目录

[一、文档概述 2](#_Toc1523)

[1.1、编写目的 2](#_Toc15594)

[1.2背景 2](#_Toc32356)

[1.3定义 3](#_Toc26500)

[1.4、参考资料 3](#_Toc16744)

[二、任务概述 4](#_Toc23073)

[2.1业务需求 4](#_Toc10807)

[2.2相关人员及用户分析 4](#_Toc735)

[2.3 用户的特点 4](#_Toc5915)

[2.4假定和约束 5](#_Toc21757)

[2.5运行环境 5](#_Toc27653)

[2.5.1设备 5](#_Toc13198)

[2.5.2支持软件 5](#_Toc31491)

[2.5.3接口 5](#_Toc16038)

[2.5.4控制 6](#_Toc2343)

[三、 系统概述 6](#_Toc32536)

[3.1需求规定 6](#_Toc7130)

[3.1.1需求概述 6](#_Toc3328)

[3.1.2对功能的规定 6](#_Toc22974)

[3.2登录 8](#_Toc28580)

[3.2.1 业务事件 8](#_Toc25470)

[3.3 数据爬取 9](#_Toc10622)

[3.3.1业务事件 9](#_Toc27711)

[3.4数据可视化 10](#_Toc14697)

[3.4.1业务事件 10](#_Toc22168)

[3.5数据库设计 11](#_Toc25147)

[3.5.1数据库概念结构设计 11](#_Toc560)

[3.5.2数据库物理结构设计 11](#_Toc21605)

[四、 非功能性需求 12](#_Toc6011)

[4.1安全性要求 12](#_Toc28999)

[4.1.1访问安全性要求 12](#_Toc18988)

[4.1.2数据安全性要求 13](#_Toc24146)

[4.2可靠性要求 13](#_Toc4175)

[4.2.1容错性要求 13](#_Toc3646)

[4.2.2 可恢复性要求 13](#_Toc31533)

[4.2.3其他可靠性要求 13](#_Toc22742)

[4.3易用性要求 14](#_Toc2158)

[4.3.1界面友好性要求 14](#_Toc15904)

[4.3.2易操作性要求 14](#_Toc13579)

[4.3.3其他易用性要求 14](#_Toc6138)

[4.4精度 14](#_Toc10510)

[4.5时间特性要求 15](#_Toc21852)

[4.6灵活性 15](#_Toc19007)

[4.7输入输出要求 16](#_Toc29917)

[4.8数据管理能力要求（针对软件系统） 16](#_Toc13601)

[4.9 故障处理要求 16](#_Toc24622)

[4.10性能要求 16](#_Toc22253)

[4.10.1 数据访问性能要求 16](#_Toc6314)

[4.10.2 数据传输性能要求 17](#_Toc30841)

[4.11可维护性要求 17](#_Toc25899)

[4.11.1 公共数据要求 17](#_Toc18722)

[4.11.2 公共语言开发要求 17](#_Toc29639)

[4.12 可移植性要求 17](#_Toc12848)

[4.12.1 适应性要求 17](#_Toc13851)

[4.12.2 易安装性要求 17](#_Toc15962)

[4.13完整性 18](#_Toc13647)

[4.14可理解性 18](#_Toc32327)

# 一、文档概述

## 1.1、编写目的

追星女孩同学是个流量明星粉，特别喜欢邓超和anglebaby。她想爬取2人的微博，以txt文件或者csv文件形式保存。还要分析微博中出现的那些人，去了哪些地方，每句微博的情感，形成词云，柱状图等献给邓超和baby作为礼物。

## 1.2背景

1. 待开发软件系统的名称：微博可视化分析系统
2. 项目任务提出者：吴迪，追星女孩

3）任务开发者：魏振东

4)软件用户：追星女孩

## 1.3定义

微博可视化分析系统： 完成微博登录，内容爬取，存储，可视化分析等功能,用python语言实现。

UML ：统一建模语言（ UML 是 Unified Modeling Language 的缩写）是用来对软件密集系统进行可视化建模的一种语言。 UML 为面向对象开发系统的产品进行说明、可视化、和编制文档的一种标准语言。

B/S 结构： Browser/Server 结构 ,即浏览器和服务器结构。它是对 C/S 结构的一种变化或者改进的结构。 在这种结构下 ,用户工作界面是通过 WWW 浏览器来实现 ,极少部分事务逻辑在前端 (Browser) 实现 ,主要事务逻辑在服务器端 (Server)实现 ,server 端访问数据库 ,形成所谓的三层架构。

## 1.4、参考资料

[1] 张海藩 .软件工程导论（第五版） [M]. 北京：清华大学出版社， 2008

[2] 刘乃丽 . 精通 ASP.NET2.0+SQLServer 2005 项目开发 .[ M]. 北京：人民邮电出版社 , 2007 ：100～150.

[3] ( 美) 沃尔特，本 - 甘，萨卡 . Microsoft SQL Server 2005技术内幕 -T-SQL 程序设计[ M]. 北京：电子工业出版社， 2007:50 ～80.

# 二、任务概述

## 2.1业务需求

追星女孩同学是个流量明星粉，特别喜欢邓超和anglebaby。她想爬取2人的微博，以txt文件或者csv文件形式保存。还要分析微博中出现的那些人，去了哪些地方，每句微博的情感，形成词云，柱状图等。

## 2.2相关人员及用户分析

1. 邓超，1979年出生于江西南昌，中国内地男演员、电影导演、投资出品人。1995年考入江西艺术职业学院话剧班，1998年考入中央戏剧学院表演系，毕业后编入中国国家话剧院。2001至02年主演喜剧《欢乐青春》、《当爱情失去记忆》等，之后在《少年天子》中饰演顺治皇帝成名，并相继在《少年康熙》《明末风云》《天下第一》中饰演皇帝被称为“皇帝专业户”。05年后主演《幸福像花儿一样》及《甜蜜蜜》饰演高干子弟。
2. 追星女孩，追星女孩

## 2.3 用户的特点

因为该系统的最终用户是追星女孩，要求使用本系统的用户对计算机能够熟练的使用，知道基本的操作。

## 2.4假定和约束

追星女孩可以据实修改停用词和替换词。

## 2.5运行环境

### 2.5.1设备

操作系统：Windows 7及以上

CPU：Intel i7酷睿双核及以上

内存：16G及以上

硬盘空间：100G以上

软件配置如下：

开发工具：MyEclipse，pycharm

数据库：Mysql等

Web服务器：Tomcat

### 2.5.2支持软件

先进可靠安全性高，可扩展性且性价比高，支持python规范。

操作系统： Microsoft Windows 2000 Advanced Server 支持环境： IIS 5.0 ;

数据库系统：mysql

### 2.5.3接口

硬件接口：本软件不需要特定的硬件或硬件接口进行支撑， 486 以上 PC 机均可运行此软件，对机器要求不太严格． 因为可能涉及数据的备份应该保持打印机和光盘刻录机的接口。

软件接口： 运行于 Windows95 及更高版本具有 WIN32 API 的操作系统之上。 因为可能

涉及一些文档、报表的处理应该保持与常用软件的办公软件的接口。

### 2.5.4控制

本系统初步决定采用 B/S 架构，用户通过本地， 在权限范围内对其所属信息和附件可增删改。

# 系统概述

## 3.1需求规定

### 3.1.1需求概述

追星女孩同学是个流量明星粉，特别喜欢邓超和anglebaby。她想爬取2人的微博，以txt文件或者csv文件形式保存。还要分析微博中出现的那些人，去了哪些地方，每句微博的情感，形成词云，柱状图等。

### 3.1.2对功能的规定

根据微博可视化分析系统的实际需求，可以将微博可视化分析系统划分为登录，数据爬取，数据存储，数据清洗，数据分析，数据可视化5个部分各个部分的具体功能的系统功能结构图如图 3-1 所示

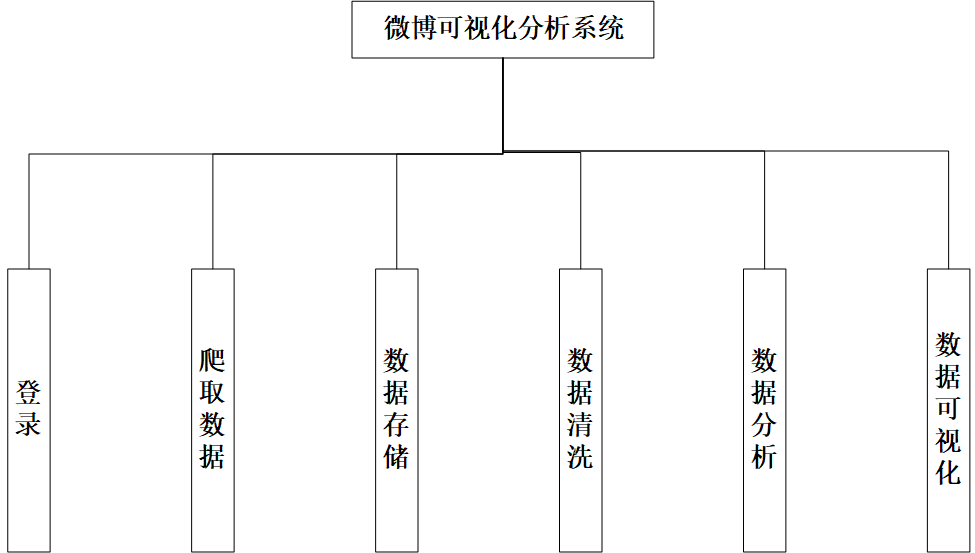


图3-1功能模块图

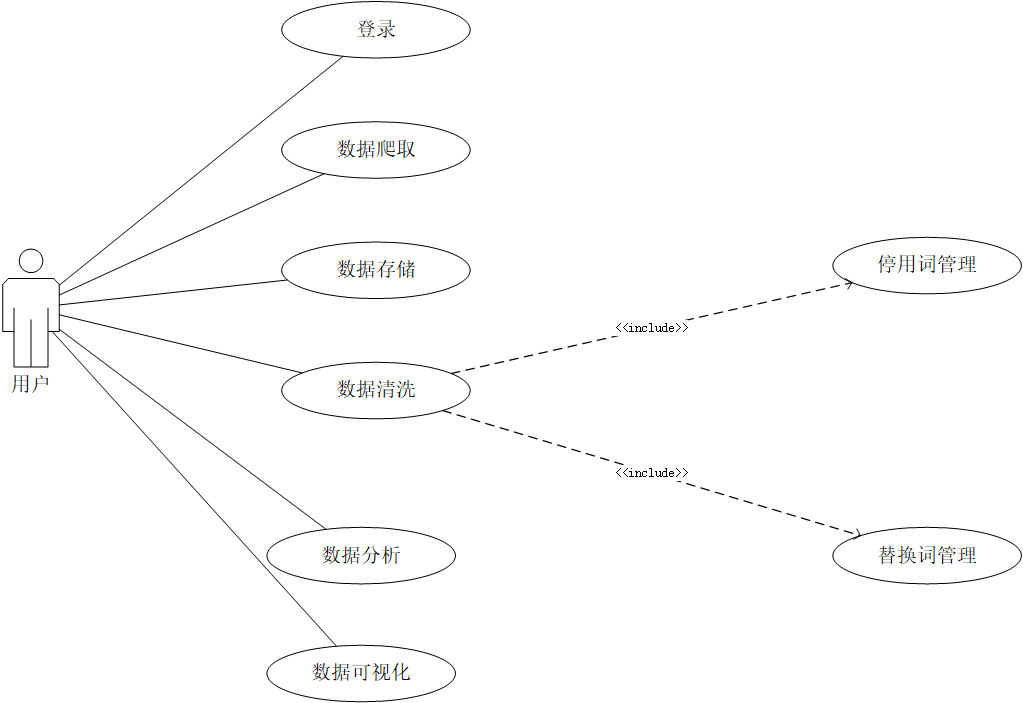


图3-2用例图

## 3.2登录

打开浏览器：使用selenium模拟打开浏览器

获取用户名和密码框：使用browser.find\_element\_by\_css\_selector获取用户名和密码框

睡眠： 模拟用户停留。

输入账号和密码：send\_keys(账号)

### 3.2.1 业务事件

#### 1、用户登录系统

业务流程分析：

用户模拟打开浏览器，然后找到输入框和密码框，输入密码和账号，然后点击登录。其流程图如图3-3

用例分析 ：

在这个流程中，只涉及到用户与系统进行交互，其用例如图3-2

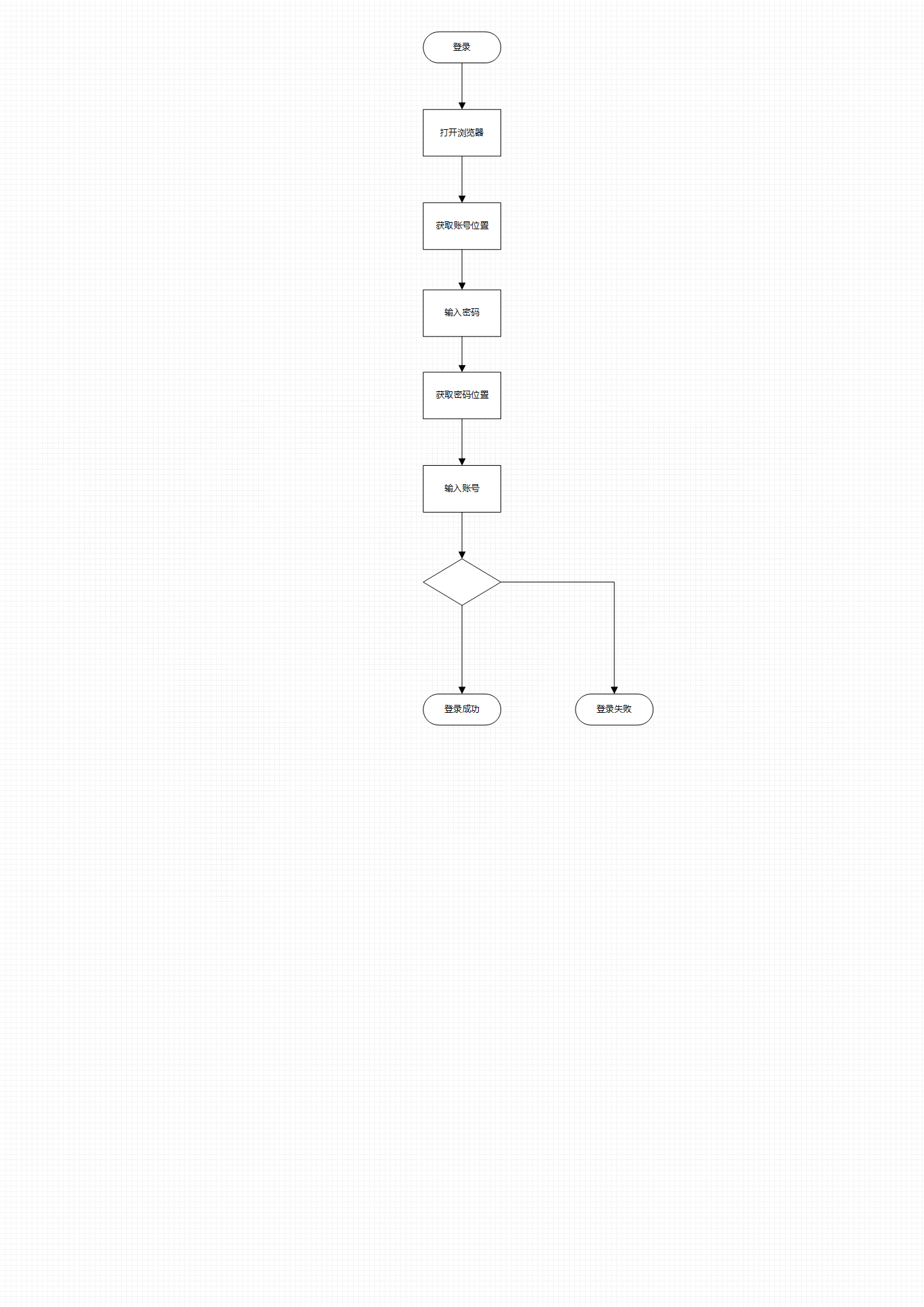


图3-3用户登录系统及操作活动图

## 3.3 数据爬取

### 3.3.1业务事件

#### 1.数据爬取系统

业务流程分析：

用户爬取数据，然后进行存储。其流程图如图3-3

用例分析 ：

在这个流程中，只涉及到用户与系统进行交互，其用例如图3-2

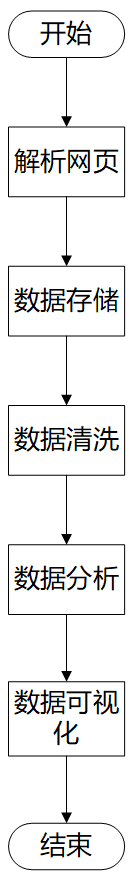


图3-4流程图

## 3.4数据可视化

### 3.4.1业务事件

#### 1.读者登录验证

业务流程分析：

用户对存储之后的数据进行清洗，然后分析可视化。其流程图如图3-3

用例分析 ：

在这个流程中，只涉及到用户与系统进行交互，其用例如图3-2

## 3.5数据库设计

### 3.5.1数据库概念结构设计

1. 文章管理，基本的增删改查，并对文章进行可视化处理，生成关系图， 饼图，柱状图，词云，趋势图等可视化图形，方便用户分析。
2. 停用词管理，对相应文章的停用词进行管理，可以进行增删改查。
3. 替换词管理，对相应文章的替换词进行管理，可以进行增删改查。
4. 文件上传，根据上传的txt文件和停用词表和替换词表，进行可视化 分析，生成关系图，饼图，柱状图，词云，趋势图等可视化图形，方便 用户分析

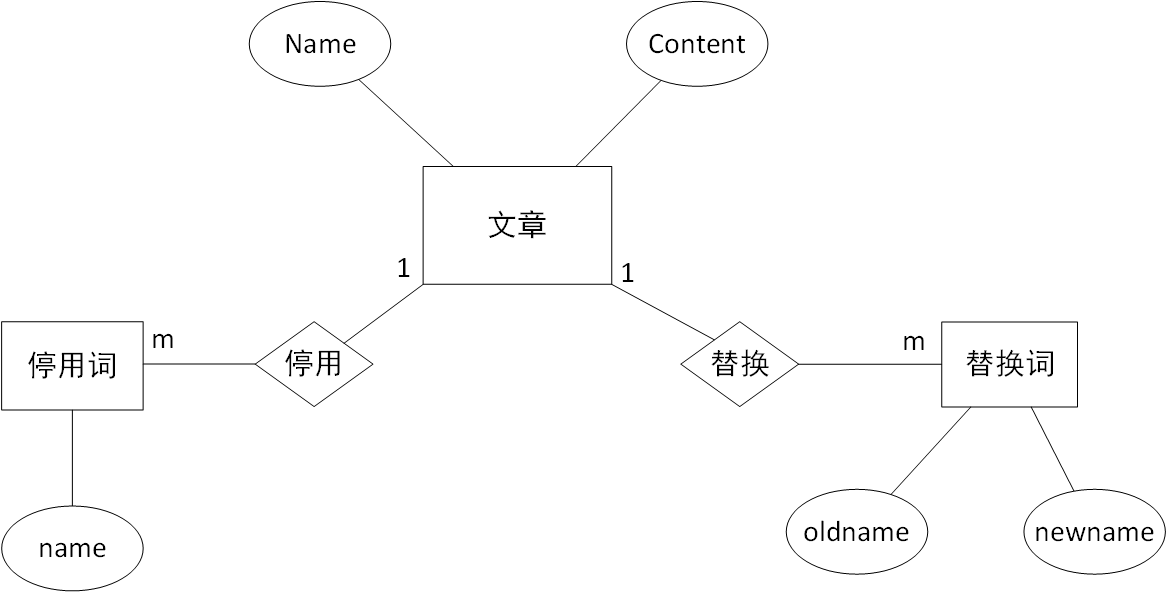


图3-16 部分实体图

### 3.5.2数据库物理结构设计

表4.1 文章信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 允许空 | 备注 | 是否为主键 |
| id | int | NOT NULL | id | 是 |
| Name | varchar | NOT NULL | 文章名称 | 否 |
| Content | longtext | NOT NULL | 内容 | 否 |

表4.2 停用词表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 允许空 | 备注 | 是否为主键 |
| id | int | NOT NULL | id | 是 |
| name | varchar | NOT NULL | 停用词 | 否 |
| a\_id\_id | int | NOT NULL | 与文章表关联 | 否 |

表4.2 替换词表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 允许空 | 备注 | 是否为主键 |
| id | int | NOT NULL | id | 是 |
| oldname | varchar | NOT NULL | 旧词 | 否 |
| newname | varchar | NOT NULL | 新词 | 否 |
| a\_id\_id | int | NOT NULL | 与文章表关联 | 否 |

# 非功能性需求

## 4.1安全性要求

### 4.1.1访问安全性要求

该微博可视化系统， 其中为登录系统的读者只可以搜索和查 看微博信息， 只有在成功登录系统后才能查看可视化图形

### 4.1.2数据安全性要求

用户成功登录后可以对自己所管辖的信息进行更改， 其他人一概没有权利进行任何更改操作。 系统内部数据在定期更新时都要求有备份

## 4.2可靠性要求

### 4.2.1容错性要求

整体系统运行稳定， 有很强的防错、 抗错能力，保证数据报送正常进行。在系统出现错 误或者异常时，可以及时的保存数据，确保重要相关数据、相关信息不会丢失。

### 4.2.2 可恢复性要求

在进行数据信息录入或更新时， 系统会间隔固定时间自动保存， 在系统出现异常时， 数 据可自动回复发生异常前的数据。

### 4.2.3其他可靠性要求

操作可靠性：用户访问网站时都能正常操作。

数据可靠性：数据信息是用户定期更新的，具有实时、准确和可靠性

## 4.3易用性要求

### 4.3.1界面友好性要求

该微博可视化系统设计的界面友好，用户操作简单容易，在操作的页面上均有操作提示，而其页面显示都是采用最便于用户使用的控件和布局方式

### 4.3.2易操作性要求

无论是对于用户微博可视化系统的操作都是简单便捷的， 即有较高可操作和易操作性，在响应时间上又较短，所以可以较大的提高操作的效率

### 4.3.3其他易用性要求

在系统中有需要时间信息的地方， 均给出了日历， 用户只需选择日期即可，不需自己再去添加

## 4.4精度

查询时应保证查全率，所有相应域包含查询关键字的记录都应能查到；查询时应保证查准率， 查到的记录应与给定的单项或组合查询条件不完全匹配的模糊查询；录入数据合法性的检验应当精确；密码允许输入 6-8 个字母或者数字：用户输入查询信息应不区分大小写。

## 4.5时间特性要求

由于此开发项目针对追星女孩， 使用频度较高， 使用性要求比较高。 为防止对信息资料和管理程序的恶意破坏，要求有较为可靠的安全性能。总之，要求稳定、安全、便捷，易于管理和操作。

1.查询速度：不超过 10 秒；

2.其它所有交互功能反应速度：不超过 3 秒；

3.可靠性：平均故障间隔时间不低于 200 小时。

4.响应时间：应在 1～2 秒内，对软磁盘和打印机的操作，以及数据的导入和导出也应在可接受的时间内完成；

## 4.6灵活性

作为独立运行的系统和其他管理系统集成的系统。本系统具有独立的服务器系统和数据库系统， 具有完善数据输入输出功能和数据维护及查询的报表生成与打印系统。 且发生故障时，能快速恢复系统和故障处理，方便系统升级和扩充，故障恢复时间不超过 5 小时。为了适应内外机构的数据要求， 本系统去采用标准界面，本身具有操作灵活的特点。可能提供鼠标选择和键盘输入双重输入功能。方便用户操作和管理。

## 4.7输入输出要求

(1)用户：用户可以添加或删除停用词，替换词，并使用可视化输出。

## 4.8数据管理能力要求（针对软件系统）

用户基本信息中的 id 号、姓名；各项信息基本上是固定不变的， 他们不会随着系统的运行和时间的变化而发生变化。他们一旦入库，一般情况下是不会改变的，所以，他们输入库的过程一定要确保正确。其他的，如图书用户基本信息中的密码， 微博id表都随时需要做一些调整，因此他们会改变。用户的权限一旦入库后基本上是不会改变的。

## 4.9 故障处理要求

正常使用时不应出错，若运行时遇到不可恢复的系统错误，也必须保证数据库完好无损。 要在项目报名时的没隔一段时间进行数据备份，以免在资料意外丢失时，无法进行恢 复。 对系统故障的处理要求区分故障的严重程度， 尽可能的对错误进行恢复。 随时监控， 在 文档、报表处理，打印机，操作系统等软硬件出现故障时、具备保 数据的功能，并及时反 映到主机中。

## 4.10性能要求

### 4.10.1 数据访问性能要求

该微博可视化系统利用数据缓存， 既保证了数据库中原始数据的可靠性， 又能够加强数据之间的交互效率。

### 4.10.2 数据传输性能要求

该微博可视化系统数据在上传时会经过部分压缩， 以加强数据的统一保存和处理， 还能节省数据所占用的空间，给数据库减小压力。

## 4.11可维护性要求

### 4.11.1 公共数据要求

在数据更新时， 不同的管理员在更新自己输入的信息时， 需要先同步其他管理员已经录入好的信息， 没有冲突才能将自己的录入。 录入的数据全部按照一定顺序进行排列储存所以维护比较容易。

### 4.11.2 公共语言开发要求

采用Python语言，该语言有较好的可维护性。

## 4.12 可移植性要求

### 4.12.1 适应性要求

系统灵活性很强，可以随时进行内容修改和界面的更新。另外也适应多种数据传输方式，能够提供灵活的配置以适应业务需求。

### 4.12.2 易安装性要求

该系统安装简单，只需将可执行程序在具备预期的使用环境所示环境的主机上运行即可。在主服务器上安装成功后客户端只需通过互联网便可登录该系统的网站，进行相关操作。

## 4.13完整性

要求在发生意外时（如断电），保证数据不丢失。

## 4.14可理解性

对于本软件提供的各种菜单、按钮，其功能应该一目了然， 易于理解