 值越大，表示词 w 具有很好的分类特征，计算公式为 $IDF_i = \log \frac{|D|}{|\{j w_i E d_j\}|}$，式中：$D$ 表示文档总数，分母表示在所有文档中包含词 w_i 的文档数之和。如词“感染”在一篇文章中出现了 10 次，该文章中共有 1 000 个词，那么 TF 的值为 10/1 000 = 0.01，如果总共有 1 000 篇文章，包含“感染”的文章有 10 篇，那么 IDF 的值为 log（1 000/10）=2，最终词“感染”的权重为 0.01 ×2 = 0.02。如果在某一类文档 c 中词 w 出现的频数很高，说明该词能够很好地预测该类文档的主题，然而由 IDF 计算公式可知，c 值越高，IDF 值越小，这就是 TF - IDF 算法的缺陷。但是通过试验发现，TextRank 算法[6]对短文本的关键字提取算法效果不如 TF - IDF 算法，因此本文在传统 TF - IDF 算法的基础上引入特征字匹配和词位置方法干预词权重的计算。如由于中文语法的特殊性，病因等词往往在“由于”、“考虑为”、“由…引起”等词的附近，因此词位置是计算权重需要考虑的因素，通过试验发现，基于 TF - IDF 算法进行改进以后能够提高关键词提取的准确率。

3 结果分析

3.1 数据预处理

3.1.1 失眠术语表 截至 2017 年 3 月，共采集数

据 14 539 条，经过数据清洗去掉无关数据 1 241 条，有效数据共 13 298 条，包含患者性别、年龄、地区、发布日期、症状描述、病因等信息。为对数据进行分类，首先依据失眠指南构建术语表，见表 2。

表 2 失眠术语

类别	划分	特征词
病程	急性	这几天，最近，这两天，这些天
	亚急性	这阵子，这段时间，近来，＊个月
	慢性	一直，＊年，每天，长期，经常，总＊，好长时间
表现形式	入睡困难	睡不着，难入睡，没＊睡意，难睡着
	睡眠维持障碍	易醒，容易醒，就醒了
	质量差	做梦，多梦，噩梦

3.1.2 分类规则树 对患者的描述文本进行分词处理后，即可依据术语表对数据进行分类。在试验过程中使用字典类型依据术语表构建分类器，程序对所有记录进行扫描，依据分类器给每条记录贴上相应标签，如患者的描述文本中出现“没睡意”，则表现形式字段标记为“入睡困难”，如果含有时间词“这段时间”，则失眠分类字段标记为“慢性”。然而分类器的效率并不高，被正确归类的数据只占总数据的 62.4%，这是由于中文表达的随意性和灵活性导致分类器无法进行正确分类，如针对文本“晚上经常要醒 23 次”，可将该患者划分为睡眠维持障碍类，因此本文依据中文表达的语法结合规则树辅助分类，规则树结构，见图 4。

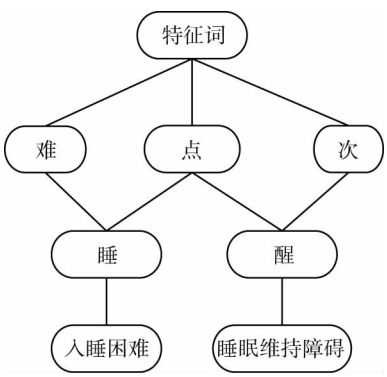

图 4 分类规则树

针对患者的问题描述，首先判断是否包含特征词“难”、“点”、“次”，如文本“晚上经常要醒 23 次”，包含“次”，接着查找“次”节点的子节点“醒”，在文本中也被包含，因此可将该记录分类为“睡眠维持障碍”。引入规则树后，分类的效果明显有所提高，被正确分类的数据占总数据的 78.7%。

3.2 结果统计

3.2.1 年龄、性别、地区分布情况 在所有患者中，青年（年龄 <45 岁）共 11 451 人，占总数的 86.1%；中年（年龄介于 45 ~ 59 岁之间）共 1 328 人，占总数的 10%；老年（>60 岁）所占比例不到 4%。在青年人中，25 岁以下共 6 628 人，占青年人总数的 57.9%；所有患者中，男性患者共 5 643 人，占总数的 42.4%，女性占 57.6%；在青年人中，男性患者共 4 868 人，占青年人总数的 42.5%。失眠患者的地区分布情况如下：排在前 3 位的为广东、江苏和河南，所占比例分别为 12.4%、11.4% 和 7.3%，其次为山东、河北、浙江、北京、四川、上海和辽宁。失眠患者最少的前 3 个省份分别是青海、宁夏和海南，所占比例之和不到 1.3%。

3.2.2 病程分类及表现形式 在所有患者中，急性失眠患者占比 27.6%，亚急性失眠患者占比 1.8%，慢性失眠患者占比 70.6%。失眠患者的主要表现形式，见表 3。

表 3 失眠患者的主要表现形式

表现形式	百分比（%）
入睡困难	35.1
睡眠维持障碍	9.8
质量差	32.5
入睡困难、睡眠维持障碍	6.0
入睡困难、质量差	8.0
睡眠维持障碍、质量差	6.2
入睡困难、睡眠维持障碍、质量差	2.4

3.2.3 伴随症状 在所有患者中，伴有头晕症状的患者占 30.6%，伴有乏力、疲累症状的患者占比 13.0%，伴有头痛症状的患者占比 22.9%，伴有身

体不适症状的患者占比 14.0%，伴有精神不振、烦躁、压抑、焦虑等不良情绪症状的患者占比 36.1%，伴有记忆力下降的患者占比 3.0%（部分患者同时伴有多种症状）。

3.2.4 失眠患者病因分析 不同年龄层次的患者，导致失眠的主要因素有所不同，各个年龄段的失眠病因，见表 4。

表 4 失眠患者的主要病因

年龄段	病因
<25 岁	心理因素、压力、气血不足、肾虚、脾气虚
25 ~ 44 岁	心理因素、压力、气血不足、心肾不交、脑供血不足
45 ~ 59 岁	心理因素、气血不足、压力、肾虚、脑供血不足
>60 岁	脑供血不足、气血不足、压力、肾虚、焦虑

值得注意的是，在任何一个年龄层次，大部分患者都伴有神经衰弱和植物神经紊乱，这两种神经内科疾病可导致失眠，而失眠会加重神经衰弱和神经功能紊乱，从而导致恶性循环。

4 结语

通过统计结果可以得知，目前失眠患者中青年人占绝大多数，通过病因分析可以得知，年龄在 45 岁以下的患者主要是由心理因素和压力引起失眠，心理因素以焦虑和紧张为主，可以推断这些不良的情绪和压力多与生活、工作和身体健康状态有关，因此对于青年人，学会释放压力、减少心理负担、控制负面情绪尤为重要，这也从侧面反映出社会竞争越来越激烈。对于 60 岁以上的患者，引起失眠的原因主要为生理原因，如气血不足、脑供血不足等，这类患者应多从养生保健方面进行调理，从而达到改善睡眠的效果。在失眠患者中，女性所占比例高于男性，这有可能是由于生理因素导致，如女性进入 40 岁后，雌激素和孕激素分泌减少，从而导致失眠；另外女性的性格相比男性更脆弱和敏感，容易受到家庭等因素的干扰而导致失眠。

参考文献

1 于娟，刘强. 主题网络爬虫研究综述 [J]. 计算机工程与科学，2015，37（2）：231 - 237.

2 杨靖韬，陈会果. 对网络爬虫技术的研究 [J]. 科技创业月刊，2010，（10）：170 - 171.

3 周宏宇，张政. 中文分词技术综述 [J]. 安阳师范学院学报，2010，（2）：54 - 56.

4 甘秋云. 基于最短路径的二元语法中文词语粗分模型的研究 [J]. 现代计算机，2013，（25）：7 - 10.

5 Onan A，Koruko，Lu S，et al. Ensemble of Keyword Extraction Methods and Classifiers in Text Classification [M]. Pergamon Press，2016.

6 张莉婧，李业丽，曾庆涛，等. 基于改进 TextRank 的关键词抽取算法 [J]. 北京印刷学院学报，2016，24（4）：51 - 55.

（上接第 58 页）

4 Michael Denny. Ontology Tools Survey [EB/OL]. [2016 - 07 - 14]. http://www.xml.com/pub/a/2004/07/14/onto.html.

5 Stanford University. Protégé [EB/OL]. [2016 - 07 - 01]. http://protege.stanford.edu/Stanford University.

6 徐国虎，许芳. 本体构建工具的分析与比较 [J]. 图书情报工作，2006，50（1）：44 - 48.

7 李景. 主要本体构建工具比较研究（上）[J]. 情报理论与实践，2006，29（1）：109 - 111.

8 李景. 主要本体构建工具比较研究（下）[J]. 情报理论与实践，2006，29（2）：222 - 226.

9 European Commission's Sixth Framework Programme. NeON Toolkit [EB/OL]. [2016 - 07 - 01]. http://www.neon - toolkit.org/wiki/Main_Page.html.

10 李晓瑛，李丹亚，夏光辉，等. 肿瘤本体构建研究 [J]. 数字图书馆论坛，2015，（8），37 - 42.

11 Stanford University. WebProtégé [EB/OL]. [2016 - 03 - 10]. https://github.com/protegeproject/webprotege.

12 Food and Agriculture Organization AGROVOC [EB/OL]. [2016 - 04 - 10]. http://aims.fao.org/vest - registry/vocabularies/agrovoc - multilingual - agricultural - thesaurus.

13 张运良，张兆锋，张晓丹，等. 使用 D3.js 的知识组织系统 Web 动态交互可视化功能实现 [J]. 现代图书情报技术，2013，（7/8）：127 - 131.