Journal of Computer Applications
计算机应用,2014,34(11):3131 -3134
ISSN 1001-9081
CODEN JYIIDU
2014-11-10
http://www.joca.cn

文章编号:1001-9081(2014)11-3131-04
doi:10.11772/j.issn.1001-9081.2014.11.3131

基于 Python 的新浪微博数据爬虫

周中华,张惠然,谢　江*
(上海大学 计算机工程与科学学院,上海 200444)
(* 通信作者电子邮箱 jiangx@shu.edu.cn)

摘　要:目前很多的社交网络研究都是采用国外的平台数据,而国内的新浪微博没有很好的接口方便研究人员采集数据进行分析。为了快速地获取到微博中的数据,开发了一款支持并行的微博数据抓取工具。该工具可以实时抓取微博中指定用户的粉丝信息、微博正文等内容;该工具利用关键字匹配技术,匹配符合规定条件的微博,并抓取相关内容;该工具支持并行抓取,可以同时抓取多个用户的信息。最后将串行微博爬虫工具和其并行版本进行对比,并使用该工具对部分微博数据作了一个关于流感问题的分析。实验结果显示:并行爬虫拥有较好的加速比,可以快速地获取数据,并且这些数据具有实时性和准确性。

关键词:新浪微博;爬虫;Python;并行;大数据
中图分类号: TP391;TP311　　**文献标志码**:A

Data crawler for Sina Weibo based on Python

ZHOU Zhonghua, ZHANG Huiran, XIE Jiang*
(*School of Computer Engineering and Science*, *Shanghai University*, *Shanghai* 200444, *China*)

Abstract: Nowadays, most of researches about social network use data from foreign social network platforms. However the largest social network platform Sina Weibo in China has no data interfaces for investors. A Sina Weibo data crawler combined with parallelization technology was put forward. It got fans information and Weibo data content of different weibo users in real-time. It also supported key words matching and parallelization. The serial data crawler and its parallel version were compared, and an experiment about flu was conducted on some Weibo data. The results indicate that, with parallelization, this tool has liner speedup and all the fetching data are with timeliness and accuracy.

Key words: Sina Weibo; crawler; Python; parallel; big data

0　引言

计算机技术的进步使人们的生活方式逐渐发生改变,社交网络就是一个非常突出的例子。越来越多的人参与到社交网络平台中去,与他人互动,分享各种内容。在大数据时代来临之际,社交网络就像一个巨大的宝库,吸引了大量的研究人员参与到相关内容的研究。在国外,人们针对 Twitter、Facebook 等知名社交平台展开了一系列的分析[1-5],但是针对国内社交网络平台的研究[6-9]还比较欠缺:一方面因为相关研究刚刚起步,缺乏相关的研究方法;另一方面缺乏相关的研究数据,使得一些研究难以开展。在国外 Twitter 等社交平台会提供一些数据接口供研究人员获取研究数据,但是在国内却无法正常访问这些接口。国内由新浪网推出的新浪微博堪称中国的 Twitter。截至 2013 年 3 月,微博用户数已达 5.56 亿,日活跃用户数超 5000 万。如此之大的一个社交平台为社交网络分析、网络数据挖掘等研究提供了强有力的大数据支持。

然而新浪微博官方并没有提供相关的数据接口,没有数据,一些研究分析工作也无法进行。目前网络中也存在一些公开的微博数据集供人们下载,但是这些数据集通常规模比较小,而且还缺乏实时性。有些技术力量强劲、资源充足的研究团队通常自己开发一些爬虫来获取研究数据。这对不熟悉爬虫技术的研究人员而言是个极大的挑战。本文提出了一款基于 Python 语言的新浪微博数据爬虫,为数据获取提供支持。

本文爬虫通过模拟客户端的操作如登录、访问好友、查看粉丝、查看微博内容等方式获取相关数据,并且将这些数据持久化保存到本地硬盘上,方便后续进一步的数据挖掘与分析。同时,本文爬虫还集成了文本匹配功能,利用该功能可以轻松地实现指定内容的数据获取。例如,可以利用本文爬虫检索包含流感、感冒、发烧、发热这四个词语中一个或多个关键字的微博,并把这些微博保存到硬盘上。使用本文爬虫能够节省分析人员的开发时间,使得他们可以将更多的精力放在数据分析上面,同时也可以对一些无用的数据起到过滤作用。

1　微博爬虫

网络爬虫是用来获取网络数据的重要工具。关于网络爬虫技术的研究[10-13]不计其数,然而普通的爬虫很难直接用来

收稿日期:2014-06-05;修回日期:2017-08-27。　基金项目:国家自然科学基金资助项目(91330116);高等学校博士学科点专项科研基金资助项目(20113108120022);上海市科委重点项目(11510500300)。

作者简介:周中华(1989-),男 江苏常州人,硕士研究生,CCF 会员,主要研究方向:生物信息、高性能计算;　张惠然(1981-),男 河南新乡人,讲师,博士,CCF 会员,主要研究方向:生物信息、高性能计算;　谢江(1971-),女,湖北恩施人,副教授,博士,CCF 会员,主要研究方向:生物信息、高性能计算。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

抓取新浪微博的相关数据。因为新浪微博有着复杂的登录机制，同时相关数据拥有统一的格式，针对这种情况，利用爬虫原理，可以开发出一款专门针对新浪微博数据的网络爬虫。

1.1 微博登录

新浪微博的数据都需要在登录的情况下才能访问到，所以微博登录是爬虫需要解决的第一个问题。通过分析网页版微博的登录相关代码可以发现微博登录主要分为三个步骤：1）客户端向微博的用户服务器发送登录请求；2）服务器接收到登录请求后会生成相应的密钥返回给客户端；3）客户端将用户的用户名、密码以及2）中服务器发回的登录密钥结合在一起再向服务器提交登录信息，服务器验证成功之后将会返回正确的登录状态以及当前用户的个人信息。成功登录之后客户端只需要保持与服务器的session会话就可以方便地访问微博中的一些数据资源。

以下是微博登录的伪代码：

```
Begin
  Step1:
    Send Login Request
  Step2:
    Get Response From Server
    Get Encrypt Information From Response
  Step3:
    Encrypt User Information
    Send Encrypt Information
  Step4:
    Get Login Status
End
```

1.2 微博关系抓取

微博爬虫成功登录之后就可以通过HTTP地址访问其他用户的相关信息，如其他用户的关注列表、粉丝列表以及微博列表。由于爬虫采用的是HTTP请求的方式获取数据，因此每次请求获取到的是一大堆复杂的HTML代码，但是用户的微博数据都具有相同的格式，可以通过正则表达式将这些数据从混乱的HTML代码中提取出来。

在新浪微博中，每个用户有自己的唯一ID与之对应，因此，可以使用ID作为用户的唯一性的判断依据，在访问相关用户的数据过程中，也只需要使用其用户ID就能访问到相关数据。本文爬虫提供了一种基于广度优先的搜索策略来抓取用户关系数据。首先选取一名种子用户，以该用户为起始点，先将他所有的好友信息收集起来。图1显示的是从1到6逐个提取每个用户的个人信息，搜索完成后再去寻找该用户的第一个好友的其他好友信息；以此类推，直到搜索完成。

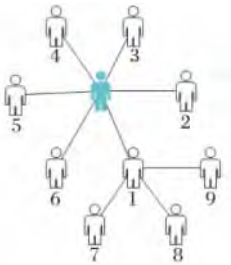

图1 广度优先的粉丝关系搜索

整个搜索过程中微博爬虫需要维护两个搜索队列：一个是已完成的队列，另一个是未完成的队列。在初始状态下已完成的队列为空，未完成的队列中只有一个种子用户，对种子用户的所有信息完成一遍搜索之后，该用户就会进入已完成队列，对该用户搜索过程中，当遇到一个新用户时，会根据其用户ID在已完成队列和未完成队列中查询该用户是否已经被记录过，如果该用户未被记录，该用户将进入到未完成队列末尾。搜索的结束条件并不是搜索完微博的所有用户才会结束，而是当已完成队列中的用户数量达到预先设置的一个最大值的时候就会结束搜索。

以下是整个搜索过程的伪代码：

```
Begin
  Initialize Waiting Query
  Initialize Finished Query
  Push Seed User Into Waiting Query
  While length(Finished Query) < MaxNum
  Begin
    Pop User From Waiting Query
    Scan User Information
    If New User Not In Waiting Query
      and New User Not In Finished Query
    Begin
      Push New User Into Waiting Query
    End
  End
End
```

1.3 微博内容抓取

微博正文是非常重要的微博数据，很多研究就是基于用户的微博内容展开的，因此，本文爬虫也针对用户的微博正文提供了相应的抓取方法。一种方法是：本文爬虫可以将用户的所有微博内容全部以文件的形式完全记录到磁盘，但是这样做就需要很多的物理存储空间才能将如此之多的用户数据保存下来。另一种方法是：本文爬虫提供了简单的字符串匹配功能，在抓取用户微博内容的过程中会根据输入的需要匹配的关键字进行匹配，如果发现匹配成功的微博爬虫会将该微博保存到磁盘。这样研究人员就可以有针对性地进行相关研究和分析。

1.4 关键字匹配

由于人们发布的微博内容各异，针对某种研究而言，一个人的全部微博中有很多无意义的内容。例如针对微博中人们得感冒与时间关系的研究就需要从所有的微博中找到与感冒相关的微博，并保存这些内容。而简单的爬虫并不能分辨哪些内容有用，哪些内容是没有用的，如果全部都保存到硬盘上会耗费大量的存储，同时还会花费更多的时间从这些数据中再次筛选与感冒相关的微博，这样既浪费了资源又浪费了时间。因此爬虫中内嵌了关键字匹配模块方便根据关键字筛选符合条件的微博，这样提高了效率，也节省了资源。

以下就是一条微博匹配多个关键字的相关代码：

```
Begin
  For i = 0; i < len(keys); i++
  Begin
    If match(content, keys [i])
    Begin
```

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
        return True
      End
    End
    return False
  End
```

2　并行爬虫

由于微博用户数量非常大,单进程的爬虫很难满足快速抓取大量数据的要求,因此本文爬虫进行了并行架构的扩展,实现了基于 MPI 的并行数据抓取功能。

本文并行爬虫主要采用主从模式,主节点负责维护整个爬虫的抓取队列以及任务分配工作,从节点负责对自己的任务列表按照第 1 章中的抓取规则进行数据抓取。每个从节点都需要维护两个队列,一个是任务队列,另一个是新用户队列。当从节点完成了其任务队列后会将自己的新用户队列交给主节点,由主节点来处理合并用户的工作,同时,主节点会将新的任务队列发送到从节点,由从节点继续抓取新数据。相关伪代码如下:

```
Begin
  If Is Master
  Begin
    Initialize Waiting Queue
    Initialize Finished Queue
    Load Some User From Disk Into Waiting Queue
    Pop 50N Users From Waiting Queue
    For i=0; i<N; i++
    Begin
      Send 50 Users to Slaver i
      Push 50 Users Into Finished Queue
    End
    While length(Finished Queue) < MaxNum
    Begin
      If Receive New Users From Slaver j
      Begin
        Foreach User in New Users
        Begin
          If User Not In Waiting Queue
            and User Not In Finished Queue
          Begin
            Push User Into Waiting Queue
          End
        End
        Pop 50 Users From Waiting Queue
        Send 50 Users to Slaver j
        Push 50 Users Into Finished Queue
      End
    End
  End
  Else
  Begin
    Initialize Job Queue
    Initialize New User Queue
    While True
    Begin
      Receive Users From Master
      Push Users Into Job Queue
      Clear New User Queue
      While length(Job Queue) >0
      Begin
        Pop User From Job Queue
        Scan User Information
        Push New Users into New User Queue
      End
      Send New User Queue To Master
    End
  End
End
```

3　实验与分析

对本文爬虫的并行效率进行分析,同时将介绍两组利用本文爬虫抓取数据并进行数据分析的实验:一个是微博社交图谱分析,另一个是与感冒相关的微博与时间关系的分析。

3.1　并行效率分析

在本实验中使用串行微博爬虫工具和其并行版本对相同的 50 名微博用户进行微博好友抓取,这 50 人的微博好友数量在 100 ~ 200 人,由此来观察该工具的运行速度和加速比(见图 2)。

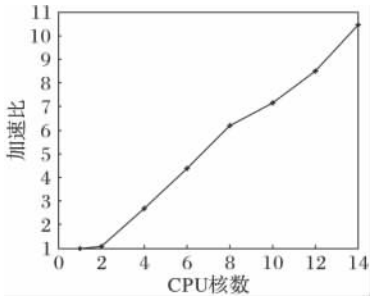

图 2　加速比曲线

表 1　并行抓取时间及加速比

CPU 核数	时间/s	加速比	CPU 核数	时间/s	加速比
1	136	1.0	8	22	6.2
2	127	1.1	10	19	7.2
4	51	2.7	12	16	8.5
6	31	4.4	14	13	10.5

从表 1 与图 2 可看出,并行的爬虫具有良好的线性的加速比。由于采用主从模式,因此当 CPU 核数为 2 时加速效果不是很好;但是,随着 CPU 核数的增加,加速比呈线性增长,这对大量数据的爬取工作具有非常好的加速效果。

3.2　用户社交图谱

在本实验中,以某个用户为中心,抓取了他所有的微博粉丝,以及粉丝之间的关系数据,并且将这些数据利用网络可视化工具 Gephi [13] 将数据显示在图 3 中。在图 3 中,每一个圆点代表一名微博用户,点的大小表示该用户的粉丝数量,粉丝数量越多,点越大,点与点之间的连线表示用户之间的粉丝关系。图中最中心的点所代表的用户就是种子用户,其他所有的点都与之有关联,但是在这些有关联的点之间还存在三个比较密集的簇,如图 4 所示。

这三个簇的出现,表明在新浪微博中该中心用户的好友中存在三个比较紧密的群体,而这三个群体分别是其大学同

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

学群体、初中同学群体，以及小学同学群体。

图 3 微博用户关系

图 4 三个用户簇

3.3 感冒数据分析

在该实验中，针对 156 235 名微博用户，从 2009 年 10 月到 2013 年 6 月这段时间内，共计 376 565 361 条微博进行了关键字匹配分析，主要匹配了咳嗽、感冒、发热、流感这四个关键字，并且记录下了 578 711 条符合匹配条件的微博所发送的时间和发送的人的数据。

图 5 显示的是不同年份每月与感冒相关的微博数量随时间的变化曲线。

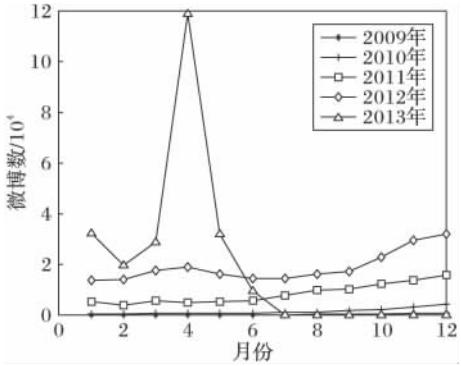

图 5 与感冒相关的微博数量与时间关系

图 6 显示的是不同年份每个月关注感冒的人数随时间变化的曲线。

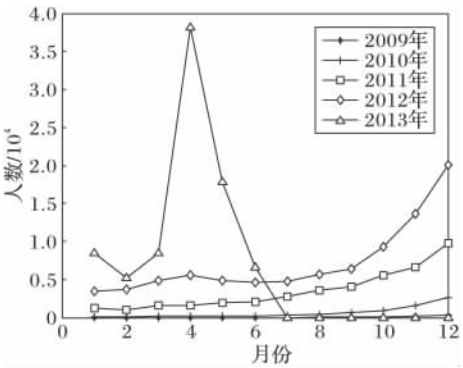

图 6 关注感冒的人数与时间关系

从图 5～6 可看出：在 2013 年度人们对于感冒的关注度明显高于其他年份，并且在 2013 年 3 月至 6 月期间存在一个巨大的波峰，这说明这一期间流感问题成为一个人们关注的热点问题。而且在这一时期刚好是国内禽流感爆发的时间，正因为禽流感的爆发引起了微博上人群的广泛关注，然而随着时间的推移人们对其关注度开始逐渐下降。

4 结语

本文爬虫从技术上为一些社交网络研究者们提供了方便快捷的新浪微博数据获取工具。其主要有以下几个特点：

1）使用方便。使用者只需提供微博账号就能利用本文爬虫抓取新浪微博中的相关数据。

2）支持关键字匹配。使用者只需自定义感兴趣的关键字本文爬虫就能自动匹配相关内容并将符合条件的微博保存到磁盘之上。

3）支持并行。对于需要大量微博数据，同时又需要快速获取数据的用户而言使用其并行功能，可以达到令人满意的效果。

参考文献：

[1] TUMASJAN A, SPRENGER T O, SANDNER P G, *et al*. Predicting elections with Twitter: what 140 characters reveal about political sentiment[C]// Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Madison: AAAI Press, 2010, 10: 178－185.

[2] WELCH M J, SCHONFELD U, HE D, *et al*. Topical semantics of twitter links[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 327－336.

[3] CARLISLE J E, PATTON R C. Is social media changing how we understand political engagement? An analysis of Facebook and the 2008 presidential election[J]. Political Research Quarterly, 2013, 66(4): 883－895.

[4] CUNLIFFE D, MORRIS D, PRYS C. Young bilinguals´ language behaviour in social networking sites: the use of welsh on Facebook [J]. Journal of Computer-Mediated Communication, 2013, 18(3): 339－361.

[5] STRAFLING N , KRAMER N C . Learning together on Facebook *et al*. The influence of social aspects and personality on the usage of social media for study related exchange [J]. Gruppendynamik und Organisationsberatung, 2013, 44(4): 409－428.

[6] DUAN J Y, DHOLAKIA N. The reshaping of Chinese consumer values in the social media era: exploring the impact of Weibo [J]. Journal of Macromarketing, 2013, 33(4): 402－403.

[7] HUANG R, SUN X. Weibo network, information diffusion and implications for collective action in China [J]. Information Communication and Society, 2014, 17(1): 86－104.

[8] MAZO J. Blocked on Weibo: what gets suppressed on China´s version of Twitter (and why) [J]. Survival, 2013, 55(6): 191－192.

[9] POELL T, de KLOET J, ZENG G, *et al*. Will the real Weibo please stand up? Chinese online contention and actor-network theory [J]. Chinese Journal of Communication, 2014,7(1): 1－18.

[10] PINKERTON B. Finding what people want: experiences with the WebCrawler[EB/OL]. [2010-10-10]. http://www.webir.org/resources/phd/pinkerton_2000.pdf.

[11] AHMADI-ABKENARI F, SELAMAT A. An architecture for a focused trend parallel Web crawler with the application of clickstream analysis[J]. Information Sciences, 2012, 184(1): 266－281.

[12] ZHOU L, LIN L. Survey on the research of focused crawling techn ique [J]. Computer Applications, 2005, 25(9): 1965－1969 (**周立柱, 林玲. 聚焦爬虫技术研究综述**[J]. **计算机应用**, 2005, 25(9): 1965－1969.)

[13] BASTIAN M, HEYMANN S, JACOMY M. Gephi: an open source software for exploring and manipulating networks[EB/OL]. [2010-10-10]. https://gephi.org/publications/gephi-bastian-feb09.pdf.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net