**网上舆情爬取系统**

# 目 录

[1 概要设计 1](#_1_概要设计)

[1.1 系统中的实体 1](#_1.1_系统中的实体)

[1.1.1 帖子实体 1](#_Toc224369792)

[1.1.2 敏感词实体 1](#_Toc224369793)

[1.2 系统主要流程描述 2](#_1.2_系统主要流程描述)

[1.2.1 系统主要处理 2](#_Toc224369801)

[1.2.2 系统流程图 3](#_Toc224369802)

[1.3 系统数据字典 6](#_1.3_系统数据字典)

[1.3.1 涉及主要数据 6](#_Toc224369804)

[2 数据库设计 9](#_数据库设计)

[2.1 帖子表 9](#_2.1_帖子实体)

[2.2 敏感词表 10](#_2.2_隐患信息)

[3 系统架构 17](#_Toc224369839)

[3.1系统体系结构 17](#_Toc224369840)

[3.2 系统架构 18](#_Toc224369841)

[3.2.1 总体架构描述 18](#_Toc224369842)

[3.2.2 软件总体架构 19](#_Toc224369843)

[4 详细设计 21](#_详细设计)

[4.1 帖子主体 21](#_Toc224369845)

[4.2 敏感词客体 22](#_Toc224369852)

[5 软件实施 32](#_Toc224369906)

[5.1 运行设备 32](#_Toc224369907)

[5.1.1 必备设备 32](#_Toc224369908)

[5.1.2 其它设备 32](#_Toc224369909)

[5.2 软件需求 32](#_Toc224369910)

[5.2.1 软件支持 32](#_Toc224369911)

[5.2.2 开发说明 33](#_Toc224369912)

# 1 概要设计

## 1.1 系统中的实体

### 1.1.1 帖子实体

帖子实体包括的属性有：发帖人、发帖时间、帖子标题、帖子内容、阅读和回复数量以及帖子链接。本系统的帖子实体属性如图1.1所示。

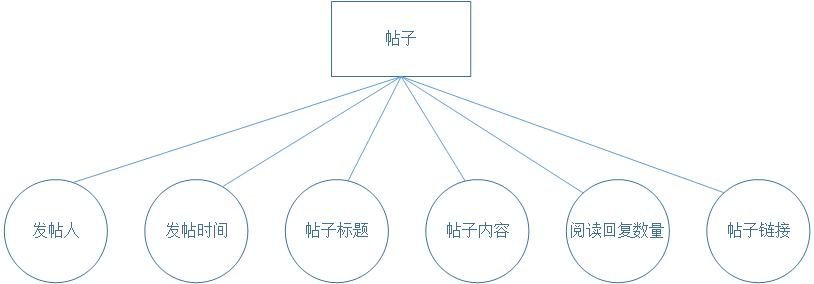


图1.1“帖子”属性描述图

### 1.1.2 敏感词实体

敏感词实体包括的属性有：敏感词名称。本系统的帖子敏感词属性如图1.2所示。

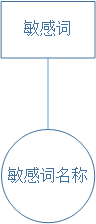


图1.2“敏感词”属性描述图

## 1.2 系统主要流程描述

### 1.2.1 系统主要处理

1. 系统的搜索流程。用户通过搜索功能对自己感兴趣的帖子进行搜索，本搜索功能支持搜索帖子标题和搜索发帖人。
2. 系统的敏感词管理流程。用户根据自己的兴趣对敏感词进行添加和删除，使得系统更加人性化。
3. 系统的设置URL流程。本系统提供了设置URL的功能，用户输入自己感兴趣的URL，点击设置，重启一下爬取系统，即可以爬取用户指定的URL。(URL在该域名之下：http://www.cxjhm.com/)
4. 系统的爬取流程。该系统后台从结合美论坛(<http://www.cxjhm.com/forum.php>)爬取帖子，并帖子进行相关的管理，在前台进行展示给用户。

### 1.2.2系统流程图

1. 系统的搜索流程。如图1.3所示。

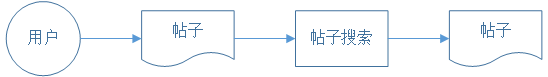


图1.3网上舆情爬取系统主要业务（一）

2. 系统的敏感词流程。如图1.4所示。

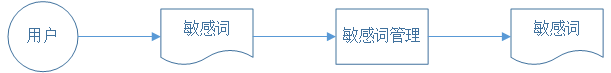


图1.4网上舆情爬取系统主要业务（二）

3. 系统的设置URL流程。如图1.5所示。

设置URL

图1.5网上舆情爬取系统主要业务（三）

4. 系统的爬取流程。如图1.6所示。

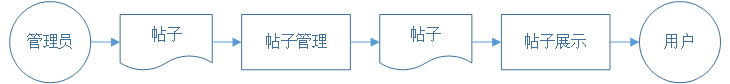


图1.6网上舆情爬取系统主要业务（四）

## 1.3 系统数据字典

### 1.3.1 涉及主要数据

通过分析可以确定系统所要处理的数据情况。

1. 帖子实体
2. 发帖人
3. 发帖时间
4. 帖子标题
5. 帖子内容
6. 阅读和回复数量
7. 帖子链接
8. 敏感词实体
9. 敏感词名称

# 数据库设计

## 2.1 帖子实体

jihemei

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 列名 | 数据类型 | 类型长度 | 主键 | 允许空 | 默认值 | 说明 |
| id | int | 11 | 是 | NO | Null | 编号 |
| postman | varchar | 20 |  | NO | Null | 发帖人 |
| firstTime | date |  |  | NO | Null | 发帖时间 |
| Title | text |  |  | NO | Null | 标题 |
| Content | text |  |  | Yes | Null | 内容 |
| readCount | int | 11 |  | NO | 0 | 阅读次数 |
| replyCount | int | 11 |  | NO | Null | 回复次数 |
| Link | text |  |  | NO | Null | 发帖链接 |

## 2.2 敏感词实体

hot\_words

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 列名 | 数据类型 | 类型长度 | 主键 | 允许空 | 默认值 | 说明 |
| id | int | 11 | 是 | NO | Null | 编号 |
| word | varchar | 20 |  | NO | Null | 敏感词名称 |

# 系统架构

## 3.1系统体系结构

该系统架构采用了B/S模式，网络程序开发模式有两种：C/S以及B/S，也叫做客户机/服务器以及浏览器/服务器模式，这两种模式用于不同的场合，各有千秋。本系统采用的是B/S模式。

B/S模式是一个具有三层结构的技术体系：第一层中，客户机扮演者整个系统和用户间接口的角色。浏览器将使HTML代码转化为具有交互功能的Web页面，能允许用户在Web页面的申请表里输入信息发送至后台；在第二层中，Web服务器会启动相关程序处理用户请求，生成HTML代码。第三层中，MySQL负责协调WEB所发的SQL请求来管理数据库。B/S模式结构如图3.1所示：

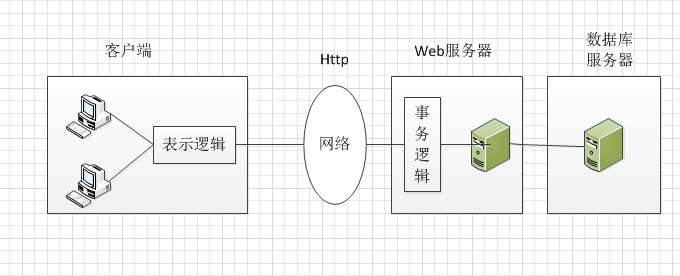


图3.1 B/S模式结构图

## 3.2 系统架构

### 3.2.1 系统后台

该系统的后台使用了scrapy框架。Scrapy是一个为了爬取网站数据，提取结构性数据而编写的应用框架。 可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。  
 所谓网络爬虫，就是一个在网上到处或定向抓取数据的程序，当然，这种说法不够专业，更专业的描述就是，抓取特定网站网页的HTML数据。抓取网页的一般方法是，定义一个入口页面，然后一般一个页面会有其他页面的URL，于是从当前页面获取到这些URL加入到爬虫的抓取队列中，然后进入到新页面后再递归的进行上述的操作，其实说来就跟深度遍历或广度遍历一样。  
Scrapy 使用 Twisted这个异步网络库来处理网络通讯，架构清晰，并且包含了各种中间件接口，可以灵活的完成各种需求。scrapy框架结构图3.2如下。



图3.2 scrapy框架结构图

### 3.2.2 系统前台

该系统的前台采用了Flask框架。Flask 是一个基于Python的微型的web开发框架。虽然Flask是微框架，不过我们并不需要像别的微框架建议的那样把所有代码都写到单文件中。毕竟微框架真正的含义是简单和短小。Flask 依赖两个外部库： Jinja2 模板引擎和 Werkzeug WSGI 工具集。Flask遵循“约定优于配置”以及合理的默认值原则。Flask框架的结构图如图3.3所示。

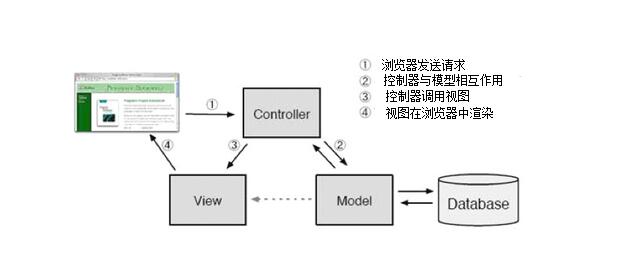


图3.3 Flask框架结构图

# 详细设计

## 4.1 帖子主体

从结合美论坛爬取的帖子数量巨大。这些帖子可以根据相关要求进行删除，但是不可以修改，这样违背了系统帖子的真实性。帖子主体包括以下几个方面。

帖子编号

帖子标题

发帖人

阅读和回复数量

帖子链接

帖子内容

其中帖子链接是唯一的，爬取帖子时可以通过链接的唯一性来进行去重。

## 4.2 敏感词客体

敏感词有两个来源，第一个是系统提供的一些敏感词，第二个是用户自行添加的。用户可以根据自己的喜好对敏感词进行添加和删除。敏感词包括以下两个方面。

敏感词编号

敏感词名称

其中敏感词编号是唯一的，敏感词的名称允许重复。

# 软件实施

## 5.1 运行设备

### 5.1.1 必备设备

* 数据库服务器，为系统提供数据支持，需要一台性能较高的商用计算机。
* Web服务器，为Web服务提供硬件支持，需要一台性能较强的商用计算机。
* 若干台工作用的计算机。

## 5.2 软件需求

### 5.2.1 软件支持

* 数据库服务器。

1. MySQL5.6.17

* Web服务器。

1. Windows 10
2. Python2.7
3. SQLAlchemy
4. Flask

* 普通客户端计算机。

1. Windows XP及以上
2. IE 6.0以上版本

此外，每台计算机上配备必要的杀毒软件及防火墙。

### 5.2.2 开发说明

系统开发所需要的软件环境如下：

操作系统：windows 10

数据库：MySQL

开发语言：Python

开发环境：PyCharm

Python2.7

Flask

SQLAlchemy

## 5.3 软件实施进度

### 5.3.1 功能模块划分

* 帖子主体

1. 帖子编号
2. 帖子标题
3. 发帖人
4. 发帖时间
5. 帖子内容
6. 阅读和回复数量
7. 帖子链接

* 敏感词客体

1. 敏感词编号
2. 敏感词名称

### 5.3.2 开发进度表

软件编码实现进度表如下图所示：

横向座标表示日期，

纵坐标表示5.3.1中对应编号模块的进展情况。

**项目开发进度表**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 日期  任务 | 3月 | | | | 4月 | | | | | | | | | |
| 25 | 27 | 29 | 31 | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 | 14 | 16 |
| 1-7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8-9 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |