PkuJava32课程项目

1501210718王乙闲

1501210890 程 权

1501210381林雄辉

本组所完成的课程项目是通过爬虫技术，以广度优先策略，从BlogJava等技术类博客网站中爬取博客内容，对博客内容进行聚类分析，构建模型参数和分析结果，从而实现对新博文的分类预测功能。

一、爬取博客内容部分类设计简洁

1、Crawler 🡺 主程序入口处；利用filter过滤博主主页、博主随笔分类、博客内容取得符合条件的网页链接，并通过限制条件对队列进行访问操作。

2、Queue 🡺 URL数据结构队列，使用LinkedList实现队列基本出栈、入栈、判断是否为空等操作。

3、LinkDB 🡺 记录URL队列，使用HashSet和Queue集合来实现已访问和待访问的URL队列的记录及基本操作。

4、LinkFilte 🡺 LinkFilter接口，使得提取出来的URL只与我们想要的网页内容相关。

5、LinkParser 🡺 使用开源工具包HtmlParser及其Filter过滤链接对Html页面进行处理，提取URL和想要的文本内容。

6、BlogContext 🡺对博客内容和标题进行抓取。

OrFilter设置过滤<a>和<frame>标签（LinkTag和htmlparser.NodeFilter）

htmlparser.filters.AndFilter、htmlparser.filters.HasAttributeFilter和htmlparser.filters.TagNameFilter等过滤网页中想要提取的内容

htmlparser.util.ParserException做异常处理

org.htmlparser.Tag、htmlparser.util.NodeList处理取到的内容，为进一步数据的存储做准备

7、SaveToLocal 🡺 本地存储BufferedWriter的io操作。

二、爬取博客内容部分程序优化

初始版本中，是通过 博客博主 🡪 博主博客 抓取过程，抓取博客链接，从而进行博客内容的获取。该方法可以实现部分博客内容爬取，但由于博主不同主页形式的内容标签不同，对filter过滤条件带了较为麻烦的补充，使得代码冗余量大，爬取过程速度受限。

优化后的版本，采取 博客博主汇总 🡪 博主主页 🡪 博主随笔分类 🡪 博客内容 的抓取过程进行链接队列的访问，采取广度优先策略，依次对指定数量的博主的随笔分类中的博客内容进行爬取。同时将初始版本中的一个filter增加至三个，分别对博主随笔分类、分类随笔文章和博客文章三个链接做过滤，进而进行条件匹配。

通过优化后，成功从BlogJava博客中爬取到122M的博客数据。

三、用到的技术简述

1、数据爬取：

采取广度优先的搜索策略，按博客博主汇总 🡪 博主博客主页 🡪 博主随笔分类 🡪 博客内容 的顺序进行URL的访问操作，以博客标题做文本文件名称，以博客内容做文本文件内容，爬取1万个博文内容，共计84.2M

HashSet数据类型：具有冗余检查功能，存储已访问的URL；

Queue数据类型：存储未访问的RUL；

LinkedList 数据结构队列：实现对URL队列的访问操作；

LinkedFilter接口：实现指定URL的提取限定；

HtmlParser库 ：对网页包含的内容进行分析、过滤和提取；

BufferedWriter和FileWriter：实现博客内容的本地存储。

2、聚类分析、分类预测

IK Analyzer分词工具包：对每篇博文进行分词；

K-Means聚类算法：聚类分析，并确定描述簇个数 （k=22）；

TF-IDF：选取特征词作为标签；