

# 深圳大学

## 本科毕业论文（设计）

题目： 个人搜索引擎的实现

姓名： 蔡汉锦

专业： 计算机科学与技术

学院： 计算机与软件学院

学号： 2007170019

指导教师：

职称：

2011 年 5 月 5 日

## 深圳大学本科毕业论文（设计）诚信声明

本人郑重声明：所呈交的毕业论文（设计），题目《个人搜索引擎的实现》是本人在指导教师的指导下，独立进行研究工作所取得的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明。除此之外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。本人完全意识到本声明的法律结果。

毕业论文（设计）作者签名：

日期：                年    月    日

# 目 录

摘要(关键词).....	1
1 引言.....	1
1.1 研究背景及意义.....	1
1.2 本课题主要工作.....	1
2 数据提取技术.....	1
2.1 Microsoft Documents 内容提取.....	2
2.2 PDF 文档内容提取 .....	3
2.3 网页数据提取技术.....	4
3 全文索引与搜索技术.....	6
3.1 全文检索.....	6
3.2 Lucene 全文检索类库.....	6
3.3 中文分词.....	7
3.4 IKAnalyzer 中文分词类库.....	7
3.5 索引与搜索核心算法.....	7
4 数据库连接技术.....	8
4.1 JDBC.....	8
4.2 本实验对数据库连接的封装.....	9
5 实验环境.....	9
6 本地文件全文搜索实验方案.....	10
6.1 系统简介.....	10
6.2 实现方案介绍.....	11
6.3 工程模块各类关系介绍.....	12
6.4 程序运行界面.....	13

7	网络数据全文搜索实验方案 .....	13
7.1	系统简介.....	13
7.2	校内公文通检索.....	14
7.2.1	系统简介.....	14
7.2.2	实现方案介绍.....	14
7.2.3	工程模块各类关系介绍.....	15
7.2.4	程序运行界面截图.....	16
7.3	腾讯微博相关数据检索.....	17
7.3.1	系统简介.....	17
7.3.2	实现方案介绍.....	17
7.3.3	工程模块各类的关系介绍.....	20
7.3.4	工程模块运行界面.....	22
7.4	人人网日志相关信息检索.....	23
7.4.1	系统简介.....	23
7.4.2	实现方案介绍.....	23
7.4.3	工程模块各类关系介绍.....	24
7.4.4	工程模块运行界面截图.....	25
8	结束语 .....	26
	参考文献.....	27
	致谢.....	28
	Abstract(Key words).....	29

**（目录为小四宋体，1.5 倍行距，目录生成到三级标题，英文和数字为 Times New Roman，小四。温馨提示：本目录为原文自动生成，目录页码为原文页码，与本文后续文本及页码并不对应，仅供格式参考。）**

# 个人搜索引擎的实现（论文标题华文中宋，小二，居中，单倍行距，段前段后 0.5 行）

计算机与软件学院计算机科学与技术专业 蔡汉锦

学号：2007170019

**【摘要】**针对通用搜索引擎无法访问内部网、SNS 网络和个人电脑文档信息的问题，本文提出并实现了基于 Lucene 与 IKAnalyzer 的个人搜索引擎。该搜索引擎主要实现的功能有：(1) 个人电脑文档信息检索。通过对 txt、word、excel、pdf 文档的正文提取，建立统一的索引。实现具有图形界面的本地文件搜索系统。(2) 封闭网络（SNS、微博、内部网络）的信息检索。突破用户权限，实时加载网页，分析网页结构并提取内容，创建索引并本地保存数据。通过实时索引与本地数据库索引相结合的方式，解决了网络空间庞大数量造成搜索缓慢的问题。以 web 的形式开发了集校内公文通、腾讯微博、人人网日志信息的一站式全文检索系统。从实际运行效果分析，该搜索系统弥补了通用搜索引擎的不足，能够满足基本的个性化搜索的需求。（楷体\_GB2312，5 号，单倍行距，段前段后 0.5 行）

**【关键词】** Lucene；网页内容提取；全文索引；IKAnalyzer （楷体\_GB2312，5 号，单倍行距，段前段后 0.5 行）



# 1 引言（一级标题三号黑体加粗，单倍行距、段前段后 0.5 行）

## 1.1 研究背景及意义（二级标题小三号黑体加粗，单倍行距、段前段后 0.5 行）

搜索技术能够为用户提供信息检索，网址导航的功能，是现在网络用户访问互联网的最主要方式。通用搜索引擎能够提供一站式的信息服务，但是存在返回结果不准确，专业性不深，个性化不强的缺点。不同于通用搜索和垂直搜索引擎，我们提出从个人信息环境出发的个人搜索引擎。个人环境最主要的即包括个人电脑文档信息和个人用户常接触的网络信息，其中最典型的个人网络环境有 SNS 网络<sup>[1]</sup>，常关注的博客网络，单位内部网络和微博信息。用户电脑数据和个人信息网络往往是其它搜索引擎无法涉及到的信息孤岛<sup>[2]</sup>。而这些信息对用户来说是最重要也是最常用的数据，因此提出个人搜索引擎具有重要的应用价值和研究意义。（正文内容五号宋体，单倍行距、段前段后 0.5 行，需指出所引用文献，用上标数字表示<sup>[1][2]</sup>，）

## 1.2 本课题主要工作

本文主要解决的问题包括兼容各种文档格式，提取文档正文和关键字，建立各种文档格式的索引，突破用户权限实现社会网络、微博、内部网信息一站式搜索。最后开发出一套集成桌面搜索和网络搜索的个人搜索引擎。满足用户对本地文件的全文检索及对校内公文通、腾讯微博、人人网日志信息的全文检索。

# 2 数据提取技术

## 2.1 Microsoft Documents 内容提取

### 2.1.1 Apache POI 类库（三级标题四号黑体加粗，单倍行距、段前段后 0.5 行）

Apache POI 是 Apache 软件基金会的开放源码函式库，POI 提供 API 给 Java 程式对 Microsoft Office 格式档案读和写的功能。

其结构如下：

HSSF — 提供读写 Microsoft Excel 格式档案的功能。

XSSF — 提供读写 Microsoft Excel OOXML 格式档案的功能。

HWPf — 提供读写 Microsoft Word 格式档案的功能。

HSLF — 提供读写 Microsoft PowerPoint 格式档案的功能。

HDGF — 提供读写 Microsoft Visio 格式档案的功能。

### 2.1.2 Microsoft Word 内容提取

.....

### 2.1.3 Microsoft Excel 内容提取

.....

## 2.2 网页数据提取技术

### 2.2.1 正则表达式

正则表达式（英语：Regular Expression、regex 或 regexp，缩写为 RE），也译为正规表示法、常规表示法，在计算机科学中，是指一个用来描述或者匹配一系列符合某个句法规则的字符串的单个字符串。在很多文本编辑器或其他工具里，正则表达式通常被用来检索和/或替换那些符合某个模式的文本内容。许多程序设计语言都支持利用正则表达式进行字符串操作。

## 3 全文索引与搜索技术

### 3.1 全文检索

.....

功能上全文检索引擎<sup>[7]</sup>需要具有建立索引，处理查询返回结果集，增加索引，优化索引结构等功能。结构上具有索引引擎，查询引擎，文本分析引擎和对外接口等。

目前，实现全文信息检索有两大基本方案，词索引和字索引。

字索引<sup>[8]</sup>，以汉语单字为索引单位的检索算法。这种方法往往会引起多查的错误。

词索引<sup>[9]</sup>，以单词为索引单位的检索算法。西文又是以单词为语言要素，每个西文单词之间都有一个空格。因此，在对全文数据库建立索引的时候，按照单词划分建立索引，是既简单又自然的。我国最开始引入全文检索技术的时候，是汉化西文的数据库系统，因此也就自然使用了词索引技术。但由于中西文环境中语素的不同特点，使得中文全文信息检索必须要解决分词的问题。

.....

### 3.2 Lucene 全文检索类库

.....

Lucene<sup>[12]</sup>不是一个完整的全文索引应用，而是一个用 Java 写的全文索引引擎工具包，它可以方便的嵌入到各种应用中，实现针对应用的全文索引/检索功能。

大部分的搜索引擎都是用 B 树<sup>[13]</sup>结构来维护索引，索引的更新会导致大的 FO 操作，Lucene 在实现中，对此稍微有所改进：不是维护一个索引文件，是在扩展索引的时候不断创建新的索引文件，然后定期的把这些新的小索引文合并到原先的大索引中(针对不同的更新策略，批次的大小可以调整)，这样不影响检索的效率的前提下，提高了索引的效率。

Lucene 采用倒排索引算法建立索引<sup>[14]</sup>，主要包括索引类 (IndexWriter)、文档对象类 (Document) 和信息字段类 (Field)。

.....

### 3.3 中文分词

.....

### 3.4 IKAnalyzer 中文分词类库

.....

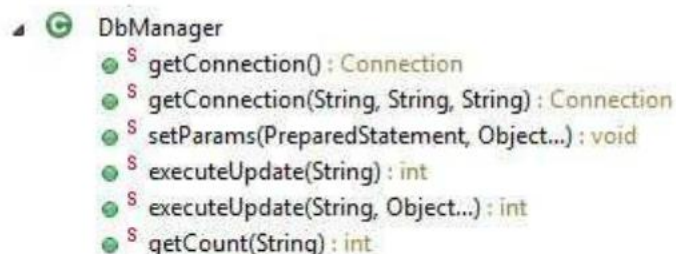
## 4 数据库连接技术

### 4.1 JDBC

.....

### 4.2 本实验对数据库连接的封装

本次实验在 JDBC 的基础写了一个管理数据库的封装类。封装了与数据库连接方法，数据库数据更新方法，类结构如图 1 所示。



```

DbManager
  S getConnection() : Connection
  S getConnection(String, String, String) : Connection
  S setParams(PreparedStatement, Object...) : void
  S executeUpdate(String) : int
  S executeUpdate(String, Object...) : int
  S getCount(String) : int
  
```

图 1 数据库管理类

(图标在图下，小五号、宋体，按顺序标号，居中，并在正文中引用，图中文字均为小五号)



## 5 实验环境

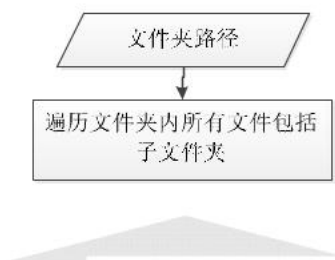
.....

## 6 本地文件全文搜索实验方案

### 6.1 系统简介

.....

### 6.2 实施方案介绍



本部分非原文内容，仅用于说明表标注格式。

在现代无线通信中，数据都是以数据包装的方式来进行传输的。对 NRF9E5 这样的无线片上系统，每次发送/接收数据也都是以数据包装的方式来进行的。数据包格式是通信协议的重要部分，NRF9E5 的无线数据包格式如表 1 所示<sup>[8]</sup>。

表 1 氨标准系列

(表标题在表上，小五号宋体，按顺序标号，居中，并在正文中引用，表中文字为小五号)

管号	标准工作液 ml	吸收液 ml	氨含量 $\mu\text{g}$
0	0	10.00	0
1	0.25	9.75	0.25
2	1.00	9.00	1.00
3	3.00	7.00	3.00
4	5.00	5.00	5.00
5	7.00	3.00	7.00
6	10.00	0	10.00

## 7 网络数据全文搜索实验方案

腾讯微博检索的设计方案如图 10 所示。

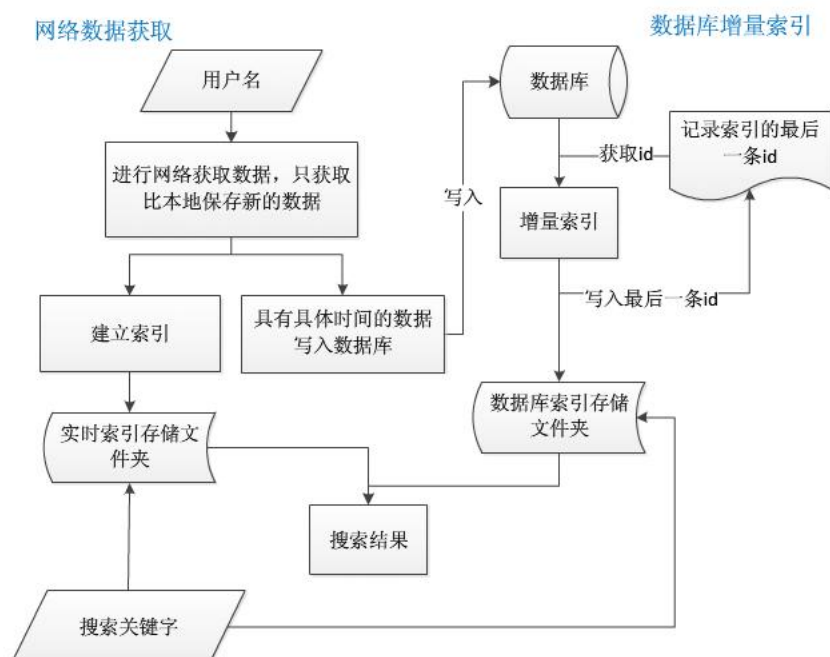


图 10 腾讯微博检索-设计方案示意图

### 7.4.4 工程模块运行界面截图

。。。。。

## 8 结束语

本次实验基于 Lucene 全文检索库与 IKAnalyzer 中文分词库，针对本地文件，实现了常用格式文件的全文检索系统。

通过穿越网络个人空间权限，对其网页内容进行提取，并进行索引，实现了对校内公文通系统的全文检索，腾讯微博的全文检索，人人网日志相关信息的全文检索系统。

在后续的工作中，将对网络个人空间的检索系统进行优化，并从数据来源上进行扩充，使之成为信息量大，效率高的检索系统，最后可以为社区网络数据挖掘提供更好的帮助。

**【参考文献】(楷体-GB2312, 5号, 加粗)**

- [1] 袁梦倩. 论 SNS 新型社交网络的传播模式与功能—基于“校内网”的现象研究[J]. 今传媒, 2009 (4): 78-80.
- [2] 王俊杰. 冲出信息孤岛, 实现数字资源共享[J]. 大学图书馆学报, 2004 (3): 16-18.
- [3] 熊雨前, 徐红轮. PDF 技术及应用. 数字与缩微影像. 2011 (1): 37-41.
- [4] 李效东, 顾毓清. 基于 DOM 的 Web 信息提取[J]. 计算机学报, 2002, 25 (5): 526-533.
- [5] 董强, 管国栋. 基于 DOM 的 Web 信息抽取方法研究[J]. 舰船防化, 2006, 3: 26-30.
- [6] 周源远, 王继成, 郑刚, 张福炎. Web 页面清洗技术的研究与实现[J]. 计算机工程, 2002, 28 (9): 48-50.
- [7] 李刚, 宋伟, 邱哲. 征服 Ajax + Lucene 构建搜索引擎[M]. 北京: 人民邮电出版社, 2006.
- [8] Yuejie Zhang, Tao Zhang, Shijie Chen. Research on Lucene-based English-Chinese Cross-Language Information Retrieval [J]. Journal of Chinese Language and Computing. 2005, 15 (1) : 25-32.
- [9] Wu Z M. Tseng Ct Chinese Text Segmentation for Text Retrieval: Achievements and Problems[J]. Journal of the American Society for Information Science, 1993, 44 (9): 532-542
- [10] 王瑞雷, 栾静, 潘晓花. 一种改进的中文分词正向最大匹配算法[J]. 计算机应用与软件, 2011, 28 (3): 195-197.
- [11] Liang Zhen, Li Yu-sheng. Reverse Backtracking Research of Chinese Segmentation Based on Dictionary of Hash Structure[C]. Information Technology and Computer Science (ITCS), 2010 Second International Conference, 2010: 265-267.
- [12] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. Lucene in Action[M]. American: MANNING PUBN, 2010.
- [13] Chien Lee-Feng. PA T-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval[J]. Information Processing and Management. 1999 (35): 501-521.
- [14] 隋丽萍, 徐承韬, 李瑞芳. 一个中文全文检索系统的设计与实现[J]. 科技资讯. 2007, 18: 244-245.
- [15] 周登朋, 谢康林. Lucene 搜索引擎[J]. 计算机工程, 2007, 33 (18): 95-118.
- [16] 房志峰. 中文搜索引擎中的分词技术研究[J]. 科学技术与工程, 2008, 8 (9): 2481-2483
- [17] 付年钧, 彭昌水, 王 慰. 中文分词技术及其实现[J]. 软件导刊, 2011, 10 (1): 18-20
- [18] linliangyi2007. 发布 IK Analyzer 3.2.8 for Lucene3.X.  
<http://linliangyi2007.iteye.com/blog/941132> (IK Analyzer 原作者官方博客)

(摘要小五号楷体, 单倍行距、段前段后 0.5 行, 图书在文献后加[M], 期刊加[J], 会议论文加[C]等, 参考文献不少于 10 篇, 其中外文文献不少于 2 篇)



## 致谢

首先衷心地感谢\*\*\*老师。在我大学生涯里，给予了很多指导与帮助。本次毕业设计，从选题到论文撰写，给予我很多宝贵的意见。\*\*\*渊博的学识、严谨的治学态度及认真负责的工作态度都使我受到鼓舞和熏陶。在此向\*\*\*老师表示崇高的敬意和衷心的感谢。

感谢研究生刘联东师兄。从毕业设计到毕业论文的撰写，刘联东师兄都给了我很大的帮助，在平时的讨论中给我提出了许多宝贵的意见和建议。他认真负责、热衷钻研的工作学习态度都使我受到很大的激励。在此，向刘联东师兄表示衷心的感谢。

感谢四年以来老师们的辛勤授课，丰富了我们的知识，拓宽了我们的视野，提高了我们发现问题、解决问题的能力，使我的思想产生了质的飞跃。

感谢开源社区，提供了很多开放源码与类库，为我提供了庞大的优秀代码资源，使我在程序开发过程中得到很多启发。

感谢一直关心我、支持我的父母和朋友们。

## To Develop Personal Search Engine

**【Abstract】** This paper proposes and implements personal search engine based on Lucene and IKAnalyzer for the deficiency of general search engines which fail to reach some isolated websites, such as internal websites of organizations and social network for the security factors. For the same reason, information of documents stored in personal computer also cannot be searched. The main functions of this system can be described as: Firstly, document information retrieval on PC. By extracting text from TXT, Word, Excel and PDF documents, a unified index is created. And the locale file search engine is implemented with a graphical interface. Secondly, retrieval system for information of closed network (SNS, microblog, internal network). By breakthrough user rights, loading webs real-time, analyzing the structure of pages and extracting text, an index is created and saved to local database. This paper solved the slowness problems due to the huge amount of data in network space by combining the real-time network index and local database index. The retrieval system is developed which integrated school board, Tencent microblog and the notes of Renren.com using Java web. In addition, this system improves the deficiencies of general search engines and satisfies the basic needs of personalized search from the results of analysis.

**【keywords】** Lucene; Web Content Extraction; Full-text Index; IKAnalyzer

指导教师: \*\*\*