PkuJava32课程项目

1501210718王乙闲

1501210890 程 权

1501210381林雄辉

本组要完成的课程项目是有关技术博客的爬取及其内容的分析，选取的目标对象是JavaBlog博客。

在经过课题讨论、基础知识的学习后，我们队实现该课题的架构、基本技术点、技术难点以及任务的人员和时间分配都有了更详细的规划。

一、架构分析

对于技术博客内容获取的爬虫部分，主要涉及到URL数据结构队列、记录URL队列、URL过滤、Html页面内容处理和本地存储五个部分。

对于技术博客内容分类分析部分，主要涉及到利用机器学习的分类算法对文章进行有指导的分类，然后预测新文章的类别，自动给用户提供新文章的分类标签。

1、URL队列的存储与提取：

首先设定种子URL，将其加入到待访问的URL队列中，待访问的种子URL出队列，放入已访问的URL队列中，由URL出发，通过过滤条件的设定，提取种子URL网页中全部符合条件的URL链接，对于每一个URL链接均判断其是否在已访问的URL集合中，即判断其是否被访问过，若未访问过，则加入待访问的URL集合中，等待出队列进行访问；若访问，过则抛弃。每一个访问过的URL均要加入到已访问的URL集合中。

2、Html页面内容的提取：

对html页面标签元素进行过滤、定位，提取所需要的内容，该课题中提取的是博客的标题和正文内容的NodeList，对其元素目标提取，放入字符串中，待保存到本地。

3、本地存储：

创建本地IO读写流，将每个博客的内容分别存入一个本地txt文件中，以博客标题作为文件名，关闭读写流。

4、博客内容分类分析：

对博客内容依次进行分词、去停词、词频计算、提取关键词，从而对博客内容进行有指导的分类分析。

5、预测分类标签：

通过分类分析，预测新文章的类别，自动给用户提供新文章的分类标签。

二、基本技术点

1、LinkedList实现队列的基本出栈、入栈和判断是否为空等操作；

2、HashSet和Queue集合实现已访问和待访问的URL队列的记录及广度优先提取链接等基本操作；

3、LinkFilter接口，使得提取出来的URL只与我们想要的网页内容相关；

4、HtmlParser进行URL链接的过滤、提取以及网页中目标文本内容的提取；

5、io读写流进行本地存储。

6、文本关键词提取；

7、关键词概率、权重分析；

8、机器学习分类算法的选取；

9、新文章的类别预测分析。

三、技术难点

1、在基于广度优先策略的队列记录和查找中，要实现URL保存的不重复性和查找的快速性，这对队列的存储方法和存储集合的选取有着较高的要求；

2、开源工具包HtmlParser中包含很多的网页标签及内容的过滤方法，需要逐一做以了解，选取适合该课题的方法；

3、由于成员对java前端知识了解较少，在设置过滤标签内容时会较为吃力；

4、对从网页中提取到的目标内容，做正确的处理，提取想要的目标文本格式。

5、文本内容的关键词提取；

6、准确、高效的机器学习分类算法的选取。

四、任务的人员和时间分配

2015.12.1—2015.12.9

王乙闲、程权、林雄辉大致了解课题内容，选取课题。

2015.12.10—2015.12.18

林雄辉对基于广度优先的URL队列操作进行架构设计和技术难点剖析；

王乙闲了解HtmlParser开源工具包的基本功能，提取完成该课题所需的方法，对URL链接的过滤、提取以及网页中目标文本内容的提取进行架构设计和技术难点剖析；研究本地存储操作方法

程权利用机器学习的分类算法对文章进行有指导的分类，然后预测新文章的类别，自动给用户提供新文章的分类标签。

该阶段中，每个人均对自己负责的技术点进行简单demo编写。

2015.12.19—2015.12.23

王乙闲 、程权、林雄辉对课题整体架构进行整合，确定技术难点及解决方案。

2015.12.24—2015.01.01

完成各自实现自己所负责的模块编码、测试。

2015.01.02—2015.01.04

完成代码的整合、测试。

2015.01.04—2015.01.06

对项目代码进行优化。