

# 基于 Lucene 的全文检索引擎研究与应用

林碧英, 赵 锐, 陈良臣

(华北电力大学 计算机科学与技术学院, 北京 102206)

**摘 要:**快速有效地索引企业累积的大量的信息资源,是提供高质量检索服务的基础。Lucene 是一个用 Java 写的全文索引引擎工具包,访问索引时间快,支持多用户访问,可以跨平台使用。文中研究了 Lucene 系统结构和数据流,分析了 Lucene 的索引文件格式,实现了一个基于 Lucene 文档检索的应用实例。

**关键词:**全文检索;索引;应用研究/Lucene

**中图分类号:**TP391.3

**文献标识码:**A

**文章编号:**1673-629X(2007)05-0184-03

## Research and Application of Full Text Search Engine Based on Lucene

LIN Bi-ying, ZHAO Rui, CHEN Liang-chen

(School of Computer Science & Technology, North China Electric Power University, Beijing 102206, China)

**Abstract:** Rapid accumulation of large enterprises effectively indexing information resources is to provide high-quality search services. Lucene is a full text indexing engine written in Java toolkit, visit indexing time fast, multi-user support visits can cross-platform use. Study Lucene system structure and data flow, analyses the Lucene index format of the document to a file based on Lucene search application examples.

**Key words:** full-text search; indexing; applied research /Lucene

### 0 引 言

随着计算机技术及网络技术的迅速发展,电子文档数目急剧膨胀,在这海量的信息里面快速、全面、准确地查找所需要的资料信息已经成了人们关注的焦点,也成了研究领域内的一个热门课题。目前,信息检索技术的最新应用是最近国内外公司相继推出的桌面搜索引擎,这是集成信息检索技术的典型代表。

信息检索的核心技术是全文检索技术。全文检索是以各种计算机数据诸如文字、声音、图像等为处理对象,提供按照数据资料的内容而不是外在特征来实现的信息检索手段<sup>[1]</sup>。在索引中创建一个包含一系列用户搜索条件的查询,它能帮助人们进行大量文档资料的整理和管理工作,并使人们能够快速方便地查到他们想要的任何信息。Lucene 是一个用 Java 写的全文检索引擎下工具包,可以方便地嵌入到各种应用中实现针对应用的全文索引/检索功能,而不是一个完整的全文检索应用。

### 1 基于 Java 的全文检索引擎—Jakarta Lucene

最初的 Lucene 是使用 Java 语言编写的一个全文索引的工具包,支持多种操作系统。随着 Lucene 的逐渐发展,2001 年年底 Lucene 成为 apache 基金会有一个子项目。并在日前推出使用 C、Delphi 等其他语言编写的版本。目前有很多 Java 项目使用 Lucene 作为其后台全文检索引擎,著名的有: Eclipse: 功能强大的 IDE 工具,全文检索部分使用 Lucene; Jive: Web 论坛系统; Conoon: 基于 XML 的 Web 发布框架,全文检索部分使用<sup>[2]</sup>。

Lucene 作为一个全文检索引擎,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。

收稿日期:2006-08-20

基金项目:中国下一代互联网示范工程(CNGI)移动奥运资助项目(CNGI-04-17-2A)

作者简介:林碧英(1955-),女,湖南长沙人,教授,硕士生导师,研究方向为网络与信息安全。

## 2 Lucene 系统结构组织

### 2.1 Lucene 系统结构

Lucene 作为一个优秀的全文检索引擎,其系统结构具有强烈的面向对象特征。首先是定义了一个与平台无关的索引文件格式,其次通过抽象将系统的核心组成部分设计为抽象类,具体的平台实现部分设计为抽象类的实现,此外与具体平台相关的部分比如文件存储也封装为类,经过层层的面面向对象式的处理,最终达成了一个低耦合高效率,容易二次开发的检索引擎系统<sup>[3]</sup>。

图 1 是 Lucene 系统结构与源码组织图。

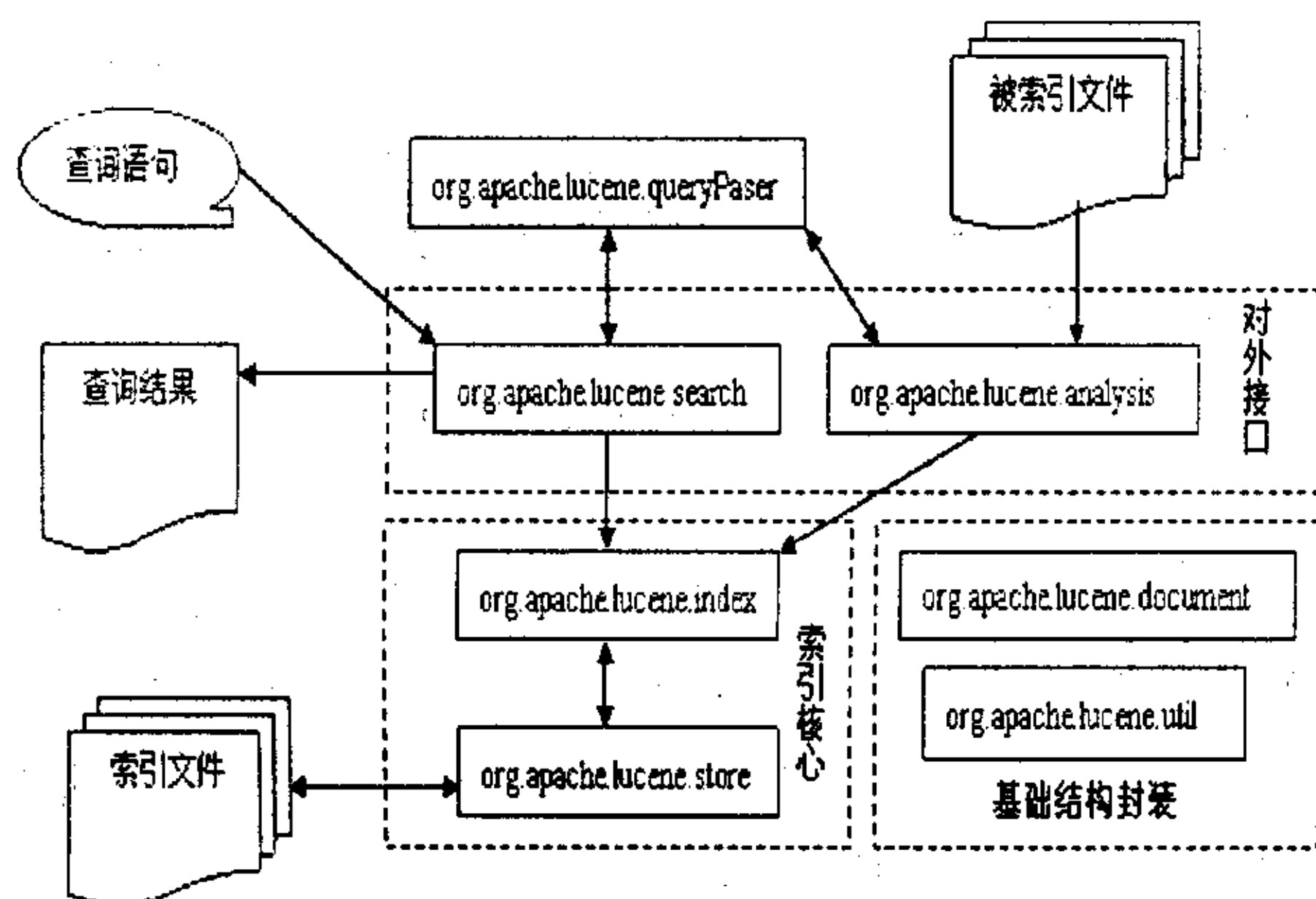


图 1 Lucene 系统结构与源码组织图

图 1 中可以清楚看到, Lucene 的系统由基础结构封装、索引核心、对外接口三大部分组成。其中直接操作索引文件的索引核心又是系统的重点。Lucene 将所有源码分为了 7 个模块(在 Java 语言中以包即 package 来表示),各个模块所属的系统部分也如上图所示。需要说明的是 org.apache.lucene.queryParser 是作为 org.apache.lucene.search 的语法解析器存在,不被系统之外实际调用,因此这里没有当作对外接口看待,而是将之独立出来。

### 2.2 Lucene 的数据流

理解 Lucene 系统结构的另一个方式是去探讨其中数据流的走向,并以此摸清楚 Lucene 系统内部的调用时序。在此基础上,能够更加深入地理解 Lucene 的系统结构组织,以方便以后在 Lucene 系统上的开发工作。对 Lucene 的数据流的分析,是深入 Lucene 系统的钥匙,也是进行重写的基础<sup>[4]</sup>。

Lucene 系统中的主要的数据流以及它们之间的关系图如图 2 所示。

图 2 很好地表明了 Lucene 在内部的数据流组织情况,并且沿着数据流方向也可对 Lucene 内部的执行时序有一个清楚的了解。现在将图中涉及到的流的类型与各个逻辑对应系统的相关部分的关系说明一下。

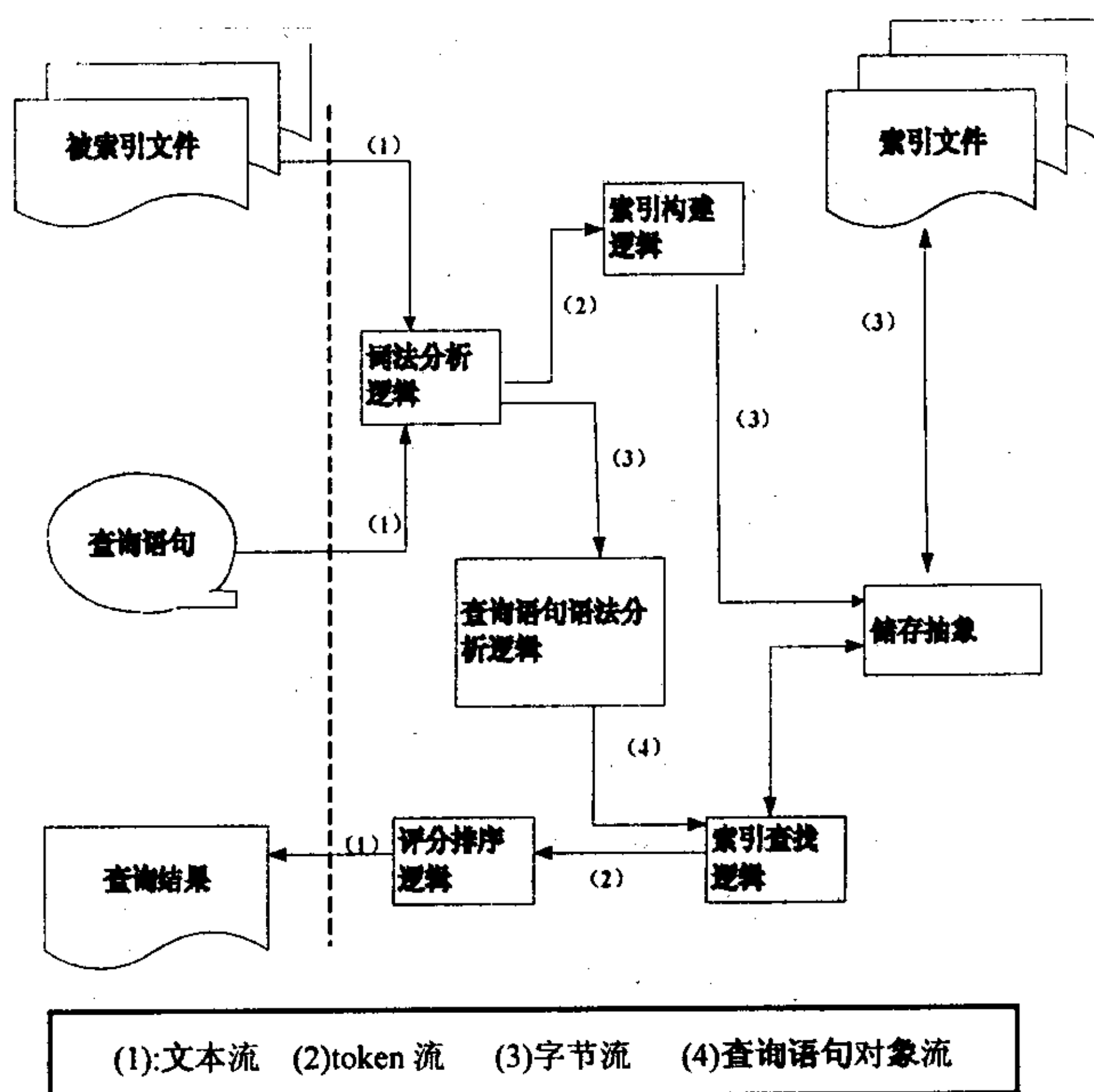


图 2 Lucene 数据流图

图 2 中共存在 4 种数据流,分别是文本流、Token 流、字节流与查询语句对象流。文本流表示了对于索引目标和交互控制的抽象,即用文本流表示将要索引的文件,用文本流向用户输出信息;在实际的实现中, Lucene 中的文本流采用了 UCS-2 作为编码,以达到适应多种语言文字的处理的目的。Token 流是 Lucene 内部所使用的概念,是对传统文字中的词的概念的抽象,也是 Lucene 在建立索引时直接处理的最小单位;简单地讲 Token 就是一个词和所在域值的组合。字节流则是对文件抽象的直接操作的体现,通过固定长度的字节(Lucene 定义为 8 比特位长)流的处理,将文件操作解脱出来,做到了与平台文件系统的无关性。查询语句对象流则是仅仅在查询语句解析时用到的概念,它对查询语句抽象,通过类的继承结构反映查询语句的结构,将之传送到查找逻辑来进行查找的操作。

## 3 Lucene 索引文件格式分析

在 Lucene 的 Web 站点上,有关于 Lucene 的文件格式的规范,其规定了 Lucene 的文件格式采取的存储单位、组织结构、命名规范等内容,但它仅仅是一个规范说明,并没有从实现者角度来衡量这个规范的实现<sup>[5]</sup>。因此,以下内容,结合了我们自己的分析与文件格式的定义规范,以期望给出一个更加清晰的文件格式说明。

首先在 Lucene 的文件格式中,以字节为基础,定义了数据类型,由于它们都以字节为基础定义而来,因此保证了是平台无关,这也是 Lucene 索引文件格式平台无关的主要原因。接下来是 Lucene 索引文件的概念组成和结构组成。

图 3 就是 Lucene 的索引文件的概念结构。Lucene 索引 index 由若干段(segment)组成,每一段由若干的文档(document)组成,每一个文档由若干的域(field)组成,每一个域由若干的项(term)组成。项是最小的索引概念单位,它直接代表了一个字符串以及其在文件中的位置、出现次数等信息。域是一个关联的元组,由一个域名和一个域值组成,域名是一个字符串,域值是一个项,比如将“标题”和实际标题的项组成的域。

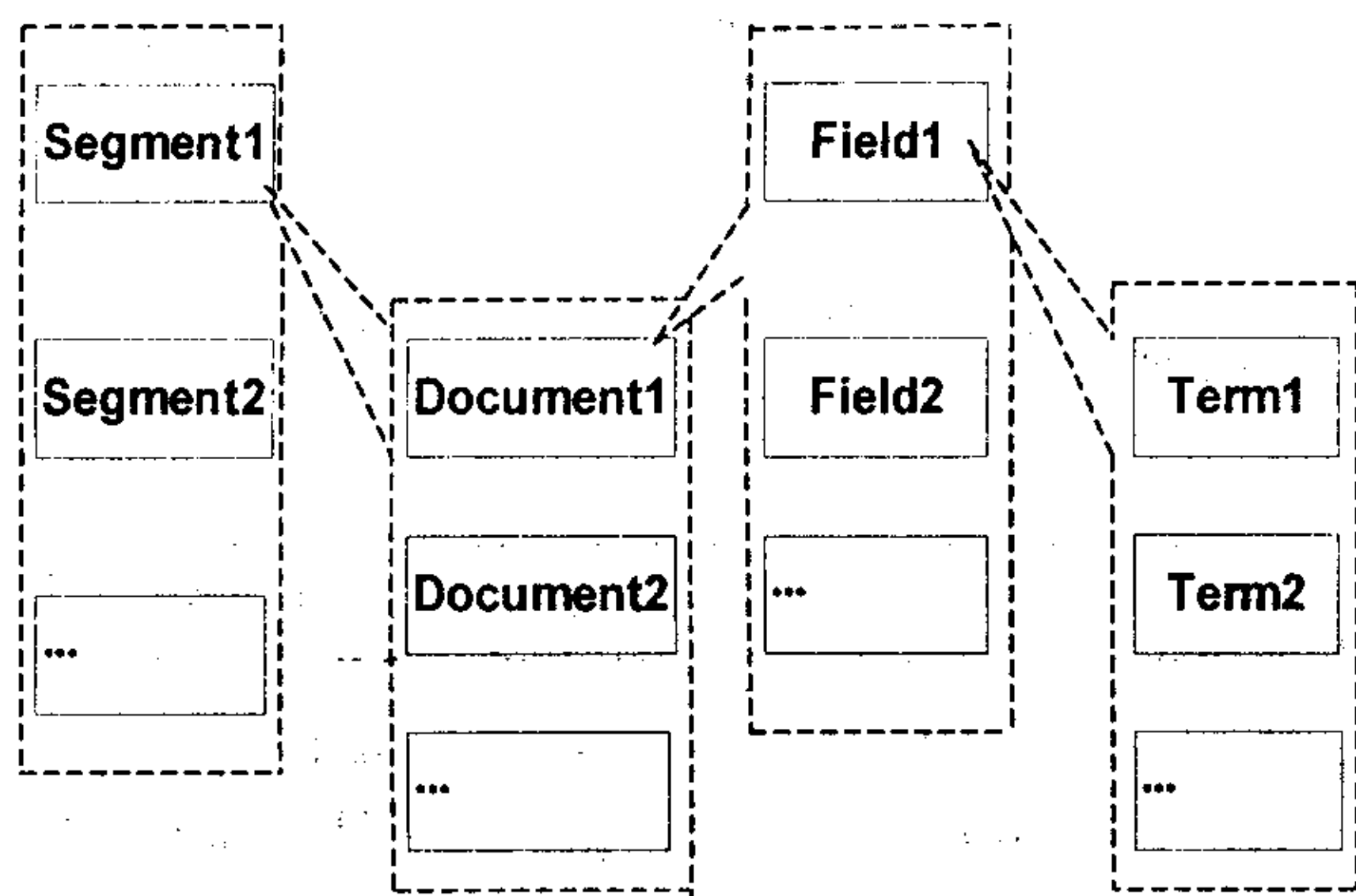


图 3 Lucene 的索引文件的概念结构

文档是提取了某个文件中的所有信息之后的结果,这些组成了段,或者称为一个子索引。子索引可以组合为索引,也可以合并为一个新的包含了所有合并内部元素的子索引。Lucene 的索引结构在概念上即为传统的倒排索引结构<sup>[6]</sup>。

#### 4 支持全文检索的一个应用的实现

对于一般的企业,长期使用计算机处理事物,形成了海量的文档,这些文档是企业的财富,如何利用好企业的这些文档显得尤为重要,由于企业没有对历史文档进行有效的索引管理,企业查找自己的资料很不方便而使用网络上提供的搜索引擎查找的资料多是网页资料,有效资料的命中率很低。如果对企业的历史资料以及现有的资料进行索引一类的管理,将会很好解决上面提到的问题。

Lucene 本身只是一个组件,若想让 Lucene 真正起作用,还得在 Lucene 基础上进行必要的一次开发。下面提到的方案是对 Lucene 的一个应用研究,中小企业可以参照此例来设计自己的文档检索应用。

对于 Lucene 组件包,为了能够支持中文,要进行修改首先将改写后支持中文的分析包 Lucene-cn.jar 加入到发布包 Analysis 包中,解开 Lucene-\*.zip,在解开的目录 src\demo\src\apache\Lucene\demo 下打开 Index-HTML.java。在第一处“import org.apache.lucene.analysis.standard.StandardAnalyzer;”下面加一行“import org.apache.lucene.analysis.cn.

ChinPSeAnalyzer;”,把第 2 处“writer=new IndexWriter(index,new StandardAnalyzer(),Create);”注释掉,换成“writer=new IndexWriter(index,new ChinPSeAnalyzer(),Create);”解开 Luceneweb.war,释放出 configuration.jsp 和 result.jsp 还有 web.xml 编辑 configuration.jsp,找到 indexLocation 这个变量 m,赋值成“/index”(或者用户自己建立的索引的目录名称);编辑成 result.jsp,找到“Analyzer analyzer=new StopAnalyzer();”,删除或者注释掉,改成“Analyzer analyzer=new org.apache.lucene.analysis.cn.ChinPSeAnalyzer();”然后,在 import 的标签下面加上:

```

<%
StringCONTENTTYPE="/text/html;charset=utf-8;
response.setHeader("/Content2Type0,CONTENT-
TYPE)
request.setCharacterEncoding("/utf-8)
%>
  
```

为了解决编码问题,需要修改 web.xml,在 webapp 这个元素中加上属性 character encoding“utf-8”(即<web-app character encoding="utf-8">)。这样就可以与 jsp 和具体的编码无关。

为了能用中文关键词搜索,修改 result.jsp 中获取表单的参数语句,进行编码转换将 queryString=request.GetParamter("query");替换为 queryString=new String(requestGetPara-.GetParameter("query").getBytes("iso8859-1"));为了要正确地实现界面显示汉字还要在每个 jsp 文件的行首加入<%@page contentType="text/html;charset=GB2312"language="java"%>,这样搜索界面和结果界面都会成为中文界面具体实现检索,首先要对文档源进行索引,索引命令为 DOS 命令:

```

d:\doc\doc1>java org.apache.
lucene.demo.Index-HTML-create-index d:\study-
Lucene\HtmlIndex.
  
```

该命令执行后,会对与工作目录 d:\doc\doc1 同级的目录(d:\doc\doc2;d:\doc\doc3)内(含以下工作目录)的文档\*.txt,\*.html,\*.htm 进行索引,并在生成索引库文件后存放在 d:\studyLucene\HTMLIndex 下,检索器在检索时候就会访问到这些库的文件。

运行随带的检索界面 http://localhost:8080/Lnce-neweb,在界面上输入关键词,点 Search 即可以进行检索查询,可以将此连接做到企业的主页上,这样

(下转第 190 页)

2 片绿化带与居民区相接,利用度量参数可以区分这些相近的拓扑关系。 $HIC_B > HIC_A > 0$ ,说明 A 和 B 为内离,且 B 比 A 更远离 P 的边界; $OC_E > OC_D > 1$ ,说明 D 和 E 为外离,且 E 比 D 更远离 R; $HIC_C = OC_F = 0$  且  $AS_F > AS_C$ ,说明 C 为内接 F 为外接且 F 与 R 之间的共有边界更多。绿化带的实际分布如图 4 所示(阴影部分为居民区)。

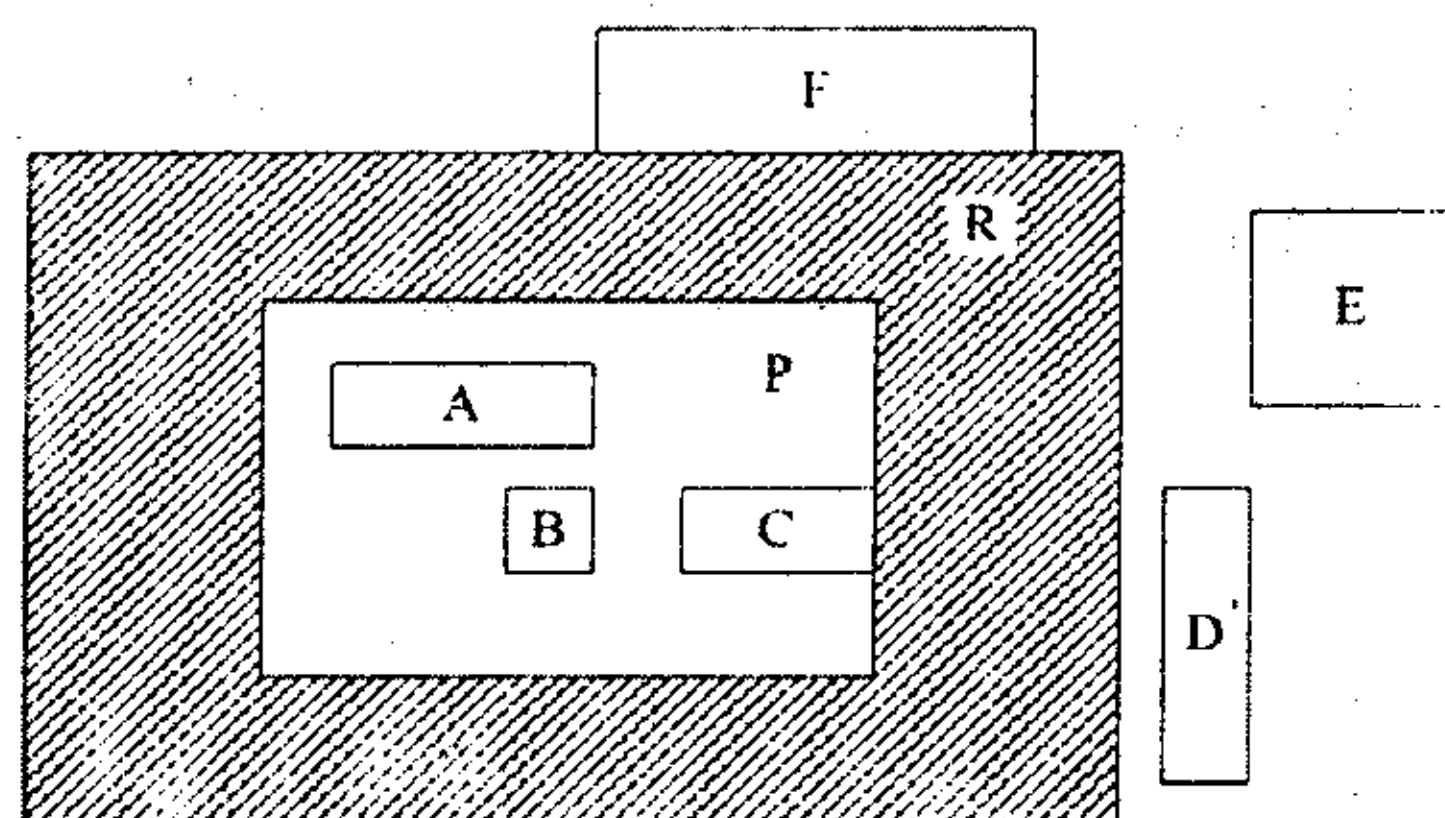


图 4 居民区与绿化带

#### 4 结束语

通过分析 9 交集模型在描述面对象间拓扑关系上的不足以及目前度量参数的不足,定义了三对细化的面面拓扑关系,提出了支持带孔面对象的度量参数,增强了面对象间空间关系的描述,并把度量参数应用在了拓扑查询中。基于文中的思想,还可以对线面对象间空间关系的度量参数进行改进。

值得提出的是,到目前为止所有的度量参数都只支持简单几何对象,支持复杂几何对象的度量参数将是下一步的研究方向。此外,如何利用度量参数定义自然语言的拓扑关系谓词,使拓扑查询变的人性化,也

是一个研究热点。

#### 参考文献:

- [1] EGENHOFER M, KUHN W. Interacting with Geographical Information System[C]//In Geographical Information Systems: Principle, Techniques, Management and Applications. London: Taylor & Francis, 1998.
- [2] EGENHOFER M, FRANZOSA R. Point - set Topological Spatial Relations[J]. International Journal of Geographical Information Systems, 1991, 5(2): 161 - 174.
- [3] 钟志农,唐征武,张帆. 一种统一的拓扑关系判断模型[J]. 国防科技大学学报, 2004, 26(5): 57 - 62.
- [4] 邓敏,李成名,刘包文. 利用拓扑和度量相结合的方法描述面目标间的空间关系[J]. 测绘学报, 2002, 32(2): 164 - 169.
- [5] EGENHOFER M, NEDAS K. Splitting Radios: Metric Details of Topological Line - Line Relations[C]//17th International FLAIRS Conference. Miami Beach, FL: [s. n.], 2004.
- [6] EGENHOFER M. Query Processing in Spatial - Query - by - Sketch[J]. Journal of Visual Languages and Computing, 1997, 8(4): 403 - 424.
- [7] EGENHOFER M, RASHID A, SHARIFF B M. Metric Details for Natural - Language Spatial Relations [J]. ACM Transaction on Information Systems, 1998, 16 (4): 295 - 321.
- [8] OpenGIS Consortium Inc. OpenGIS Simple Features Specification for SQL[S/OL]. Revision 1. 1. 1999 - 05 - 05. Http://www.opengis.org.
- [9] SHEKHAR S, CHAWLA S. 空间数据库[M]. 谢昆青, 马修军, 杨冬青译. 北京: 机械工业出版社, 2004: 34 - 36.

(上接第 186 页)

就可以很方便地实现对企业信息资源的快速检索,出于安全考虑,可以对文档设置访问权限。

Lucene 并没有规定数据源的格式,而只提供了一个通用的结构(Document 对象)来接受索引的输入,因此输入的数据源可以是数据库,Word 文档,PDF 文档,HTML 文档……只要能够设计相应的解析转换器将数据源构造成 Document 对象即可进行索引,该实例默认支持 \*.txt, \*.html, \*.htm 类型的电子文档的索引和检索,如因实际应用需要,还可以再开发设计 Word,PDF 等相应文档的解析转换器,将数据源构造成 Document 对象,就可以实现对 Word,PDF 等文档的检索支持。

#### 5 结束语

文中提出了一种解决全文检索的方法,可以应用到搜索引擎、中小企业网站站内检索、个人用户桌面搜

索引擎建立、特定文档检索数据库建立等,从而实现对目标文档方便的检索管理,提高检索效率。

#### 参考文献:

- [1] 苏新宁. 信息检索理论与技术[M]. 北京: 科学技术文献出版社, 2004.
- [2] 徐宝文, 张卫丰. 搜索引擎与信息获取技术[M]. 北京: 清华大学出版社, 2003.
- [3] 肖创柏. 基于全文检索技术的商业信函处理系统的设计与实现[J]. 计算机应用研究, 2004(1): 150 - 152.
- [4] 曹元大, 贺海军. 全文检索索引技术的研究与实现[J]. 计算机工程, 2004, 28 (6): 21 - 23.
- [5] Bookstein A, Swanson D R. A Decision Theoretic Foundation for Indexing[J]. Journal of the American Society for Information Science, 1975(26): 76 - 77.
- [6] Crossman D A, Frieder O. Information Retrieval: Algorithms and Heuristics [M]. Boston: Kluwer Academic Publishers, 1998: 49 - 51.