网络舆情监控系统的实现方法

何　佳[1]，　周长胜[1]，　石显锋[2]

(1.北京信息科技大学 计算机学院　北京 100192;
2.中国矿业大学(北京)　机电与信息工程学院　北京 100083)

摘要: 结合信息通信技术,使用中文信息处理和文本挖掘中的关键技术对舆情监控设计流程进行分析.通过相关功能模块建立网络舆情监控系统,实现网页信息采集和 Web 挖掘基础上的热点发现与跟踪.

关键词: 网络舆情; 文本挖掘; 监控
中图分类号: TP 391.1　　文献标识码: A　　文章编号: 1671－6841(2010)01－0082－04

0　引言

据中国互联网络信息中心(CNNIC)统计数据显示:截至 2009 年 6 月 30 日,中国网民规模达到 3.38 亿人,普及率为 25.5%,网民规模依然保持快速增长趋势[1].目前,网络已成为信息的主要载体,互联网上存储和传输的信息能够在很大程度上反映一定时期社会各领域人们所关注的热点.鉴于网络中个人信息隐蔽性的特点,大量良莠不齐的信息充斥其中,在对社会发展产生积极作用的同时,也有大量的舆情信息不符合甚至违背社会的发展.网络舆情以"舆论多元"为最大特点,信息的数量极为庞大,且类别繁多,背景信息复杂.在这种情况下,要人工甄别每个意见的具体情况并加以分类统计是不现实的,只有采用计算机技术自动地对网络舆情语料进行分析整理,才能够建立起全面、有效、快速的网络舆情预警机制.Web 挖掘分为 Web 内容挖掘和 Web 使用记录挖掘[2],本文属于 Web 内容挖掘的研究领域,针对网络舆情监控系统的功能和设计过程进行研究,实现网页信息采集和 Web 挖掘基础上的热点发现与跟踪.

1　舆情监控

目前我国对舆情还没有形成统一的认识,王来华[3]对舆情进行了定义,即舆情是"舆论情况"的简称,是指在一定的社会空间内,围绕中介性社会事件的发生、发展和变化,民众对社会管理者产生和持有的社会政治态度.网络舆情是社会舆情的直接反映.

作为网上信息交流的主体,网民们对一些突发事件和社会流行所持的态度和发表的言论可能在短时间内对整个社会产生巨大影响,需要及时采取措施,实施舆情监控,以控制和引导事态发展.舆情监控是通过网页自动抓取舆情信息,通过文本挖掘等技术对舆情信息进行分析处理,再将处理后得到的热点信息上报跟踪来实现,舆情监控实现了对舆情信息由被动防堵向主动疏导的转化.

2　系统功能实现分析

近年来,国内外学者普遍关注网络舆情信息发现、热点信息跟踪机制等方面的关键技术.这一领域的研究涉及 2 方面理论:一是基于自然语言处理(NLP)技术;二是从数据挖掘的角度考虑热点信息的发现.这 2

收稿日期: 2009-12-17
基金项目: 北京市教育委员会科技发展计划项目,编号 KM200910772020.
作者简介: 何佳(1982－),女,硕士研究生,主要从事 Web 文本挖掘研究,E-mail: he_jia_2009@163.com;通讯联系人:周长胜(1961－),男,副教授,博士,主要从事数据挖掘研究,E-mail: wrz12@yahoo.cn.

个领域从不同的角度对舆情进行研究，同时又相互交叉和借鉴. 网络舆情监控系统用到的关键技术有信息检索、自然语言的识别和处理、文本分类和聚类、观点倾向性识别、主题检测与跟踪等. 整个系统结构功能模块如图 1 所示.

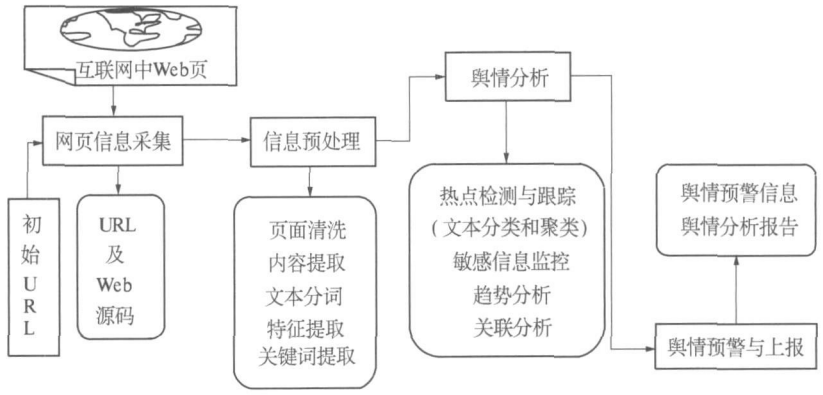

图 1　舆情监控系统结构功能模块

Fig. 1　The structure and function module of public opinion monitoring system

2.1　舆情信息采集

舆情信息采集模块是整个系统数据分析的信息源. Web 页面包含大量的数据信息，可以看作是一个巨大的数据库. 但由于 Web 页面的数据是半结构化或者非结构化的，加上 Web 页面极快的增长速度，其信息还在不断地发生更新，作为一个动态性极强的信息源，对 Web 页面进行信息采集是比较复杂的. Web 舆情信息采集流程如图 2 所示.

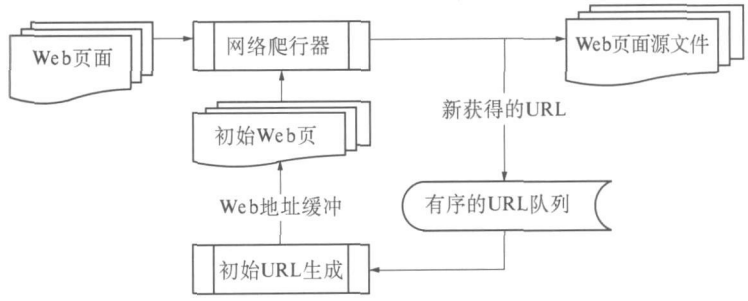

图 2　舆情信息采集流程

Fig. 2　Information collection module flow chart on public opinions

舆情信息采集是指对 Web 网页抓取和相关数据存储，网页抓取类似于搜索引擎中的页面爬行机器人. 首先，通过 Web 信息采集器，从一个初始集出发，将这些 URL 全部放到一个有序的待采集队列里[4]，然后按次序取出 URL，获取它所指向的页面，返回页面的 HTML 文件. 通过页面间的链接关系，获取新的页面的 URL，并将它们放到待采集的队列里. 重复上述过程，直到整个网站的全部网页都被采集完为止，也可以根据用户的需要下载一定层数的网页. 为了提高效率，系统可以设计几个信息采集器并行采集数据，即多线程地爬行多个网页并存储 Web 网页源码.

这种通用的网络爬虫的目标是尽可能多地采集信息页面，并不太在意页面采集的顺序和被采集页面的相关主题，这样消耗了很多的系统资源和网络带宽，但并没有换来采集页面的较高利用率. 为了解决这一问题，建议使用定向抓取相关网络资源的主题网络爬虫. 主题网络爬虫就是根据一定的网页分析算法过滤与主题无关的链接，保留与主题相关的链接并将其放入待抓取的 URL 队列中，按照事先给出的主题，分析超链接和已经下载的网页内容，预测下一个待抓取的 URL 以及当前网页的主题相关度，保证尽可能多的爬行，下载与主题相关的网页，尽可能少地下载无关网页. 当然，基于主题的网络爬虫的应用也提出了新的问题：如何定义感兴趣的主题，如何决定待爬行的 URL 的访问次序等，这也对实际的研究与应用提出了更深层次的思考.

2.2 信息预处理

与普通的文本文档相比,Web 文档包含了除正文以外的大量的其他信息,如广告链接、导航链接和版权信息等.与传统的数据库中结构化的数据相比,Web 文档中的数据结构极其复杂,因此计算机很难对抓取到的数据进行直接处理.信息预处理模块作为信息采集后的一个重要的模块,所做的工作包括:页面清洗与内容提取,文本分词,特征提取和关键词提取.舆情信息预处理模块的流程如图 3 所示.

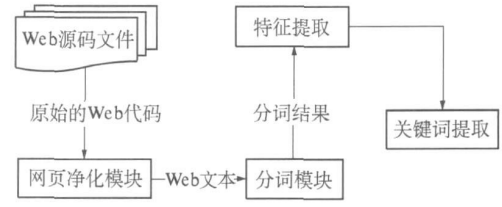

图 3 信息预处理模块流程

Fig. 3 Information pre-processing module flow chart

页面清洗与内容提取:由于 Web 页面不像传统的文本文档一样整齐干净,含有大量的噪声,同时 Web 页面在语义内聚性上难以得到保证,一个 Web 页面通常包括几个语义无关的部分,因此页面清洗对挖掘的效果具有重要意义.Web 页面清洗的目的是从页面中划分出更精确的信息单位,主要工作是:网页清洗模块对网页的注释、导航、广告、版权说明等噪声信息去噪.页面进行清洗之后,通过内容提取将半结构化的 Web 数据转化成具有模式的、可操作的信息.经过若干年的研究,学者们提出了很多种 Web 信息提取的方法,如基于规则的提取方法,基于应用本体的提取方法等.这些网页内容提取方法都可以通过 2 个步骤来实现:①网页逻辑结构(DOM 模型)的解析与表示;②指定元素的过滤与选择.

特征提取与关键词的提取:特征提取是一种从分词后的文档中提取有效和关键信息的方法,其目的是从噪音数据中分离出有用的信息以及减少数据的维数.常用的特征抽取的方法包括基于词性和词义[5] 的特征提取方法,这里介绍一种基于词性的特征提取的基本思想:首先,提取中文文本中的名词和动词作为文本的一级特征词,通过计算这些一级特征词的文本频数和文档频数来计算其权重.然后,根据各个特征词的权重,对这些一级特征词进行排序.给定一个阈值 K,在这些一级特征词中,选取 K 个权重较大的一级特征词作为文本的核心特征词,组成表示文本的特征向量.这里的文本频数是特征词在文本中的频数,文档频数是在训练库中,特征词在其中至少出现一次的文档的数目.

2.3 舆情分析模块

舆情分析模块是系统中最为关键的处理模块,利用文本分类和聚类等方法对预处理后的舆情素材信息进行分析、挖掘,实现舆情信息的热点发现和跟踪.舆情分析模块的流程如图 4 所示.

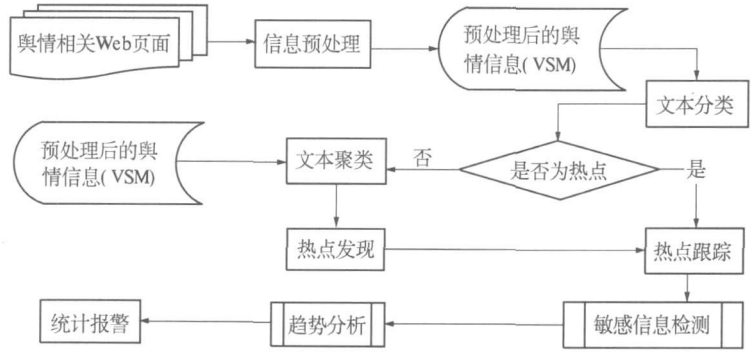

图 4 舆情分析功能模块流程

Fig. 4 Analysis module flow chart on public opinions

热点发现算法从本质上来说是属于数据挖掘中的文本聚类算法.算法的实现过程如下:将预处理后的文本信息归入不同的话题,并在需要的时候建立新的话题,热点发现的目的就是要按照话题将文档进行聚类,从一组文档集中发现新热点,由于没有关于新热点的先验知识,需要建立新的主题簇.

热点事件跟踪是为了用户能够跟踪自己所关心的类型事件而进行的操作,用户可以将已获得的事件的样本信息通过系统学习的方式交给系统,然后系统通过文本挖掘技术对不断到来的信息进行分类,判断是否为用户感兴趣的内容,将判断为是的信息交给用户.同时系统可以通过用户对获得的信息的反馈,不断地修正系统的学习结果,使得系统可以获得越来越接近用户所希望的信息.因此,热点事件跟踪是一种特殊的二

元分类问题.

敏感话题识别就是分析某个主题在不同的时间段内人们所关注的程度. 网络中的话题随着时间的推进, 以及某些相关事件的发生, 往往呈现出一定的波动和变化. 文献[6]中使用了观点对立度这一概念, 是指参与评论的网民之间评论倾向性的离散程度, 用户往往关注在过去某个时间段之内观点对立度上升较快的话题. 研究发现, 对于较小规模的话题, 即使其观点对立度在一段时间内上升较快, 但是由于参与的规模不大, 不能代表大多数网民的观点, 因此在进行敏感话题发现时需要考虑其规模因素, 结合主题关注度[7]的概念, 找出舆情在一段时间内的相关网页数, 只有同时满足规模和观点对立度 2 方面的阈值, 才能触发预警.

2.4　舆情预警与上报

舆情预警与上报是系统与用户的交互界面, 通过图表等方式将分析后的结果反馈给用户, 舆情展示包括热点排序显示、话题敏感性趋势变化显示等. 通过这些直观展示可以使用户对热点信息和敏感信息实施在线分析, 确定舆情在网站中的变化情况, 必要时系统将自动触发预警信息实施预警.

3　结束语

网络舆情监控是一个较新的研究领域, 实施网络舆情信息监控存在以下困难: ①网络上的信息源较多且网页结构复杂, 很难全面有效地收集到基于同一主题的所有信息; ②中文有其自身的特点, 基于中文信息处理和文本挖掘的热点发现算法有待进一步研究改进. 本文针对网络舆情监控系统的功能和设计过程做了一定的研究, 具体算法实施还需要进一步的分析与改进.

参考文献:

[1]　中国互联网络信息中心. 第 24 次中国互联网络发展状况统计报告[R]. 北京: CNNIC, 2009.

[2]　Eirinaki M, Vazirgiannis M. Web mining for Web personalization[J]. ACM Transactions on Internet Technology, 2003, 3(1): 12-13.

[3]　王来华. 舆情研究概论[M]. 天津: 天津社会科学院出版社, 2003.

[4]　刘尚喜, 蔡开裕, 卓琳. 内网舆情信息监测系统研究与设计[J]. 电脑应用技术, 2009(1): 32-33.

[5]　谢飞, 吴信东, 胡学钢, 等. 基于语义联系的新闻网页关键词提取[J]. 广西师范大学学报: 自然科学版, 2009, 27(1): 145-146.

[6]　张超. 文本倾向性分析在舆情监控系统中的应用研究[D]. 北京: 北京邮电大学, 2008.

[7]　钱爱兵. 基于主题的网络舆情分析模型及其实现[J]. 现代图书情报技术, 2008(4): 51-52.

Implementation Method for Network Public Opinion Monitoring System

HE Jia[1],　ZHOU Chang-sheng[1],　SHI Xian-feng[2]

(1. *School of Computer, Beijing Information Science & Technology University, Beijing* 100192, *China*; 2. *School of Mechanical Electronic & Information Engineering, China University of Mining & Technology, Beijing* 100083, *China*)

Abstract: Based on information communication technology, the Chinese information processing and the key technologies in text mining are used to analyze the design flow of public opinions on Internet. In order to establish network public opinion monitoring system, several function modules are designed. The system is built to search and trace the hot spot based on information collection of the Web page and text segmentation.

Key words: network public opinion; text mining; monitoring