武汉理工大学毕业设计（论文）

**人脸跟踪与识别系统实现**

学院（系）： 计算机科学与技术学院

专业班级：软件工程专业 软件zy1202班

学生姓名： 杨晨

指导教师： 岑丽

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：

年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于1、保密囗，在 年解密后适用本授权书

2、不保密囗 。

（请在以上相应方框内打“√”）

作者签名： 年 月 日

导师签名： 年 月 日

摘要

在这个互联网飞速发展的时代，计算机技术已经深入到每个人的生活中。而计算机图形图像技术在生活中被也越来越多的运用起来。人脸跟踪与识别系统，涉及到人脸的检测与人脸的识别。将其运用到Android系统之中，从而实现了通过手机摄像头识别人脸的功能。

本文首先阐述了人脸检测Haar分类器、人脸识别LBP算法的相关内容，然后根据实际的Android开发内容编写了需求报告。首先通过JavaCV中Camera相关类获取视频流的一帧，然后使用Haar分类器进行人脸检测，获取人脸矩阵信息。对录入的人脸进行训练后再使用LBP算法进行人脸识别，最后在屏幕上绘制出人脸的矩阵与识别出的人的姓名。在下一帧中再进行如上的操作，从而实现了对视频流的人脸识别。

关键词：人脸识别；haar分类器；LBP算法；Android

**Abstract**

In this era of rapid development of Internet,computer technology has been deep into everyone’s life.The computer graphics technology is more and more used in our life.Face tracking and recognition system,related to face recognition and face detection.As it applied to Android system,enabling moblile phone camera face recognition function.

The paper describes the face detection Haar classifier,the content LBP face recognition algorithm,and then write a report based on the actual needs of the Android development content.Get through the first JavaCV Camera Class a video stream,and then use the Haar classifier for face detection,face matrix to obtain information.On the entry face training before using LBP face recognition algorithm,and finally draw the human face of the matrix and identified the person’s name on the screen.In the next frame,then the above operations,in order to achieve the recognition of the video stream.

**keywords**：face recognition；haar classifier；LBP algorithm；Android

目录

第1章 绪论1

1.1 研究背景及国内外研究现状1

1.2 研究目的及意义3

1.3 研究内容及目标4

1.4 论文组织结构4

第2章 相关技术及开发工具简介6

2.1 搜索引擎6

2.2 RSS介绍6

2.3 开发工具7

2.4 开发技术7

第3章 新闻搜索引擎需求分析9

3.1 功能需求9

3.1.1 爬取子系统功能需求9

3.1.2 索引子系统功能需求10

3.1.3 web客户端功能需求10

3.2 非功能需求11

第4章 系统数据库与详细设计12

4.1 数据库设计12

4.1.1 实体属性图12

4.1.2 数据库设计12

4.2 系统详细设计13

4.2.1 爬虫子系统模块设计13

4.2.2 索引子系统模块设计14

4.2.3 web客户端模块设计15

第5章 系统实现17

5.1 系统总体结构17

5.2 爬虫子系统实现流程17

5.3 索引子系统实现流程19

5.3.1 关键词提取算法分析19

5.3.2 搜索引擎排序算法分析21

5.3.3 建立索引的实现流程23

5.4 搜索引擎框架搭建25

5.4.1 B/S架构设计25

5.4.2 UI设计25

5.4.3 前端框架开发流程25

第6章 关键问题及解决方法27

6.1 爬虫子系统部分27

6.2 索引子系统部分27

6.3 web客户端部分27

第7章 总结与展望29

参考文献30

致谢32

# 

# 第1章 绪论

## 1.1 研究背景及国内外研究现状

在最近10年中，计算机科学技术取得了巨大的发展，人们的生活水平得到了很大的提升，人们越来越多的关注社会信息。在很多行业中，对人员进行的信息进行确认与身份的辨别的需求越来越大。例如公安机关设置摄像头，来监控地铁，商场，火车站等人流量打的地方，根据画面中出现的人脸来寻找犯罪嫌疑人。在以前这样的工作通常是由人工完成的，不仅消耗很多的人力，也消耗很多的时间；在机场，登机时，工作人员会对将人脸与身份证进行核对；在银行取钱时，银行需要对取钱的用户进行人脸识别，判断其是否是本人；在海关，同样也需要人脸识别技术来确认出入境人员是否为本人，这些在生活中切实存在的需求极大的推动了近些年人脸识别技术的飞速发展与应用。

人脸识别技术是通过计算机图形学，提取人脸的特征，通过这些特征，与训练好的样本进行对比，从而进行身份验证的一种技术。人脸特征与人身上其他的特征一样，例如指纹和虹膜，都具有唯一性和不容易被伪造的特性，这是作为身份鉴别依据的前提；对比其他生物特征的识别技术，人脸识别技术在操作上更简单，结果更直观，更加隐蔽。因此，在刑事案件的侦破，信息安全领域等都有广泛的应用前景。

人脸识别技术在最早的时候限制很大，只能对单一背景的正面灰度图像进行识别，也就是只能在二维中进行人脸识别。随着对多种姿态下人脸识别的研究，包括正面和侧面，人脸识别正在向三维方向发展，在维度提升的同时，识别率也在逐步提高。在这个过程中，人脸识别技术越来越成熟，但是还是存在着很多的问题，比如在复杂的背景下，系统很难识别出其中的人脸，跟踪也存在比较大的问题。人脸识别技术并非纯粹的数字技术，它需要研究人员拥有计算机图形学、计算机视觉、生物特征技术等学科的知识，是一个融合了多种学科的技术，因此对研究人员有非常高的要求。环境和人脸本身都有很大的不确定性，在环境方面，光照和图像采集工具会对采集到的图像有比较大的影响，而人脸本身由于表情，身体姿态和脸上戴的物件的不同，会对人脸识别造成影响。综上所述，人脸识别是非常有挑战性的课题。

人脸识别有悠久的研究历史。高尔顿分别在1888年和1910年在《Nature》杂志上发表了两篇关于利用人脸来识别身份的文章。这两篇文章对人类自身如何进行人脸的识别进行了分析。在当时的科学文化环境下，并没有涉及到人脸的自动识别。

1965年陈和布莱德索在Panoramic Research Inc.发表了关于自动人脸识别的研究论文。布莱德索建立了一个半自动的人脸识别系统。该系统识别人脸是以人脸的特征点的间距和比率等参数特征。这种思想也成为之后一段时间的主流。这种人脸识别思想的特点有：①将人脸拆分成若干个部件，作为特征进行识别，主要利用的是各个部件的信息和部件之间的几何关系。这是一种很直观的方法，对人脸图像有很高的要求。如果人脸图像中的人脸不是正面照或者表情有变化的话识别率就会很低。为了弥补这种方法的缺点，后来出现了更好的人脸识别方法。根据样本库中的样本和需要识别的人脸的灰度图的比较来对人脸进行鉴别，这种方法叫做模板匹配方法。②人脸识别对环境的要求非常高，在单一背景或者没有背景的情况下，人脸的位置非常容易获得，相反，在复杂的背景下，人脸的位置非常难获得。在这段时间，人脸识别由于这个原因，并没有被应用到现实场景当中。

1964年至1990年是人脸识别的第一个阶段，这一阶段人脸识别通常作为一个普通的模式识别问题来研究。这个阶段采用的技术方案是基于人脸几何特征的方法。研究者们花费了很多时间在对面部剪影曲线的结构特征的提取和分析方面。而人工神经网络也曾被用于人脸识别的问题中。早期，除了布莱德索在进行自动人脸识别之外还有戈登斯泰因、金出武雄等。金出武雄是人脸识别领域较活跃的人物，它在1973年，于京都大学完成了自动人脸识别方面的博士论文。如今，他成为了卡内基-梅隆大学机器人研究院的教授。

第二个阶段是1991年至1997年。在短短的8年时间内，人脸识别的研究进入了高潮期，取得了很多的成就。许多具有代表性的人脸识别算法在段时间内诞生。比如著名的FERET人脸识别算法，麻省理工学院媒体实验室的特克和潘特提出的“特征脸”方法。“特征脸”方法是这一时间段内的代表作，衍生出了之后许多的算法。如今特征脸已经与归一化的协相关量方法一起成为人脸识别的性能测试基准算法。麻省理工学院人工智能实验室的布鲁内里和波基奥进行了基于结构特征的方法与模板匹配方法的对比，得出了模板匹配要优于基于特征的方法这个结论，促进了基于表现的线性子空间建模和基于统计模式识别技术的人脸识别方法的发展。这一时期主要的成果有：贝尔湖米尔等提出的Fisherface人脸识别方法；麻省理工学院的马哈单提出的基于双子空间进行贝叶斯概率估计的人脸识别方法。

第三个阶段是从1998年至现在。为了解决在光照，姿态等非理想的情况下人脸识别鲁棒性较差的问题，研究者么注重研究了这一部分。吉奥盖迪斯等人提出基于光照锥模型的多姿态、多光照条件人脸识别方法。布兰兹和维特等提出了基于3D变形模型的多光照和姿态条件人脸图像分析与识别的方法。它是基于合成的分析技术。这种方法使得人脸可以用简单的举证特征作为特征，将大量弱分类器组合成了强分类器，并且使用联级技术提高检测速度，为现在实时人脸识别打下了很好的基础。沙苏哈在2001年提出了一种基于熵图像的人脸图像识别和绘制技术。为解决光照问题提供了一种重要的思路。

现在人脸识别的算法主要分成3种。

(1)基于肤色划分的人脸检测方法。首先确定肤色模型，检测出其中的肤色区域，获得可能存在的人脸区域。之后是对人脸区域的检测，通过区域特征的方法对人脸区域进行人脸检测，从而区分出具有类肤色的其他物体。这种方法在复杂背景的环境下会有比较大的误差。在复杂背景下，人脸区域可能和其他肤色区域混杂在一起，从而造成混淆，在获得肤色区域后任然无法判断出此区域是人脸区域。另一种情况是光照和面部表情造成的影响。光照和面部表情会对人脸检测造成影响，区域特征会受到影响，在这种情况下会使用聚类，归并，验证的方案来减少其带来的影响。首先将人脸肤色像素按照严格的几何关系去拆分，再将其按照一定规则合并到一起，在合并过程中使用一些其他特征进行验证。

(2)基于人工神经网的方法。由于人脸轮廓的复杂性，人脸很难用数学模型表示，而神经网是一种可以表示复杂模型的方法。人工神经网的优势是方便建模和较高的鲁棒性。但是人工神经网在运算速度上有比较大的欠缺，这是制约它发展的一个因素。

(3)基于启发式模型的方法。这种方法主要是通过抽取图像的若干特征进行人脸检测。首先对人脸局部的特征进行判断，再对人脸整体布局进行判断。也就是从人的五官的检测到人的五官的相对位置的检测。使用人脸五官分布特征的知识模型进行检测。由于抽取的特征较少，所以这种方法有比较快的检测速度。这种方法的主要障碍也是由于抽取特征较少，在复杂环境下的人脸检测成功率较低。

## 1.2 研究目的及意义

人脸识别在生活中有广泛的应用场景。比较常见的有安检，监控，身份认证等。随着人脸识别技术渐渐成熟，人脸识别会更多的出现在我们生活中的各个角落。人脸识别的优势是新颖的交互方法和人脸的不变性，劣势是复杂的环境下识别率低。研究的目的便是实现人脸识别，并且提高人脸识别的准确率，并且将其运用到实际生活中。

随着支付宝刷脸支付，FaceU检测人脸实时添加表情，腾讯QQ人脸登录等人脸识别技术在移动端的应用，人脸识别技术在移动端的应用也是越来越广泛。学习人脸识别技术在移动端的应用具有重要的意义。

## 1.3 研究内容及目标

本文主要的研究内容是人脸检测与人脸识别在安卓终端的应用，其中主要内容有：

(1)人脸检测算法涉及到Haar分类器的研究。

(2)人脸识别算法涉及到LBP人脸识别算法与样本训练。

(3)研究使用基于opencv的JavaCV库。

(4)研究安卓客户端开发技术,。

项目的研究目标有：

1. 实现80%的人脸检测成功率。
2. 实现在非复杂背景下50%的人脸识别成功率。
3. 实现安卓客户端录入人脸样本并且完成训练。

## 1.4 论文组织结构

本文一共分为7章，各章节具体介绍如下：

第一章：绪论。主要介绍了人脸识别的发展前景与历史以及当前人脸识别主要使用的技术，讲述了研究人脸识别的目的和意义，最后列出了研究的具体内容与目标。

第二章：相关技术及开发工具简介。介绍了搜索引擎的组成和RSS新闻源的特点，然后对此次系统的开发工具和开发技术做了大体的介绍。

第三章：新闻搜索引擎需求分析。对系统的每一个模块进行功能需求说明，对整个系统进行非功能需求说明。

第四章：系统数据库与详细设计。主要说明了系统的设计思路与功能结构，并对每一个模块进行说明。并且介绍了数据库的相关设计。

第五章：系统实现。首先介绍了代码整体结构，然后介绍了每一个子系统的具体实现流程，并附上核心代码与截图，同时对核心算法进行了分析与比较。

第六章：关键问题及解决方法。主要说明了系统开发中遇到的关键问题和处理方法。

第七章：总结与展望。对基于web的新闻搜索引擎系统的开发工作进行了总结，分析了有待改进与完善的部分。

# 第2章 相关技术及开发工具简介

## 2.1 人脸识别技术

人脸检测技术是指从图像或视频流中通过一定策略检测到人脸的位置，大小等信息的技术。比较常见的人脸检测方法有参考模型法、人脸规则法、样品学习法、肤色模型法、特征子脸法。人脸检测的难点在于人脸内在的变化和人脸外部环境的变化。本文使用了基于Haar分类的人脸检测方法进行人脸检测。

## 2.2 人脸识别技术

人脸识别技术通常与人脸检测技术捆绑在一起。人脸识别技术一般分为3个过程：(1)通过一定方法获取人面信息，生成面纹并且存储到面纹库。(2)通过摄像机或相机获取当前人脸信息，生成当前面纹。(3)将获取到的当前面纹与面纹库中的面纹通过一定方法进行对比，找到最相近的面纹，从而获取到识别出的对象。现在主要的人脸识别方法有：几何特征的人脸识别方法、基于特征脸的识别方法、神经网络的人脸识别方法、弹性图匹配的人脸识别方法、线段Hausdorff距离的人脸识别方法、支持向量机(SVM)的人脸识别方法。人脸识别技术具有非接触，非强制性，并发性等优点。人脸识别技术的缺点是人脸所在背景对识别的准确度影响较大，人脸本身的表情和穿戴的装饰对人脸识别的结果也影响较大。

本文使用了LBP算法进行人脸识别。

## 2.3 开发工具

（1）Android Studio

Android Studio是谷歌推出的一个集成开发环境，提供了Android开发所需要的开发和调试工具。Android Studio使用Gradle进行项目管理，它是在Intellij IDEA的基础上进行开发的。Gradle是一种高级的构建工具，用于管理依赖性，允许自定义构建逻辑。构建系统可以根据不同的配置为同一个项目生成多个APK。构建系统同时支持本地文件系统和远程存储库支持的依赖，这样就不用把依赖库下载到本地了。Android Studio分成三个模块。Java库模块，Android库模块，Android应用程序模块。Android Studio 2.0提供了运行时修改代码，并且极大的提高了虚拟机的运行速度。使用Android Studio可以方便的进行Android开发。

## 2.4 开发技术

（1）Java语言介绍

Android开发主要使用的Java语言。Java语言是由Sun Micrososystems公司与1995年推出的。Java是一种简单的、面向对象的、分布式的、解释型的、健壮安全的、结构中立的、可移植的、性能优异、多线程的动态语言。Java语言的这些优良特性使得Java应用具有无比的健壮性和可靠性。Java是编译性语言和解释型语言的合集。一个java文件首先被编译成class文件，然后再被解释成0和1组成的二进制指令并被执行。使用Java语言需要安装JDK和JRE。JDK是Java开发包，是一个开发工具的合集。JRE是Java的运行环境包含了JVM的标准实现及Java核心类库。

（2）JavaCV库介绍

JavaCV库是一个开源库，它是基于OpenCV的Java封装。OpenCV是一个跨平台的计算机视觉库。在Windows，Linux和Mac OS上都可以运行OpenCV使用C++编写，它主要的接口是面向C++语言的。但它也给其他的语言提供了接口，比如说:Python,Ruby,MATLAB。实现了图片处理和计算机视觉方面的很多算法。

# 第3章 人脸识别需求分析

## 3.1 功能需求

人脸识别系统划分为4个模块。第一个是人脸样本获取模块，第二个是当前图像获取模块，第三个是人脸检测模块，第四个是人脸识别模块。这4个模块相对独立，但是需要为彼此提供接口，最终实现从图像获取到人脸识别的功能。

## 3.1.1 人脸样本获取模块

人脸样本包含的要素有两个，第一个是人脸图片，第二个是人的姓名。在存储时，需要把这两个要素结合起来。一个人的人脸信息需要多张人脸图进行训练得到，因此在取样时需要同一个人拍摄多张人脸图像。

因为拍摄时人脸所处的环境可能比较复杂，所以需要在拍摄完成后对图片进行裁剪。裁剪时需要统一图片的尺寸，质量，图片格式。

需要有给用户输入姓名的编辑框，在用户输入姓名后开始拍摄图片，再对拍摄好的图片进行裁剪。

## 3.1.2 当前图像获取模块

当前图像是从视频流中获取的。在Android设备上的表现是：通过前置或后置摄像头获取视频流，再从视频流中获取单帧图像。

需要注意的是屏幕的方向以及最终单帧图像的数据结构。应当与人脸检测模块的输入一致。

需要实现实时预览摄像头获取到的画面。

## 3.1.3 人脸检测功能需求

人脸检测模块接收获取到的视频单帧图像，然后对图像中的人脸进行检测，最终将检测到的人脸位置大小显示在预览画面中。检测结果要求显示清晰，明了，一眼可以看出检测到的人脸位置和大小。

每一帧要进行检测，如果检测到人脸及实时绘制出人脸所在位置和大小，如果未检测到人脸则不绘制任何东西。

人脸检测的结果需要作为人脸识别模块的输入，因此要统一人脸数据的数据结构。

## 3.1.4 人脸识别功能需求

当人脸检测检测到人脸的时候，将人脸数据传递给人脸识别模块，人脸识别模块需要通过一定的算法与本地人脸面纹进行比对，得到人脸信息。如果匹配到了结果，则在预览页面显示出匹配到的人的姓名，如果未匹配到，则显示未匹配到身份。

## 3.2 非功能需求

为了提高用户体验，人脸识别系统也需要有以下的非功能需求：

（1）可用性：样本收集，人脸识别的过程应当减少用户的操作，通过尽量少的步骤完成整个人脸识别过程。检测和识别结果应当简单明了的显示出来。

（2）可靠性：人脸检测和人脸识别的成功率应当比较高，可以适应复杂环境下的人脸识别需求。

（3）健壮性：可以适配不同安卓版本，不同机型的终端，对异常有优秀的处理。

（4）可维护性：人脸识别系统的每一个模块应当相对独立，减少相对之间的依赖，提取公共函数。保证可以方便的更换人脸检测和人脸识别算法，可以方便的修改系统的各个部分。

# 第4章 算法分析

## 4.1 Haar分类器

目前主要有两类人脸检测方法：基于知识和基于统计。基于知识的人脸检测方法是利用先验知识，根据人脸器官的特征，以及他们之间的关系来检测人脸的。基于统计的人脸检测是把人脸当成一个整体的模式，通过大量的人脸样本，使用统计的方法来构造人脸模式空间，利用相似度来判断是否存在人脸。目前很多方法是把基于知识和基于统计结合在一起使用的。

Haar分类器包含了Adaboot算法。分类器是数据挖掘的一种重要概念。分类器把有限数据分成若干个类，任何一个新的数据都可以映射到某一个类中，从而可以对数据进行预测。比如在这里，分类器将所有的样本分成两类，第一类是人脸，第二类是非人脸。

Haar分类器实际上是Boosting算法的一个应用，Haar分类器用到了Boosting算法中的Adaboot算法。Haar分类器将AdaBoost算法训练出的强分类器进行了联级。Haar分类器是由Haar-Like特征、积分图方法、AdaBoost组成的。

## 4.1.1 Haar-Like特征

Haar-like特征在最早的时候是用于人脸表示的，由Papageorgiou等提出。Haar特征分为3类：边缘特征、线性特征、中心特征和对角线特征，共同组合成特征模板。在这个特征模板中分成了两种颜色的矩形，白色和黑色，定义模板的特征值为白色矩形像素减去黑色矩形像素和。Haar特征值反映的是图像在灰度上的变化。例如其在人脸上的表现为：眼睛比鼻子的颜色深；嘴巴颜色比周围颜色深。由于特征是矩形的，所以其只能描述特定走向的边缘或线段结构。

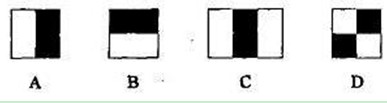


图4.1特征模板

对于图中A,B和D这几种特征，特征数值计算公式是：v=白色的和-黑色的和。但是对于C来说，计算公式为：v=白色的和-2\*黑色的和。黑色区域乘二是为了使两种矩形区域中的像素数量相同。

矩形特征可以在图像的任意位置，可以设置任意大小，可以选定不同的矩形模板类别。因此，在极小的检测窗口也可能会有大量的矩形特征。

## 4.1.2 AdaBoost

AdaBoost实际上是一种具有一般性的用来提升分类器的算法它并不局限使用某一特定的分类器。AdaBoost可以帮助我们更好地选择矩阵特征的组合，矩阵特征组合将以二叉决策树的形式存储起来。

AdaBoost的核心思想是针对同一个训练集训练不同的若干个弱分类器，最终将这些弱分类器组合成一个强分类器。

Adaboost算法是通过迭代实现的，其中重要的一步是更改数据的分布。根据每次训练集中样本是否正确来给样本设置权值，然后把划分权值的分类器传递给下层分类器进行训练，将每次训练得到的弱分类器融合，得到最终的强分类器。具体过程如下：

1. 对N个训练样本进行学习，得到第一个弱分类器。
2. 使用测试样本进行测试，将分错的样本和其他的新样本构成一个新的样本数为N的训练样本，学习得到第二个弱分类器。
3. 把（1）和（2）测试出来分错的样本加上若干新的样本组成N个训练样本，学习得到第三个弱分类器。
4. 反复进行上面的步骤，学习得到若干个弱分类器，最终组合成一个强分类器。

AdaBoost算法对分错的样本进行了加权，使得训练的焦点放在了比较难区分的训练样本上。该算法还对弱分类器进行了加权处理，使得分类效果较好的分类器具有比较高的权重，分类效果差的分类器具有较小的权重。

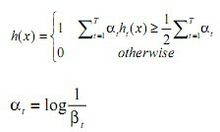


图4.2权重处理

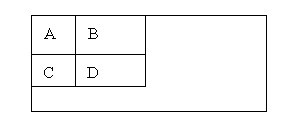
## 4.1.3 积分图

积分图是只遍历一次图像就可以求出图像中所有区域像素和的快速算法，大大提升了图像特征值计算的效率。在进行haar-like分类器训练和检测的过程中，每当遇到图片样本的时候都会需要计算某一个窗口的特征值，这样的计算是非常大的，而积分图就可以解决haar-like计算量过大的问题。积分图是一种可以描述全局信息的矩阵表示方法。积分图的构造方法是位置（i，j）处的值ii（i，j）是原图像（i，j）左上角方向所有像素的和：

http://images.cnblogs.com/cnblogs_com/ello/%E5%9B%BE%E7%89%8720.png

积分图的构建算法如下：

1. 用s（i , j）表示行方向的累加和，初始化s(i,-1)=0;
2. 用ii(i , j)表示一个积分图像，初始化ii(-1,i)=0；
3. 逐行扫描图像，递归计算每个像素(i , j)行方向的累加和s(i , j)和积分图像ii(i ,j)的值
4. s(i ,j)=s(i ,j-1)+f(i , j)
5. ii(i,j)=ii(i-1,j)+s(i,j)
6. 扫描图像一遍，当到达图像右下角像素时，积分图像ii就构造好了。
7. 积分图构造好之后，图像中任何矩阵区域的像素累加和都可以通过简单运算得到如图所示。



1. 设D的四个顶点分别为α、β、γ、δ，则D的像素和可以表示为
2. Dsum = ii( α )+ii( β)-(ii( γ)+ii( δ ));

## 4.2 LBP算法

LBP算法，Local Binary Patterns，局部二值模式。是一种使用局部特征作为判别依据的识别算法。使用LBP时，图像必须为灰度图。每张灰度图都是由一个一个像素点组成，每个像素点都有自己的灰度。

LBP算法最初是定义在一个3×3的一个邻域内的，以中间的点作为阈值，将邻近的8个像素的灰度值与中间这个像素的灰度值进行对比，如果大于中心点的灰度值，则记为1，如果小于中心点的灰度值，则记为0。这样在每个点对比过后，可以得到一个8位的2进制数，即这个邻域的LBP值。这个值表现的是这个邻域的纹理信息。

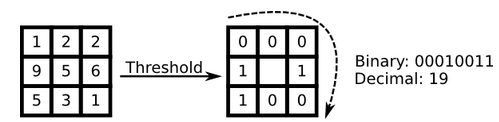


图4.3 LBP算法示意图

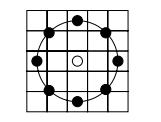
正式公式为：

http://img.blog.csdn.net/20140409102028328

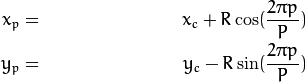
其中http://img.blog.csdn.net/20140409102051484表示中间的那个点。ic表示中间点的灰度值，ip表示邻域的点的灰度值。S（x）是符号函数，定义如下：

http://img.blog.csdn.net/20140409102057984

最初的LBP算法存在着缺陷，它无法应对不同尺寸和不同频率纹理以及旋转的情况。因此Ojala对LBP算法进行了改进。他将3乘3的邻域改成了任意大小的邻域，并且并且由原来的正方形改成了圆形。类似于下图：



假设圆的半径为R，其中有P个采样点，其中每个采样点的值可以用下面的公式表示：



其中http://img.blog.csdn.net/20140409102904234为中心点，http://img.blog.csdn.net/20140409102917984为某个采样点，通过上式可以计算出每个采样点的坐标值，但是坐标并不一定是整数值，可以通过双线性插值计算得到采样点的像素：

http://img.blog.csdn.net/20140409103242671

在获取到人脸的LBP特征后还需要进行特征匹配，例如下面这张人脸图像，将其划分

为7×7的子区域并计算得到每个邻域的LBP值并且统计其直方图。这样做可以避免人脸没有完全对准的情况，也对LBP特征进行了降维处理。在得到直方图后，有多种方法可以判别其相似性。可以使用：

1. 直方图交叉核算法

http://img.blog.csdn.net/20140409110315812

1. 卡方统计方法

http://img.blog.csdn.net/20140409110426046

其中Mi为已知人脸直方图，Si为待匹配人脸直方图。

# 第5章 系统实现

## 5.1 系统总体结构

系统总体结构如图5.1

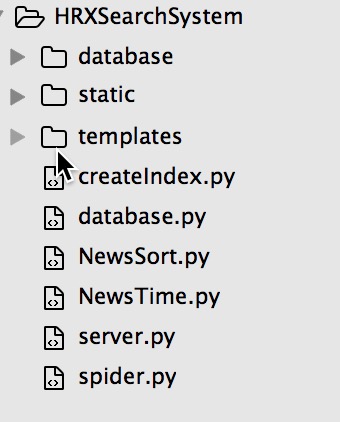


图5.1 总体结构图

本系统使用的是flask框架，文件结构和通用web应用类似。database文件中存储的是和数据库操作有关的文件。

static文件夹中存储的是静态文件，包括CSS，fonts，images，JavaScript等文件。

template中存储的是页面模板，即待jinja2模板引擎渲染的html页面。

后面的py文件是服务器端运行的python文件，server.py是服务器主进程，负责监听端口，响应客户端请求，处理用户输入，从数据库索引表中得到新闻条目计算权值并且渲染页面等。

Spider.py 是爬虫子系统的主要代码，负责将新闻爬取下来并且调用createIndex.py中的方法为新闻建立索引记录，将相关信息存入数据库中。

NewsSort.py、NewsTime.py、datebase.py分别处理有关相关度排序，新闻事件处理，数据库操作相关的一些业务逻辑代码，将这部分重用性较高的代码分离出来有利于降低系统内代码的耦合度，便于后期对这部分代码进行修改和维护。

## 5.2 爬虫子系统实现流程

首先在腾讯和网易的RSS新闻订阅源上，按照体育、国际、国内、财经、科技和娱乐等类别收集RSS源并放在一个字典中。feedparser是较为通用的用于解析rss格式的python模块，它的强大不仅是因为它对各种rss版本的兼容性好，而且使用起来非常简单，只用feedparser的parse函数就可以方便的对rss源进行下载并解析，d = feedparser.parse(url)，这样rss的所有内容就都在d里面了。用feedparser遍历所有的rss源，对每一个rss源的所有新闻项目进行分析，获取到日期、标题、提要、链接和类别这些信息之后，将这些信息一条一条地存入数据库中。

网络爬虫从RSS新闻源提取新闻相关信息并存入数据库的核心代码如下：

#将所有rss源分类放在一个字典中

rssUrls = {'体育':['http://sports.qq.com/basket/rss\_basket.xml',

'http://sports.qq.com/isocce/rss\_isocce.xml',

'http://sports.163.com/special/00051K7F/rss\_sportsyc.xml'],

'国际':['http://news.qq.com/newsgj/rss\_newswj.xml'],

'国内':['http://news.qq.com/newsgn/rss\_newsgn.xml',

'http://news.163.com/special/00011K6L/rss\_sh.xml'],

'财经':['http://finance.qq.com/financenews/domestic/rss\_domestic.xml'],

'科技':['http://tech.qq.com/web/webnews/rss\_11.xml',

'http://tech.163.com/special/000944OI/hulianwang.xml'],

'娱乐':['http://ent.163.com/special/00031K7Q/rss\_entmovie.xml']}

#提取每个rss源的每条新闻的类别,日期,标题,提要,链接

def getItemsInfo(category,rssUrl):

itemList = []

d = feedparser.parse(rssUrl)

items = d.entries

for item in items:

date = item.published

if "GMT" in date:

date = NewsTime.wyNewsTimeFormat(date)

title = item.title

summary = item.summary

link = item.link

#为每条link在数据库建立索引

# createIndex.createIndex(link,title+summary.split('<')[0])

# createIndex.createIndex(link,title)

itemList.append((category,date,title,summary,link))

return itemList

#爬虫主进程

def spider(rssUrls):

for category,urls in rssUrls.items():

for url in urls:

itemList = getItemsInfo(category,url)

#把抓取的内容存入数据库

database.saveToDB(itemList)

该部分运行截图如图5.2:

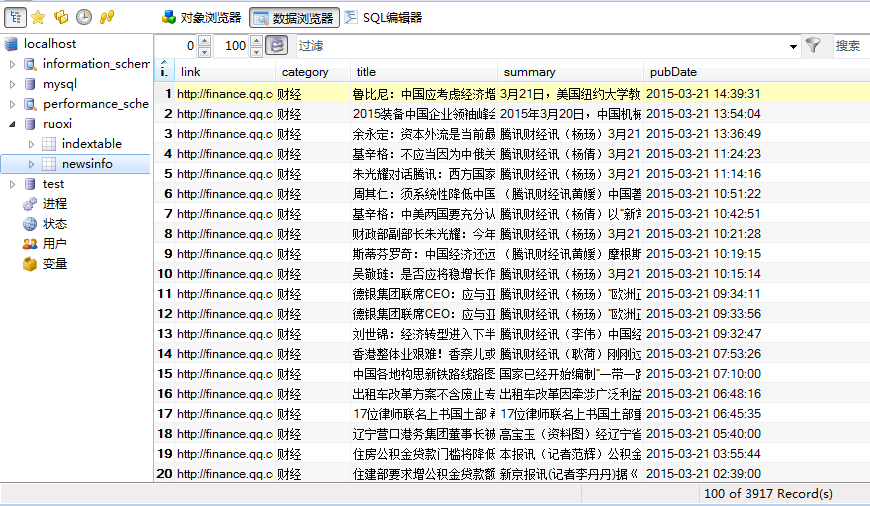


图5.2 爬取子系统截图

## 5.3 索引子系统实现流程

## 5.3.1 关键词提取算法分析

关键词提取就是从文本里面把与这篇文章意义最相关的一些词抽取出来。这个可以追溯到文献检索初期，当时还不支持全文搜索的时候，关键词就可以作为搜索这篇论文的词语。因此，目前依然可以在论文中看到关键词这一项。除此之外，关键词在文本聚类、分类、摘要等领域中也有着重要的作用。比如从某天所有新闻中提取出这些新闻的关键词，就可以大致知道那天发生了什么事情。

关键词提取从方法来说大致有两种：一种叫做关键词分配，就是给定一个关键词库，然后对于一篇文章，从词库里面找到几个词语作为这篇文章的关键词。另一种是关键词抽取，就是对于一篇文章，从文章中抽取一些词语作为这篇文章的关键词。目前大多数领域无关的关键词抽取算法（领域无关算法的意思就是无论什么主题或者领域的文本都可以抽关键词的算法）和它对应的库都是基于后者的。从逻辑上说，后者比前者在实际使用中更有意义。另外，从结果的角度来说，关键词抽取也可以分为两种。一种是仅仅把词语抽取出来，这个相对简单，容易实现，比如FudanNLP、jieba、SnowNLP。另外一种则是连词和短语一起抽取出来，这个还需要增加短语抽取这一个步骤，这一类的实现包括ICTCLAS、ansj\_seg等。常见的关键词抽取算法有TextRank算法和TF-IDF算法。

关键词特征之一就是在文本中反复出现且关键词附近出现关键词的概率非常大，因此就有了TextRank算法[10]。它基于PageRank算法，将文本中每个词看成一个页面，认为文本中某一个词语与之周围N个词存在一个link，然后在这个网络中使用PageRank算出每个词语的权值，把权值最高的几个词作为关键词即可。TextRank典型的实现包括FudanNLP和SnowNLP等。

使用TextRank提取关键字，首先将原文本拆分为句子，然后在每个句子中过滤掉停用词，并只保留指定词性的单词，由此可以得到句子的集合和单词的集合。每个单词作为pagerank中的一个节点。设定窗口大小为k，假设一个句子依次由下面单词组成：，，，，，…，。其中，，…，、，，…，、，，…，等都是一个窗口，在一个窗口中的任两个单词对应的节点之间存在一个无向无权的边。基于上面构成图，可以计算出每个单词节点的重要性，最重要的若干个单词就可以视作关键词。

最为常见且最易实现的关键词提取算法就是TF-IDF算法，它可以找到文本中常见但是在别的文本中不常出现的词语，这正好符合关键词的特点。字词的重要性随着它在文件中出现的次数成正比增加，但同时也会随着它在索引库中出现的频率成反比下降。TextRank实际应用效果并不比TF-IDF有明显优势，而且由于涉及网络构建和随机游走的迭代算法，效率极低。TF-IDF具有较强的普适性，该算法基本能应付大部分关键词抽取的场景。使用这种方法典型的例子是jieba。

本系统采用的便是TF-IDF算法，其算法原理如下：

1.先给本聚类内所有文档进行分词，然后用一个字典保存每个词出现的次数，即词频(term frequency, TF)。由于同一个词语在长文件里可能会比在短文件有更高的词频，而不管该词语重要与否，所以该数字通常会被归一化，以防止它偏向长文件。

2.遍历每个词，得到每个词在文档里的逆向文件频率 (inverse document frequency, IDF)，即在所有文章里出现的倒文档频率。某一特定词语的IDF值，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。

3.用一个key为词，value为TF\*IDF权重的字典来保存所有词信息，然后按value对字典排序，取权重高的若干个词作为关键词。

对于在某一特定文件里的词语来说，它的重要性可表示为：

其中，是该词在文件中出现的次数，分母是文件中所有字词的出现次数之和。

其中，是语料库中的文件总数，||是包含词语的文件数目。如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用1+||。

然后

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

## 5.3.2 搜索引擎排序算法分析

排序算法往往是搜索引擎的核心竞争力之一，用户行为研究显示，通常用户只会关注搜索结果的前几条，越往后面的结果越少有人关注。因此改进排序算法，让用户最感兴趣最关注的内容排在前面是提升用户满意度的关键。常用的搜索引擎排序算法有：PageRank算法、HITS算法、Direct Hit算法等[11][12]。

PageRank 算法是当前最为主流的搜索引擎排序算法之一，它是一种和查询词无关的算法，在用户查询之前进行离线计算网页等级，不占用查询时的计算时间。现已成功运用于Google搜索引擎。PageRank 算法的中心思想是一个网页的重要性决定并且依赖着其他网页的重要性，一个网页被许多其它网页引用，则它可能是重要页面；一个网页尽管没有被多次引用，但如果它被一个重要的网页引用，则它也可能是重要页面；一个网页的重要性被平均的传递到它所引用的网页。

PageRank算法的优点在于它对互联网上的网页给出了一个全局的重要性排序，并且算法的计算过程是离线完成的，在用户查询时仅仅根据关键字匹配获得网页集合，然后排序推荐给用户，这样有利于迅速响应用户的请求。但该算法还存在不少缺陷：主题无关性，无法区分页面内的导航链接、广告链接和功能链接等，容易对广告页面有过高评价，从而导致出现“主题漂移”问题。另外，旧的页面等级会比新页面等级高，因为新网页通常拥有很少的链入网页数量，所以其PageRank值会大大低于其他网页。对于时间因素比较敏感的页面，如本文讨论的新闻页面，实时性是一个重要相关因素，因此不适合运用此算法。

HITS算法也是利用被链接网页的质量和数量来确定搜索结果的权重值。该算法将网页分为两种类型：一种是表达某一主题的权威页面，称为Authority页面，依赖于指向它的页面；另一种是能把Authority页面联结在一起的页面，称为Hub页面，依赖于它指向的页面。由于网页的这种分类，HITS算法也就涉及了两个重要的权值，Authority值和Hub值。Authority表示一个权威网页被其它网页所引用的加权数量，即该权威网页的加权入度值，若某网页被引用的数量越大，则该网页的加权入度值越大，Authority越大；Hub表示一个Web页面指向其它网页的加权数量，即该Web页面的加权出度值，它提供了指向权威页面的链接集合，某网页的加权出度值越大，则该网页的Hub值越大。通常情况下，好的Hub页面指向很多好的Authority页面；好的Authority页面也会有许多Hub页面所指向。HITS算法正是利用了页面间的这种关系，从而大大提高了搜索质量。

HITS算法的优点是它能更好地描述互联网的组织特点，由于它只是对互联网中很小的一个子集进行分析，所以它需要的迭代次数更少，收敛速度更快，减少了时间复杂度。但HITS算法也存在一些问题：第一，中心网页之间的相互引用以增加网页评价，某些网页可能通过恶意大量引用其它页面，来提高自身排名；第二，主题偏离，由于该算法只是根据链接关系确定权重，不会确认页面是否有效，忽略了页面链接之间的差异性，把所有的链接都看作具有同等的重要性，当查询很多广义主题时，HITS算法会把一些和搜索主题无关的链接给予很高的价值度，尤其当这些链接自身也是中心网页或权威网页时。例如，当用户输入查询关键词“电影人物”时，一些电影公司的网站主页很有可能会排在最前面，因为电影公司网站商业性存在的原因，许多链接会在这些公司网站的主页之间产生，从而导致“主题偏离”现象。

PageRank算法和HITS算法都是客观的描述了网页之间的本质特征，但它们都很少考虑到用户浏览习惯时的主题相关性。与这些算法相比，Direct Hit算法是一种重视信息质量和用户反馈的排序方法。它的基本思想是，系统将查询的结果提交给用户，接着开始跟踪用户在检索结果中的点击行为，如果网页被用户点开进行浏览，并且浏览的时间较长，则可认为该网页的受欢迎程度就高，相应地，系统将增加该网页的相关度。反之，若用户点开网页后停留时间较短，并且重新返回点击其他网页，那么认为该网页与查询词具有很小的相关性，系统将降低该网页的相关度。由此可见，采用这种算法，相关度在不断变化，对于同一个词在不同的时间进行检索，得到的排序结果可能不尽相同，网页的排名取决于被点击的次数和被浏览的时间长度，用户点击反馈的多少决定了算法的精确度，所以说Direct Hit算法是一种取决于用户检索行为的动态排序算法。

Direct Hit算法的优点是直接融入用户的反馈信息，考虑了用户的主观因素，能够保证页面的质量。然而缺点是，该算法只适用于检索关键词较少的情况，因为它实际上并没有进行排序，而是一种筛选和抽取。用户行为具有很强的随意性，在检索数据库很大、关键词很多的时候，返回的结果有很多页，用户不可能一条条查阅，多数用户只会关注前几页显示的内容，因此一些排名较后的网站获得的点击率就会很低，这样很难提升网页的排名。所以，Direct Hit算法不能作为主要的排序算法来使用，它是一种很好的辅助排序算法，目前在许多搜索引擎中仍然在使用。

分析与比较了多种相关度排序算法后，本系统采用的相关度排序方法是使用TF-IDF模型[13][14]。TF-IDF是一种统计方法，用以评估单词对于一个文件集或一个语料库中的其中一份文件的重要程度。TF-IDF模型的主要思想是：如果词w在一篇文档d中出现的频率高，并且在其他文档中很少出现，则认为词w具有很好的区分能力，适合用来把文章d和其他文章区分开来。信息检索时，对于每个文档，都可以分别计算一组搜索词的TF-IDF，将它们相加，就可以得到整个文档的TF-IDF。这个值最高的文档就是与搜索词最相关的文档。

TF-IDF算法的优点是简单快速，结果比较符合实际情况。缺点是单纯以“词频”衡量一个词的重要性，不够全面，有时重要的词可能在文章中出现的次数并不多。此外，这种算法无法体现词的位置信息，出现位置靠前的词与出现位置靠后的词都被视为重要性相同，这是不合理的。

## 5.3.3 建立索引的实现流程

Jieba中文分词是python的中文分词组件，它不仅具有分词功能，还具有添加自定义词典、关键词提取、词性标注等多种功能。首先利用jieba的关键词提取功能对每条新闻的正文进行关键词提取并计算其TF-IDF值，keywordsList = jieba.analyse.extract\_tags(content,withWeight = True)，content为新闻正文，即待提取的文本，withWeight为是否一并返回关键词权重值，将其设置为True。然后将每条链接—关键词—权值记录存储到数据库的索引表中。在用户输入搜索词后，先利用jieba分词功能的搜索引擎模式对用户输入的检索词进行分词处理，keywordslist = jieba.cut\_for\_search(inputSentence)，分词后得到n个关键词，然后从数据库的索引表中得到这n个关键词对应的结果序列，将这些序列中链接相同的记录的权值进行相加，最后将所有记录按总权值从大到小进行排列，即作为最终返回结果的顺序。

由一组关键词列表得到排好序的新闻条目列表的核心代码如下：

#根据keywordsList返回sorted linkslist

def getSortedList(keywordsList):

linkDict={}

for each in keywordsList:

print each

cursor.execute('''select link,wValue from indexTable where keyword=%s''',[each])

values = cursor.fetchall()

for each in values:

link = each[0]

wValue = each[1]

if link in linkDict.keys():

print "hi"

linkDict[link] = (linkDict[link] + wValue)\*100

print link

else:

linkDict[link] = wValue

sortedTuplelist = sorted(linkDict.items(),key=lambda a:a[1])

newlist = [i[0] for i in sortedTuplelist]

return newlist[::-1]

#派生HTMLParser基类,解析html，提取关键词

class myHtmlParser(HTMLParser.HTMLParser):

url = ''

newsId = ''

def handle\_starttag(self,tag,attr):

if tag == 'meta':

if ('name','keywords') in attr:

keywords = attr[1][1]

titleKeywordsList = jieba.analyse.extract\_tags(keywords,withWeight = True)

finalList = titleKeywordsList

if len(finalList) is not 0:

database.saveToIndexDB(self.url,finalList)

该部分运行截图如图5.3

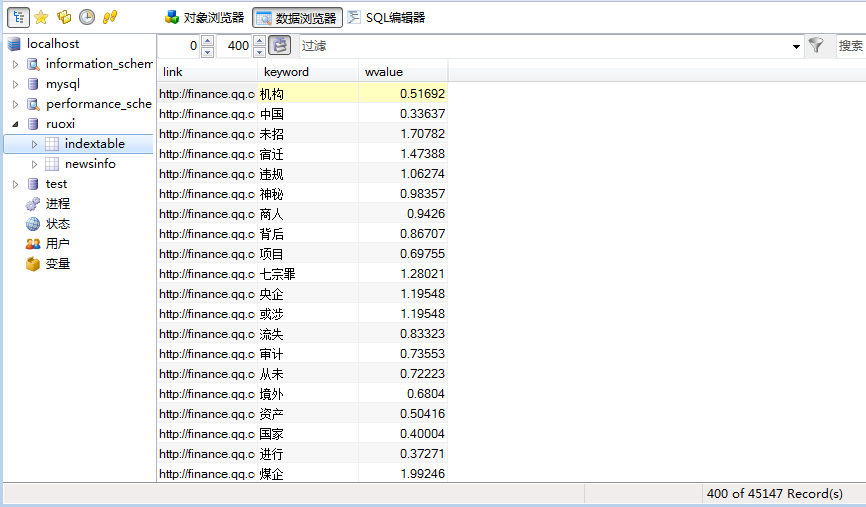


图5.3 索引子系统截图

## 5.4 搜索引擎框架搭建

## 5.4.1 B/S架构设计

本次客户端开发采用的是典型的B/S架构，服务器主要为客户端提供两个页面，一个是搜索框页面，用户在这个页面进行输入操作，还有一个是结果页面，服务器将搜索结果显示在这个页面。服务器要实现的核心业务逻辑仅包括根据用户输入的关键词来返回对应结果。

## 5.4.2 UI设计



图5.4 UI设计图

## 5.4.3 前端框架开发流程

开发客户端时，首先用easy\_install工具通过命令行安装好了flask框架，在Sublime Text中编写好服务器主程序后，测试运行了服务器端代码，在浏览器中，访问http://127.0.0.1:5000 成功访问到了服务器提供的“Hello Flask”页面。接下来编写系统主页面，即搜索入口，用户访问该地址时，服务器端将主页面返回给客户端。根据设计好的主页面，用html、css编写好主页面，将css文件、Jquery库等静态文件分别放入系统根目录的/static/css和/static/js文件夹下，此时访问该地址，便可以成功显示出主页。接下来要能够使服务器端接收到用户的请求参数，包括搜索关键词，搜索种类和排序方式。通过form表单形式向服务器端发送带参数get请求，然后在服务器端主程序中，编写好处理该请求的路由方法/search，根据参数得到对应的结果列表，然后根据结果列表，动态生成结果页面，生成结果页面的方式是通过Jinja2模板引擎，将结果列表传递给Jinja2模板引擎之后，Jinja2通过自身的循环控制代码及html代码，动态生成每一项结果，包括新闻标题、日期、简述和链接等。最后在Sublime Text中运行服务器主程序，即可通过浏览器访问该地址，使用到该系统的检索服务。

运行截图如图5.5



图5.5 程序运行截图

# 第6章 关键问题及解决方法

## 6.1 爬取子系统部分

为了突出新闻搜索引擎的特点，在爬取新闻信息时加入了新闻分类这个属性，用户在使用该系统时，可以先选择自己感兴趣的新闻类别，然后在所选类别下进行进一步的查找，这样可以缩小查询范围，节省查询时间，提高查询结果的准确性。

由于选取的新闻源来自腾讯和网易两个网站，因此爬取到的新闻时间格式不一致。腾讯的时间格式形如2015-01-01 13:52:20，网易的时间格式形如Thu，19Mar 2015 01:12:11 GMT，网易采用的是格林威治时间。通过上网查询到解决方法，首先要将格林威治时间转换成北京时间，先用切片选取网易时间GMT以外的部分，然后转换成时间戳，再转换成datetime格式，这样时间上就可以进行加减，让小时部分加8小时，从而转换成北京时间。最后统一格式，先用time.strptime()将时间的数字部分以元组形式存放，然后用time.strftime()函数将时间统一改成纯数字的形式，这样方便之后对新闻的时鲜度进行比较。

该模块进行测试时，发现爬虫有时候耗时较长，长时间无响应。这是因为某些网站自身原因或网速不稳定，对某些新闻链接的访问请求耗时较长，使程序长时间阻塞。为了防止这类链接破坏爬虫的工作效率，设置socket的超时时间为10s，若10s内未得到响应，则抛弃该链接，继续请求下一个链接。这样保证了爬虫的工作效率和稳定性。

## 6.2 索引子系统部分

在对系统进行测试时，输入单关键词和输入多关键词时检索到的结果几乎一致。打印权值之后发现，该关键词的权值异常的高。调试后发现这是因为索引表中的某些记录有大量重复。由于当前的索引子系统每次都会在爬虫爬取了RSS新闻后马上为它建立索引，然而很多RSS源上的新闻长期不更新，因此定期爬取时可能会多次爬取到该新闻，以至于多次为该新闻建立了索引，计算权值时也计算了多次，导致再对这些词进行检索时，第二个关键词的权值几乎不起作用。

对数据库进行去重操作后，再修改了索引部分的代码，确定爬取的新闻链接成功存入数据库时（表明之前未爬取过），然后再对新闻建立索引。问题解决后，输入多关键词时，检索的结果表现出很高的相关度。

## 6.3 web客户端部分

最开始开发前端页面时，将所有的html、css文件都放在了根目录下，通过使用Chrome浏览器的开发者工具调试后发现，服务器并没有将css文件发送给浏览器，查阅资料后了解到，flask框架使用的Jinja2模板引擎对文件路径有特殊规定，它会默认按照该规定进行路径搜索，静态文件必须放入根目录的static文件夹下，而js，images，css等静态必须放入文件指定的文件夹下才能被正确获取并发送给客户端。按如上所述建立了文件目录后，重新把所有文件放入指定文件夹内并修改了代码内相关路径后，问题成功解决。

在服务端运行程序时有时出现socket.error: [Errno 48] Address already in use的错误。这种错误的原因是因为出现了端口占用，因为flask运行时一直在监听5000端口，而上次程序没关闭时再次运行就会端口被占用，程序无法运行的状况，在命令行中通过“lsof -i tcp:5000”的命令找到占用5000端口的进程pid，然后“kill pid”杀死该进程，即可正常运行现有的服务器程序。

# 第7章 总结与展望

基于互联网和信息技术的发展，网络信息检索与分享具有重要影响，搜索引擎受到越来越多用户的关注，因此研究和实现一个搜索系统显得很有意义。本文研究了基于Web的新闻搜索技术，并在此基础上实现了一个Web新闻搜索平台。现对本文所做工作做个总结：

本文首先对搜索引擎的概念和分类进行介绍，特别是在新闻方面做了进一步了解。在系统开发时，查阅资料后采用抓取RSS新闻源的方法，从网上爬取新闻信息。然后在关键词提取技术上分析了TextRank算法和TF-IDF算法，同时分析与比较了多种排序算法，并联系当前互联网现状浅谈了这几种算法的优点和所面临的不足。最后对本次新闻搜索引擎系统的开发流程进行了详细的介绍。

由于时间和精力有限，本论文在对搜索引擎系统的研究中，只是对它的基本理论和基本框架进行了学习，所开发的新闻搜索引擎系统虽然具有新闻搜索的基本功能，但还有很大的改进空间：

1.没有实现索引与数据库数据自动更新功能。自动更新对每日都要进行数据更新的新闻搜索引擎来说是必不可少的，要在后期阶段进行实现。

2.对于用户个性化信息的获取方面，除了使用用户输入检索词外，还可以通过用户对页面的停留时间和链接的收藏等情况来获取用户兴趣偏好。

3.新闻分类方面，可将TF-IDF值和余弦相似度一同使用于向量空间模型中，以此判断两份文件之间的相似性。当然，除了使用文本分类技术进行分类外，还可以通过聚类技术对用户进行聚类，这样可以优化用户兴趣所属类别的划分结果。

4.相关度排序方面，TF-IDF算法中并没有体现出单词的位置信息，对于Web文档而言，权重的计算方法应该体现出HTML的结构特征。特征词在不同的标记符中对文章内容的反映程度不同，其权重的计算方法也应不同。因此应该对于处于网页不同位置的特征词赋予不同的系数，然后乘以特征词的词频，以提高文本表示的效果。

5.新闻的相关度排序很大程度上和用户平时的搜索行为与搜索爱好相关，因此，如果能考虑到用户的兴趣偏好与行为习惯，将能够更好的为用户展示出最匹配的内容。

总之，互联网的高速发展在给广大用户提供便利的同时，也给信息搜索领域带来了很大的挑战。为了满足用户日益变化的要求，相关领域的技术也需要不断改进与创新。搜索引擎系统的真正内涵是非常丰富的，还有许多问题需要解决，许多方法需要论证。深入研究和探讨相关内容也是我以后需要继续努力去做的工作。

# 参考文献

[1]朱健琛.个性化新闻搜索引擎的研究与设计[D].成都：成都电子科技大学，2012.

[2]张玉宝.中文搜索引擎的原理及相关技术分析[J].福建电脑，2008(1)：7.

[3]张文静.搜索引擎的分类及发展趋势[J].焦作大学学报，2006（3）: 76- 78.

[4]常晓燕.基于Java的新闻搜索引搜索的设计与实现[D]. 成都：西南交通大学，2004.

[5]王益明，刘菲.中文搜索引擎搜索结果重合率研究报告[R].2007.

<http://www.searchlab.com.cn/thesis.php>.

[6]赵敏涯.基于主题的新闻搜索引擎的研究与实现[D].扬州：扬州大学，2006.

[7]林子熠.基于模板的Web新闻搜索技术的研究与发现[D].上海：上海交通大学，2009.

[8]钱爱兵.基于RSS的Web新闻主题聚合系统的设计与实现[J].现代图书情报技术，2007(4).

[9]Martin C.Brown. Python：The Complete Reference[M]. McGraw-Hill Company，2001.

[10]Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]//Proceedings of EMNLP. 2004,4(4): 275.

[11]文军舰.基于Nutch的Web结构挖掘算法研究[D].天津：天津财经大学，2011.

[12]黄言之.浅谈搜索引擎的核心算法[Z].2013.<http://www.williamlong.info/archives/3526.html>.

[13]图灵社区.TF-IDF模型的概率解释[Z].2012.<http://www.ituring.com.cn/article/15381>.

[14]博客频道.tf-idf详解.2013[Z].<http://blog.csdn.net/july_2/article/details/9698755>.

[15]王启户.面向远程教育的搜索引擎系统设计与实现[D].西安：西安电子科技大学，2008.

[16]尹浩.基于WWW的新闻搜索引擎的设计与实现[D].成都：西南交通大学，2003.

[17]王晶.基于Web信息获取的新闻数据分析研究[D].上海：华东师范大学，2009.

[18]徐宝文，张卫丰.搜索引擎与信息获取技术[M].第1版.北京：清华大学出版社，2003.

[19]胡小睿.基于Web挖掘的搜索引擎技术研究[D].武汉：武汉大学，2005.

[20]林伟业.基于爬虫的Sohu新闻搜素引擎设计与实现[D].广州：中山大学，2012.

[21]曾小芹.基于领域本体的新闻搜索引擎的研究与实现[D].南昌：南昌大学，2012.

[22]Christos Bouras，Vassilis Poulopoupos，Panagiotis Silintziris. Personalized News Search in WWW：Adapting on user’s behavior. [Fourth International Conference on](http://61.183.148.149:8000/rewriter/IEEE/http/hdddwoknqd9hddd9nqf/xpl/mostRecentIssue.jsp?punumber=5072479) Internet and Web Applications and Services[Z],2009.

# 致谢

时光飞逝，大学四年的学习生涯即将结束，首先我要向我的母校——武汉理工大学致以诚挚的谢意。在这里，我度过了人生中一段美好而又难忘的时光。值此论文完成之际，我要感谢我的导师岑丽副教授，论文从选题、撰写、修改到定稿，系统从开发、调试、修改到优化，每一步都得到了岑老师的精心指导，这对我毕业设计的顺利完成起到了重要的作用。同时，谨向所有关心我学业的老师、同学、朋友表示衷心的感谢。

本论文从开题到完成经历了一个学期，看着这厚厚的文稿，轻松之余，更多的是一种由衷的感动。完成毕业设计过程中，岑老师对我的论文提出了许多宝贵的建议，我遇上难题向她询问时，她也耐心解答。岑老师渊博的知识，严谨的治学态度，优良的工作作风，谦逊的人格给我留下了深刻的印象，这会是我一生学习的典范。除了学业的悉心教导和严厉鞭策外，生活上岑老师也经常关心我们。有幸成为岑老师的一名学生，是我一生受用不尽的财富。

我要感谢计算机学院的领导和老师们。作为软件工程的学生，他们传授的宝贵知识是使我论文得以顺利完成的坚实基础，我们学院丰富的教学资源开拓了我的眼界。这一切都是我今后从事专业工作的强力后盾。然后，我要感谢四年里和我朝夕相处的同学们，我们从陌生到相识相知，一起上课、一起学习讨论、一起生活，这是一种缘分。他们给予的真挚的友谊、真心的关爱、真诚的鼓励、真切的微笑也会伴随我的一生。

特别要感谢我的父母，感谢他们对我从小到大的关爱与教诲，对我学习工作的支持与鼓励，对我无知幼稚行为的宽容，感谢他们一直以来的无私奉献与不求回报。在学业完成之际，向含辛茹苦的父母表示由衷的谢意和崇高的敬意。

面对如此之多的关心与帮助，常感无可回报，唯有在未来的工作和生活中时时敦促自己，取得更加丰富的成果，才不负亲人、师长和朋友的殷殷期望之情和谆谆教诲之意！