网络舆情监测系统的设计与实现

刘小强，苟元琴

（三门峡职业技术学院 信息传媒学院，河南 三门峡 472000）

摘 要：针对当前网络舆情监测引导方面出现的一系列问题，本文设计实现了网络舆情监测系统，通过系统设计和应用，为地方政府及时高效的进行网络舆情监测分析、进一步做好网络形象构建与传播发挥积极的作用。

关键词：网络舆情；舆情监测；引导机制

中图分类号：TP 393.07 文献标志码：A 文章编号：1671-2153(2015)03-0065-04

0 引 言

伴随着信息技术的迅猛发展，互联网已成为广大人民群众抒发民意、表达愿望、参政议政的重要场所，也是政府职能部门收集民意、了解民情、监测互联网活动的重要场所。面对互联网上每天迅速增长的海量互联网信息和产生的网络舆情，人工方式已经远不能实现对互联网信息处理和网络舆情的监测[1]。因此，设计与实现网络舆情监测系统是目前迫切的需要，本文对此进行了研究。

1 系统的设计思路、目标及框架结构

1.1 基本思路

网络舆情监测系统能够为政府部门全面掌握网络舆情、争取处置主动权提供有效分析依据，实现政府部门对网络舆情监测和新闻专题追踪等需求[2,3]。本系统将利用整合互联网搜索技术及信息智能处理技术，通过对互联网海量信息自动抓取、自动分类聚类、热点发现和分析、专题聚焦等，形成简报、报告、图表等分析结果。

1.2 系统建设目标

网络舆情监测系统要能实现及时快捷的对互联网信息进行整合，包括网站新闻、微博言论、论坛帖子等，并运用先进的中文网站过滤技术[4]、信息处理技术、文本聚类技术、热点追踪挖掘等技术对互联网海量进行自动筛选获取、自动分析产生监测预警，从而实现对网络舆情信息的收集与监测，并最终将处理结果以图形、图表等多种方式显示出来，为政府主管职能部门全面掌握互联网上民众的思想动态、社会活动，及时有效的做出正确舆论引导，避免恶性事件发生，有利于维护社会稳定、构建和谐社会。本文设计的网络舆情监测系统包括三大功能模块，即：舆情数据采集模块，数据处理模块和舆情分析模块。

1.3 系统框架结构

本文所设计的网络舆情监测系统采用了面向对象的方法，实现手动进行信息采集、信息数据分析、数据索引建立以及舆情发现与跟踪等功能；用户使用时可自行设置运行参数，服务器按照设置要求定时运行，完成用户设定参数的信息采集、分析以及话题发现和追踪等功能并将运行结果进行存储，方便以后比较使用；数据处理完成后，系统

收稿日期：2015-03-18

基金项目：河南省教育厅科学技术研究重点项目(14B520043)

作者简介：刘小强(1982-)，男，陕西咸阳人，三门峡职业技术学院讲师，硕士，研究方向为计算机软件设计与开发、网络系统开发。

以图表图形等方式展现统计结果，具体系统框架结构如图 1 所示。

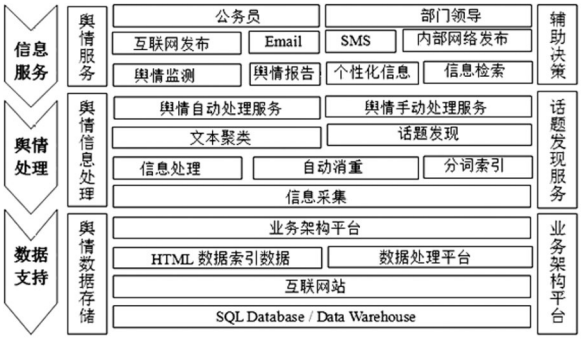

图 1　网络舆情监测系统的结构

2　系统建设方案

网络舆情监测系统的数据流来自系统数据采集模块，数据采集模块从设置定义的网站采集海量互联网数据信息，并将采集结果存储到服务器上；数据处理模块负责解析 HTML 文件，将服务器上的文件读取出来，对文件中的文本进行分词，同时把分词前后的语句及相关信息分别保存到数据库中[5]；舆情分析模块对保存到数据库中的文本信息进行分析，确定其事态的强弱程序，以此完成舆情发现和深度挖掘分析；舆情预警模块按照预先设定的报警机制读取数据库中形成的舆情信息，并以直观的图表方式将结果反馈给用户。

本系统主要分为三大模块，即：舆情数据采集模块、数据处理模块和舆情分析决策模块。其中数据处理模块又包含信息处理、话题发现两大功能模块。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。

3　网络舆情监测系统实现

3.1　舆情数据采集

舆情信息采集模块为整个舆情监测系统的基础模块，该模块主要完成了对指定数据采集空间内的信息资源进行采集与存储，该模块所采集的信息资源，将作为舆情分析的有效文本集合[6]。采集过程中运用网络蜘蛛技术实现多线程蜘蛛同时进行抓取，大大提高抓取效率。

在采集过程中，为了提高系统性能，系统采用了如下处理方案：

（1）系统采用文本数据库进行数据存储，降低了关系型数据库读写消耗；

（2）设置缓存区，对常用数据进行缓存，定期将过期信息写入文本数据库，降低输入输出读写消耗；

（3）采用多线程数据采集，以提高采集速度。

采集模块主要代码如下：

```
protected void search()
{
    DateTime start = DateTime.Now;
    _indexDirectory = Server.MapPath ("index");
//词库路径;
    dictPath = Server.MapPath("App_Data"); //词库路径
    highanalyzer = new LuceneNetAnalysis.Standard.StandardAnalyzer();
    _searcher = new IndexSearcher(_indexDirectory);
    Analyzer KTDanalyzer = new KTDictSegAnalyzer(dictPath);
    PerFieldAnalyzerWrapper wrapper = new PerFieldAnalyzerWrapper(highanalyzer);
    simpleAnalyzer = new WawaSimpleAnalyzer(); //按分隔符语汇单元化的分析器
    wrapper.AddAnalyzer("title",simpleAnalyzer);
    wrapperAddAnalyzer("content",simple Analyzer);
    string[] fields = new string[] { "title" };
    if (keyssidt == "2")
    { fields = new string[] { "content" }; }
    _mfqp = new MultiFieldQueryParser (fields, wrapper);
    string nowq1 = reqs(this.Query);
    if (nowq1.Length<2)
    return;
    q1 = _mfqp.Parse(reqs(this.Query));
    BooleanQuery m_BooleanQuery = new BooleanQuery();
    Query query1 = new TermQuery(new Lucene.Net.Index.Term("soid", keyssid)); //词语搜索
    m_BooleanQuery.Add(q1, BooleanClause.Occur.MUST);
    ……
    Repeater1.DataBind();
    DataBind();
}
```

采集模块主要结构如图 2 所示。

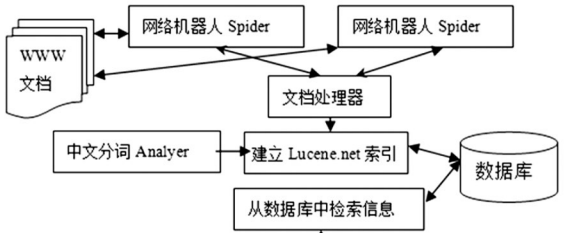

图 2　舆情数据采集结构

3.2　数据信息处理模块

数据信息处理模块实现分词与建立索引的功能。该模块主要对信息采集过程采集到的文本数据库进行读取,逐条进行数据清理,去除文本中的脚本等无用信息,提取出文本的标题、内容,并利用 Lucene.Net 对文本进行分词索引,为热点话题发现模块创建文本模型提供数据资源[6]。

分词索引功能的实现,主要是利用了采集回的 HTML 信息,进行信息处理,并对有效数据进行分词,建立索引[7]。

实现步骤如下:第一步,系统读取文本数据库,将每条数据的内容读入信息预处理模块,在信息预处理模块中,对文本内容进行分析,根据<title>标签及<body>标签等,分别获取文本的标题及内容;第二步,利用中文分词系统,建立本系统中使用的 Analyzer 及 Tokenizer 类,将分词器跟分析器进行合理的组合,使之产生对文本分词和过滤效果,同时将文本规则切分为一个个可以进入索引的最小单元;第三步,构造 indexWriter 对象,并将负责把索引文件写入存储介质,是控制逻辑存储转换为物理存储的纽带;第四步,建立索引,并对索引结构进行优化。信息处理模块主要代码如下:

```
public class HTMLDocParser
{    private String htmlPath;
     private HTMLParser htmlParser;
     public HTMLDocParser(String htmlPath)
     {    this.htmlPath=htmlPath;
          initHtmlParser();
     }
private void initHtmlParser()
{    InputStream inputStream=null;
     try
     {inputStream=new FileInputStream(htmlPath);}
     catch (FileNotFoundException e)
     {e.printStackTrace();}
     if(null! =inputStream)
     {    try
          {
           htmlParser=new HTMLParser (new InputStream Reader(inputStream,"utf-8"));
          catch (UnsupportedEncodingException e)
               { e.printStackTrace();}
          }
     }
……
}
```

信息处理模块主要结构如图 3 所示。

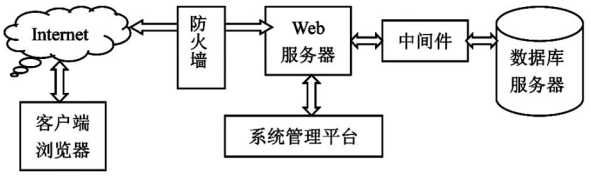

图 3　信息处理模块结构

3.3　舆情分析模块

使用 B/S 访问方式作为舆情分析模块的主要开发模式,更加有利于分析人员及时有效获取当前舆情信息,其实现步骤为:第一步,抽取阶段主要完成数据源数据的链接、数据访问等工作;第二步,清洗阶段则完成了对列属性的清洗和增补、对数据结构的清洗和增补以及对数据规则和业务规则的清洗和增补工作,并为下一步准备数据;第三步,一致性处理完成了维度表的建立、度量及性能指标的建立,去除重复数据,并为下一步准备数据;第四步,交付阶段则主要完成了维度表数据的加载及处理,并将处理好的数据加载到数据仓库,然后利用 Reporting Service 功能实现舆情信息的统计、分析。

4　结 论

本文针对网络舆情监测引导进行了系统开发研究,提出了舆情数据采集模块、数据处理模块和舆情分析模块相结合的设计思路,并设计实现了网络舆情监测系统。通过系统设计和应用,实现政府部门对网络舆情监测和新闻专题追踪等需求,形成简报、报告、图表等分析结果显示出来,从而更好的为地方政府开展网络舆情监测分析、做好

社会管理工作发挥积极的作用；并且在地方政府掌握网络舆情、引导网络舆情发展、构建地方政府良好形象方面提供了有效分析依据。

参考文献：

[1] 高洪杰. 互联网舆情监测分析系统实现 [D]. 上海:复旦大学,2009.
[2] 杨涛. 智能信息处理技术在互联网舆情分析中的应用 [D]. 上海:同济大学,2008.
[3] 潘文富,郭友实. 网络舆情监测技术研究综述[J]. 福建电脑,2011(8):39-41.
[4] 刘小强，廖建锋. WSN 中一种基于网格的并行数据收集方案[J]. 计算机应用与软件,2014(7):127-131.
[5] 张显江，刘小强. 一种参数联合优化的网络流量非线性预测模型[J]. 计算机工程与应用,2014(6):64-67.
[6] 于琨，孙新领. 基于信息挖掘的高校网络舆情监测系统开发[J]. 河南机电高等专科学校学报,2012(1):24-26.
[7] 刘小强. 二手转让及房产租售垂直搜索引擎的设计与实现[J]. 三门峡职业技术学院学报,2010(3):118-121.

Network Public Opinion Monitoring System Of Design and Implementation

LIU Xiao-qiang, GOU Yuan-qin
(Information and Media Institute, Sanmenxia Polytechnic, Sanmenxia 472000, China)

Abstract: Aiming at a series of problems in current network public opinion monitoring and guidance, this paper presents the design and implementation of network public opinion monitoring system, through the design and application for network public opinion monitoring system, analysis, further improve and communication play an active role to construct the network of local government image timely and efficient.

Key words: network public opinion; public opinion monitoring; guidance mechanism

(责任编辑:徐兴华)