第五周工作报告

丘骏鹏

2013-03-29 Fri

1 主要工作内容

这一周主要开始搭建程序的基本框架和实现数据的一些预处理。

1.1 简单的数据统计

首先对数据做了一些基本的统计：

	blog.zip	news.zip	other.zip
文件个数	59998	81561	183635

新浪博客的目录页和详细页可以用 URL 区分，比如某个博主的目录页为 http://blog.sina.com.cn/u/1439351555 ，他的某篇文章的 URL 格式为 http://blog.sina.com.cn/s/blog_55cac30301016yb1.html 。因此对于博客数据可以用 URL 正则进行过滤。通过 shell 命令进行统计后，得到的结果为：blog 中不带 html 后缀的文件有 23430，带有 html 后缀的文件有 36568。可以初步判断 blog 数据中有用的详细页数据为 36568。

blog 数据中还有少量的 404 页面，不符合上面的 url 组成规律的大部分是 404 页面。我用简单的"404 Not Found" 字符串进行过滤，以下的 shell 命令：

```
ls blog/ | xargs -I{} grep "404 Not Found" -c {} | awk '{sum+=$1};END{print sum}'
```

结果为 174，即有 174 个没用的 404 页面。这一步也可以先排除一些没用的 404 页面。（当然，如果严格来说包含“404 Not Found”的页面不一定就是 404 页面，但是这个简单的筛选方法对于这个实验来说应该够了）

结合以上方法可以做一次粗粒度的筛选，然后将可用的实验页面初步筛选出来。

我从 blog 和 news 中各抽取了 1000 个文件作为样本，用于测试使用。初步打算先用 blog 中的数据进行后续的实验。

1.2 实现

- 实现语言：打算采用 Java+Scala，Scala 是 JVM 上的静态类型语言，可以和 Java 之间无缝操作，支持面向对象和函数式等编程范式。我之前写过 Java+Scala 的大作业，对两者都比较熟悉。
- 关于实现方面的一些库的选择：

  - 关于字符集探测库：上网搜了一些相关的库和相应的评价，决定选择icu4j。这个库目前仍在活跃开发中，对各个字符集的支持很成熟。目前的测试结果来看可以对下载的 HTML 的字符集进行正常进行分辨。
  - HTML parser：目前初步决定采用Jsoup，写了一些基本的程序进行初步测试，比较符合需求。
  - 日志系统：用的是 twitter 包装的 util-logging 包，是在 java.util.logging 上的一个简单包装。
  - 配置文件读取：基于 java 实现的一个配置文件读取库，支持 java 原生的 properties 文件格式、JSON 格式和 HOCON(Human-Optimized Config Object Notation) 格式的配置文件。维护者是http://typesafe.com, 项目地址:https://github.com/typesafehub/config

- 目前已经实现的部分：目前写好了简单的预处理部分，包括先将一些不关心的标签去掉，如 script,style,link,br,img,strong,em,font。这样可以先将文档的结果简化，后期 DOM 的解析会快，因为 DOM 解析一般会比较耗内存，而且速度较慢，因此这些预处理是必要的。另一方面，从我们需要关心的模板细度来说，我们也不需要这种细粒度的标签。不过这部分还需要不断修改验证，可以多去掉一些标签，但是不能将一些必要的去掉了。

  同时系统的日志系统和配置文件解析部分也已经基本完成。

2 下周工作

下一周打算将网页的相似度计算和聚类部分完成，先采用一些基于 path 和 tag 的简单算法计算相似度。