**分布式爬虫**

**概要设计说明书**

组内成员： 祖浩然、邵文政、陈鹏

**目录**

[1.引言 1](#_Toc1361)

[1.1编写目的 1](#_Toc285)

[1.2项目背景 1](#_Toc18305)

[1.3定义 2](#_Toc13619)

[1.4参考资料 2](#_Toc7920)

[2.任务概述 2](#_Toc10477)

[2.1目标 2](#_Toc22092)

[2.2运行环境 2](#_Toc18782)

[2.3条件与限制 3](#_Toc9992)

[3.总体设计 3](#_Toc12308)

[3.1处理流程 3](#_Toc6850)

[3.2总体结构和模块外部设计 3](#_Toc17781)

[3.3功能分配 4](#_Toc20633)

[3.3.1下载模块 4](#_Toc27432)

[3.3.2 解析模块 4](#_Toc3699)

[3.3.3 调度模块 4](#_Toc5848)

[4接口设计 4](#_Toc1355)

[4.1内部接口 4](#_Toc18365)

[5数据结构设计 5](#_Toc10483)

[5.1逻辑结构设计 5](#_Toc21979)

[5.1数据结构与程序的关系 6](#_Toc16119)

[6.运行设计 6](#_Toc20441)

[6.1运行模块的组合 6](#_Toc9600)

[6.2运行控制 6](#_Toc1236)

[6.3运行时间 6](#_Toc6761)

[7.出错处理设计 7](#_Toc424)

[7.1出错输出信息 7](#_Toc19429)

[8.安全保密设计 7](#_Toc17779)

# 

# 1.引言

## 1.1编写目的

对本爬虫的概要设计进行描述。将工作原理简要说明。预期读者为，软件小组全体成员、指导老师、评委老师。

## 1.2项目背景

爬虫系统，是对海量的分散的互联网数据进行采集的系统，是搜索引擎系统的基础。大数据近年来快速发展，炙手可热，不仅是数据的容量大，更是强调对全样本的数据的分析。互联网数据中包含了大量有价值信息，是大数据的重要数据来源。

而互联网上的数据内容丰富，组织形式也灵活多样。传统的爬虫系统，对所有的网页采用同样的办法处理，利用深度优先或广度优先的办法获取网页链接，下载网页，对网页中的所有的文本数据建立倒排索引。这种方式没有对网页数据的信息进行组织、归类。

应大数据的需求，分布式爬虫系统是解决这一问题的方案。分布式爬虫，对同一个网站的同类数据，进行结构化。同时，能利用分布式的软件设计方法，实现爬虫的高效采集。

## 1.3定义

## 1.4参考资料

1. 《中国软件杯大学生软件设计大赛》题目要求
2. 《分布是网络爬虫开发计划》
3. 《基于Hadoop的分布是网络爬虫系统的研究与实现》 -----=-孙瑞---2015.1.5

# 2.任务概述

## 2.1目标

实现分布式网络爬虫，能够快速高效的爬取网络资源，能够自适应的对网络信息进行爬取筛选整合，得到有价值的信息；能够绕过反扒机制实现自动化爬取；

## 2.2运行环境

操作系统：ubuntu 16.04

支持环境：MyEclipse + JDK1.7

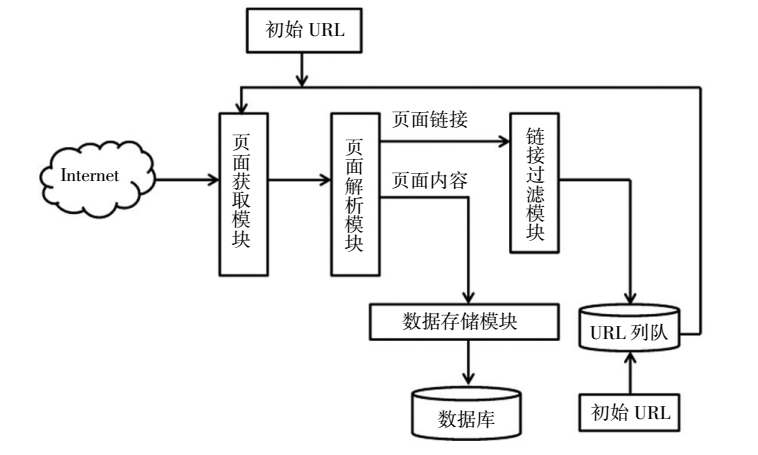
数 据 库：MySQL （5.0版本）

## 2.3条件与限制

a.计算机在网络通畅，网速良好的环境下。

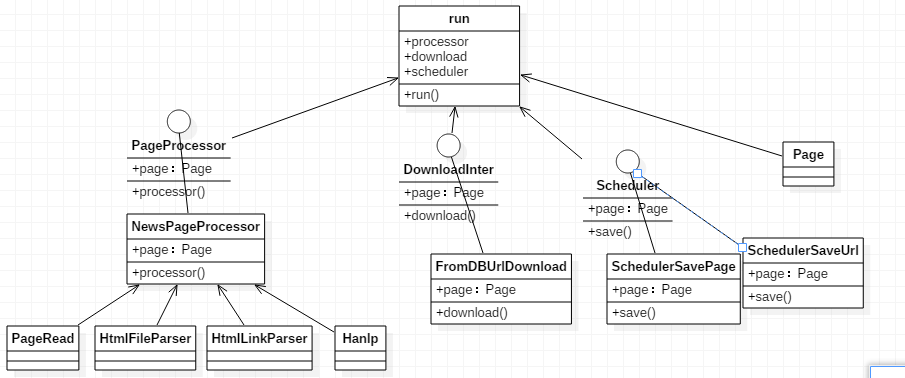
# 3.总体设计

## 3.1处理流程



## 

## 3.2总体结构和模块外部设计

程序的入口run调用接口downloader进行下载，将下载的数据传入PageProcessor进行解析，解析后的数据传入Scheduler进行存储

## 3.3功能分配

### 3.3.1下载模块

爬取网页模块采用socket通信方式实现客户端与服务器的通信：首先将客户端与服务器进行三次握手后建立连接，客户端发送HTTP请求头，服务器端收到客户端请求后，进行HTTP响应，发送相应的网页信息，客户端收到服务器的响应后将所获得网页文件交给网页分析模块进行处理并提取URL，存到相应的数据库及文档中供前台调用显示；

### 3.3.2 解析模块

网页分析模块的网页分析处理主要在于对抓取到的HTML文件的内容进行URL的提取，和网页内容的分析。Url的提取采用的是htmlparser库。网内容的分析采用htmlparser库和hanlp文本分析库，分别进行内容的提取过滤，和自动关键字、自动摘要。

### 3.3.3 调度模块

将分析好的数据存入数据库，并可以通过接口进行调取。存入url时进行布隆过滤。然后存入数据库。存入网页信息时，将信息转换为二进制，然后存入数据库。

# 4接口设计

## 4.1内部接口

a.网页下载接口

b.网页解析接口

c.调度接口

# 5数据结构设计

## 5.1逻辑结构设计

pagesave原始网页表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 中文含义 | 数据类型 | 允许空 | 约束条件 |
| pageURL | 网页url | nvarchar(20) | N | 主键 |
| document | 网页源代码的二进制 | blob | N | 无 |

urlwait带抓取网页url

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 中文含义 | 数据类型 | 允许空 | 约束条件 |
| Url | 网页url | nvarchar(20) | N | 主键 |

page网页数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 中文含义 | 数据类型 | 允许空 | 约束条件 |
| Url | url | nvarchar(20) | N | 主键 |
| time | 时间 | nvarchar(20) | N | 无 |
| title | 标题 | nvarchar(20) | N | 无 |
| h1 | 第一段 | nvarchar(20) | N | 无 |
| p | 段落 | blob | N | 无 |
| keyname | 关键字 | nvarchar(20) | N | 无 |
| zhaiyao | 摘要 | nvarchar(20) | N | 无 |

## 5.1数据结构与程序的关系

下载模块将原始网页存入pagesave表中，调度模块将解析出来的url放入urlwait表中，将解析出的数据放入page中。

# 6.运行设计

## 6.1运行模块的组合

调度模块从带抓取库中提取带爬取的url传给下载模块，下载模块接收url，并下载url对应的网页，然后将下载的网页分别保存数据库和传入解析模块。解析模块将url和有用的信息提取出来，传入调度模块。调度模块将数据存入数据库。

## 6.2运行控制

运行时有较友好的界面，基本能够实现网页的爬取和信息处理。按照模块之间的关系进行调用。

## 6.3运行时间

系统响应时间在2s以内。爬取时间因网络差异不定。在网速500k/s时爬取网页每分钟10个。

# 7.出错处理设计

## 7.1出错输出信息

当用户输入不合法的url时，系统弹出警告，无法爬取

# 8.安全保密设计

爬取的主要数据以二进制流进行存取。