单位代码 10006

学 号 10231007

分类号 TP311

****

毕业设计(论文)

基于结构和内容的Web页面分析

|  |  |
| --- | --- |
| 院（系）名称 | 计算机学院 |
| 专业名称 | 计算机科学与技术 |
| 学生姓名 | 张凯 |
| 指导教师 | 李舟军 |

2013年6月

基于结构和内容的Web页面分析 张凯 北京航空航天大学

北京航空航天大学

**本科毕业设计（论文）任务书**

Ⅰ、毕业设计（论文）题目：

基于结构和内容的Web页面分析

Ⅱ、毕业设计（论文）使用的原始资料（数据）及设计技术要求：

使用原始数据：

新浪、搜狐、百度、人民网等网站网页

设计技术要求：

1、基于.Net 4.0/4.5平台，开发环境Visual Studio 2010/2012

2、编码格式规范，代码可读性好，有适当的注释

Ⅲ、毕业设计（论文）工作内容：

对Web信息抽取技术进行深入研究，并在此基础上完成如下工作：

1.设计一种简单有效的页面分块算法，能完成对多种类型页面的高效分割。

2.在页面分块的基础上，设计页面块类型的识别算法实现对导航、正文、评论等类型的识别。

Ⅳ、主要参考资料：

[1]《Microsoft .NET框架程序设计》 清华大学出版社

[2]《C#高级编程(第7版)》 清华大学出版社

[3]《信息检索》 清华大学出版社

[4]《智能Web算法》电子工业出版社

[5]《Web数据挖掘》清华大学出版社

计算机 学院（系） 计算机科学与技术 专业类 390615 班

学生 张凯

毕业设计（论文）时间： 2013 年 3 月 \* 日至 2013 年 6月 \* 日

答辩时间： 2013 年 6 月 \* 日

成 绩：

指导教师： 李舟军

兼职教师或答疑教师（并指出所负责部分）：

系（教研室） 主任（签字）：

本人声明

我声明，本论文及其研究工作是由本人在导师指导下独立完成的，在完成论文时所利用的一切资料均已在参考文献中列出。

作者：黄绍晗

签字：

时间：2013年6月

基于结构和内容的Web页面分析

学 生：张 凯

指导教师：李舟军

摘 要

随着Internet及其技术的迅速发展，Web已经成一个拥有海量资源的信息库，给人们的日常生活带来了极大的便利。

。

关键词： Web信息抽取，Web页面分块，多特征分析，页面类型识别

Based on

Author: ZHANG kai

　　 Tutor: LI Zhoujun

Abstract

With the rapid development of Internet technology, Web has become a repository of information,

**Key words**：Web information extraction, page segmentation algorithm, multi-feature analysis, page type identification

**目 录**

[1 绪论 1](#_Toc387788761)

[1.1 课题背景及目的 1](#_Toc387788762)

[1.2 国内外研究状况 2](#_Toc387788763)

[1.3 课题研究内容 3](#_Toc387788764)

[1.4 论文构成及研究内容 4](#_Toc387788765)

[2 首页登录与权限控制的设计与实现 5](#_Toc387788766)

[2.1 基本技术的介绍 5](#_Toc387788767)

[2.1.1 SSL/HTTPS 5](#_Toc387788768)

[2.2 首页登录的设计与实现 5](#_Toc387788769)

[2.3 权限控制的设计与实现 5](#_Toc387788770)

[3 用户管理模板的设计与实现 5](#_Toc387788771)

[3.1 基本技术的介绍 5](#_Toc387788772)

[3.1.1 Bootstrap 5](#_Toc387788773)

[3.1.2 JQuery 5](#_Toc387788774)

[3.1.3 HttpServlet 5](#_Toc387788775)

[3.1.4 Restful 5](#_Toc387788776)

[3.2 用户管理模块的设计 5](#_Toc387788777)

[3.2.1 用户管理模块的前端设计 5](#_Toc387788778)

[3.2.2 用户管理模块的后台逻辑设计 5](#_Toc387788779)

[3.3 用户管理模块的实现 5](#_Toc387788780)

[4 环境运行模板的设计与实现 6](#_Toc387788781)

[4.1 基本技术的介绍 6](#_Toc387788782)

[4.1.1 CGI 6](#_Toc387788783)

[4.2 环境运行模块的设计 6](#_Toc387788784)

[4.2.1 环境运行模块的前端设计 6](#_Toc387788785)

[4.2.2 环境运行模块的后台逻辑设计 6](#_Toc387788786)

[4.1 环境运行模块的实现 6](#_Toc387788787)

[结论 6](#_Toc387788788)

[致谢 7](#_Toc387788789)

[参考文献 9](#_Toc387788790)

# 绪论

## 课题背景及目的

高性能计算能力是国家综合实力和创新能力的重要体现。随着我国高性能计算机研制水平和高性能计算中心规模的不断发展，建立国家级高性能计算基础设施并投入良性运行已成为迫切需要解决的问题。国家级高性能计算基础设施的建立有利于充分有效的组织国内优秀的计算资源，提升资源整体利用情况，避免了不必要的浪费；同时有利于为科研人员提供优质、不间断的计算服务，提高科研人员的科研效率，加速其成果产生。

本课题为国家高性能计算环境运行支持平台前端的设计与实现，运行支持平台将对整个高性能计算环境的计算资源、基础设施服务、用户进行监控和管理提供帮助与服务，是高性能计算环境在实际生产性运维期间的环境监控和管理工具，提供对环境运维决策制定的数据支持。本课题作为环境运行支持平台前端的设计与实现，主要是使用Nagios/CGI和SCE/RESTful接口作为数据调用接口，设计实现用户界面的展示和web应用服务。

运行支持平台的目的是主要研究资源监控和运行管理技术、统一认证与系统安全技术，构建环境的运行支持平台，为国家高性能计算基础服务的稳定运行和长期发展提供支持，最大程度的降低运维成本。而本课题作为运行支持平台的前端，可以完成更好的运行支持平台与用户的交互，更加完善、更加人性化的数据展示和数据可视化。

平台用户界面

平台Web应用服务

SCE/RESTful接口

Nagios/CGI接口

## 国内外研究状况

中国国家高性能计算基础设施的建立要追溯到20世纪九十年代后期。1998年，国家863计划“智能计算机系统”主题将研究重点从研制单台高性能计算机转向从技术上支持我国高性能计算基础设施的构建。这个项目支持建立了由5个高性能计算中心构成的国家高性能计算环境，形成了我国高性能计算环境的雏形。

2002-2005年，863计划实施了“高性能计算机及其核心软件”重大专项，在网格环境、网格软件和网格应用等方面完成了一批重要的研究课题；基于自主研制的网格系统软件，将分布的8个高性能计算中心连接起来，向全国的科学研究用户和行业用户提供开放共享的高性能计算服务，为我国的科学研究和信息化建设提供了新型环境和平台。

2006-2011年，国家863计划实施了“高效能计算机及网格服务环境”重大项目，其战略目标是研制千万亿次高效能计算机系统，构建中国国家网格服务环境。在这个重大项目支持下，建成了14个结点组成的国家级高性能计算服务环境，聚合计算能力超过3000万亿次以上，存储能力超过15PB，部署了450多个软件与服务，支持了1100多项国家与地方科技项目。

经过十余年的不懈努力，国家高性能计算环境积累了丰富的基础软件建设经验和环境运行管理经验，在实践中发挥了重要的作用，但仍存在着亟需解决的问题，比如环境建设缺乏规划、实际应用推广还有待普及、既懂计算又具有应用领域知识的跨学科人才严重不足等等。

美国是世界上最重视高性能计算机的国家，同时美国也是在高性能计算基础环境上投入最多的国家。在美国有很多高性能计算的重大计划和项目，比如：美国能源部的加速战略计算ASCI计划，ASCI计划的目标是构造具有100万亿次计算能力的超级计算机；白宫直属的HECC（High-End Computing and Computations）计划，对高性能计算的关键技术进行研发，并构建高性能基础设施；Petaflops计划开发构造千万亿次级系统的技术。

欧洲的强项则主要体现在高性能计算机的应用方面。欧盟投入巨资建设欧洲数据网格Europena DataGrid(EDG)，面向卫星观测和数字地球、生物和气象、高能物理等应用联合开展研究并构建计算和数据网格。欧盟还资助了EuroGrid的研究和技术开发计划，目标是使用高性能计算为科学和工业群体提供一个广域分布的信息处理环境。

## 课题研究内容

利用已有的工具，给定一个实体，把与他相同类别实体的属性当做他的属性，并将这些属性放到已有工具中进行值的抽取。在抽取的过程中，如果信息抽取算法返回的可信度得分高，则将此属性-值对放入infobox。

经过上述三个部分之后，正文部分可以成功生成。

## 论文构成

本文的内容组织如下：

第一章 绪论，介绍了课题背景及目的，概述了运行支持平台对于高性能计算环境的重要性，同时简单介绍了国内外对于高性能计算环境基础的研究，最后介绍了本课题的研究内容。

第二章 相关技术，对在运行支持平台前端设计和实现中涉及到的相关技术进行了介绍。

第三章 运行支持平台前端的整体结构，对运行支持平台前端的整体结构进行了分析与介绍。

第四章 首页登录与权限控制。介绍了首页登录的设计和权限控制的具体算法和实现思路。

第五章 用户管理模块。

第六章 环境运行模块。

第七章 结论语。对本文工作进行了总结，并讨论了之后的工作目标。

# 相关技术

在这一章节中主要介绍了本课题运行支持平台在实现过程中使用到的一些关键技术，其中有bootstrap、jQuery、RESTful和SSL/HTTPS等，对这些技术的背景、特点以及在课题中使用情况进行了简单的介绍。

## Bootstrap

Bootstrap是由著名的社交网站Twitter在2011年8月推出的开源WEB前端框架，用于前端开发的工具包。它由Twitter公司的设计师Mark Otto和Jacob Thornton合作开发，集合CSS和HTML，使用了最新的浏览器技术，为快速WEB开发提供了一套前端工具包，包括布局、网格、表格、按钮、表单、导航、提示等等。使用Bootstrap可以构建出非常优雅的前端界面，而且占用资源非常小。

Bootstrap是基于jQuery框架开发的，它在jQuery框架的基础上进行了更为个性化和人性化的完善，形成一套自己独有的网站风格，并兼容大部分jQuery插件。Bootstrap的主要特点是：

* 丰富的组件，其中下拉菜单、按钮组、按钮下拉菜单、导航、导航条、面包屑、分页、排版、缩略图、警告对话框、进度条、媒体对象。
* 人性化的Javascript插件，Bootstrap自带了13个jQuery插件，包括模式对话框、标签页、滚动条、弹出框。
* 可制定的框架代码，可以对Bootstrap中所有的CSS变量进行修改，依据自己的需求裁剪代码。

在本课题中，运行在浏览器中的html文件使用的就是Bootstrap2.0，Bootstrap2.0把现有框架进行了清晰的功能划分，主要分为框架（Scaffolding），基础CSS，构件库和jQuery插件库。在平台前端开发过程中框架（Scaffolding）、基础CSS，构件库和jQuery插件库都有所使用。

Scaffolding主要提供基于网格的各种布局，包括普通网格系统、嵌入式网格，固定布局、自适应布局，同时可以对网格和布局进行自定义。Bootstrap2.0提供了响应式设计，可以通过单个文件支持各种手持设备，自适应不同的设备和屏幕变化。

基础CSS包括各种排版样式（标题、段落、引用块、列表、内联标签等），代码展示方面提供了基于code标签的内嵌代码，基于pre的块代码和基于Google Prettify的代码样式。同时提供各种表格、表单、按钮、图标的展示方式。

构件库提供了基于按钮、导航、标签、排版、警告、进度栏、图像网格等控件。

jQuery插件库则提供了十几种插件实现动态效果，例如Modal、Dropdown、Tab、Tooltip、Popover、Carousel等，开发者可以根据自己的业务需求使用不同的插件实现各种动态效果。

## JQuery

jQuery是一个优秀的Javascript框架。它是轻量级的js库，jQuery 是一个优秀的 JavaScript 框架，它能使用户更方便地处理 HTML 文档、events 事件、动画效果和 Ajax 交互。

本课题中静态页面和CSS效果通过使用bootstrap框架完成，而前台的绘制、事件与服务器端的数据交互都是通过jQuery完成。在本课题中主要使用了jQuery以下的几个功能：

1. 访问页面框架的局部。这是 DOM 模型所完成的主要工作之一，DOM获取页面中某个节点或者某一类节点有固定的方法，而jQuery则大大地简化了其操作的步骤。 jQuery的元素选择器和属性选择器通过标签名、属性名或内容对 HTML 元素进行选择，要比Javascript简单得多。这也是使用jQuery最主要的原因。
2. 修改页面的表现。CSS 可以通过样式风格来修改页面的表现。然而由于各个浏览器对 CSS3 标准的支持程度不同，使得很多CSS 的特性没能很好地体现。

可以通过jQuery获取到元素的位置，对元素的CSS属性进行修改，完成对页面的表现工作。

1. 更改页面的内容。同修改页面的表现原理一致，jQuery获取到元素的位置，可以对元素的属性进行修改，同样可以任意添加、删除某些元素。通过强大而方便的 API括文本的内容、插入新的图片、表单的选项，甚至整个页面的框架。
2. 响应事件。JavaScript 有处理事件的相关方法，而引入 jQuery 之后，可以更加轻松地处理事件，而且开发人员不再需要考虑讨厌的浏览器兼容性问题。 在本课题中的用户管理模块，有许多按钮需要添加点击事件，通过jQuery的on方法，监听click事件，可以简化处理事件的相关方法，提高开发效率。
3. 与服务器异步交互。Ajax框架可以简化代码的编写，jQuery也提供了一整套Ajax 相关的操作，大大方便了异步交互的开发和使用。jQuery与服务器异步交互功能同上面的功能一样重要，在平台前端实现过程中，需要与后台进行大量的数据交互，Ajax框架简化了代码的编写，而jQuery使得后台的操作简单化。

## RESTful

REST (REpresentation State Transfer) 描述了一个架构样式的网络系统，比如 web 应用程序。它首次出现在 2000 年 Roy Fielding 的博士论文中，他是 HTTP 规范的主要编写者之一。REST 指的是一组架构约束条件和原则。满足这些约束条件和原则的应用程序或设计就是 RESTful。

Web 应用程序最重要的 REST 原则是，客户端和服务器之间的交互在请求之间是无状态的。从客户端到服务器的每个请求都必须包含理解请求所必需的信息。如果服务器在请求之间的任何时间点重启，客户端不会得到通知。此外，无状态请求可以由任何可用服务器回答，这十分适合云计算之类的环境。客户端可以缓存数据以改进性能。

在服务器端，应用程序状态和功能可以分为各种资源。资源是一个有趣的概念实体，它向客户端公开。资源的例子有：应用程序对象、数据库记录、算法等等。每个资源都使用 URI (Universal Resource Identifier) 得到一个惟一的地址。所有资源都共享统一的界面，以便在客户端和服务器之间传输状态。使用的是标准的 HTTP 方法，比如 GET、PUT、POST 和 DELETE。Hypermedia 是应用程序状态的引擎，资源表示通过超链接互联。

另一个重要的 REST 原则是分层系统，这表示组件无法了解它与之交互的中间层以外的组件。通过将系统知识限制在单个层，可以限制整个系统的复杂性，促进了底层的独立性。

当 REST 架构的约束条件作为一个整体应用时，将生成一个可以扩展到大量客户端的应用程序。它还降低了客户端和服务器之间的交互延迟。统一界面简化了整个系统架构，改进了子系统之间交互的可见性。REST 简化了客户端和服务器的实现。

## SSL/HTTPS

HTTPS（Secure Hypertext Transfer Protocol）安全超文本传输协议

　　它是由Netscape开发并内置于其浏览器中，用于对数据进行压缩和解压操作，并返回网络上传送回的结果。HTTPS实际上应用了Netscape 的完全套接字层（SSL）作为HTTP应用层的子层。（HTTPS使用端口443，而不是象HTTP那样使用端口80来和TCP/IP进行通信。）SSL 使用40 位关键字作为RC4流加密算法，这对于商业信息的加密是合适的。HTTPS和SSL支持使用X.509数字认证，如果需要的话用户可以确认发送者是谁。。

　　https是以安全为目标的HTTP通道，简单讲是HTTP的安全版。即HTTP下加入SSL层，https的安全基础是SSL，因此加密的详细内容请看SSL。

　　它是一个URI scheme(抽象标识符体系)，句法类同http:体系。用于安全的HTTP数据传输。https:URL表明它使用了HTTP，但HTTPS存在不同 于HTTP的默认端口及一个加密/身份验证层（在HTTP与TCP之间）。这个系统的最初研发由网景公司进行，提供了身份验证与加密通讯方法，现在它被广 泛用于万维网上安全敏感的通讯，例如交易支付方面。

HTTPS是以安全为目标的HTTP通道。HTTP下加入SSL层，HTTPS的安全基础是SSL，提供了身份验证与加密通讯方法。

在平台首页用户登录模块中，会用到HTTPS（全称：Hypertext Transfer Protocol over Secure Socket Layer）。用户登录需要提交用户名和密码，如果使用HTTP连接，使用明文传输，将十分不安全，因此采用HTTPS协议，由SSL+HTTP协议构建的可进行加密传输。

# 运行支持平台前端的整体结构

# 首页登录与权限控制

## 相关技术的介绍

## 首页登录的设计与实现

## 权限控制的设计与实现

# 用户管理模块

## 相关技术的介绍

## 用户管理模块的设计

### 用户管理模块的前端设计

### 用户管理模块的后台逻辑设计

## 用户管理模块的实现

# 环境运行模块

## 相关技术的介绍

## 环境运行模块的设计

### 环境运行模块的前端设计

### 环境运行模块的后台逻辑设计

## 环境运行模块的实现

结论

致谢

在本论文即将完成之际，我想把心里最诚挚的感谢送给所有关心我帮助我的人们。

首先， 我要感谢我毕业设计的导师李舟军教授。李老师是一位学识渊博、治学严谨、和蔼可亲的老师。在教学方面，李老师总能用最朴实的语言把高深的理论知识讲的惟妙惟肖、深入浅出，李老师的课让我受益匪浅。他为本科生开设的课程《离散数学2》在同学心中评价很高。在实验室，我深深地感到这里有一种浓厚的科研氛围，每个师兄、师姐都很热情、很友善，对我的帮助很大。李老师倡导定期开展小组讨论，而且每次都亲自主持讨论，提出了很多宝贵的意见，指引我们朝着正确的道路不断探索、研究。在我做毕业设计期间，李老师也给予了我很多帮助和指导，并且时刻提醒我要珍惜时间，按计划完成任务。在李老师的督促和指引下，我顺利完成了毕业设计。在此再次感谢李老师。

同样也要感谢张小明老师和陈铮师兄。张老师和陈师兄在我做毕设的整个过程中，从选题、系统设计与开发、论文完成，他们都给了我很大的帮助。在张老师的细心讲解和指导下，我们积累了很多和毕业设计相关知识。陈师兄也总是耐心地与我们讨论，使我们能够在很多关键技术和难点的地方得到突破。没有张老师和陈师兄的帮助，就很有可能没有这篇论文的顺利完成，在此向他们表示最真挚的谢意。毕业设计是用C#语言编写的，由于之前对这门语言不是非常熟悉，因此刚开始编写时有些困难。在刘康师兄的细心点播和讲解下，程序的编写按照原先的计划一步一步顺利开展。有些代码，我调试了好久都没有找到问题的所在，正当我感到迷茫的时候，刘悦师兄帮助了我，为我找到了程序的问题，并指出了问题可能的解决方案，让我去尝试。在他们的帮助下，我的毕业设计顺利地进行，编码能力和解决问题的能力得到了很好的锻炼，收获颇丰。

在实验室对着电脑奋战毕业设计的日子总是让人难以忘怀，老师们一丝不苟的工作作风是我学习的榜样，同学们的珍贵友情是人生最大的财富。在我做毕业设计的过程中，他们也多次提出宝贵的意见，有些提醒总能令我茅塞顿开。在此，送上我对他们的感谢。

感谢计算机学院的各位老师，是他们辛勤的付出，才换来了我们文化知识的不断提高。感谢任远、柴巧珍两位辅导员，无论是在学习还是在生活上，你们都给了我很多帮助，你们默默的付出，让我们感到3906这个大班的温暖。四年中，每当我们遇到困难，你们总是能够细心地察觉，组织开展大班会，给我们鼓励和勇气，指引我们不断向前。感觉3906大班的所有同学，缘分让我们相聚，四年的朝夕相处让我们相知，我会好好珍惜这份来之不易的友谊。特别感谢390615小班，我们是一个团结、融洽的集体，我很荣幸能够在这个集体中生活四年。还有809和406寝室的兄弟们，是你们让我的四年变得异常精彩。

在这里，我还要把最真诚的感谢献给我的父母。从小到大，父母对于我的关心总是无微不至的。上大学是第一次离开父母，独自一人在外漂泊，他们不仅在生活上给了我莫大的关怀，在精神上也给予了很多支持和鼓励。每当我生活上或者学习上遇到问题时，我总是会找父母咨询，每次父母都是耐心地听我述说，告诉我该如何解决。正是带着他们的期盼和鼓励，我不断地提高自己。

最后，衷心感谢为评阅本文而付出辛勤劳动的各位老师和评委！

参考文献

[1] Freitag, Dayne. "Toward general-purpose learning for information extraction."*Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1998.

[2] Knoblock, Craig A., et al. "Accurately and reliably extracting data from the web: A machine learning approach." *Intelligent exploration of the web*. Physica-Verlag HD, 2003. 275-287.

[3] Kushmerick, Nicholas. *Wrapper induction for information extraction*. Diss. University of Washington, 1997.

[4] Banko, Michele, et al. "Open information extraction for the web." *IJCAI*. Vol. 7. 2007.

[5] Wu, Fei, and Daniel S. Weld. "Open information extraction using Wikipedia."*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

[6] Li, Peng, Jing Jiang, and Yinglin Wang. "Generating templates of entity summaries with an entity-aspect model and pattern mining." *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

[7] Zhou, Liang, Miruna Ticrea, and Eduard H. Hovy. "Multi-Document Biography Summarization." *EMNLP*. 2004.

[8] Biadsy, Fadi, et al. "An Unsupervised Approach to Biography Production Using Wikipedia." *ACL*. 2008.

[9] Sauper, Christina, and Regina Barzilay. "Automatically generating wikipedia articles: A structure-aware approach." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

[10] Yao, Conglei, et al. "Autopedia: automatic domain-independent Wikipedia article generation." *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011.

[11] Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR '98, pages 335{336, New York, NY, USA. [12, 14, 15]

[12]Radev, D. R., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management 40 (2004), 40:919{938. [16, 17]

[13] Erkan, Günes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." J. Artif. Intell. Res.(JAIR) 22.1 (2004): 457-479.

[14] Lin, Chin-Yew, and Eduard Hovy. "Automatic evaluation of summaries using n-gram co-occurrence statistics." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

[15]Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[16] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.

[17] Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." *Advances in neural information processing systems* 2 (2002): 841-848.

[18] Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.