文章编号:1673 - 0062(2013)03 - 0071 - 07

石鼓网络舆情监控系统的设计与实现

徐 强,万亚平,雷龙艳,阳小华

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:分析了高校网络舆情监控的现状,提出了基于高校网络舆情监控系统的设计方案. 该系统通过网络爬虫将相关网络新闻、博客和论坛的信息采集下来,采用关键词和敏感词匹配的方式对信息进行过滤,将用户关注的信息呈现出来,从而实现对舆情信息及时准确的发现.

关键词:网络舆情;网络爬虫;舆情监控

中图分类号:TP393.06 **文献标识码**:B

The Design and Implementation of Scool Network Public Opinion Monitoring System

XU Qiang, WAN Ya-ping, LEI Long-yan, YANG Xiao-hua

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: In this paper we analyze the present situation of the internet public sentiment monitoring work in colleges, then put forward the system design scheme of the college public opinion monitoring system. This system collects the information of the related network news, posts it on forum and BBS using the Web Crawlers, then filters the collected information by the way of keywords and sensitive words matching, and finally reveals the information which the user is focusing on, therefore, the public opinion information can be found timely and accurately.

key words: public opinion on network; web crawlers; public opinion monitoring

0 引 言

根据中国互联网络信息中心(CNNIN)发布的《第 31 次中国互联网络发展状况统计报告》显示,截至 2012 年 12 月底,中国网民规模达到 5.64 亿,互联网普及率为 42.1%. 互联网拥有了越来越大的

收稿日期:2013 - 04 - 13

基金项目:湖南省自然科学基金资助项目(11JJ6047);湖南省科技计划基金资助项目(2011FJ3087)

作者简介:徐 强(1986 -),男,山西大同人,南华大学计算机科学与技术学院硕士研究生. 主要研究方向:信息检索.

用户群,而在这些用户群中,大学生是网络利用率极高的一个群体,是学网用网的主力军之一.

网络舆情是由于各种事件的刺激而产生的通过互联网传播的人们对于该事件的所有认知、态度、情感和行为倾向的集合[1].

近年来,高校群体性事件频繁的发生,而所有这些事件的产生和发展都离不开网络的传播,网络舆情已成为事件发展的"导火索"和"催化剂"[2],因而如何防范网络不良舆情的快速扩散和演化,实现对网络舆情的有效监控与处理,已成为高校网络信息安全和宣传部门工作的重点[3].

国内外关于网络舆情的技术的研究相类似,以 BBS、博客、新闻网站等信息源为研究对象,集成信息采集、内容分析、话题提取、事件聚类等多种技术,对网络舆情监控进行了研究. 国外,1996 年 DARPA、卡内基 - 梅隆大学、Dragon 系统公司以及麻萨诸塞大学的研究者将对新闻媒体信息流在识别和跟踪方面的研究深入到了话题的领域,也就是对未知话题的识别和对已知话题的跟踪,从而诞生了国外最有名的与网络舆情监控方面相关的系统——TDT(Topic Detection and Tracking)系统[4-7]. 国内北京交通大学朱旭对高校群体突发事件预警平台相关技术和相关模块进行了研究,设计和实现了告警发布子系统[8],国防科技大学的叶昭晖等人设计并实现了基于搜索引擎的网络舆情监控系统[9].

石鼓舆情监控系统是一个以 MySql 作为后台数据库,使用 SSH 架构开发的针对普通高等院校的舆情监控系统,通过它实现对高校网络舆情信息的收集、监控和预测.

1 系统的设计与实现

本系统针对网络舆情的特性,采用计算机技术实现了自动对网络舆情进行分析、处理等服务,建立起全面高效的舆情监控机制,并且提供了相关信息的统计分析,从而帮助舆情工作人员能够及时的了解舆论动向,及时的采取相应的措施.

1.1 系统的功能描述

如图 1 所示:这是系统的主页面,管理员可以通过点击"管理员设置"(一般用户的主页面没有这个链接)来审核注册用户和更新系统敏感词库以及对目标网站信息和用户信息的维护,针对与一般的用户,首页上会将用户所关注的当天最新舆情信息显示出来,点击论坛监控、新闻监控等标签会出现相应网站类型下的舆情记录,方便用户对信息的所在网站的定位,点击更多舆情的链接,用户可以查看到截至当前日期的任何时候的舆情记录,用户也可以通过舆情搜索来实现搜索功能.

图 1 石鼓舆情监控系统主界面

Fig. 1 The main interface of Scool Public Opinion Monitoring System

系统还提供舆情统计分析的功能,会将所监控的几类门户网站的舆情信息以"一周内每天数据"、"一月内每周数据"和"一年内每月数据"的形式以柱状图的方式形象直观的呈现在用户面前,方便用户对相关舆情信息的统计分析,以便做出相应的策略. 用户通过点击"用户设置"对个人的关键词和监控网址的设置,可以实时的参与到舆情信息的采集和处理过程中来,从而高效、有针

对的获取相关的舆情信息.

1.2　系统设计

系统是基于 Java 开源的 SSH 框架实现的,三个框架各司其职,将系统的体系结构分为四层,即表示层、控制层、业务逻辑层和数据访问层,另外实体类用于各层之间数据的传递. 系统架构图如图 2 所示:表现层使用 JSP 来构建,为客户端提供对应用程序的访问和相关结果的显示;控制层接受客户端的请求,根据不同的请求来调用不同的事务逻辑,并将处理结果返回到相关的页面. 这两层的功能是由 Struts2 框架实现的. 业务逻辑层是系统的核心层,它主要提供系统所需要的业务逻辑方法,完成系统的业务需求. 数据访问层主要是处理来自业务逻辑层的数据请求,提供对持久化对象的增、删、查、改操作. 这层的功能由 Hibernate 框架实现,通过框架用注解的方式实现了实体类与数据库表结构之间的映射,生成了可供操作的持久化对象,并通过 DAO 接口提供的操作方法来实现对底层数据的操作. Spring 框架工作在控制层、业务逻辑层和数据访问层,用它作为系统业务对象的工厂,集中的为系统配置和装配业务对象. 在这部分通过 IoC 技术和面向接口的编程方式实现了各层之间的"松耦合设计". 另外,为了减少应用程序对物理数据源访问的次数,提高系统的运行性能,对系统的用户关键词和行业敏感词这两个持久化对象进行了二级缓存的设计.

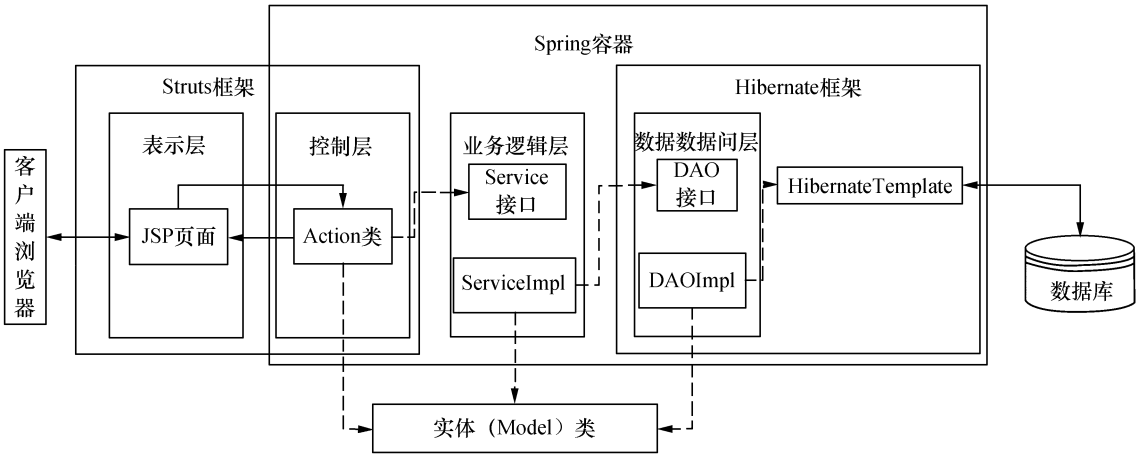

图 2　石鼓舆情监控系统体系架构图

Fig. 2　The architectures of Scool Public Opinion Monitoring System

总体上系统分为前后和后台两大模块. 前台用于对舆情记录以及热点事件进行展示以及对系统的用户信息进行维护;后台负责对所监控的目标网站的数据信息进行周期性的抓取并进行相关的分析.

1.2.1　后台设计

后台是一个面向目标网站的包含下载、解析、索引、过滤、自然语言处理等功能的模块,它的整体设计是基于数据缓冲池,多线程之上的,结构设计如图 3 所示.

由图 3 可知后台模块整体上分为数据缓冲池,线程池,数据库. 缓存池中的队列负责缓存数据,关键处理流程分拆到线程池中的各个线程中去实现,数据库负责存储经文档解析线程和查询过滤线程处理之后的数据,数据缓存池中数据的走向交由线程控制,各线程本身的并发运行交由线程池进行统一管理. 以上结构是基于生产者—消费者的二维队列实现的,各个线程和队列之间的关系和数据流走向如图 4 所示.

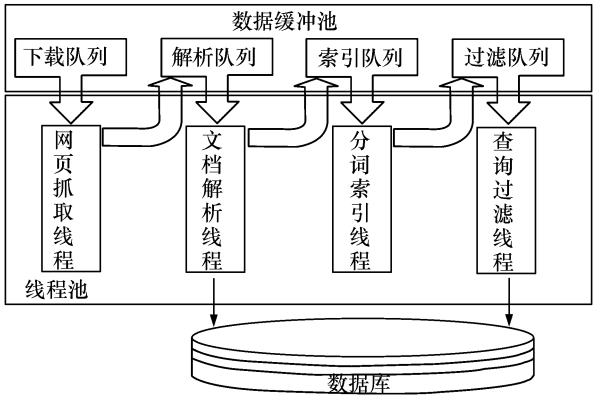

图 3　后台模块结构框图

Fig. 3　The background module structure diagram

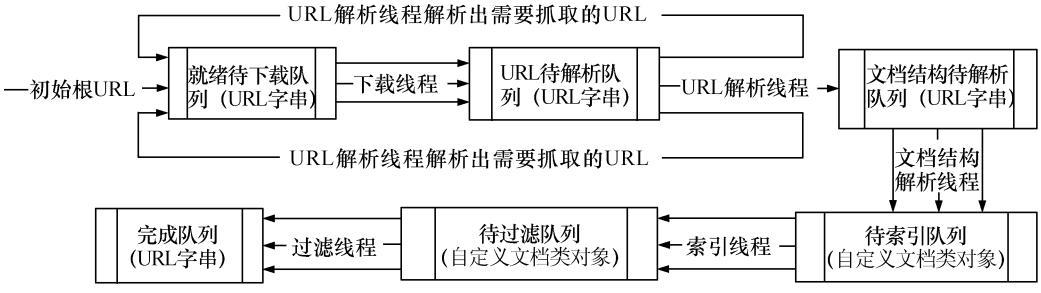

图 4 基于生产者—消费者的二维队列数据流图

Fig. 4 Two-dimensional queue data stream diagram based on producer/consumer

如图 4 所示:按处理前后顺序看有六个队列,依次是:就绪下载队列,队列中元素为待下载的目标 URL 标识字串;URL 待解析队列,队列中元素为已下载完 URL 标识字串,它等待进一步解析以获取其中链接进行进一步下载;文档结构待解析队列,队列中元素为已完成 URL 链接解析后待进行页面结构解析的 URL 标识字串;待索引队列,队列中元素为已完成文档结构解析后得到的系统定义文档对象;待过滤队列,队列中元素为已完成索引的文档对象;完成队列,队列中元素为已完成舆情过滤的页面 URL 标识. 队列的出入由生产者—消费者模式的多线程控制. 就绪待下载队列初始化时会加入要抓取的网站根链接 URL,一般是域名或者某版块地址,下载线程作为就绪待下载队列的消费者和 URL 待解析队列的生产者;URL 解析线程作为 URL 待解析队列的消费者和就绪待下载队列、文档结构解析队列的生产者;文档结构解析线程作为文档结构待解析队列的消费者和待索引队列的生产者;索引线程作为待索引队列的消费者和待过滤队列的生产者;过滤线程作为待过滤队列的消费者和完成队列的生产者. 除了完成队列,其余都是有界队列,当这个链任务完成后,完成队列的元素将会得到清理.

1.2.2 前台设计

前台模块主要是提供舆情信息展示、舆情搜索、舆情统计以及系统用户信息的维护等功能,设计框图如图 5 所示.

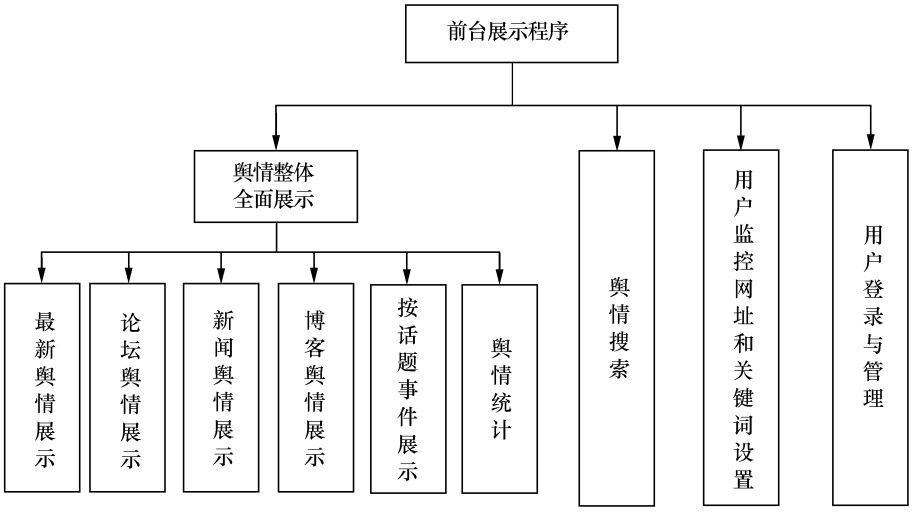

图 5 前台设计框图

Fig. 5 The design diagram of the system front module

1.3 系统实现

如图 6 所示,系统通过使用网络爬虫作为定时对信息采集的基本手段,将系统定制和由用户设置的新闻、论坛、博客等门户网站的数据信息采集下来,然后对网页进行解析和去重,定制系统的文档模型,通过信息过滤技术,将带有行业敏感词

和用户关键词的信息采集到数据库,针对与这些信息,一方面对其建立索引,另一方面将其作为事件聚类的数据来源,同时将当天最新过滤出的信息展现给用户. 其中用户关键词是由用户根据自己的关注点设定并可随时更新的,而系统敏感词库是专门针对高校舆情本身的特征所建立的,大学生作为高校的主体,校园里与之学习生活息息相关的一切事务都是其关注的焦点,另一方面大学生作为思想觉悟较高和爱国热情强烈的社会群体,对社会各种焦点热点问题和国内外的重大事件也有着极大的兴趣和关注度. 对信息的关注自然会引发情绪、意愿、态度和意见的产生,进而会产生相关舆情的主题. 通过文献资料的查阅和对近年来各大高校产生的重大事件进行梳理,并结合其相关的网络舆情信息,来建立敏感词库. 文献[10]把高校舆情产生的主题类归结为时事政治类、社会民生类、高校内部管理类、高校安全类、大学生权益类、高校声誉类和大学生心理健康类. 系统的敏感词也是围绕着这些主题类筛选的,具体明细如表 1 所示,并且系统提供对敏感词的管理模块,可以根据当前形式实时的更新敏感词库.

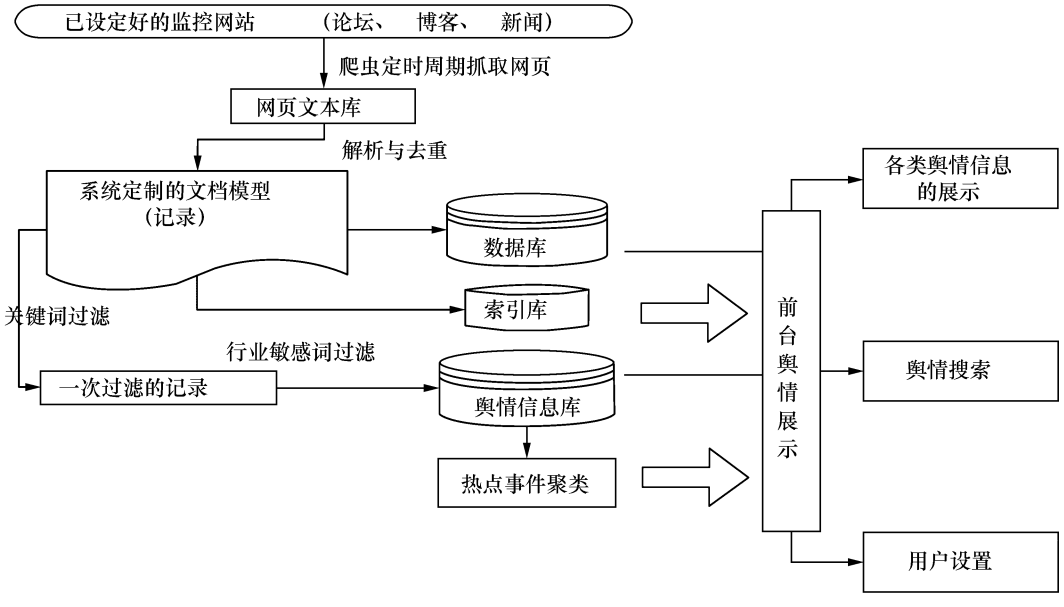

图 6　系统的实现流程图

Fig. 6　The system process diagram

表 1　石鼓网络舆情监控系统敏感词列表

Table 1　The sensitive word list of the scool network public opinion monitoring system

主题	敏感词
时事政治类	钓鱼岛、西藏问题
社会民生类	衡阳高校
高校内部管理类	收费、课改、裸考
高校安全类	中毒、疾病、着火
大学生权益类	后勤、网络、宿管、留级、伙食、医疗
高校声誉类	学术不端、贪污腐败、贿赂
大学生心理健康类	自杀、报复、跳楼、出走、猥亵

系统采用原型法开发,开发环境为 JDK1. 6 加上 MyEclipse9. 0,应用服务器为 Tomcate7. 0,使用开源数据库 MySQL 进行相关数据的存储,用于对普通高等院校舆情信息的监控.

2　系统测试及使用效果说明

由于功能测试可以直接反映系统是否能够正常运行并实现了在需求阶段所定义的各项功能[11],测试工作主要是针对系统功能模块的测试和系统的整体运行测试. 系统的功能一方面表现在后台模块对目标网页的周期性抓取并做分析,另一方面表现在系统用户在界面的交互操作中. 在本次测试中,采用手工测试的方式,以“南华”、“船山”、“衡阳高校”、“政府”、“大学教师”作为用户关键词,对系统进行为期一月的测试,时间段为 2013 - 04 - 11 00:00:00 ~ 2013 - 05 - 11 00:

00:00,在这段时间,系统总的采集记录为 3460 条,其中筛选出的舆情记录为 1311 条.

在测试阶段期间返回功能调整缺陷有 9 处,这些缺陷主要体现在脚本语言对各种浏览器的兼容问题. 现在均已修复,针对用户可以使用的各种功能和操作的测试都已通过,达到了预期的效果.

系统的两大重要工作是对目标网页的周期性抓取并做分析和对当天最新舆情记录的及时显示,因而我们就这两个方面对系统的使用效果进行简要的说明. 图 7 是系统针对红网相关版块的周期性抓取并分析的过程图.

图 8 是当天舆情记录显示页面,在这个页面系统会动态的将当天过滤出来的舆情记录显示出来,并将用户关键词和系统敏感词以高亮显示.

图 7 系统网页抓取分析效果图

Fig. 7 The effects diagram of crawling and analysing the webpages

图 8 系统当天舆情记录显示效果图

Fig. 8 The effects diagram of the show about the latest public opinion records

3 结 论

目前系统正处在试用的阶段,对南华人,红网,新浪湖南新闻等门户网站的一些版块进行监控,基本上实现了预期的设计要求,运行情况稳定,没有出现问题或者故障. 在与系统用户的交流反馈中也看到了系统的一些不足,其一,针对与系统所监控的目标网站,不同的网站对应与特定网页解析类,随着网站的不定期的改版网页结构,这种方式就得不到相应网页的内容,另一方面,用户可以随时更新自己的监控目标网站,这就要求开发人员必须实时的为用户设计对应目标网站的解析类,因而如何减少对于特定网页结构的依赖,设计一个通用的解析方案使其能够灵活的适应与用

户自己的监控网站的网页结构也是下一步的研究方向. 其二,就是系统没有把发言者的 IP 显示出来,不方便用户对恶意发言者的屏蔽和追踪. 其三,目前系统对数据信息的分析只是基于用户关键词和行业敏感词的匹配,这就导致由系统分析出的舆情信息数目较多,也没有突出具体的事件或者话题. 因而设计一个高效的文本聚类机制也是下一步的研究重点. 另外系统的界面也不够精美,影响用户体验,也需要进一步的完善.

下一步的主要工作就是针对大量的信息进行自动分类,将有关事件的 Web 信息资源进行聚类分析,全面掌握与各种热点事件的相关的信息,对事件进行追踪以及进行发展趋势的预测,帮助舆情工作人员时刻掌握事件的发展状况,并提供可靠的数据理论支撑.

参考文献:

[1] 曾润喜,徐晓林. 网络舆情突发事件预警系统、指标与机制[J]. 情报杂志,2009,28(11):51-54.

[2] 邓燕. 论网络舆情对高校群体性事件的影响[J]. 重庆邮电大学学报,2009,21(3):123-126.

[3] Pena-Shaff J B, Nicholls C. Analyzing student interactions and meaning construction in computer bulletin board discussions[J]. Computers & Education, 2004, 42(3):243-265.

[4] 刘德鹏. 互联网舆情监控分析系统的研究与实现[D]. 成都市:电子科技大学,2011.

[5] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking[C]//Proc. of SIGIR Conference on Research and Development in Information Retrieval, 1998:37-45.

[6] Allan J, Carbonell J, Doddington G, et al. Topic detection and tracking pliot study: final repot[M]. Proceedings of the DAPRA Broadcast News Transcription and Understanding Workshop. San Francisco, CA: Morgan Kaufmann Publishers, Inc, 1998:194-218.

[7] Yang Y, Carbonell J G, Brown R D, et al. Learning approaches for detecting and tracking new events[J]. Intelligent Systems and their Applications, IEEE, 1999, 14(4):32-43.

[8] 朱旭. 基于网格研判的高校群体突发事件预警平台的研究[D]. 北京:北京交通大学,2009.

[9] 叶昭晖,曾琼,李强. 基于搜索引擎的网络舆情监控系统设计与实现[J]. 广西大学学报(自然科学版),2011,36(1):302-307.

[10] 董亚倩. 高校网络舆情演变规律及安全评估指标体系构建研究[D]. 淄博:山东理工大学,2012.

[11] 李婷. 校园 BBS 舆情分析系统的设计与实现[D]. 武汉:华中科技大学,2009.