doi:10.16652/j.issn.1004-373x.2017.24.005

基于大数据的网络舆情分析系统

谌志华
（中国软件与技术服务股份有限公司，北京　100081）

摘　要：针对互联网数据快速增长和舆情信息飞速传播的问题，提出一种基于大数据的网络舆情分析系统。该系统包括数据采集、预处理、分析和报告汇总四个模块，实现舆情信息的全网自动搜索与采集，大规模舆情数据的格式化存储以及舆情信息的分析、统计汇总等功能。该系统还使用Hadoop平台进行数据处理，并使用HDFS分布式文件系统存储舆情数据，使用MapReduce技术完成舆情分析和报告。仿真结果表明，该系统有助于及时、准确地分析网络舆情，能较好地满足网络舆情分析的需求。

关键词：大数据；网络舆情；舆情分析；Hadoop；HDFS；MapReduce
中图分类号：TN711-34；G206.3　　**文献标识码：**A　　**文章编号：**1004-373X(2017)24-0015-03

Network public opinion analysis system based on big data

SHEN Zhihua
(China National Software & Service Company Limited, Beijing 100081, China)

Abstract: In allusion to the rapid growth of Internet data and the rapid spread of public opinion information, a network public opinion analysis system based on big data is proposed. Four modules of data collection, preprocessing, analysis and report aggregation are included in the system to realize the automatic search and collection of the overall network public opinion information, the formatted storage of large-scale public opinion data, and the analysis and statistical summary of public opinion information. In the system, the Hadoop platform is used for data processing, the HDFS distributed file system is used to store public opinion data, and the MapReduce technology is used to complete public opinion analysis and report. The simulation results show that the system can help analyze network public opinion timely and accurately, and meet the requirement of network public opinion analysis well.

Keywords: big data; network public opinion; public opinion analysis; Hadoop; HDFS; MapReduce

0　引　言

目前，我国互联网普及率[1]已超过全球平均水平4.6个百分点，达到54.3%。网民规模占全球网民总数的1/5，达到7.51亿，并有超过70%的网民使用微博、博客等参与话题讨论并发表观点。互联网已逐渐成为热门话题和事件讨论的重要平台以及舆情事件的放大器[2-3]。

网络舆情[4]是指网络媒体或网民使用互联网对热门话题和事件进行讨论，所产生的具有一定倾向性与影响力的言论或意见，通常具有开放性、迅速性、丰富性、互动性和落地性等特点。虽然正面积极的舆情信息具有示范效应并能带来良好的社会影响力，然而消极负面的舆情信息将严重威胁社会的稳定和安全。因此，如何利用并控制网络舆情已成为相关管理部门与政府机关所关注的核心问题。

传统的舆情分析系统由舆情搜索和舆情分析两部分组成，并使用B/S模式将舆情分析系统分为功能层、数据访问层和业务逻辑层三层架构。其中，功能层用于响应用户的请求、展现请求结果和转发控制；数据访问层实现数据库的封装访问；业务逻辑层用于分离业务和逻辑。然而，当前互联网数据急剧增长，且具有价值巨大但密度低的特点，如何全面抓取信息，并及时、准确地分析网络舆情已成为当前网络舆情分析亟需解决的问题[5]。

本文针对互联网数据急剧增长和舆情信息传播速度快的问题，提出一种基于大数据的网络舆情分析系统，将大数据及数据挖掘技术应用到网络舆情分析中。该系统包括舆情信息采集、预处理、分析和报告四个模块，实现了全网自动搜索、采集舆情信息、大规模舆情数据的格式化存储以及舆情信息的分析、统计汇总等功能。

收稿日期：2017-09-07

1 网络舆情分析系统架构

本文将大数据和数据挖掘技术应用到网络舆情分析中，实现了基于大数据的网络舆情分析系统。该系统使用Hadoop平台进行数据处理，使用HDFS文件系统存储舆情数据，并使用MapReduce技术完成舆情分析。系统整体包括数据采集、预处理、分析和报告汇总四个模块，系统整体架构如图1所示。

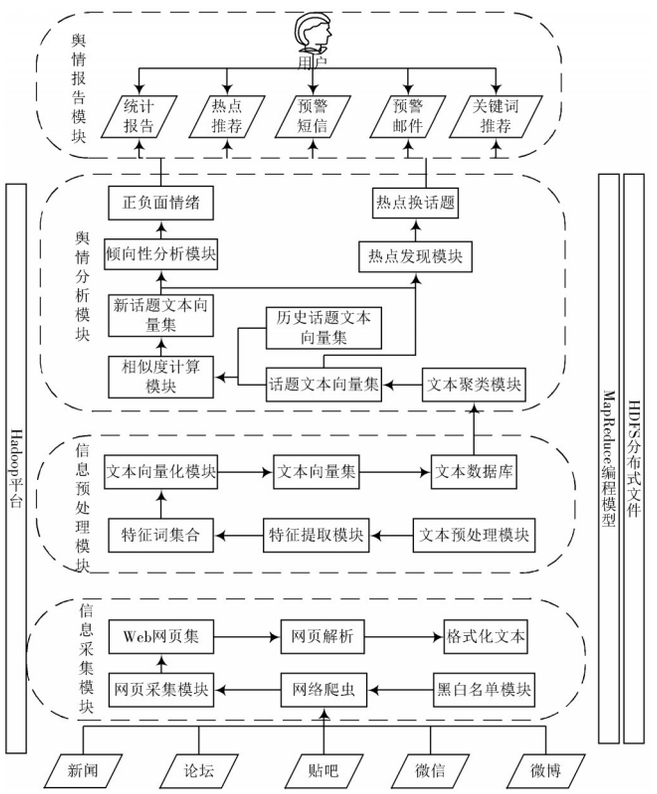

图1 系统整体架构

2 系统实现

2.1 数据采集模块

舆情数据采集模块是本文舆情分析系统的基础模块，主要负责使用网络爬虫从新闻、论坛、贴吧、微信和微博等Web页面采集舆情信息，具体流程如图2所示。

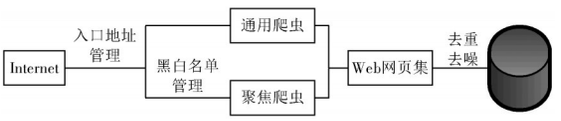

图2 数据采集流程

基于大数据的舆情分析系统不仅需要使用传统搜索引擎爬虫保证所下载网页的全面性，且还需要使用聚焦爬虫保证所采集信息的精确性。通过设置黑白名单，保留有用的URL链接，并依据确定的搜索策略重复搜索，直至达到停止条件。在抓取Web信息时，主要采集网页的文章内容和版块列表两种信息。其中，文章内容采集即通过分析网页的HTML源码抓取和保存网页内容，版块列表采集即通过确定初始网页的URL、设定爬行深度、制定爬行参数和采集规则等操作抓取初始网页源文件[6]。

2.2 预处理模块

舆情信息预处理模块是本文舆情分析系统的数据准备阶段，该模块先将采集到的各种网页信息进行去重、去噪等预处理。然后，选择文本特征并格式化为文本向量，最终得到文本向量集。其工作流程如图3所示。

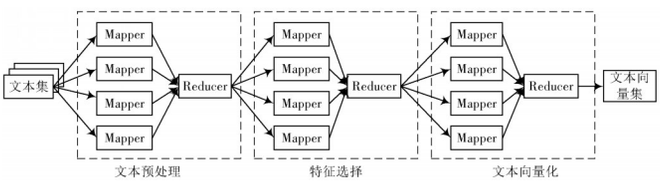

图3 预处理模块流程

由于新闻、论坛和微博等的网页结构各不相同，因此需要清洗与文本无关的HTML源码，并保留网页标题、内容摘要、发布时间以及评论等与舆情相关的信息。过滤掉无意义或重复的网页信息后，为了避免噪声干扰并保证数据的完整性需要剔除或填补缺失数据。

为了便于后续的文本分析，本系统使用MapReduce技术和分词工具并行处理格式化文本，提取词频特征，构造文本向量集。同时，将其保存到HDFS分布式文件系统中。

2.3 舆情分析模块

舆情分析模块是本文舆情分析系统的核心模块，主要完成识别、跟踪舆情话题和评估舆情情感，其具体工作流程如图4所示。

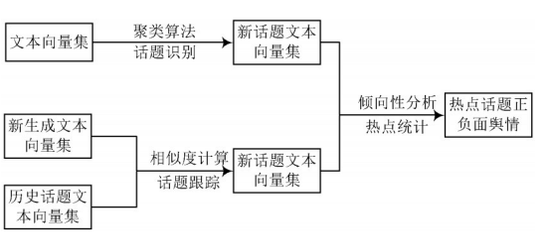

图4 舆情分析模块工作流程

舆情分析模块先使用聚类算法将预处理模块得到的文本向量集进行汇总，并识别出主要舆情话题；然后检测后续更新的向量化文本，判断其与已存在的话题的相关性，如果相关性达到一定的阈值则将其归类到该话题中；最后分析各话题的情感倾向性。

本系统使用Hadoop平台Mahout机器学习库中MapReduce的K-means算法实现文本聚类[7-8]。只需要输入文本向量集、聚类中心数和迭代终止条件即可得到归类

文件及中心点。其中,Map函数将文本向量集划分为小块并发送到各子节点的执行程序中,并行执行计算任务,计算得到键值对形式的中间结果后传递给Reduce服务器;Reduce汇总各子节点的结果,并求和平均后得到聚类中心。

2.4 舆情报告模块

为了满足不同用户的需求,本系统使用舆情报告模块自动推送舆情热点、统计汇总相关内容、关键词推荐和辅助采编。当某一热点或负面舆情达到预先设定的报警阈值后,舆情报告模块可使用邮件、短信等方式通知检测人员。

3 实验与结果分析

基于大数据的舆情分析系统使用1台交换机和6台普通PC机来搭建Hadoop集群,分别在6台PC机上安装Ubuntu 16.04系统,并设置1台Maste服务器和5台Slave服务器。

为了验证本文提出的基于大数据技术的文本预处理效率,使用一份160 MB的预料文档在不同规模的集群中运行预处理程序,得到如表1所示的实验结果。

表1 不同节点数目下的文本预处理时间和加速比

节点数	时间/s	加速比
0	567.8	—
1	642.6	0.87
2	335.4	1.07
3	240.3	2.43
4	201.3	2.98

从表1可以看出,增加节点的数目可以加快预处理的速度,表明节点数越多,任务分块数越多,具有更高的并发运行程度。同时,加速比并不与节点数成正比,这是因为节点数增加,节点间的通信所消费的时间也在增加,从而影响了系统并行运行的效率。

如图5所示为文本预处理、特征提取和向量化三步骤的加速比对比。从图5可以看出,文本向量化的加速比较小,原因是在计算词频时启动各子任务需要占用一定的系统开销。而特征选择将计算分配在Mapper中并行执行,故具有较大的加速比。

综上所述,基于大数据的舆情分析系统使用分布式并行化处理技术,能大幅提高舆情分析的速度和数据处理能力。

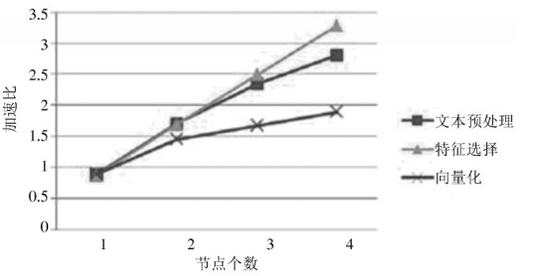

图5 文本预处理、特征提取和向量化三步骤的加速比对比结果

4 结 语

互联网数据快速增长和舆情信息飞速传播给舆情分析带来了较大的挑战,本文使用分布式并行化处理技术,提出一种基于大数据的网络舆情分析系统。该系统实现了舆情信息的全网自动搜索和采集,大规模舆情数据的格式化存储以及舆情信息的分析、统计汇总等功能。仿真结果表明,该系统有助于及时、准确地分析网络舆情,能较好地满足网络舆情分析的需求。

参 考 文 献

[1] 周红福,贾璐,张婷婷,等.微博舆情分析中信息转发路径提取方法研究[J].信息网络安全,2016(4):61-68.
[2] 张昕,孙江辉.舆情监测系统设计[J].现代电子技术,2015,38(11):98-102.
[3] 马梅,刘东苏,李慧.基于大数据的网络舆情分析系统模型研究[J].情报科学,2016,36(3):25-28.
[4] 孙彬,王东.微信息舆情的主动介入导引模式[J].沈阳工业大学学报,2016,38(5):584-589.
[5] 宫泽林,徐艳红.大数据时代网络舆情分析与研究[J].黑龙江科技信息,2016(17):169-169.
[6] 冯登国,张敏,李昊.大数据安全与隐私保护[J].计算机学报,2014,37(1):246-258.
[7] 苏毅娟,邓振云,程德波,等.大数据下的快速KNN分类算法[J].计算机应用研究,2016,33(4):1003-1006.
[8] 刘若冰.面向大数据云存储系统的关键技术研究[J].现代电子技术,2016,39(6):21-24.

作者简介:谌志华(1971—),男,江西南昌人,高级工程师,硕士。研究方向为计算机应用。