|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 西华大学  毕业设计说明书     |  |  |  | | --- | --- | --- | | 题目 | : | 跨平台轻量级诗词系统设计与开发  —— 手机端 | | 学院 | : | 计算机与软件工程学院 | | 年级专业 | : | 2018级计算机科学与技术 | | 姓名 | : | 王荣 | | 学号 | : | 3120180901114 | | 指导教师 | : | 魏冬梅 | | 完成时间 | : | 2020年05月24日 | |

# 摘要

随着移动互联网的高速发展以及移动智能设备的普及，为移动学习等新的学习理念提供了技术以及社会基础。互联网中充斥着多元化的信息，使得学习者可以随时随地的获取学习资源，人们也逐渐开始接受新的学习模式。

本文论述了古诗词鉴赏网站的设计与实现过程，提出了一种基于Web的古诗词鉴赏网站。进入该网站的用户无需注册，用户可以浏览诗词、欣赏诗词。进入搜索网页利用关键词搜索本网站内的诗词内容，也可以进入登录页，登录后收藏喜欢的诗人和诗歌。本网站的设计采用前后端分离模式，前端采用vue框架，后端采用express框架，适用于Internet发布。此网站分为前台界面和后台管理两部分，前台是友好的浏览界面，用户可以自己查找喜欢的诗词，并且对诗词进行收藏；后台管理是提供给管理员的，其中包括：古诗词的管理和诗人的管理。本网站适应现代社会的时间观念，花费低廉的传输成本在网络上收集信息。这样不仅节约了古诗词传播的成本，同时加快了信息的更新，更为广大喜爱古诗词的用户提供了一个交流的平台。该网站组建后经过反复的调试和修改，界面友好、操作简洁、使用方便、运行稳定。

从理论方面，通过对移动学习，古诗词学习和移动学习平台的现状阐述，分析移动学习的优缺点，并且对相关概念进行界定，为后续研究和设计提供支撑。

从设计方面，分析并设计支持移动学习的初中古诗词教学软件的功能模块，对内容及数据库进行设计，搭建了网页的开发和服务器环境，并实现了相关功能。根据线上运行和数据进行分析评价，了解了整个软件开发中的不足，方便后续改进。

【关键词】诗词管理，数据库，vue，node，javascript

# Abstract

With the rapid development of mobile Internet and the popularity of mobile intelligent devices, mobile learning and other new learning concepts have provided a technical and social basis. The Internet is full of diversified information, so that learners can access learning resources anytime and anywhere, and people gradually begin to accept the new learning mode.

This paper discusses the design and implementation process of ancient poetry appreciation website, and puts forward a kind of ancient poetry appreciation website based on Web. No registration is required to enter the site, which allows users to browse and enjoy poems. Enter the search page to search the poetry content in this website by keywords, you can also enter the login page, after logging in to collect favorite poets and poems. The design of this website adopts the front and back end separation mode, the front end adopts VUE framework, the back end adopts Express framework, which is suitable for Internet publishing. The website is divided into two parts: front desk interface and background management. The front desk is a friendly browsing interface. Users can find their favorite poems and collect them. Background management is provided to administrators, including: ancient poetry management and poet management. This website ADAPTS the modern society time concept, spends the low transmission cost to collect the information on the network. This not only saves the cost of the dissemination of ancient poetry, but also speeds up the update of information, and provides a platform for communication for the vast number of users who love ancient poetry. After the establishment of the website through repeated debugging and modification, friendly interface, simple operation, easy to use, stable operation.

In terms of theory, the advantages and disadvantages of mobile learning, ancient poetry learning and mobile learning platform are analyzed, and relevant concepts are defined to provide support for subsequent research and design.

From the aspect of design, analyze and design the function modules of the ancient poetry teaching software supporting mobile learning in junior middle school, design the content and database, set up the development and server environment of the web page, and realize the related functions. According to the online operation and data analysis and evaluation, understand the shortcomings of the whole software development, convenient for subsequent improvement.

【Key words】 Poetry management, Database, VUE, The node, javascript

**目录**

[1．绪论 1](#_Toc19827)

[1.1 研究背景 1](#_Toc19556)

[1.2 研究意义 1](#_Toc28295)

[1.3 研究内容 2](#_Toc18918)

[2. 系统概述 3](#_Toc22112)

[2.1 用户的特点 3](#_Toc26974)

[2.2 系统基本信息 3](#_Toc2710)

[2.3 系统开发工具 4](#_Toc19975)

[2.4 系统可行性分析 5](#_Toc11604)

[3．需求分析 6](#_Toc30145)

[3.1 系统功能需求 6](#_Toc32751)

[3.2 性能设计需求 6](#_Toc13231)

[3.3 条件与限制 7](#_Toc27524)

[3.4 系统用例图 7](#_Toc11516)

[4．概要设计 9](#_Toc18205)

[4.1 处理流程 9](#_Toc20685)

[4.1.1 情感分析处理流程 9](#_Toc4209)

[4.1.2 情感可视化处理流程 11](#_Toc19614)

[4.2 总体结构和模块外部设计 12](#_Toc13118)

[4.3 功能分配 13](#_Toc18584)

[4.4 接口设计 14](#_Toc7367)

[4.4.1 外部接口 14](#_Toc21424)

[4.4.2 内部接口 14](#_Toc1030)

[4.5 数据结构设计 15](#_Toc25986)

[4.5.1 逻辑结构设计 15](#_Toc21637)

[4.5.2 物理结构设计 16](#_Toc19528)

[4.6 运行设计 17](#_Toc22791)

[4.6.1 运行机制 17](#_Toc20327)

[4.6.2 运行控制 18](#_Toc27743)

[4.7 出错处理设计 19](#_Toc21339)

[4.7.1 出错输出信息 19](#_Toc27183)

[4.7.2 出错处理对策 20](#_Toc29195)

[4.8 维护设计 20](#_Toc9825)

[5. 详细设计 21](#_Toc21203)

[5.1 软件结构 21](#_Toc10642)

[5.2 数据采集与预处理模块 22](#_Toc10179)

[5.2.1 新闻爬取模块 22](#_Toc8660)

[5.2.2 新闻评论爬取模块 24](#_Toc8673)

[5.3 情感分析模块 25](#_Toc12905)

[5.3.1 新闻情感分类模块 26](#_Toc31858)

[5.3.2 评论情感分类模块 30](#_Toc25189)

[5.3.3 文本聚类模块 34](#_Toc32661)

[5.4 用户交互模块 36](#_Toc10587)

[5.4.1 后台用户管理模块 36](#_Toc7351)

[5.4.2 后台新闻管理模块 39](#_Toc21124)

[6. 运行测试 41](#_Toc7084)

[6.1 测试计划 41](#_Toc29180)

[6.2 测试结果分析 41](#_Toc17765)

[6.2.1 数据爬取与预处理测试 41](#_Toc17675)

[6.2.2 情感分类功能测试 42](#_Toc8617)

[7. 运行结果 44](#_Toc23956)

[7.1 数据爬取与预处理运行结果 44](#_Toc18228)

[7.2 情感分类运行结果 45](#_Toc14880)

[7.3 文本聚类运行结果 46](#_Toc31271)

[总结 47](#_Toc14113)

[致谢 48](#_Toc29366)

[参考文献 49](#_Toc27484)

# 1．绪论

## 1.1 研究背景

目前，网络正以一种非同寻常的的冲击力影响着人类的活动，网络的出现和发展，颠覆了传统信息的传播方式，冲破了传统交流方式中时间和空间上的种种限制，使人类从形式到内容，从物质到精神，从生产到生活的各种活动发生了极大的转变，并且给人类带来了新的机遇和挑战。随着Internet在全球的迅速发展，社会信息化强度迅速增强，尤其是在电子商务、网上购物、信息查询和动态发布等商业化浪潮的推动下，Web站点提供的信息服务要求用户和Web服务器之间能够进行实时、动态、高效的交互，传统的静态页面无法实现这一功能，这就要求Web与数据库相结合。22世纪作为Internet和Web的世界，传统的信息处理方式已经满足不了信息技术产业发展的要求，对于运用计算机来辅助协调和管理的需求正逐步提高。

本次项目的目的，就是来制作一个古诗词鉴赏网站，该网站属于学习交流型站，众多的用户访问网站不仅可以学习到古诗词，还能收藏喜欢的诗词，诗人。网站精选优秀的古诗词，大多数人都能在这里找到自己喜欢的古诗词，每个古诗词后面还有经典的赏析。还可以通过搜索功能，以关键字进行搜索，查询你需要的诗词。

## 1.2 研究意义

移动互联网技术的日益成熟和移动智能设备的逐渐普及，人们可以更加便捷地获取到信息和学习，人们也开始接受移动学习这种新的学习方式。一份网络新媒体报告中指出:“移动设备的普及改变着人们和周围环境互动的方式，移动技术支持下的新型方式可以增加人们接受知识的机会，其便利性推动着人们对移动学习战略的需求。移动诗词的实现基于移动智能设备，青少年也习惯了这些设备的使用。

这也使得古诗词移动学习的理念，应用于网页传播成为可能。相对于传统的古诗词而言，基于网页平台的移动传播可以结合信息技术，使诗词更好地传播。 同时，信息技术诗词资源的多元化，也可以成为人们对学习古诗词兴趣的方法之一。激发用户的使用兴趣。

## 1.3 研究内容

本文是围绕着诗词和诗人的文化内涵展开的研究，研究内容包括以下几点：

1. 数据获取：本文通过管理员上传诗词信息，诗人信息后保存到数据库与本地。
2. 诗词诗人分类：为了方便用户使用和查找信息，本系统将诗词按风格分类，有田园诗，壮志诗，思想诗等，将诗人按朝代分类，并且将诗词和诗人分类信息通过数据可视化直观地展示给用户。
3. 关键词分析：本系统通过对关键词搜索功能，能根据用户的需求将相应的诗词信息展示给用户，提高用户体验。
4. 趣味模块：本系统单独增加了趣味模块，包含歇后语，藏头诗，视频展示等模块，让内容丰富多样。

# 2. 系统概述

## 2.1 用户的特点

网站中的用户按系统权限可分为三种角色——管理员、普通用户、游客，他们所拥有的权限各不相同，其中管理员拥有管理系统中的诗词信息、增加，删除，修改数据的权限；普通用户享有查看诗人信息、查看诗词及评论，收藏诗词等功能；游客只享有诗词信息功能。

表 2-1 用户描述表

|  |  |
| --- | --- |
| **用户类** | **描述** |
| 管理员 | 管理员拥有管理系统中的诗词信息、增加，删除，修改数据的权限，还可以增加管理员，上传视频 |
| 普通用户 | 普通用户可以对诗词进行查询及评论，对喜欢的诗词和诗词可以收藏，还可进入趣味版块享受特色功能，也可进入聊天室参与聊天。 |
| 游客 | 游客只能查看诗人和诗词信息，不能收藏诗人和诗词。 |

## 2.2 系统基本信息

本系统主要的运行环境为网页，因此此系统是跨平台的能满足苹果和安卓各大平台。系统功能需求有如下几点：

（1）基本模块(用户注册、 登录、查询 、个人中心)；

（2）诗词及评论信息上传与处理；

（3）聊天室模块(所有人可在同一个平台聊天)；

（5）推荐系统(根据用户浏览的诗人和诗词推荐给用户对应的诗词诗人信息)；

## 2.3 系统开发工具

此次设计主要采用vsCode加nginx后台服务器进行，设计过程中页面主要使用html，css，javascript技术完成。

vsCode是一个优秀的用于前端开发的集合，vsCode的功能非常强大，支持也十分广泛，尤其是对各种开源产品的支持十分不错。vsCode有很多免费的插件，利用它们可以在数据库的开发，发布，以及应用程序服务器的整合方面极大的提高工作效率。它是功能丰富的前端集成开发环境，包括完备的编码，调试，测试和发布功能，完整的支持HTML, CSS， Javascript。

Nginx是一款[轻量级](https://baike.baidu.com/item/%E8%BD%BB%E9%87%8F%E7%BA%A7/10002835" \t "_blank)的[Web](https://baike.baidu.com/item/Web/150564" \t "_blank) 服务器/[反向代理](https://baike.baidu.com/item/%E5%8F%8D%E5%90%91%E4%BB%A3%E7%90%86/7793488)服务器及[电子邮件](https://baike.baidu.com/item/%E7%94%B5%E5%AD%90%E9%82%AE%E4%BB%B6/111106" \t "_blank)（IMAP/POP3）代理服务器，在BSD-like 协议下发行。其特点是占有内存少，[并发](https://baike.baidu.com/item/%E5%B9%B6%E5%8F%91/11024806" \t "_blank)能力强，事实上nginx的并发能力在同类型的网页服务器中表现较好，中国大陆使用nginx网站用户有：百度、[京东](https://baike.baidu.com/item/%E4%BA%AC%E4%B8%9C/210931" \t "_blank)、[新浪](https://baike.baidu.com/item/%E6%96%B0%E6%B5%AA/125692" \t "_blank)、[网易](https://baike.baidu.com/item/%E7%BD%91%E6%98%93/185754" \t "_blank)、[腾讯](https://baike.baidu.com/item/%E8%85%BE%E8%AE%AF/112204" \t "_blank)、[淘宝](https://baike.baidu.com/item/%E6%B7%98%E5%AE%9D/145661" \t "_blank)等。

开发环境：操作系统window10。

运行环境：vsCode和nginx工具。

表 2-2 系统开发工具表

|  |  |  |
| --- | --- | --- |
| **系统名称** | | 跨平台轻量级诗词系统设计与开发 |
| **系统功能** | | 跨平台轻量级诗词系统 |
| **系统负责人** | | 王荣，张玉 |
| **系统开发工具** | **a.编程语言工具** | VSCode |
| **b.数据库及管理工具** | MongoDb，MongoDb Compass |
| **c.文档编写工具** | WPS 2019 |
| **d.版本控制** | Git |

## 2.4 系统可行性分析

中华文化源远流长，博大精深，古诗文作为中华传统文化的一部分，有着自身的历史意义和现实意义，古诗文作为中华传统优秀文化的标志，有利于被世界各民族所接纳、借鉴，有利于使中华优秀传统文化走向世界。学习古诗文是对优秀传统文化的继承和发展；是陶冶情操，使人们得到美的熏陶的重要途径。传统文化的学习，是走进古人生活的途径，而文字最能走进一个人的内心。古人通过书写来传达自己的情感，传达着自己内心的观念和人文情怀的载体。随着科技的进步，互联网技术的推广，古诗文在网站上呈现，使之传承，被越来越多的人接受。因此，设计实现传统美文学习网站，为人们提供古诗文学习服务变得非常必要。

（1）理论分析：由于本诗词管理系统设计的目标明确, 功能强大, 且系统建成后可以直接供广大人员使用, 通过对本系统的使用, 可以起到传播文化的作用, 也可以带来一定的流量效益。从经济上分析, 本系统是可行的。

（2）功能模块分析：该系统最根本的目的是实现诗词的管理，展示。传统美文学习网站是供广大人群学习，欣赏的网站，因此该网站应能满足用户的常规需求，如用户注册，登录，阅读古诗文，根据关键字查找古诗文、名句、通过视频学习古诗文，提供成语接龙，藏头诗，歇后语等趣味小游戏。根据时代查看诗人，根据诗词分类查看诗词。诗词下方展示推荐诗词，诗人下方展示推荐诗人。用户可收藏，取消收藏诗人和诗歌。对于网站管理者能添加、纠正，删除古诗文，古诗文分类和诗人信息，管理管理员信息，设置网页轮播图等。

（3）非功能需求分析：

1）本系统主要以诗词为核心，每首诗词的信息与诗人，时代等多个信息关联

2）考虑到系统的安全性，对数据的访问通过数据库独立实现，对登录操作进行审核，对非法的用户进行限制。只有被允许的用户才能操作数据。

3）系统界面美观，友好，布局合理，窗口信息简明，有层次感，分类清晰明了，用户操作简单，把复杂的操作隐藏在了后台，用户能轻松使用。

4）诗词和用户的数据信息需要定期进行安全备份，防止丢失造成不必要的损失。

# 3．需求分析

## 3.1 系统功能需求

经过详细分析，确定了该项目有如下功能需求：用户登陆/注册，首页诗词展示，诗词详情展示，趣味界面，聊天室，推荐功能。具体包括：

1. 用户登陆/注册：老用户直接登陆，对比数据库数据正确后显示登陆成功，新用户填写相关信息，然后保存到数据库；
2. 诗词信息：展示诗词详情及其解析；
3. 用户管理：管理员管理用户信息，用户管理个人信息；
4. 诗词数据管理：管理员对诗词数据增删改。

## 3.2 性能设计需求

(1)系统的稳定性需求：系统应具备稳定性，在系统运行期间不出现自动终止，程序bug等；

(2)系统的可维护性需求：系统的可维护性需求与系统的稳定性需求是分不开的，当系统运行不稳定，出现数据丢失错误、运行中断时，要保证能够恢复或者还原[5]；

(3)系统的易用性需求：系统要界面美观、功能适用、使用简洁；

(4)系统的完整性需求：系统除了有必要的功能界面，还应该有数据管理界面，如：用户信息界面，后台管理界面等。

(5)系统的可扩展性需求：后期有必要改进系统的开发和建设，在系统设计和开发之前应该充分考虑到系统的可扩展性需求。

## 3.3 条件与限制

只有注册用户才能收藏诗人诗词或进行评论，游客通过注册之后才能使用这些功能。只有系统的管理员拥有数据库中的用户信息与诗词及诗人数据增删查改操作的权限，注册用户只能查看、修改部分个人信息。

## 3.4 系统用例图

本系统共有三种参与者——管理员、游客与注册用户。管理员拥有用户信息的增删查权限与诗人诗词数据增删查改的权限；游客只有通过注册之后，才能拥有收藏，评论等权限，注册用户拥有查看、修改个人信息等功能。

图示

描述已自动生成

图3-1 管理员用例图

图示

描述已自动生成

图3-2 游客、注册用户用例图

# 4．概要设计

## 4.1 处理流程

本部分主要介绍用户使用流程，主要包括诗词赏析，诗词收藏，诗词评论，趣味模块和聊天室。

### 4.1.1 用户使用全流程

该系统最根本的目的是实现诗词的管理，展示。传统美文学习网站是供广大人群学习，欣赏的网站，因此该网站应能满足用户的常规需求，如用户注册，登录，阅读古诗文，根据关键字查找古诗文、名句、通过视频学习古诗文，提供成语接龙，藏头诗，歇后语等趣味小游戏。根据时代查看诗人，根据诗词分类查看诗词。诗词下方展示推荐诗词，诗人下方展示推荐诗人。用户可收藏，取消收藏诗人和诗歌。对于网站管理者能添加、纠正，删除古诗文，古诗文分类和诗人信息，管理管理员信息，设置网页轮播图等。

该网站属于学习交流型站，众多的用户访问网站不仅可以学习到古诗词，还能收藏喜欢的诗词，诗人。网站精选优秀的古诗词，大多数人都能在这里找到自己喜欢的古诗词，每个古诗词后面还有经典的赏析。还可以通过搜索功能，以关键字进行搜索，查询你需要的诗词。

图示

描述已自动生成

图4-1 系统流程图

## 4.2 总体结构和模块外部设计

考虑到系统的简单易操作性，本文将系统的总体功能结构拆分为：用户注册、用户登录、诗词浏览、诗词查询、诗词诗人分类、分类结果展示、后台管理等模块，详见表4-2。

表 4-2 总体结构及模块表

|  |  |  |
| --- | --- | --- |
| **编号** | **功能名称** | **功能描述** |
| **1** | **用户注册** | 检验和完成用户注册并保存用户注册信息 |
| **2** | **用户登录** | 完成用户登录和检验用户登录信息 |
| **3** | **用户信息** | 用户登录以后可以对自己的个人信息进行维护、修改 |
| **4** | **诗词浏览** | 首页展示所有诗词列表 |
| **5** | **诗词查询** | 对数据库内诗词根据关键字进行查询 |
| **6** | **诗词评论** | 用户登录后可在诗词下方发表自己的评论 |
| **7** | **诗词收藏** | 用户登录后可收藏喜欢的诗人和诗词，方便查找，还可取消收藏 |
| **8** | **诗词推荐** | 用户浏览诗词时系统能根据诗词种类推荐相关的诗词 |
| **9** | **诗词分类** | 将诗词根据类别进行分类，选择相应种类可查看对应诗词 |
| **10** | **诗人分类** | 将诗人根据时代进行分类，选择相应时代可查看对应诗人 |
| **11** | **后台管理** | 对诗词和诗人进行增删改查 |
| **12** | **成语接龙** | 用户输入四字成语，系统自动回复并计算相应分值 |
| **13** | **藏头诗** | 用户输入相应的句子，系统自动生成五言或七言诗句 |
| **14** | **视频展示** | 通过视频的形式介绍相应的诗人诗词文化 |
| **15** | **聊天室** | 用户可加入聊天室，参与话题讨论 |

## 4.3 功能分配

根据上一节的系统功能需求设计，现把基于跨平台轻量级诗词系统的功能需求具体划分为用户注册、登录、自助服务、新闻查询、新闻或评论分析、后台管理六大模块，其与程序结构的关系如下表4-2所示。

表 4-2 各项功能与程序结构的关系表

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **程**  **功 序**  **能 结**  **需 构**  **求** | **用户注册** | **用户登录** | **用户信息** | **诗词查询** | **诗词浏览** | **诗词评论** | **成语接龙** | **藏头诗** | **视频展示** | **诗人诗词分类** | **聊天室** |
| **用户注册** | √ |  |  |  |  |  |  |  |  |  |  |
| **用户登录** |  | √ |  |  |  |  |  |  |  |  |  |
| **用户中心** |  |  | √ |  |  |  |  |  |  |  | √ |
| **分类模块** |  |  |  |  |  |  |  |  |  | √ |  |
| **诗词模块** |  |  |  | √ | √ | √ |  |  |  |  |  |
| **趣味模块** |  |  |  |  |  |  | √ | √ | √ |  |  |

## 4.4 接口设计

## 判断一个接口是否优秀，并不是简单地根据第一个版本给出判断的，而是要看随着时间的推移，该接口是否还能存在，是否仍旧保持得不错。槽糕的接口各种各样，但是好的接口对于用户来说必须满足以下几个点：

1. 易学习：有完善的文档及提供尽可能多的示例和可copy－paste的代码，像其他设计工作一样，你应该应用最小惊讶原则。

2. 易使用：没有复杂的程序、复杂的细节，易于学习；灵活的接口允许按字段排序、可自定义分页、 排序和筛选等。一个完整的接口意味着被期望的功能都包含在内。

3. 难误用：对详细的错误提示，有些经验的用户可以直接使用接口而不需要阅读文档。

4. 易阅读：代码的编写只需要一次一次，但是当调试或者修改的时候都需要对代码进行阅读。

5. 易开发：个最小化的接口是使用尽可能少的类以及尽可能少的类成员。这样使得理解、记忆、调试以及改变接口更容易。

本系统接口分为外部接口和内部接口两类，外部接口是调用的第三方接口，内部接口是自己开发的接口

### 4.4.1 外部接口

(1) 图片/视频上传

图片上传主要借助阿里云的对象存储。因为浏览器本身的限制，浏览器是不能直接操作文件系统的，需要通过浏览器所暴露出来的统一接口，由用户主动授权发起来访问文件动作，然后读取文件内容进指定内存里，最后执行提交请求操作，将内存里的文件内容数据上传到服务端，最后服务端解析前端传来的数据信息后存入服务器里。

(2) 成语接龙/歇后语接口

此接口借助于第三方接口，可实现通过此接口可拿到丰富的成语

### 4.4.2 内部接口

所有关于用户，诗人，诗词等其他接口都是系统内部实现，能保证安全性也方便后期维护

## 4.5 数据结构设计

该部分主要给出系统所使用的各数据结构的设计及其关系，包括逻辑结构设计、物理结构设计。

### 4.5.1 逻辑结构设计

本系统主要包括管理员、一般用户、新闻数据三个实体，其中管理员与用户和新闻数据之间存在着管理关系，一般用户与新闻数据存在查看关系，具体设计如图4-3所示。

图示

描述已自动生成

图4-3 系统主要组成E-R关系图

图示

描述已自动生成

图4-4 广告E-R关系图

图示

描述已自动生成

图4-5 管理员E-R关系图

### 4.5.2 物理结构设计

本系统主要设计了八张数据库表——用户表（users）、管理员表（adminusers），诗歌分类表(poetrytypes),诗歌表(poems),广告表(ads),诗人表(poets),时代表(longs),视频表(videos),具体结构设计如表4-4、4-5所示。

表 4-3 数据库表单

|  |  |
| --- | --- |
| **数据表名** | **注释** |
| users | 用户表 |
| adminusers | 管理员表 |
| poetrytypes | 诗歌分类表 |
| poems | 诗歌表 |
| ads | 广告表 |
| poets | 诗人表 |
| longs | 时代表 |
| videos | 视频表 |

数据库名：vue\_poetry

表4-4 adminusers（管理员表）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| **\_id** | String | 否 | 无 | 主键 管理员id |
| username | String | 否 | 无 | 管理员登录名 |
| password | String | 否 | 无 | 管理员登录密码 |
| description | String | 是 | 无 | 管理员描述 |
| phone | String | 否 | 无 | 管理员电话 |

表4-5 ads（广告表）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 广告id |
| name | String | 否 | 无 | 广告标题 |
| image | String | 否 | 无 | 广告轮播图 |
| body | String | 否 | 无 | 广告内容 |
| url | String | 否 | 无 | 广告跳转链接 |

表4-6 poetrytypes（诗歌分类表）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 诗歌分类id |
| name | String | 否 | 无 | 诗歌分类名称 |
| banner | String | 否 | 无 | 诗歌分类背景图片 |
| description | String | 否 | 无 | 诗歌分类描述 |

表4-7 poets诗人表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 诗人id |
| name | String | 否 | 无 | 诗人名字 |
| title | String | 是 | 无 | 诗人称号 |
| description | String | 否 | 无 | 诗人描述 |
| head\_portrait | String | 否 | 无 | 诗人头像图片 |
| time | ObjectId | 否 | 无 | 诗人所属时代的id |

表4-8 poems诗歌表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 诗歌id |
| name | String | 否 | 无 | 诗词名字 |
| banner | String | 否 | 无 | 诗词背景图 |
| poetry\_types | Array | 否 | 无 | 诗歌所属分类数组 |
| poet | ObjectId | 否 | 无 | 诗歌所属诗人id |
| content | String | 否 | 无 | 诗歌内容 |
| description | String | 否 | 无 | 诗歌描述 |
| time | ObjectId | 否 | 无 | 诗歌所属时代id |

表4-9 longs时代表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 时代id |
| name | String | 否 | 无 | 诗词名字 |

表4-9 users用户表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 用户id |
| username | String | 否 | 无 | 用户名 |
| password | String | 否 | 无 | 用户密码 |
| icon | String | 否 | 无 | 用户头像 |
| like\_poems | Array | 否 | 无 | 收藏诗歌数组 |
| like\_poets | Array | 否 | 无 | 收藏诗人数组 |

表4-9 videos视频表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 是否允许空 | 默认值 | 备注（主外键、字段说明、值域） |
| \_id | String | 否 | 无 | 主键 视频id |
| name | String | 否 | 无 | 视频名称 |
| video | String | 否 | 无 | 视频链接地址 |

## 4.6 运行设计

该部分描述在一个系统中施加不同的外部操作时所引起的各种不同的运行设计。包括运行机制、运行控制两个部分。

### 4.6.1 运行机制

用户首先访问网站，进入首页后系统会自动查询数据库中所有诗词，将诗词数据用列表展示，只展示关键信息，用户如果对诗词感兴趣可以点击查看详情，诗词详情界面有诗词解析，诗词推荐，还可以发表自己的评论，用户如果想分享可直接点击复制链接将网址发给别人，喜欢诗词还可登陆后进行收藏，方便下次查找，系统还将诗人和诗词进行分类展示，给用户提供良好的交互，为了使网站更有意义，还增加了趣味模块，用户可进入趣味模块，体验成语接龙，歇后语，藏头诗的体验，还有视频方便传播知识，个人中心可查看可更新个人信息，收藏的诗人和诗词信息可在个人中心查看，还增加了独特的聊天室功能，供用户在网上随时交流自己的想法

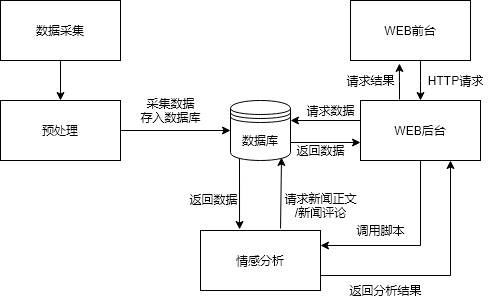


图4-4 系统运行机制图

### 4.6.2 运行控制

本文的系统运行控制是按照事务所属的模块来划分的，包括登录、注册、新闻显示、新闻或评论分析结果展示等，详情如表4-6所示。

表 4-6 运行控制表

|  |  |
| --- | --- |
| **事务名** | **事件流** |
| **登录模块** | |
| 登录 | 1. 用户登录：通过输入用户名，密码，连接数据库查询是否存在此用户。若有，则记住登录用户的类型并转到前台页面，若无，则提示出错信息 2. 管理员登录：通过输入管理员名，密码，验证码，连接数据库查询此登录名是否为管理员。若是，则转到后台管理页面 |
| **注册模块** | |
| 注册 | 游客按要求填写注册信息，系统判断输入的用户名是否已被注册，若未被注册则判断用户的注册信息是否符合要求，如果符合则注册成功，否则注册失败 |
| **普通游客模块** | |
| 新闻显示 | 用户进入系统以后，系统自动从数据库中读取信息，并且在首页新闻显示部分按新闻类别显示新闻标题信息 |
| **用户模块** | |
| 新闻分析选择及结果展示 | 用户可选择新闻分析其正文并查看分析结果 |
| 评论分析及结果展示 | 用户可选择新闻分析其评论并查看分析结果 |
| 用户自助服务管理 | 1、用户对个人信息进行修改  2、用户对个人登录密码进行修改 |
| **管理员模块** | |
| 1. 用户管理 | 分配给操作系统的用户不同的角色，包括管理员及用户，不同的角色可进行不同的操作 |
| 1. 新闻管理 | 管理员修改、添加、删除新闻信息 |

## 4.7 出错处理设计

### 4.7.1 出错输出信息

表 4-7 出错信息表

| **类别** | **发生情况** | **系统输出信息** | **处理方法** |
| --- | --- | --- | --- |
| 用户注册 | 用户输入的用户名已经存在 | 页面提示“该用户名已经存在！” | 跳转到注册页面，用户更改用户名 |
| 用户未将必须的注册相关内容填写完整 | 页面提示“……不能为空，请重新输入！” | 跳转到注册页面，需要用户将个人注册信息填写完整 |
| 两次输入的密码不相同 | 页面提示“两次输入的密码不一致，请重新输入！” | 跳转到注册页面，用户重新输入一致的登录密码 |
| 用户登录 | 验证码错误 | 页面提示“验证码错误，请重新输入！” | 跳转到登录页面 |
| 用户输入的用户名或密码不正确 | 页面提示“用户名或密码不正确，请重新输入！” | 跳转到登录页面，用户重新输入正确的用户名和密码 |
| 用户信息修改 | 两次输入的密码不相同 | 页面提示“两次输入的密码不一致，请重新输入！” | 跳转到用户修改密码页面 |
| 用户没有将必须需要的注册信息填写完整 | 页面提示“……不能为空，请重新输入！” | 停留在修改页面，需要用户将个人信息填写完整 |
| 用户未登录 | 未登录用户进分析结果展示页面 | 页面提示“您还未登录，不能查看分析结果！” | 跳转到首页 |
| 用户或管理员密码修改 | 提供错误的当前密码 | 页面提示“当前密码不匹配，请重输入！” | 跳转到用户或管理员密码修改页面 |
| 两次输入的密码不相同 | 页面提示“两次输入的密码不一致，请重新输入！” | 跳转到用户或管理员密码修改页面 |
| 未提供新设定的当前密码 | 页面提示“新密码不能为空！” | 跳转到用户/管理员修改密码页面 |
| 管理员登录 | 验证码错误 | 页面提示“验证码错误，请重新输入！” | 跳转到登录页面 |
| 管理员输入的账号或密码不正确 | 页面提示“用户名或密码不正确，请重新输入！” | 跳转到登录页面，重新输入正确的管理员账号和密码 |
| 管理员添加新闻信息 | 没有填写所有必填的新闻相关内容 | 页面提示“待添加新闻……内容不能为空！” | 跳转到添加新闻页面 |
| 管理员删除新闻信息 | 没有填写所有必须需要的新闻信息 | 页面提示“待删除新闻……不能为空” | 跳转到删除新闻页面 |
| 管理员添加用户基本信息 | 没有填写所有必须需要的用户信息 | 页面提示“待添加用户……不能为空” | 跳转到添加用户页面 |
| 管理员删除用户基本信息 | 没有填写所有必须需要的用户信息 | 页面提示“待删除用户……不能为空” | 跳转到删除用户页面 |
| 用户查看分析结果 | 没有选择待分析文本 | 页面提示“待分析文本不能为空” |  |

### 4.7.2 出错处理对策

数据采集模块采用的是网络爬虫技术采集新浪网数据，在请求网页数据时会出现请求失败，数据获取不全等情况，管理员应在数据采集后查看错误记录文件，对出问题新闻手动重新爬取信息。

## 4.8 维护设计

必须按照软件产品设计规范的步骤进行开发，并充分考虑系统的可维护性，在各个阶段编写详细的文档，以方便系统的移植或二次开发。定期维护数据库，根据信息更新修改数据库信息，保持数据库最近最新。通常使用改正性维护、适应性维护、完善性维护、预防性维护这4种维护活动来完成系统维护。

# 5. 详细设计

## 5.1 软件结构

本系统大致分为数据采集与预处理、情感分析、用户交互三大模块。

数据采集与预处理模块包括新闻正文采集、新闻评论采集、数据清洗、数据存储四个功能，新闻正文采集、新闻评论采集的功能是通过网络爬虫对新浪网新闻及其评论网页内容进行抓取，数据清洗的功能是提取出抓取内容里的新闻id、新闻标题、新闻内容、新闻评论文本、评论时间、评论点赞数等信息，数据存储的功能是将清洗过的新闻数据信息保存到数据库，评论信息以JSON文件的形式保存。情感分析模块包括新闻正文情感分析与新闻评论情感分析，是本系统的主要模块。用户交互模块包括新闻数据展示、数据分析结果展示、用户管理界面、新闻数据管理界面等功能。

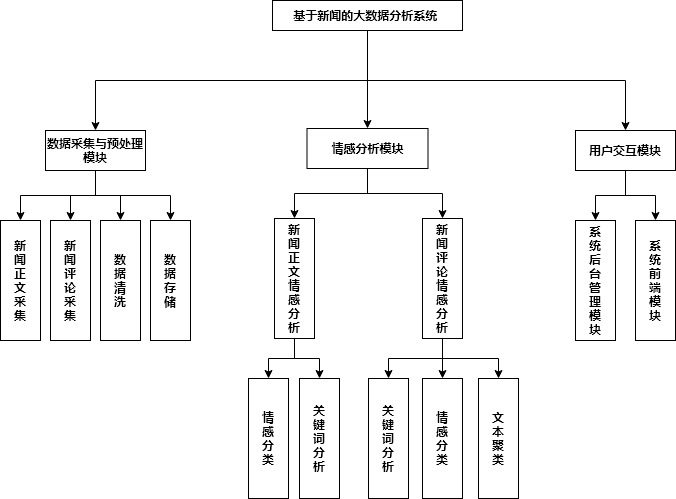


图5-1 软件结构图

## 5.2 数据采集与预处理模块

该模块的功能是从网络中采集新闻信息及其评论信息并进行数据清洗与保存，其数据来源是新浪网新闻中心全部类型的新闻及其评论。

### 5.2.1 新闻爬取模块

(1) 程序描述

该模块的功能是采集新闻信息，包括新闻的标题、正文内容、关键字等。

新闻正文获取：通过requests模块获取网页信息，然后使用Beautiful Soup库的select()方法定位并获取新闻正文。

新闻其他信息获取：通过Request模块获取网页js文件脚本信息然后将获取的json数据转换为字典，提取出所需信息。

(2) 输入项

表5-1 输入项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **数据来源** | **数据处理** |
| 新闻类别 | news\_kind | String | 新浪网 | 添加到url |
| 新闻页面URL | news\_url | String | 新浪网 | 请求网页信息 |
| 新闻其他信息URL | parse\_url | String | 新浪网 | 请求网页信息 |

(3) 输出项

表5-2 输出项表

| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| --- | --- | --- | --- | --- |
| 新闻id | n\_id | String | 写入数据库 | 数据库 |
| 新闻标题 | title | String | 写入数据库 | 数据库 |
| 新闻内容 | article | String | 写入数据库 | 数据库 |
| 新闻关键字 | keyword | String | 写入数据库 | 数据库 |
| 发布时间 | ctime | datetime | 写入数据库 | 数据库 |
| 发布媒体 | media\_name | String | 写入数据库 | 数据库 |
| 新闻链接 | url | String | 写入数据库 | 数据库 |
| 新闻类型 | kind | int | 写入数据库 | 数据库 |
| 评论文件地址 | comment | String | 写入数据库 | 数据库 |

(4) 逻辑流程

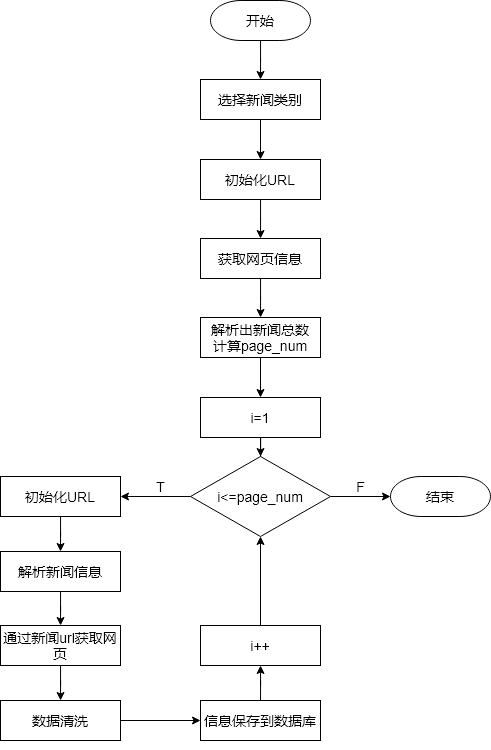


图5-2 新闻爬取流程图

(5) 测试计划

点击新闻分类选择模块按钮，确保新闻类别列表能正确显示，点击某个分类，确保脚本能正确爬取新闻信息及新闻内容。

### 5.2.2 新闻评论爬取模块

(1) 程序描述

通过Request模块获取新闻详情网页的评论js文件脚本信息，将获取的json数据转换为字典，解析出该新闻的总评论数，若小于200条则一次性请求数据，若大于200则按200条/页再次请求数据，然后将获取的json数据转换为字典，提取出所需信息。当数据量大于200时，由于要多次请求网页数据，为避免频繁访问，设置每次请求后将休息5秒。

(2) 输入项

表5-3 输入项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **数据来源** | **数据处理** |
| 新闻id | news\_id | String | 新浪网 | 添加到url |
| 新闻频道 | channel | String | 新浪网 | 添加到url |
| 解析地址 | parse\_url | String | 新浪网 | 请求网页信息 |

(3) 输出项

表5-4 输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 评论id | comment\_id | String | 写入文件 | json |
| 评论内容 | content | String | 写入文件 | json |
| 评论设计 | time | String | 写入文件 | json |
| 分布位置 | area | String | 写入文件 | json |
| 点赞数 | agree | String | 写入文件 | json |

(4) 逻辑流程

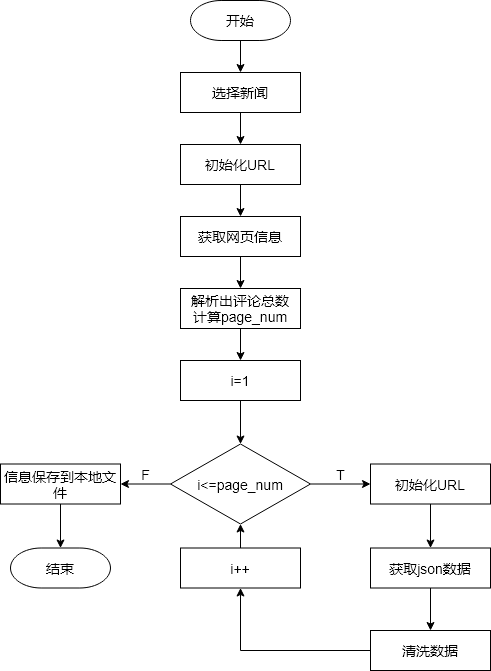


图5-3 新闻评论爬取流程图

(5) 测试计划

指定一新闻URL，确保脚本能正确爬取新闻评论信息及新闻评论内容。

## 5.3 情感分析模块

该模块主要是研究对新闻及其评论文本的情感分析，其包括情感分类与关键词分析。

情感分类方面，本系统采用了基于依存句法分析的方法与基于深度学习的方法。

关键词分析方面，本系统采用的是北京大学开源分词工具pkuseg、TextRank算法与AP聚类算法。pkuseg主要用在分词阶段，pkuseg与常用的jieba分词对比，在中文新闻文本分词中有着更高的准确率；TextRank算法主要用于文本关键词提取，它是一个基于图排序的、用于文本处理的算法，考虑了相邻词的语义关系，可以更好的表达文本，更准确的提取文本中关键词[7]； AP聚类算法主要用于某一个新闻或者某一类新闻的评论文本的聚类，AP聚类算法是一种基于数据点间的"信息传递"的聚类算法[8]， 每个数据点都被视为图的一个节点，并且图的有向边被视为节点之间消息的传递，目的是发现到所有数据点的总相似度最大的最优类代表点集合。

### 5.3.1 新闻情感分类模块

(1) 程序描述

新闻文本的情感分类是通过依存句法分析的方法计算文本情感得分来实现的，得分为0代表中性，得分在[-1,0)之间为负面，得分在（0，1]之间为正面，其处理流程如下：

用户登录系统，进行新闻选择界面，选择想要分析的新闻，新闻分析脚本接收待分析新闻的id，脚本从数据库读取新闻信息进行情感分类、关键词分析，将分析结果返回到前台展示。

(2) 输入项

表5-5输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **数据处理** |
| 新闻id | n\_id | String | 传参 | 网页爬取 | 查询数据库 |
| 新闻内容 | content | String | 传参 | 数据库查询 | 分句分词 |

(3) 输出项

表5-6 输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 情感得分 | score\_result | int | 写入txt文件 | 文本 |
| 词云 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 情感曲线 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 关键词排名 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 情感比例图 | n\_id(.HTML) | 文件 | HTML文件 | 网页 |

(4) 情感词典设计

如图5-4所示，本文的情感词典由知网情感词典、清华大学李军中文褒贬义词典、BosonNLP情感词典、程度副词词典四部分组成，词典内容包含两个部分——词语，权重。

如图5-5所示，本文所使用的情感词典中积极词占比约为71%，消极词占比约为29%。

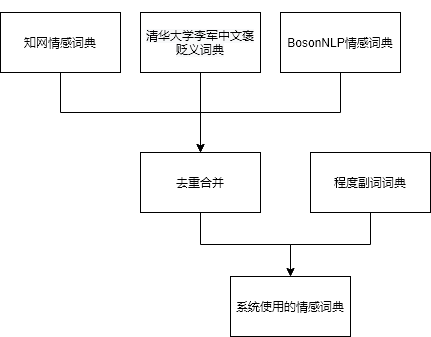
****

图5-4 情感词典组成图

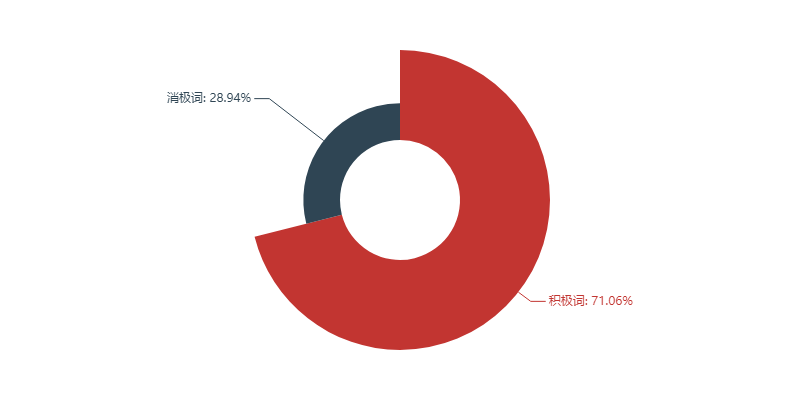
****

图5-5 情感词比例图

(5) 算法设计

依存句法分析经过剖析句子的成分之间的依存关系来展示其句法结构[9]，与将句子分为主语、形容词和语法元素的传统句法分析相比，依存句法认为句子的不同组成部分（词语）之间存在依存关系，其首先专注于句子的动词部分，然后分析动词与句子中其他成分之间的关系。依存关系是指句子中词与词之间支配与被支配的关系，其既可以是词与词之间的句法关系，也可以是词与词之间的语义关系。

基于依存句法分析的情感分类方法基本流程如下：

1）通过pkuseg工具对文本进行分句分词操作，并借助情感词典完成情感句过滤；

2）通过哈工大 LTP 语言云平台对过滤后的情感句进行词性标注、依存句法分析，最终从待处理文本提取出[词，词性，依存弧的父节点词，父节点词词性，关系]列表；

3）借助情感词典与程度副词词典计算情感句情感得分，然后通过加权整合，作为文本的情感分数。

(6) 逻辑流程

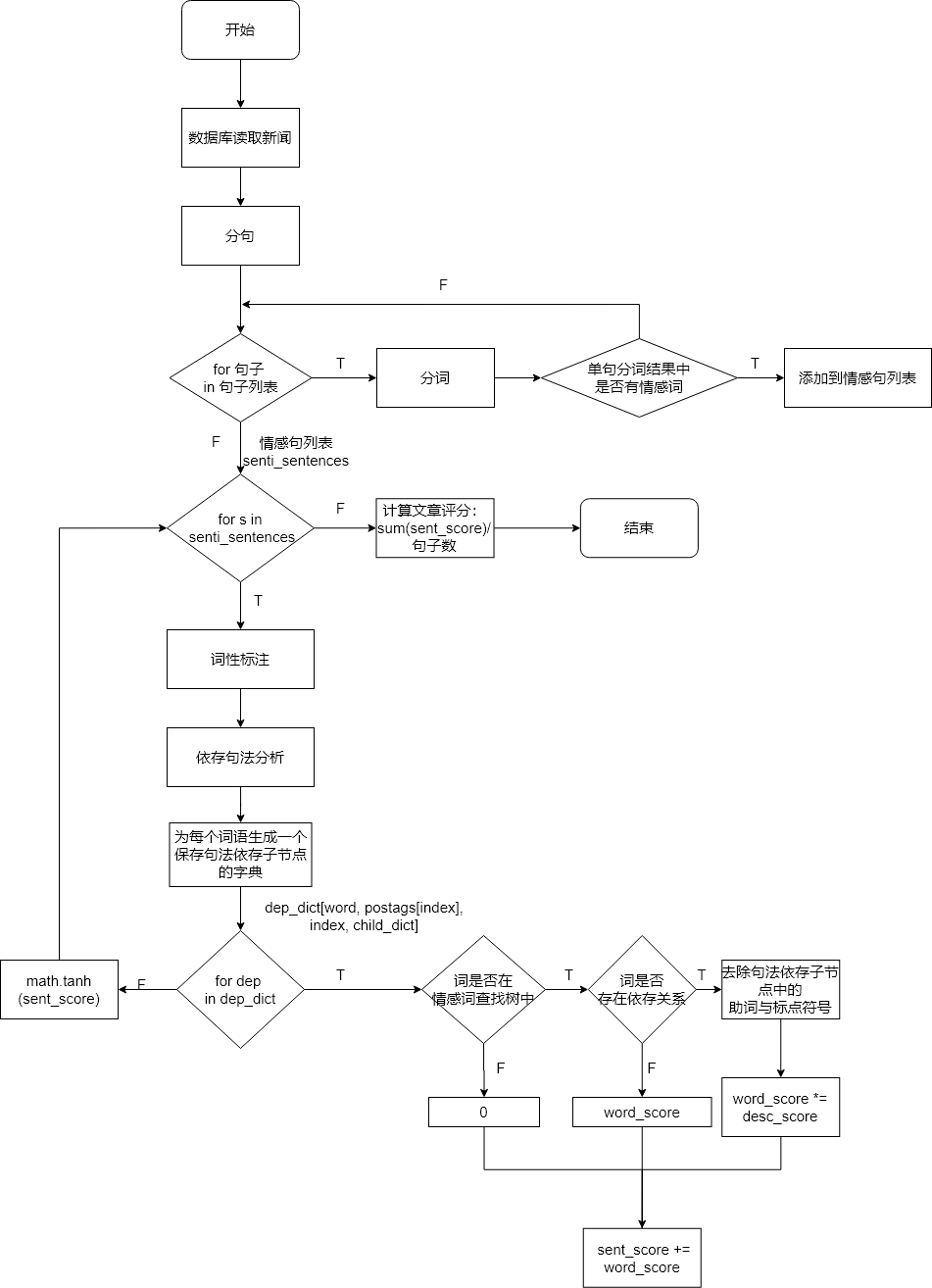


图5-6 新闻情感分类流程图

(7) 测试计划

挑选5到10篇情感趋向明显的新闻报道进行情感分类，判断结果与人工分类结果是否一致，计算算法准确率。

### 5.3.2 评论情感分类模块

(1) 程序描述

本文的评论情感分类使用的是基于深度学习模型的情感分析方法，是基于LSTM神经网络搭建的新闻评论文本情感分类的深度学习模型。

处理流程：用户登录系统，进行新闻选择界面，选择想要分析的新闻，评论分析脚本接收待分析的新闻id，脚本从数据库读取新闻对应评论信息文件存储位置，读取评论信息文件信息进行情感分类、关键词分析，将分析结果返回到前台展示。

(2) 输入项

表5-7 输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **数据处理** |
| 新闻id | n\_id | String | 传参 | 网页爬取 | 查询数据库 |
| 评论文件位置 | comment\_file | String | 传参 | 数据库查询 | 文件读取 |

(3) 输出项

表5-8 输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 情感趋向 | predict\_result | int | 写入txt文件 | 文本 |
| 词云 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 情感曲线 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 关键词排名 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |

(4) 训练词向量

本文借助Word2Vec方法训练词向量，以网络抓取的新闻评论作为训练语料，使用jieba分词对语料中的每个句子分词并组成词列表，然后使用gensim工具包的word2vec开始训练词向量。鉴于训练语料较大，故本文采用的是更为成熟的jieba分词工具进行分词，可大大节省时间。本文训练忽略词频低于10的词的统计，得到的是100维的词向量。训练步骤如下：

1.建立一个空的模型对象；

2.遍历一次语料库建立词典；

3.第二次遍历语料库建立word2vec模型。

(5) 训练分类模型

本文使用的分类模型是基于LSTM的，基于神经网络的分类模型需要给模型提供已人工分类的标注语料做训练，本文采用的训练语料是网络中开源的评论标注语料，包括谭松波老师整理的酒店评论语料集以及其他开源的评论语料。其中正面评论共8000条，负面评论共8700条，中性评论共4300条。本文的语料使用划分为70%作为模型的训练数据集，30%作为模型测试数据集。

LSTM是一种特殊的RNN网络，该网络设计出来是为了解决长依赖问题。如图5-7为LSMT的一个单元。如图5-8所示，LSTM的细胞状态由三个门进行控制——遗忘门、输入门和输出门[10]；

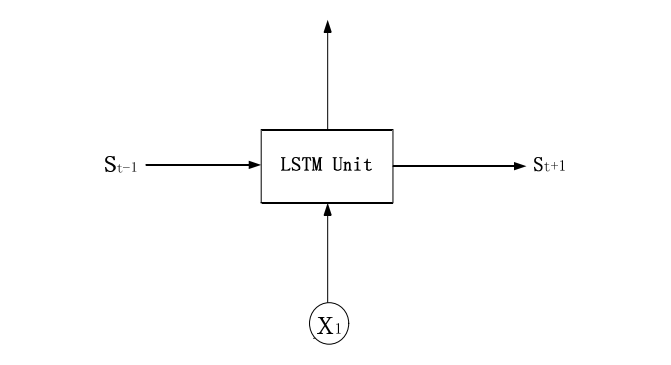


图5-7 LSTM网络单元

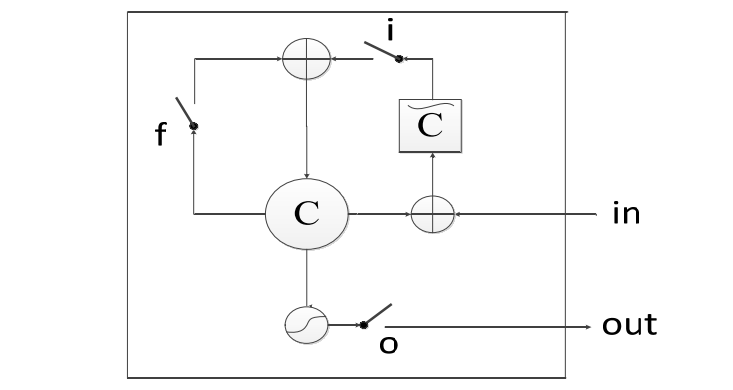


图5-8 LSTM网络结构

遗忘门决定cell需要丢弃哪些信息，其计算公式为：



输入门决定需要更新哪些信息，更新的规则就是通过遗忘门选择遗忘旧细胞的一部分信息，通过输入门选择添加候选细胞信息的一部分得到新的细胞信息，其计算公式为:



输出门判断输出细胞的哪些状态特征，其计算公式为



表示t-1 时刻LSTM 单元的输出，W 表示系数矩阵，b 为偏置向量，σ为标准 sigmoid 函数[11]。

本文使用的LSTM模型结构如图5-8所示，其输入的向量维度为100维，输出的为50维。本文使用的分类模型损失函数选取的是categorical\_crossentropy，优化方法选取的是adam，其创建流程如下：

1)将所有词的向量集传入 embedding层，由于训练集数量较大，使用嵌入式词向量来表示单词更有效率。

2)通过embedding 层, 将新的词表示传入 LSTM cells。

3)LSTM cells连接一个softmax output layer 。

模型训练中，本文训练轮数选取的是5，批次大小选取的是32，经过5轮训练之后，语料数据在验证集上的准确率达到了88%。

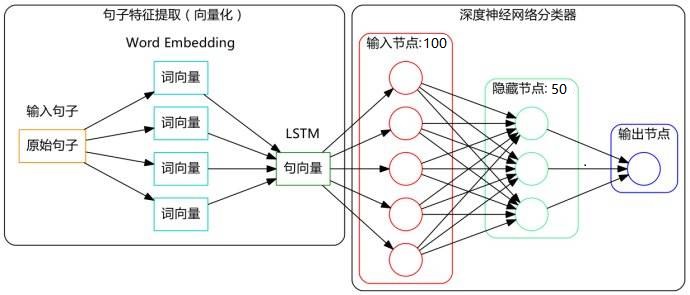
****

图5-8 LSTM模型结构图

(6) 逻辑流程

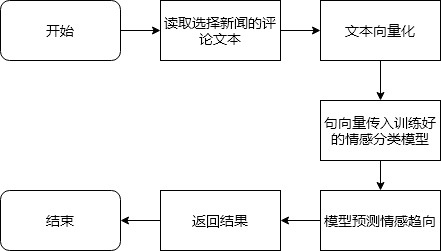
****

图5-9 评论情感分类流程图

(7) 测试计划

选择新闻，点击评论分析按钮，确保能跳转到数据分析结果页面，并能完整展示分析结果。

### 5.3.3 文本聚类模块

本文通过AP聚类算法实现某一个新闻或者某一类新闻的评论文本聚类，其基本思想是将所有的数据点都视为潜在的聚类中心，然后将数据点两两相连以形成网络(相似性矩阵)，再通过网络中各条边的消息的吸引度与归属度来计算各样本的聚类中心[12]。

1. 程序描述

本文通过词向量和词向量对应的tf-idf 值构建文本向量，通过计算两个文本向量的余弦相似性得到文本之间的相似性，最后通过AP聚类算法得到观点数量，并从观点中筛选典型评论[13]。

AP聚类算法基本流程如图5-10所示：

1）读取评论文本，并根据已训练好的word2vec模型生成文本向量；

2）计算相似度矩阵；

3）调用sklearn包中的AP算法在相似度矩阵上执行聚类并返回聚类标签。

典型评论筛选基本流程如图5-11所示：

1. 从已近分好类的评论中读取评论文本；
2. 分别提取每一个分类下的文本关键词并向量化；
3. 将对应分类中的每句评论分别分词并向量化；

4）计算每句评论与其对应的类标签关键词的相似度；

5）挑选出每个分类下与其关键词最大相似的评论作为典型评论。

(2) 输入项

表5-9 评论文本聚类输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **处理数据** |
| 评论文本 | comment | String | 文件读取 | 数据爬取 | 添加到实体 |
| word2vec模型 | word\_model | 模型 | 模型加载 | 模型训练 | 模型加载 |

(3) 输出项

表5-10 评论文本聚类输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 评论标签 | label | int | 写入文件 | 文件 |
| 典型评论 | typical\_comments | String | 写入文件 | 文件 |

(4) 逻辑流程

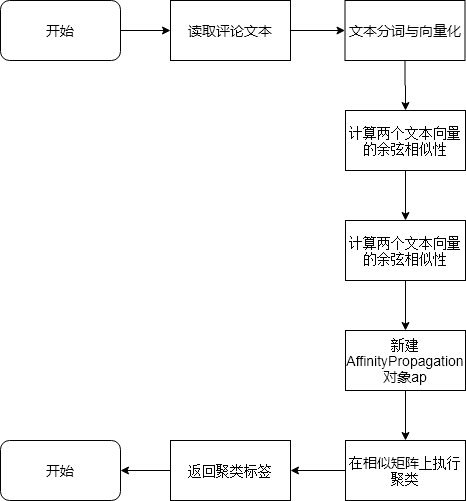
****

图5-10 AP聚类流程图

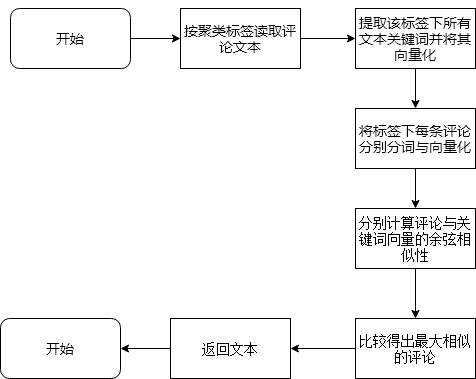
****

图5-11 典型评论筛选流程图

(5) 测试计划

人工分类好一批评论文本作为测试数据，将测试数据传入到程序，比对程序的分类结果，计算准确度。

## 5.4 用户交互模块

### 5.4.1 后台用户管理模块

(1) 程序描述

该模块实现系统管理员对用户信息的管理的设计，包括添加用户信息、修改用户信息和删除用户信息三个功能点。管理员登录系统，进入后台管理首页，点击菜单栏中的用户管理，弹出下拉选项，点击下拉选项中的添加用户信息，进入用户信息添加页面，输入用户信息，进行用户信息的添加，添加成功后页面跳转到用户列表页面；点击下拉选项中的用户信息修改，跳转到用户列表页面，点击列表中的修改按钮，跳转到修改页面，对用户信息进行修改。

(2) 输入项

表5-11 用户信息添加输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **处理数据** |
| 用户id | id | int | 键盘 | 用户 | 添加到实体 |
| 用户名 | username | varchar | 键盘 | 用户 | 添加到实体 |
| 登陆密码 | password | varchar | 键盘 | 用户 | 添加到实体 |
| 手机号 | phone | varchar | 键盘 | 用户 | 添加到实体 |
| 添加 | add\_bt | -- | 点击按钮 | -- | 添加到数据库 |
| 取消 | Cancel\_bt | -- | 点击按钮 | -- | 清空数据 |

表5-12 用户信息修改输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **处理数据** |
| 用户id | id | int | 键盘 | 用户 | 添加到实体 |
| 用户名 | username | varchar | 键盘 | 用户 | 添加到实体 |
| 登陆密码 | password | varchar | 键盘 | 用户 | 添加到实体 |
| 手机号 | phone | varchar | 键盘 | 用户 | 添加到实体 |
| 权限 | level | int | 键盘 | 用户 | 添加到实体 |
| 确定 | update\_bt | -- | 点击按钮 | -- | 替换库中数据 |
| 取消 | Cancel\_bt | -- | 点击按钮 | -- | 清空数据 |

(3) 输出项

表5-13 用户信息修改输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 用户id | id | int | 写入数据库 | 数据库 |
| 用户名 | username | varchar | 写入数据库 | 数据库 |
| 登陆密码 | password | varchar | 写入数据库 | 数据库 |
| 手机号 | phone | varchar | 写入数据库 | 数据库 |
| 权限 | level | int | 写入数据库 | 数据库 |

(4) 逻辑流程

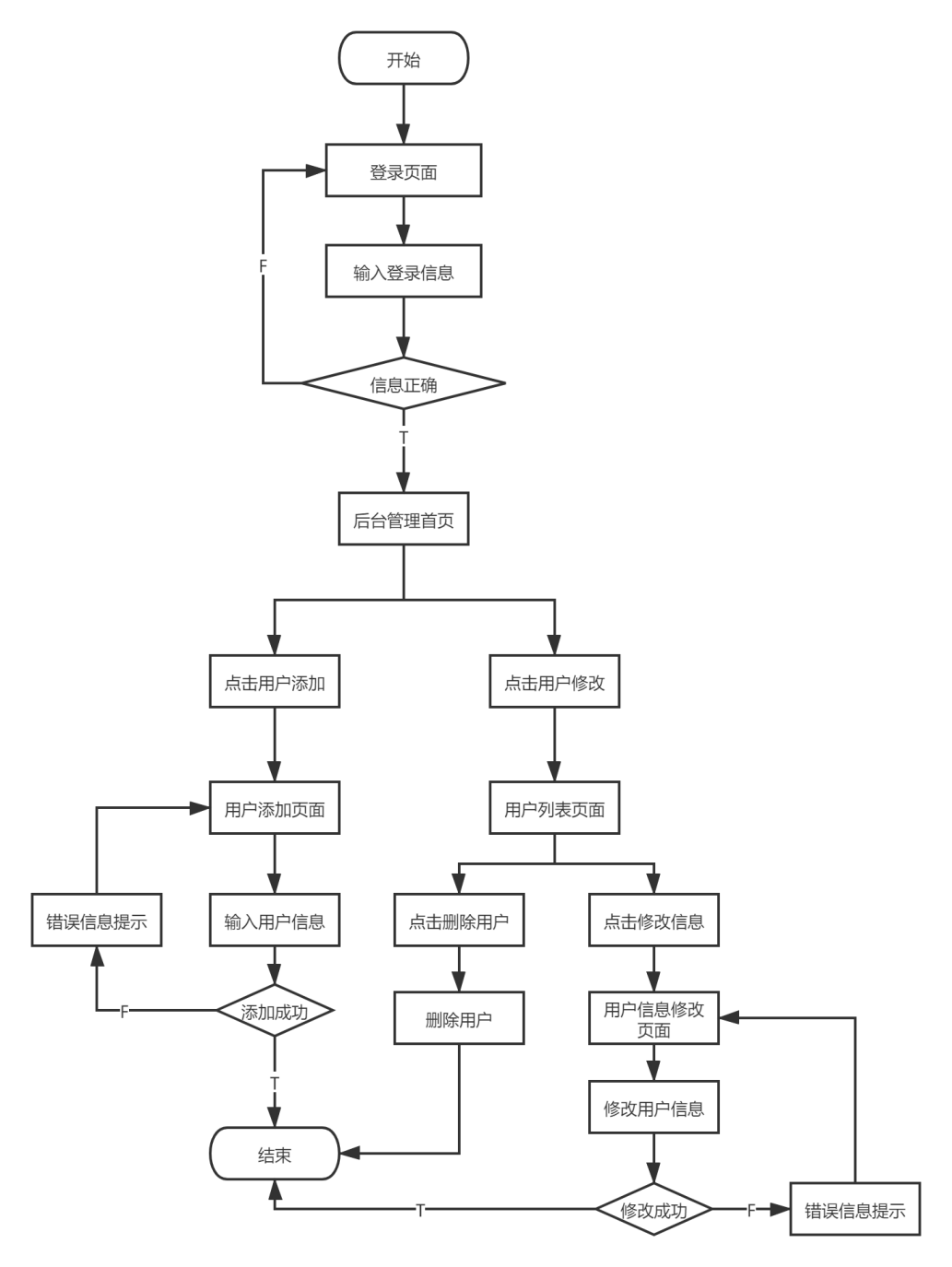


图5-12 后台用户管理模块流程图

### 5.4.2 后台新闻管理模块

(1) 程序描述

该程序完成管理员对新闻的管理，包括添加、修改和删除新闻信息。管理员登录系统，进入后台管理首页，点击新闻管理，弹出下拉列表。点击新闻添加，进入新闻添加页面，输入新闻URL，进行新闻的添加，添加成功后系统跳转到新闻列表页面，或点击新闻列表进入列表页面，在新闻列表中可点击删除进行新闻的删除，也可点击编辑，进入新闻修改页面，进行新闻的修改。

(2) 输入项

表5-14 新闻添加/修改输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **数据处理** |
| 新闻标题 | title | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻内容 | article | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻关键字 | keyword | String | 网页爬取 | 新浪网 | 添加到实体 |
| 发布时间 | ctime | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻类型 | kind | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻评论  文件位置 | comment | String | 网页爬取 | 新浪网 | 添加到实体 |
| 提交 | BtnAdd | -- | 点击按钮 | -- | 添加到数据库 |

(3) 输出项

表5-15 新闻添加/修改输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 新闻ID | NewsId | int | 写入数据库 | 数据库 |
| 信息标题 | NewsTitle | varchar | 写入数据库 | 数据库 |
| 发布人 | ReleaseMan | varchar | 写入数据库 | 数据库 |
| 发布时间 | ReleaseDate | varchar | 写入数据库 | 数据库 |

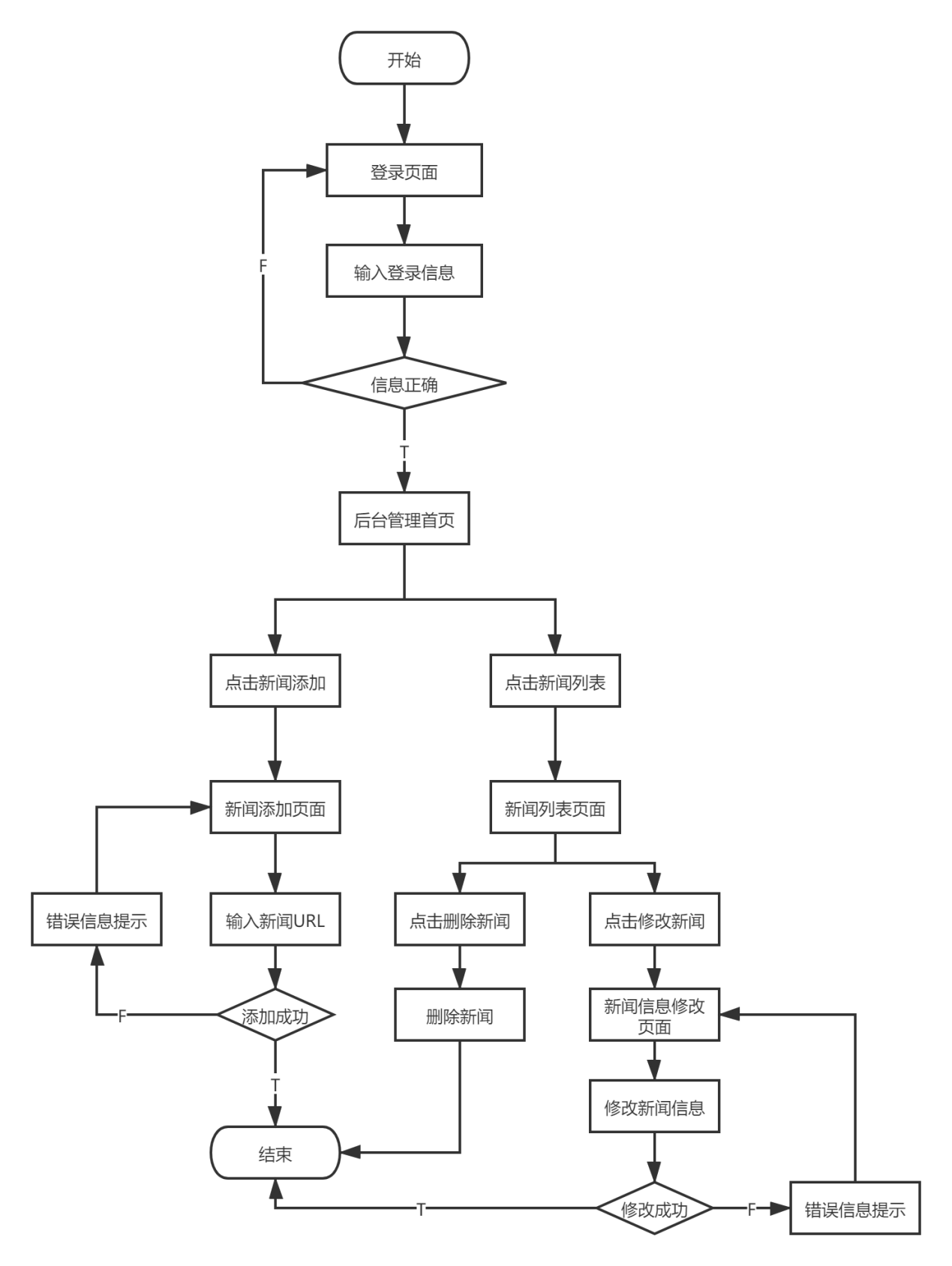
(4) 逻辑流程

图5-13 新闻管理流程图

# 6. 运行测试

## 6.1 测试计划

本文现已完成系统的设计与实现，为了确定系统的可行性和可靠性，需要进一步对其性能与功能进行测试然后对结果进行分析，本文主要就系统的数据爬取与预处理、情感分析两大模块进行功能测试与准确性测试。测试计划如下：

1. 为验证数据爬取与预处理的准确性，本文选取了多篇新闻对其进行数据抓取与预处理，判断是否能准确的抓取网页与提取出所需内容。
2. 为验证情感分类结果的准确性，本文选取五篇新闻及其评论，先对其情感趋向进行人工标注，然后与系统的分析结果一一比对，计算准确度。

## 6.2 测试结果分析

### 6.2.1 数据爬取与预处理测试

数据爬取与预处理功能测试主要是测试系统能否准确的抓取所需数据。本次测试选取的数据是2020年5月17日当天的所有国内新闻。测试结果如图6-1、6-2所示，新闻信息抓取与预处理无任何报错，评论信息共抓取1000条，抓取出现两次报错，两次都是网页请求失败，原因为请求过于频繁。

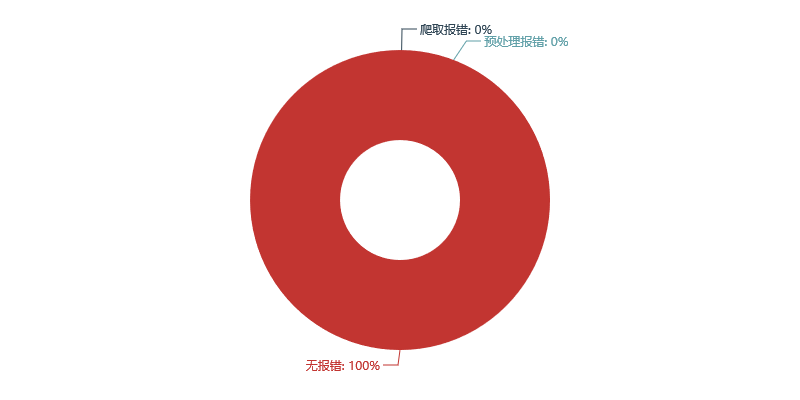


图6-1新闻爬取与预处理测试结果

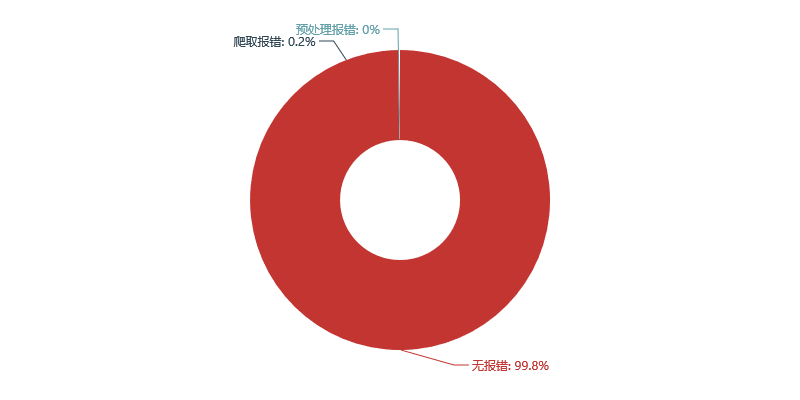


图6-2新闻评论爬取与预处理测试结果

### 6.2.2 情感分类功能测试

情感分类功能测试主要是测试系统情感分析的准确度。本次测试选用的数据是2020\05\17当天所有新闻与新浪网“嫦娥四号和“玉兔二号”结束休眠 已奋战500天”、“河底“抹水泥”，大理苍山五溪生态治理惹争议”、“武汉十答全民核酸检测：消除社会恐惧和偏见，个人免费”、“3月份以来，中国已出口口罩509亿只”、“中共中央政治局常务委员会召开会议 习近平主持”五篇新闻的评论。测试结果如图6-3、6-4所示，新闻文本准确度在90%以上，评论文本准确度在86%以上，从结果图中还可以分析出，评论分析中的中性文本判断失误的频率明显减低，说明针对中性评论的优化，起到了预期效果。

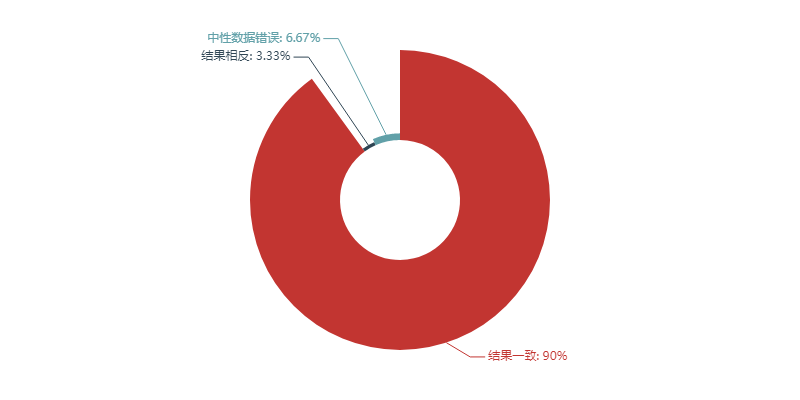


图6-3新闻文本情感分类测试结果

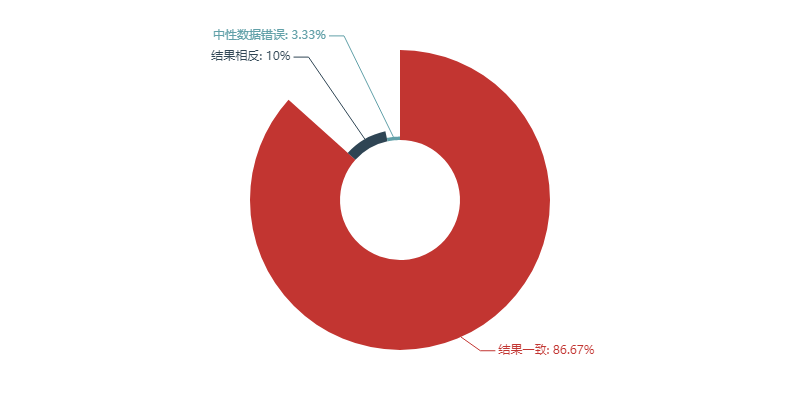


图6-4新闻评论情感分类测试结果

# 7. 运行结果

## 7.1 数据爬取与预处理运行结果

数据爬取过程如图7-1所示。管理员选择爬取新闻种类，开始爬取，按每页50条新闻爬取信息，默认爬取当前种类的200条新闻及其评论。数据预处理结果如图7-2所示，清洗爬取的网页信息，获取所需要的新闻及评论相关数据，然后插入数据库。

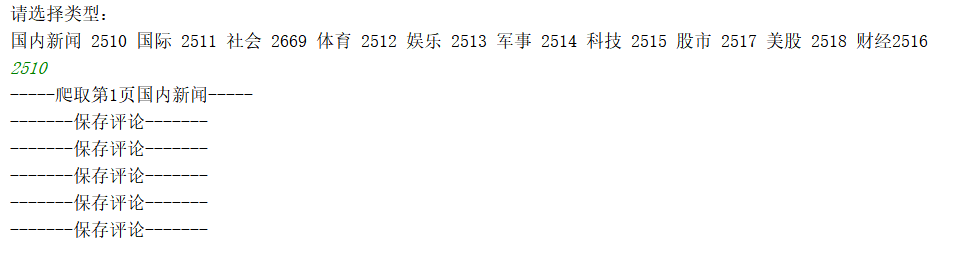


图7-1 数据爬取过程图



图7-2 预处理结果图

## 7.2 情感分类运行结果

用户传入新闻id，系统自动分析文本，得出文本每句情感分数，计算文本情感分数，并判断文本情感趋向。结果如图7-3、7-4所示，该新闻文本在8-14句、16-21句之间情感波动较小，情感得分在0.5~1之间的句子占33.33%，情感得分在-1~-0.5之间的句子占66.67%，文章总体偏负面。

用户选择新闻评论文本，系统自动分析文本，得出每句评论情感趋向，结果如图7-5所示，正面评论占39.47%，中性评论占2.63%，负面评论占57.9%

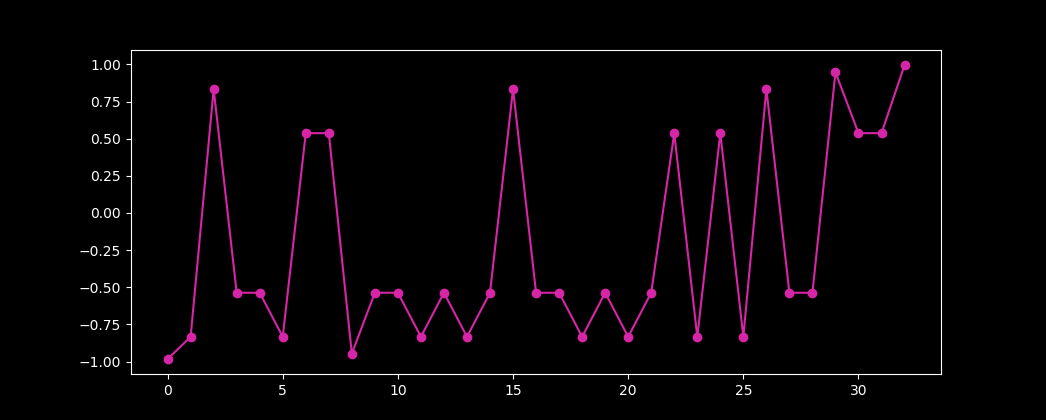


图7-3 新闻文本情感曲线图

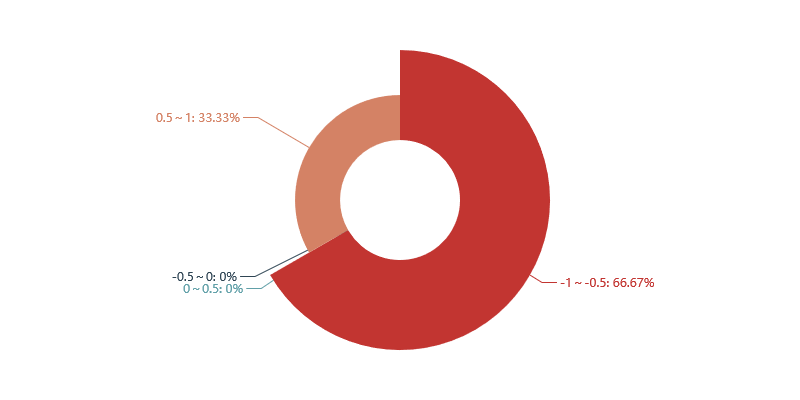


图7-4 新闻文本情感比例图

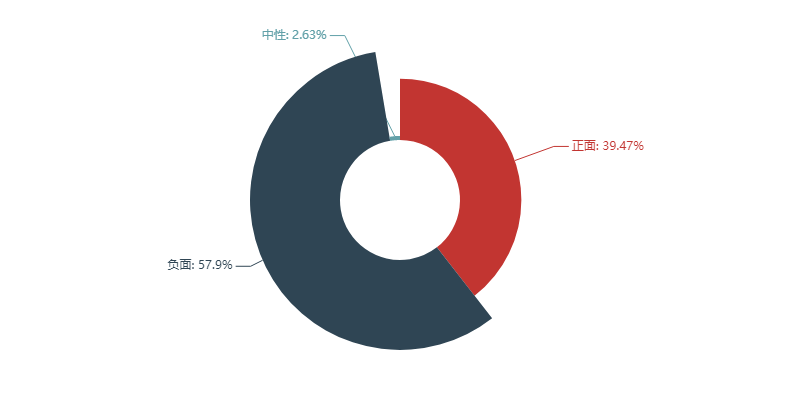


图7-5 新闻评论情感比例图

## 7.3 文本聚类运行结果

用户选择新闻，系统自动将新闻评论文本分类，并得出每个分类里面的典型评论。评论文本聚类结果如图7-6所示，关键评论提取如图7-7所示。

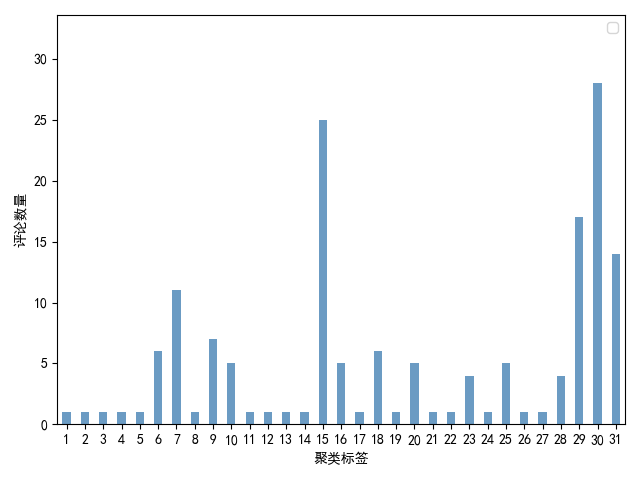


图7-6 评论聚类结果图

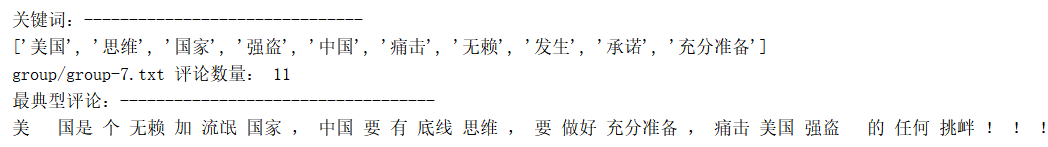


图7-7 关键评论提取结果图

总结

本文主要是对网络新闻文本包括新闻正文及其评论的情感分析的设计与实现。相较于其他研究主要有以下几点不同与优势：

1. 由于本文研究对象为新闻正文与评论文本，两者篇幅差距较大，故在情感分析方法上面，本文选择了两种方法——依存句法分析与基于LSTM的深度学习模型，从分析速率与准确度考虑，最终选择了在新闻正文的情感分析研究上使用依存句法分析，新闻评论的情感分析研究上使用深度学习模型。
2. 在深度学习模型的研究中，与传统的二分类模型差别的是，本文采用的是三分类模型，三分类除了涉及到二分类中的正面和负面情感外，还有一种中性情感，从原始数据中，提取具有语义转换的句子，从而获得3个具有不同情感的数据。
3. 在新闻正文中文分词中，本文选取了更具领域性的北京大学开源分词工具pkuseg，pkuseg提供了在MSRA（新闻语料）上训练得到的模型，在新闻文本分词中更为准确。但考虑到其分词耗时长在大数量的评论分词中本文选择的是较稳定、快速的jieba分词工具。
4. 本系统的模块符合高内聚、低耦合的原则。

基于以上分析和算法研究，结合python网络爬虫技术和Java web技术，设计并完成了一个基于新闻的大数据分析系统，并对系统进行功能测试，测试结果表明，系统符合预期的要求。

但是本系统还是存在不足，比如本系统的评论情感分类准确度不高，主要原因是，网络新闻的标记语料的缺乏；中性评价出现的概率不高，主要原因是中性数据集的数量要比其他两个数据集少一半多，并且通过简单规则“然而”，“但”提取出来的中性数据集质量也不是很高，所以才会出现偏差。

致谢

毕业设计是四年大学生活的结束标志，经过五个多月的努力，此次毕业设计终于完成了，这也标志着我美好而充实的大学生活即将结束。四年的大学生活中，老师、同学与身边的朋友们给予了我许多真诚的帮助与关怀。在此，我表示万分感谢。

这次毕业设计的完成，首先要感谢魏冬梅老师的精心指导。疫情居家期间，魏老师仍坚持每周的指导，在每周的指导时，魏老师督促与检测我们毕业设计的完成度，并对我们已完成部分提出意见与建议，让我及时了解到自己毕业设计中的不足与下一步目标。在本次毕业设计中，如果没有魏老师的悉心指导，我无法轻松地完成程序的编写与相关文档的准备。

除了感谢魏老师外，我还要感谢那些帮助与关怀我的同学、朋友们。我要感谢陈晓玮同学在相关文档的编写过程中的提醒与建议；还有我的朋友周伍，在神经网络学习方面，他给我提供了许多学习资料与指导，得益于他的帮助，我才能在几周之内完成神经网络的构建。感谢我的大学同学以及舍友们，我很荣幸能跟大家一同度过难忘而又可贵的四年大学生活。

参考文献

[1] 中国互联网络信息中心（CNNIC）.第45次中国互联网络发展现状统计报告[R].2020年4月28日.

[2] 陈敬,刘欢欢,李寿山,周国栋.基于主动学习的新闻读者情绪分类方法 [J]. 山西大学学报(自然科学版), 2017,40(155), 01 26-31.

[3] 赵妍妍.句子级情感信息抽取与极性识别的方法研究[D].哈尔滨工业大学, 2011.

[4] 唐立婷.关于网络舆情生态系统的构建与运行机理研究[J].现代经济信息,2019,15 411+413.

[5] 李雪.面向网络新闻的舆情检测与分析系统设计与实现[D].山东师范大学,2014.(4)

[6] 段传明.传统情感分类方法与基于深度学习的情感分类方法对比分析[J].软件导刊,2018,17(01):22-24.

[7] 韦修喜,黄华娟,周永权. 基于AP聚类的约简孪生支持向量机快速分类算法 [J]. 计算机工程与科学, 2019,41(298), 10 189-194.

[8] Frey B J, Duech D.Clustering by passing messages between data points[J].Science, 2007, 315 (5814):972-976.

[9] 陶杨明. 基于深度学习的细粒度文本情感分析研究[D].浙江工商大学,2020.

[10] 张原溥.基于LSTM长短期记忆网络的情感分析[J].电子世界,2018(17):16-17+20.

[11] 张英. 基于深度神经网络的微博短文本情感分析研究[D].中原工学院,2017.

[12] 韦修喜,黄华娟,周永权. 基于AP聚类的约简孪生支持向量机快速分类算法 [J]. 计算机工程与科学, 2019,41(298), 10 189-194.

[13] 季蕾娜. 基于用户权威度的中文微博话题检测研究[D].昆明理工大学, 2014.