·信息资源开发与利用·

网络舆情监测系统的研究与实现

邓凯英[1,2] 彭 超[1]
（1. 西北民族大学数学与计算机科学学院，甘肃 兰州 730030；
2. 东北师范大学地理科学学院，吉林 长春 130024）

〔摘 要〕网络舆情作为一种重要的舆情形式，具有形成速度快，受众人群广等特点，对国家和社会的影响越来越重大。互联网用户可以自由地在微博、论坛、博客等中发表有关社会中各类现实问题的态度和意见。监测网络舆情的主要手段就是利用网络爬虫对目标网络的页面数据进行挖掘，然后对挖掘的数据进行分类处理，并科学地统计舆情信息。本文主要分析网络舆情的特征和处理对策，并利用网络爬虫、全文检索、关键词评分、以及科学数理统计等手段对网络舆情监测系统的原理进行探索与系统实现。

〔关键词〕网络舆情；爬虫；关键字排名
DOI：10.3969/j.issn.1008－0821.2013.11.009
〔中图分类号〕TP301 〔文献标识码〕A 〔文章编号〕1008－0821（2013）11－0038－04

The Research and Implement of Network Public Opinion Monitoring System

Deng Kaiying[1,2] Peng Chao[1]
（1. Mathematics and Computer Science Institute, Northwest University for Nationalities, Lanzhou 730030, China;
2. School of Geographical Science, Northeast Normal University, Changchun 130024, China）

〔**Abstract**〕As an important form of network public opinion, fast generation, wide audience, network pubic opinion plays an increasingly key role in the nation and society. Internet users can freely present their opinion in micro－blogs, forums, and blogs. The primary means of monitoring the network public opinion is the use of a Web crawler page of the target network, data mining, classification, and scientific statistical information of public opinion information. This paper mainly focused on the features of the network public opinion and treatment measures. Using Web crawler, text search, explore and system the Keywords rating, as well as scientific and mathematical statistics, it explored the principle of network public opinion system and realize the system.

〔**Key words**〕network public opinion; spider; keyword ranking

舆情是民众关于现实社会中各种现象、问题所表达的政治信念、态度、意见和情绪的总和[1－2]。网络舆情信息是指社会民众通过互联网这一媒介所表达的情绪、态度、信念、意识、思想、意见、要求和行为方式等方面的综合表现，是对现代社会物质、政治、精神和社会 4 个文明建设活动的各种反映[3－5]《2012 年互联网舆情分析报告》蓝皮书指出，2012 年微博成为社会舆情的发动机，在本年网民重点关注的是社会转型、环境问题、钓鱼岛问题、南海问题等热点话题。据统计 2012 全年关于“钓鱼岛与反日游行”话题的网络博文合计17 742万篇，“伦敦奥运”话题博文7 583万篇，“神舟九号与天宫一号对接”的话题博文 3 923万篇。由此可见，网络舆情基本都是在短期爆发的，且影响范围广泛，都是些对国家、对社会意义深远的热门话题。因此，对网络舆情进行监测分析是十分必要的。

随着科技的发展，计算机技术的迅速普及与推广，网络为社会各阶层的人们提供了广阔、自由的交流平台[6]，

收稿日期：2013－07－12
基金项目：教育部人文社科青年基金项目（项目编号：12YJCZH027，13YJCZH029）规划基金项目（项目编号：11YJAZH053）、中央高校基本科研业务费专项资金项目（西北民族大学，项目编号：31920130007）。
作者简介：邓凯英（1982－），女，讲师，博士研究生，研究方向：智能信息处理。

互联网成为了社会网络舆情传播的主要平台。而网络舆情主要来自于 BBS、博客、微博、点评等，在网络中网民平等的表达着自己的观点，可以说真话，也可以说假话，言论相对自由，网络的开放性直接决定了网络舆情的直接性、突发性、偏差性。网络舆情的独立属性，信息流和环境会影响舆论的传播[7]。

网络舆情主要来自 BBS、博客、微博、点评等，在网络中网民平等的表达着自己的观点，言论相对自由，网络的开放性直接地决定了网络舆情的直接性、突发性和偏差性。

本文设计的网络舆情监测系统，主要考虑以下几个方面：

（1）对主流的社交网站、门户网站的网页、帖吧、文本文件、新闻评论、微博、博客等近期发布的信息，进行分类存储处理。

（2）对指定的网站上的近期信息，包括网页、帖吧、文本文件、新闻评论、微博、博客等数据进行采集与归类存储处理。

（3）对采集到的各种数据进行关键字分词处理，分词存储，分词评分，分词排名等处理。

（4）建设关键字检索系统，检索的结果按照标题与内容的综合评分进行合理的排序。

（5）本网络检测系统，采用合理的框架，预留更多未来开发的扩展接口，方便开源与二次开发。

1 系统的主要功能模块

1.1 网络舆情的采集模块

根据设置的检索条件，如限定域名的http://***.sina.com/***的所有页面的信息，将采集的数据适当的过滤，留下有用的关键数据。爬取的对象为重点新闻网站、知名社交网站、各大论坛，博客，以及政府网站等。

1.2 数据处理模块

对从网络上采集而来的数据进行处理，处理的手段包括：归类、分词、标注、加权、存储优化等。

1.3 关键字检索

为本网络舆情监控系统提供一个搜索引擎的功能，方便对网络蜘蛛爬取的数据进行查看管理。在一次检索的基础上，提供二次检索。提供智能的检索方案，按字索引、按词索引以及字词混合索引，对检索结果进行排名与统计。

1.4 舆情分析与统计

舆情分析是对舆情进行深层次的思维加工和分析研究。主要包括内容分析法和实证分析法。内容分析法对信息内容进行客观系统的定量分析，提示信息所含有的隐性情报内容，对事物发展做情报预测。实证分析法是通过分析大量案例和相关数据从而得出结论的一种研究方法。

经过分析后，可以自动提取关键字，提取一段完整的内容进行智能提取摘要，也可以根据已经设置的检索条件进行动态地提取摘要。对标题进行分词检索与排名。智能识别数据并归档到本地数据源。

网上数据的表示可以采用“点”与“线”组成的模型图，来表示互联网中的各类数据。用“线”来表示各个页面之间的 URL 链接关系，用“点”来表示网络中的各个页面。在这样的一个由点线组成的网状结构的图形中，每一个点与线都表达了非常重要的信息。所以互联网中的文本类型的数据可以简单的划分成由页面标题、页面的内容、页面的超文本标记以及页面之间的 URL 链接等构成。

一般的 HTML 页面由 Head 标签和 Body 表组成，主要的元素有标题 Title，表格 Table，层 div 等信息标签组成。然而每当用户浏览器收到数据时，去掉多媒体信息数据，如视频数据、flash 动画、图片数据、音频数据等非文本文件数据，其余的文本文件所包含的信息可以分为两类：一类是用于结构控制的 HTML 标签，HTML 由“〈”和“〉”构成一个标签，如〈div〉、〈head〉等标签；另外一类就是内容信息了，这些信息就提供给我们可以直接阅读的文字。也就是我们最终需要分词处理，存储处理的，建立索引的文本数据。

在页面设计的时候，为了方便搜索引擎搜录其页面的信息，通常会在页面添加关键字，在页面的〈head〉标签中，可以添加〈meta name=”关键字 1，关键字 2，关键字 3” content=”页面摘要描述……”〉的标签信息来描述本页面的主要信息，方便搜索引擎的网络蜘蛛爬取信息。

1.5 关键字高亮显示

在查询检索结果中，对关键字进行统计并高亮显示，虽然是一个小功能，但是技术实现的难度大，对用户体验有较高的提升，使得在检索结果中对关键的信息对用户一目了然。

1.6 网络舆情的预测

通过分析近期捕获的网络舆情，对这些数据进行自动分类，进一步聚类，并统计出各个关键字的数据图表，周期升降率，从而预测未来的舆情演化与趋势。

2 系统架构

本网络舆情监测系统采用 MVC 的设计模式。MVC 的全

称就是 Model View Controller 的缩写，意思为模型 model——视图 view——控制器 controller；MVC 是最常用的一种程序基本结构的设计，使用 MVC 架构可以使业务逻辑模块、数据链路模块、UI 界面模块具有良好的分层，这 3 个模块在具体的实现内容上彼此分离，在关系上又彼此调用，可以使各个模块的负责人集中精力编写各自的模块，只需要对彼此的调用关系提供接口，以便降低程序关系的耦合度，达到高内聚低耦合的目的，在 MVC 架构发展的近些年中，许多有经验的程序员习惯用 Java 的反射特性来更好地控制 UI 界面模块和业务逻辑模块的耦合性。

利用 MVC 独特的界面层、控制层、数据模型层的良好解耦的特点，本系统基本架构为：

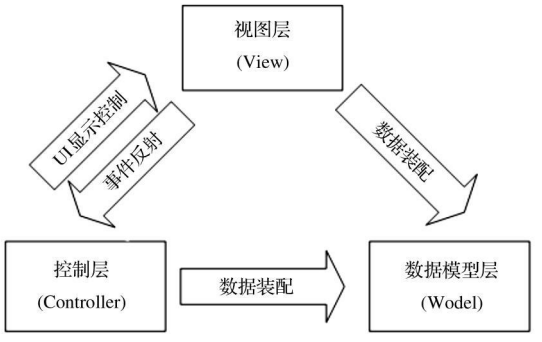


图 1 MVC 系统架构示意图

2.1 系统的功能架构

本网络舆情监测系统按功能模块划分，可大致划分为网络爬虫采集模块、中文分词系统、UI 界面管理模块、索引文件管理模块、内容搜索及搜索显示模块、中文全文检索系统、关键字智能评分系统、关键字高亮显示模块等八大模块。

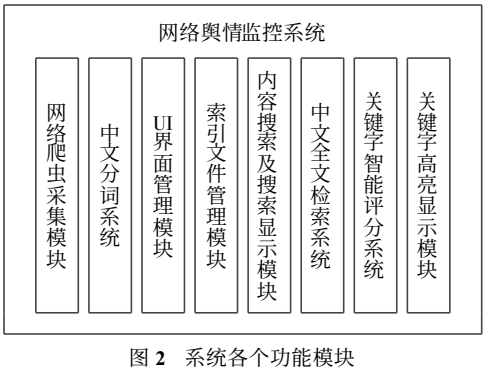


图 2 系统各个功能模块

2.1.1 网络爬虫

网络蜘蛛（Web Spider），也翻译为网络爬虫（Web Crawler），不管用其中的哪一个翻译都是一个非常形象的名称。其实，网络就好比蜘蛛网一样，上面有无数个节点，爬虫 Crawler 就好比是在网络中爬来爬去的一只虫子。网络蜘蛛在搜寻的网页中检索一个个超链接 URL，再对各个 URL 进行判断是否曾经检索过，如果没有，则通过该链接进行信息爬取，并且一直循环爬取，一直到把该网站所有的页面都爬取完为止。

2.1.2 中文分词系统

英文单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，惟独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂的多，困难的多。中文分词系统用于将一个又一个的单个汉字进行分词。一般中文分词是先判断前面和后面的几个汉字能否和本汉字组成为一个词语，并把前后连续的几个汉字，按照一定的顺序和语法进行重新排列或组合成为一个词序列的过程。中文分词最重要的是把最相关的结果排在最前面，这也称为相关度排序。

2.1.3 中文全文检索系统

中文全文检索是指把一个中文的文件中的全部的文本和检索项，进行全文式的匹配检索文本文件的方法。中文的全文检索可以把一个数据库或者一些文本文件，一个 Web 页面的内容进行全文查找检索。该系统还能分析文中的相关字、词、句、段、篇等内容，并带有统计功能，如果我们给一本书的每一个分词都加上一个分字标签，那么就可以统计分析全文的内容了。比如，我们要统计“中国名著《西游记》这本书中，‘孙悟空’一词在本书中共出现多少次”就可以通过这个检索方法实现。

2.1.4 UI 界面触发的事件反射到逻辑的处理

事件的反射处理是利用 Java 的反射原理将 View 层中的事件反射到逻辑中来执行，UI 响应反射事件时，需要通过事件动作配置数据 Relation.java 类，判断事件的类型，事件分为“无条件跳转”和“执行逻辑函数”两种类型。

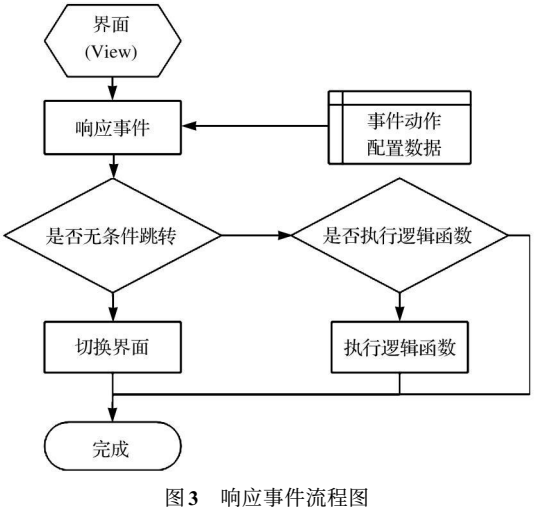
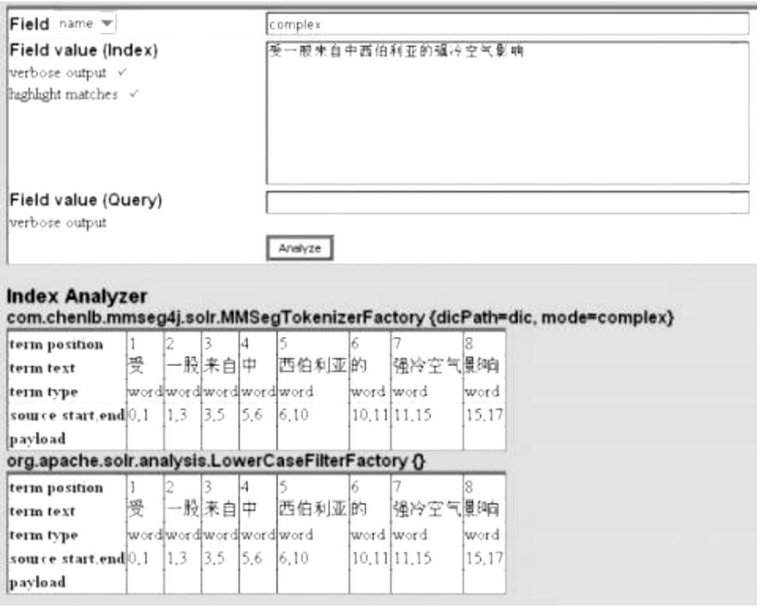


图 3 响应事件流程图

3 网络舆情的统计与分析

如果人工采集互联网上的信息，这个工作量将会是巨大的，因此需要研究如何在网络上进行自动实现信息采集，并及时的对采集来的信息进行处理，由人工采集信息的防拥塞，变为自动采集的自动归类，梳理，建立索引。



term position	1	2	3	4	5	6	7	8
term text	受	一股	来自	中	西伯利亚	的	强冷空气	影响
term type	word	word	word	word	word	word	word	word
source start,end	0,1	1,3	3,5	5,6	6,10	10,11	11,15	15,17
payload								

term position	1	2	3	4	5	6	7	8
term text	受	一股	来自	中	西伯利亚	的	强冷空气	影响
term type	word	word	word	word	word	word	word	word
source start,end	0,1	1,3	3,5	5,6	6,10	10,11	11,15	15,17
payload								

图 4 中文分词的输入输出

网络舆情分析系统是处理已采集信息的核心功能模块，具体功能如下：

（1）可以对热门话题与敏感词汇进行标识。

（2）可以根据新闻发布机构的权威度、回复数量、评论的频率，对信息进行评分加权，使得检索时排位靠前。

（3）可以识别出采集的信息在某一段时间内是否是最热门的话题，使用关键字的分词、排序、语法分析和语义分析，来辨别各类文章中是否包含敏感话题。

互联网页面上的数据不仅包括页面的内容数据，还含有一些 HTML 超文本标签主要用来对网页的结构进行设计。目前，部分国际化组织制定 HMTL5 协议对页面上的数据的格式进行统一的标记，但是这一类协议仅仅用于内容信息的表述形式上，这样做的原因是让浏览页面的用户能够更好地阅读页面信息。

4 结 论

本文在现有网络舆情研究的基础上，依据系统性、科学性、可靠性及可操作性原则，对如何采集监测网络舆情信息进行深入剖析，这有助于了解网络舆情发展规律，并据此设计了网络舆情监测系统，当然，该系统的功能还需进一步完善以便推广使用。

参 考 文 献

[1] 董亚倩，邓尚民. 基于社会网络分析的网络舆情主体挖掘研究[J]. 情报资料工作，2011，(6)：45－49.

[2] 石彭辉. 基于社会网络分析的网络舆情实证研究[J]. 现代情报，2013，33(2)：27－31.

[3] Xiao Qiang. The Rising Tide of Internet [R]. International Journalism Nieman Reports，2004：103－104.

[4] Guo Liang. The Internet is Changing China [EB/OL]. http://china.usc.edu/app－images/guo_liang.Pdf.

[5] 戴媛，姚飞. 基于网络舆情安全的信息挖掘及评估指标体系研究[J]. 情报理论与实践，2008，31(6)：873－876.

[6] 陈新杰，呼雨，兰月新. 网络舆情监测指标体系构建研究[J]. 现代情报，2012，32(5)：4－7.

[7] Suo Shuguang，Chen Yu. The Dynamics of Public Opinion in Complex Networks [J]. Journal of Artificial Societies and Social Simulation，2008，11(4)：2.

（本文责任编辑：王 涓）