FineSearch

解析 Google 等搜索引擎结果、提取网页正文利器

版本	0.1
作者	张知临
联系方式	zhzhl202@163.com
最后更新	2012/04/5
系统主页	https://code.google.com/p/fine-searcher/

第1章 FineSearch系统介绍

主要功能点：

1、向 Google 或百度等搜索引擎输入 Query，将结果页面中的 URL 解析出来。

2、输入指定的 URL，利用行密度算法将网页正文抽取出来。

3、利用 XML 动态配置搜索引擎。

1.1 提取搜索引擎结果页面中的URL

核心计划一句话就可以概括：

1、读取由用户设定的配置文件中搜索引擎的参数，构造所要搜索的 URL。通过这些 URL 将用户输入的关键字传递给各搜索引擎。

2、解析搜索引擎的结果页面，通过页面中特定的 Tag，将结果的 url 扣取出来。

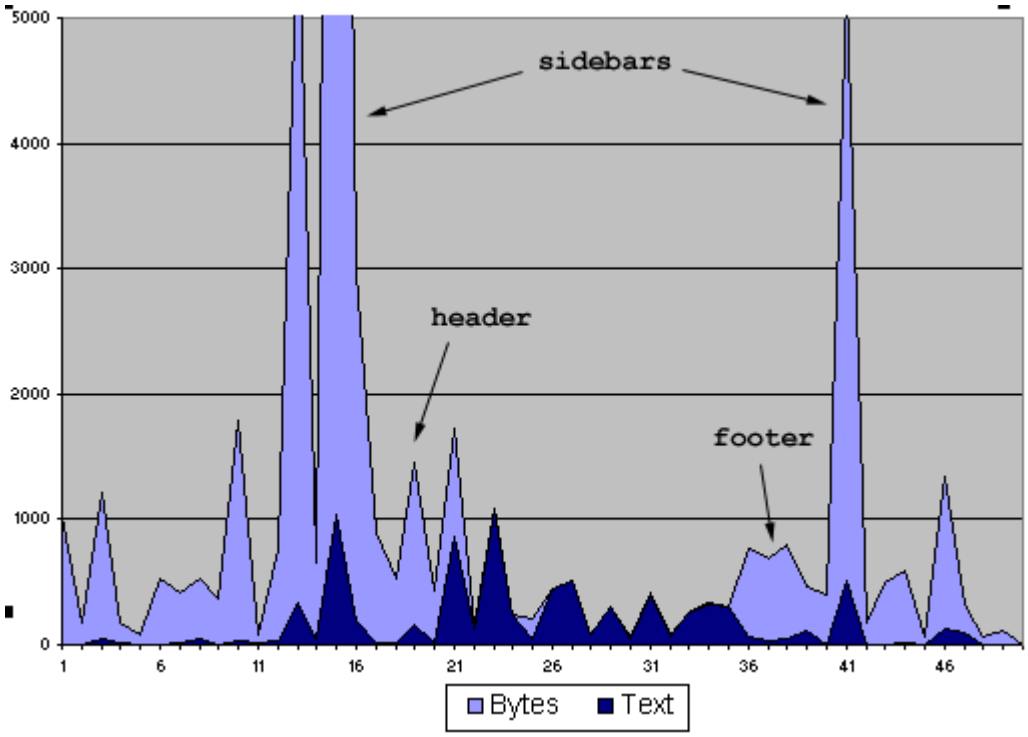
所用工具：HtmlParser

1.2 抽取网页正文

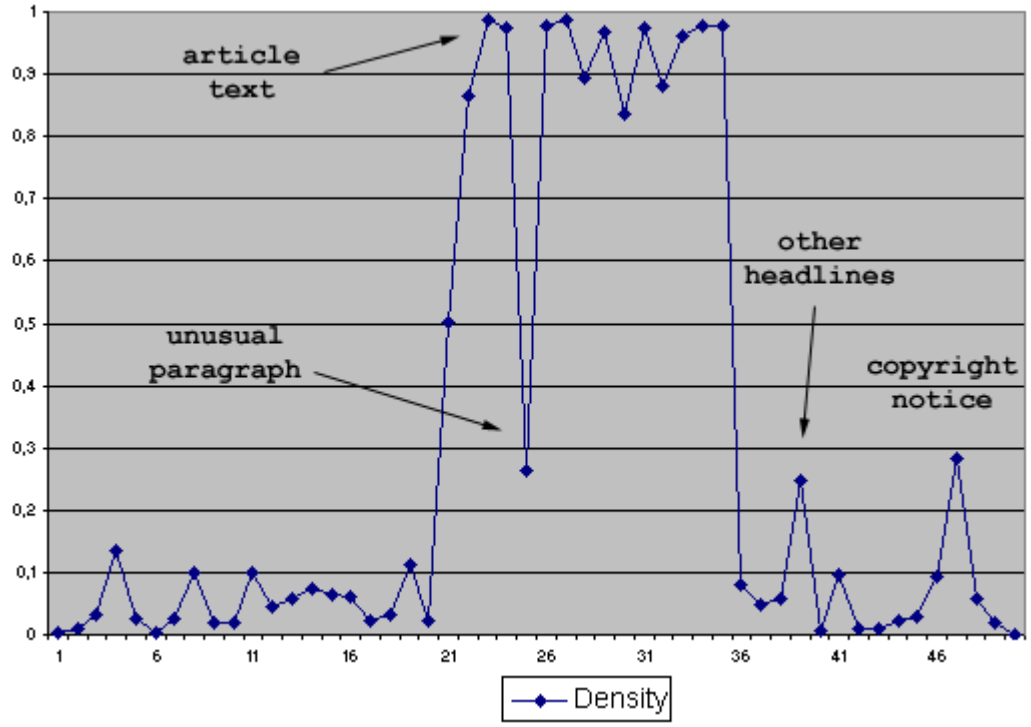
引自：http://blog.csdn.net/lanphaday/article/details/1741185

每个人手中都可能有一大堆讨论不同话题的 HTML 文档。但你真正感兴趣的内容可能隐藏于广告、布局表格或格式标记以及无数链接当中。甚至更糟的是，你希望那些来自菜单、页眉和页脚的文本能够被过滤掉。如果你不想为每种类型的 HTML 文件分别编写复杂的抽取程序的话，这里有一个解决方案：

首先来看两张图



从上面的原始输出你可以发现有些文本需要大量的 HTML 来编码，特别是标题、侧边栏、页眉和页脚。虽然 HTML 字节数的峰值多次出现，但大部分仍然低于平均值；我们也可以看到在大部分低 HTML 字节数的字段中，文本输出却相当高。通过计算文本与 HTML 字节数的比率（即密度）可以让我们更容易明白它们之间的关系：



密度值图更加清晰地表达了正文的密度更高，这是我们的工作的事实依据

因此算法的主要步骤如下：

1、解析 HTML 代码并记下处理的字节数。

2、以行或段的形式保存解析输出的文本。

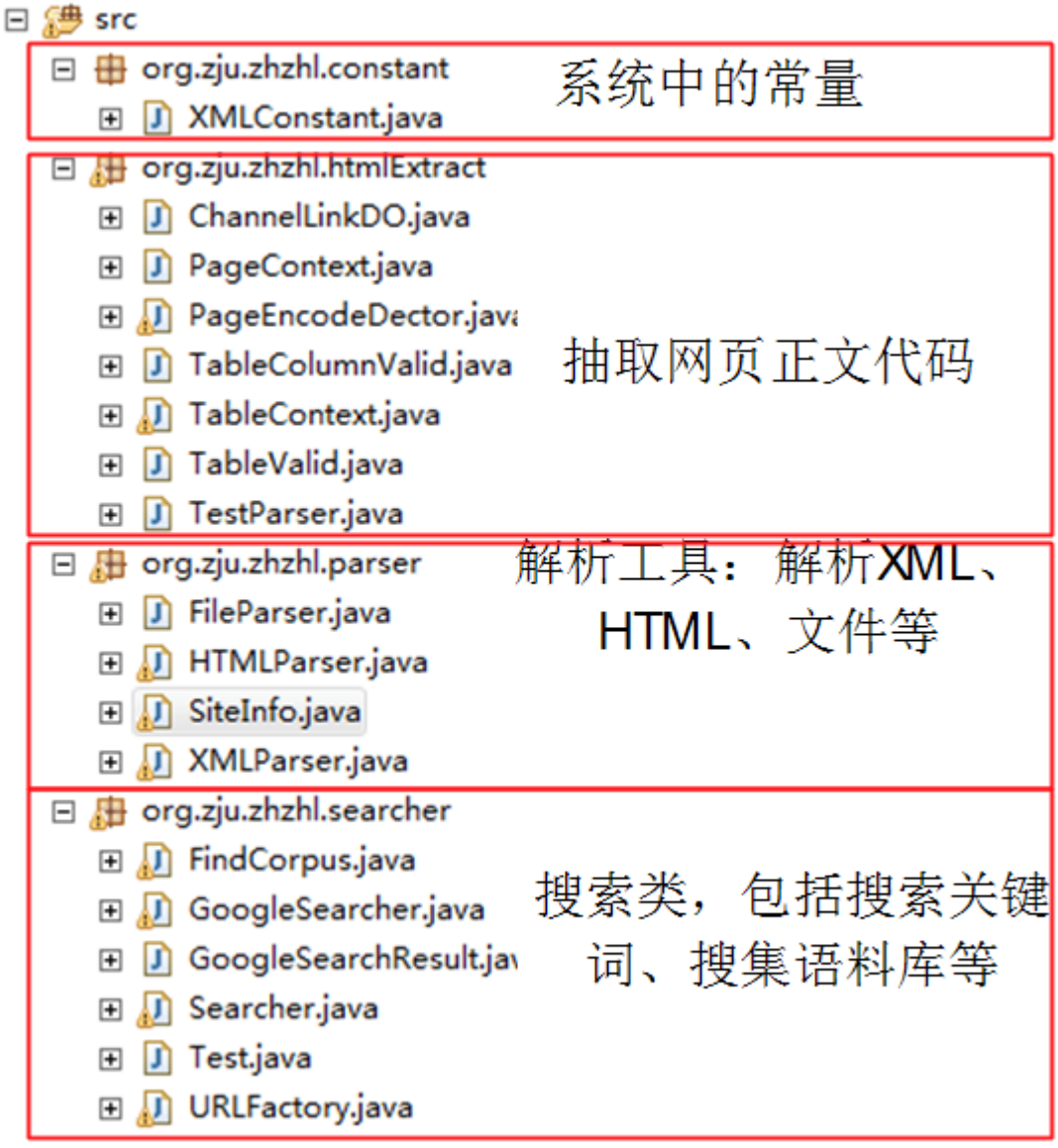
3、统计每一行文本相应的 HTML 代码的字节数

4、通过计算文本相对于字节数的比率来获取文本密度

5、最后用通过阈值或者是神经网络来决定这一行是不是正文的一部分。

1.3 系统模块

以下为系统程序模块图。其中



1.4 参数配置

以上为配置文件，这里我们以 Baidu 为例子，看一下看其源文件

```
1   <Sites>
2   <site name="baidu">
3       <status>true<status>
4       <baseUrl>http://www.baidu.com/s?</baseUrl>
5       <keyWord>wd</keyWord>
6       <encoding>gb2312</encoding>
7       <filters>
8           <attribute name="target">_blank</attribute>
9           <attribute name="onmousedown"></attribute>
10      </filters>
11      <options>
12          <option name="pn">3</option>
13      </options>
14  </site>
```

其中第 2 行：`<site name="baidu">` 这里的 name 唯一标示一个搜索引擎，由用户来输入。

第 3 行：标示该搜索引擎的状态：是否可用。设置成 true，就代表在搜索时使用该搜索引擎，而 false 则为不使用该搜索引擎。

第 4 行：为搜索引擎的基地址

第 5 行为搜索关键字在搜索引擎的 URL 中的表示符号。

第 6 行为关键字的编码格式。

第 7-10 行为在网页中定位目标搜索结果的属性。其中过滤 URL 采用的过滤器为 HasAttributeFilter，即找到能唯一表示 URL 所在的 Tag 中各个属性以及其 value。所以在配置时，需要在 filter 中填入相应的 Attribute 的名称以及其 value，如果 value 不确定，可以不填。

第 11-13 行为搜索引擎中一些可选的参数设置。如每页显示的数目。

1.5 各搜索引擎的参数意义

1.5.1 百度

url 很长一大段,都是些参数,cl=3 表示网页搜索,tn 表示来源站点,word 是关键词[有线怪谈 20090214],ie 表示编码方式,这里是 utf-8 编码.

百度搜索命令中的参数

必备参数:

wd——查询的关键词(Keyword)

pn——显示结果的页数(Page Number)

cl——搜索类型(Class),cl=3 为网页搜索

可选参数:

rn——搜索结果显示条数(Record Number),取值范围在 10--100 条之间,缺省设置 rn=10

ie——查询输入文字的编码(Input Encoding),缺省设置 ie=gb2312,即为简体中文

tn——提交搜索请求的来源站点

tn=baidulocal 表示百度站内搜索,返回的结果很干净,无广告干扰.

tn=baiducnnic 想把百度放在框架中吗?试试这个参数就可以了,是百度为 Cnnic 定制的

si——在限定的域名中搜索,比如想在 sofuc.com 的站内搜索可使用参数 si=sofuc.com,要使这个参数有效必须结合 ct 参数一起使用.

ct——此参数的值一般是一串数字,估计应该是搜索请求的验证码

si 和 ct 参数结合使用,比如在 sofuc.com 中搜索"wordpress",可用 :http://www.baidu.com/s?q=&ct=2097152&si=sofuc.com&ie=gb2312&cl=3&wd=wordpress

bs——上一次搜索的关键词(Before Search),估计与相关搜索有关

1.5.2 google 搜索参数

q－查询的关键词(Query)，百度对应的参数为 wd

hl－Google 搜索的界面语言(Interface Language)

hl=zh-CN 简体中文语言界面，我们用的 Google 中文就是这个参数。

hl=zh-TW 繁体中文语言界面，港台地区常使用

hl=en 英文语言界面

start－显示结果的页数,百度对应的参数为 pn

lr－搜索内容的语言限定(Language Restrict),限定只搜索某种语言的网页。如果 lr 参数为空，则为搜索所有网页。

常用的有：

lr=lang_zh-CN 只搜索简体中文网页

lr=lang_zh-TW 只搜索繁体中文网页

lr=lang_zh-CN|lang_zh-TW 搜索所有中文网页

lr=lang_en 只搜索英文网页

ie－查询输入文字的编码(Input Encoding),Google 缺省设置 ie=utf-8,即请求 Google 搜索时参数 q 的值是一段 utf-8 编码的文字，如果要直接使用中文，可以设置 ie=gb2312,即为简体中文编码

oe－搜索返回页面的编码(Output Encoding),Google 缺省设置 oe=utf-8

num－搜索结果显示条数(Number),取值范围在 10－100 条之间，缺省设置 num=10,百度对应的参数为 rn

newwindow－是否开启新窗口以显示查询结果。 缺省设置 newwindow=1，在新窗口打开网页

safe－安全搜索选项(SafeSearch),设置该参数可以过滤成人内容，缺省设置 safe 为空，即不过滤成人内容，设置为 safe=vss，即过滤成人内容。

q－你要查询的词

start－基于零的第一个期望的搜索结果的索引.

maxResults－每次期望查询结果数，最大查询值为 10。

注意: 如果你的查询结果没有更多的匹配项，真实的查询结果数可能小于你的请求数。

lr－语言限定- 限定在一种或多种语言的范围内的搜索.

ie－输入编码 － Google 已经不赞成大家使用这个这个参数了，而且这个参数已经被忽视了。所有对 API 请求都必须是 UTF-8 编码。

oe－输出编码 － Google 已经不赞成大家使用这个这个参数了，而且这个参数已经被忽视了。所有对 API 请求都必须是 UTF-8 编码。

1.5.3 Soso

“w=” 搜索关键词

“bs=” 上次搜索关键词

组合关键词搜索链接符号为“+”

http://www.soso.com/q?gid=&c … &w=%B5%E7%D3%B0

http://www.soso.com/q?pid=s.idx&w=%BF%DA%B1%AE+%C9%FA%BB%EE

1.5.4 youdao

“q=” 搜索关键词

组合关键词搜索连接符号“+”，顺次链接下来。如下 refer 链接

http://www.youdao.com/search?q=%E7%94%B5%E5%BD%B1 %20%E9%98%BF%E5%87%A1%E8%BE%BE&keyfrom=web.suggest

1.5.5 sougou

“query=” 搜索关键词

组合关键词搜索链接符号为“+”

http://www.sogou.com/web?query=% … p=40040100&dp=1

1.5.6 yahoo

“p=” 搜索关键词

组合关键词搜索链接符号为“+”

http://one.cn.yahoo.com/s?p=%E5% … mp;v=web&pid=hp

参考资料

1) http://code.google.com/intl/zh-CN/apis/customsearch/v1/overview.html
2) http://www.cnblogs.com/elric/archive/2010/07/25/1784522.html
3) http://www.cnblogs.com/elric/archive/2010/07/27/1786484.html 利用 Google AJAX SearchAPI 并获取搜索结果
4) http://hi.baidu.com/lck0502/blog/item/2336663684fb92380b55a96d.html 用 JAVA 调用 Google AJAX SearchAPI 并获取搜索结果

5) http://www.path8.net/tn/archives/2383 百度 Google 搜索引擎 url 查询参数详解

6)