Java 课程项目

期末大作业

PKUJAVA07 组　　刘翀 1501210948

黄义珊 1501210919　　普筱越 1501210940

微博舆论爬虫

阶段性报告二

2015/12/22

一、架构中的技术点填充

项目架构中的技术点分为以下：

1.从网站上获取网页

2.收集网页上的微博链接Link

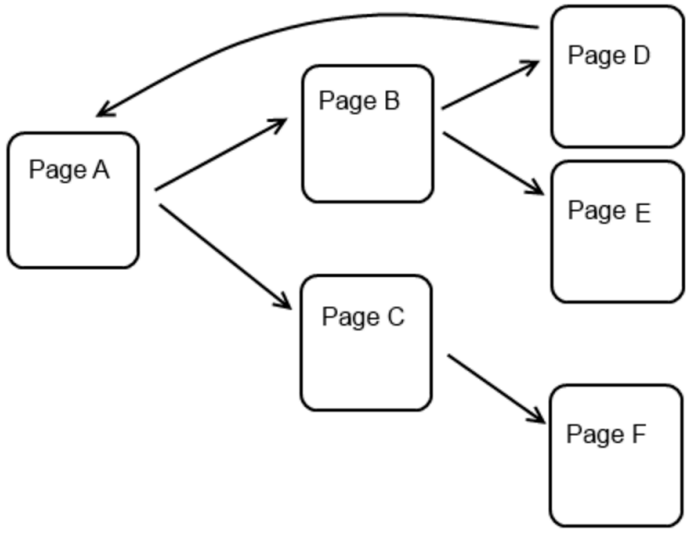
3.收集网页上的文字信息

4.检查我们需要的关键字是否在文档里。

5.访问下一条微博

二、明确技术难点

技术难点：



如图，如果从PageD访问回PageA的话会造成死循环。

本组经过讨论需要注意：

1. 纪录爬虫之前访问过的页面（微博）。不可能只访问一次，如PageB那个页面有两个链接。

2. 在访问时设置一个阀值，使得访问数量达到可控而非无限。阀值大概是一个页面大致的有效链接数。（有效链接：含有有效微博内容的链接。）

3. Http请求，下载下来页面（文档）的解析将使用开源项目Jsoup。

4. 存储将使用MySQL数据库。

三、分配任务

刘翀：网页抓取。

黄义珊：文档解析。

普筱越：数据持久化。

四、确定时间点

12月28日：第一个版本的爬虫实现完毕，第一次抓取实验。

12月30日：根据第一次实验的问题，对第一个版本迭代。

1月2日：第二个版本的爬虫实现完毕，第二次抓取实验。

1月3日：设计数据分析与处理。

1月5日：文档撰写，准备报告。

五、Demo

1. 项目的规模

经资料查找，2013年有70多个账户的微博分析超过1000万。

我们的项目规模为70个左右意见领袖的微博账户。爬取转发量最多的微博，微博中的意见领袖爬取，我们的爬虫将抓取转发量和点赞数超过阀值的微博内容。

2. 项目流程及重点

（1）找到影响力较高的意见领袖链接URL

（2）根据这个URL爬取其近1周的微博。

（3）将微博进行转发量和点赞量的统计和解析，当转发量和点赞量超过一定阀值时，将该微博存储到MySQL数据库中