软件需求规格说明书

Project-Spider

1 引言

1.1 编写目的

为明确软件需求、安排项目规划与进度、组织软件开发与测试，撰写本文档。预期读者为参与本项目的开发人员。

1.2 背景

(1) 项目的名称

Project-Spider 项目开发。

(2) 项目的委托单位

四川大学 2015 级软件学院项目开发小组。

(3) 项目的用户（单位）：

插画家、专业的漫画师或者漫画爱好者。

(4) 项目的任务提出者

周林。

(5) 项目的主要承担部门

项目的承担部门主要有：开发部门、测试部门、集成部门等三个部门。

(6) 项目建设背景

从业务环境来看本项目能够使广大喜爱漫画的人员能够更简单的寻找到当前热点漫画和高点击量漫画，而且本爬虫项目还能够完善相关网站的搜索功能，让使用者能够更加方便的获取自己喜爱类型的相关漫画。

(7) 软件系统与其他系统的关系

本系统依赖于 node.js 平台、腾讯云、Mongodb 数据库等服务

(8) 软件系统与机构的关系

由于本项目是基于爬虫的一类项目，因此，对于目标网页的原式数据和相关程序不能篡改和破坏，而且对于我们获取的目标网页的数据我们要严格保护，在没经过对方同意的情况下坚决不用于商业行为。

本项目不需要外包，且不用经过专业测试机构的相关测试。

1.3 定义

网络爬虫：

1.定义：网络爬虫（又被称为网页蜘蛛，网络机器人，在 FOAF 社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

2.英文表示：web crawler

1.4 参考资料

【GB8567－88 软件需求说明书】

2 任务概述

2.1 目标

产品目标：该项目产品是网站，通过对爬到的数据进行的分析为对插画、原画有浓厚兴趣的人以及热爱美术的人提供检索、下载收集的便利。比如，可以为感兴趣的人员提供当前网站最火热的画师，或者提供又需要的问题答案。

产品范围：该项目产品是依据爬虫技术以及数据分析，通过爬虫技术获取数据，进而对收集到的数据进行分类、计算，从而为用户提供检索便利，同时能为用户做决策提供帮助。

2.2 用户的特点

本软件最终用户：插画家、专业的漫画师或者漫画爱好者。

特点：该系统用户主要面向插画家、专业的漫画师或者漫画爱好者，在次元文化流行的现在，该人群基数庞大，具有热情，狂热、消费能力高等特点。

2.3 假定和约束

项目进度安排：

编号	任务阶段	开始时间	结束时间
1	需求分析	2017/03/05	2017/04/4
2	系统开发	2017/04/5	2017/04/20
3	前后台联调、功能测试	2017/04/21	2017/05/02
4	部署测试、Bug Fix	2017/05/03	2017/05/15
5	Release	2017/05/16	2017/05/23

约束：若数据来源网站有反爬虫机制，更换数据来源网站。

3 需求规定

3.1 对功能的规定

IPO 表如下：

系统名称：project-Spider	设计人：周林
模块名：fetch	日期：2017/4/1
上层调用模块：requestHeader, log	下层被调用模块：parsePage, User, Util
输入数据：URL 地址	输出数据：执行回调函数
处理：传入目标 url,取得网页并将网页解析成支持 JQuery 选择器的$，并将待处理的数据传递给回调函数处理	
注释：设计模式：策略模式;根据不同的回调函数执行不同的方法,降低代码的耦合度	

系统名称：project-Spider	设计人：周林
模块名：download	日期：2017/4/1
上层调用模块：log，requestHeader	下层被调用模块：
输入数据： url, fn, author	输出数据： .JPG 或者.png 后缀的图片
处理： request(options) .pipe(fs.createWriteStream('./'+author+'/'+filename)) .on('error',function(err){console.log('error in download:'+err); return;}) .on('close', function(){console.log(filename+'下载完成');}); 通过请求头，伪装浏览器得到原网站的资源图片，并下载到本地	
注释： 传入三个参数：url 为原图地址，fn 为文件名，author 为文件保存的路径； url format: http://i2.pixiv.net/c/600x600/img-master/img/2014/08/14/23/06/55/45358677_p0_master1200.jpg; 通过 reference 解决图片防盗链问题	

系统名称：project-Spider	设计人：周林
模块名：database	日期：2017/4/1
上层调用模块：log	下层被调用模块：search，spider
输入数据： JSON 格式的图片、用户信息	输出数据： JSON 格式的图片、用户信息

处理：

将爬虫得到的数据保存到数据库中，用户模型如下：

```
var User = mongoose.Schema({
    uid:Number,//作者 ID 号
    works:[Number],//存储作品的 ID 号
    fans:Number,//粉丝数
    date：date//记录日期
});
```

作品模型如下：

```
var Works = mongoose.Schema({
    id:Number,//作品 ID 号
    uid:Number,//作者 ID
    agree:Number,//赞数
    grade:Number,//总分
    tags:[String],//作品 TAG
    date：date//记录日期
});
```

注释： 作者 primary key (uid); 作品 primary key (id), foreign key (uid) references User(uid)

系统名称：project-Spider	设计人：周林
模块名：spider	日期：2017/4/1
上层调用模块：log，requestHeader，fetch，database，parsePage	下层被调用模块：database，router
输入数据： URL，target	输出数据： 作者信息，图片信息
处理： 是一个方法集合的模块，可进行各种操作	

系统名称：project-Spider	设计人：周林
模块名：router	日期：2017/4/1
上层调用模块：spider，download，database	下层被调用模块：
输入数据： 客户端用户的 POST 以及 GET 请求	输出数据： 返回相应的页面

处理：

根据用户的请求，将其分配到不同的路由中进行处理，

```
routes(app);
app.get('/',require('./home'));
app.use(function(req,res){
    res.send('404');
});
```

功能及路由设计如下：

1. 主页
    1. 主页：`GET /`
2. 搜索
    1. 搜索：`POST /search/:keyword=???`
    2. 搜索结果:`GET /search/results`
3. 下载
    1. 下载：`POST /download`

注释： 该模块用于管理客户端和服务器的交互

> 路由设计采用`REST`风格

系统名称：project-Spider	设计人：周林
模块名：search	日期：2017/4/1
上层调用模块：router，database，spider	下层被调用模块：router
输入数据： URL，keyword	输出数据： JSON 格式的作者、图片信息
处理： 服务端解析客户端传递过来的数据，并返回相应的数据给客户端	
注释： 当出现错误时，返回 404 界面	

3.2 对性能的规定

3.2.1 精度

a. 对于约束条件，采用整数类型

b. 输出数据精度的要求，保留六位有效数字的整数

3.2.2 时间特性要求

在网络通畅的情况下，有以下时间限制要求：

a． 用户得到网页的响应时间：50ms 以内

b． 图片加载完全的时间：5 秒以内

3.2.3 灵活性

a． 当在 winxp 或 win7 运行时能够兼容；

b． 保证对 ie（ie8 及以上版本）或其他一些主流非 ie 内核浏览器的支持；

c． 当输入整数能自动转换为两位小数的精度，遇错误类型数据及时报错；

d． 软件系统进行升级时保证用户数据的安全性。

3.3 输人输出要求

	类型	格式	数值范围	精度
输入	查询	字符串	无	无
	表单查询	JSON 表单	整数	无
输出	图片	JPG/PNG	无	无
	作者/作品信息	字符串	无	无
	错误信息	HTML	无	无

3.4 数据管理能力要求

用户表：

字段名	类型	主键
uid	Number	yes
works	[String]	
fans	Number	
date	date	

作品表

字段名	类型	主键
id	Number	yes
uid	Number	foreign key (uid) references User(uid)
agree	Number	
grade	Number	
tags	[String]	
date	date	

3.5 故障处理要求

正常使用时不报错，若运行时遇到不可恢复系统错误，也必须保证数据库完好无损。遇到问题及解决方案：

a. 若遇到不可识别路由请求，返回 404 页面并提示用户

b. 若服务器错误，返回 500 页面并提示用户

c. 关于数据库若无法得到特定数据，抛出错误并继续执行

d. 下载图片时若由于网络等问题无法得到完整图片，监听错误写入日志并继续执行任务，伪代码如下：

```
Download.on('error',function(err){log(err); return;})
```

3.6 其他专门要求

用户在搜索时会有搜索提示

4 运行环境规定

4.1 设备

◎硬件：
服务器 CPU：1 核 1Mbps 或更高
内存：1GB 或更高
硬盘：50G 或更高
网络：至少一台服务器
使用 TCP/IP 协议的网络服务

4.2 支持软件

◎软件：
服务器操作系统：centOS6.5 或以上
开发平台操作系统：Windows10
后端平台：Node.js
开发工具：Sublime Text 3
数据库系统：Mongodb v3.4.1
配置管理工具：Git 2.10.2.windows.1

4.3 接口

由于该网站为依赖爬虫技术提供服务的资源类网站，因此需要与 http://www.pixiv.com/ 保持通信。

协议：TCP/IP，http 协议

4.4 控制

运行方法：输入网址，进入网站即可访问

控制信号：鼠标点击触发链接

控制信号来源：鼠标点击