目 录

## 绪论

#### 课题背景及意义

#### 发展历史

#### 发展现状

#### 论文研究内容

## 网络爬虫技术基础

#### 2.1爬虫的分类与作用

###### 2.1.1爬虫的定义

###### 2.1.2爬虫的作用与分类:

#### 2.2HTTP协议

###### 2.2.1Http协议定义

###### 2.2.2 POST与GET的请求

###### 2.2.3 Cookie和Session

###### 2.2.4 HTTP响应报文

#### 2.3Rebots协议

###### 2.3.1网络爬虫引发的问题

###### 2.3.2Rebots协议

#### 2.4网站的分析

###### 2.4.1微博移动版网站的分析

#### 2.5 爬虫伪装

###### 2.51User-agent伪装

#### 2.6信息过滤规则-正则表达式

###### 2.6.1 HTML页面解析:

###### 2.6.2正则表达式简介

###### 2.6.3正则表达式相关注解

###### 2.6.4Chrome编写正则表达式

## 第3章 爬虫系统架构设计

#### 3.1.非关系型数据库

###### 3.1.1NOSQL的介绍

###### 3.1.2Mongodb简介

###### 3.1.3Mongodb搭建

###### 3.1.4Redis简介

###### 3.1.5Redis搭建

#### 3.2 Scrapy框架

###### 3.2.1Scrapy框架简介

###### 3.2.2Scrapy框架解析

###### 3.2.3Scrapy创建

#### 3.3 Scrapy+redis架构

###### 3.3.1 Scrapy+Redis架构定义

###### 3.3.2 组件间关系:

## 第4章 爬虫重难点

#### 4.1爬虫搜索策略-防止环路的出现

###### 4.1.1网站的树结构

###### 4.1.2网络爬虫图的搜索策略

#### 4.2去重

###### 4.2.1去重原理

###### 4.2.2 Redis去重

#### 4.3反扒技术

###### 4.3.1企业中常见的反爬技术手段

###### 4.3.2对抗AJAX技术

###### 4.3.3 JSON格式爬取

###### 4.3.4验证码处理

#### 4.4 Cookies池

###### 4.4.1 Cookies池简介

###### 4.4.2构建Cookies池

## 第5章 数据模型设计与分析

#### 5.1数据模型

###### 5.1.1数据库概念设计——E-R模型

###### 5.1.2数据库逻辑设计

###### 5.1.3数据表设计

#### 5.2数据的简单分析

###### 5.2.1数据分析

## 第6章 总结与展望

###### 6.1总结

###### 6.2不足与展望

## 参考文献

## 致谢

## 附录