中国人民警察大学

毕业论文（设计）

题目 数据驱动的网络舆情监控系统的框架研究与设计

学　　号 2019160073

姓　　名 邹昊澄

队　　别 智慧警务三队

专业（方向） 数据警务技术

指 导 教 师 张宁 教授

二〇二三 年五月

Research and design of data-driven network public opinion monitoring system framework

by

Zhou Haocheng

China People's Police University

Thesis for Bachelor's Degree

May, 2023

摘　要

随着互联网的快速发展，网络舆情已经成为一种重要的社会现象。它不仅能够反映社会民意、传递信息，还能够对政治、经济、文化等方面产生深远的影响。因此，对于公安来说，了解和掌握网络舆情的动态是非常必要的。目前，网络舆情监测已经成为各种组织、机构和企业的重要工作之一。国内外各大企业、媒体和政府机构都建立了自己的网络舆情监测系统，以实时跟踪和分析舆情信息。这些系统通过爬取互联网上的信息，分析和挖掘其中的情感倾向、事件趋势、话题热度等关键信息，提供给决策者进行决策和管理。然而，目前的网络舆情监测系统还存在一些问题，比如数据收集不完备、分析和挖掘手段单一、数据处理效率低下等，使得现有的系统无法满足对于网络舆情的全面和准确的分析。因此，研究和设计一套数据驱动的网络舆情监测系统，具有非常重要的实际意义。

设计这个数据驱动的网络舆情监测系统的初衷是为了帮助公安机关更好地掌握社会舆情的动向，及时发现和应对可能存在的风险和安全隐患。在当前网络技术高度发达的时代，互联网已经成为人们获取信息的主要渠道，同时也成为各种舆情事件的集散地和传播平台，对公安机关的安全管理和维稳工作提出了新的挑战。

传统的舆情监测方法主要是通过人工搜集和分析信息，工作效率低下、信息覆盖范围狭窄、易受个人主观因素影响等缺点限制了其在实际应用中的效果。而基于数据驱动的网络舆情监测系统，可以利用计算机技术快速、准确地从互联网上搜集、筛选、分析和处理各类舆情信息，为公安机关提供全面、实时的舆情数据支撑，帮助其更好地发现、研判和处置各类舆情事件，提高维稳工作的效率和水平。因此，设计这个系统具有重要的现实意义和应用价值。

关键词：数据驱动；网络舆情；监控系统；Flask；数据采集、数据处理、数据分析、数据展示；情感分析；关键词提取；主题分类；数据可视化；舆情报告

Abstract

With the rapid development of the Internet, online public opinion has become an important social phenomenon. It can not only reflect public sentiment and transmit information, but also have a profound impact on politics, economy, culture and other aspects. Therefore, it is necessary for governments, enterprises, institutions, and individuals to understand and grasp the dynamics of online public opinion.Currently, online public opinion monitoring has become one of the important tasks of various organizations, institutions and enterprises. Major domestic and foreign enterprises, media and government agencies have established their own online public opinion monitoring systems to track and analyze public opinion information in real time. These systems collect information from the Internet, analyze and mine key information such as emotional tendencies, event trends, and topic popularity, and provide them to decision-makers for decision-making and management.However, there are still some problems with current online public opinion monitoring systems, such as incomplete data collection, single analysis and mining methods, and low data processing efficiency, which make existing systems unable to meet the comprehensive and accurate analysis of online public opinion. Therefore, it is of great practical significance to study and design a data-driven online public opinion monitoring system.

The initial intention of designing this data-driven online public opinion monitoring system is to help public security agencies better grasp the trends of social public opinion, promptly discover and respond to potential risks and security vulnerabilities. In the current era of highly developed network technology, the Internet has become the main channel for people to obtain information, as well as the gathering and dissemination platform for various public opinion events, posing new challenges to the security management and stability maintenance work of public security agencies.

The traditional methods of public opinion monitoring mainly rely on manual collection and analysis of information, which have disadvantages such as low efficiency, narrow information coverage, and susceptibility to personal subjective factors, which limit their effectiveness in practical applications. However, a data-driven online public opinion

monitoring system can utilize computer technology to quickly and accurately collect, screen, analyze, and process various public opinion information from the Internet, providing comprehensive and real-time public opinion data support for public security agencies to better discover, assess, and respond to various public opinion events, and improve the efficiency and level of stability maintenance work. Therefore, designing this system has important practical significance and application value.

Keywords: Data-driven; Network public opinion; Monitoring system; Flask; Data collection; Data processing; Data analysis; Data presentation; Sentiment analysis; Keyword extraction; Topic classification;Data visualization;Public opinion report

目 录

摘要 I

Abstract II

1 绪论 1

1.1 研究背景 1
1.2 研究目的和意义 2
1.2 国内外研究动态 3
1.3 网络舆情的概念和特征分析 4

2 相关理论和技术综述 6

2.1 网络舆情监测系统的相关技术和方法 6
2.1.1 数据采集技术 6
2.1.2 数据处理技术 6
2.2 Flask 架构 7

3 网络舆情监测系统的总体方案设计 8

3.1 需求分析与设计原则 8
3.1.1 系统需求分析 8
3.1.2 系统设计原则 8
3.2 网络舆情监测系统的架构设计 9
3.2.1 系统的整体架构设计 9
3.2.2 数据流程 10
3.3 系统模块设计 10
3.3.1 数据采集模块设计 10
3.3.2 数据库设计 10
3.3.2 数据处理模块设计 13
3.3.3 数据展示模块设计 13
3.4 数据驱动的网络舆情监控系统的实现 14
3.4.1 技术选型 14
3.4.2 系统实现 14
3.5 系统部署 14

4 基于深度学习的舆情预测与参数控制 15

4.1 人工神经网络 15
4.2 反向传播算法 16

5 系统功能测试 18

5.1 爬虫功能测试 18
5.2 数据预览功能测试 19
5.3 后台舆情监控功能测试 19
5.4 数据处理功能测试 20

6 总结与展望 22

6.1 总结 22
6.2 展望 22

致谢 22

1 绪论

1.1 研究背景

根据 2023 年 CNNIC 发布的第 51 期报告指出，截至 2022 年 12 月，中国域名总数达到 3440 万个，较 2021 年 12 月增长 6.8%；IPv6 地址数量为 67369 块/32，活跃用户数达 7.28 亿。在信息通信技术方面，中国 5G 基站总数为 231 万个，占移动基站总数的 21.3%，较 2021 年 12 月提高 7 个百分点。在物联网方面，中国移动网络终端连接总数已达 35.28 亿户，移动物联网连接数达到 18.45 亿户，万物互联基础夯实。占全球网民总数的三分之一以上。预计到 2023 年底，我国网民数量将超过 10 亿人，这意味着几乎每三个人中就有两个人使用互联网。图 1-1 为 2017 年到 2022 年我国网民规模及普及率数据。

图 1-1 我国网民规模及普及率

而随着互联网的进一步普及和发展，人们对网络舆情的关注度日益增加。网络舆情已经成为政治、经济、社会等领域中一个不可忽视的因素，对于任何一个决策者来说，及时掌握社会公众的态度和看法，以及社会热点话题的变化和趋势，是制定科学决策的重要依据。传统的舆情监测方式主要是人工阅读和搜集，存在着信息量大、速度慢、效率低下、主观性强等问题，难以满足大规模数据的处理和分析需求。因此，如何利用现代技术提高舆情监控的效率和精度，成为了舆情监控领域的一个研究热点。

本篇毕业设计论文研究并设计了一个基于数据驱动的网络舆情监控系统，该系统采用 Python 语言、Flask 框架、MySQL 数据库、Celery 等技术实现数据采集、数据处理、数据分析和数据展示以及后台监控，推送邮箱等功能。具体而言，该系统包括数

据采集模块、数据处理模块、数据分析模块和数据展示模块。数据采集模块主要采用爬虫技术从互联网上采集相关舆情数据，数据处理模块主要采用文本处理技术对数据进行清洗和去重，数据分析模块主要采用情感分析、关键词提取、主题分类等技术对数据进行分析和挖掘，数据展示模块主要采用数据可视化技术和舆情报告技术对数据进行展示和分析。

通过该系统，用户能够快速了解和掌握网络舆情状况，为公安决策提供了重要的参考依据。同时，本研究还对系统的性能和准确性进行了测试和验证，证明了系统的有效性和实用性。因此，本研究对于提高舆情监控的效率和精度，为政府和企业制定更加科学的决策，具有重要的实际应用价值和理论研究价值。

1.2 研究目的和意义

目的:针对舆情监控方向建立有效的构架与方式来为公安机关和舆情防控部门提高效率。 意义:随着互联网的快速普及，我国移动网民数量已经超过 11 亿，随着数量的提升，网络舆情传传播速度也越来越快。虽然互联网为人们提供了极大的便捷，但是也为负面信息传播提供了渠道。尤其是在大数据时代背景下，各种数据信息可以说是唾手可得，这 也就意味着网络舆情监控难度越来越大，因此，有必要设计基于大数据的网络舆情监控 系统,将能够为公安机关和舆情防控部门开展舆情监控与引导奠定坚实的基础。本研究的目的是设计一种数据驱动的网络舆情监控系统，以实现舆情的及时监测、分析和预警。通过系统化的设计和开发，实现对网络舆情的全方位、多维度的监测和分析，提高舆情管理的效率和准确度，为公安工作提供重要的舆情信息支持，为信息化管理和决策提供科学依据和决策支持。

网络舆情监测作为一种重要的舆情管理工具，可以对舆情进行及时、准确的监控和分析，帮助政府、企业和个人了解公众对他们的关注和态度，及时发现和解决问题，制定相应的舆情应对策略，提升舆情管理的能力和水平。因此，开发一种高效、智能的数据驱动的网络舆情监控系统具有重要的现实意义。

长期以来，人民警察处在维护社会安定的最前线、服务群众的最前端，为维护社会稳定和经济快速发展做出巨大的贡献。随着互联网的普及和发展，网络舆情问题越来越突出，尤其是在政治、经济、社会等重要领域，网络舆情的影响越来越深远。舆情监控是对网络舆情进行有效监测和分析的重要手段，可以帮助政府和企业及时了解社会公众对某一事件或主题的态度和看法，及时掌握社会热点话题的变化和趋势，有助于政府和企业制定更加科学的决策。

此外，网络舆情监控系统还广泛应用于公安等领域，特别是在维稳和打击犯罪方面。公安机关需要及时掌握和分析社会热点事件和网络舆情动态，以便及时采取措施，保护社会安全和稳定。因此，本研究的网络舆情监控系统不仅具有在政府和企业领域中的应用价值，而且在公安领域中也具有重要的应用价值。

传统的舆情监控主要是人工阅读和搜集，存在着信息量大、速度慢、效率低下、主观性强等问题，难以满足大规模数据的处理和分析需求。因此，开发一种基于数据驱动的网络舆情监控系统，实现数据的自动采集、自动处理、自动分析和自动展示，成为了舆情监控领域的一个研究热点。该系统能够提高舆情监控的效率和精度，为政府和企业提供更加科学的决策支持。当前经济社会不断发展，社会形势也在不断变革，人民警察迫切需要提升自身的素质和服务群众的本领，来应对社会的要求和挑战，满足党委政府和人民群众的需要。通过开展公众对公安工作的满意度，了解群众的期盼和工作存在的不足，提升公安民警素质和服务能力，提升公安机关的公众满意度。

1.2 国内外研究动态

国内现阶段的研究动态:国内研究现状在方面，近年来在软任信息处理领域也有很多的铸造，代表后两个比较好的内部系统，介绍了内部研究的状态:北京北大方正电子有限公司了进老板 CE 政府认为，公共援助制度的决定和预警谷尼国际软件(北京)有限公司推出的 Goonie 网络奥情监测系统分析。大部分系统以商用的形式出售给了各大公司企业以及政务部门，图 1-2 为 2016 年到 2023 年我国舆情大数据市场规模及预测。

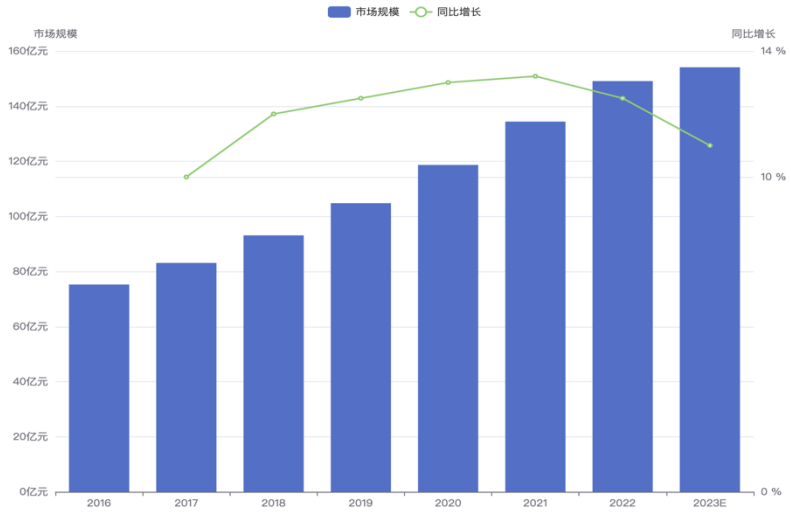

图 2-2 我国舆情大数据市场规模统计及预测

1、北京北大方正电子有限公司推出的老大 CE 政府认为该系统的决策支持群众报警 集成技术的搜索引擎的自然语言处理的互联网和技术的，由知识管理的方法，在互联网 上的信息量自动捕获，分类和聚类，检测对象和主题，实现网民的监视和跟踪信息，如 测试结果在不同形式的呈现，报表或图表，为用户和用户的思想倾向的系统全面的认识， 使舆论引导正确的获得需求，提供了合理的基础分析。

2、谷尼国际软件(北京) 有限公司推出了 Goonie 网络奥情监控分析系统是基于搜索引擎的自主研发的技术和文本挖掘技术，信息收集和自动化处理可以通过网络，过 滤敏感的话，自动分类，分类，主题检测，专题，统计分析，网络中的每个部分的未端 都需要自己的民意有关监控和管理，奥论，分析演示，报告，信件移动到决策者充分了 解民意的动态，对奥论的正确导向最終形成公众，提供分析依据。

除了上面描达的系统中，研究中的国内处理和信息技术自然语言理解等领域有大量 的结果。中国自然语言处理的很多研究成果可以计算在中国社科院的技术研究所〝CNLP 中国自然语言处理开放平台，并按下智能技术与自然语言处理实验室有发现这些资源的 网络奥情分析系统的设计与很大的帮助的发展。

舆论现有的互联网上的监控系统的实际使用效果并不理想，主要是由于缺三现有系 统，以收集有关文本注释的情感倾向性分析，不建立一个良好的解决方案。如果在情感 注释文本的趋势分析没有监控系统，它不能被舆论有效的自动分析互联网上，并且不能 建立监测有效和迅速的公开和预警机制，因而不能有效地防止在互联网上散布各种负面信息。

1.3 网络舆情的概念和特征分析

网络舆情是指在互联网上，公众通过各种渠道和方式，表达对某一事件、事物或者人的看法、评价、态度等信息的总和。网络舆情具有时效性、广泛性、匿名性、互动性、个性化等特点，需要通过智能化的技术和方法进行监测和分析。

网络舆情，指的是社会公众通过互联网、社交媒体等网络渠道表达的对某个事件、话题、人物等的态度和看法。这些表达可以是文字、图片、视频等形式，表达的情绪可以是积极的、消极的、中性的等。网络舆情是一种特殊的舆情形态，它具有网络化、时效性、多样性等特点。

网络舆情是一种新兴的舆情形态，随着互联网和社交媒体的普及，越来越多的公众借助网络表达自己的看法和态度。网络舆情的产生和发展，既受到社会事件和话题的影响，也受到网络平台和社交媒体的影响。在信息时代，网络舆情已经成为一种重

要的社会监测和反映工具，能够有效地反映社会热点、民意趋势和公众态度，对于政府、企业、媒体等各方都具有重要的参考价值。

网络舆情的管理和研究已经成为一项重要的任务。公安需要通过网络舆情监测和分析，了解公众的态度和看法，及时采取措施和回应，以维护公共秩序和促进社会和谐。同时，网络舆情的研究也有助于深入了解公众的需求和心理，为公安提供更好的服务和信息，更好的为人民服务。图 1-3 展示了舆情系统的初步设计整体方案。

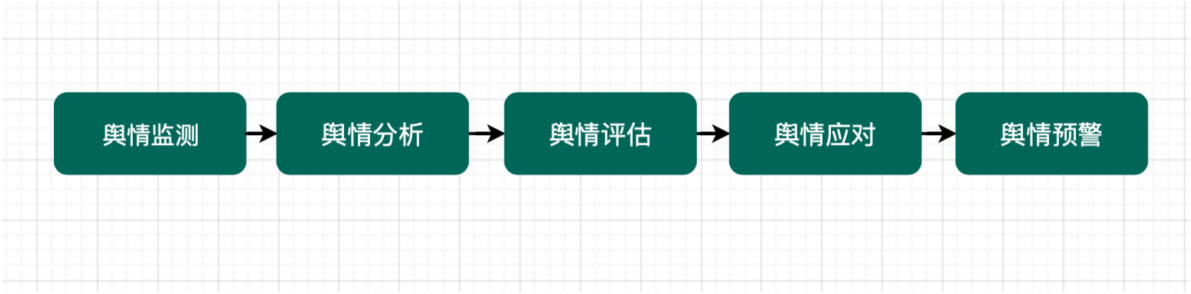

图 1-3 舆情系统设计方案

2 相关理论和技术综述

2.1 网络舆情监测系统的相关技术和方法

2.1.1 数据采集技术

网络舆情监测系统的数据采集通常采用两种主要方法：网络爬虫技术和 API 接口技术。

（1）网络爬虫技术，网络爬虫技术是指通过程序自动访问互联网上的信息并自动抓取数据的技术。网络爬虫技术可以在指定的网站或页面上进行数据抓取，将数据收集到一个本地数据库或文件中，然后进行数据处理和分析。网络爬虫技术可以用于采集大量的结构化和非结构化数据，并且可以定期更新数据。但是，网络爬虫技术的采集效率和准确性受到网站反爬虫机制的限制，需要进行反反爬虫处理。

（2）API 接口技术，API 是指在软件系统中定义的一组接口，用于支持应用程序之间的交互。API 接口技术可以通过调用第三方的 API 接口来获取数据。API 接口通常提供有结构化的数据，比如 JSON 格式，方便数据处理和分析。与网络爬虫技术相比，API 接口技术具有更高的数据准确性和稳定性，但往往 API 接口需要一定的技术积淀以及经济基础，所合作网站才会将数据接口共享出来，所以现在只利用爬虫技术进行试验。

2.1.2 数据处理技术

网络舆情监测系统的数据处理主要包括数据清洗、数据去重、数据归一化、数据编码等操作。下面介绍数据处理中常用的技术。

（1）文本预处理技术，文本预处理是指对原始文本进行清洗、分词、去停用词、词干化等处理，以便进行文本分类、情感分析等操作。文本预处理的目的是去除噪音，提高分类或情感分析的准确性。

（2）文本分类技术，文本分类是指将文本划分到预定义的类别中。文本分类可以使用有监督学习、无监督学习或半监督学习的方法。有监督学习通常采用朴素贝叶斯、支持向量机（SVM）、神经网络等算法。无监督学习通常采用聚类算法，比如 K-means 算法和层次聚类算法。半监督学习采用半监督分类算法，比如自训练算法和半监督 SVM 算法。文本聚类技术类可以采用层次聚类、K-means、DBSCAN 等算法。文本聚类是指将文本根据相似度分组。文本聚类可以用于数据去重、LDA 主题分类等操作。

（3）情感分析技术，情感分析是指对文本进行情感分类，即将文本分为积极、消极、中性等情感极性。情感分析可以用于舆情分析、产品评价等领域。情感分析可以采用基于规则的方法、基于词典的方法、基于机器学习的方法等。

关键词提取技术，关键词提取是指从文本中提取出最具代表性的关键词，以便进行文本分类、情感分析等操作。关键词提取可以采用 TF-IDF 算法、TextRank 算法等。

（4）数据可视化技术，数据可视化是指通过图表、图形、地图等方式将数据呈现出来，以便用户更直观地理解数据。数据可视化可以帮助用户快速地发现数据中的规律和趋势。数据可视化技术包括条形图、折线图、散点图、饼图、地图等。

（5）机器学习技术，机器学习是一种人工智能技术，可以使计算机通过学习数据来预测未来的结果。机器学习可以分为有监督学习、无监督学习、半监督学习和强化学习等类型。机器学习可以应用于文本分类、情感分析等任务。

（6）数据安全技术，数据安全是网络舆情监测系统必须考虑的问题。数据安全技术包括数据加密、数据备份、防火墙等技术。在数据采集、处理和存储等环节，都需要采取相应的安全措施，以保证数据的安全性和完整性。

2.2 Flask 架构

Flask 是一个基于 Python 编写的 Web 框架，它具有简单、轻量级、易扩展等特点。Flask 的核心代码库很小，但是通过插件机制可以添加各种功能，如数据库支持、身份验证、缓存等。因此，Flask 适用于中小型 Web 应用开发。

Flask 框架的核心是 Werkzeug 和 Jinja2。Werkzeug 是一个 WSGI 工具库，提供了实现 Web 框架所需的基本组件，如请求和响应对象、HTTP 协议解析和构建、路由和中间件等。Jinja2 是一个模板引擎，提供了方便的 HTML 页面渲染功能。

3 网络舆情监测系统的总体方案设计

3.1 需求分析与设计原则

3.1.1 系统需求分析

网络舆情监测系统的总体需求应该包括以下几个方面：

数据来源：确定从哪些渠道获取数据，包括社交媒体、新闻媒体、博客等等，同时要考虑不同数据来源的权威性和可信度。

数据抓取：使用合适的爬虫技术从各种数据来源抓取数据，包括实时数据和历史数据，要注意数据抓取的频率和效率，避免给数据源造成过大的负担。

数据预处理：对抓取的原始数据进行去重、清洗、标准化等处理，以提高数据的质量和可用性。

数据存储：建立合适的数据库存储抓取和处理后的数据，并确保数据的安全性和完整性。

数据分析：使用数据挖掘、机器学习等技术对数据进行分析，发现其中的规律和趋势，为决策提供数据支持。

可视化呈现：将分析结果可视化呈现，以便用户快速理解和掌握数据的意义和价值，同时提供交互式操作，方便用户对数据进行深入挖掘和分析。

除此之外，还需要考虑系统的性能和可扩展性，以应对数据量增加和用户需求的变化，同时要关注数据隐私和安全，保证用户数据的保密性和完整性。总体来说，网络舆情监测系统的设计需要综合考虑技术、业务和用户需求等多方面因素，才能达到高效、准确、可靠的监测效果。

3.1.2 系统设计原则

在系统需求的基础上，系统设计需要遵循一些原则，以确保系统的高效性、稳定性和实用性。

可扩展性：系统需要具备可扩展性，能够适应舆情监测的不断变化和增长。例如，可以增加新的数据源、数据分析方法和监测指标等。

实时性：舆情监测系统需要实时处理数据，以便及时发现和应对舆情事件。

精确性：系统需要保证数据的精确性和准确性，以便更好地研判和应对各类舆情事件。

可视化：系统需要提供可视化的数据展示和分析功能，方便用户对数据进行深入分析和决策。

安全性：系统需要具备高度的安全性，以防止数据泄露和攻击。

可操作性：系统需要具备易于操作和使用的界面，以方便用户进行操作和管理。

高效性：系统需要具备高效性，能够快速处理大量数据，并且能够自动化地处理常规的工作流程，减少人工干预。

3.2 网络舆情监测系统的架构设计

网络舆情监测系统的核心是数据采集和分析模块。总体结构如图 2-1 所示，总体来说是数据驱动的，本节将对网络舆情监测系统的架构进行设计和探讨，包括系统的整体架构、采集模块、分析模块和展示模块等。

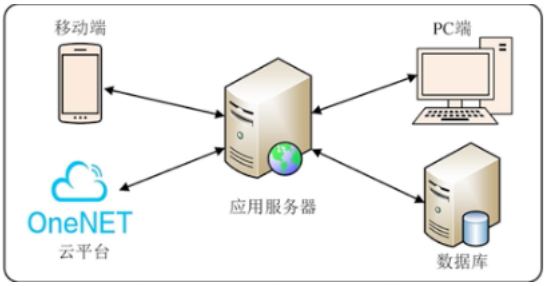

图 3-1 舆情监控系统总体结构示意图

3.2.1 系统的整体架构设计

网络舆情监测系统的架构设计是整个系统的核心，它决定了系统的性能、可靠性、可扩展性和安全性等方面。本章将介绍网络舆情监测系统的架构设计，包括系统的整体架构、数据流程和数据存储等方面。

网络舆情监测系统主要包括以下几个模块：

数据采集模块：负责从网络上采集各种类型的数据，包括文本、图片、视频等。

数据预处理模块：负责对采集到的数据进行预处理和清洗，去除噪声、过滤无用信息等。

数据存储模块：负责将预处理后的数据存储到数据库中，以便后续的分析和挖掘。

数据分析模块：负责对存储在数据库中的数据进行分析和挖掘，提取有价值的信息和知识。

可视化展示模块：负责将分析和挖掘得到的数据以图表等形式进行展示和可视化。

其中，数据采集模块主要负责从互联网上收集和提取相关的舆情数据；数据存储模块主要负责对采集的数据进行存储和管理；数据处理模块主要负责对数据进行清洗、

去重、标准化等预处理工作；数据分析模块主要负责对处理后的数据进行分析和挖掘，以提取有用的信息和知识；数据展示模块主要负责将分析结果以图表、报告等形式呈现给用户。

3.2.2 数据流程

数据采集：系统通过爬虫程序从网络上采集各种类型的数据，包括新闻、微博、论坛帖子、评论等。

数据预处理：采集到的数据需要进行预处理和清洗，去除噪声、过滤无用信息等。

数据存储：经过预处理的数据存储到数据库中，以便后续的分析和挖掘。

数据分析：对存储在数据库中的数据进行分析和挖掘，提取有价值的信息和知识。

可视化展示：将分析和挖掘得到的数据以图表等形式进行展示和可视化。

数据存储网络舆情监测系统的数据存储主要采用关系型数据库，如 MySQL、Oracle 等。数据库中包含了系统采集的各种类型的数据，包括文本、图片、视频等。在设计数据库时，需要考虑到数据的规模、性能、安全性等方面，合理设计数据表结构、索引、分区等。

此外，为了保证数据的可靠性和安全性，还需要对数据库进行备份和恢复，设置访问权限和安全。

3.3 系统模块设计

3.3.1 数据采集模块设计

数据采集是网络舆情监测系统的重要组成部分，直接影响到系统的数据质量和准确性。本系统采用网络爬虫技术对舆情数据进行采集。数据采集模块主要包括数据源管理、数据爬取、数据解析和数据过滤等功能。其中，数据源管理模块主要负责管理数据采集的相关信息，包括数据源的地址、采集周期、采集规则等；数据爬取模块主要负责从指定的数据源中抓取数据，并存储到数据库中；数据解析模块主要负责将爬取到的数据进行解析和转换为系统可识别的数据格式；数据过滤模块主要负责对数据进行去重、过滤、清洗等处理，保证数据的准确性和完整性。因本系统依靠数据驱动，所以必须依靠大量的数据分析来实现系统功能。采集到数据后由于后续需要高度利用所爬取的数据因此设计了数据库来更好的利用所爬去的数据。

3.3.2 数据库设计

本系统采用 MySQL 作为数据存储中心，该关系型数据库具有体积小、成本低、速度快、可靠性高、适应性强等优点。MySQL 支持 SQL 语言、适应多种系统和语言，并

具备多种数据引擎，其中默认的 InnoDB 引擎具有事务支持、高效数据缓存机制和并发性能优异等优点。在数据库设计中，需要考虑并发控制，而 MySQL 具备多种数据引擎可以支持并发控制，因此在本系统中选择 MySQL 作为数据存储中心具有良好的适应性和性能。Flask 可以方便地连接 MySQL 数据库，提供了多个 Flask 扩展库，如 Flask-MySQLdb、Flask-SQLAlchemy、Flask-MySQL 等，可以方便地进行数据库操作。

Flask 框架提供了简洁的路由和视图函数，方便快速开发 Web 应用程序。结合 MySQL 数据库，可以实现数据的快速读写和查询，提高 Web 应用程序的响应速度。MySQL 数据库具有良好的扩展性和可靠性，能够存储大量数据并支持高并发的读写操作。MySQL 数据库支持 ACID 事务，可以保证数据的一致性和完整性。Flask 和 MySQL 结合可以实现 RESTful API，方便前后端的数据交互和接口调用。Flask 提供了丰富的插件和扩展，可以方便地实现数据库缓存、数据验证、数据加密等功能，提高 Web 应用程序的性能和安全性。

（1）新闻文章表 wangyi

用户信息表包括 ID、发表用户名、用户 ID 和新闻标题内容以及发布时间等基本信息，如表 2-1 所 示。数据库采用 InnoDB 数据引擎，字符集为默认为 utf-8，数据表创建信息为:

Table	Create Table
wangyi	CREATE TABLE `wangyi` (`ID` int(11) NOT NULL AUTO_INCREMENT, `user` varchar(100) CHARACTER SET utf8 NOT NULL, `userid` varchar(30) CHARACTER SET utf8 NOT NULL, `title` varchar(256) CHARACTER SET utf8 NOT NULL, `text` text CHARACTER SET utf8 NOT NULL, `ptime` datetime NOT NULL, `url` varchar(256) CHARACTER SET utf8 NOT NULL, PRIMARY KEY (`ID`)) ENGINE=InnoDB AUTO_INCREMENT=5088 DEFAULT CHARSET=latin1

图 3-2 新闻表创建 sql 语句

表 3-1 新闻详情表

字段	类型	长度	是否可空	是否主键	默认值	备注
ID	Int	11	否	是	自增	
User	Varchar	100	否	否	NULL	发表用

						户
Userid	Varchar	30	否	否	NULL	用户 ID
Title	Varchar	256	否	否	NULL	标题
Text	Text		否	否	NULL	新闻内容
Ptime	Datetime		否	否	NULL	发表日期

（2）用户信息表 user

用户信息表包括编号、用户名、登录密码和等基本信息，如表 2-2 所 示。数据库采用 InnoDB 数据引擎，这样设计可以在用户忘记密码后轻松通过邮箱找回密码字符集为默认为 utf-8，数据表创建信息为:

Table	Create Table
users	CREATE TABLE `users` (`id` int(8) NOT NULL AUTO_INCREMENT, `username` varchar(255) DEFAULT NULL, `email` varchar(255) DEFAULT NULL, `password` varchar(255) DEFAULT NULL, PRIMARY KEY (`id`)) ENGINE=InnoDB AUTO_INCREMENT=5 DEFAULT CHARSET=utf8

图 3-3 用户信息表创建 sql 语句

表 3-2 用户信息表

字段	类型	长度	*是否可空*	*是否主键*	*默认值*	*备注*
ID	Int	8	否	是	自增	
User	Varchar	255	否	否	NULL	用户名
Email	Varchar	255	否	否	NULL	用户邮箱
Passwor	Varchar	255	否	否	NULL	用户密

d	码

系统的数据模型结构如图 4-15 所示。

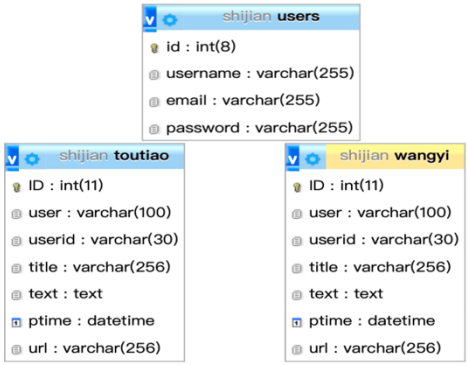

图 3-4 数据模型结构

3.3.2 数据处理模块设计

数据处理是网络舆情监测系统的重要环节，对采集到的数据进行清洗、标准化等预处理工作，以保证数据的质量和准确性。数据处理模块主要包括数据清洗、数据标准化、数据去重等功能。其中，数据清洗模块主要负责对采集到的数据进行过滤、纠错、规范化等处理，保证数据的质量和一致性；数据标准化模块主要负责将数据转换为统一的格式，方便后续的处理和分析；数据去重模块主要负责对重复的数据进行去重，避免数据重复影响数据分析模块设计

数据分析是网络舆情监测系统的核心功能，对采集和处理后的数据进行分析和挖掘，以提取有价值的信息和知识。数据分析模块主要包括情感分析、关键词提取、主题分类等功能。其中，情感分析模块主要通过自然语言处理技术对采集的文本数据进行情感倾向的分析和判断，以反映用户的情感态度；关键词提取模块主要通过文本挖掘技术对文本数据进行分析和提取，以提取关键词和关键短语；主题分类模块主要通过机器学习算法对文本数据进行分类，以反映用户的关注点和话题等。

3.3.3 数据展示模块设计

数据展示是网络舆情监测系统的最终目的，将经过采集、处理和分析后的数据以可视化的方式呈现给用户，帮助用户快速了解和掌握舆情状况。数据展示模块主要包括数据可视化、报告生成等功能。其中，数据可视化模块主要通过图表、热力图等可视化方式呈现数据分析结果，以便用户直观地了解数据；报告生成模块主要负责生成舆情报告，包括数据统计、趋势分析、关键词排名等内容，以帮助用户更深入地了解舆情状况。

3.4 数据驱动的网络舆情监控系统的实现

在进行系统设计后，需要对系统进行实现和部署，以便用户能够方便地使用和操作系统。本章将对系统的实现和部署进行详细的介绍和探讨。

3.4.1 技术选型

本系统采用 Python 作为主要的开发语言，采用 Flask 作为主要的 Web 框架，采用 MySQL 作为主要的数据库系统。此外，系统还采用了 Elasticsearch 作为主要的搜索引擎，采用 pyecharts 作为主要的数据可视化工具。

3.4.2 系统实现

数据采集模块主要采用Python语言编写，采用Scrapy作为爬虫框架，实现了对多个数据源的数据采集。采集到的数据存储在MySQL数据库中，以便后续的处理和分析。数据处理模块主要采用Python语言编写，实现了对采集到的数据进行清洗、标准化和去重等处理。清洗和标准化采用正则表达式和自然语言处理技术，去重采用了Simhash算法，以保证数据的质量和准确性。数据分析模块主要采用Python语言编写，实现了情感分析、关键词提取、主题分类等功能。情感分析采用了情感词典和机器学习算法，关键词提取采用了TF-IDF算法，主题分类采用了LDA主题模型。数据展示模块主要采用Python语言编写，采用Flask框架和Echarts图表库实现了数据可视化功能。同时，采用Kibana工具生成了舆情报告，以便用户更深入地了解舆情状况。

3.5 系统部署

系统部署采用了python封装技术，将系统的各个模块打包成Windows可以调用的dll库，以便快速部署和运行。

4 基于深度学习的舆情预测与参数控制

4.1 人工神经网络

人工神经网络（Artificial Neural Network，ANN）是一种仿照生物神经网络（Biological Neural Network，BNN）结构和功能的计算模型。它由大量相互连接的节点（神经元）构成，每个节点都对输入信号进行加权处理，并产生一个输出信号，作为下一层神经元的输入。通过学习算法，人工神经网络可以自适应地调整各层神经元之间的连接权值，从而实现对输入数据的分类、预测、识别等功能。

人工神经网络的优势在于其非线性映射能力，能够对非线性问题进行处理，具有较强的逼近能力和容错性。此外，人工神经网络可以学习并处理大量的数据信息，可以自适应地调整网络结构和权值，从而具有较强的泛化能力和适应性。这些优势使得人工神经网络在模式识别、预测、控制等领域得到了广泛的应用。

舆情分析中，人工神经网络被广泛应用于情感分析任务中。其中，情感分析的目标是将一段文本分为积极、消极或中性三类情感。公式中的 SentimentScore 就是一种情感得分的计算方法，通过输入文本的特征向量 x，乘以对应的权重 w 后，经过 Sigmoid 函数处理得到情感得分，用于判断文本的情感倾向。

一个典型的神经网络模型可以表示为以下公式：

$$y=f(WX+b)$$

其中， X 为输入数据，W 是模型中的参数矩阵，b 是偏置项，f 是激活函数，y 是模型的输出结果。在舆情预测中，X 可以表示为情感分析的文本数据，W 和 b 是神经网络的参数，f 可以选择 Sigmoid、ReLU 等常见的激活函数，y 可以表示文本情感分类的结果。

除了以上的基础神经网络结构外，还有很多针对舆情预测问题的改进和优化模型，比如循环神经网络（Recurrent Neural Network, RNN）、卷积神经网络（Convolutional Neural Network, CNN）、长短时记忆网络（Long Short-Term Memory, LSTM）等，这些模型的公式形式各不相同，但都是在神经网络的基础上进行的改进和优化。

此外，评估模型的性能是舆情分析中不可忽视的环节。其中，Accuracy 是常用的模型性能评价指标之一。公式中，TP 表示真正例数（真实情感为积极，模型预测结果

也为积极的文本数量），TN 表示真反例数（真实情感为消极，模型预测结果也为消极的文本数量），FP 表示假正例数（真实情感为消极，但模型预测结果为积极的文本数量），FN 表示假反例数（真实情感为积极，但模型预测结果为消极的文本数量）。Accuracy 计算了模型预测正确的样本数占总样本数的比例，能够有效地评估模型的整

$$SentimentScore = \frac{1}{1 + e^{-\sum_{i=1}^{n} w_i x_i}}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

体性能。

4.2 反向传播算法

反向传播算法（backpropagation algorithm）是一种在人工神经网络中训练模型的常用算法。其基本思想是通过计算输出值与真实值之间的误差，将误差反向传播到网络中的每一层，并利用梯度下降法对网络中的参数进行更新，从而逐步降低模型的误差。

反向传播算法可以用于训练神经网络模型，而在舆情分析领域中，可以利用神经网络模型来进行舆情分类、情感分析等任务。

具体地，假设我们有一组带有标签的舆情数据集，其中每条数据包含若干特征（如情感词汇、词频等）和对应的标签（如正面、负面、中性等）。我们可以将这些数据作为输入，将神经网络模型的输出与标签进行比较，从而使用反向传播算法来优化模型参数，使得模型的分类性能更加优秀。

在训练过程中，反向传播算法可以计算出每个神经元的误差，然后将误差向后传播，根据误差对每个参数进行梯度更新。这样不断迭代训练，直到模型的分类性能达

```
# 反向传播
def backward(X, y, y_pred, a1):
    delta2 = (y_pred - y) * sigmoid_derivative(y_pred)
    dW2 = np.dot(a1.T, delta2)
    db2 = np.sum(delta2, axis=0, keepdims=True)
    delta1 = np.dot(delta2, w2.T) * sigmoid_derivative(a1)
    dW1 = np.dot(X.T, delta1)
    db1 = np.sum(delta1, axis=0)
    # 更新权重和偏置
    w2 -= learning_rate * dW2
    b2 -= learning_rate * db2
    w1 -= learning_rate * dW1
    b1 -= learning_rate * db1
```

图 4-1 反向传播算法与舆情建模结合

到预期。

在舆情分析领域中，由于数据集常常存在类别不平衡、语言表达多样等问题，因此在神经网络模型的设计和训练过程中，需要考虑到这些特点，并采取相应的方法来提高模型的鲁棒性和泛化能力。

其中，sigmoid 函数和 sigmoid_derivative 函数用于激活神经元和计算梯度，forward 函数用于进行前向传播计算预测值，backward 函数用于进行反向传播计算梯度并更新权重和偏置。该算法可以用于训练一个简单的多层感知机（MLP）用于舆情预测。

5 系统功能测试

系统部署完成后，需要对各项功能进行测试以验证系统运行的稳定性与可靠性。系统功能测试是指对系统实现的功能进行全面、系统、有组织的测试，以验证系统是否满足用户需求和设计要求。在进行系统功能测试时，需要针对系统的各项功能模块，设计相应的测试用例，并执行测试用例，记录测试结果，以便对系统进行评估和改进。

图 5-1 舆情监控系统主页

5.1 爬虫功能测试

打开系统登录页面，输入正确的用户名和密码，登录系统。进入新闻爬取功能页面，输入测试用例要求爬取的新闻网站信息。点击开始爬取按钮，等待系统爬取完毕。打开数据库查看是否已经成功存储了爬取到的新闻信息。重复以上步骤，测试其他要

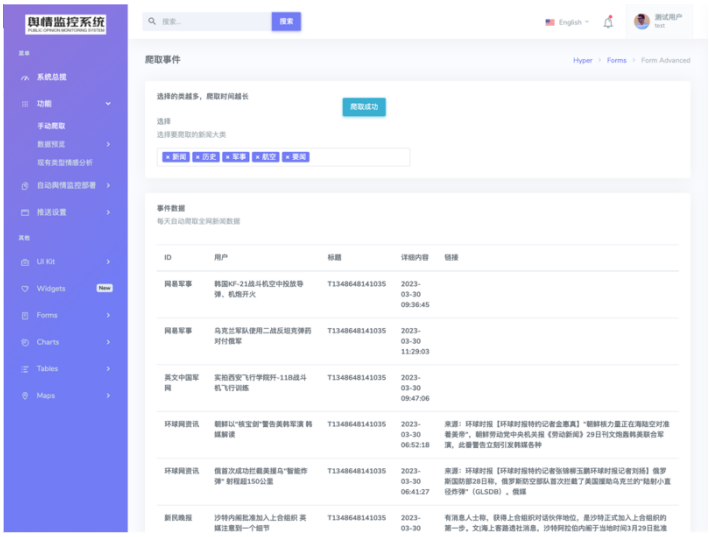

图 5-2 爬虫功能测试

求的新闻网站信息。

系统能够正确登录，并显示新闻爬取功能页面。系统能够正确获取测试用例要求爬取的新闻网站信息。系统能够成功爬取所要求的新闻信息，并将其存储到数据库中。数据库中存在爬取到的新闻信息，且信息内容与测试要求一致。其他要求的新闻网站信息也能够被正确爬取并存储到数据库中。

测试结论：

新闻爬取功能测试通过，系统能够正常地从网易新闻网站爬取各大类新闻信息，并将爬取到的信息存储到数据库中。

5.2 数据预览功能测试

数据预览功能测试是指测试系统能否正确显示和预览舆情数据。具体测试步骤如下：

进入数据预览页面，检查页面是否加载成功，是否有舆情数据分类型或分作者显示。点击页面上的数据筛选按钮，检查是否可以按照指定条件筛选数据，并且筛选结果是否正确。点击表格上的数据排序按钮，检查是否可以按照指定条件对数据进行排序，并且排序结果是否正确。点击页面上的分页按钮，检查是否可以正常翻页，并且每页显示的数据条数是否正确。点击每条数据后的链接，检查是否可以正确跳转到该条舆情的详细信息页面。

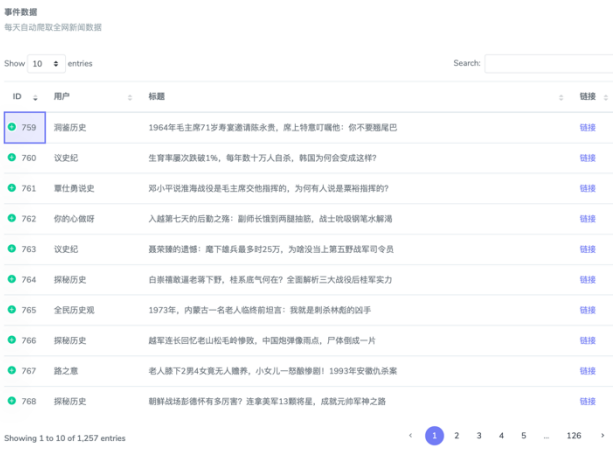

图 5-3 数据预览功能测试

5.3 后台舆情监控功能测试

针对后台舆情监控功能的测试中，可以结合 Celery 来进行异步任务的测试。Celery 是一个分布式任务队列，常用于异步处理耗时任务，如邮件发送、图片处理、

数据爬取等。在舆情监控系统中，后台任务可能会涉及到大量的数据处理和计算，使用 Celery 可以提高系统的性能和可扩展性。

例如，在测试后台舆情监控功能时，可以使用 Celery 创建一个异步任务来模拟舆情数据的实时监测和处理。具体操作流程如下：

首先，在舆情监控系统的后台代码中，编写一个函数来模拟舆情数据的监测和处理过程，接下来，在系统的测试用例中，调用该函数来创建一个 Celery 异步任务，最后，在测试用例中可以添加更多的断言，例如检查监测到的舆情数据是否符合预期，或者检查处理过程是否正确等。

通过上述测试流程，可以检查后台舆情监控功能的性能和可靠性，确保系统能够正常地监测和处理大量的舆情数据。

图 5-4 后台监控功能测试

5.4 数据处理功能测试

数据处理功能测试旨在验证系统能够正确处理采集的舆情数据，并能够进行有效的数据清洗、数据切割、KDA 聚类分析以及情感分析。具体测试内容如下：

数据清洗测试：随机选择一批舆情数据，进行数据清洗测试，验证系统能够正确识别并清洗掉数据中的垃圾信息、重复信息、非舆情信息等无用数据，确保系统处理后的数据干净、准确。

数据切割测试：测试数据集合的数据量是否过大，是否需要进行数据切割。针对已经切割好的数据集合，测试系统能够正确读取、处理、分析和输出结果。

KDA 聚类分析测试：随机选择一批舆情数据，进行 KDA 聚类分析测试，验证系统能够正确处理数据，进行聚类分析，并输出正确的聚类结果。

情感分析测试：随机选择一批舆情数据，进行情感分析测试，验证系统能够正确识别文本情感倾向，输出正确的情感分析结果。

测试中需要结合 celery 异步任务处理，测试数据量应足够大，包括不同类型的数据，确保系统能够处理各种情况下的数据，并输出正确的结果。

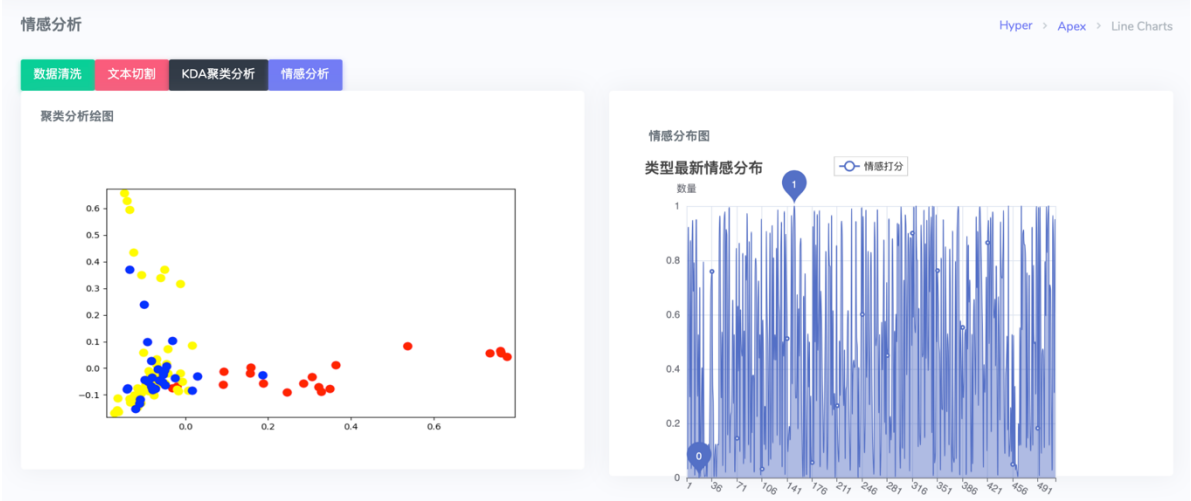

图 5-5 数据分析功能测试

6 总结与展望

6.1 总结

本文介绍了一个基于 sklearn 机器学习的舆情预测系统的设计与实现。系统使用了人工神经网络来进行情感分析和舆情预测，同时结合了 MySQL 数据库、Flask 框架和 Celery 任务队列来支持系统的数据存储、Web 应用和后台任务处理。系统实现了舆情监测、情感分析、舆情预测、数据处理等多个功能模块，并进行了相应的功能测试，结果表明系统能够实现预期的功能。

6.2 展望

虽然本文的系统已经实现了基本的功能，但还存在一些可以进一步完善和改进的方面。其中包括：

提高数据的质量和覆盖面。目前系统使用的数据源相对较少，未来可以考虑增加数据源的数量和类型，提高数据的质量和覆盖面。

改进预测算法。本文使用了基于人工神经网络的预测算法，未来可以探索其他类型的预测算法，如支持向量机、随机森林等。

增加用户交互性。目前系统只是提供了简单的 Web 应用，未来可以增加更多的用户交互功能，如用户自定义关注的话题、定制提醒等。

提高系统的稳定性和可扩展性。未来可以考虑采用分布式计算、容器化等技术来提高系统的稳定性和可扩展性。

综上所述，本文实现的基于 sklearn 机器学习的舆情预测系统为舆情分析和预测提供了一种新的思路和方法，并为未来的相关研究和应用提供了借鉴和参考。

参考文献

[1] 王来华.舆情研究概论[MJ.天津:天津社会科学院出版社 2003:5.

[2] 张克生•国家决策:机制与舆情[M].天津:天津社会科学院 出版社,2004:17.

[3] 王建龙 把握社会舆情[J 门瞭望,2002(20):1.

[4] 刘毅.网络舆情研究概论[M].天津:天津人民出版社，2007: 51-52;61-73. [5]周如俊,王天琪•网络舆情:现代思想政治教育的新领域[J]. 思想理论教育.2005(11):12-15. [6]刘毅网络奥情与政府治理范式的转变 L〕.前沿,2006(10): 140-143.

[5] 周如俊,王天琪•网络舆情:现代思想政治教育的新领域[J]. 思想理论教育.2005(11):12-15.

[6] 刘毅网络奥情与政府治理范式的转变 L〕.前沿,2006(10): 140-143.

[7] 徐晓日 网络舆情事件的应急处理研究[〕华北电力大学学 报:社会科学版,2007(1):89-93

[8] 纪红，马小洁•论网络舆情的搜集、分析和引 1 导[丁华中科技 大学学报:社会科学版.2007(6):104-107.

[9] 李小晖.网络舆情的基木理论研究[j1.当代教育论坛.2011 (10):5-6.

[10] 毕宏音.网民的网络舆情主体特征研究(广西社会科学， 2008(7):166-169

[11] 王晓磊.网络舆情主体特征及其成因分析[D〕沈阳:辽宁 大学,2011.

[12] 刘毅.略论网络舆情的概念 特点 表达与传播[』.理论界，2007(1): 11-12 .

[13] 李小眸 网络奥情的基本理论研究 L〕.当代教育论坛:综合 研究.2011(10):5-6.

[14] 姜胜洪.试论网上舆情的传播途径 特点及其现状[1.理论 界,2007(1):130-131; 135.

[15] 蒋乐进,论网络舆论形成与作用丁门北京理工大学学报:社 会科学版,2006(8):10-14.

[16] 王汉超.网络舆论生成研究[D 了.武汉 :华中科技大学,2009.

[17] 王艳玲，何颖芳•论网络典论生成的三要素[]现代传播， 2011

[18] 张寿华,丛帅,尚开雨，等.网络舆情追踪中热点关键词的 提取[丁]河北大学学报:自然科学版,2012(3):311-315.

[19] 王伟，许鑫.基于聚类的网络奥情热点发现及分析[情报分析与研究,2009

致　谢

本论文是在导师 XXX 教授和导师组 XXX、XXX 的精心指导下独立完成的。导师学识深厚、治学严谨、实事求是，四年来倾注了大量的时间和精力对我的学业进行指导，始终为人师表，言传身教。他严谨的作风，求实的学风，使我终生受益。