|  |  |  |  |
| --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | |  |  | | --- | --- | | 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

阶段性报告一

2015/12/11

1. 选题结果

微博舆情分析。通过用户发布的微博获取到最近的热点话题。

1. 技术调研

目前较为流行的Java爬虫框架由Nutch，Heritrix和WebCollector等等。本组欲采用具备扩展性的WebMagic框架进行爬虫设计。Apache HttpComponents 4.3 提供HTTP接口，用来向目标网址提交HTTP请求，以获取网页的内容。 [HTML Parser 2.0](http://htmlparser.sourceforge.net/" \t "_blank) 用来解析网页，从DOM节点中提取网址链接。MySQL Connector连接Java程序和MySQL，然后就可以用Java代码操作数据库。

1. 项目架构

若把微博页面的所有链接看作一个图的话，为了避免微博链接过分深度的抓取，我们准备选用广度遍历图来抓取信息。

1. 项目流程

本组准备爬取新浪微博的评论数、转发数和点赞数以及相应的微博。存入数据库或文本文件。最后选取评论数或转发数点赞数较高的微博内容作为社会舆情即近期的热点话题。

1. 现有产品比对

现有微博爬虫产品爬取内容相对较为繁杂，并无太大的针对性，如weibobee爬取了用户的id与简介等数据，但是冗余的数据为后期的舆论关键词带来了数据处理与过滤的困难。本产品专门针对微博的舆论热点设计。