基于爬虫技术的校园网络舆情分析和监测系统

栾 悦

（西北民族大学 甘肃 兰州 730030）

[摘 要] 互联网时代下的今天，随着新媒体的兴起，高校在应对网络舆情危机方面面临新的机遇和挑战。网络舆情是在互联网上传播的公众所关注的现实生活中某些热点、焦点问题，是网民通过互联网来表达和传播不同思想和态度的总和。网络舆情源于现实生活，但它是不同于现实社会的虚拟现实，它将人们传播渠道和表达舆情的方式拓展到了互联网上。而本文介绍的此系统将爬虫技术和现有的技术结合起来设计了一种关注大学生心理健康的系统。这个系统以聚焦爬虫为主，民意调查为辅，通过对学校同学浏览网站的情况以及浏览网站内容做出相关的分析研究，提出相对应的策略帮助学校了解学生心理健康情况，使得大学校园出现更少的意外事件，减少校园暴力的出现。

[关键词] 舆情；爬取；聚焦爬虫；大学生；Python

0 引言

计算机网络的普及应用，为信息的广泛传播提供了前所未有的便利途径。与传统的传播媒介相比，高校校园网络的发展一定程度上容易给大学生的身心健康带来消极的影响，存在一定的安全隐患。因此，高校管理层应及时加强大学生身心健康教育工作，发掘正确的网络舆情引导机制，有效控制网络给学生带来的负面影响，控制舆情的蔓延。所以，如何及时化解、实现对高校网络舆情的引导和控制，维护高校大学生生活的和谐健康，已经成为构建社会主义和谐校园迫切需要解决的问题。大学生是众多高中生的佼佼者，他们自己往往会有一种优越感，但是这种优越感往往在一些困难面前会变成自卑，从而引发一些极端心理。现阶段正是大四毕业季，很多即将毕业的学生由于自身抗压能力不强而出现逆反心理，最终导致自己无法毕业并且后悔终生。基于这个问题，我们发现可以利用爬虫技术以及民意调查可以了解大学生心理状况，此技术一旦成型，将可以为学校的关注学生心理健康问题提供很好的便利。我们通过爬虫技术爬取网页，并且运用基于网页内容的分析算法对网页进行分析和过滤，算法主要是针对以文本和超链接为主的结构或简单的网页，可以快速有效的对网页进行分类和聚类。然后根据过滤得到的结果进行分析，得出最有可能的结论，进而为关注高校大学生的心理健康提供便利。

1 系统的设计和实现

该系统的服务对象目前只适用于高校老师和学生，意在帮助学生减少心理问题和高校老师更加了解学生，减少校园暴力的发生。

本系统主要分为三个模块：网络舆情的捕获模块，舆情信息的处理模块，分析模块。

1.1网络舆情的捕获模块：主要分为聚焦爬虫和网络民意调查两个部分。

在网络舆情的捕获模块，我们主要以聚焦爬虫为主，民意调查为辅。聚焦爬虫主要分为三个方面：对抓取目标的描述与定义、对网页的分析与过滤、对URL的搜索策略。

我们爬虫所爬取的对象一般为大学生浏览的网页为主，通过用户行为确定的抓取目标样例。再运用基于网页内容的分析算法对网页进行分析和过滤，为了提高爬取内容的准确性，我们将基于内容评价的搜索策略和基于Web链接结构的搜索策略相结合，从网页内容相关性分析角度确定网页与主题的相关性，从链接分析角度确定网页的权威性和主题资源搜索的覆盖率，对待爬行队列中的网页进行排序，确定主题爬虫爬行URL的优先级。URL的搜索方法我们主要采用的是广度优先算法，广度优先可以尽可能多的覆盖网页。

在网络民意调查当中，我们制定信息搜索计然后通过spss，matlab等技术对获取的信息进行分析处理，在动态的变化中捕捉到有力的信息。

1.2舆情信息的处理模块：主要是网络去重去噪模块。

网页去重分为同源网页去重和内容去重两个方面，同源网页的去重是对网页进行哈希散列实现的。通过对网页进行去复制，可以得到更加清晰的结构化舆论信息，最终将舆论网页存储在数据库中。网页中的噪声是指网页中包含的广告、版权信息、页面导航、注释和脚本以及一些与主题不相符的广告、无用图片、无用链接。这些噪声的存在将直接影响所捕获的有价值的信息，导致主题偏离，影响主题的正确识别。网页净化是对网页进行去噪，以获得网页的标题和这些有价值的信息的主体。在去噪之后，网页的主题内容成为处理的直接对象。文本处理程序提高了处理过程的准确性。因此，网页去噪已成为信息预处理过程中的重要环节，也是舆论分析模型。数据块提供数据的主要来源。

网页经过去噪后进行对页面的结构分析，网页是使用超文本标记语 言编写而成的，用户可以发布带有文本、表格、图像等资源的网络文档，可以点击超文本链接进行浏览。文档是一种半结构化的，用标记符来分隔文本各个组成部分的页面。而文档对象模型作为一个接口和平台，在存储和更新文件内容、结构以及风格的时候允许程序或脚本动态编译。

1.3舆情分析：分词识别模块，分词分析模块，舆情走向分析。

分词的实现采用了聚类算法，多中心的表现形式能形象的体现舆情话题的变化动态；使用双重或多重关键词赋予更高权值法更能准确地识别话题；我们在该模块的设计思想是：利用多中心，使用标题向量及正文向量来进行双向量比较，比较过程的原则是采用了双重或者多重关键词赋予更大权值，将分词聚类成分词树。

对于分词分析模块，首先要建立一套舆情分析评估指标体系，该体系利用专家评估、调查统计等方法，建立起相关数据表。其次，要建立评估模型，该模型是基于层次分析法的权值计算模块，在此对其进行简单的流程描述：通过构建一个判断矩阵，对其进行一致性的校验，该方法即可判断出矩阵是否合理，如果合理，建立层次机构模型，同时利用求和方法对各层的指标进行权值计算。最后建立评估计算模块，对以上所有的指标体系进行综合计算，采用加权平均法，将结果存于数据库中。

结束语

目前大学生心理问题已经受到了广泛的关注，越来越多的大学生由于自己压力大而造成了无法挽回的后果。我们这个系统主要针对大学生心理状况，防止学生“表面微笑”的现象出现。目前我们的项目预期成果是完成一款能实现相关功能的系统以及一篇学术论文，如今，项目所需要完成的系统已基本完成，具体功能正在完善中，预计在今年正式推出。

参考文献

[1]惠莹.基于爬虫技术的校园网络舆情监测元数据管理研究[J].电脑编程技巧与维护，2018（01）：116-118.

[2]艾舟.你看到了优点还是缺点[J].思维与智慧，2007（12）：54.

作者简介：

栾悦，1998年1月，女，辽宁铁岭市，本科，研究方向：数据库等。