

网络舆情监测系统的分析与设计

洪小娟^{1,2}, 宗江燕¹, 于建坤¹, 黄卫东¹

(1.南京邮电大学管理学院, 江苏 南京 210003;

2.南京邮电大学江苏现代信息服务业决策咨询研究基地, 江苏 南京 210003)

摘要:大数据时代下,网络舆情监测对政府合理控制舆情方向、进行舆情治理具有重要意义。网络舆情监测系统主要根据网络舆情需求,构建在.NET平台下基于Entity Framework模型的网络舆情监测系统的C/S和B/S框架体系。系统对信息采集、信息过滤、关键词挖掘、舆情主题分类等模块进行优化分析设计和实现。应用马尔可夫模型,将舆情发展态势划分为生成期、发展期、极速发展期和衰退期,通过对历史数据的计算实现了舆情的未来发展区间。

关键词:舆情监测;网页文本分析;网络爬虫;舆情预警;预测

中图分类号: TP399 **文献标识码:** A

Analysis and Design of the Network Public Opinion Monitoring System

HONG Xiaojuan^{1,2}, ZONG Jiangyan¹, YU Jiankun¹, HUANG Weidong¹

(1. Management School, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. Jiangsu Information Service Decision Research Base, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Under the era of big data, network public opinion monitoring is of great significance for governments to reasonably control the public opinion direction and public opinion governance. To meet the requirements in the network public opinion, the network public opinion monitoring system mainly constructs the C/S and B/S framework system of the network public opinion monitoring system based on the Entity Framework model under the .NET platform. The system optimizes the analysis and design of modules such as information collection, information filtering, keyword mining, and lyric topic classification. The Markov model is applied to the system to divide the development situation of the public opinion into the generation period, the development period, the rapid development period and the recession period. The calculation of the historical data realizes the future development range of the public opinion.

Keywords: public opinion monitoring; web page text analysis; web crawler; public opinion warning; prediction

1 引言(Introduction)

互联网作为一种新的信息传播形式迅速发展,对人们的日常生活产生了巨大影响^[1],已然引起学术界的广泛关注,目前研究方向包括网络舆论的传播、控制及相关问题^[2,3]。据《中国互联网络发展统计报告》显示,截至2018年上半年,我国网民数量已达8.02亿人^[4],互联网已被公认为是继报纸、广播、电视之后能够反映社会舆情的重要载体之一^[5]。此外,超过六成的中国网民经常在网上就各种话题发表言论并进行讨论^[6],以充分表达自身的思想观点和利益诉求。

由于网络中的舆情在一定程度上能够代表现实世界中人们的观点,并且对现实世界的稳定产生一定的影响,因此如何才能监测和发现网络舆情,为政府或者企业提供决策上的

数据支持成为舆情问题研究的一个重要课题,本文旨在通过对网络爬虫、中文分词、信息存储方式、马尔可夫模型等的研究,来实现一个可以及时、准确的舆情监测和预测系统,为相关领域的工作人员提供数据上的支持。

2 网络舆情监测系统需求分析(Requirements analysis of the network public opinion monitoring system)

运用互联网平台进行交流具有匿名、及时、参与程度广、影响面宽、破坏面大等特点,这些特点给舆情监督的工作人员带来了极大的困难和挑战。因此,国内外普遍重视舆情监测关键技术的研究,目前,舆情监测涉及的技术非常多,其核心多为网络信息抓取技术、网络信息提取技术、自

然语言处理技术。其中，网络信息抓取技术多指利用网络爬虫工具进行信息抓取工作^[7]；网络信息提取技术则指将文本里的信息进行结构化处理，多被处理为表格形式^[8]；自然语言处理技术主要研究人与计算机交互的语言问题，通过分词、关键词提取等一系列操作对舆情状况进行分析，从而达到舆情监测的目的。

国内对网络舆情的监测起步较晚，又由于中西文的差异造成中文分词技术、文本挖掘技术等许多先进技术无法借鉴，同时研究机构与应用机构又严重脱节，直接导致我国的舆情产品比较昂贵同时发展也较为迟缓。不过，随着电子计算机和互联网络在中国的普及，网络舆情对社会生活的影响越来越大，网络舆情监控的产品也越来越多，其中以网智天元、北大方正、西盈信息、人民网舆情为代表的软件公司纷纷推出了自己的产品，并且都实现了24小时实时监控、关键词监测设置、舆情分析报告等基本功能。虽然国内的软件可以完成对网络舆情的监测任务，并且提供完整的分析报告，但是还没有提供对于舆情事件的趋势预测功能。

3 网络舆情监测系统概要设计(Overview of network public opinion monitoring system)

3.1 系统总体功能介绍

网络舆情监测系统主要由舆情数据处理系统和舆情趋势预测系统两个部分组成，在舆情数据处理系统中，又包括信息采集、信息处理、关键词提取等模块。系统的功能模块图如图1所示。

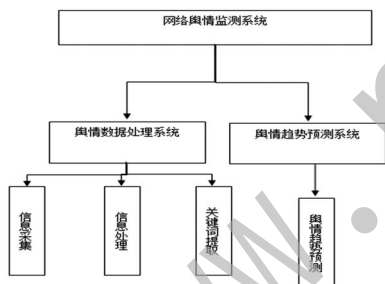


图1 系统功能模块图

Fig.1 System function module diagram

系统采用C/S和B/S相结合结构进行设计，C/S客户端主要负责自动采集指定网站上的信息，并且对信息进行过滤和分析，最后对信息进行存储；B/S系统主要用于数据展现、数据查询以及舆情趋势预测等功能，系统的架构图如图2所示。

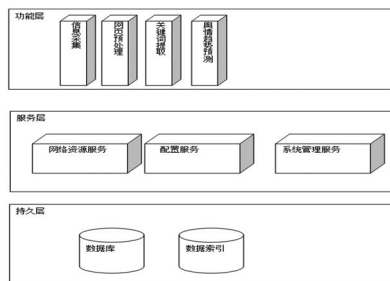


图2 系统架构图

Fig.2 System architecture diagram

3.2 系统数据流程图

系统的数据流来自采集模块，从自定义的采集网站中开始采集网络信息，采集的结果直接通过信息过滤模块，将用户所需信息从采集的网页源代码中分离出来，保存到本地数据库。随后，关键词提取模块读取采集信息的正文，利用中科院的ICTCLAS 2013版分词系统提取正文信息中的关键词及关键词在文本中权重，并将其保存到数据库中，系统的数据流程图如图3所示。

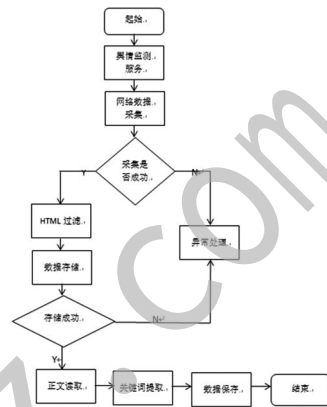


图3 系统数据流程图

Fig.3 System data flow chart

4 网络舆情监系统关键模块的设计与实现(Design and implementation of the key modules in the network public opinion monitoring system)

网络舆情监测系统主要分为两大部分，即舆情数据处理系统以及数据查询显示系统。其中舆情数据处理系统又包括信息采集、信息处理、趋势预测等三个主要模块。该系统主要用于数据的采集分析与处理，为用户提供有效的舆情信息。数据查询显示系统主要用于信息查询与趋势预测。该系统主要为用户提供舆情信息、舆情分析统计，以及舆情趋势判断等功能。其中，信息采集、信息处理、趋势预测也是网络舆情监测系统中的重点与难点。

信息采集模块的实现步骤如下：

Step1：在数据库中准备好需要抓取信息的网站的URL地址，并且配置好每个网站中的信息过滤规则。

Step2：根据用户预设的采集空间信息，初始化信息采集模型，并且通过该采集模型，对URL地址列表进行循环采集，下载当前页面的HTML代码到本地，然后由HTML解析模型进行数据解析。

Step3：在HTML解析过程中，将网页的HTML代码根据相应的网站信息过滤规则进行解析，并且将有效的信息保存到本地数据库中。

Step4：循环上述过程，24*7小时执行数据采集的过程，保持采集的数据的及时性。

信息处理模块的实现步骤如下：

Step1：系统读取数据库中为被处理的帖子的信息。

Step2: 利用中科院的ICTCLAS 2013分词系统, 提取出文本中的关键词, 以及关键词所占有的权重。

Step3: 将关键词信息保存到数据库中, 并且以此作为相应帖子的内容标签。

趋势预测模块的实现步骤如下:

Step1: 将热度趋势划分为四个状态区间, 大于0的分为两个区间, =急速上升, =缓慢上升, 小于0的也分为两个区间 =缓慢下降, =快速下降。

Step2: 统计出热度趋势值从当前状态到下一刻状态的数目。

Step3: 计算初始状态概率向量以及转移矩阵, 预测对象在任何一个时期处于任何一个状态的概率。

5 网络舆情监测系统的测试(Testing of the network public opinion monitoring system)

网络舆情监测系统的测试主要分为功能测试和性能测试。功能测试中, 主要对系统的UI界面操作, 以及查询显示功能进行测试, 确保系统能够给用户提供的简洁、准确的数据以及良好的用户体验。性能测试主要对系统的舆情主体分类和舆情趋势预测的准确性进行测试, 测试的结果表明, 本系统可以准确地对信息进行舆情事件的划分和对舆情趋势进行预测。

系统包括首页、舆情监测、舆情管理、舆情站点这四个栏目, 涵盖了系统需求部分所提及的所有功能。主页为用户提供了关键词搜索界面, 用户在搜索的文本框中, 输入想要监测的舆情的关键词, 并且选择想要监测的时间段和监测的网站范围, 就可以得到与关键词相关的帖子数量日均变化图、帖子热度日均变化图、帖子各站点比例图, 以及帖子在未来一段时间内的趋势预测, 搜索的结果页面如图4至图6所示。

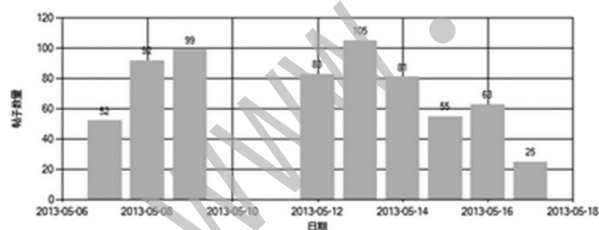


图4 帖子数量日均变化图

Fig.4 Daily change in the average number of posts

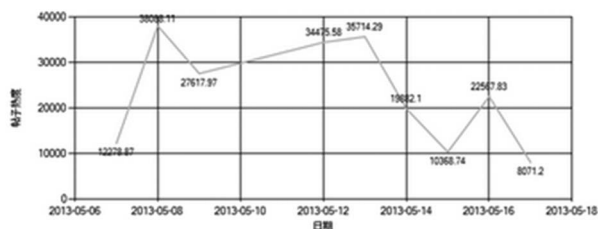


图5 帖子热度日均变化图

Fig.5 Post heat daily change chart



图6 帖子热度论坛对比图

Fig.6 Post heat forum comparison chart

6 结论(Conclusion)

综观本文的研究过程和结果, 存在以下几点不足与改进:

(1)在舆情趋势预测方面, 通过将马尔可夫模型运用到舆情监测的机制当中, 有效地对大区间内的舆情事件趋势进行了预测。

(2)在舆情数据获取方面, 系统所建立的规则不仅可以对单个论坛使用, 而是可以应用于多个论坛当中, 保证了系统所监测的论坛的普遍性。

(3)信息处理过程未详细研究, 只是借鉴了中科院的ICTCLAS分词系统, 对中文分词等过程及方法还需要进一步的分析和研究。

(4)由于时间仓促, 本文对于舆情信息倾向性分析, 时候评估分析等技术及应用未进行深入研究。

根据本文总结的改进之处和不足, 本课题展望如下:

通过实验论证, 系统可以进一步完善中文分词模块; 对于本文未深入研究的技术和应用, 将进行进一步的研究, 旨在全面提高网络舆情监测系统的准确性、合理性以及实用性, 为舆情监测领域提供优秀的技术平台。

参考文献(References)

- [1] ZHANG Le-jun, TONG Wang, JIN Zi-long, et al. The research on social networks public opinion propagation influence models and its controllability[J]. 中国通信, 2018, 15(07): 98-110.
- [2] WANG Qi-yao, JIN Yue-hui, ZHEN Lin, et al. Influence maximization in social networks under an Independent cascade-based model[J]. Physica A: Statistical Mechanics and Its Applications, 2016(444): 20-34.
- [3] FEI Xiong, YUN Liu, CHENG Jun-jun. Modeling and predicting opinion formation with trust propagation in online social networks[J]. Communications in Nonlinear Science and Numerical Simulation, 2017(44): 513-524.
- [4] 李朋朋, 李英武. “互联网+”背景下降低员工网络闲散行为的思考[J]. 决策探索(下), 2019(01): 87.
- [5] 尉译心. 网络舆情监控系统的关键技术[J]. 电子技术与软件工程, 2018(07): 26-27.
- [6] 杨华. 网络言论失范与政府管理[J]. 采写编, 2011(05): 54-56.
- [7] 胡亚楠. 社交网络数据获取技术与实现[D]. 哈尔滨工业大学, 2011.

(下转第13页)