

天津外国语大学国际商学院

本科生毕业论文

|  |  |
| --- | --- |
| **课程名称：** | 网络安全与管理 |
| **论文题目：** | 毕业设计 |
| **姓 名：** | 尚思羽 |
| **学 号：** | 1307114019 |
| **专 业：** | 信息管理与信息系统 |
| **年 级：** | 13级 |
| **班 级：** | 13708 |
| **任课教师：** | 王斌 |
|  |  |
| **2017 年 02 月** | |

内容摘要

随着网络日益发达，网络搜索功能已经渐渐融入我们的生活，成为不可缺少的一部分。一方面，我们可以利用搜索功能增长自己的知识范围，另一方面也可以解决生活和学习、工作中遇见的问题。目前，国内外搜索引擎数量庞大、种类繁多，所提供的功能与采用的技术也各不相同，本文通过对搜索引擎的分类、原理、功能、国内外发展趋势等方面的介绍，结合所学知识，根据搜索引擎的搜索功能的工作原理，开发了一个小型的demo来帮助理解搜索引擎的工作过程，最后总结开发过程中遇到的问题和自己的收获。

**关键词：** 搜索引擎；爬虫；nodejs

目 录

[一.绪论 1](#_Toc476337017)

[（一） 搜索引擎起源 1](#_Toc476337018)

[（二） 搜索引擎分类及工作原理 1](#_Toc476337019)

[1.分类 1](#_Toc476337020)

[2.工作原理 2](#_Toc476337021)

[二、需求分析 3](#_Toc476337022)

[（一）项目背景 3](#_Toc476337023)

[（二）开发目标 3](#_Toc476337024)

[（三）开发工具及使用技术 3](#_Toc476337025)

[（四）开发工具介绍 3](#_Toc476337026)

[1.代码编辑工具 3](#_Toc476337027)

[2.版本控制工具 5](#_Toc476337028)

[三、搜索引擎实现相关技术 6](#_Toc476337029)

[（一）Web超链分析算法 6](#_Toc476337030)

[1.PageRank算法 6](#_Toc476337031)

[2.HITS算法 7](#_Toc476337032)

[四、系统分析 8](#_Toc476337033)

[（一）可行性分析 8](#_Toc476337034)

[（二）业务流程分析 9](#_Toc476337035)

[（三）数据流程分析 9](#_Toc476337036)

[五、简易搜索引擎的设计与实现 9](#_Toc476337037)

[（一）模块设计 10](#_Toc476337038)

[1.模块设计的原则 10](#_Toc476337039)

[2.爬虫模块的设计 10](#_Toc476337040)

[3.结论 10](#_Toc476337041)

IDS与IPS的分析与选择

一.绪论

（一） 搜索引擎起源

在当今的网络环境下，说起搜索引擎几乎无人不晓，但在20多年前，却几乎没有人了解所谓“搜索引擎”的概念。搜索引擎的出现源于人们对信息获取的需求，最早的搜索引擎应该是1990 年由Montreal的McGill University 三名学生（Alan Emtage、Peter Deutsch、Bill Wheelan）发明的Archie，它是第一个自动索引互联网上匿名FTP网站的程序，它通过定期搜集并分析系统中存在的文件名信息,提供查找分布在各个主机中文件的服务，用户能在只知道文件名的前提下,为用户找到这个文件所在的服务器的地址。实际上是一个大型的数据库,再加上与这个大型数据库相关联的一套检索方法。该数据库中包括大量可通过下载的文件资源的有关信息,包括这些资源的文件名、文件长度、存放该文件的计算机名及目录名等。尽管所提供服务的信息资源对象和本文所讨论搜索引擎的信息资源对象网页不一样,但基本工作方式是相同的[1]，被公为是所有搜索引擎的祖先。

（二） 搜索引擎分类及工作原理

1.分类

搜索引擎的分类并没有明确的界限，有关搜索引擎的分类方式也有许多的原则和标准。原则如：发展原则、标准一致原则、科学分类原则、优先原则等，标准如：按检索语言分、按检索功能分、按检索范围分、按组合方式分等[2]。目前比较有代表性的分类方式是：全文搜索引擎、分类目录、元搜索引擎、垂直搜索等。全文搜索引擎如：百度、谷歌等，这种搜索引擎搜索范围较广、采用“机器人”或“蜘蛛”进行广泛的搜索，将搜索的数据收入自己的索引数据库中，每隔一段时间会更新自己的网页数据库，以备用户查询；分类目录如：Yahoo、新浪等，这种搜索引擎会与全文搜索引擎相比，有许多不同之处，全文搜索会自动搜索网页，而分类目录则会加入大量的人工操作的部分，想要被收录进分类目录搜索引擎的网站主需要把自己的网站信息提交给目录编辑人员，由目录编辑人员决定是否要将你的网站收录，其次，普通的全文搜索引擎只要在不违反有关规则的情况下都能轻松登录，而在分类目录对网站要求很高，有时很难登录成功。如果目录编辑人员认为网站提交的目录不合适，会自主改变你的网站信息；元搜索引擎如：InfoSpace、Dogpile、Vivisimo等，指的是接受用户输入的查询请求后，在多个搜索引擎上搜索，并返回结果给用户的搜索引擎；而垂直搜索引擎则属于专业性较强的搜索引擎，是对网页库中的某类专门的信息进行一次整合,定向分字段抽取出需要的数据进行处理后再以某种形式返回给用户。

2.工作原理

搜索引擎的工作主要分为三步，即爬虫、存储和检索。

爬虫是搜索引擎的重要组成部分，主要依靠一个叫“网络机器人”（也叫“网络蜘蛛”）的软件，对互联网上的网页进行爬取，一般情况下，爬虫会根据一批初始URL列表开始搜索，解析出这些网页中包含的URL，然后进一步爬取新搜索到的URL，一步一步同时将爬取到的网页按照一定的规则存储到索引数据库中。由于Web上的网页数量巨大，并且始终呈量级增长，对URL爬取需要遵循一定的策略[3]。

1. IP地址搜索策略：这种方式的爬虫主要基于同一个站点的URL爬取，当爬虫工作的时候，会忽视来自其他站点的URL，而是专注于本站点搜索，直到下载完本站点的全部URL为止。这种搜索比较全面，但不适合大规模搜索，因此更多用于自主开发定制的搜索。
2. 深度优先搜索策略：这种方式是以深度搜索为优先，在一个 HTML 文件中，当一个超链被选择后，被链接的 HTML 文件将执行深度优先搜索，也就是说在搜索其余的超链结果之前必须先完整地搜索单独的一条链。深度优先搜索沿着 HTML 文件上的超链走到不能再深入为止，然后返回到某一个 HTML 文件，再继续选择该HTML 文件中的其他超链。当不再有其他超链可选择时，说明搜索已经结束。
3. 广度优先搜索策略：基于广度优先搜索策略编写的爬虫会将当前正在搜索的HTML文件中所有的超链接全部搜索完毕后，再处理下一层的链接，逐层递增，直到底层为止。比如：在一个HTML文件A上搜索，得到三个链接a,b,c，选择其中的一个链接a开始搜索，搜索完毕之后，并不继续探究深层，而是返回HTML页面A，开始搜索链接b，以此类推，当搜索完页面A的全部链接后，再开始处理搜索三个超链接a,b,c,得到的下一层页面。如此可以保证对浅层的优先处理，避免陷入深层的WWW文件中。

二、需求分析

（一）项目背景

虽然如今全文搜索引擎已经极大程度地融入了我们的生活，但仍有一些网站（如百度云网盘等）并没有提供搜索的功能，这种情况下就需要我们自己开发爬虫来爬取这些网站中的内容，又或者当我们需要获取某些基于个人要求的信息时，定制自己的搜索爬虫变得很有必要，因此我尝试编写一个基于自己需要的爬虫。通过观察身边接触的网络环境，我发现我校教务处经常会有很多通知，但由于同学们平时工作或生活忙碌，并不会及时地查看通知，由于通知更新较快，当需要某条通知再去教务处寻找的时候会比较麻烦且耗费时间，因此我决定写一个爬虫根据输入的关键词来爬取教务处的通知，以便及时获取有效信息。

（二）开发目标

本次开发的搜索引擎将模拟搜索引擎的工作流程，通过爬取数据，存储数据，检索数据的方式呈现开发结果，开发内容为：一个爬虫软件，一个利用具有中文分词的搜索功能和一个存储媒介存储爬取的数据，一个用户界面供用户搜索并显示搜索结果。

（三）开发工具及使用技术

开发过程中使用到的工具有：Sublime Text3、Git、Eclipse、WebStrom。

开发过程中使用到的技术有：HTML5、CSS3、javascript、jQuery、nodejs、json、java。

（四）开发工具介绍

1.代码编辑工具

一个好的开发工具对程序员这种工作量巨大的工作来说绝对是一种高效的辅助，合理地使用开发工具甚至能够间接地减少开发的工作量。对于前端开发者来说，开发工具有很多，但最后终究归结为两类：文本编辑器和IDE（集成开发环境）。

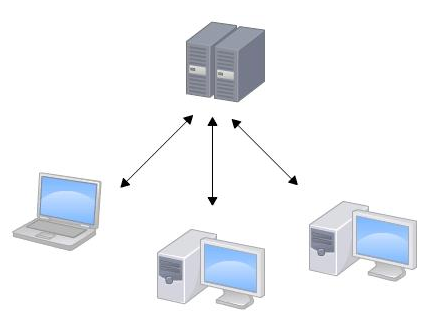
比较著名的文本编辑器有：sublime、notepad++等。相比于功能强大的各种IDE，文本编辑器的优点在于快速地打开及使用，加上现在市面上琳琅满目的开源插件，有些文本编辑器甚至可以与IDE媲美。回想前几年，notepad++无疑是一个优秀的开发编辑器，它是一个优于Windows的文本编辑器，完全免费且开源，对于不同的编程语言可以实现语法高亮，代码折叠以及宏，可定制性非常强，缺点是只能运行在Windows上。而Sublime Text3是ST的第三代产品，对于程序员来说绝对是一个编码神器，相较于ST2，ST3具有更加优美的界面，更加丰富的插件支持，其中最重要的插件就是Package Control，安装这个插件之后可以很方便地管理和一键安装其他插件。ST3是一款跨平台的软件，也就是不仅限于Windows系统，而是可以多系统使用。ST3的开启速度极快，拥有强大的代码补全、自动缩进等功能，其自身携带的代码地图更是一大亮点，类似于看图软件的图片地图，可以一键快速定位到想看的代码块，省去了麻烦的滚动时间。使用像Emmet这种自带Zen coding功能的插件，可以极大提高编写HTML和CSS效率，它可以接受能够被动态解析的语法。例如：想要写一个包含5个li的ul，可以写“ul>li\*5”，然后按下Tab键，Emmet会为你智能生成想要的DOM结构；又或者每次写一个新的页面之前，需要写出规范的HTML页面结构，使用文本编辑器不会智能地根据文件类型生成页面，这时候Emmet又要发挥作用了，写出html:x（x代表HTML版本号），再按下tab键，即可生成HTML页面结构，当所用版本是HTML5时，甚至可以只用一个叹号来生成页面，十分快捷。

在IDE方面，主要使用了Eclipse和WebStrom。IDE的特点是智能提示强大，但占用内存较高，开启也比较慢，对电脑性能要求略高。

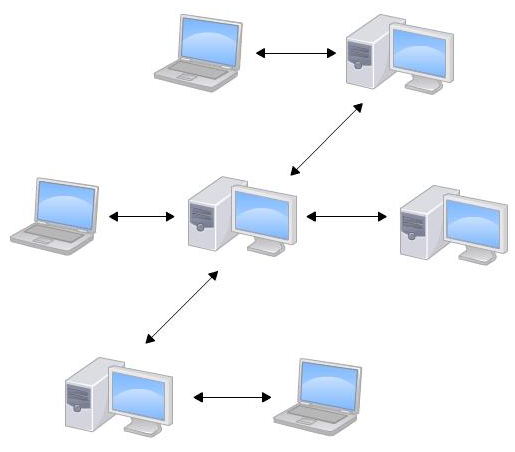
Eclipse的前身是IBM的Visual Age，是目前比较普遍的JAVA开发工具，相比较myEclipse，它的运行速度更快，并且完全开源免费有汉化包的支持，从编写、查错、编译、帮助等各方面为JAVA语言提供良好的支持。Eclipse也拥有许多强大的插件帮助代码编写，缺点是GUI界面设计有不完善、更新速度太快以至于插件更新速度难以跟上。另一个支持js编写的IDE就是JetBrains公司的WebStrom，WS专门为 JS 开发做了很多优化，例如图片显示宽高、内联CSS重构、深度自定义的格式化规则、自带版本控制等。

2.版本控制工具

目前业界比较有名的版本控制工具当属SVN和Git了，任何一个多人写作的公司都离不开版本控制功能，否则多人协作将无法进行。由于SVN的知名度较高，且出现较早，因此市场占有率更高，作为CVS的继任者[4],SVN应该是现在大多数公司使用的工具，他的分支策略是优于CVS的新亮点，但仍无法逃开集中式管理工具的弊病，即必须联网和一台中央服务器，在网速慢的情况下会严重限制上传效率，并且如果“中央服务器”瘫痪，则会导致所有人无法继续干活。



而Git作为分布式的管理控制工具，它的诞生可谓十分传奇，是由Linux的创始人Linus为了赌气花了两周时间自己用C写的，用来管理Linux开源社区的代码。Git继承了Linux的开源精神，不仅免费而且提供了Github网站提供代码托管，让我们不必搭建自己的Git服务器也可以使用Git。Git会在每一台工作的电脑上创建一个完整的版本库，通常也有一台充当“中央服务器”的电脑，但这个服务器的作用仅仅是用来方便“交换”大家的修改，没有它大家也一样干活。本次课题虽然没有多人协作，但仍然选择了github作为代码托管工具，原因是如果在不同的电脑上写了代码，使用git可以很方便的合并到一起，省去了文件传输的麻烦。



三、搜索引擎实现相关技术

（一）Web超链分析算法

1.PageRank算法

著名的搜索引擎Google就是使用了这种算法，Google最初是斯坦福大学的博士研究生Sergey Brin和Lawrence Page实现的一个原型系统，现在已经发展成为WWW上最好的搜索引擎之一。Google的体系结构类似于传统的搜索引擎，它与传统的搜索引擎最大的不同处在于对网页进行了基于权威值的排序处理，使最重要的网页出现在结果的最前面。Google通过PageRank元算法计算出网页的PageRank值，从而决定网页在结果集中的出现位置，PageRank值越高的网页，在结果中出现的位置越前，也就是越受欢迎。

PageRank算法的基本思想[5]是：

面很可能就是重要页面；一个页面尽管没有被多次引用，但被一个重要页面引用，那么这个页面很可能也是重要页面；一个页面的重要性被均分并将传递到它所引用的页面中。其中页面的重要程度量化后的分数就是 Page Rank 值。

计算网页的 Page Rank值 PR(A)的计算公式为:  


其中：

PR(A):页面 A 的网页级别

PR(T1)：页面 T1 的网页级别，页面 Ti 链向页面 A

C(T1)：页面 T1 链出的链接数量

d：阻尼系数，取值在 0－1 之间

由上面的公式可以看出，一个网页链入的数量越多，这些链入网页的Page Rank值越高，这些链入网页本身的链出数量越少，则该网页的Page Rank 值越高。和Alexa(一个统计网站排名的一个很重要的算法)算法不同，Page Rank 算法并不是将整个网站排名而是以单个页面计算的。其次，页面 A 的 Page Rank 的值取决于那些连接到A的Page Rank的递归。网页想提高自己的排名度可以根据这个公式提高PR值

也可以用矩阵来描述此算法，设A为一个方阵，行和列对应网页集的网页。如果网页i有指向网页j的一个链接，则 ，否则 ＝0。设V是对应网页集的一个向量，  
有V=cAV，V为A的特征根为c的特征向量。实际上，只需要求出最大特征根的特征向量，就是网页集对应的最终PageRank值，这可以用迭代方法计算。

2.HITS算法

HITS（Hyperlink－Induced Topic Search）算法是利用Hub/Authority方法的搜索方法，首先获取用户的搜索q，然后将q提交给传统的基于关键字匹配的搜索引擎。取得得到的满足如下条件的网页集S：

1．S中网页数量相对较小

2．S中网页大多数是与查询q相关的网页

3．S中网页包含较多的权威网页。

通过向S中加入被S引用的网页和引用S的网页将S扩展成一个更大的集合T．

以T中的Hub网页为顶点集Vl，以权威网页为顶点集V2，Vl中的网页到V2中的网页的超链接为边集E，形成一个二分有向图SG＝(V1，V2，E)。对V1中的任一个顶点v，用h(v)表示网页v的Hub值，对V2中的顶点u，用a(u)表示网页的Authority值。开始时h(v)＝a(u)＝1，对u执行I操作修改它的a(u)，对v执行O操作修改它的h(v)，然后规范化a（u），h（v），如此不断的重复计算下面的操作I，O，直到a（u），h（v）收敛。（证明此算法收敛可见 ）

I 操作： （1） O操作： （2）

每次迭代后需要对a(u),h(v)进行规范化处理：

式(1)反映了若一个网页由很多好的Hub指向，则其权威值会相应增加(即权威值增加为所有指向它的网页的现有Hub值之和)。式(2)反映了若一个网页指向许多好的权威页，则Hub值也会相应增加(即Hub值增加为该网页链接的所有网页的权威值之和)。

和PageRank算法一样，可以用矩阵形式来描述算法，这里省略不写。

HITS算法输出一组具有较大Hub值的网页和具有较大权威值的网页。

四、系统分析

（一）可行性分析

搜索引擎是一种比较复杂的系统，根据搜索量级的不同，会有不同的系统架构。目前，多种语言具有比较成熟可用的搜索引擎框架，大到基于java的Lucene、ElasticSearch等框架，小到nodejs的npm库中的插件和许多封装好的组件，都可以提供搜索功能。本次课题为了实现可用流程，根据预先制定的URL列表采用广度搜索优先策略，对制定的站点进行爬虫并爬取信息，存储到固定的媒介中以备用户检索取用。由于我本身具有javascript的基础，且应用的爬虫方式是使用nodejs，许多语法可复用，因此本课题的可行性较高。

（二）业务流程分析

本次项目的开发目的，是通过编写一个自动提取网页的程序，根据一个设定好的URL开始爬虫，该程序对界面的结构进行解析，通过解析后的结果，分析出想要的DOM节点，获得其中的信息并获得所需内容的相关链接。由于此爬虫程序依据广度优先搜索策略，因此将优先搜索页面中的所有符合要求的链接，之后再向上进一步搜索。具体流程如下图所示：

（三）数据流程分析

一个搜索引擎的运行过程中，会有许多数据的形成和改变，最终才能将原本在URL服务器上的数据以结构化的方式呈现在用户的屏幕上。本项目是一个简单搜索引擎的模拟，数据在整个的获取、存储和搜索的过程中经历了六个过程：

在进行爬虫软件的编写之前，需要人工定向分析页面结构，按照HTML的DOM结构进行解析，爬虫程序向目标URL发送请求后，会得到服务器返回的结构化的页面内容，这是数据的第一个流向。

爬虫程序获取页面内容后，按照事先准备好的解析方案，对页面内容进行解析，利用类名（class名）或id名或者更直接的HTML标签名，获取指定DOM节点，得到的内容以既定的json结构存入json对象中，这是数据的第二个流向。

通过对json对象的整理，利用fs数据存取框架写入本地的json文件中，以备取用，这是数据的第三个流向。

利用开源框架Lucene，创建索引数据库，对output.json进行分词，分词后对内容提取，进行索引的存储，这是数据的第四个流向。

当用户发出搜索请求后，Lucene会创建一个搜索器，获取搜索的目标词和搜索范围，向索引数据库发出请求，得到一个类似于SQL的ResultSet的结果集，并返回，这是数据的第五个流向。

对结果集进行分析，提取出关键词所在文件位置和出现次数，返回到GUI页面上，呈现给用户，这是数据的第六个流向。

五、简易搜索引擎的设计与实现

（一）模块设计

1.模块设计的原则

一个系统的可维护性，是评价一个系统好坏的重要因素。设计一个搜索引擎也应该遵循高内聚、低耦合的原则。

2.爬虫模块的设计

本次课题的爬虫利用了nodejs和javascript来实现，爬虫本质上是一个大循环和字符串的提取，因此理论上可以使用任何语言来爬，在语言的选择上，我曾经做过一个技术选型的比较，备选的爬虫语言主要有以下三种：

JAVA：作为传统的编译型语言，JAVA更适合于项目厚重的爬虫，性能较高。

Python： python语言简洁，框架丰富，十分擅长做字符串的处理，且编辑方便，作为后端的脚本语言，即使不用IDE也能随时编辑。

nodejs：node是一个服务器端 JavaScript 解释器，利用谷歌V8引擎，特点有事件驱动、异步编程等，非阻塞I/O的异步爬虫可以大大降低爬虫时的等待时间。

三种语言各有优劣，但基于个人原因最终选择了nodejs作为爬虫语言。

我的爬虫主要分三步，一是分析页面的DOM结构，二是编写爬虫，三是利用Lucene将爬取到的数据存入索引数据库。

3.搜索模块的设计

搜索模块共有三个过程，首先由HTML接收到用户输入的数据，然后利用Ajax向后台发送异步请求，将查询数据以字符串的形式随着Ajax请求传到后台，服务器端接收到数据后，创建搜索器，然后进行一个类似SQL的关键字查询，Lucene会到之前创建好的索引数据库中进行查询并返回结果，服务器再通过Ajax将查询到的结果通过json的形式传回前端，再由前端通过jQuery解析后显示给用户。

关于数据结构，之所以向后端发送请求的时候，使用字符串形式，是因为通常情况下用户输入的查询条件都是一个字符串，而服务器端向前台传输数据的时候，由于数据中通常会包括标题、作者、时间、甚至是图片路径，因此会嵌套多层数据，有着比较复杂的结构，而json是一种轻量级的数据结构，在各种语言下都有良好的解析方式，避免了麻烦的xml或其他格式数据的解析。

（二）软件实现

1.html静态页面实现

2.爬虫项目实现

3.搜索模块实现

3.结论

。

参考文献

[1] 尹传勇,刘寿强,蒋建勋. 从IDS到IPS的主动防御体系研究[J]. 计算机安全. 2003(09).

[2] 吴穷,崔升广,乔英久. 浅谈入侵防御系统IPS[J]. 科技信息(学术研究). 2008(09)

[3] 余慧. IPS入侵防御系统[J]. 湖北教育学院学报. 2006(08)

[4] 裘锋. IDS、IPS技术研究[J]. 计算机与现代化. 2004(12)

[5] 王秋华. 浅析IDS及IPS网络安全深层防护策略[J]. 网络安全技术与应用. 2005(09)

[6] 卿昊,袁宏春. 入侵防御系统(IPS)的技术研究及其实现[J]. 通信技术. 2003(06)

[7] 崔健. IDS与IPS将长期共存——启明星辰IDS/IPS应用与发展趋势研讨会[J]. 信息安全与通信保密. 2007(10)

[8] 汪松鹤,任连兴. 入侵检测系统(IDS)与入侵防御系统(IPS)[J]. 安徽电子信息职业技术学院学报. 2004(Z1)

[9] 李安国. 计算机网络防护的主要手段[J]. 陕西师范大学学报(自然科学版). 2003(S1)

[10] 杨拥军. IDS与IPS将长期共存[J]. 信息安全与通信保密. 2007(07)