文章编号：2095-2163(2019)04-0016-05 中图分类号：TP277 文献标志码：A

基于贴吧的高校网络舆情预警和引导系统研究

来纯晓，李艳翠，金松林
（河南科技学院 信息工程学院，河南 新乡 453003）

摘 要： 随着移动互联网的发展和大数据、云计算、网络爬虫等技术的成熟，高校师生可以方便地从网上获取和发布信息。将大学生通过互联网表达出来的情绪和思想动态综合起来可以构成高校网络舆情。本文主要是对高校贴吧的帖子和评论进行采集，提出了一种基于 B / S 模式的网络舆情预警和引导系统构建的方法，用来监督高校网络舆情的状态，实现了对贴吧舆情信息的预警和引导，为高校网络舆情危机事件提供决策支持。

关键词： 网络爬虫；贴吧；高校网络舆情；B/S 模式

Research on college network public opinion early warning and guidance system based on post bar

LAI Chunxiao, LI Yancui, JIN Songlin
(School of Information Engineering, Henan Institute of Science and Technology, Xinxiang Henan 453003, China)

【Abstract】 With the development of mobile Internet and the maturity of technologies such as big data, cloud computing and Web crawler, college teachers and students can easily get and publish information from the Internet. Integrating the emotional and ideological dynamics expressed by college students through the Internet can form the network public opinion of colleges. This paper mainly collects the relevant network public opinion data of colleges and universities through posts and comments posted by colleges and universities. A method based on B/S mode is proposed to construct an early warning and guidance system for network public opinion of colleges and universities, which is used to monitor the state of network public opinion of colleges and universities, to realize the early warning and guidance of public opinion information of post bars, and to provide decision support for crisis events of network public opinion of colleges and universities.

【Key words】 Web crawler; post bar; network public opinion in colleges; B/S mode

0 引 言

随着网络和信息技术的快速发展，互联网在人们日常生活中发挥越来越重要的作用。人们通过互联网表达出来的态度、情感和意见共同构成了网络舆情。在校大学生作为一个特殊的群体，不仅会注意到普通网民所关心的热点和焦点问题，同时还将接触到自身知识层面人群所关注的特殊问题，比如各高等院校的百度贴吧[1]、微博[2-3]以及各大媒体正面或负面的网络舆情。这些信息在一定意义上反映出校园文化的健康程度，形成了高校网络舆情[4]。

高校网络舆情是对在校大学生在社会事件上观点的采集、整理、分析和反馈。舆情可以反映出部分在校生关注的热门话题，通过对反馈的舆情信息、特别是消极信息进行研究和透彻剖析，对负面信息做出及时预警，可以及时掌握学生的思想动态，提早开展思想政治教育工作。通过对高校网络舆情中的敏感词处理予以引导和控制，有利于维护高校的和谐稳定发展。

加强对高校网络舆情预警和引导系统研究，对网上言论进行有效及时的监测，对敏感信息进行预警，快速定位不良信息的源头，避免在舆情危机事件[5]处置工作中陷入被动局面，及时进行网络舆论的正面引导显得尤为重要，本次研究具有重要的实践意义与应用价值。

1 系统分析

1.1 系统导航功能分析

分析可知，系统导航则用于向用户显示各种功能的导航入口。为保证系统的功能和质量，系统导航主要实现注册登录功能。注册时，在前端页面加入验证，校验用户名是否已经注册，若为“是”，即将

基金项目：河南科技学院基层党建工作研究课题(ZDK-2018-06)。

作者简介：来纯晓(1991-)，男，硕士研究生，主要研究方向：自然语言处理、农业信息化；李艳翠(1982-)，女，博士，讲师，硕士生导师，主要研究方向：自然语言处理、农业信息化；金松林(1983-)，男，硕士，讲师，主要研究方向：农业信息化研究与应用。
通讯作者：李艳翠 Email：liyancui@ hist.edu.cn

收稿日期：2019-05-15

用户信息存入数据库，对密码进行算法加密，增强用户信息的安全性。登录时，用户将输入用户名和密码，同时还会设定登录规则和权限，确保系统操作的安全性。

1.2　舆情信息采集与储存功能分析

针对设定的目标贴吧，采用 Scrapy 爬虫[6-7]和 Path 构建数据爬取的框架，实现对目标贴吧的数据信息的爬取。信息采集系统能够方便、及时、有效地对指定贴吧的重要信息进行采集，并将采集到的信息保存到数据库或本地的储存设备中。本文的信息采集系统既能给用户返回一个 URL 地址，而且还能够实现对全部内容（发帖人、发帖时间、发帖内容和其它人的回复等）的提取。由于在网络上采集到的信息内容不尽相同，格式也是多种多样，本文拟对采集到的信息进行格式规范化整理，使得保存到数据库的内容更易于识别、以及后续的分析处理。

1.3　舆情热点功能分析

移动网络的发展、尤其是 Android 系统的智能手机的普及，使得人们可以方便地使用各种工具轻松搜索或发布相关信息。此现象会引起很多不同观点的交流和碰撞，个人意见在网络的迅速传播、并吸引相应人群的关注，就推动了高校网络舆情的产生、且日趋活跃。舆情热点功能主要就是按照舆情的热点进行展示，显示出当前时间用户关注度最高的 n 条数据。

1.4　舆情预警功能分析

按照发帖的时间先后排序，剔除掉当天发帖数量小于一定数目的数据，同时选取最近几天的所有有效帖子进行数据的情感分析，判断帖子情感信息是积极、中立还是消极，对内容消极帖子的用户将发送一个友好提示，起到舆情预警[8-9]的作用。在此基础上，用户可以根据关键词搜索，搜索出含有关键词的所有的帖子，包括发帖人、发帖时间、帖子地址等。系统不会仅局限于贴吧，而是适用于各种数据的监测分析，只要在数据处理后、再转而存储于数据库中，就可以进行数据分析及数据展示。

1.5　舆情引导功能分析

网络的发展使得网络舆情的传播越来越迅速，网络舆情成为了高校不容重视的重要问题，网络舆情的出现在很大程度上引导着人们的情感变化。因此需要设计有针对性的舆情引导[10-11]系统，从外部获取舆情信息并将其展示到页面中，为了实现对舆情信息敏感词的有效过滤，将舆情信息中的负面影响进行引导处理，防止负面信息的影响扩散，最终将正确展示所有信息，给网站营造一个良好的环境，从而为高校对于舆情管理的监管控制提供有效的解决方案。

2　系统设计

2.1　系统的总体设计

本次设计将分为高校网络舆情信息采集处理模块、网络舆情预警模块和网络舆情引导模块三个部分进行研究。系统利用 B/ S 结构[12]，前端浏览器部分主要用于实现舆情数据的呈现以及系统和用户之间的交互，服务器端负责搜集网络舆情数据、批量发布舆情引导信息以及响应用户的执行动作。系统前后端的架构设计如图 1 所示。

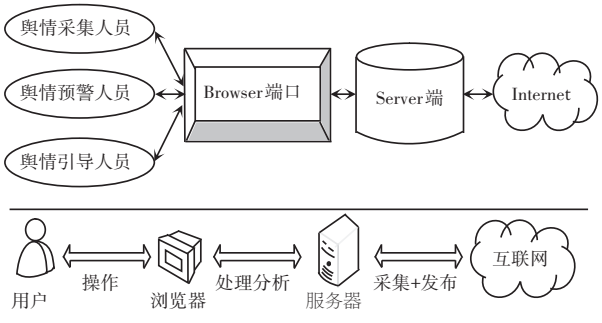

图 1　系统前后端架构设计图

Fig. 1　Front end and back end diagrams of the system

本文研发的是一套综合系统，设计层面既包含软件开发、又包含硬件配置，而且需用到信息发布服务器、数据库服务器、数据采集抓取服务器以及舆情信息处理服务器，限于目前条件，这些内容都将部署在同一台服务器上。

2.2　系统功能模块的设计

结合实际情况，本文设计了高校网络舆情信息采集处理、高校网络舆情预警和高校网络舆情引导三个模块。系统总体功能模块设计如图 2 所示。对此可做阐释分述如下。

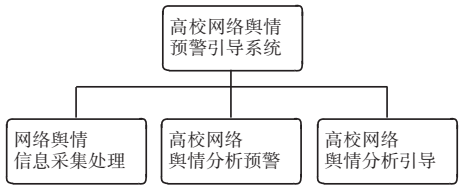

图 2　系统总体功能模块图

Fig. 2　System overall function module diagram

2.2.1　网络舆情信息采集处理模块

本文中界面模块包括 3 个按键，分别是：搜索框、显示数据和数据详情。界面右面的信息框旨在显示爬取贴吧的主题，界面左面空白框则用于显示

主题的详细信息和回复的详细内容。选中左边帖子信息的主题、再按数据详情,帖子的详细内容就在左边的文本框中显示出来。采集贴吧信息首先是要分析贴吧网页的具体内容。剔除广告和一些杂乱的信息,提取贴吧中发帖的全部内容(发帖人、帖子主题、内容和别人回复等详细信息)。在网络上采集到的信息格式是多种多样的,将采集到的信息进行格式转换,预处理的首要任务就是把搜集得到的信息进行格式处理,使之成为纯文本格式。

网络舆情信息预处理主要是对采集系统从贴吧获取到的数据通过去重、去噪、分词、编码等操作环节,并提交给分析系统做进一步分析与研究。舆情信息分析系统作为本系统的核心部分,主要任务是完成帖子主题的检测与追踪、焦点主题的识别和预警、敏感信息的替换和引导。舆情分析可依据用户的要求展示各种主题帖子的产生、发展、移除的整个过程周期,同时对分析信息进行析取、转换和存储,最终构建出主题信息分析数据库。

网络舆情信息分析功能是将所采集的数据进行分类整理,按照对应的类别属性描述,通过一段时间的积累,系统可以从多角度掌握采集到的初始信息源的状况,更有利于后续对初始信息的研究分析。分析流程过后,系统展示功能是形成各种规则统计、图表,并整合为分析报告,经过专家审查合格后,形成最后的结构化信息,提交给对应的管理部门,作为高校网络舆情分析和最终决策的重要依据。

2.2.2 高校网络舆情预警模块

情感分析[13-14]也叫做意见挖掘、倾向性分析等。简单来说,即是对带有感情色彩的主观性文本进行分析、处理、归纳、推理的过程,分析这句话表达的是积极、还是消极的情绪。目前,情感分析方法主要分为2种:基于情感词典的方法和基于机器学习的方法。本文使用基于情感词典的方法。需要尽量庞大完备的词库,每个词库基于不同领域效果是不同的,所以需要找到适用于本系统需求的词库,将文本进行分词处理(本文使用结巴分词),再根据情感词典对分词结果予以评判打分,制定一个标准,使其分类为积极、中立和消极三大类。

提取数据库中的数据,并对所有帖子的回复进行分词处理,研究中使用结巴分词,将使用情感词典对分词后的结果做出打分,并计算得分情况,判断情感为积极、中立或者消极。用户可以通过此地址进行帖子的查看,以及对消极情绪的帖子采取一定的措施。对发生过的紧急事件,系统还要随时监测其网络舆情的发展态势,并在达到警戒值之前发出预警。同时还要对可能诱发突发事件的因素进行实时监控,探究发现突发事件发生的前兆,进行先兆预警,做到防患于未然。

2.2.3 高校网络舆情引导模块

舆情引导系统要具有网站最基本的用户登录和注册功能,能够查看本系统的人员变化以及权限分配,系统需要对舆情信息中的负面消极信息进行及时引导,因而就要通过相应页面展示该类信息方便后续的查看。模块的核心功能是过滤用户回复,过滤信息时要匹配敏感词,还要用户将其添加到数据库中。当公开发布信息含有敏感词汇时,就要将该条评论进行替换或者屏蔽的引导处理,确保发布信息的正面性。

调研后分析可知,高校网络舆情的特点总是依据其所处阶段的不同而发生变化,因此应对策略也各有差异。此时,就要依据网络舆情预警指标的不同,提供差异化的舆情引导。建立动态专题,添加近段敏感问题专题,扩充信息量,增强此类信息在平台存在量。及时发布信息,高校师生的微信、贴吧关注度非常高,可以发动广大师生的群体力量,将其身边的党政宣传工作、尤其是富含正能量的校园信息、甚至是生活趣事进行有选择的发布,提升发布信息的鲜活度和生命力。

3 系统实现

3.1 系统开发环境

本文系统的开发运行环境详见表1。

表1 运行环境表

Tab. 1 Running environment table

功能	软件环境
后台显示界面	x-admin2.1
数据采集	Scrapy1.5.1/Requests2.19.1
信息发布	贴吧 api: comments/create
数据存储	MySQL5.7
数据分词	Jieba0.39
后台环境	Django1.11.3

3.2 网络舆情信息采集处理模块构建

利用 python 爬虫中的 request 进行请求,爬取百度贴吧,利用 Xpath 提取感兴趣的数据并将其保存在 mongodb 数据库中。而在清洗爬虫采集到的数据过程中,将除去无效的信息内容,接下来开启数据预处理环节,即将句子拆分为词语集合,使用正则表达式和中文结巴分词进行分词处理,处理后的数据则

保存到 MySQL 数据库中。信息采集处理流程如图 3 所示。

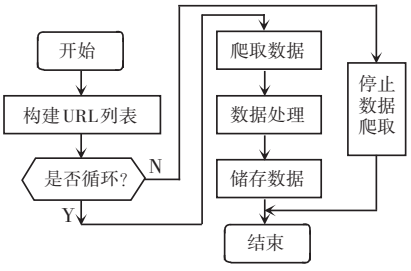

图 3　信息采集处理流程

Fig. 3　Information collection and processing process

在爬取时，主要采集的是百度贴吧的数据。考虑到贴吧中数据较多，例如河南科技学院贴吧共有主题数 260 316 个，帖子数 7 291 694 篇，科院精英数 218 598，时间从 2006 年开始，由于针对同一 IP 爬取数据较多将被封，只爬取了 2018 年的主题约 8 400 个。目前共收集了河南的 10 所高校的数据，主题约 80 000 个。

经过对所采集数据进行简单分析后发现，大多数主题是与学生自身利益相关的事项、国内国际重大新闻时事事项、以及各类突发事项。事件又有一定的特点，对此可概述如下。

(1)与学校主要工作日程相关，如开学新生报到、毕业生离校、期末考试、运动会等大型校园文化活动事项。

(2)与节假日或重大纪念日等相关，如国庆节、劳动节、国家公祭日等。

(3)与各类长期性问题相关，如大学生权益表达与利益诉求问题、校园安全等。

3.3　高校网络舆情预警模块构建

研究最初，即为数据准备阶段。信息采集处理模块已经将所需数据准备好存入 MySQL 数据库中，结合本文使用的 BosonNLP 的情感词典，该字典中包含了大量的社交词语。这样就为本系统进行数据分析奠定了良好基础。因为回复中有大量的空格、图片、表情等。这类信息是无效的。因此需要进行数据处理。本次研究中，使用正则表达来除去这些无用的空格，从而保障数据分析的准确性。接下来，进行分词处理时，使用的是结巴分词。因为采集数据把所有回复放入了同一个字段，就是一个大字典中包括了很多列表，每一个列表就是一条回复。这里将有效遍历该字典，将每一个列表(每一条回复)进行结巴分词。对数据库中的分词信息和情感词典运算后进行打分操作，再对打分值做出判断，由此得到最终的情感分析结果。近 10 天贴吧数据采集信息数量和情感分析结果分别如图 4 和图 5 所示。

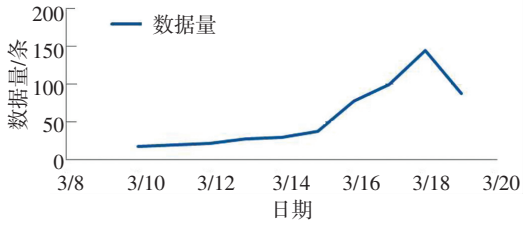

图 4　近 10 天贴吧数据采集信息数量图

Fig. 4　Data collection information quantity chart of tieba in recent 10 days

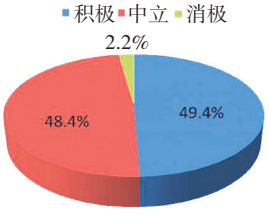

图 5　情感分析结果图

Fig. 5　Emotional analysis result diagram

在使用情感分析进行情感判断的过程中，基于机器学习的方法比基于词典的情感分析方法的表现更为客观，同时还发现用于训练和测试的数据集分别来自不同领域且本文获取到的有效信息有限，因此未来将会考虑扩充训练集以提升准确率。综上分析可知，情感分析属于机器学习，需要大量的数据进行机器训练，因而目前只做了一些简单的分析，与实际情感表达方向会存在一定的误差。在简单处理了数据后，提供数个网页用来实现数据展示、表达及一定程度的预警，对负向贴子的地址给出提示，用户可以根据地址访问负向帖子。

3.4　高校网络舆情引导模块构建

舆情引导模块由 5 部分组成，且有 1 个核心。其中，5 个模块分别是登录及注册模块、帖子及评论展示模块、发帖模块、回复模块、敏感词添加删除模块，而核心则是过滤算法。高校网络舆情引导系统如图 6 所示。文中，对此可得研究论述如下。

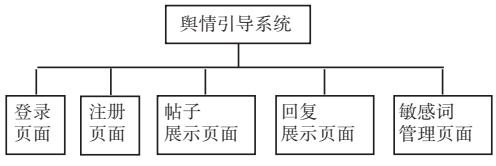

图 6　舆情引导系统图

Fig. 6　Public opinion guidance system diagram

登录及注册模块用于用户的登录注册使用，可以详细地确定用户是否在登录状态；展示模块用于信息呈现，主要是对帖子的信息进行展示，每一个主

题下可以显示多条回复,对此用户可做到全面查看;发帖模块可以让用户自己发表话题;回复模块是由用户对于帖子来发表评论;敏感词模块用于对数据库的敏感词进行添加/删除操作。过滤算法使用 Java 的 HashMap 来研发设计,构建敏感词库。为此,将扫描文章中的所有字符,当某一个字符匹配到敏感词的第一个字时,开始检索下一个字符,如果与敏感词的第二个字不匹配则退出,开始匹配下一条回复,如果匹配到相同字符,就针对从第一个字符到目前扫描的字符之间的字符串,求其 HashMap 值,看能否从对应的表中检出敏感词,如果检出就可以展开下一步操作,如果找不到或者超长,就从开始的字符继续扫描。这样就实现了敏感词过滤。在帖子回复页面中,如果有信息包含数据库敏感词时,系统将会自动选择用户,将已经存储在数据库中的正能量信息或者无关的普通信息发布出来,从而替换这条敏感回复,最终达到了对负面信息的引导消除作用。

4 结束语

高校网络舆情对建设和谐校园有重要的意义。当前互联网技术发展迅速,及时地引导校园网络舆情走向,防止舆情泛滥,将负面影响降至最低,如此才能使学生建立良好的价值观。本文基于 B/S 架构的高校网络舆情监控引导系统,将高校网络舆情的采集、分析、处理、引导实现了一体化,为高校在网络舆情引导工作管理和执行上提供技术和决策支持。本文构建的网络舆情监控与引导系统可为作者院校的正常发展起到保驾护航作用,出现危机情况时可做到及时的预警和正方向引导。但目前本系统尚处于初期起步阶段,后期还将面临科研工作任务,诸如完善系统、优化界面、增加功能等。

参考文献

[1] 赵芬,雷珍臻,杨晓云,等. 基于百度贴吧大学生网络舆情分析[J]. 电脑知识与技术,2018,14(28):227-229.
[2] 王杰. 基于微博大数据的舆情监测系统的设计与实现[D]. 天津:中国民航大学,2017.
[3] 罗咪. 基于 Python 的新浪微博用户数据获取技术[J]. 电子世界,2018(5):138-139.
[4] 陈艳红,向军,刘嵩. 高校网络舆情分析的 K-Means 算法优化研究[J]. 湖北民族学院学报(自然科学版),2018,36(4):442-447.
[5] 黄小熳. 高校危机公关中的网络舆情分析及有效引导[J]. 科教文汇(中旬刊),2019(1):7-9,39.
[6] 安子建. 基于 Scrapy 框架的网络爬虫实现与数据抓取分析[D]. 长春:吉林大学,2017.
[7] 陈涛,栾禹鑫,谭英杰,等. 基于爬虫技术的校园网络舆情分析和监测系统[J]. 网络安全技术与应用,2018(12):54-55.
[8] 杜艳,杜华. 新媒体时代高校网络舆情预警机制构建[J]. 山东工会论坛,2018,24(6):106-109.
[9] 连芷萱,连增水,张秋波,等. 面向突发事件的网络衍生舆情预警模型与实证研究[J]. 情报杂志,2019,38(3):133-140.
[10] 朱胜楠,李师,高蓉蓉. 自媒体时代下高校网络舆情引导研究[J]. 科技经济导刊,2019,27(9):180.
[11] 李巧芳. 新时代背景下医学院校网络舆情引导策略研究[J]. 锦州医科大学学报(社会科学版),2019,17(2):29-31.
[12] 吴晓珊,曹旭东,王森,等. 基于 B/S 架构的管理系统软件开发[J]. 计算机测量与控制,2019,27(2):123-128.
[13] 彭浩,朱望鹏,赵丹丹,等. 面向多源社交网络舆情的情感分析算法研究[J]. 信息技术,2019(2):43-48.
[14] 张鹏,崔彦琛,兰月新,等. 基于扎根理论与词典构建的微博突发事件情感分析与舆情引导策略[J]. 现代情报,2019,39(3):122-131,143.

(上接第 15 页)

能够很好地反映出症状和疾病之间的关系,降低了模糊证据推理的不确定性。

(4)在对模糊推理的隶属函数进行设置时,应该做到阈值的动态变化,对不同的疾病科室设定不同的阈值,以保证诊断的准确性。

(5)致病因素与疾病之间存在着多对多的关系,在知识库的设计中应考虑添加致病因素。

参考文献

[1] 曾梅. 浅谈人工智能在医疗器械领域的应用[J]. 科技广场,2017(12):57-60.
[2] 张德政,彭嘉宁,范红霞. 中医专家系统技术综述及新系统实现研究[J]. 计算机应用研究,2007,24(12):6-9.
[3] 周仲宁. 眼科疾病诊断专家系统的研究与实现[J]. 计算机工程与应用,1998(6):80-81.
[4] 潘军杰,张文清,周震. 口腔电子病历及辅助诊疗系统[J]. 深圳中西医结合杂志,2003,13(4):252-254.
[5] 戴细华. 多值逻辑语义博弈[D]. 广州:中山大学,2006.
[6] 胡宝清. 模糊理论基础[M]. 2 版. 武汉:武汉大学出版社,2010.