文章编号:1671-4598(2019)07-0128-04 DOI:10.16526/j.cnki.11—4762/tp.2019.07.028 中图分类号:TP391.3 文献标识码:A

针对微博的免登录分布式网络爬虫的研究

王 林,刘星辰
(西安理工大学 自动化与信息工程学院,西安 710000)

摘要:微博作为优质的数据源,其中的数据非常适合做舆情分析等;新浪官方提供的API限制数据采集速度,而利用模拟登录的网络爬虫采集数据又相对复杂且会降低效率;针对这些问题,设计了一个免登录的微博网络爬虫;通过实验表明,该爬虫可以更快的对微博数据进行完整稳定的采集;随着对数据需求量越来越大,单机网络爬虫已经不足以满足要求,将Hadoop分布式计算平台与免登录爬虫相结合,设计了一个基于MapReduce的分布式网络爬虫系统,利用多台计算机组成的集群,实现短时间内免登录抓取海量微博数据;通过实验证明,该爬虫系统可以每天稳定抓取近千万条微博。

关键词:免登录网络爬虫;分布式网络爬虫;Hadoop;MapReduce

Research on Distributed Web Crawler without Login for Microblog

Wang Lin, Liu Xingchen
(College of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710000, China)

Abstract: Weibo is a good source of data, and the data is very suitable for public opinion analysis. The API provided by Sina officially limits the speed of data collection, and the network crawler using simulated login is relatively complicated and reduces efficiency. For these problems, a crawler without login for Weibo is designed. Experiments show that the crawler can perform complete and stable collection of Weibo data more quickly. With the increasing demand for data, the single network crawler can't meet the requirements. The Hadoop distributed computing platform is combined with the crawler without login to design a distributed network crawler system based on MapReduce. Using a cluster of multiple computers, you can capture massive amounts of Weibo data in a short period of time. Through experiments, the crawler system can stably capture nearly 10 million micro blog per day.

Keywords: crawler without login; distributed web crawler; Hadoop; MapReduce

0 引言

随着人类社会进入互联网时代,数据已经成为了一种新的资源。人们在互联网上挖掘数据资源后,进行大数据分析,可以产生巨大的社会和经济价值。

新浪微博作为国内最大的微博平台已经深入人们生活,2018年6月月活跃用户数增至4.31亿,日活跃用户数增至1.9亿,每天新产生微博数千万条。微博[1]具有传播速度快,实时性高,覆盖面广等特点,使得其中的数据具有很高的价值。尤其是它的实时性,已经让其成为舆情分析最好的数据源。

目前获取微博数据通用的解决方案是新浪官方提供的API和网络爬虫。但是官方API严格限制访问频率,再加上新浪设置了诸多反爬虫障碍,使得快速获取微博中的海量数据成为了难题。

廉捷[2]等人提取采用官方API和普通网络爬虫的方法采集数据,但是由于API的访问限制,获取大数据量时速率明显较慢;黄延炜、刘嘉勇[3]提出将微博官方API和基于网络数据流的微博采集方法相结合的方案,虽然数据抓取速度相对更快,但是依然没有突破官方API的访问限制,还牺牲了一定的数据完整性;孙青云[4]等人提出了基于模拟登录的网络爬虫采集方案,打破了API的访问限制,解决了传统的网络爬虫需要身份验证的问题,但是由于增加了模拟登录操作以及单机计算能力的限制,数据获取速度依然不足以满足对海量数据获取的要求。

另外,在对分布式网络爬虫的研究方面,斯坦福大学的Cho J和Garcia-Molina[5]提出了多个分布式网络爬虫架构并且首次给出了分布式网络爬虫的分类方法和评价标准等一系列基本概念,认为分布式网络爬虫与单机爬虫相比,具有高扩展性和减少网络负载的优势,为分布式网络爬虫的后续研究打下了基础。DL Quoc等人提出了一种地理分布式的网络爬虫系统UniCrawl[6];还有Apache基金会资助的开源网络爬虫项目Nutch。

本文以新浪微博这个优质的社交平台为数据源,先设计了一个免登录的网络爬虫,又将Hadoop大数据平台与该爬虫相结合,设计了一个免登录的分布式网络爬虫系统,可以很好解决海量数据挖掘的问题。主要工作如下:

1)对于新浪微博,设计了免登录的爬虫程序,实现了比模拟登录爬虫更快的数据抓取,并且保证数据的完整性

收稿日期:2018-12-24; **修回日期**:2019-01-15。
基金项目:陕西省科技计划重点项目资助(2017ZDCXL-GY-05-03)。
作者简介:王 林(1963-),男,江苏东台人,博士,教授,主要从事大数据、数据挖掘、计算机视觉方向的研究。

和程序的稳定性。

2）设计了一个分布式爬虫系统。利用 Hadoop 分布式计算平台，将（1）中设计的爬虫程序 MapReduce 化，利用多台计算机的计算能力，实现更加快速的信息获取。

1　免登录微博爬虫的设计

1.1　新浪官方 API

新浪微博开放平台开放了包括微博、评论、用户及关系在内的二十余类接口，通过 Oauth2.0 用户授权后即可在任意开发环境下使用。虽然新浪微博 API 提供的功能丰富齐全，但是由于对访问速度有限制，见表 1，这样的限制不满足我们想要快速抓取海量用户微博数据的要求，所以我们选择利用网络爬虫来获取微博数据。

表 1　新浪微博官方 API 访问限制表　　次/小时

针对一个服务器 IP 的请求次数限制					
授权方式	测试授权	普通授权	中级授权	高级授权	合作授权
总限制	1000	10000	20000	30000	40000
针对一个用户在使用一个应用请求次数限制					
总限制	150	1000	1500	2000	4000

1.2　免登录稳定抓取微博数据

用户访问微博时需要登录才能完整获取信息。网络爬虫为了获取完整数据通常也需要设置模拟登录操作。但是这个操作相对复杂，而且需要与 Web 服务器多一次交互，所以速度会因此减慢。

我们发现了一个更好的数据采集方法，即通过解析动态网页 XHR 的 URL，来获取微博动态网页的源代码，这样可以巧妙地实现免登录抓取微博动态网页，同时保证了数据的完整性。

我们发现微博用户首页 XHR 的 URL 为固定格式，如下：

https://m.weibo.cn/api/container/getIndex?type=uid&value=+ID+&containerid=107603+ID+&page=1

只要通过用户 ID 和该固定格式，就可以生成每个用户的微博首页 XHR 的 URL，然后直接利用爬虫解析这个 URL，就可以跳过登录操作，免登录的抓取微博数据。

另外，由于微博互动量巨大，服务器压力较大，如果快速大规模爬取微博数据时，会触发微博预设的爬虫检测机制，服务器一旦检测访问为爬虫，则拒绝其访问，让爬虫难以长时间稳定运行。所以，我们利用付费 IP 代理建立 IP 代理池，让我们的爬虫程序随机切换 IP 代理池中的 IP，将爬虫“伪装”成不同地点的用户进行访问，规避微博的反爬虫系统的检测。

1.3　设计针对微博的免登录网络爬虫

首先，由于微博 PC 端网页结构复杂，不利于信息抓取，而其移动端网页结构相对简单，主要数据如微博内容、评论、时间等信息并没有缺失，所以我们选择对微博的移动端网页进行爬取。

用户是微博的基本单元，微博用户通过互相“关注”形成了如图 1 网状结构。所有的数据采集都需以用户微博首页为起点，但是微博并不提供所有的用户列表，所以我们首先需要尽可能多的获取微博用户列表。

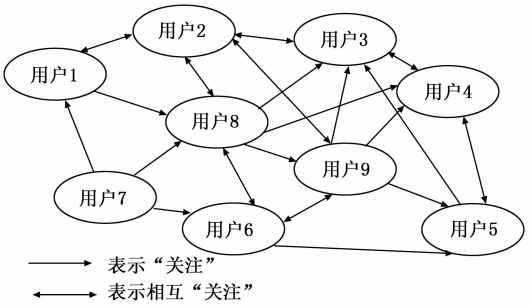

图 1　微博用户网络结构图

我们可以通过这个网络，利用爬虫，获取大量的用户 ID。由于用户形成的网络结构复杂，深度较大，很容易形成抓取“黑洞”，而且随着深度的不断增加，爬取到的用户重复度会越来越高，再加上我们设计的网络爬虫无需进行特定主题搜索，所以我们设计了一个基于广度优先策略的通用网络爬虫来获取微博用户列表。

在获取到微博用户列表后，再设计一个通用网络爬虫，遍历用户列表中的 URL，解析出网页源代码，利用正则表达式或网页标签抓取每个用户的数据。综上，我们设计了如图 2 的免登录网络爬虫系统。

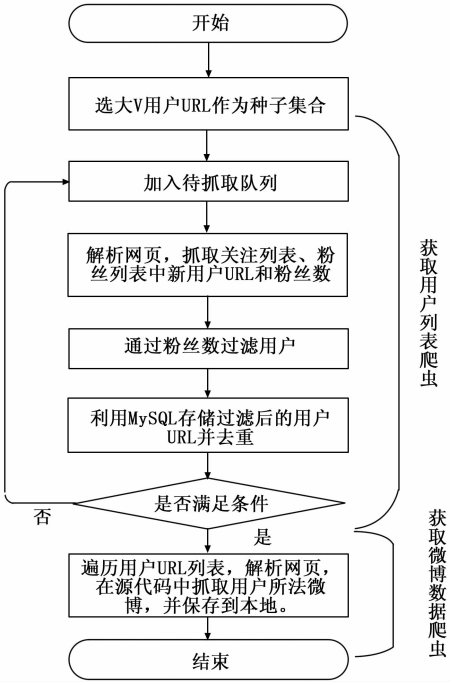

图 2　免登录微博网络爬虫工作流程

具体过程如下：

1）我们从微博上选取各个领域一些大 V 用户首页的 URL 作为爬虫的种子集合，具体如表 2。

表 2 大 V 用户表

央视新闻	头条新闻	新浪新闻	新华网
人民日报	共青团中央	新浪娱乐	何炅
新浪体育	央视财经	财经网	谢娜
平安北京	北京那些事儿	汽车之家	易车
局座召忠	新浪军事	新浪科技	梨视频
……	……	……	……
日本零君	英国那些事儿	玩乐美国	……

2）然后将 URL 放入到待爬取的初始队列中。

3）根据待爬取列表中的 URL 访问爬取网页源代码，解析网页，利用正则表达式提取粉丝列表中的用户 ID 和该用户的粉丝数。

每个用户微博首页的粉丝列表和每条微博的评论中含有用户链接，但相比之下微博评论中的用户链接数量较少，且存在大量重复。为了高效获取用户列表，我们只抓取用户粉丝列表中的新用户连接。

4）用户过滤。

微博用户中存在许多“僵尸”用户，这些用户只发广告或充当“水军”，他们所发微博价值很小，有时甚至产生副作用，因此我们对用户粉丝数设定阈值 $F=10$，忽略粉丝小于 F 的用户来规避“僵尸”用户。

5）利用用户 ID 生成微博用户首页 URL，保存到 MySQL 数据库中并利用 MySQL 去重。

在 MySQL 数据库中建表时，对用户 ID 这一列使用 UNIQUE 约束，SQL 语句如下：

```
CREATE TABLE userlist(
……
id INT(20),
url VARCHAR(100),
……
UNIQUE(id)
)
```

这样，当发现一个用户 ID，经过过滤并生成该用户微博首页 XHR 的 URL，将 ID 和 URL 存入 MySQL 数据库时，如果是重复 ID 则无法存入数据库，以此达到去重的目的。同时，将去重后的新用户连接放入待爬行的队列中。

6）判断，如果用户数大于等于 1000 万或者没有发现新用户连接，则进行 7）；否则重复步骤 2）、3）、4）、5），继续获取该新用户的粉丝列表中的用户 URL，扩充用户列表。

7）设计通用网络爬虫，获取微博数数据。

设计一个通用网络爬虫，遍历前几步获得的用户列表，根据每个用户微博 URL 获取网页源代码，利用正则表达式或网页标签抓取用户信息、所发微博、时间、地点、所用设备等数据，并保存到 MySQL 中。

通过这个免登录微博爬虫系统，我们可以避免设置复杂的模拟登录操作，规避掉诸多反爬虫策略，实现更加快速的免登录长时间稳定的数据采集，同时保证了数据的完整性。

2 设计分布式免登录微博网络爬虫系统

第一节中，我们设计了采取广度优先策略的网络爬虫获取用户列表，然后再利用通用网络爬虫遍历用户列表获取微博数据。但是随着用户列表的不断增大，达到千万级、亿级，以及对于数据需求量的不断扩大，单机遍历用户列表爬虫的速度已经不能满足需求，所以我们将 Hadoop 分布式计算平台与免登录网络爬虫相结合，设计出可以满足海量数据采集需求的免登录分布式网络爬虫。

2.1 Hadoop 分布式计算平台

Hadoop[7] 是 Apache 下的开源分布式计算平台。Hadoop 可以将大量的普通计算机搭建成集群，整合这些计算机的运算能力和存储能力，解决了大数据并行计算、存储、管理等关键问题。

HDFS（Hadoop Distributed File System）和 MapReduce 是 Hadoop 分布式系统的核心。HDFS[8] 是分布式计算中数据存储管理的基础。MapReduce[9] 是一种高性能的分布式计算框架，可将一个大的任务分配给数千台普通计算机的集群，并且高可靠性和高容错性并行处理海量数据集。HDFS 在集群上实现分布式文件系统，MapReduce 在集群上实现分布式计算和任务处理。他们相互依赖，共同完成了 Hadoop 分布式计算平台的主要任务。

另外，Hadoop 还有为它量身打造的非关系数据库 HBase[10]。利用 HBase 技术可在大量廉价普通计算机上搭建起大规模结构化存储集群。本次我们设计的分布式网络爬虫系统利用 HBase 数据库存储。

2.2 基于 MapReduce 的网络爬虫系统的设计

在进行爬虫设计前，需要先将用户列表从之前的 MySQL 中转移到 Hbase 中。Hadoop 平台提供了一个组件 Sqoop，它的功能是在 Hadoop 和关系数据库之间传送数据，可以将数据在 MySQL、Oracle 数据库和 HBase 数据库之间进行传递。我们先利用 Sqoop 将 MySQL 数据库中的用户列表迁移到 HBase 数据库中，具体命令如下：

```
sqoop import
——connect jdbc:mysql://192.168.1.12:3306/ll ——username xxxx ——password yyyy
——query "SELECT id FROM userlist"
——hbase-table userlist ——hbase-create-table
——hbase-row-key id
——column-family user
```

在 HBase 接收到之前的用户列表后，我们将第 1 节中设计的通用爬虫 MapReduce 化，这样就可以将免登录抓取微博数据的任务分配给多台普通计算机共同并行完成，大大加快了数据采集速度。

由于该过程中只需要 Map 过程，不需要 Reduce 过程。所以网络爬虫细节如下：

Map 过程：

输入：HBase 中的用户 ID；

输出：用户所发微博文本；

```
1. map(ImmutableBytesWritable key, Result value){
2. id = value.get(); //提取 Hbase 中的用户 ID
3. create xhr.url from id; //根据用户 ID 生成 XHR 的 URL
4. crawl html without login from url;
5. get data from html;
6. save data to HBase;
7. }
```

该分布式爬虫的输入为 HBase 数据库，而不是通常的 HDFS，因此对主函数做特殊说明。具体如下：

```
1. main(){
2. create configuration; //根据集群生成 Mapreduce 任务配置
3. create job(configuration); //根据配置建立 Mapreduce 任务
4. set input HBase column and columnfamily;
5. set input HBase Table,Mapperclass and job;
6. set output HBase column,columnfamily,Table,Mapperclass and job;
7. //不需要像以 HDFS 为输入输出时设置 Map 过程输出的键、值类型;Mapreduce 任务输出的键、值类型和输出的目录等。
8. }
```

这样设计好的分布式爬虫系统就可以在 map 过程时，把对千万级、亿级用户采集的大任务分发到各个节点，各个节点共同完成任务，快速采集海量数据。

3　实验与分析

3.1　实验环境

实验硬件：I7 CPU，16 G 内存，2 TB 硬盘服务器。

实验软件：CentOS 6.5、MySQL5.7、Hadoop－2.7.3、HBase－1.2.4、Zookeeper－3.4.6、Sqoop－1.4.6

根据文献［11］搭建 Hadoop 分布式计算集群，共包含 1 个 Mater 主节点，7 个 Slave 从节点。

3.2　实验结果与分析

在大规模爬取数据时，爬虫的稳定性是最基本的要求。首先对比本文免登录爬虫与模拟登录爬虫的稳定性，利用两种爬虫分别对微博进行 10、15、20、25、30 小时的抓取，比较程序是否可以稳定运行。结果如表 3。

表 3　两种爬虫稳定性比较

时间/h	10	15	20	25	30
免登录网络爬虫	稳定	稳定	稳定	稳定	稳定
模拟登录网络爬虫	稳定	稳定	稳定	稳定	稳定

由此可见，本文设计的免登录爬虫与模拟登录操作的爬虫均可实现对于新浪微博的长时间稳定抓取。

然后我们比较官方 API、免登录爬虫与模拟登录爬虫获取数据的完整性。验证它们是否可以抓取用户 ID，所在身份城市，个人描述，性别，粉丝列表，关注列表，所发微博详情（微博内容、时间、地点、设备等）等。结果如表 4。

接下来，比较两种爬虫在普通单机情况下的数据采集效率，让两种爬虫在 10、15、20、25、30 小时内稳定抓取微博数，对比结果如图 3。

表 4　两种爬虫与官方 API 数据完整性对比

	官方 API	免登录爬虫	模拟登录爬虫
用户 ID	可以	可以	可以
所在城市	可以	可以	可以
个人描述	可以	可以	可以
性别	可以	可以	可以
粉丝列表	可以	可以	可以
关注列表	可以	可以	可以
微博内容	可以	可以	可以
发微博时间	可以	可以	可以
发微博地点	可以	可以	可以
发微博设备	可以	可以	可以

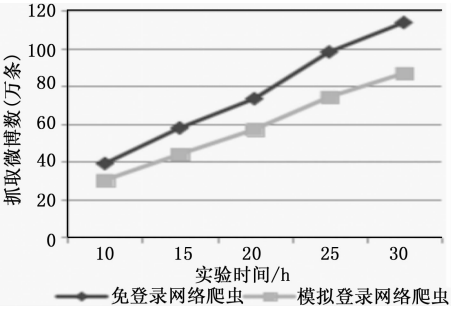

图 3　免登录爬虫与模拟登录爬虫速度对比

由此可见，本文设计的免登录爬虫，由于没有了复杂的模拟登录操作，更加简单而且少了一次与 Web 服务器的交互，因此数据抓取速度更快。

最后，比较单机免登录网络爬虫与分布式网络爬虫的数据采集效率。通过 5 次连续 24 小时的抓取，对比抓取到的微博数，结果如图 4。

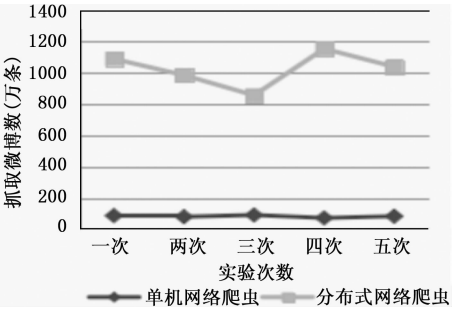

图 4　单机爬虫和分布式网络爬虫速度对比

结果表明：虽然我们设计的单机网络爬虫可以突破 API 访问次数的限制，实现免登录长时间的稳定抓取，但是单机的计算能力还是限制了其抓取速度，结合 Hadoop 大数据平台后，设计出的分布式网络爬虫，利用集群的运算能力，可以大大加速微博抓取效率，实现速度的 10 倍增长，满足了人们对海量数据抓取的需求。

4　结束语

本文设计的针对微博的网络爬虫，以 XHR 的 URL 为入

（下转第 136 页）