**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. **框架**

本系统拟定用JSP来做页面显示，用JAVA来做后台处理。MySQL来存储数据。

数据流图：

操作数据库

MySQL

调用

JAVA

JSP

返回记录

返回结果

系统总体框架为：

注册页面

会员注册页面

聊天系统首页

聊天界面

聊天区选择页面

会员登录页面

系统管理页面

管理员登录页面

模块图：

在线聊天系统

聊天模块

管理模块

注册登录模块

数据库管理模块

传输文件模块

文件分享模块

音乐模块

在线用户列表模块

截图模块

聊天区选择模块

聊天信息显示模块

信息输入模块

1. **关键技术**
   1. **JSP技术**

JSP（JAVA Server Pages）动态网页，是微软公司推出的一种用以取代CGI（Common Gateway Interface）通用网关接口的技术。JSP本身不是脚本语言，它只是提供了一种镶嵌在HTML页面中的脚本程序得以运行的环境。我们可以通过JSP结合HTML语言、JSP指令和JAVAX元件以及数据库等方面的知识，使用自己的WEB服务器创建并运行动态式的交互式WEB站点。用户端只要使用可执行HTML码的浏览器，即可浏览JAVA Server Pages所设计的网页内容。JAVA Server Pages所使用的脚本语言均在WEB服务器端执行，用户端的浏览器无需执行这些脚本语言。

目前，JSP（JAVA Server Pages）技术已经相当的普及，在Internet上几乎处处可以看到它的身影。应用JSP技术，无需复杂的编程，就可以开发出专业的动态WEB站点。JSP的出现，给你静态为主的WEB带来了全新的动态效果，使站点具有更加灵活和方便的交互性，在Internet中实现信息的传递和检索越来越容易。JSP构建的网站有以下的特点：

1. 使用JavaScript和VBScript等简易的Script语言，结合HTML代码，就可以快速的完成网站的应用程序。
2. 不需编译和链接就可以直接执行，并整合于HTML当中。
3. 使用一般的文本编辑程序，如记事本就可以编辑设计。
4. 与浏览器无关。客户只要使用执行HTML 代码的浏览器就可以浏览JSP的网页内容。不存在浏览器兼容的问题，由于JSP是在服务器端运行的，当客户端浏览器浏览JSP网页时，服务器会将该网页文件重新解释一遍，并将生成的标准的HTML文件发送给客户端浏览器，因为送出的是标准的HTML文件，当然不会存在浏览器的兼容问题了。
5. JSP与任何JAVAX Script语言完全兼容。除了可使用JavaScript和VBScript语言来设计外，可通过任何Plug-in 方式，使用厂商提供的其它脚本语言。
6. JSP的源程序不会传到用户的浏览器上，因而可以保护源程序不会外漏。
7. 采用了面向对象技术。
8. 可通过JAVAX服务器组件来扩充增加功能。。
   1. **JAVA技术**

Java技术是由美国Sun公司倡导和推出的，是一种简单易用、完全面向对象、具有平台无关性且安全可靠的主要面向Internet的开发工具。随着Java Servlet的出现和JSP的渐渐成熟，让Java正式的走向了Web开发技术的舞台。Java以其强大的移植能力，多线程处理和网络处理能力已成为研究人员，开发人员瞩目的焦点，它与Web技术的有机结合尤其适宜进行网络计算和动态多媒体信息的开发处理。Java技术带来的是一场革命，它是第一个真正独立于平台的语言，基于Java语言开发的软件可以实现“一次设计，到处运行”，即Java程序不经改变即可部署到任何运行Java虚拟机的计算机结构和操作系统，使得整个Web世界发生了翻天覆地的变化。

* 1. **MySQL数据库**

常用的网络数据库管理系统有SQL Server、Oracle、IBMDB2、MySQL、Sybase及Access等。其中MySQL由瑞典的T.c.X公司负责开发和维护，是一个快速、多线程、多用户和强壮的SQL数据库服务器。SQL是世界上最流行的标准化的数据库语言。MySQL是具有客户机/服务器体系结构的分布式数据库管理系统，是完全网络化的跨平台关系型数据库系统，是目前运行速度较快的SQL语言数据库。MySQL具有功能强，使用简便、管理方便、运行速度快、安全可靠性强等优点。

* 1. **Ajax技术**

信息化的迅猛发展使人们生活方式发生深刻变革。人们之间的对话不再局限在面对面的方式，网络聊天室提供交流信息的另一种场所。传统的聊天室为了实时显示用户在线情况和用户发送的消息，需要定时刷新页面。页面刷新不仅要占用一定的系统资源和网络带宽，而且出现“白屏”现象，对良好的用户体验很不利。Ajax是Web应用的一种新方法，利用Ajax技术开发的聊天室系统不仅解决了页面刷新带来的白屏问题，而且能降低服务器和网络传输负荷。

AJAX不是一种新的编程语言，而是一种用于创建更好更快以及交互性更强的Web应用程序的技术。使用Javascript向服务器提出请求并处理响应而不阻塞用户！核心对象为XMLHTTPRequest。通过这个对象，您的 JavaScript 可在不重载页面的情况与Web服务器交换数据。

AJAX 在浏览器与 Web 服务器之间使用异步数据传输（HTTP 请求），这样就可使网页从服务器请求少量的信息，而不是整个页面。从而使因特网应用程序更小、更快，更友好。

* 1. **Tomcat应用服务器**

目前支持JSP的应用服务器是较多的，Tomcat是其中较为流行的一个Web服务器，被Java World杂志的编辑选为2001年度最具创新的java产品，可见其在业界的地位。

Tomcat是一个免费的开源的Serlvet容器，在Tomcat中，应用程序的部署很简单，你只需将你的WAR放到Tomcat的webapp目录下，Tomcat会自动检测到这个文件，并将其解压。你在浏览器中访问这个应用的JSP时，通常第一次会很慢，因为Tomcat要将JSP转化为Servlet文件，然后编译。编译以后，访问将会很快。另外Tomcat也提供了一个应用：manager，访问这个应用需要用户名和密码，用户名和密码存储在一个xml文件中。通过这个应用，辅助于Ftp，你可以在远程通过Web部署和撤销应用。当然本地也可以。

     Tomcat也具有传统的Web服务器的功能：处理Html页面。但是与Apache相比，它的处理静态Html的能力就不如Apache。我们可以将Tomcat和Apache集成到一块，让Apache处理静态Html，而Tomcat处理JSP和Servlet。这种集成只需要修改一下Apache和Tomcat的配置文件即可。

Tomcat是一个很好的工具，不仅仅因为其免费，功能强大，更因为其开放性，越来越收到人们的重视。

* 1. **JQuery**

[Jquery](http://baike.baidu.com/view/1020297.htm)是继[prototype](http://baike.baidu.com/view/1217697.htm)之后又一个优秀的[Javascript](http://baike.baidu.com/view/16168.htm)库。它是轻量级的JS库 ，它兼容[CSS3](http://baike.baidu.com/view/1713027.htm)，还兼容各种[浏览器](http://baike.baidu.com/view/7718.htm)（IE 6.0+, FF 1.5+, Safari 2.0+, Opera 9.0+），jQuery2.0及后续版本将不再支持IE6/7/8浏览器。jQuery使用户能更方便地处理[HTML](http://baike.baidu.com/view/692.htm)（[标准通用标记语言](http://baike.baidu.com/view/5286041.htm)下的一个应用）、events、实现动画效果，并且方便地为网站提供AJAX交互。jQuery还有一个比较大的优势是，它的文档说明很全，而且各种应用也说得很详细，同时还有许多成熟的[插件](http://baike.baidu.com/view/18979.htm)可供选择。jQuery能够使用户的html页面保持代码和html内容分离，也就是说，不用再在html里面插入一堆JS来调用命令了，只需要定义id即可。

本系统的截图功能就是采用JQuery的Jcrop插件截图工具来完成的。

1. **总结**

当今社会基于C/S结构的程序客户端日益庞大，而互联网在迅猛发展，带宽越来越大，网速也越来越快。

而B/S结构统一了客户端，将系统功能实现的核心部分集中到服务器上，简化了系统的开发、维护和使用。因为这些，基于B/S结构的程序正日益成为主流。用户可以在不用客户端的情况下，只通过一个浏览器就可以完成大部分想要完成的工作。本系统也紧跟时代潮流，采用了JSP技术来实现B/S结构。旨在完成一个人性化的，简洁的，易于操作的在线聊天系统。

1. **参考文献**

[1]（美）Bruce Eckel著 陈昊鹏 译，Java编程思想，机械工业出版社，2007

[2]武延军,黄飞跃.精通JSP编程技术.人民邮电出版社,2001.

[3]赛奎春.JSP工程应用与项目实践.机械工业出版社，2005.

[4]贾素玲，王强.JSP应用开发技术.清华大学出版社, 2008.

[5]王珊，萨师煊.数据库系统概论.高等教育出版社，2006.

[6] 汤一平，Java语言程序设计，科学出版社，2006

[7] 孙卫琴，java面向对象编程[M]，电子工业出版社， 2006

[8] （美）Cay S. Horstmann / Gary Cornell著 陈昊鹏 / 王浩 / 姚建平【译】

JAVA核心技术卷2:高级特征，机械工业出版社，2008