学校编码：10384　　分类号_____密级_____

学号：X2012231191　　UDC _____

厦门大学

工 程 硕 士 学 位 论 文

基于云平台的网络舆情监测系统的分析与设计

Analysis and Design of Network Public Opinion Monitoring System Based on Cloud Platform

张永光

指 导 教 师： 林 坤 辉 教 授

专 业 名 称： 软 件 工 程

论文提交日期： 2014 年 10 月

论文答辩日期： 2014 年 10 月

学位授予日期： 年 月

指 导 教 师：_____

答辩委员会主席：_____

2014 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

随着移动互联网技术的发展，网络传播也随之兴起，以往在传统新闻媒介上无法实现的个人表达自由的言论得到空前的发展，而这种言论往往具有突发性、非理性等特点，因此对网络舆情监测系统及相关技术进行研究设计具有很好的实际意义。

网络舆情监测需要对网络舆情进行采集、分析、提取并保存，以供查询、展示及预测等，数据抓取、数据解析、云存储是其中的三个关键技术。为了有效进行网络舆情监测，本文对以上三方面的关键技术展开研究，主要创新内容包括：

1．设计网络信息获取过程中的各种关键算法，包括：分布式 bloom 过滤器设计，解决爬虫系统的 url 重复爬取问题；高效 http 代理算法设计，解决网站对 ip 限制的问题；基于双消息队列的 P2P 分布式网页下载算法设计，实时响应各种请求，可有效的突破网站限制等。

2．设计网络信息预处理的各种关键算法，包括：一种海量 web 数据模式发现和数据结构提取，解决了网页模板自适应的技术瓶颈；采用“基于 DOM 树的可适应性信息抽取算法”，结合并应用“重复信息块路径集法”、“样本页面集信息块路径归纳学习算法”、“Web 信息规则抽取学习算法”等多种技术手段，具有自动化程度高、适应性和学习能力强的优点；

3．设计一种基于云平台的存储架构，解决海量数据存储的问题。采用云计算架构，对各种异构数据进行统一管理，能够有效防止单点故障，确保数据的可用性；同时可随意插拔节点，高度可扩展,以普通 PC 当数据节点，具有低功耗、低成本等优点。

最后基于本研究所提的理论及方法，开发一套基于云平台的网络舆情监测系统，并集成应用于厦门超算中心取证云平台中的互联网舆情分析与控制模块。

关键词：网络舆情；网络爬虫；云计算

Abstract

With the development of mobile internet, network communication is popular and people can exchange opinions with each other easily. Unfortunately, some of these opinions always seems to be unexpected and irrational. Thus, it is necessary to do some researches on how to detect network public opinions.

In order to view and forecast network public opinions, it needs to collect all kinds of network information and extract the network public opinions, and save them somewhere at last. During the procedure, network data capturing, data analysis and cloud storing are the three key technologies. To detect network public opinions effectively, this paper carries out some researches on the three technologies.

Firstly, this thesis design some key algorithms for network data capturing. Distributed bloom filter algorithm is to exclude the some url out of network spider system. Effective http agency algorithm is to solve ip resource limited on some websites. P2P website download algorithm based on double message queue response to various spider requests and so on.

Secondly, this thesis design some key algorithms for network data analysis. Huge web data pattern discovery and data structure extracting algorithm is to solve the self-adaption problem of webpage template. It makes up of self-adaption data extracting algorithm based on DOM tree, duplicated-data path set algorithm, data path of sample webpages inductive learning algorithm and web information extracting learning algorithm and so on. All these algorithms are self-adaption and learning-ability.

Finally, this thesis design a storage system based on cloud computing platform to store huge data of network public opinion. It manages all kinds of heterogeneous data and is immune single point of failure. Besides, each data node in this system can be removed or inserted optionally. And general PCs can be inserted as a data node into the system, it reduces the cost of this system.

At the end, this thesis design a network public opinion detecting system based on cloud platform which is based on all the algorithms designed in this paper, and

integrate it into the evidence cloud platform in Xiamen supercomputer center.

Keywords: Network Public Opinion; Network Spider; Cloud Computing

目 录

Contents

第一章　绪论

随着移动互联网技术的发展，网络舆情得到空前的发展，但其往往具有突发性、非理性等特点，因此对网络舆情监测系统及相关技术进行研究设计，有利于网络舆情是疏导，具有很好的实际意义。

1.1　研究背景与意义

随着互联网技术、特别是移动互联网技术的发展，网络传播也随之兴起，以往在传统新闻媒介上无法实现的个人表达自由的言论得到空前的发展。由于网络本身的虚拟性、隐蔽性、发散性、渗透性以及随意性等诸多特点，更多的人愿意采用网络这种渠道对自己关心的、与自身利益相关的某一焦点或者热点问题发表具有倾向性的言论，而这种言论往往具有突发性、多变性、情绪化、非理性等特点[1]；近年来由网络舆情引爆的政府形象危机越来越多，地方政府也相应面临着由此带来的新挑战，研究如何疏导和解决网络舆情危机问题，如何维护地方政府的形象，督促其加强社会公共管理，都具有非凡的意义：

1. 理论指导。随着移动互联网的发展，越来越多的人通过网络关注社会热点问题，从而引发网络舆情事件。许多研究人员对社会管理者在网络舆情监管中的角色作用等展开了研究，同时也发表了许多应对理论。
2. 现实参考。网络的普及，使得网络晋升为四大媒体之一，其特有的优势是民众可以参与到其中，这也是其迅速发展的原动力。然而，由于引导及监管的缺失，网络舆论也会引发一系列的社会问题，如何构建舆情信息化管理平台，实现网络舆情信息的收集、管理以及疏导，促进社会的和谐发展，具有非凡的现实参考意义。

网络舆情理论研究是新兴的社会科学与自然科学交叉的研究领域，收到众多学者的积极关注，网络舆情监测是指利用互联网技术对上述网络舆情的一种监视和预测行为。由于互联网舆情技术是以信息内容安全技术为研究和应用基础，而美国、法国、以色列、英国、丹麦等海外国家的信息内容安全市场和产品已步入成熟期；虽然国内近年来也取得了一定的成绩，但从整体技术发展方面来看，仍然处于跟随及模仿阶段，而且缺乏创新性的核心技术，因此，从单

纯的技术角度来看，我国互联网舆情技术的研究水平同国外相比，仍具有一定的差距[2]。随着网络民众数量的增加以及网络规模的不断扩大，网络舆情的规模变得空前庞大，需要处理的信息量也呈海量式增长，因此研究一套基于云平台、高可用、高扩展性、高效实时的网络舆情监测系统，以满足日益增大的舆情监测需求，了解及引导网络民意，维护社会的和谐稳定，具有非凡的意义。

1.2 国内外研究现状

国外在网络舆情监控方面的研究开展较早，且技术积累厚重，其在网络舆情领域的研究远优于国内，其中比较重要的舆情相关会议及论坛包括[3]：Topic Detection and Tracking、Special Interest Group on Information Retrieval 以及 Text Retrieval Conference 等；而比较权威的网络舆情研究机构包括：Canterbury 大学欧洲舆情研究中心[4]、美国舆情研究协会[5]及欧盟舆情分析官网[6]等，并形成了调查问卷、文本数据自动分析及 Web 数据自动分析等三种类型的网络舆情分析系统，其中 Web 数据自动分析型[7-16]是现今及今后的主流研究方向,目前国内大部分的分析系统大多基于该类型。

国内关于网络舆情研究方向可分为两方面：一方面是基于社会心理学、人文教育学等理论方面的研究[17-22]，为网络舆情的控制和疏导其他理论基础；另外一方面是基于信息技术的智能系统研究[23-26]，为网络舆情的监测及分析提供数据基础；这两方面的完美结合能够有效引导级及疏通由网络舆情所引发的社会问题等。

基于信息技术的智能系统研究又可以分为 3 种类型：

1. 关于网络舆情研判指标体系的研究：如网络舆情监测及预警指标体系构建[25]、网络舆情灰色预警评价及模式识别[27, 28]、我国网络舆情安全评估指标体系的构建研究[22]等研判系统的研究；
2. 关于网络舆情预测与预警模型、运算的研究：如基于贝叶斯网络建模的非常规危机事件网络舆情预警研究[16] 、基于直觉模糊推理的网络舆情预警方法倡[29]、基于数据挖掘技术的网络舆情智能监测与引导平台设计研究[30]等挖掘算法研究；
3. 关于网络舆情的信息系统建设的研究：如Goonie 互联网舆情监控系统

[31]、方正智思互联网舆情监控系统[32]、乐思舆情监测系统[33]、军犬网络舆情监控系统[34]等数据系统的研究；

信息系统的建设研究为网络舆情提供必要的数据基础，而预测及预警、挖掘算法模型为网络舆情的事态发展提供必要的参考，研判模型则为网络舆情提供了严格的风险评估，这三者是相辅相成的，缺一不可；然而目前的学术研究大多集中在预警及研判这 2 个环节，企业更多的侧重于网络舆情数据的收集及提取。鉴于目前的研究现状，本文的研究基于信息技术的智能系统研究，侧重点放在网络舆情数据的收集及提取，能够高效地为网络舆情的监测及分析提供完备的数据基础，为学术研究提供必要的技术及算法支持，特别是网络舆情数据爆炸的云计算时代。

1.3 关键问题及主要研究内容

由前所述可知，数据是一切算法及模型的基础，因此网络舆情数据的收集是网络舆情监测过程中的重中之重；该过程需要对网络舆情进行采集、分析、提取并保存，以供查询、展示及预测等，其最终的目的就是高效的提取舆情数据及有效存储，那么就需要解决以下几个关键问题：

首先需要解决的问题是如何快速又准确地抓取感兴趣的舆情数据；在以往的互联网时代，数据规模较小，系统设计几乎偏简单化。对于数据量暴增的云计算时代，如何设计一套高性能的数据抓取系统，是本文首要的关键问题；

其次，舆情数据可来源于论坛、微博、网页等，格式多样，因此需要进行格式化解析工作。因此，如何设计自适应数据处理算法，以解析各种类型的网络数据并进行格式化，是本文的第二个关键问题

最后，解析后的数据需进行固化存储，才能供后续的查询展示等。因此，对于抓取的海量数据，如何设计海量的存储平台，并具有高可扩展、高可用、低成本等特性，是本文的最后一个关键问题；

针对以上的关键问题，本论文从以下几方面展开研究：

1. 网络信息获取过程中的关键算法设计，包括：分布式bloom过滤器设计，提供了一种自适应、高性能、高扩展的去重机制，解决爬虫子系统的url重复爬去问题；高效http代理算法，则充分利用了有限的代理资源，解决网站对

ip限制的问题，使得爬虫子系统能够在尽量断的时间间隔内访问对应网站；基于双消息队列的的P2P分布式网页下载算法，提供了一种高度并行的网页下载机制，实时响应各种请求，可有效的突破网站限制；以上算法系统可扩展性强，适合大规模于大规模的分布式网页爬虫系统应用。

2. 网络信息预处理中的关键算法设计，以处理海量web数据模式发现和数据结构提取问题，解决了传统方法依赖特定网页模板、无法适用多变web信息形式的技术瓶颈，包括：页面去噪算法提供了页面解析时去除噪声的关键算法及处理流程；采用“基于DOM 树的可适应性信息抽取算法”，结合并应用“重复信息块路径集法”、“样本页面集信息块路径归纳学习算法”、“Web 信息规则抽取学习算法”等多种技术手段，实现正文信息的定位；模板变化自适应算法则解决系统同各种格式页面的耦合性问题，具有自动化程度高、适应性和学习能力强的优点
3. 基于云平台的存储架构设计，解决海量数据存储的问题。包括：并行索引生产引擎设计，解决传统索引数据生成效率低下及扩展性差等问题；采用云计算架构，对各种异构数据进行统一管理，能够有效防止单点故障，确保数据的可用性；同时可随意插拔节点，高度可扩展，几乎没有存储容量限制；普通PC可当数据节点，代替传统服务器，具有低功耗、低成本等优点。

1.4 论文结构安排

本论文共包括七章：

第一章 综述，对系统研究的背景和意义，以及现状，研究内容作了简单的介绍。

第二章 介绍网络舆情监测的相关原理，主要包括舆情定义、舆情特征、舆情发展态势等网络舆情理论基础；爬虫技术原理及工作流程等信息获取技术基础；网页去噪、文本分词、文本聚类等信息预处理技术基础以及云储存相关技术基础。

第三章 介绍网络信息获取过程中的各种关键算法，包括：分布式bloom过滤器设计，解决爬虫子系统的url重复爬去问题；高效http代理算法设计，解决网站对ip限制的问题；基于双消息队列的的P2P分布式网页下载算法设计，实时响