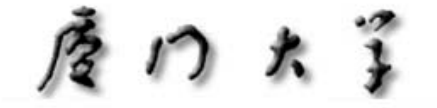
学校编码：10384　　分类号_____密级_____
学号：X2012231190　　UDC _____



工 程 硕 士 学 位 论 文

网络舆情分析系统的设计与实现

Design and Implementation of the Network Public Opinion Analysis System

王毅宏

指 导 教 师：王 备 战 教 授
专 业 名 称：软 件 工 程
论文提交日期：2014 年 9 月
论文答辩日期：2014 年 月
学位授予日期：　　年　月

指 导 教 师：_____
答辩委员会主席：_____

2014 年　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

网络舆情是指互联网上的社情民意，是指众多网民对现实社会和虚拟社会上的各种政治和社会现象及问题所表达的有较强影响力、倾向性的言论和观点。以论坛、贴吧、博客、微博、微信等为代表的网络媒体，由于其用户量大、快速传播、便捷的信息流转、互动性的信息交换等特点，已经成为我国舆情爆发的主要源头和发酵地。目前对网络舆论进行及时监测和分析,同时进行正确引导，对国家和社会的稳定有重大作用。

网络舆情分析系统，应实现在海量的互联网信息中采集舆情信息，达到用户的舆情及时发现、及时预警、跟踪舆论发展、协助用户消除舆情危机、形成舆情简报等分析结果的需求，建设一个这样的网络舆情分析系统手段来加强舆情管理的质量和效率。

论文在对舆情分析系统的相关技术进行分析研究的基础上，针对当下互联网舆情分析面临的挑战，设计并实现了一个网络舆情分析系统。首先，依据网络舆情信息的不同来源，系统设计了一个综合数据采集平台。数据采集平台采用了元搜索技术和多样化的垂直搜索子系统完成互联网上不同舆情信息来源的自动数据采集，并设计了自适应的 Web 信息数据抽取算法来过滤、抽取关键信息，最后进行统一的存储和索引，为后续的舆情分析提供有力的支撑。其次，舆情分析平台子系统采用基于权重的多级综合评判的舆情分析相关技术对数据进行分析，结合热点话题的发现算法，实现了舆情时间的舆情聚焦、事件聚焦、人物聚焦、热点聚焦、主题跟踪及传播分析、趋势和倾向性分析、图形统计分析展示等功能，满足了当下社会网络舆情分析工作的实际需要。最后是该系统的实现，并给出了实例进行分析。

关键字：舆情；垂直搜索；数据挖掘

Abstract

Public opinion is the aggregate of individual attitudes or beliefs held by the adult population in some area in a period. It refers to the number of users on a variety of political and social phenomena and problems of the real world and the virtual community expressed a strong influence, tendentious remarks and perspectives. Forum, news, blog, micro-blog and SNS as the main representative of the Internet media, due to its high user capacity, spread rapidly, information transfer conveniently, information exchange interactively and other characteristics, has become the main source of public opinion outbreaks and fermentation ground.

Rapid dissemination of information and interaction of opinions, expression of increasingly diverse demands for public opinion, some of the irrational arguments, gossip or negative reports often will inspire a sense of crisis of users. It will be the impact of social stability of the major risks if lacks effective guidance and control. Strengthen the timely monitoring of public opinion, to guide the development of public opinion effectively, to resolve the crisis of public opinion actively; it has important implications for national development for network monitoring and analysis of public opinion timely, guiding the development of public opinion on the network actively and effectively to avoid and resolve crisis of public opinion. Public opinion analyzing and mining on the Internet information become more and more important.

Analysis system of Internet public opinion should collect information of public opinion from vast amounts of Internet information. It should help users timely founding public opinion, timely warning, tracking public opinion development, analyzing spreads of public opinion, looking for the roots of crisis public opinion information , preventing and treating the propagation and diffusion of the negative public opinion timely, guiding the development trend actively, assisting users to eliminate the crisis, creating statistical reports and dynamic chart analysis results. It is imminent to construct such a analysis system of Internet public opinion to enhance the quality and efficiency of the hand in the management of public opinion.

This dissertation introduces related technology about public opinion analysis, and aim at giving responses to present challenges; it can analyze and design a public opinion analysis system of internet. And it implements a data collecting platform, first we imply various of vertical search components to collect multiple types of HTML pages, and then it design a self-adjust algorithm to extract the necessary information from web pages, finally the dissertation indexs and stores them at a unified data

format. Analysis module apply information mining technologies base on weight, and the system support functions such as hot topic analysis, event analysis and character analysis etc. At the end of this dissertation it will introduce the implemented system, and give an analysis example about a public issue.

Key Words： Public Opinion**;** Vertical Search**;** Data Mining

目录

Contents

第一章 绪论

网络舆情分析能够及时发现、预警、防范处理负面舆情泛滥和扩散、积极引导事态发展、消除舆情危机，是有效掌控网络，促进社会稳定和谐发展的重要手段。以微博为代表的新型社交网络平台的广泛应用和发展，其快速、便捷的信息发布和传播特性给舆情分析系统的信息抽取和分析提出了更高的要求。本章将对论文的研究背景、现状做一个介绍和分析。

1.1 项目的背景及选题意义

舆情即社情民意，是公众对各种政治和社会现象及问题的认识或看法，是公众的意愿和态度，是群众心理、情绪、意见、要求和思想的综合表现。网络舆情是指网络上的社情民意，是指在互联网背景之下，众多网民对现实社会和虚拟社会上的各种现象、问题所表达的信念、态度、意见和情绪表现的总和，其中混杂着理智和非理智的成分[1]。

网络舆情即在互联网的环境下的社情民意，它以互联大网为传播载体，具有信息多元化、表达便捷、媒体和网民互动、海量信息等新的特点。在数以百亿计网页的互联网海洋中，转载分享等形成的大量重复性信息，每日每刻都有着新的网页的产生，收集和处理舆情信息已经不是人工可以解决的问题，建立一个完善的网络舆情信息分析系统，及时应对网络舆情，及时发布权威信息，澄清事实，强化主流言论，孤立负面言论，有效引导促成正确舆论的形成，迫在眉睫。建设一个具有信息获取及时、分析准确、按需监控和信息可适时发布的互联网舆情分析系统，对于积极化解网络舆论危机，掌控网络安全，对维护社会安全、和谐、稳定发展具有重要的意义。

2014 年 1 月，中国互联网络信息中心(CNNIC)发布的报告显示，截至 2013 年 12 月，中国网民规模达 6.18 亿，互联网普及率为 45.8%（其中手机网民规模达 5 亿，网民中使用手机上网的人群占比提升到 81.0%）。截至 2013 年 12 月，我国拥有 IPv4 地址 3.3 亿，IPv6 地址 16670 块/32，域名总数 1844 万，网站总数 320 万，中国网页数量为 1500 亿个（相比 2012 年同期增长 22.2%）。

从统计报告可以看出，我国互联网资源在整体上都得到了相当的提升。不论

是 IPV4 和 IPV6 的地址数量，还是域名数量和网站数量都增长迅速，尤其是中国网民数量和网页数量。博客、论坛、微博，社交网络等新兴的交互式应用，智能终端的普及应用，3G 移动通信快速接入，极大的改变网民的上网行为。随时随地的获取最新资讯并进行评论转发，发布、共享信息已成为网民生活的主要内容。越发多样化的网络应用，快捷方便的上网方式，日益增强的参与性和互动性，使得网络信息主要载体的网页数量也呈现爆炸式增长。

1.2 研究现状及存在问题

互联网舆情数据挖掘的目的是从基于网络的海量异构数据集合中发现焦点和热点信息，通过焦点和热点信息的分析，获取其历史发展轨迹，预测其趋势变化，掌握其发展规律。网络舆情数据挖掘涉及技术面甚广，包括网络与信息表示、海量数据组织存储及管理、信息内容智能挖掘等多个领域。本节对舆情监测分析技术的研究现状进行综述。

目前，国内对于网络舆情的概念有两种比较主流的说法:一、网络舆情是以事件为核心，以网络为载体，是广大网民以及网民的态度、意见、观点的表达、传播与互动，以及后续影响力的集合。综合以上两个定义，无论网络舆情的定义是什么，网络事件、网民情感、传播互动、影响力等关键词总是不可缺少的，并且贯穿于整个研究活动的始终。二、网络舆情起源于现实社会或网络中的某个事件，这些事件因为受到某种刺激，并通过互联网进行广泛传播和发酵，集合了人们对于该事件的所有认知、态度、情感和行为倾向。

国外对于网络舆情的监管和分析研究比国内起步较早，应用也更加广泛。1995 年韩国首个开始进行网络审查，2002 年推出网络实名制度。2005 年英国科波拉软件公司推出可以判断一篇文章对政党的政策是持肯定还是否定态度的名为“感情色彩”的舆情分析软件。如美国的舆情研究协会、欧洲舆情分析中心、欧盟的舆情分析中心等国外的舆情研究机构也陆续设立。目前，国内已经有许多机构开始从事公开网络舆情研究，已经初步实现各领域覆盖。创建于 1999 年的天津市社会科学院舆情研究所是国内第一个正式研究网络舆情的科研机构，自成立以来在舆情领域的基础理论方面取得了丰硕的成果，在国内相关领域中一直处于领先地位。国内网络舆情的研究自 2005 年开始得以快速发展，以北京交通大学、