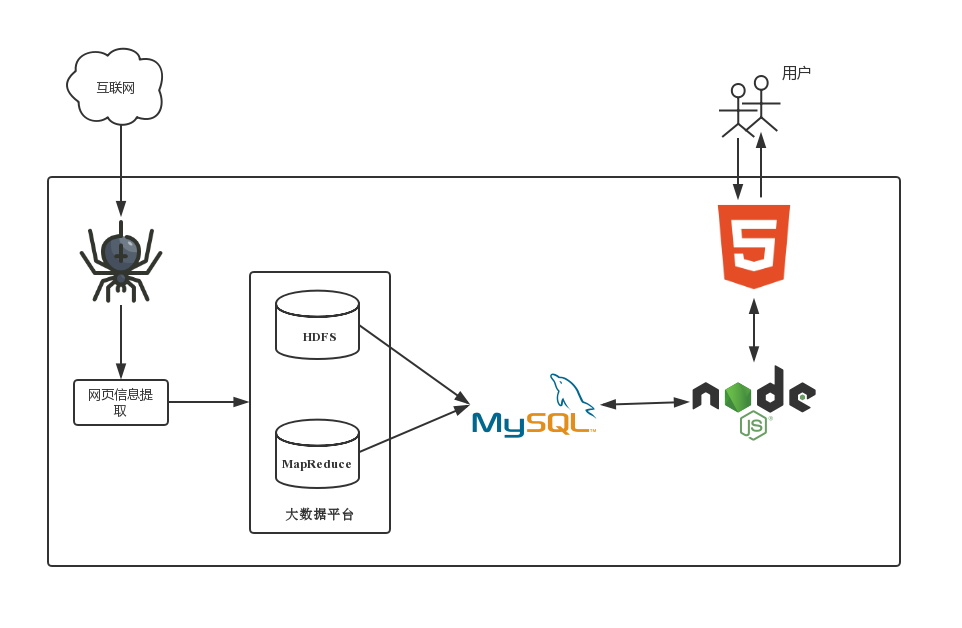
**文献综述**

1. **引言**

随着信息社会的高速发展，移动个人数据、社交网站、科学计算、证券交易、网站日志、传感器网络等产生的数据量迎来爆发式增长。其速度远远超出了单机计算和存储能力提升的速度。相对于传统的数据处理方法，大数据处理方法采用多机器、多节点来处理海量数据。该方法特点是随着数据量的不断加大，可以增加机器数量，水平扩展。一个大数据系统，可以多达几万台机器甚至更多。

本文主要阐述了面向高校学者信息大数据平台的分布式爬虫设计与实现。利用爬虫爬取海量的高校学者信息，借助大数据平台对学者信息进行清洗和统计，最后以web形式将统计结果可视化。

1. **系统架构**



1. **关键技术**
   1. **Python**

Python是一种计算机程序设计语言。是一种动态的、面向对象的脚本语言，最初被设计用于编写自动化脚本(shell)，随着版本的不断更新和语言新功能的添加，越来越多被用于独立的、大型项目的开发。Python语言主要有如下特点：

1. 简单：语法简洁，语义清晰，符合一般人的思维逻辑。对于新手而言，python学习成本低，极易上手。
2. 第三方库丰富：得益于python胶水语言的特性，python的第三方库非常丰富。它可以帮助你处理各种工作，包括正则表达式、文档生成、单元测试、线程、数据库、网页浏览器、CGI、FTP、电子邮件、XML、XML-RPC、HTML、WAV 文件、密码系统、GUI（图形用户界面）、Tk 和其他与系统有关的操作。本系统就用到了requests、redis、pymysql、flask、beautifulsoup等这些库。
3. 可移植性强：由于它的开源本质，Python 已经被移植在许多平台上（经过改动使它能够工作在不同平台上）。如果你小心地避免使用依赖于系统的特性，那么你的所有 Python 程序无需修改就可以在下述任何平台上面运行。这些平台包括 Linux、Windows、FreeBSD、Macintosh、Solaris、OS/2、Amiga、AROS、AS/400、BeOS、OS/390、z/OS、Palm OS、QNX、VMS、Psion、Acom RISC OS、VxWorks、PlayStation、Sharp Zaurus、Windows CE 甚至还有 PocketPC、Symbian 以及 Google 基于 Linux 开发的 Android平台。
4. 交互性强：支持互动模式，您可以从终端输入执行代码并获得结果。
   1. **网络爬虫**

网络爬虫是一个自动提取网页信息的程序。爬虫从一个或若干初始网页的URL开始，根据一定的网页分析算法提取出有效信息，并且保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。

* 1. **Docker**

Docker 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中，然后发布到任何流行的 Linux 机器上，也可以实现虚拟化。容器是完全使用沙箱机制，相互之间不会有任何接口。

* 1. **Redis**

Redis 是一个开源（BSD许可）的，内存中的数据结构存储系统。它可以用作数据库、缓存和消息中间件。它支持多种类型的数据结构，如字符串（strings），散列（hashes），列表（lists），集合（sets），有序集合（sorted sets）与范围查询，bitmaps， hyperloglogs和地理空间（geospatial）索引半径查询。Redis 内置了复制（replication），LUA脚本（Lua scripting），LRU驱动事件（LRU eviction），事务（transactions）和不同级别的磁盘持久化（persistence），并通过Redis哨（Sentinel）和自动 分区（Cluster）提供高可用性（high availability）。

* 1. **HDFS**

Hadoop Distributed Filesystem：Hadoop分布式文件系统(HDFS)被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统。HDFS有着高容错性的特点，并且设计用来部署在低廉的硬件上。它提供高吞吐量来访问应用程序的数据，适合那些有着超大数据集的应用程序。HDFS放宽了POSIX的要求这样可以实现流的形式访问（streaming access）文件系统中的数据。

* 1. **MapReduce**

MapReduce是一种编程模型，用于大规模数据集（大于1TB）的并行运算。概念"Map（映射）"和"Reduce（归约）"，是它们的主要思想，都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。当前的软件实现是指定一个Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的Reduce（归约）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

* 1. **MySQL**

MySQL是一个关系型数据库管理系统。关系数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。MySQL所使用的 SQL 语言是用于访问数据库的最常用标准化语言。MySQL 软件采用了双授权政策，分为社区版和商业版，由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点，一般中小型网站的开发都选择 MySQL 作为网站数据库。目前MySQL8.0已经支持文档存储、事务性数据字典、通用关系表达式等符合现代化需求的功能。

* 1. **Node.js**

Node.js采用Google开发的V8运行代码，使用事件驱动、非阻塞和 异步输入输出模型等技术来提高性能，可优化应用程序的传输量和规模。这些技术通常用于数据密集的即时应用程序。Node.js大部分基本模块都用JavaScript语言编写。在Node.js出现之前，JavaScript通常作为客户端程序设计语言使用，以JavaScript写出的程序常在用户的浏览器上运行。Node.js的出现使JavaScript也能用于服务端编程。Node.js含有一系列内置模块，使得程序可以脱离Apache HTTP Server或IIS，作为独立服务器运行。本系统采用Node.js来开发后台。

* 1. **Express**

Express 是一个简洁而灵活的 node.js Web应用框架, 提供了一系列强大特性帮助你创建各种 Web 应用，和丰富的 HTTP 工具。使用 Express 可以快速地搭建一个完整功能的网站。Express 框架核心特性：可以设置中间件来响应 HTTP 请求。定义了路由表用于执行不同的 HTTP 请求动作。 可以通过向模板传递参数来动态渲染 HTML 页面。

* 1. **Echarts**

ECharts开源来自百度商业前端数据可视化团队，基于html5 Canvas，是一个纯Javascript图表库，提供直观，生动，可交互，可个性化定制的数据可视化图表。创新的拖拽重计算、数据视图、值域漫游等特 性大大增强了用户体验，赋予了用户对数据进行挖掘、整合的能力。

1. **总结**

本系统主要分为三大部分：分布式爬虫的设计与实现、大数据平台的运算以及web前后端的开发。在实现这三大部分的过程中，遇到过很多困难。比如对Redis并发访问的控制、大数据平台的搭建、MapReduce程序的编写、前端样式适配等等。

通过爬虫可以为科研实验提供训练数据；通过Hadoop可以提高算力，缩短实验运行时间；通过echarts可以将数据可视化，展现数据之美。

整个系统的完成极大地提高了本人的工程实践能力，也为以后的科研生涯打下了基础。希望在以后的学习生活中再接再厉，勇攀科研高峰。

1. **参考文献**

[1]崔庆才.Python3网络爬虫开发实战.人民邮电出版社，2018.

[2]余明辉，张良均.Hadoop大数据开发基础.人民邮电出版社,2017.

[3]林大贵.Python+Spark2.0+Hadoop机器学习与大数据实战.清华大学出版社，2017.

[4]安俊秀，王鹏，靳宇倡.Hadoop大数据处理技术基础与实践.人民邮电出版社，2015.

[5]Thomas H.Cormen，Charles E.Leiserson，Ronald L.Rivest Clifford Stein.算法导论.机械工业出版社，2013.

[6]Hector Garcia-Molina，Jeffrey D.Ullman，Jennifer Widom.数据库系统实现.机械工业出版社，2010.

[7]Marc Harter.Node.js实战.人民邮电出版社，2014.

[8]Silberschatz，Galvin，Gagne.操作系统概念.高等教育出版社.2007.

[9]谢希仁.计算机网络.电子工业出版社.2017.