目 录

## 摘要

## ABSTRACT

## 绪论

#### 课题背景

#### 发展历史

#### 发展现状

* 1. **研究意义**
  2. **研究目标**
  3. **研究中遇到的关键问题**
  4. **论文主要的组织结**

## 网络爬虫基本知识与相关技术

#### 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

#### 2.3Rebots协议

###### 2.3.1网络爬虫引发的问题

###### 2.3.2Rebots协议

#### 2.4爬虫搜索策略-防止环路的出现

###### 2.4.1网站的树结构

###### 2.4.2网络爬虫图的搜索策略

## 第3章 爬虫系统架构设计

#### 3.1需求分析

###### 3.1.1需求分析

#### 3.2非关系型数据库

###### 3.2.1NOSQL的介绍

###### 3.2.2Mongodb简介

###### 3.2.3Mongodb搭建

###### 3.2.4Redis简介

###### 3.2.5Redis搭建

#### 3.3 Scrapy框架

###### 3.3.1Scrapy框架简介

###### 3.3.2Scrapy框架解析

###### 3.3.3Scrapy创建

#### 3.4 Scrapy+redis架构

###### 3.4.1 Scrapy+Redis架构定义

###### 3.4.2 组件间关系:

## 第4章 爬虫遇到的关键问题解决

#### 4.1网站的分析

###### 4.1.1微博移动版网站的分析

#### 4.2 爬虫伪装

###### 4.2.1User-agent伪装

#### 4.3信息过滤规则-正则表达式

###### 4.3.1 HTML页面解析:

###### 4.3.2正则表达式简介

###### 4.3.3正则表达式相关注解

###### 4.3.4Chrome编写正则表达式

#### 4.4去重

###### 4.4.1去重原理

###### 4.4.2 Redis去重

#### 4.5反扒技术

###### 4.5.1企业中常见的反爬技术手段

###### 4.5.2对抗AJAX技术

###### 4.5.3 JSON格式爬取

###### 4.5.4验证码处理

#### 4.6 Cookies池

###### 4.6.1 Cookies池简介

###### 4.6.2构建Cookies池

## 第5章 实验测试

#### 5.1测试环境

#### 5.2运行状态及测试

#### 5.3测试过程其他组件过程展示

#### 5.4数据抓取结果展示与分析

## 第6章 数据库设计与数据分析

#### 6.1数据模型

###### 6.1.1数据库概念设计——E-R模型

###### 6.1.2数据库逻辑设计

###### 6.1.3数据表设计

#### 6.2数据的分析

###### 6.2.1数据分析

## 第7章 总结与展望

###### 7.1总结

###### 7.2不足

###### 7.3展望

## 参考文献

## 致谢

## 附录