**基于Lucene的搜索**

**摘 要**

随着互联网的普及，数据量的增大，搜索引擎对日常生活产生巨大的影响。如何从海量的数据中准确找到自己需要的信息成为现今重要的话题。搜索引擎解决了用户检索的难题。

随着互联网的不断发展，搜索引擎成为现今上网的查找相关信息的必备工具。一般站点的各类信息都是存于数据库，再利用数据库的检索功能实现站点的搜索功能，但是随着数据量的剧增，使用数据库的模糊查询将导致检索功能的急剧性下降，而lucene搜索引擎的实现则可以避免这种问题。lucene正是这样一个开放源码的全文检索引擎工具包，可以在目标系统中实现全文检索的功能，甚至于建立起完整的全文索引引擎。

目 录

[1 绪论 1](#_Toc518392141)

[1.1系统开发意义以及背景 1](#_Toc518392142)

[1.1.1 开发背景 1](#_Toc518392143)

[2 开发工具及开发技术 2](#_Toc518392144)

[2.1开发环境及主要工具 2](#_Toc518392145)

[2.1.1系统开发环境 2](#_Toc518392146)

[2.1.2系统开发工具 2](#_Toc518392147)

[2.2系统实现技术介绍 2](#_Toc518392148)

[2.2.1 java编程技术 2](#_Toc518392149)

[2.2.2 Lucene介绍 3](#_Toc518392150)

[3 需求分析 4](#_Toc518392151)

[3.1需求陈述 4](#_Toc518392152)

[3.2系统可行性分析 4](#_Toc518392153)

[3.2.1时间可行性 4](#_Toc518392154)

[3.2.2用户可行性 4](#_Toc518392155)

[3.2.3技术可行性 5](#_Toc518392156)

[3.4 系统用例图 5](#_Toc518392157)

[3.4.1 用户用例图 5](#_Toc518392158)

[3.4.2 界面用例图 5](#_Toc518392159)

[总结与展望 6](#_Toc518392160)

# 1 绪论

## 1.1系统开发意义以及背景

### 1.1.1 开发背景

搜索引擎并不是真正搜索互联网，它搜索的实际上是预先整理好资料索引数据库。真正意义上的搜索引擎，通常指的是搜集了Internet上几千万到几十亿个网页并对网页中的每一个词（即关键词）进行索引，建立索引数据库的全文搜索引擎。当用户查找某个关键词的时候，所有在页面中包含了该关键词的网页都将作为搜索结果被搜出来 。在经过复杂的算法进行排序后，将这些结果按照与搜索关键词的相关度高低依次排列。

Lucene是apache软件基金会4 jakarta项目组的一个子项目，是一个开放源码的全文检索引擎工具包，是一个全文检索引擎的架构，提供完整的索引引擎和查询引擎。本系统将基于lucene工具包及相关的其他框架搭建一个小型的B/S结构的全文搜索，提高对搜索引擎技术的了解及应用。

# 2 开发工具及开发技术

## 2.1开发环境及主要工具

### 2.1.1系统开发环境

开发硬件环境：英特尔 第二代酷睿 i3-2310M @ 2.10GHz 双核，4G RAM，320G硬盘

开发软件环境：Windows 7 旗舰版 64位 SP1 ( DirectX 11 )

### 2.1.2系统开发工具

开发语言：Java 1.7.x

编程工具：Myeclipse 10.1

数据库工具：MySQL

开发框架：spring mvc 3.2.x、Spring3.2.x、Lucene 3.5 、mybatis3.2.x

其他工具：Microsoft Office Excel、Start UML 10.5、Microsoft Visio 2003

## 2.2系统实现技术介绍

### 2.2.1 java编程技术

1、JSP简介

JSP技术使用[Java编程语言](http://baike.baidu.com/view/53201.htm" \t "_blank)编写类XML的tags和scriptlets，来封装产生动态[网页](http://baike.baidu.com/view/828.htm" \t "_blank)的处理逻辑。网页还能通过tags和scriptlets访问存在于服务端的资源的[应用逻辑](http://baike.baidu.com/view/2646378.htm" \t "_blank)。JSP将网页逻辑与[网页设计](http://baike.baidu.com/view/8972.htm" \t "_blank)和显示分离，支持可重用的基于组件的设计，使基于Web的[应用程序](http://baike.baidu.com/view/330120.htm" \t "_blank)的开发变得迅速和容易出错。

JSP页面由HTML代码和嵌入其中的Java代码所组成。服务器在页面被客户端请求以后对这些Java代码进行处理，然后将生成的HTML页面返回给客户端的浏览器。[Java Servlet](http://baike.baidu.com/view/3618433.htm" \t "_blank)是JSP的技术基础，而且大型的Web应用程序的开发需要Java Servlet和JSP配合才能完成。JSP具备了Java技术的简单易用，完全的[面向对象](http://baike.baidu.com/view/125370.htm" \t "_blank)，具有平台无关性且安全可靠，主要面向[因特网](http://baike.baidu.com/view/1706.htm)的所有特点[4]。

### 2.2.2 Lucene介绍

Lucene是一个高性能、可伸缩的信息搜索(IR)库。它可以为你的应用程序添加索引和搜索能力。Lucene是用java实现的、成熟的开源项目，是著名的Apache Jakarta大家庭的一员，并且基于Apache软件许可 [ASF, License]。同样，Lucene是目前非常流行的、免费的Java信息搜索(IR)库。其具有如下突出的优点：

(1) 索引文件格式独立于应用平台，Lucene定义了一套以8位字节为基础的索引文件格式，使得兼容系统或者不同平台的应用能够共享建立的索引文件。

(2)在传统全文检索引擎的倒排索引的基础上，实现了分块索引，能够针对新的文件建立小文件索引，提升索引速度。然后通过与原有索引的合并，达到优化的目的。

(3)优秀的面向对象的系统架构，使得对于Lucene扩展的学习难度降低，方便扩充新功能。

(4)设计了独立于语言和文件格式的文本分析接口，索引器通过接受Token流完成索引文件的创立，用户扩展新的语言和文件格式，只需要实现文本分析的接口。

(5)已经默认实现了一套强大的查询引擎，用户无需自己编写代码即使系统可获得强大的查询能力，Lucene的查询实现中默认实现了布尔操作、模糊查询（Fuzzy Search）、分组查询等等。

# 3 需求分析

## 3.1需求陈述

作为一个以功能为主系统，本系统的立足点不仅仅是技术开发层面，而是整个项目的一体化进程，管理人员可以通过系统设置文档类型，设置用户信息。普通用户可以使用该系统的搜索功能，这些会得到对应的内容。系统还可以记录下用户的搜索记录，用来自动生成小时榜，二十四小时榜，周榜，月榜。这些搜索有利于用户便于了解当前最热门的词语，热门的情况。当然，当有人开始上传内容的时候，服务器会推送上传内容给在线的每个用户。用户还可以查询自己的账号情况，对于自己上传的内容还可以下载查看等。针对管理员用户提供了添加分类等功能。

## 3.2系统可行性分析

### 3.2.1时间可行性

该系统于2012年2月底启动，3月中上旬收集相关资料。估计3月下旬完成概要设计，于4月份前完成详细设计和数据库设计，4月份和5月上旬进行编码及测试工作，5月份开始整理资料和编写论文，5月下旬答辩。时间比较充分，所以在时间上是可行的。

### 3.2.2用户可行性

该系统的用户多为普通用户，没有专业训练，所以在交互方面我要做的更加人性化。从设计上我都会征询非计算机的同学对功能的需求，然后综合起来，得出一套比较符合他们的设计。系统拥有很多提示功能，就是用户出现任何错误的时候，系统都会出现一些提醒，告诉用户该如何去正确使用。进过一段时间的开发，得出一套操作方便的系统，所以在用户可行性上也是可行的。

### 3.2.3技术可行性

系统使用的主要是lucene技术，这套技术已经有多年的历史了，所以很完善，尤其是资料很丰富。当出现问题时候，去一些论坛上查找，都能很快找到一些解决的方法。当然里面有一些算法要我自己去设计，这个我也是先去论坛找一些类似的算法研究，然后开发出适用的一套算法。系统的界面开发，里面有使用到jquery的ui，这样页面上的设计业更加简单很多，总而言之，技术上是没有问题的。

## 3.4 系统用例图

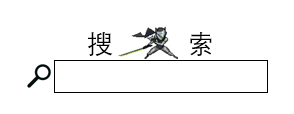
### 3.4.1 用户用例图

用户用例图如图3.2所示：



图3.2 用户用例图

### 3.4.2 界面用例图



# 总结与展望

经过不断的编码与调试，本系统基本实现了所有需求分析所需要的功能，预期目标也全部达到。目前系统已经可以正常交付使用，暂无发现严重BUG。

在本系统中，管理员通过后台系统操作系统信息，而用户则通过前台实现搜索功能，前后台管理互相呼应，构成整个全文搜索的核心。

通过对本系统的设计和实现，对计算机软件的设计有了进一步的了解和认识，也对此类的软件的应用前景及未来发展有了新的认识，希望在以后的工作中，能进一步研究和探讨：

(1)进一步完善本系统的功能，如：增加管理员操作日志的操作。在系统的起初需求设计时，这一功能是有增加进去的，数据库表和相应的dao层操作也已经建立完毕，本来打算利用spring的面向切面技术来实现，后来由于时间的关系只好作罢，后面有时间应该能添上此功能，还有一些类似的日志操作。

(2)提高效率，如：搜索效率和存储效率，在文档管理中，效率一直是个很重要的话题，尤其是这种面对大量用户的系统。如何提高效率就成了整个项目最重要的部分，因为现在热门的搜索引擎的效率都很快，这样才能提高用户体验。在之后的开发中，我会将效率视为最先的开发要求。

由于时间紧迫，还有很多功能没有开发出来，但是我会在后面的学习中继续完成，使之成为一个完整点的系统。