开放源代码:开放源代码搜索引擎 (转贴)

疯狂代码 http://www.crazycoder.cn/ ĵ:http:/www.crazycoder.cn/Arithmetic/Article30850.html

开放源代码搜索引擎 (转贴)开放源代码搜索引擎为人们学习、研究并掌握搜索技术提供了极好的途径与素材，推动了搜索技术的普及与发展，使越来越多的人开始了解并推广使用搜索技术。使用开源搜索引擎，可以大大缩短构建搜索应用的周期，并可根据应用需求打造个性化搜索应用，甚至构建符合特定需求的搜索引擎系统。搜索引擎的开源，无论是对技术人员还是普通用户，都是一个福音。 搜索引擎的工作流程主要分为三步：从互联网抓取网页→创建抓取网页的索引库→从索引库中进行搜索。 首先需要一个能访问网络的爬虫器程序，依据URL之间的关联性自动爬行整个互联网，并对爬行过的网页进行抓取收集。当网页被收集回来后，采用索引分析程序进行网页信息的分析，依据一定的相关度算法（如超链接算法）进行大量计算，创建倒排序的索引库。索引库建好后用户就可以通过提供的搜索界面提交关键词进行搜索，依据特定的排序算法返回搜索结果。因此，搜索引擎并不是对互联网进行直接搜索，而是对已抓取网页索引库的搜索，这也是能快速返回搜索结果的原因，索引在其中扮演了最为重要的角色，索引算法的效率直接影响搜索引擎的效率，是评测搜索引擎是否高效的关键因素。 网页爬行器、索引器、查询器共同构成了搜索引擎的重要组成单元，针对特定的语言，如中文、韩文等，还需要分词器进行分词，一般情况下，分词器与索引器一起使用创建特定语言的索引库。 而开放源代码的搜索引擎为用户提供了极大的透明性，开放的源代码、公开的排序算法、随意的可定制性，相比于商业搜索引擎而言，更为用户所需要。目前，开放源代码的搜索引擎项目也有一些，主要集在中搜索引擎开发工具包与架构、Web搜索引擎、文件搜索引擎几个方面，本文概要介绍一下当前比较流行且相对比较成熟的几个搜索引擎项目。 开源搜索引擎工具包 1．Lucene Lucene是目前最为流行的开放源代码全文搜索引擎工具包，隶属于Apache基金会，由资深全文索引/检索专家Doug Cutting所发起，并以其妻子的中间名作为项目的名称。Lucene不是一个具有完整特征的搜索应用程序，而是一个专注于文本索引和搜索的工具包，能够为应用程序添加索引与搜索能力。基于Lucene在索引及搜索方面的优秀表现，虽然由Java编写的Lucene具有天生的跨平台性，但仍被改编为许多其他语言的版本：Perl、Python、C++、.Net等。 同其他开源项目一样，Lucene具有非常好的架构，能够方便地在其基础上进行研究与开发，添加新功能或者开发新系统。Lucene本身只支持文本文件及少量语种的索引，并且不具备爬虫功能，而这正是Lucene的魅力所在，通过Lucene提供的丰富接口，我们可以根据自身的需要在其上添加具体语言的分词器，针对具体文档的文本解析器等，而这些具体的功能实现都可以借助于一些已有的相关开源软件项目、甚至是商业软件来完成，这也保证了Lucene在索引及搜索方面的专注性。目前，通过在Lucene的基础上加入爬行器、文本解析器等也形成了一些新的开源项目，如LIUS、Nutch等。并且Lucene的索引数据结构已经成了一种事实上的标准，为许多搜索引擎所采用。 2．LIUS LIUS即Lucene Index Update and Search的缩写，它是以Lucene为基础发展起来的一种文本索引框架，和Lucene一样，同样可以看作搜索引擎开发工具包。它在Lucene的基础上作了一些相应的研究及添加了一些新的功能。LIUS借助于许多开源软件，可以直接对各种不同格式/类型的文档进行文本解析与索引，这些文档格式包括MS Word、MS Excel、MS PowerPoing、RTF、PDF、XML、HTML、TXT、Open Office及JavaBeans等，对Java Beans的支持对于进行数据库索引非常有用，在用户进行对象关系映射（如：Hibernate、JDO、TopLink、Torque等）的数据库连接编程时会变得更加精确。LIUS还在Lucene的基础上增加了索引更新功能，使针对索引的维护功能进一步完善。并且支持混和索引，可以把同一目录下与某一条件相关的所有内容整合到一起，这种功能对于需要对多种不同格式的文档同时进行索引时非常有用。

3．Egothor Egothor是一款开源的高性能全文搜索引擎，适用于基于全文搜索功能的搜索应用，它具有与Luccene类似的核心算法，这个项目已经存在了很多年，并且拥有一些积极的开发人员及用户团体。项目发起者Leo Galambos是捷克布拉格查理大学数学与物理学院的一名高级助理教授，他在博士研究生期间发起了此项目。 更多的时候，我们把Egothor看作一个用于全文搜索引擎的Java库，能够为具体的应用程序添加全文搜索功能。它提供了扩展的Boolean模块，使得它能被作为Boolean模块或者Vector模块使用，并且Egothor具有一些其他搜索引擎所不具有的特有功能：它采用新的动态算法以有效提高索引更新的速度，并且支持平行的查询方式，可有效提高查询效率。在Egothor的发行版中，加入了爬行器、文本解析器等许多增强易用性的应用程序，融入了Golomb、Elias-Gamma等多种高效的压缩方法，支持多种常用文档格式的文本解析，如HTML、PDF、PS、微软Office文档、XLS等，提供了GUI的索引界面及基于Applet或者Web的查询方式。另外，Egothor还能被方便地配置成独立的搜索引擎、元数据搜索器、点对点的HUB等多种且体的应用系统。
4．Xapian Xapian是基于GPL发布的搜索引擎开发库，它采用C++语言编写，通过其提供绑定程序包可以使Perl、Python、PHP、Java、Tck、C#、Ruby等语言方便地使用它。 Xapian还是一个具有高适应性的工具集，使开发人员能够方便地为他们的应用程序添加高级索引及搜索功能。它支持信息检索的概率模型及丰富的布尔查询操作。Xapian的发布包通常由两部分组成：xapian-core及xapian-bindings，前者是核心主程序，后者是与其他语言进行绑定的程序包。 Xapian为程序开发者提供了丰富的API及文档进行程序的编制，而且还提供了许多编程实例及一个基于Xapian的应用程序Omega，Omega由索引器及基于CGI的前端搜索组成，能够为HTML、PHP、PDF、PostScript、OpenOffice/StarOffice、RTF等多种格式的文档编制索引，通过使用Perl DBI模块甚至能为MySQL、PostgreSQL、SQLite、Sybase、MS SQL、LDAP、ODBC等关系数据库编制索引，并能以CSV或XML格式从前端导出搜索结果，程序开发者可以在此基础上进行扩展。 5．Compass Compass是在Lucene上实现的开源搜索引擎架构，相对比于Lucene而言，提供更加简洁的搜索引擎API。增加了索引事务处理的支持，使其能够更方便地与数据库等事务处理应用进行整合。它更新时无需删除原文档，更加简单更加高效。资源与搜索引擎之间采用映射机制，此种机制使得那些已经使用了Lucene或者不支持对象及XML的应用程序迁移到Compass上进行开发变得非常容易。 Compass还能与Hibernate、Spring等架构进行集成，因此如果想在Hibernate、Spring项目中加入搜索引擎功能，Compass是个极好的选择。 开源Web搜索引擎系统 1．Nutch Nutch是Lucene的作者Doug Cutting发起的另一个开源项目，它是构建于Lucene基础上的完整的Web搜索引擎系统，虽然诞生时间不长，但却以其优良血统及简洁方便的使用方式而广收欢迎。我们可以使用Nutch搭建类似Google的完整的搜索引擎系统，进行局域网、互联网的搜索。
2．YaCy YaCy是一款基于P2P(peer-to-peer)的分布式开源Web搜索引擎系统，采用Java语言进行编写，其核心是分布在数百台计算机上的被称为YaCy-peer的计算机程序，基于P2P网络构成了YaCy网络，整个网络是一个分散的架构，在其中所有的YaCy-peers都处于对等的地位，没有统一的中心服务器，每个YaCy-peer都能独立的进行互联网的爬行抓取、分析及创建索引库，通过P2P网络与其他YaCy-peers进行共享，并且每个YaCy-peer又都是一个独立的代理服务器，能够对本机用户使用过的网页进行索引，并且采取多机制来保护用户的隐私，同时用户也通过本机运行的Web服务器进行查询及返回查询结果。 YaCy搜索引擎主要包括五个部分，除普通搜索引擎所具有的爬行器、索引器、反排序的索引库外，它还包括了一个非常丰富的搜索与管理界面以及用于数据共享的P2P网络。 开源桌面搜索引擎系统 1．Regain regain是一款与Web搜索引擎类似的桌面搜索引擎系统，其不同之处在于regain不是对Internet内容的搜索，而是针对自己的文档或文件的搜索，使

用regain可以轻松地在几秒内完成大量数据（许多个G）的搜索。Regain采用了Lucene的搜索语法，因此支持多种查询方式，支持多索引的搜索及基于文件类型的高级搜索，并且能实现URL重写及文件到HTTP的桥接，并且对中文也提供了较好的支持。 Regain提供了两种版本：桌面搜索及服务器搜索。桌面搜索提供了对普通桌面计算机的文档与局域网环境下的网页的快速搜索。服务器版本主要安装在Web服务器上，为网站及局域网环境下的文件服务器进行搜索。 Regain使用Java编写，因此可以实现跨平台安装，能安装于Windows、Linux、Mac OS及Solaris上。服务器版本需要JSPs环境及标签库（tag library），因此需要安装一个Tomcat容器。而桌面版自带了一个小型的Web服务器，安装非常简单。 2．Zilverline Zilverline是一款以Lucene为基础的桌面搜索引擎，采用了Spring框架，它主要用于个人本地磁盘及局域网内容的搜索，支持多种语言，并且具有自己的中文名字：银钱查打引擎。Zilverline提供了丰富的文档格式的索引支持，如微软Office文档、RTF、Java、CHM等，甚至能够为归档文件编制索引进行搜索，如zip、rar及其他归档文件，在索引过程中，Zilverline从zip、rar、chm等归档文件中抽取文件来编制索引。Zilverline可以支持增量索引的方式，只对新文件编制索引，同时也支持定期自动索引，其索引库能被存放于Zilverline能够访问到的地方，甚至是DVD中。同时，Zilverline还支持文件路径到URL的映射，这样可以使用户远程搜索本地文件。 Zilverline提供了个人及研究、商业应用两种许可方式，其发布形式为一个简单的war包，可以从其官方网站下载（http://www.zilverline.org/）。Zilverline的运行环境需要Java环境及Servlet容器，一般使用Tomcat即可。在确保正确安装JDK及Tomcat容器后只需将Zilverline的war包（zilverline-1.5.0.war）拷贝到Tomcat的webapps目录后重启Tomcat容器即可开始使用Zilverline搜索引擎了。 2008-12-6 22:24:58