学校编码：10384　　分类号_____密级_____
学号：X2010230436　　UDC_____

厦门大学

工 程 硕 士 学 位 论 文

网络舆情信息处理系统的设计与实现

Design and Implementation of the Public Opinion Information Processing System

王 威

指 导 教 师：李贵林副教授
专 业 名 称：软 件 工 程
论文提交日期：2014 年 6 月
论文答辩日期：2014 年 7 月
学位授予日期：　　年　　月

指 导 教 师：_____
答辩委员会主席：_____

2014 年 6 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

如今，互联网飞速发展，移动时代已经到来。移动媒体、互联网平台渐渐的融入到人们的生活中，越来越多的人通过互联网平台来表述的自己的思想、见解和感情。由于不同人的人对同一个事物的看法存在不一致性，那么在互联网上就会有积极向上的思想，也会有消极负面的看法。当前互联网仍然在快速发展，网络接触的人群面更加广泛，从学生到青年以及中老年人，互联网舆论对人们的思想和行为的影响越来越重要。因此，对于网上很多舆论信息进行监测分析，及时有效地维护社会稳定，具有重要的现实意义。

本文将首先对网络舆情的特性进行介绍，针对其特点设计实现了相应的解决方案。通过该网络舆情信息处理系统对网络上的热点敏感文本信息实施监测和预警。本系统的功能作用要求其必须在较短的时间内对互联网中的舆情热点信息做出监测和定位，因此采用具有指定功能的搜索引擎。同时，系统建立一个统一的互联网文本信息库，对获取的文本主题按照其热度排列，从而反应出热点话题。对于特殊的最新消息和热点以可视化的界面展示出来，随时随地对互联网上的敏感信息建立预警监控，准确和有效地管理网络信息。

本文采取当前前言流行的自然语言处理技术，在对以前互联网舆情分析算法的分析基础上，设计和实现了网络舆情信息处理系统。该系统对当前舆情分析系统中的话题识别和预测方法技术进行了优化整合，能过对互联网当中的热点敏感主题进行预测，促使互联网平台的进一步健康安全的发展。

关键词：网络舆情；监测分析；关联规则库

Abstract

With Today, the rapid development of the Internet , mobile era has arrived. Removable media, the Internet platform gradually integrated into people's lives , more and more people through the Internet platform to express their own thoughts, opinions and feelings. Due to the different people who inconsistency on the same view of things , then they would have positive thoughts on the internet , there will be a negative negative views . The current Internet is still in rapid development of the network is more extensive surface contact with people , from students to young people and the elderly, the impact of the Internet on people's thoughts and opinions behavior is increasingly important. Thus, for many online media monitoring and analysis of information, timely and effective manner to maintain social stability , has important practical significance.

This article first introduced the characteristics of network public opinion for its characteristic design and implement appropriate solutions. Through the network of public opinion sensitive information processing system for hot text on implementation, monitoring and early warning network information. The functional role of this system must be within the required shorter time in the information made public opinion monitoring Internet hotspots and location , so a search engine with the specified function . Meanwhile, the establishment of a unified system of Internet text information database, the text arranged according to their topics for heat , thus reflecting the hot topic. For the latest news and hot special visual interface to show up anywhere, anytime to sensitive information on the Internet early warning monitoring, accurate and effective management of network information.

In this paper, taking the current preface popular natural language processing technology, the Internet public opinion in the previous analysis algorithm based on the analysis , design and implementation of a network of public opinion information processing system . The system analyzes the current system of public opinion topic identification and prediction methods were optimized technology integration, which can be too hot for the sensitive topic of Internet predict further promote healthy and safe development of the Internet platform.

Keyword: Network public opinion; Monitoring analysis; Association rule library

目　录

Contents

第 1 章 绪论

1.1 研究背景和目的

纵观人类发展和传播信息的历史，相比人类文明的发展传播是以闪电般的速度创造奇迹。飞鸽传书起始于公元前 3000 年，到 1830 年开始的电报时代，以及打手摇电话到现在的 3G 、4G 时代。互联网改变了人类的生活方式同样改变了整个世界。人类通信发展的奇迹伴随着网络的出现而发生，人类文明终于在不断挑战自我的同时实现了超越[1]。从开始的连接四个主机装置，APPA 军事安全网络发展到现在每个人都可以分享的、上网冲浪的大型网络。在人类文明发展的史册上写下了一段辉煌的互联网历史。互联网吹到世界各地，是通信领域的一个激进的革命。作为一个高科技信息基地和生活的新方式，互联网给人们带来了太多，它不仅改变了人们的生活方式和实践本身，而且也改变了大家的学习，就业，生活和思想形式。但任何事情都是物极必反，当然，互联网在带给人们无限的恩典，便利和财富的时候，很多网页信息内容涉及黄色，反动和网络犯罪也在侵蚀着虚拟世界[2]。这不仅是纵容了一组无法无天的野心和企图，同时也引发了很多人的不理性的判断和行为上的障碍，网络的快速发展已经成为社会舆论信息和不良内容的来源和网络犯罪的工具。与此同时，随着网络覆盖的深入，基本上人人都参与到现今的网络虚拟世界中，网络上的任何微小事件或热点的公共讲座的发生可能会导致消息迅速蔓延[3]。因此在当今舆论自由，网络高度普及的时代，如何能够对网络中的热点和敏感话题进行监测和预测，是十分非常必要的。

1.2 研究内容和意义

本研究的主要内容有对数据的采集以及对采集回来的数据进行分词，便于搜索；追踪最新的话题与热点，网络舆情的追踪；

详细的研究内容如下：

(1)舆情信息浅层语义分析研究。主通过训练和测试的方法，语义角色标注

工具实现对文本的语义角色进行标注。

(2)舆情敏感话题如何识别研究。对于网络媒体中一阶段时期内大量出现的一些文字、图片、媒体信息实现抓与跟踪，通过 ICTCLAS 分词技术，文档频率特征提取方法等技术实现对热点话题的检测和识别。

(3)舆情信息倾向分析研究。提取主体的情感、观点等有价值的文本信息，包括特色数据库建设，情感词典构建，情感和语料库研究的知识发现算法计算等。本系统设计了一种方法来解决互联网舆情信息的这个问题。这个方法思想就是建立系统的针对互联网上的舆论收集，监测。并通过设计的搜索引擎技术方法快速的获得其相关网站服务器的数据信息。系统的建立可以对网络媒体反映的舆情主体系统的进行分类、排序、聚类和分析的目的。为了准确、有效地管理网络信息，快速掌握和了解主要机关单位的网络舆论，可以用可察看的形式对热点信息及热点问题进行专题凸显，并使系统全天候的监控互联网上的较为敏感信息，形成预警机制。

1.3 国内外研究现状

1.3.1 国外研究现状

英国科波拉软件公司设计实现了一个称之为“感情色彩”的软件[4]。该软件可以从互联网网站页面中自动提取文本信息，从文本信息中文章作者的基本思想，从网站和其他电子自动分析文章中得到用户的基本思路，进而做出一个报告，判断其是正面还是侧面，或者是中性观点。这对于有关政府机构提供公共关系的想法，可以帮助一些企业知道他们的舆论。该软件可以识别句子名词，动词和形容词，能够自动识别，如语法成分的句子，也可以判断动词的主体和对象，或者即使来了句代词分析，找出是指内容。该软件可以根据语法成分分析，删除的内容和主要内容没有关联，达到去噪和减少角色的复杂性分析的目的。

在外国，自然语言处理学科起步比较的早，随着时间的推移有部分学者和专家都在自然语言处理中设计的方法和思想提出了一系列独特的见解。其中基于关

键词统计分析技术是其中十分成熟的方法，然而在执行效率方面还有很大的提升空间[5]。

1.3.2 国内研究现状

国内近几年在信息处理领域内成型的软件产品也很多，如北京北大方正电子政务技术有限公司发起创办的智慧民意预警决策支持系统，可以结合知识管理方法，整合的互联网搜索引擎技术，自然语言处理技术，在互联网上的海量信息自动抓取，分类和聚类，主题检测和重点项目[6]。目前实际运用的网络媒体分析和处理系统的使用效果并没有让大多用户满意，因为缺乏对文字内容的情感倾向分析不够完善。如果没有文字情感倾向分析引擎，那么将会使舆情处理系统缺乏对网络民意分析的高效性。此外，大部分系统也没有建立一个有效和快速的舆情监测和预警机制，不能有效制止在网上流传各种负面新闻[7]。

1.4 论文组织

本文组织结构的具体安排为：

第 1 章，是论文的绪论。该章主要是介绍网络舆情信息处理系统的研究的背景和意义；然后说明了本文的研究内容和意义；最后介绍了舆情信息处理系统在国外和国内的研究现状。

第 2 章，主要介绍了网络舆情分析的相关技术和技术特点。分别叙述了系统采用的技术介绍、舆情系统采用的相关技术。

第 3 章，是系统的需求分析。本章分析了舆情信息系统的需求，对系统的需求获取、可行性分析方面和舆情系统定义做出了详细的分析。

第 4 章，给出了舆情信息处理系统技术详细设计。本章节详细了介绍了数据的采集，中文分词以及关联库的详细规则设计。

第 5 章，给出了本系统的详细实现方法，描述了如舆情采集模块、中文分词模块、关联规则库模块、热点话题识别和热度分析模块的具体实现。

第 6 章，系统测试。本章通过软件测试方法、通过测试用例完成对网络舆情信息处理系统的测试工作。

第 7 章，总结和展望。本章对论文进行了总结，展望了下一步的研究工作，并对系统的进一步提升提出了改进意见。

通过上述的结构安排，本文开头对网络舆情信息处理系统的分析，研究的背景和意义，然后实现技术的分析，提出了切实可行的技术选择，然后解释实现技术的可行性，从而实现网络舆论系统。经过和其他网络舆情系统的比较，该系统的效率具有一定的优势性。

1.5 本章小结

本章的只要内容对网络舆情信息处理系统的研究背景、内容和意义进行了介绍和分析，并且介绍了国内外在网络舆情系统分析中的关键技术，主要包括 Web 新闻网页上热门话题的检测、跟踪、分析、存储、预警等技术。

第 2 章 技术介绍和特点分析

本系统选择基于.NET 框架开发，为了方便系统的维护，降低业务逻辑接口与数据接口之间的耦合，系统选用 MVC(模型-视图-控制器)设计模式。

2.1 .NET 框架和 MVC 模式

(1).NET 框架

.NET 框架是指微软用以实现 XML，Web Services，SOA 和敏捷性的技术手段。如果想深入了解.NET，那么就要首先对.NET 出现的技术原因和所能解决的问题进行了解，就要知道为什么需要 XML,Web Services 和 SOA。使用框架搭建技术平台，而在应用系统在这个技术平台上得以展现。.NET 是新一代的应用平台，这所实现的系统是标准,联通,适应变化,稳定和高性能的。从技术方面来说,.NET 应用是一个基于.NET Framework 之上的应用程序。若应用程序跟.NET Framework 无关，它就不能叫做.NET 程序。.NET 是基于 Windows 系统运行的操作平台，应用于互联网的分布式。

(2)MVC 模式

该模式是一个基于服务器表达层的模型将应用分开，并改变应用之间的高度粘合,MVC 简称是 Model-View-Control 就是模型、视图、控制器。MVC 设计模式是建立在编程语言 Smalltalk-80 基础上的一种软件模式，它把应用的输入、处理、输出流程按照 Model、View、Controller 的方式进行分离，即将一个应用分成模型层、视图层、控制层。

2.2 网络舆情的相关技术

2.2.1 采集与提取技术

互联网上有着海量的信息，如何获取这些海量的信息是本系统要分析舆情的首要前提。采集海量的互联网信息并对其进行有效的提取既是本系统的基础，有