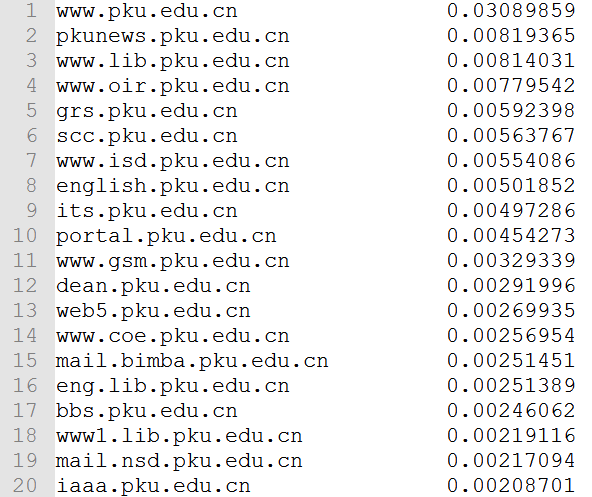
**网页采集与PageRank计算**

学号：1300012758 姓名：张闻涛

**数据汇报与分析：**

alpha=0.95 （重新从某一个网站开始浏览概率为0.05）



前十名分析：

[www.pku.edu.cn](http://www.pku.edu.cn) its.pku.edu.cn portal.pku.edu.cn [www.lib.pku.edu.cn](http://www.lib.pku.edu.cn) pkunews.pku.edu.cn 是北大不管什么专业的人都常用的网站，上榜理所当然。

[www.ori.pku.edu.cn](http://www.ori.pku.edu.cn) [www.isd.pku.edu.cn](http://www.isd.pku.edu.cn) 是北大国际合作部港澳台办公室、汉语国际推广办公室和留学生办公室，体现出北大对国际交流合作方面十分重视，建设国际一流大学有望。

grs.pku.edu.cn 是北大研究生院网站，涉及到硕士、博士报名、研究生信息管理、以及发布各种招收、评奖、交流等信息。官方性很强，等我们本科毕业后应该是一个重要的网站。

scc.pku.edu.cn 学生就业指导服务中心，找工作确实是大学生、研究生的重要环节。

[www.gsm.pku.edu.cn](http://www.gsm.pku.edu.cn) 光华管理学院，名次领先其他分院，光华在北大的地位果然很高。

留学、就业、研究生；学习、上网、看新闻；这个结果很贴近北大学生实际生活中的使用状况。

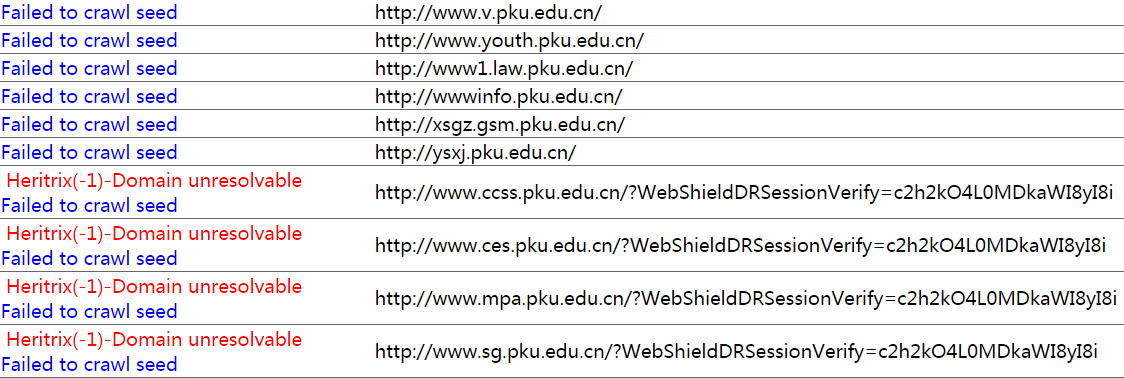
**完成这个作业的过程：**

对于这个作业，开始我有两种不同的想法。第一种是在Crawler中，对网页之间的跳转关系直接进行构图，然后在Crawler中进行PageRank算法。第二种是将这两部分开进行，首先把域名为pku.edu.cn的网页全都爬取下来，然后再用另外一个程序根据爬取的网页建立关系，计算PageRank。

由于我曾经用Java实现过一个简易的爬虫，所以我首先采用了第一种做法，在当时的代码上进行改动。但是遇到了一个小麻烦：没法获取https网页的内容。为了提高效率，我就没有再去改进，转而使用第二种做法。

我用的开源爬虫工具是Heritrix，根据作业要求，设置seed为[www.pku.edu.cn](http://www.pku.edu.cn)，在过滤条件中加入判断网页链接中是否有.pku.edu.cn子串，并且只爬取html和htm文件，这些设置方法我参考了网上的教程。设置完毕后，爬了几次测试后，能够成功运行了。

然而，我发现，爬虫运行至50%以上时，效率变得非常低下，一天都没有明显进度增长（这个百分比是根据当前已爬取网页除以队列中和已爬网页和算得的）。于是我就把进程掐掉了，写了小程序来遍历那些已经爬到的html文件，寻找满足条件但是没有被爬到的网页结点。发现确实有好多网页没有爬到，我就把他们一起作为种子，再放到Crawler中爬一遍。

这一遍我发现，这些Url中，确实有一些没法通过Crawler访问到。 

我尝试设置忽略机器人采集协议，和有个masquerade的选项，也没有效果。这些网页有些在浏览器中不能访问，有些能正常访问，也许是对Crawler识别比较严吧。

将两次采集到的网页合并后，作为计算PageRank的数据。通过文件夹操作和文件读入分析网页之间的关系，建立邻接矩阵。我选取了各个主网页作为结点，跳转链接作为邻接矩阵中的边，各个子网页中的链接也算作从主网页跳转至另一个主网页的边，但是子网页跳转的链接加入图中时，将其乘上从主网页到这个子网页的概率（我觉得这样比较符合随机游走模型）。建图完成后，通过直接迭代计算出PageRank值。

**改进和建议：**

排名中有不少实际上重复的网页，比如english前缀的英文版网页，又如web5，www1前缀的网页，若加以识别，可能可以改进效果。

另外，一些应用、娱乐性，更贴近生活的网站：比如bbs，排名不高而实际生活中访问量很高。我猜可能是因为种子网页官方性太强了。如果从更大范围的网页进行PageRank计算，可能未名bbs的排名会有所提升。

**参考资料:**

Java8 API文档

[**http://blog.sina.com.cn/s/blog\_6cc084c90100nf39.html**](http://blog.sina.com.cn/s/blog_6cc084c90100nf39.html)

[**http://www.ibm.com/developerworks/cn/opensource/os-cn-heritrix/**](http://www.ibm.com/developerworks/cn/opensource/os-cn-heritrix/)