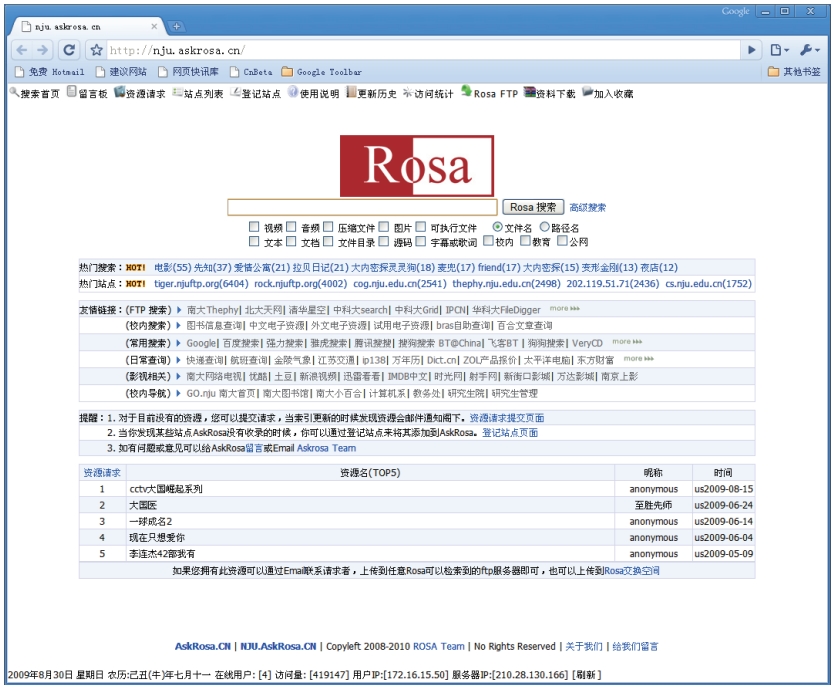
ROSA FTP 搜索引擎综述

2009-8-28

1 简介

ROSA 是一个通用意义上的开源 FTP 搜索引擎，我们借助 Lucene, Struts 及其他开源软件的帮助，设计实现了一个基于当前搜索引擎技术的 FTP 搜索引擎。相对于校内以前的搜索引擎，ROSA 的搜索灵活性更高，更符合当前用户的自定义搜索需求。并且 ROSA 采用分布式设计思路，扩展性更好。面向对象的设计和前后台设计分离，使 ROSA 的部署，升级和后续功能的实现变得较为容易。图1给出了 ROSA 首页的截图。



1: ROSA 首页

2 设计动机

现有的校内搜索引擎 Thephy 存在的问题：

1. 不支持复杂搜索，只是简单的字符串匹配。这样的搜索匹配效率不高，最严重的缺点是不能支持用户的复杂自定义搜索。

2. 搜索结果冗余。比如对于”/movie/ 蓝莓之夜 /1.rmvb”这样的资源，当搜索“蓝莓之夜”时结果会有两条：

   - /movie/ 蓝莓之夜

- /movie/ 蓝莓之夜 /1.rmvb

当用户输入的关键词匹配条目很多的时候，这种冗余会带来很大的混乱，迫使用户浪费很多时间来区分条目。

3. 现有的搜索引擎不支持任何形式的排序。对于很多资源，比如软件，电影，用户往往喜欢选择最新的资源下载，在没有排序支持的情形下需要用户人工区分资源更新时间。

4. 考虑到同一关键字可能会同时出现软件，电影等资源名称中，所以分类搜索也是用户十分需要的功能。

5. 现有的搜索引擎用专门的论坛负责站点登记，当用户想把自己的站点登入的时候，要在论坛发帖，然后由管理员手工添加。这种方式过于浪费人力和时间。我们需要一个 FTP 服务器站点的自主管理平台。

6. 现有的搜索引擎无法确定结果页中的站点是否可以访问。我们希望能在搜索结果页面中给出站点连接性能提示，以便用户选择。

ROSA 搜索引擎要实现的改进:

1. 使用当代搜索引擎技术，抛弃简单的字符串匹配，而使用基于分词的倒排表保存索引数据和提供搜索服务。

2. 细致的可用搜索域，可以满足用户各种搜索需求。

3. 针对常用搜索域的升序降序排列。方便用户排序搜索资源。

4. 提供针对路径名和文件名的多种搜索方式，以使搜索结果简洁美观。

5. 对资源根据文件后缀进行自动分类，并提供相应的搜索接口，方面用户进行分类搜索。

6. 自主的站点登录、管理和登出系统。方面 FTP 管理员管理站点。

7. 资源访问和站点访问统计，方便用户分享各自的兴趣和选取热点站点。

8. 提供搜索建议功能，方便用户选择搜索关键词。

9. 站点可连接性提示，方便用户选择站点下载资源。

10. 多点下载程序，减轻用户站点选取的负担，因为用户关注的是资源而非站点。所以我们需要一个多点下载程序可以根据用户选取的资源信息同时从多个站点来分片下载资源。

3 开发环境

开发语言 ROSA 后台使用 Java 语言开发，前台使用 JSP 和 Struts 架构构建。

开发工具 开发工具为 MyEclipse 6.0.1，VisualSVN Server 2.0.6 和 Subclipse 1.6.5。

服务器 使用的服务器为 Apache Tomcat 6.0.18 和 MySql 5.0。

软件规模 后台纯 Java 代码共 222 个源文件，29587 行代码。JSP 代码共 31 个文件，3000 行代码。

4 使用的开源软件

4.1 Lucene

Lucene 是一套用于全文检索和搜寻的开源程式库，由 Apache 软件基金会支持和提供。Lucene 提供了一个简单却强大的应用程式接口，能够做全文索引和搜寻，在 Java 开发环境里 Lucene 是一个成熟的免费开放源码工具[1]。

4.2 Struts

Struts 是 Apache 软件基金会（ASF）赞助的一个开源项目。它最初是 Jakarta 项目中的一个子项目，并在 2004 年 3 月成为 ASF 的顶级项目。它通过采用 Java Servlet / JSP 技术，实现了基于 Java EE Web 应用的 Model-View-Controller〔MVC〕设计模式的应用框架〔Web Framework〕，是 MVC 经典设计模式中的一个经典产品[2]。

4.3 Log4j

log4j 是一个开源的日志系统，它允许开发者以任意的粒度控制日志语句的输出。通过使用外部配置文件，log4j 在运行时也可以自由的配置。最重要的是 log4j 简单易学，非常容易上手[3]。

4.4 JDom

JDOM 是一个开源项目，它基于树型结构，利用纯 JAVA 的技术对 XML 文档实现解析、生成、序列化以及多种操作。JDOM 直接为 JAVA 编程服务。它利用更为强有力的 JAVA 语言的诸多特性（方法重载、集合概念以及映射），把 SAX 和 DOM 的功能有效地结合起来[4]。

4.5 Commons Net

Jakarta Commons Net 实现了很多客户端的网络协议。这个程序库的目的在于提供基本的协议使用而非高层抽象[5]。它支持的协议包括：

- FTP/FTPS
- NNTP
- SMTP
- POP3
- Telnet
- TFTP
- Finger
- Whois
- rexec/rcmd/rlogin
- Time (rdate) and Daytime
- Echo
- Discard
- NTP/SNTP

4.6 Jakarta Oro

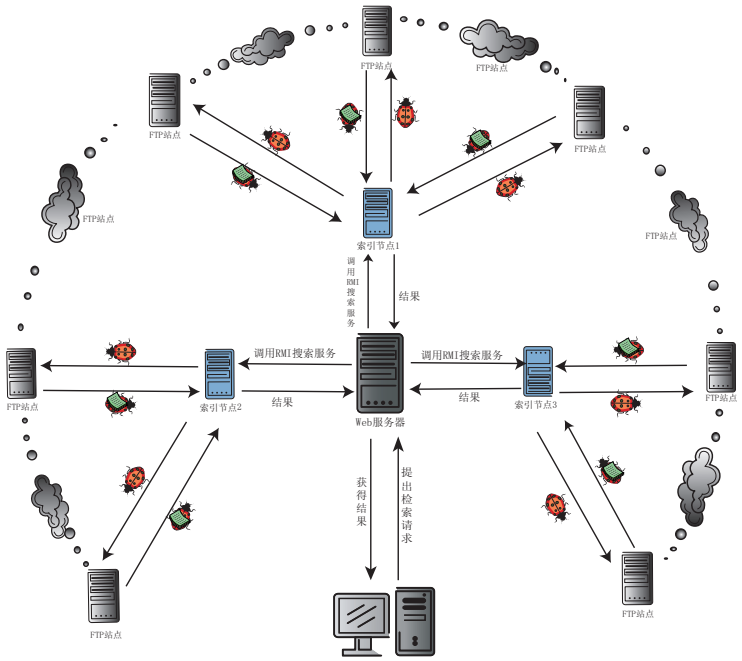
Jakarta-ORO 是一套兼容 Perl5 和类 AWK 正则表达式的文本处理 API[6]。

5 软件架构

考虑到缺乏专用的服务器，如果将所有的爬虫程序放在某一台机器上的话会影响正常的使用，所以我们考虑了分布的方式来进行索引和提供搜索服务。我们通过站点编号 mod 站点数 来给每个索引节点分配任务。每个节点都同时提供搜索服务，采用 RMI 的实现方式。之所以选择 RMI，是考虑到 RMI 的如下三个优点：

- 对其他框架的依赖性弱，RMI 框架一直都是 JDK 的标准组件
- 经济性好，通过整合开源工具，避免购买昂贵的授权
- 运行效率高，具有一定的可扩展性。比如 RMI 的序列化和压缩机制都可以进行自由的自定义扩展。

独立设计的爬虫程序，采用线程池并发的从多个 FTP 服务器站点获取资源列表，然后提交给索引器建立索引。图2给出了 ROSA 的三层架构。



2: ROSA 架构

6 开发进度

1: ROSA 开发进度

时间	进度
2008/2-2008/3	项目规划设计
2008/4-2008/5	编码实现
2008/5-2008/12	部署测试和修改
2009/1-2009/4	前端界面改进和系统功能添加

更详细的系统变更参看表2的 ChangeLog 记录。

2: ROSA ChangeLog

增加搜索建议功能，支持 Firefox,IE,Opera 等浏览器	2009-04-10
增加基于文件后缀的搜索	2009-04-10
增加基于站点位置的搜索，比如教育网，校园网，公网	2009-04-10
修改了 FtpSearch 的代码，修正了原来当用户点击页码数超过一定数值抛出异常的 bug	2009-04-04
更新了 Lucene 版本到 2.4.1	2009-04-04
新增 ROSA 的源码和文档下载页面	2009-03-18
更改资源请求页面	2009-03-18
FTP 站点、留言板、资源请求三个页面增加分页和搜索功能	2009-03-18
更改留言板页面	2009-03-18
增加快照功能	2009-03-18
增加多点下载功能	2009-03-18
修改搜索结果显示页面	2009-03-18
更改站点登记页面	2009-02-08
增加友情链接	2009-02-08
增加留言板和资源请求页面	2008-10-25
首页增加热点搜索排名	2008-07-19
增加支持 FlashFxp 导入的站点列表文件下载	2008-07-19
更改高级搜索页面查询解释，方便用户理解	2008-06-06
更改高级搜索页面日期大小对比操作，增加页面农历显示	2008-06-05
增加页面底部信息提示	2008-06-03
增加历史访问量统计	2008-06-01
申请了域名 http://askrosa.cn	2008-05-29
增加 ChangeLog 页面	2008-05-27
首页布局调整	2008-05-27
增加站点列表中的站点快照和单点搜索功能	2008-05-27

7 功能特色

本搜索引擎从用户使用角度出发，结合开源的 Lucene 搜索引擎，将校园内包括公网的 FTP 站点资源整合起来，方便了校园用户的检索和资源共享。相较于现存的校内搜索引擎，我们的搜索引擎设计的更为灵活，支持用户各种搜索需求。此外三层的设计结构，对系统的扩展性保留了很大的空间，使的系统可以快速的适应需求变化。

7.1 搜索关键字过滤

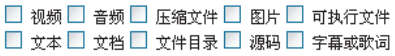
有些关键字会导致 Lucene 分析器抛出异常，比如“[先知].Knowing.2009.BDRip.X264-TLF”这样的关键字由于“[先知]”的存在会导致程序抛出异常，这是因为“[]”这样的符号对在 Lucene 中表示范围搜索。一旦出现异常，搜索结果将为空，这是用户不愿看到的，所以我们队用户输入进行了过滤。比如[]以及~这样的 Lucene 关键字字符都会被过滤掉。但高级搜索页面中没有进行过滤，这里假定使用高级搜索页面的用户知道一些更高级的搜索方法。

7.2 文件名或文件路径查询

用户可以根据路径或者文件名查询。路径就是 FTP 服务器上的完整路径，文件名就是最后一级的文件或文件夹名称。默认是根据文件名查询。举例来说：对于 /movie/ 蓝莓之夜 /1.rmvb。如果根据路径搜索，会有下面两条结果：
/movie/ 蓝莓之夜
/movie/ 蓝莓之夜 /1.rmvb
而根据文件名搜索只有一个结果:
/movie/ 蓝莓之夜
默认根据文件名查询，参看图3。

3: 文件名或路径名搜索

7.3 分类查询



这个功能使用起来比较简单，只需要在页面上选择你需要的分类就可以了。多选表示或者关系，不选表示任意分类。比如你搜电影：老无所依，而发现结果有一堆的分卷压缩包，那么你选择文件夹会使得结果变得简洁很多。使用这个功能是比较方便的，不需要给出文件后缀，因为我们都已经帮你考虑了。分类是基于文件后缀，具体的分类依据参看表3。

3: 分类说明

分类	意义	相应的文件后缀
video	视频	rm; rmvb; mpg; mpeg; mov; mtv; dat; wmv; avi; 3gp; dmv; divx; asf; vob; flv; mkv; swf
subtitle	字幕或歌词	srt; idx; sub; smi; ssa; lrc
audio	音频	wav; mp3; ra; rma; wma; asf; mid; midi; rmi; ogg; mod; ape; aiff; au; voc; vox
document text	文档 文本	pdf; doc; docx; chm; ppt; pptx; xls; xlsx; rtf tex; txt; html; htm; xml; log; ini; properties; prop; url; css
program	源码	java; c; cc; c++; cpp; cxx; h; hh; hpp; hxx; lisp; perl; pas; sh; pyo; a; so; lo; la; sql; class; js; o
image	图片	gif; jpg; jpeg; jpe; ico; mng; pbm; pgm; psd; png; pnm; ps; ppm; tif; tiff; bmp; xpm; eps; dcm; dicom
compress	压缩文件	gz; bz2; tar; zip; rar; ar; ear; jar; war; iso
executable	可执行文件	bat; exe; msi; dll; cab; com; sys; deb
unknown	未知	
directory	文件目录	

7.4 搜索结果排序

通过点击页面上的排序方式链接可以分别根据相关度（默认排序方式），文件大小或日期进行排序。参看图4。

4: 搜索排序

7.5 后缀名查询

通过使用 关键字 *+ext:*后缀名 这样的查询条件可以针对特定的后缀名进行过滤。此处的后缀名全为小写。使用大写的后缀名将导致搜索结果为空。

7.6 逻辑与或非

这个其实也是比较简单的，就是说搜索的时候可以用 AND,OR,NOT，注意此处我使用的是大写形式，也就是说你在搜索里使用的时候也要是大写形式，小写形式不做处理，就当作一般的搜索词使用。ROSA 中默认的连接词是 AND。

7.7 单点搜索

如果你只想对某些站点的内容进行搜索的话，你可以使用这个功能，使用的方式也是非常简单的。比如 friend server:172.16.65.79 这样搜索语句就可以将对关键词 friend 的搜索限制在 172.16.65.79 这个站点内。如果你希望在多个指定站点内搜索，使用上面的逻辑组合就可以了，比如 movie (server:172.16.65.79 OR server:172.16.65.105) 就可以将搜索限制在站点 172.16.65.79 或 172.16.65.105 上。

7.8 通配符

这个大家可能不是很熟悉，但都应该听过。其实就是使用 *，? 这些具有特定意义的字符。? 代表一个或零个字符，* 代表 0 个或多个字符。我们对这些通配符提供了一定的支持，比如：movie server:172.16.*.* 这样的搜索请求我们是支持的，它讲搜索限制在以 172.16 为前缀的地址上。当然也可以在搜索关键词中加入通配符，比如 fri?nd server:172.16.*.*。这对于搜索在同一网段的资源十分有益，因为同一网段的资源往往下载速度可观。

特别提示：不建议搜索以 * 开头的关键词。虽然我们支持这种查询，但由于这种通配符生成的枚举项过多会导致搜索性能的下降。我们使用的是 1024 个枚举项的默认设置，如果你的搜索关键词导致枚举项的数目超过 1024 那么你的到的将是一个异常，当然我们没有将这个异常展示在你的面前而是返回一个空的结果集。

7.9 模糊查询

我们的系统对这个功能有可以容忍的支持。之所以说可以容忍是因为不支持中文，但我想使用此功能的大多数情况应该是英文，所以模糊查询还是值得一提的。当你记不清某个单词的时候是不是就没办法查询了呢，你可能会考虑使用通配符，但你可能发现通配符要求你的输入字母有一定的顺序性，比如 frien* 和 frine* 的查询结果就很不相同，很可能后一个的查询结果为空（我测试了一下，结果

确实为空）。而使用模糊查询，你的这个问题就很容易解决了，你可以尝试搜索 frine~或者 frined~ 这样的关键词，你会发现你真正想要的 friends 也在搜索结果中，并且是比较靠前的。由于关键字过滤，这个功能只能在高级搜索页面使用。

7.10 短语查询

这个功能其实大家也都在 Google 上使用过，比如你想查 love me love my dog 或 to be or not to be 这样的短语，直接输入可能结果不是很理想，但加上引号之后结果就比较符合你的需要了。这就是短语查询，在我们的系统中也提供了这样的支持，但这要建立在你对文件名或者文件路径比较熟悉的情形。比如："YOU ARE TEH LOVE.mp3"这样的搜索结果只有一条，而去掉括号后结果条数为 1944 条。短语查询的另一个用法是匹配单词间隔。比如你输入"Say Love"，查询结果为空，但输入"Say Love"~1 输出结果如下：

这样的输入匹配了 Say 和 Love 的间隔小于一个词的结果。而不加上~默认的是 0 个间隔，故匹配结果为空。所以在你不能很准确的把握某个短语时，使用这种加上间隔的查询会帮助你找到你想要的资源。同样由于关键字过滤，这个功能只能在高级搜索页面使用。

7.11 范围查询

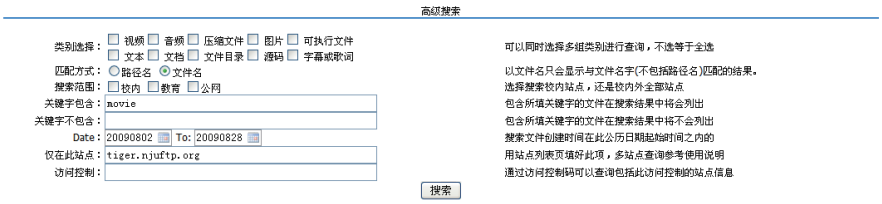
这个功能在我们的搜索中主要提供时间范围的查询，比如你想查询修改时间在 2008 年 3 月以来的有 movie 关键词的条目，可以这样搜索 movie date:[20080301 TO null]。某一端为 null 表示在这以端不加限制。同样由于关键字过滤，这个功能只能在高级搜索页面通过选择起始结束时间来使用。

7.12 访问控制

如果你有一个 FTP 站点，但你不想将这个站点注册上来然后能被所有人搜索，那么你可以在我们的注册页面上将访问控制设定为你想要的字串，那么在高级搜索页面中可以输入这个字串来搜索你的站点，而不输入这个字串的用户是搜索不到你的站点的。默认的字串是 anybody，所以如果你想让你的站点只在小范围共享的话请确保你的字串不要为 anybody。这个功能只能在高级搜索页面使用。

7.13 高级搜索

在高级搜索页面中，不对关键字进行非法字符过滤。此外用户可以组合多种搜索条件来减少结果范围，达到快速定位资源的目的。参看图5。功能7.9,7.10,7.11和7.12都只能在这个页面中使用。

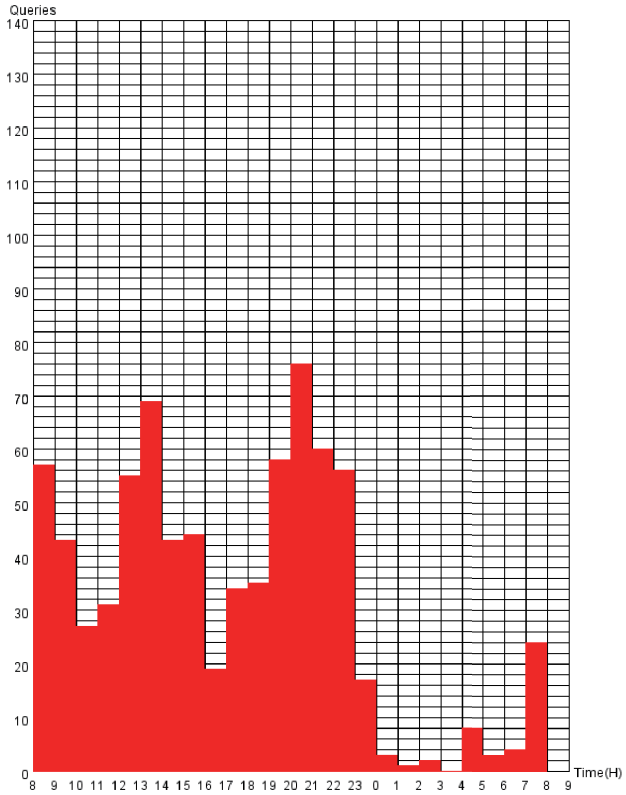


5: 高级搜索

7.14 搜索统计

通过记录用户搜索历史和点击历史，我们在首页和单独的系统负载页面给出了相关的统计信息。一方面可以让用户了解群体的搜索喜好，另一方面也可以让管理员了解系统的负载情况。图6中的资源名称后的数字表示此资源的一周搜索次数；站点名称后的次数表示此站点的历史用户点击数。图7给出了系统 24 小时的负载直方图，横坐标表示 24 时制的时间，纵坐标表示系统每小时的总搜索次数。

6: 热门资源和站点统计



7: 24 小时系统负载统计

7.15 搜索建议

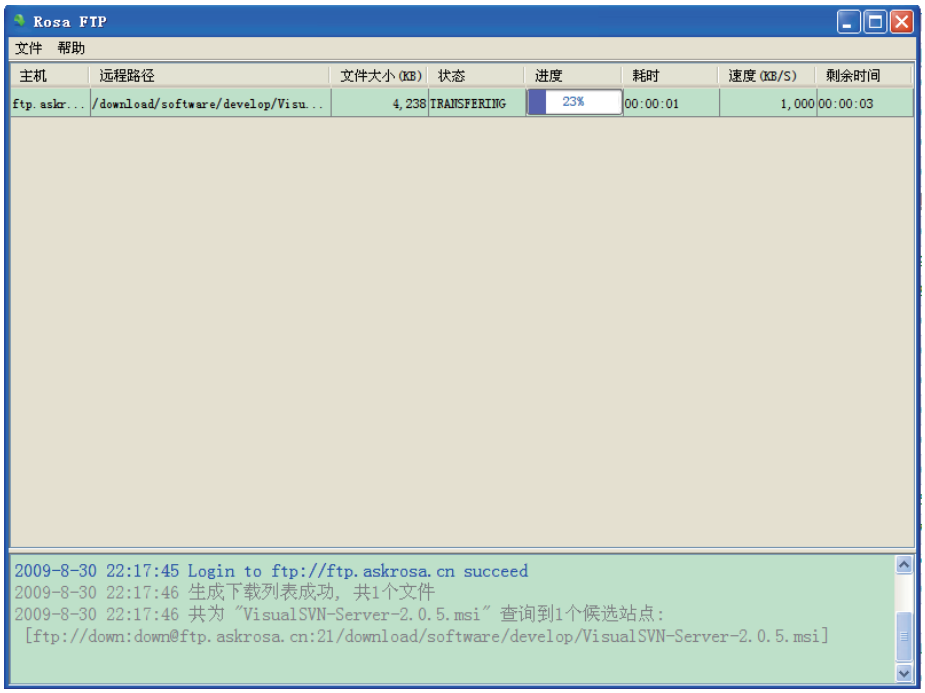
根据用户搜索历史，设计实现了搜索建议功能。当用户输入关键词时，根据输入给出最相关的历史搜索记录作为提示。参见图8。

8: 搜索建议

7.16 ROSA FTP

用户对搜索到的资源可以选择复制链接地址后使用常用的 FTP 下载软件下载，同时也可以选择使用 ROSA 提供的 ROSA FTP 多点下载工具下载。通过点击图9中标记的“启动 ROSA FTP”来启动下载工具，然后点击“操作”下方的箭头图标就可以添加下载项。图10给出了 ROSA FTP 下载过程的截图。

9: ROSA FTP 启动



10: ROSA FTP 下载

7.17 离线站点快照

由于 ROSA 保存了索引站点的资源信息，所以用户可以用离线的方式通过 ROSA 浏览 FTP 站点上的资源。见图11。

11: 站点快照

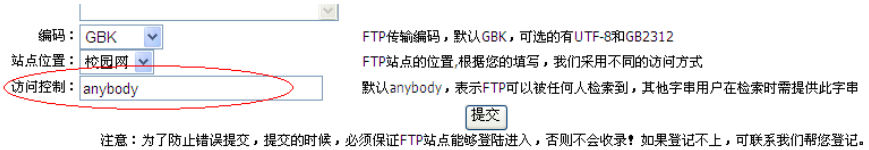
7.18 FTP 站点管理

1. 登记 FTP 站点

2. 更改站点信息

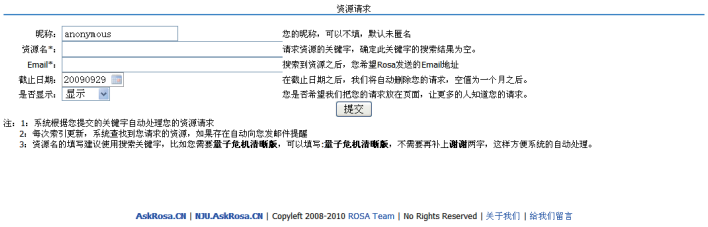
3. 自定义共享圈

FTP 站点管理员为自己的站点设定一个访问控制码，只有拥有这个访问控制码的用户才能检索到此站点的资源，用于用户之间私有资源的共享。



7.19 资源请求

用户可以在系统资源请求页面（见图12）提交资源请求信息。系统每次更新索引后检查用户请求资源是否已经存在，如果发现用户请求资源，则自动向用户发送请求资源已存在的邮件通知。



12: 资源请求

7.20 高度可配置

ROSA 在设计之初就考虑到了系统参数的可配置性，所以我们将系统的配置参数全放入到了配置文件中。用户可以通过修改配置文件而不需要修改代码来实现系统的配置。可配置的参数包括：线程池大小，代理服务器，数据库连接，可用缓存，文件日期格式，索引文件存放地址以及网络连接超时时间等。

8 问题与解决

- 系统运行中发现有些站点不支持 ls -lR 命令。对于这些站点目前的方法是检查在根目录发送 ls 命令和发送 ls -lR 命令返回条目数的关系，如果接近则认为不支持 ls -lR 命令，并将此信息存入数据库，下次就使用程序手动递归方式进行索引。但这种程序手动递归方式存在一个严重的偶发性超时问题，目前无法完美解决。

- Lucene 默认分词器 StandardAnalyzer 在使用中发现分词效果不理想，在此基础上我们加入了自己的分词规则。比如对于 [先知].Knowing.2009.BDRip.X264-TLF 这样的字符串，原分词器的分词结果为：“先” “知” “knowing.2009.bdrip.x264-tlf”，修改后的分词器的分词结果为：“先” “知” “know” “2009” “bdrip” “x” “264” “tlf”。通过这样的修改，能更针对地将 FTP 资源名称和路径分词，也便利了用户使用多种变化和自己喜好或擅长的搜索条件。

- ROSA FTP 多点下载程序的下载文件的并发写入会导致文件损坏，通过使用 FileChannel 解决了并发写入的问题。

- ROSA FTP 多点下载程序中的从 FTP 服务器传送文件分片还存在严重的 bug，目前尚未解决。

9 运行历史

目前 ROSA 搜索引擎自 2008 年六月开始运行，为大家提供了无偿高效的搜索服务，并且会一如既往地继续为大家提供更优质的搜索体验。在运行初期系统出现严重的不稳定停机，经过长达半年的运行测试，ROSA 系统与 2009 年初达到稳定状态。目前系统已可以无故障的长时间运行。相关的运行数据统计参看表4。

4: ROSA 运行统计

运行时间	2008/6/3---现在
收录站点	222[1]
索引资源文件总数	1781961[2]
历史访问量	459988
日平均检索次数	2000 左右
历史检索次数	828490
单次搜索耗时	500 毫秒左右

[1]校园网 167 个，教育网 54 个，公网 1 个（由于缺乏专属网络，目前只是使用代理索引公网站点，所以没有加入更多的公网站点）。

[2]有 13 个站点由于不明原因索引文件总数均为 0。

[1] Wikipedia. http://zh.wikipedia.org/zh-cn/Lucene.

[2] Wikipedia. http://zh.wikipedia.org/zh-cn/Struts.

[3] Ceki Gülcü. http://logging.apache.org/log4j/1.2/manual.html, 2002.

[4] JDom Organization. http://www.jdom.org/.

[5] Apache. http://commons.apache.org/net/.

[6] Apache. http://jakarta.apache.org/oro/.