基于Lucene的本地搜索引擎研究与实现

秦 杰[1] 宋金玉[1] 张广星[2]
(解放军理工大学指挥信息系统学院 南京 210007)[1] (中国人民解放军66329部队 北京 101402)[2]

摘 要 为了改善计算机传统搜索在效率和返回结果上的不足,基于Lucene全文检索高效、准确的特点,采用非结构化文档结构化的思想,提出了文档内容自然分片索引的方法,实现了一个针对本地资源的个性化搜索引擎。

关键词 Lucene,全文检索,本地搜索,内容分片
中图法分类号 TP393 **文献标识码** A

Research and Application of Local Search Engine Based on Lucene

QIN Jie[1] SONG Jin-yu[1] ZHANG Guang-xing[2]
(Colleage of Command Information System, PLA University of Science and Technology, Nanjing 210007, China)[1]
(Unit 66329 of PLA, Beijing 101402, China)[2]

Abstract In order to improve the deficiencies of traditional search on efficiency and return results, based on full-text Lucene efficient and accurate characteristics, we adopted the idea of unstructured documents structured, proposed a method to slice document content naturally and index, and realized an individuation search engine for local resource.

Keywords Lucene, Full-text search, Local search, Content slicing

随着计算机存储资源的急剧增加,如何快速、准确地从大量本地资源中找到所需要的信息成为个人办公急需解决的问题。计算机操作系统自带的搜索工具功能十分有限,搜索缓慢,常常无法满足用户的需求。

Lucene是目前使用非常广泛的全文检索[2]工具包,可以方便地嵌入到应用程序中为用户提供全文检索服务。本文设计实现了一款本地搜索引擎,通过使用第三方工具包将WORD、PDF、HTML等常见文档资源转换构造为Lucene可识别的Document对象,再进一步对文档内容进行分片处理,最后使用Lucene索引/检索,实现了对本地资源的高效、个性化检索。

1 本地搜索现况

本地计算机一般没有专门的搜索工具,通常使用操作系统自带的搜索工具或者部分软件内嵌的简单搜索模块。常见的搜索方式有两种,分别是搜索文件名信息和搜索具体内容。

"搜索文件名"是操作系统自带的一个非常简单的搜索服务,最典型的体现是Windows的资源管理器搜索服务。搜索可以根据全部或部分文件名,也可根据文件内容中一个字或词组对指定路径下的文件进行搜索,并返回文件的文件名、路径等信息。它的特点是很简单、速度慢。

"搜索文件内容"一般是在文件中完成的搜索服务。这种服务通常是由相应的安装程序提供的。例如,使用word时,可以在"查找"框中输入需要查找的关键词,执行查找后,系统就可以自动找到所要的关键词,并高亮显示。

这种搜索方式是一种简单的线性"匹配"查找方法。查找是通过线性匹配驻留内存的文本信息来实现的,这种方法被称为顺序查找(Serial Scanning)[3]。它无需对文档集合中的信息进行预处理,或者只需要很简单的处理。这种方法适合文档较少或者文档信息经常变动的情况,优点是结构简单、易于实现;缺点是检索速度比较慢。

2 Lucene概述

2.1 Lucene简介

Lucene是一个基于Java的开源全文检索工具包,是一个全文检索引擎的框架,对外提供非常完全的索引引擎和查询引擎[1]。它为数据的处理和管理提供了简单的类调用接口,经过简单的二次开发就可以嵌入到应用中实现全文检索。

2.2 Lucene全文检索引擎

Lucene共有7个子包(Java中包用package表示),各个包都有着不同的功能[3],如表1所列。

表1 Lucene源码子包功能表

包名	功能
Org. apache. lucene. analysis	语言分析器,用于文本分词
Org. apache. lucene. document	逻辑文件(document)管理
Org. apache. lucene. index	索引管理
Org. apache. lucene. queryParser	查询分析器,对约束关键词处理计算
Org. apache. lucene. search	查询管理,根据查询条件搜索结果
Org. apache. lucene. store	存储数据
Org. apache. lucene. util	公共类

秦 杰(1990—),男,硕士,主要研究方向为军事信息学,E-mail:qinjie_823@163.com;宋金玉(1967—),女,副教授,主要研究方向为数据工程;张广星(1984—),男,硕士生,主要研究方向为软件工程。

其中，核心的包有 3 个：Org. apache. lucene. analysis（语言分词）、Org. apache. lucene. index（索引创建）和 Org. apache. lucene. search（执行搜索）。3 个核心包保证了 Lucene 作为一个检索引擎的功能，如图 1 所示。

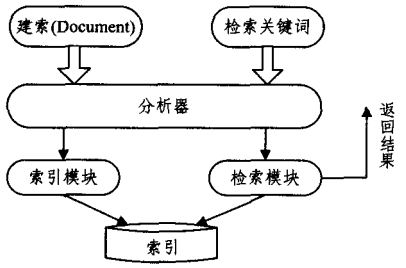

图 1　分词、索引创建和检索

(1)分词。Lucene 的分析工具位于 Org. apache. lucene. analysis 中。分词是搜索引擎的核心模块，分词的效果将直接影响搜索引擎的搜索精度和速度。Lucene 的分析器由两部分构成：分词器（Tokenizer）和过滤器（TokenFilter）[3]。过滤器主要是用于过滤从 Tokenizer 切分出来的词，如去除感叹词、无意义词等。

(2)索引。Lucene 索引相关 API 位于 Org. apache. lucene. index 包中。Lucene 读取需要索引的文件信息传给分析器（Analyzer），进行过滤、分词等处理后，按照倒排[4]的方式创建索引，并写入索引库。比如索引一本书，书作为逻辑文件（Document），其编号、书名、作者、出版社、摘要等信息分别作为域（Field）。一个 Document 类似于关系数据表中一条记录，而 Field 相当于属性项，如表 2 所列。

表 2　Document 与普通关系记录对照表

Document	Field1	Field2	Field3	Field4	Field5	…
记录	编号	书名	作者	出版社	摘要	…

实际索引的对象为每一个 Field，分词器根据分词算法分析处理后，存储相应的词（term）。如“北京欢迎你”用分词工具（MMAnalyzer）[3]处理后，出现有“北京”、“京欢”、“欢迎”、“迎你”，对照词库信息筛选后，存储“北京”，“欢迎”作为关键词。

(3)搜索。Lucene 搜索相关方法包含在 Org. apache. lucene. search 中。常用的基础类有 4 个：Searcher、Term、Query 和 Hits 类[3]。Searcher 是抽象的搜索类，包含多种搜索方法；Term（词）是搜索的基本单元；Query 是一个包装 Term 的抽象基类；Hits 用于接收搜索结果，通过 Hits 可以访问 Document 的各域（Field）。搜索与索引要用同样的分词器，搜索语句需要指定搜索域和关键词，如：search（file_name，钢铁）表示搜索文件名包含关键词“钢铁”的文件。

3　本地搜索引擎设计与实现

3.1　系统需求

针对目前本地搜索引擎的不足，本文所描述的本地搜索引擎的系统设计满足如下需求：

(1)用户可确定搜索目录，并自动完成索引创建；

(2)对选定目录下的文件进行文件名搜索，并给出所有检索结果文件的文件名和路径信息，供进一步查看。

(3)对选定目录下的 word、pdf 等常见非结构化文档内容信息进行全文索引/检索，给出文档所在位置和查询关键词的上下文信息。

(4)对检索结果文件，可以打开文档查看文档完整内容或执行文件。

3.2　系统实现

计算机文档资源常常包括 WORD、PDF、HTML、EXCEL、TXT 等格式，为了能够使用 Lucene 检索，需要将普通的文档资源转换构造成 Document 文档对象，再创建索引。索引之前，系统对 Document 文档对象内容项进行了分片处理。最后从文件名和内容两个方面定制用户的查询请求，通过检索得出结果。如图 2 所示。

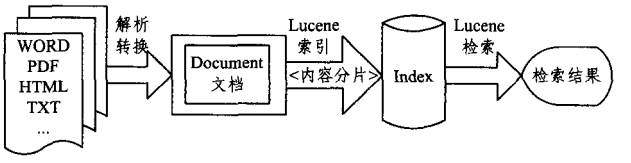

图 2　系统实现过程

3.2.1　文档解析转换

本文针对不同格式文档，用不同的第三方工具包来处理文档的解析，如表 3 所列。

表 3　文档解析第三方工具包表

文件格式	解析工具包
WORD，EXCEL，PPT	poi-3. 6-20091214. jar； poi-ooxml-3. 6-20091214. jar； poi-ooxml-schemas-3. 6-20091214. jar；
PDF	PDFBox-0. 7. 3. jar
HTML	htmlexer. jar；htmlparser. jar；
XML	dom4j-1. 6. 1. jar；xmlbeans-2. 3. 0. jar；

其中，PDF 是一种电子格式文件，它将文字、字型、格式、颜色、图片等信息封装成一个文件[5]，文中采用 PDFBox 读取文字信息；对于 Word、Excel 等 Office 产品，采用 POI 进行解析处理，POI 包含一系列的 API 可用于操控 MicroSoft OLE 2 Compound Document Format 的各种格式文件；采用 Java 编写的 HTML 解析库 htmlparser 处理 html 文档[6]；解析 XML 文档用到了 DOM（文档对象模型）[7]，DOM 将 XML 文档转换成为 DOM 树，为应用的访问提供了很大的灵活性。

对同一路径下的不同格式文件，系统调用不同的模块进行解析处理，并返回一致的 Document 对象以供下一步的 Lucene 索引。

3.2.2　内容分片索引

3.2.2.1　分片算法

搜索引擎中为了对数据进行快速检索，需要对数据建立索引。对于较大的数据块，常常先对数据进行分片，再建立索引。文献[3]提到对数据按照恒定大小（如：MAX_SIZE）进行分片的方法，结果导致片段断句不合理，一句话被分割到了不同的片中。

在按恒定大小（MAX_SIZE）进行分片的基础上，本文通过定义“断点”，设计了自然分片算法，其中“断点”即‘。’和‘\n’符号。算法在利用指向内容的指针（j）叠加 MAX_SIZE 基础上，判断并寻找后续最近的“断点”进行分片。算法流程如图 3 所示。

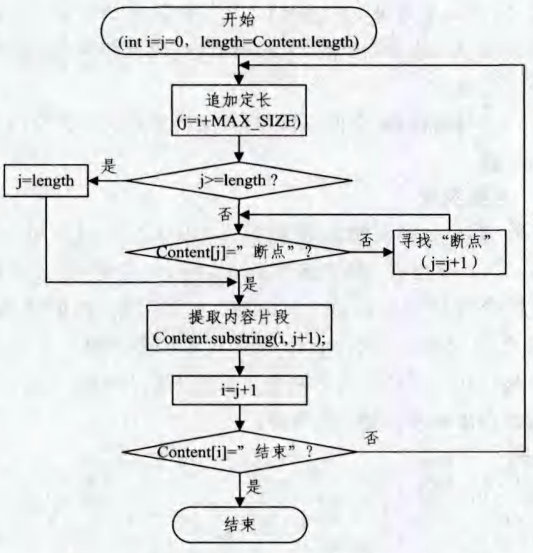

图 3　分片算法流程图

流程图中 i,j 为两个指针,Content 表示内容,MAX_SIZE 表示指定的分片大小,Content. substring(i,j)表示截取 Content 两个"断点"之间的字符内容。

3.2.2.2　索引创建

本文系统设计的核心是对文档资源的内容进行快速检索。对于一个 Document 文档对象,提取其文件名、路径和内容作为信息源进行索引。其中,内容项往往是很大的数据块,若将内容项整体返回,搜索就没有任何意义。

系统采用文中提出的分片算法,将内容划分为相对较少的片段,再将片段内容与文件名、路径信息分别作为属性项(field)加入一个新的 Document 对象,最后通过索引器(IndexWriter)的 addDocument 方法将 Document 写入索引,如图 4 所示。需要说明的是,在申请一个 Document 时,Lucene 会自动为其添加唯一 ID。

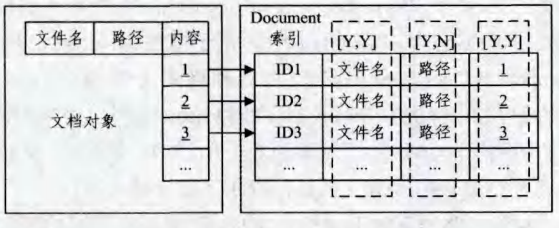

图 4　内容分片索引示意图

在将属性项(Field)加入 Document 时,可以为属性项指定是否存储和分词处理,图 4 中[Y,Y]表示要存储和分词,[Y,N]表示只存储不分词。存储表示将 field 值存储,分词表示对 field 值进行分词处理,提取 field 中的"词(Term)"并存储,"词"即为 Document 可以被检索的关键词。

3.2.3　检索定制

检索模块是直接与用户交互的部分,接收用户的查询请求,返回给用户需要的信息。文中系统提供本地资源的文件名和文件内容查询服务。查询模块在布尔型方法(BooleanQuery)的基础上进行模糊查询(FuzzyQuery)[3],实现了对文件名和内容的模糊查询。

对图 4 所示的索引方式生成的索引文件进行文件名为空(null)和文件内容为"lucene"的检索,结果如图 5 所示。其中文件 A 共有 3 条 Document 记录,文件 B 有 2 条 Document 记录。A 的前两个 Document 内容项(File_content)中含关键词"lucene",B 的第一个 Document 内容项中含"lucene"。当含有"lucene"的内容分片加入 Document 时,"lucene"会被存储并作为可检索的关键词。即 ID1、ID2 和 ID4 3 条 Document 可被内容关键词为"lucene"的检索语句找到。所以,在设定查询条件文件名空(null)和内容关键词(lucene)情况下可以得出图 5"检索结果"的部分内容。

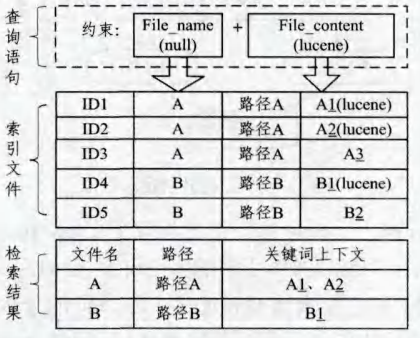

图 5　检索示意图

3.2.4　结果呈现

本文的搜索引擎在用户选定好查询路径后自动完成对资源信息的索引创建,用户可以对文件的文件名和文件内容进行联合查询,使用方便、高效。

系统返回查询结果的数量、用时和与关键词相关的文档内容上下文供用户浏览查看,用户还可根据浏览结果判断是否需要进一步查看的文档,在"结果文件"中选中该文件,调用相关的应用程序打开该文件。如图 6 所示,给出了在指定资料库(E:\资料\)下设定文件名关键词为空(null)和内容关键词"Lucene"的情况下检索出的相关结果,图 7 为调用系统 pdf 阅读器查看一个 pdf 格式的检索结果文件。

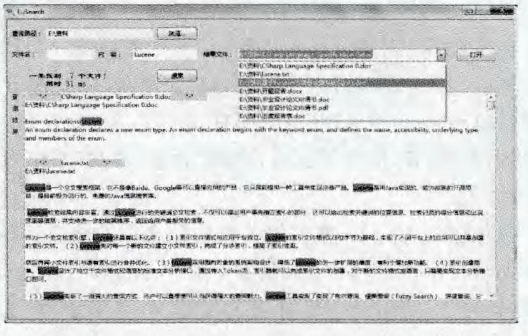

图 6　检索"Lucene"结果

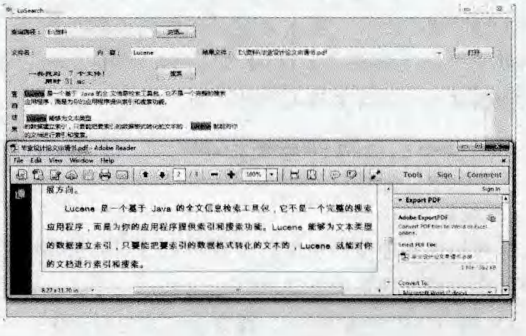

图 7　查看 pdf 格式结果文件

(下转第 407 页)

结束语 在过去的十年中，可视化编程语言为学生学习编程概念和技能提供了简单的方法，成功地激励了学生。然而，可视化语言并未集中于研究学生创建这些游戏学到了什么类型的知识。计算思维模式图为评估学生获取的具体知识提供初始方法。据观察，对于使用可视化语言进行教育的学校教师和学生，发现教师思维模式的能力是非常重要的。计算思维模式图为我们要回答的问题即“现在学生编程实现俄罗斯方块，学生可以编写科学模拟程序么?”提供了初始方法。此外，计算思维模式图有能力使以人为中心的计算成为可能，如同教师可以获得学生进步的即时反馈。

计算思维模式图的局限性包括指定的计算思维模式的任意性、区分类似模式的难度、计算思维模式图中选择计算思维模式的数量。已选的计算思维模式中，某些如扩散，描述得没有其它准确，如爬山法。尽管这种异常需要进一步研究，但是它没有破坏计算思维模式图的相对精确度，也没有减少检测这些情况下知识转移存在的价值。

分析多种组合的计算思维模式更接近于显示学生知识的深度和广度。计算思维模式图的语义性可让我们评估和可视化一个程序的实际基本意义。一个学生学习的句法评价只显示学生非常有限的背景知识。此外，学生将先前学到的计算思维模式应用到一个科学背景，更清晰地描绘了学生如何将新知识转移到新情况，通过比较 CTP 图可以证明知识转移的存在。在大多数的学习情境中，知识转移常常是假设的，不能保证这些转移实际发生。计算思维模式图是评价知识转移的更好的工具，因为计算思维模式图代表计算思维模式组合为一个可观察的可定义的结果。在一个学期的时间内，通过计算思维模式图检测知识转移的能力，是有效的第一步测量其它领域的转移以及可能的其它形式的学习。

进一步将研究计算思维模式识别的附加验证。目前的模型已经过手动验证，通过比较计算思维模式图的输出和根据玩游戏/模拟以及查看源代码来人工分级评价计算思维模式。计算思维模式图表现相当好，其潜在的误报问题可以通过深化分析水平降低。目前的分析水平停留在个别条件和行为上。分析没有将这些行为和条件分解成参数，这些参数可以更有效地区分类似模式。

参 考 文 献

[1] Repenning A, Webb D, Ioannidou A. Scalable game design and the development of a checklist for getting computational thinking into public schools[C]// Proc. SIGCSE'10. ACM Press, WI, USA, 2010

[2] 周以真. 计算思维[J]. 中国计算机学会通讯, 2007, 3(11)

[3] Wing J M. Computational Thinking[J]. Communications of the ACM, 2006, 49(3)

[4] 董荣胜，古天龙. 计算思维与计算机方法论[J]. 计算机科学, 2009, 36(1)

[5] Kozaczynski V, Ning J, Engberts A. Program Concept Recognition and Transformation[J]. IEEE Trans. On Software Eengineering, 1992, 18(12): 1065-1075

[6] Duscasse S, Rieger M, Demeyer S. A Language Independent Approach for Detecting Duplicated Code[C]// Int'l Conf. on Software Maintenance. 1999: 109-118

[7] 苏舟. 基于向量空间范围搜索的大型软件相似度检测[D]. 杭州：浙江大学, 2008

[8] 李亚军，徐宝文，周晓宇. 基于 AST 的克隆序列与克隆类识别[J]. 东南大学学报, 2008, 38(2): 228-232

[9] Krinke J. Identifying Similar Code with Program Dependence Graphs[C]// Proceedings Eighth Working Conference on Reverse Engineering. 2001: 301-309

[10] Lewis C M. How programming environment shapes perceptionlearning and goals: Logo vs. Scratch[C]// Proc. SIGCSE'10. ACM Press, WI, USA, 2010

[11] 朱国强，刘真，李宗伯. 对计算机系统中程序行为的分析和研究[J]. 计算机应用, 2005(12)

[12] 陈浩，王广南，孙建华. 一种基于图的程序行为相似性比较方法[J]. 计算机应用研究, 2010(2)

[13] 汪应洛，李勖. 知识的转移特性研究[J]. 系统工程理论与实践, 2002(10)

(上接第 370 页)

从检索结果可以看出，在对指定资料库进行文件名(null)和内容项(Lucene)关键词搜索时，一共找出 7 个相关文件，包括 WORD、PDF、TXT 格式文档。在“查询结果”输出框中，给出了相关文件内容关键词的上下文信息，信息断句自然合理。整个检索用时 31ms，效率较高，达到了预期研发目的。

结束语 本文阐述了当前计算机本地资源搜索存在的问题，介绍了全文检索引擎 Lucene 的基本原理、源码和功能结构，实现了一个利用 Lucene 来解决本地搜索问题的搜索引擎。引擎有效地运用了 Lucene 全文检索的特性，通过将 WORD、PDF 等常见文档解析构造成 Document 对象并对文档内容进行分片索引，实现了对文档内容的全文检索。下一步工作将对引擎的动态索引进行研究，定期检测路径下文件的变动信息，以保证检索的时效性。

参 考 文 献

[1] Gospodnetic O, Hatcher E. Lucene in action[M]. Manning Publications Co., 2005

[2] 孙西全，马瑞芳，李燕灵. 基于 Lucene 的信息检索的研究与应用[J]. 情报理论与实践, 2006, 29(1): 521-528

[3] 邱哲，符滔滔，王学松. 开发自己的搜索引擎 Lucene+Heritrix (第 2 版)[M]. 人民邮电出版社, 2005

[4] 林洁. 个性化综合倒排索引在 Lucene 中的应用[J]. 电脑知识与技术, 2010, 6(4): 932-934

[5] PDF-百度百科[OL]. http://wapbaike.baidu.com/view/15963.htm

[6] 包宇宁. 使用 Java 编程解析 HTML 文档[J]. 福建电脑, 2004(9): 86-87

[7] 周筱媛. 用 Java 集合类处理 XML 文档[J]. 西安科技学院学报, 2002, 22(3): 318-320