

**软件工程综合课程设计**

--概要设计

**题 目** 中标统计分析系统

**院 系**  计算机科学与技术学院

**专 业**  软件工程

**学生姓名** 吴未名,王祥,曹立群,张桔,柳鹏磊

**学 号**  161330119

**指导教师** 张德平

**二0 一五 年 十二 月 二十五 日**

**目录**

[1 引言 4](#_Toc438710084)

[1.1 编写目的 4](#_Toc438710085)

[1.2 背景 4](#_Toc438710086)

[1.3 定义 4](#_Toc438710087)

[1.4 参考资料 4](#_Toc438710088)

[2 总体设计 4](#_Toc438710089)

[2.1 简述 4](#_Toc438710090)

[2.2 架构设计 5](#_Toc438710091)

[2.2.1 系统逻辑架构图 5](#_Toc438710092)

[2.2.2 系统物理架构图 5](#_Toc438710093)

[2.2.3 顶层系统包图 6](#_Toc438710094)

[2.2.4 业务类包图 6](#_Toc438710095)

[3 概要设计 7](#_Toc438710096)

[3.1 类图 7](#_Toc438710097)

[3.1.1 爬虫数据 7](#_Toc438710098)

[3.1.2 爬虫管理 8](#_Toc438710099)

[3.1.3 总体结构 8](#_Toc438710100)

[3.2 类说明 9](#_Toc438710101)

[3.2.1 爬虫数据处理 9](#_Toc438710102)

[3.2.2 状态数据处理 9](#_Toc438710103)

[3.2.3 中标数据处理 10](#_Toc438710104)

[3.2.4 爬虫管理 10](#_Toc438710105)

[3.2.5 爬取规则 11](#_Toc438710106)

[3.2.6 爬虫设置 11](#_Toc438710107)

[3.2.7 视图 11](#_Toc438710108)

[3.2.8 数据库 12](#_Toc438710109)

[3.2.9 安全检测 12](#_Toc438710110)

[3.3 界面设计 13](#_Toc438710111)

[3.3.1 网站首页 13](#_Toc438710112)

[3.3.2 登陆界面 13](#_Toc438710113)

[3.3.3 注册界面 14](#_Toc438710114)

[3.3.4 后台首页 14](#_Toc438710115)

[3.3.5 用户信息界面 14](#_Toc438710116)

[3.3.6 爬虫管理界面 15](#_Toc438710117)

[3.3.7 爬虫状态界面 15](#_Toc438710118)

[3.3.8 数据查询界面 16](#_Toc438710119)

[3.3.9 中标详情界面 16](#_Toc438710120)

[3.3.10 企业中标情况统计界面 17](#_Toc438710121)

[3.3.11 时间中标情况统计界面 18](#_Toc438710122)

[4 系统数据结构设计 18](#_Toc438710123)

[4.1 逻辑结构设计 18](#_Toc438710124)

[4.1.1 总数据表（SE\_RAWDATA） 18](#_Toc438710125)

[4.1.2 爬取规则表（RULES） 18](#_Toc438710126)

[4.1.3 运行状态表（STATUS） 19](#_Toc438710127)

[4.1.4 用户表（USER） 19](#_Toc438710128)

[5 系统出错处理设计 20](#_Toc438710129)

[5.1 出错信息 20](#_Toc438710130)

[5.2 补救措施 20](#_Toc438710131)

[5.3 系统维护设计 20](#_Toc438710132)

**中标统计分析系统概要设计说明书**

# 引言

## 编写目的

此概要设计说明书是为了说明整个系统的体系架构，以及需求用例的各个功能点在架构中的体现，为系统的详细设计人员进行详细设计时的输入参考文档。

## 背景

软件系统的名称：中标统计分析系统

开发工具：Python、PHP、Scrapy

开发者：吴未名、王祥、曹立群、张桔、柳鹏磊

## 参考资料

1. 吴小竹等，基于Python的多线程SPIDER的设计与实现
2. 胡宏涛等，基于网络的信息获取技术浅析
3. 李学勇等，网络蜘蛛搜索策略比较研究
4. 高克宁等，支持Web信息分类的高性能蜘蛛程序
5. 吴强等，Python线程，Reactor平台程序设计和实战
6. 荣传湘等，搜索引擎中数据获取的设计与实现
7. 叶允明等，分布式Web Crawler 的研究: 结构、算法和策略

# 总体设计

## 简述

系统采用流行的B/S结构模式。系统的分析设计采用面向对象的技术，应用Visio等工具进行辅助设计。

## 架构设计

### 系统逻辑架构图



### 系统物理架构图



### 顶层系统包图



### 业务类包图



# 概要设计

## 类图

### 爬虫数据



### 爬虫管理

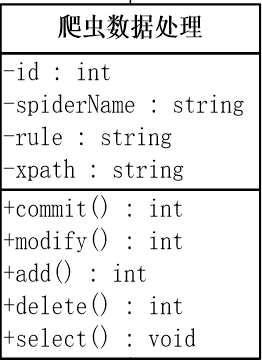


### 总体结构



## 类说明

### 爬虫数据处理



ID：标识符

spiderName：爬虫名称，是爬虫唯一的标识符

rule：爬取规则，是一个字符串，用逗号分割

xpath：是一组规则的表示形式

commit()：提交一个请求

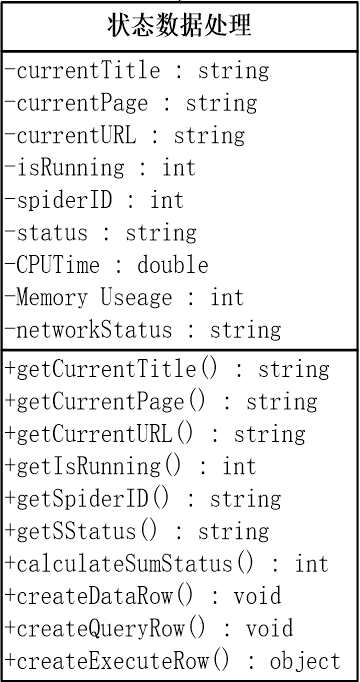
modify()：修改数据表内容

add()：添加一条数据

delete()：删除一条或者若干条数据

select()：查询一条或若干条记录

### 状态数据处理



currentTitle：当前爬取的标题

currentPage：当前爬取的页码

currentURL：当前爬取的网址

isRunning：是否正在运行

spiderID：爬虫的标识符ID

status：状态对象

CPUTime：CPU占用时间

MemoryUseage：内存占用

networkStatus：网络状态

getCuttentTitle()：获取当前的标题

getCurrentPage()：获取当前的页码

getCurrentURL()：获取当前爬取的标题

getIsRunning()：获取是否运行的状态

getSpiderID()：获取爬虫标识符ID

getStatus()：获取状态对象

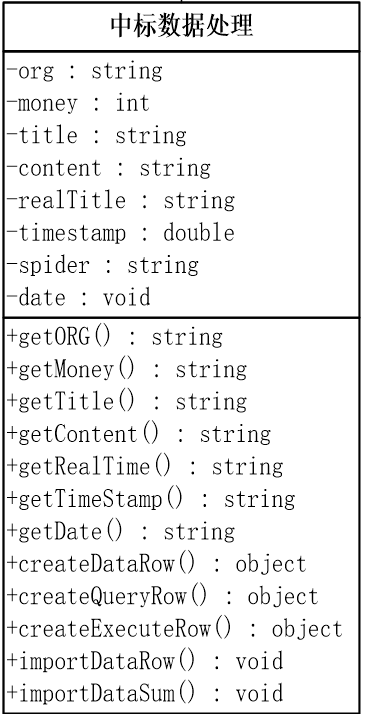
calculateSumStatus()：获取汇总状态

createDataRow()：创建数据一条

createQueryRow()：创建查询行

createExecuteRow()：创建执行行

### 中标数据处理



Org：中标单位

Money：中标金额

Title:项目名称

Content：项目内容

realTitle：项目真实名称

timestamp：时间戳

spider：爬虫名

date：中标时间

getORG()：获取中标人

getMoney()：获取中标金额

getTitle()：获取项目标题

getContent()：获取项目内容

getRealTIme()：获取项目名称

getTimeStamp()：获取时间戳

getDate()：获取中标时间

createDataRow()：创建数据行

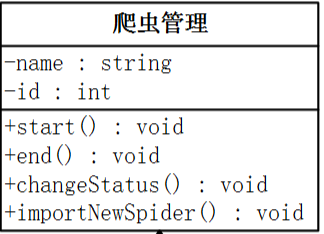
createQueryRow()：创建数据查询行

createExecuteRow()：创建数据执行行

importDataRow()：导入数据行

importDataSum()：导入数据统计

### 爬虫管理



Name：爬虫名称

Id：标识符ID

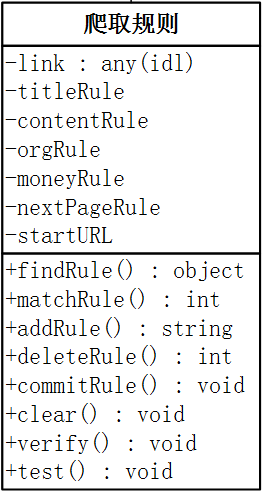
Start()：开始

End()：结束

changeStatus()：改变状态

importNewSpider()：导入新爬虫信息

### 爬取规则



Link:超链接规则

titleRule:标题规则

contentRule：内容规则

orgRule：中标人规则

moneyRule：中标金额规则

nextPageRule：写一页规则

startURL：开始地址

findRule()：查找规则

matchRule()：匹配规则

addRule()：添加规则

deleteRule()：删除规则

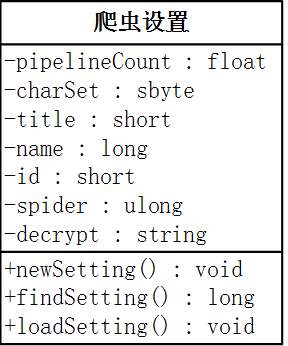
commitRule()：提交规则

clear()：清除所有规则

verify()：验证规则

test()：测试，不输出到视图

### 爬虫设置



pipelineCount：同时运行的网络线程数

charset：文字编码

title：标题

name：名称

id:标识符ID

spider:爬虫名

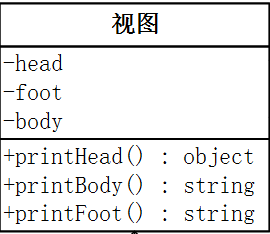
decrypt:加密方法

newSetting()：创建新设置

findSetting()：寻找已经存在的设置

loadSetting()：读取设置

### 视图



head:头信息

foot：网页尾部信息

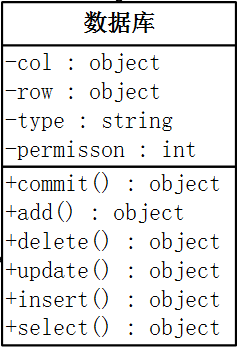
body：网页主题部分信息

printHead()：打印头信息

printBody()：打印主题部分信息

printFoot()：打印尾部信息

### 数据库



Col：列数

Row：行数

Type：类型

Permisson：访问权限

Commit()：提交

Add()：添加

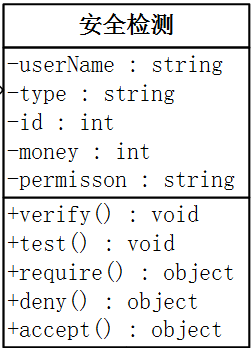
Delete()：删除

Update()：更新

Insert()：插入

Select()：查询

### 安全检测



userName：用户名

type：用户类型

id：用户标识符ID

money：积分

permisson：权限

verify()：验证

test()：测试

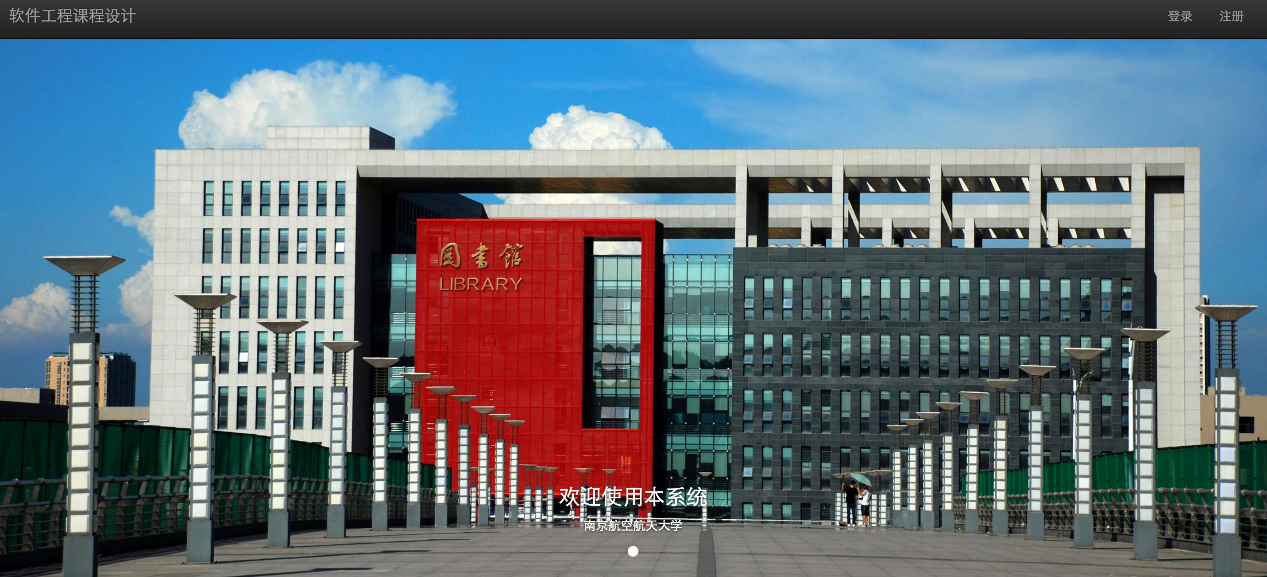
require()：需求

deny()：拒绝访问

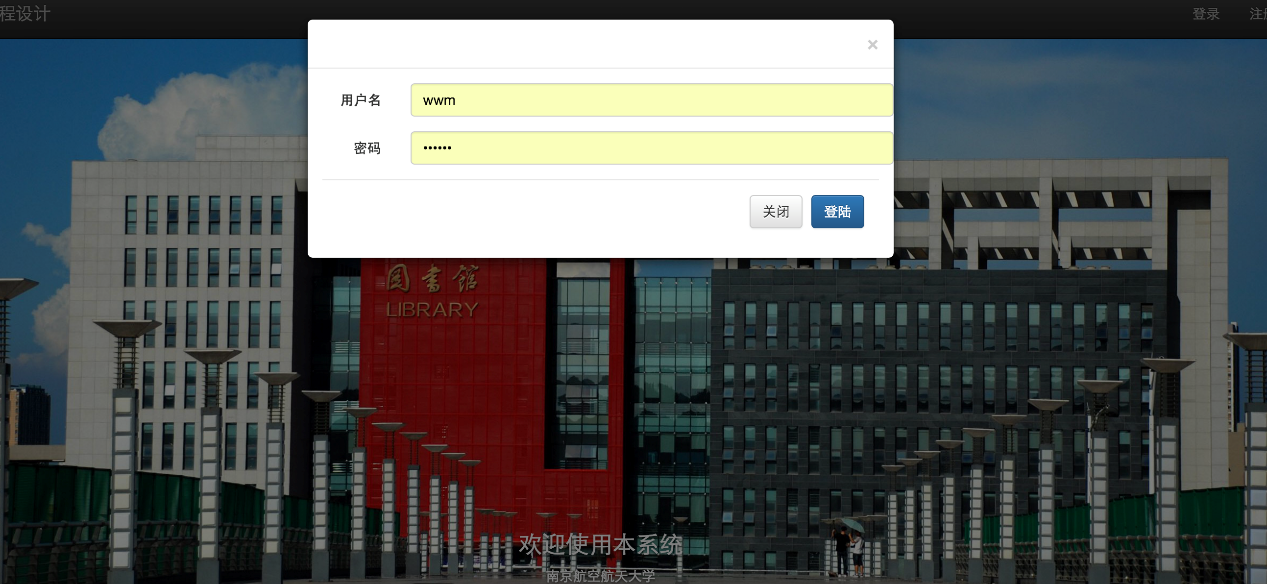
accept()：同意访问

## 界面设计

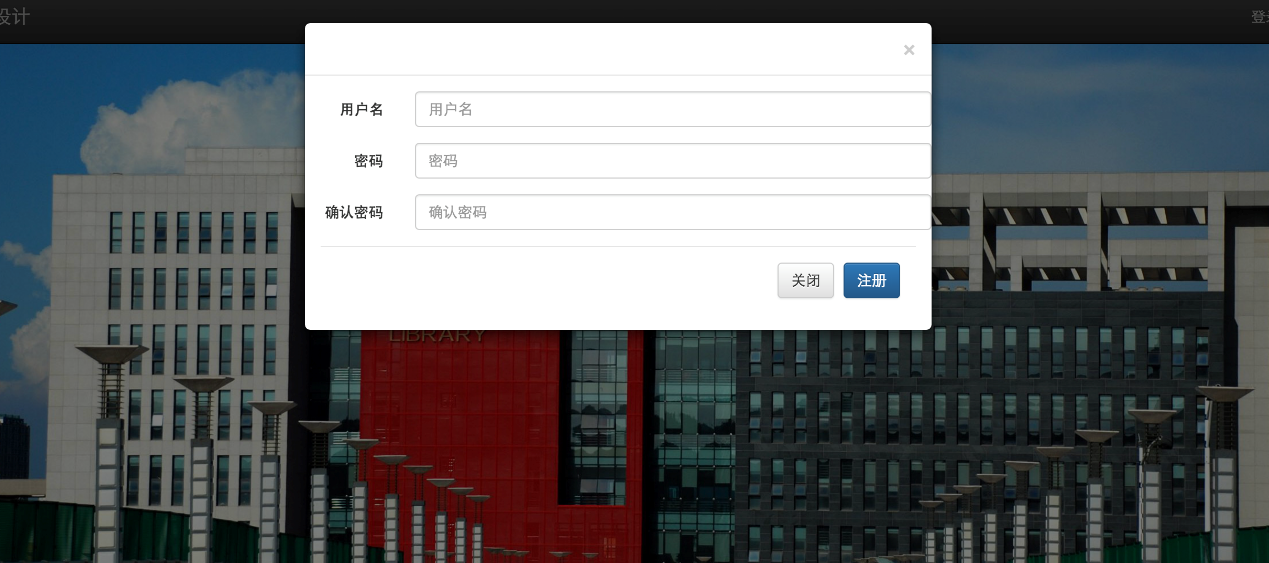
### 网站首页



### 登陆界面



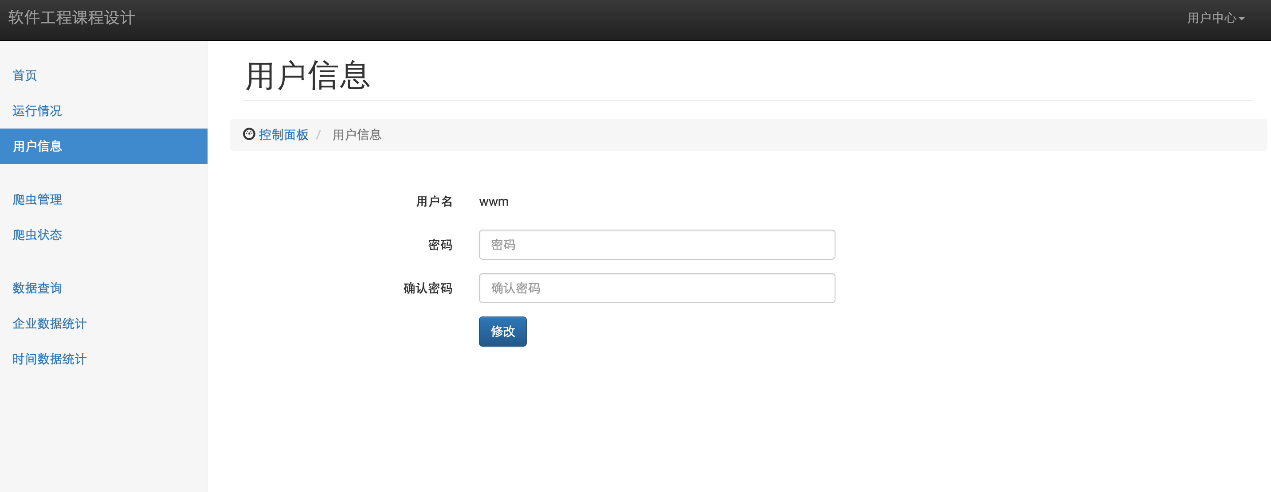
### 注册界面



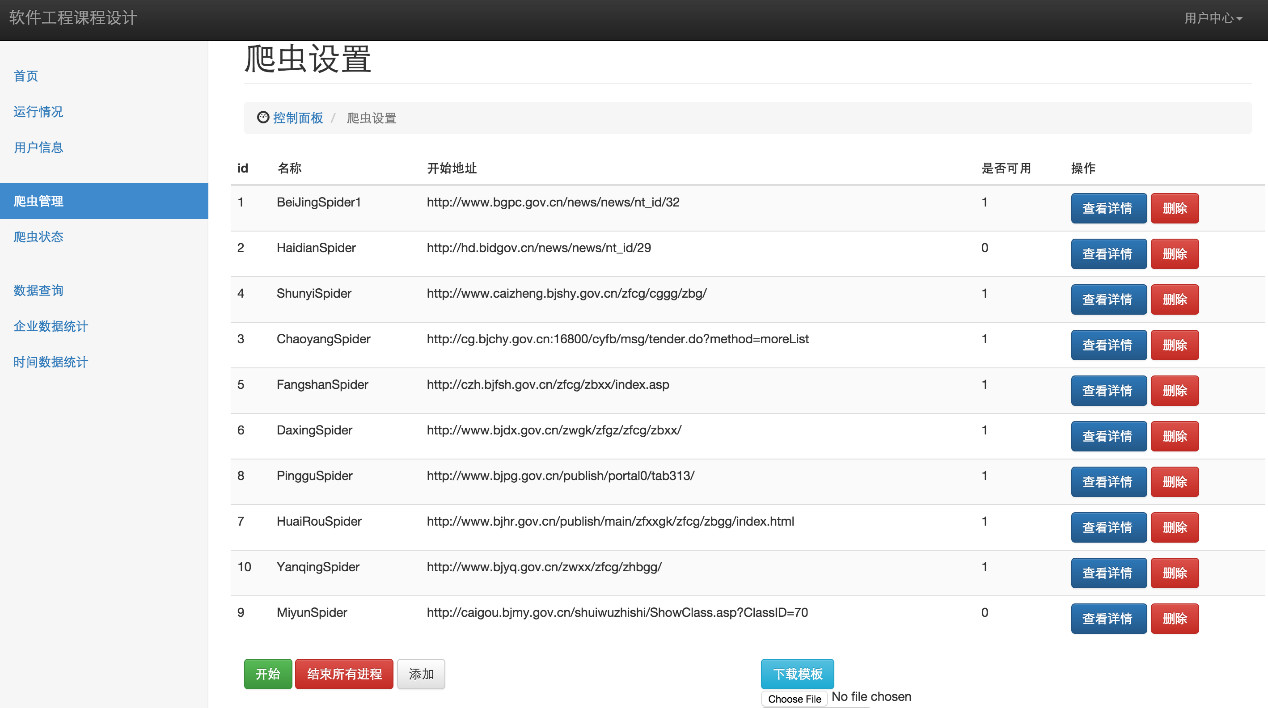
### 后台首页



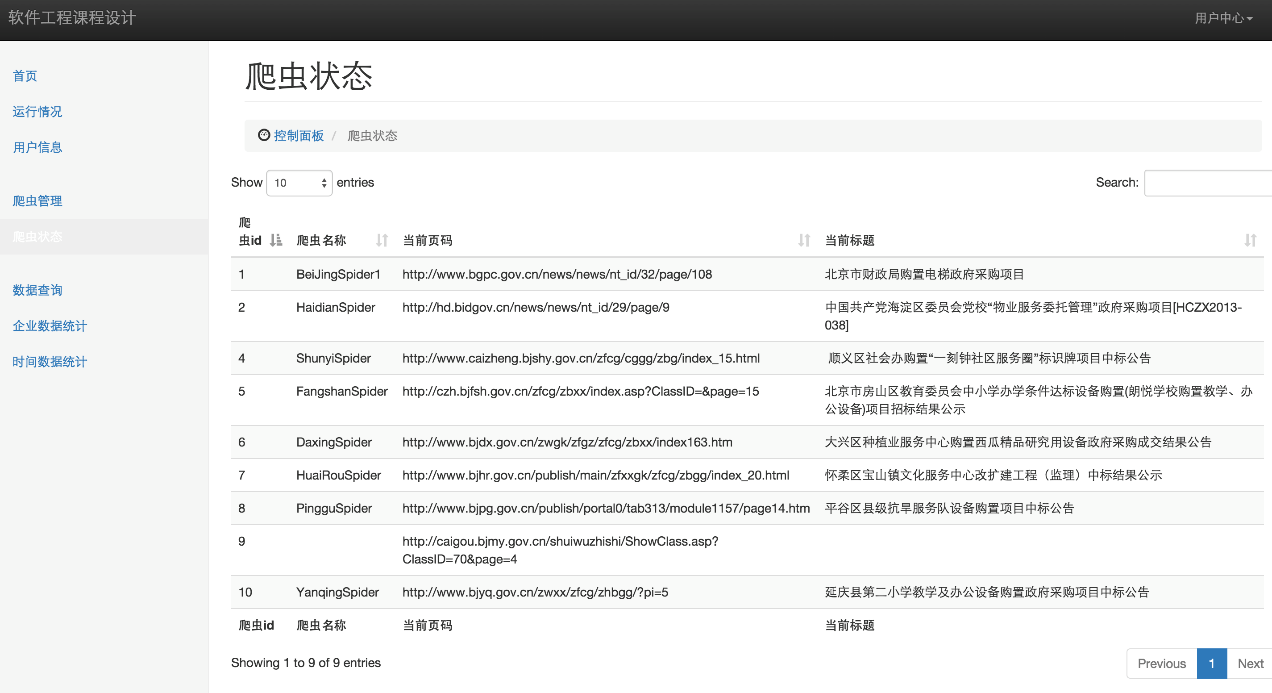
### 用户信息界面



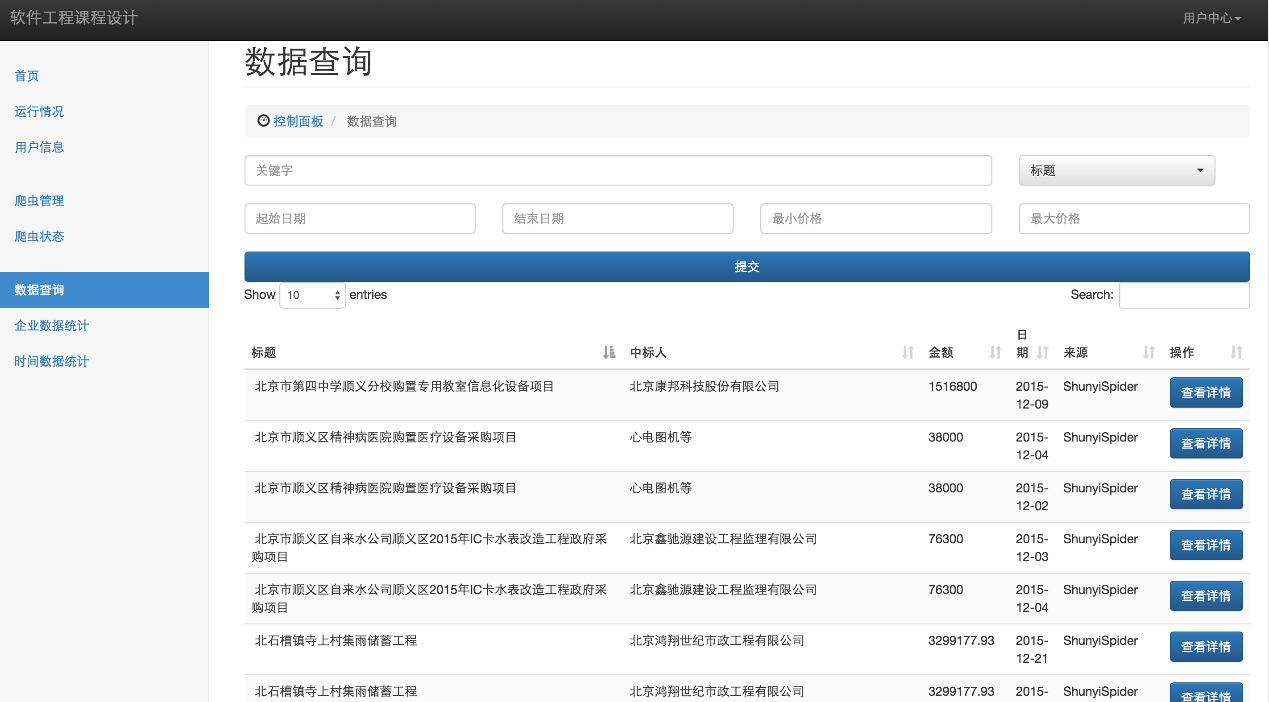
### 爬虫管理界面



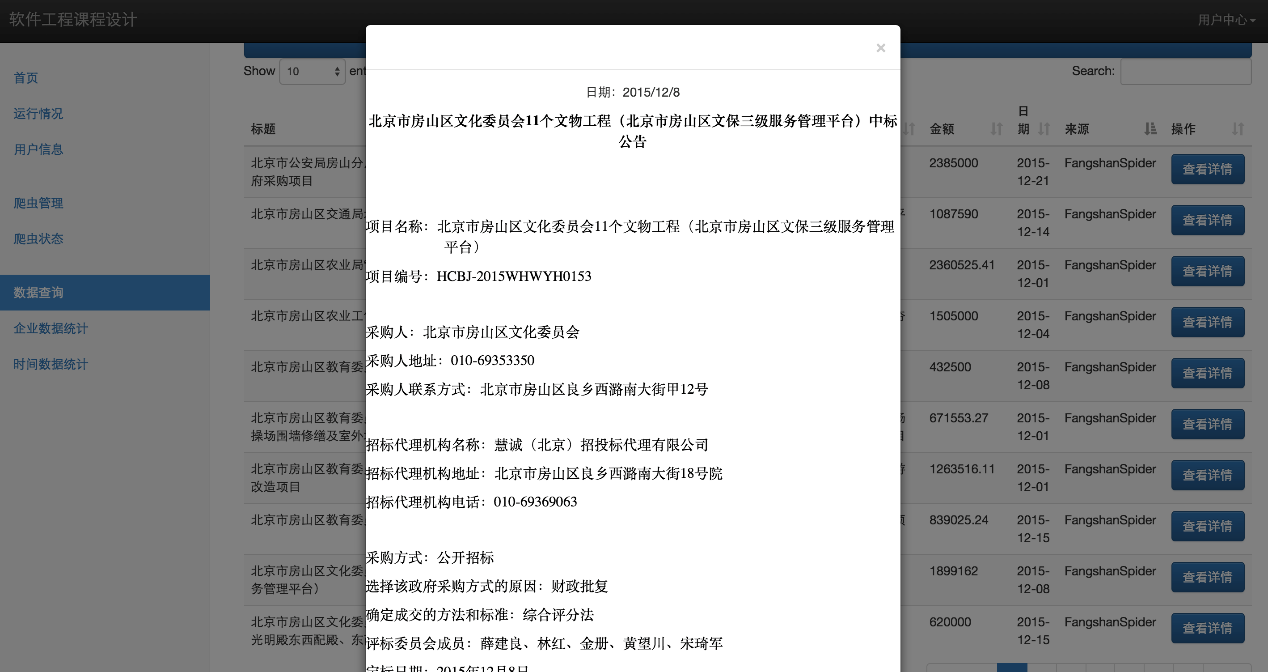
### 爬虫状态界面



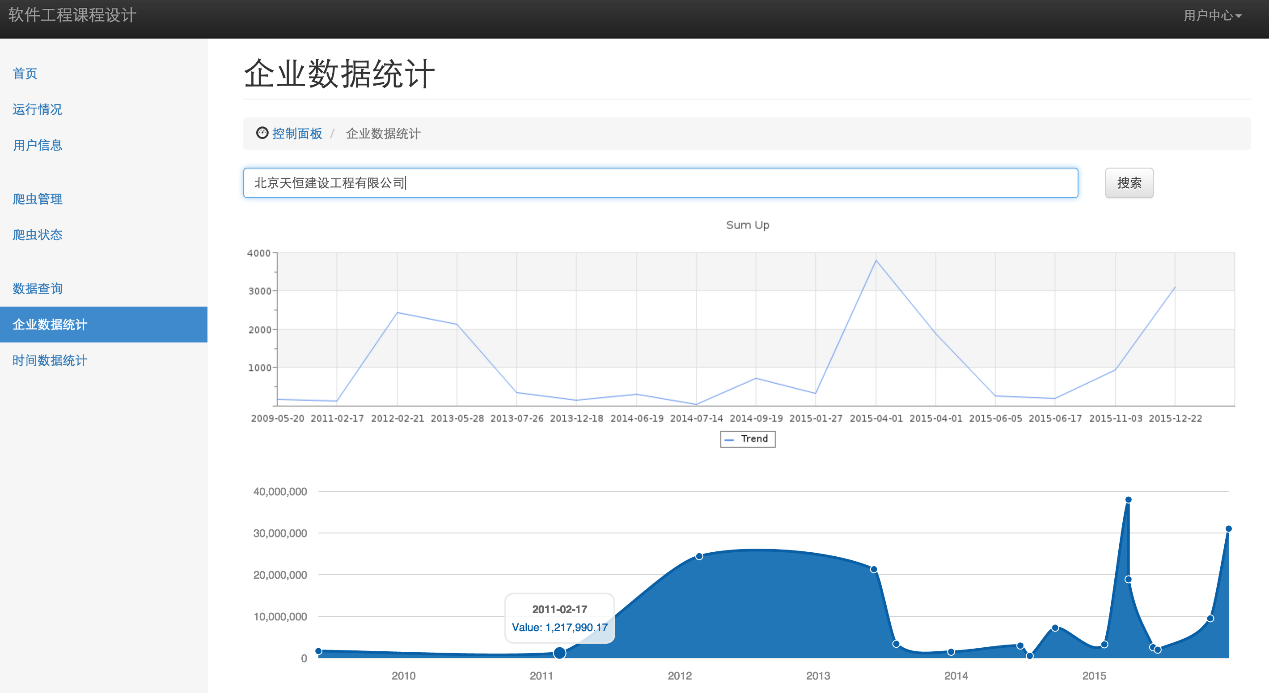
### 数据查询界面

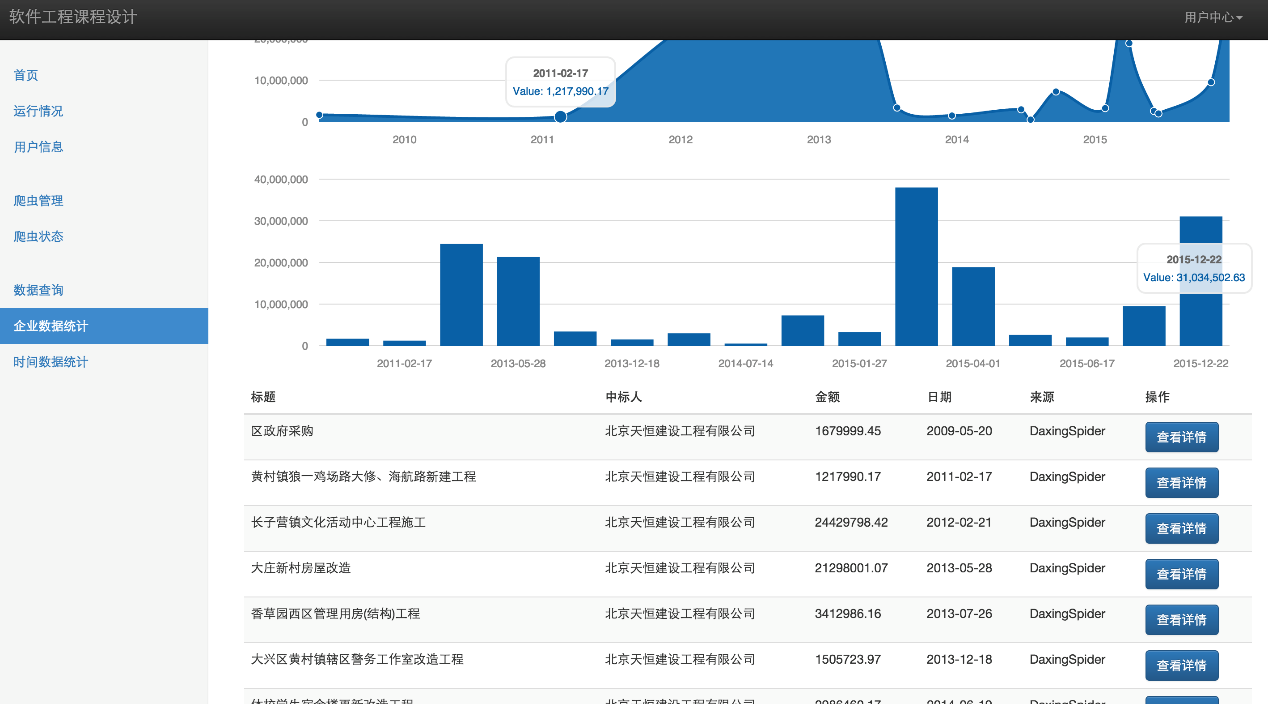


### 中标详情界面



### 企业中标情况统计界面





### 时间中标情况统计界面



# 系统数据结构设计

## 逻辑结构设计

### 总数据表（SE\_RAWDATA）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **字段名称** | **中文说明** | **键别** | **是否空** | **数据类型** | **长度** | **备注** |
| DataID | 标识Id | PK | NO | int | 32 | AI |
| Title | 项目名 |  | NO | VARCHAR | 32 |  |
| Content | 项目内容 |  | NO | TEXT |  |  |
| Spider | 来源网站 |  | NO | VARCHAR | 32 |  |
| Org | 中标单位 |  | YES | TEXT |  |  |
| Date | 中标日期 |  | YES | DATE |  |  |
| Money | 中标金额 |  | YES | Int |  |  |

### 爬取规则表（RULES）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **字段名称** | **中文说明** | **键别** | **是否空** | **数据类型** | **长度** | **备注** |
| RuleID | 权限标识Id | PK | NO | Int | 32 | AI |
| Name | 名称 |  | NO | VARCHAR | 32 |  |
| AllowedDomain | 允许的域名 |  | NO | VARCHAR | 64 |  |
| StartURL | 开始地址 |  | NO | VARCHAR | 64 |  |
| NextPageRule | 下一页匹配规则 |  | NO | VARCHAR | 64 |  |
| ArticleLinkRule | 超链接规则 |  | NO | VARCHAR | 64 |  |
| TitleRule | 标题匹配规则 |  | NO | VARCHAR | 64 |  |
| ContentRule | 内容匹配规则 |  | NO | VARCHAR | 64 |  |
| Enable | 当前爬虫是否可用 |  | NO | Int | 4 |  |

### 运行状态表（STATUS）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **字段名称** | **中文说明** | **键别** | **是否空** | **数据类型** | **长度** | **备注** |
| StatusID | 状态ID | PK， | NO | int | 32 | AI |
| Crawer\_lID | 爬虫ID | PK， | NO | int | 32 |  |
| Curren\_Page | 当前页码 |  | NO | TEXT | 64 |  |
| Current\_URL | 当前地址 |  | NO | TEXT | 64 |  |
| Current\_Title | 当前标题 |  | NO | TEXT | 64 |  |
| Name | 名称 |  | NO | TEXT | 64 |  |
| isRunning | 是否运行 |  | NO | Int | 4 |  |
| TimeStamp | 时间戳 |  | NO | DATETIME | 32 |  |

### 用户表（USER）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **字段名称** | **中文说明** | **键别** | **是否空** | **数据类型** | **长度** | **备注** |
| UserID | 标识Id | PK | NO | Int | 32 | AI |
| UserName | 用户名 |  | NO | VARCHAR | 32 |  |
| Password | 密码 |  | NO | VARCHAR | 32 |  |
| Type | 类型 |  | NO | VARCHAR | 32 |  |
| Money | 积分 |  | NO | Int | 32 | 默认为0 |