社交网络数据采集技术研究与应用

徐雁飞 刘 渊 吴文鹏
(江南大学数字媒体学院 无锡 214122)

摘 要 随着社交网络的快速发展,对其研究也逐步深入。显然,社交网络基础数据的获取对研究具有非常重要的意义。针对目前已有的数据采集方案,根据新浪授权标准以及最新的微博加密方式,研究了两种采集方案:1)经 OAuth2.0 认证后,通过微博 API 接口获取数据;2)在 RSA2 加密方式下模拟登录微博,再通过网络爬虫抓取数据。同时,还研究了通过网页采集器针对微博编写适当的采集规则进而实现对数据的获取。3 种数据采集方案都能有效地对数据进行采集且各具特点,针对数据的采集需求,提出融合不同的采集方案的策略。经实验研究,方案的融合策略可快速、高效地实现大数据量的采集。

关键词 Python,微博 API,模拟登录,网络爬虫,采集器,融合策略
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.051

Research and Application of Social Network Data Acquisition Technology

XU Yan-fei LIU Yuan WU Wen-peng
(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract With the rapid development of social networks, the study on it is also gradually deepening. Obviously, the acquisition of basic data of social networks has very important significance to the study. In this paper, aiming at the existing data acquisition programs, according to the Sina authorization standards and the latest microblog encryption, the paper studied two kinds of acquisition programs. One obtains data through the API interface after the OAuth2.0 certification, and another crawls data through the Web crawler after being simulated by the RSA2 encryption. At the same time, it also studied the acquisition of the data by using the appropriate acquisition rules for the microblog. Three kinds of data acquisition programs are able to collect the data effectively and they have their own characteristics. According to the requirements of data acquisition, the fusion of different acquisition programs were proposed in this paper. Through the experimental study, the fusion strategy can quickly and efficiently obtain vast amount of data.

Keywords Python, Microblog API, Simulated login, Web crawler, Collector, Fusion strategy

作为 Web2.0 时代的典型应用,社交网络服务正在世界各地以极快的速度流行起来[1]。社交网络的出现使互联网上的社交形态和用户行为向现实社会推进,虚拟社会与现实社会开始不断交叉。随着社交网络的不断发展以及注册用户的迅猛增加,越来越多的研究人员参与其中来进行多方面内容的研究。根据《第 35 次中国互联网络发展状况统计报告》[2]中的数据显示,截止 2014 年 12 月,我国网民规模达 6.49 亿,手机网民规模达 5.57 亿。而根据有关报告预估新浪微博的注册用户已亿突破 6 亿,2015 年新浪微博第一季度财报[3]数据显示第一季度月活跃用户达到 1.98 亿,日均活跃用户也已达到 8900 万。

国外研究者多针对 Facebook[4]、Twitter[5,6]等进行数据采集,研究者在 Twitter 平台上建立网络模型,对网络特性进行研究,Twitter 底层的很多接口和库是公开的,研究者多利用这一特性对 Twitter 数据进行统计;Facebook 也逐渐开放了 API 接口,研究者利用陆续开放的接口对 Facebook 用户数据进行采集与研究应用。新浪微博成为国内研究者进行数据挖掘研究以及分析[7]的一个强大的数据提供平台。谭廉等人[8]提出了结合基于微博 API 与基于网络爬虫的页面解析的采集方法,但该方法在获取数据的完整性上不及网络爬虫,且 API 获取大数据量用户 ID 的速率低于页面采集器;黄延炜等人[9]研究了微博 API 与基于网络数据流的信息获取技术,在该技术中简单采用 API 接口进行数据采集会出现访问受限,且采集的用户微博数据量少。

本文以新浪微博作为实验研究对象,所涉及实验均在 Python 语言环境下实现。研究了通过微博 API 接口以及在模拟登录下通过网络爬虫进行数据获取的方案,提出了在采集器下通过进行规则设置来采集微博数据并进行数据处理的方案。最后通过分析各采集方案的利弊,进一步提出采集方案融合的策略,得出 API 接口与网络爬虫融合、采集器与网络爬虫融合都能明显地提高采集效率,且后者效率更高。

1 微博 API 数据获取

微博开放平台[10]是一个基于微博客系统的开放的信息

收稿日期:2015-12-26 返修日期:2016-03-05 本文受国家自然科学基金项目(61103223)资助。
徐雁飞(1991-),女,硕士生,主要研究方向为社交网络分析,E-mail:1274998390@qq.com;刘 渊(1967-),男,教授,CCF 会员,主要研究方向为网络信息系统及网络安全;吴文鹏(1991-),男,硕士生,主要研究方向为社交网络用户行为建模。

订阅、分享与交流平台。研究人员可以登录平台并创建应用，使用微博平台提供的 API 接口[11]获取大量的微博信息、粉丝关系、用户信息等。

1.1 OAuth2.0 认证

若第三方服务要获取用户在微博平台上的信息则需要经过用户的授权，目前微博已全面采用 OAuth2.0 作为应用认证方式。OAuth 是一种开放的授权标准。它允许用户将自己存放在一个站点上的资源分享给另一个资源，而在这个过程中用户不需要将他在资源站点上的证书提供给另一站点，如用户名、密码。具体认证过程如下。

(1)用户编写开发者基本信息并提交身份认证，通过身份认证后，应用可以申请应用广场以及高级接口权限。

(2)用户向微博开放平台 OAuth 服务商提出创建应用的申请，在其中填写应用名称、回调地址和其他的一些必要信息便可完成应用注册。注册完成后将会获得该应用的 App Key 和 Secret Key，App Key 是应用的唯一标识，开放平台通过 App Key 来鉴别应用的身份，App Secret 是分配给应用的密钥，这个密钥用来保证应用来源的可靠性。可使用微博提供的默认回调页：https://api.weibo.com/oauth2/default.html 来填写回调地址。

(3)利用认证程序，授权用户到如下页面：https://api.weibo.com/oauth2/authorize?redirect_uri=YOUR_REGISTERED_REDIRECT_URI&response_type=code&client_id=YOUR_CLIENT_ID。其中所提到的 YOUR_CLIENT_ID 为应用的 AppKey，YOUR_REGISTERED_REDIRECT_URI 为之前填写的授权回调页，两者一定要完全相同。

(4)用户授权成功，跳转到回调页。在浏览器返回的 url 地址中，提供由 32 位十六进制数组成的认证码，例：f97605c0518fdd5b314570a348c8fbec，将所提供的认证码提交给认证服务器，服务器便颁发微博授权的 API 调用令牌 Access_Token 与对应的密钥。微博为 Access_Token 设置的默认有效期为 24 小时。

(5)将 Access_Token 令牌作为参量，调用相应的 API 接口获取或者发布数据，Access_Token 过期后需要重新授权。在步骤(4)中，使用代码正确的模拟浏览器发包，即可自动获取 code 参数值，无须再手动输入。

1.2 数据获取

微博 API 提供 REST 风格的基础数据接口，包括：获取下行数据集接口、微博接口、用户接口、标签接口、话题接口、OAuth 接口等，这些接口为第三方开发者提供了诸如获取用户信息、获取好友关系、发送微博等功能。

若要获取某个用户的信息，可以利用用户接口、标签接口、微博接口等。服务器同意用户调用接口后，微博返回的结果是 JSON，但是 SDK 将其封装成 JSON 对象，直接通过返回结果调用相应的属性即可。

文中通过调用 users/show 接口返回 JSON 格式的数据：

```
{'statuses':[{'created_at':'Thu Feb 19 00:01:16 +0800 2015',…
'users':{
'id':1812321643,…
'province':'37',
'city':'1000',
'location':'山东',…
'gender':'f',
'followers_count':115,
'friends_count':139,
'statuses_count':279,
'created_at':'Mon Sep 06 01:52:09 +0800 2010',
…}}]}
```

返回数据包括：用户 id、用户昵称、友好显示名称、用户所在省级 id、用户所在城市 id、用户所在地、用户个人描述、用户博客地址、用户头像地址、用户个性化域名、用户性别、粉丝数、关注数、微博数、收藏数、用户创建(注册)时间、是否是微博认证用户、认证原因、用户的在线状态(0：不在线，1：在线)、用户互粉数等。若 r 为返回结果，那么通过 $r.statuses$ 可获取 JSON 中的 *statuses* 列表，$r.statuses[i]['text']$ 可获取第 i 个用户的更新的微博内容。则 $r.statuses[i]['user']['id']$ 便可获取第 i 个用户的微博 id，以此为例获取所需的用户微博数据，并将其存入数据库。通过用户关系接口中的 friendships/followers 接口获取用户的粉丝列表，请求方式为 get，获取的粉丝数据示例如表 1 所列。

表 1 粉丝用户数据示例表

昵称	粉丝	关注	微博	用户 id	性别
周明伟	192	129	426	1728860912	m
傅俊鹏	187	215	490	1298113144	m
丿访旋	525	1830	2751	3780828351	f
喵小懒	253	373	2439	1879908483	f

利用微博 API 进行数据获取时存在返回结果条数的限制：1)用户不登陆时基于 IP 的限制(1000 次/h)；2)用户登录时基于用户的限制(1000 次/h)。针对接口限制，文中实验调用 API 接口后添加程序控制部分，程序线程控制 API 被调用的次数，检测到 API 被调用 100 请求后，程序休眠 3min，以防止超过 API 的调用上限。此外，还对数据存储进行处理，添加存储控制部分，有效地防止了数据的重复存储，提高了数据存储效率。完整的程序结构流程图如图 1 所示。

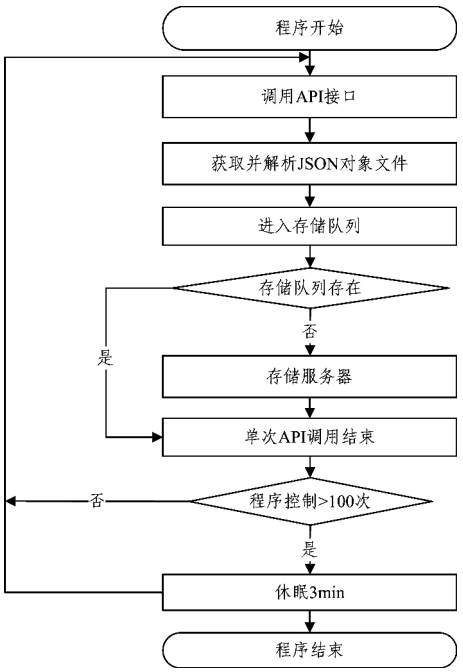

图 1 API 数据获取流程图

由于微博提供多种特色 API 接口，如获取最新的公共微博、获取学校列表、获取某条微博的评论列表等，因此在需要某些特定条件下的数据且数据量不大时应采用微博 API 获

取方法进行实验采集与研究分析，方便快捷。但在需要数据量较大和需长时间进行微博获取的情况下，由于频繁调用接口使得采集量出现较大的波动，稳定性不高，且数据的完整性也降低。

2　基于网络爬虫的微博数据抓取

由于调用微博 API 获取数据可受到未通过审核应用的测试账号的限制[12]以及高级接口不开放等因素的影响，针对研究需求，除了通过微博 API，还可在成功模拟登录微博的前提下通过网络爬虫实现数据的采集。

2.1　模拟登录

模拟登陆微博的方式在手机微博与浏览器微博中不同。选择任意一种方式采集数据都需通过模拟登陆成功访问微博。通过分析发送的报头信息得知，手机微博上用户名密码登陆时是明文传输，报头信息中的主要参数为 password 的 name 值及 vk 的 value 值。登陆成功后，返回的 cookie 中包含一个 gsid 字段，以 get 方法发送请求获取此参数，便可访问微博，页面返回的编码格式为 UTF-8。浏览器微博登陆比手机微博登陆更复杂，具体步骤如下：

(1)添加用户名(username)，用户名经过 base64 计算：username＝base64.encodestring(urllib.quote(username))[:－1]；base64 编码[13]按照字符串长度，以每 3 个 8bit 的字符为一组，针对每组，首先获取每个字符的 ASCII 编码，然后将 ASCII 编码转换成 8bit 的二进制，得到一组 3＊8＝24bit 的字节，再将这 24bit 划分为 4 个 6bit 字节，并在每个 6bit 的字节前面都填充 2 个高位 0，得到 4 个 8bit 字节，将这 4 个 8bit 字节转换成 10 进制，对照 base64 编码表，得到对应编码后的字符。例如，用户名 xyfjiang@163.com 经过 base64 计算后得到：eHlmamlhbmdAMTYzLmNvbQ。

(2)请求 prelogin 链接地址：http://login.sina.com.cn/sso/prelogin.php?entry＝sso&callback＝sinaSSOController.preloginCallBack&su＝%s&rsakt＝mod&client＝ssologin.js(v1.4.4)。

使用 get 方法得到以下类似内容：

sinaSSOController.preloginCallBack({"retcode":0,"servertime":1436208292,"pcid":"ja-b615644f62068425f335ccb07c8284c231f5","nonce":"UAWG0D","pubkey":"EB2A385…D6442443","rsakv":"1330428213","uid":"1812321643","exectime":3})

从中提取所需要的 servertime，nonce，pubkey 和 rsakv。由于 pubkey 和 rsakv 是固定值，便可直接写入代码中。

(3)对 password 加密，先前是采用 SHA1 加密算法[14]进行加密。目前新浪微博采用的是 RSA2 加密方式。首先创建一个 RSA 公钥，对于公钥的两个参数，新浪微博都给了固定值，但均为 16 进制的字符串，分别为 pubkey 和加密文件中的'10001'，将两个值转换成 10 进制，最后将加密信息转换为 16 进制。

具体实现：

```
rsaPublickey＝int(pubkey,16);
key＝rsa.PublicKey(rsaPublickey,65537);
message＝str(servertime)＋'\t'＋str(nonce)＋'\n'＋str(password);
password＝rsa.encrypt(message,key)。
```

(4)请求通行证：http://login.sina.com.cn/sso/login.php?client＝ssologin.js(v1.4.4)

需要发送的报头信息如下：

postdata＝{'entry':'weibo','gateway':'1','from':'','savestate':'7','userticket':'1','ssosimplelogin':'1','vsnf':'1','vsnval':'','su':username,'service':'miniblog','servertime':servertime,'nonce':nonce,'pwencode':'rsa2','sp':password,'encoding':'UTF-8','prelt':'115','rsakv':rsakv,'url':'http://weibo.com/ajaxlogin.php?framelogin＝1&callback＝parent.sinaSSOController.feedBackUrlCallBack','returntype':'META'}

使用 POST 方式发送请求，检验是否登录成功，参考 POST 后得到的内容，若 retcode＝101，则表示登录失败；若 retcode＝0，则表示登录成功。

(5)登录成功后，在服务器返回的内容中提取出要使用的 url 地址，然后对该 url 地址使用 get 方法向服务器发送请求，并且保存这次请求的 cookie 信息，这便是需要的登录 cookie。

2.2　数据抓取

网络爬虫[15,16]按照一定的逻辑和算法从互联网上抓取和下载互联网的网页，是搜索引擎中一个重要的组成部分。网络爬虫[17]原理：网络看作有向图，网站上的网页作为有向图上分布的节点，网页间的相互关系看作图的有向边。网络爬虫的工作流程即是根据这些有向边(url 链接)，从预先设定的一个或多个节点开始爬取页面，获取网页上的内容并从内容中获得网页中其他的 URL 链接，然后根据 url 链接继续遍历网络中的其他节点。

微博用户有与其相对应的唯一的 id，则使用 id 作为微博用户唯一性的判断依据是可行的，访问用户时，使用其 id 即可访问成功。这样就可以通过使用 id 形成的 url 地址进行页面内容的获取。例如：通过 http://weibo.com/＋id 便可访问该 id 所对应的微博用户页面。

本文实验采用一种广度优先[18]的爬行策略，它将首先处理初始页面上的所有链接，当把该层页面处理完毕后才处理该页面对应的下一层页面，直到完成遍历才终止。这样可以根据所需数据量有效地控制页面的爬行深度，可避免在爬行深层次页面时出现不能终止爬行的情况。微博数据抓取工作过程：

(1)选取一名种子用户(即提供初始微博页面的 url)作为爬虫起始点。

(2)处理待访问的 url 集合中的 url，然后采用多线程或并行技术下载 url 指向的网页，借助于 HTTP 等 Web 协议采集微博网页数据，将采集到的微博页面进行存储。

(3)对采集的微博页面进行分析，提取用户微博 id，对 id 形成的 url 地址进行规范化处理，统一格式并存储，以做下一步处理。

(4)对获取的 url 地址进行分析，剔除重复及无效的微博 url，然后对获取的微博 url 进行整理存储。对于微博 url 的维护和管理[19]，在程序执行的过程中，程序将从队列中选出下一个待处理的 url，当 url 为空时爬虫程序会终止。在实际的应用中，通常会适当控制这些队列的大小。如果队列中的 url 数量过多，会增加服务器的负担，反之，会减小爬虫执行过程中服务器的压力，使爬虫的速度以及效率提高。待下载的 url 队列不能规定得太小，若太小会使得队列迅速为空，若没有新的 url 程序会停止，这样会影响程序执行的效率。

(5)存储下载到本地的页面数据，供后期处理。

如果某条微博的出现的时间段恰巧在用户好友活跃人数最多的时间段，那么该条微博有较高的可能性得到关注、转发或者评论；反之，如果出现的时间段是用户微博活跃人数最少的时候，那么该条微博很有可能被忽略。那么活跃时间段对于微博消息的传播将有很大的影响。实验中通过网络爬虫爬取 500 个用户所发布的微博总数，并统计微博发布时间。横轴为单位时间(h)，纵轴为单位时间内发布的微博数占总微博数的比例，结果如图 2 所示。可得出：在 12 点－14 点、16 点－18 点、20 点－22 点 3 个时间段内用户发布微博数量所占比例最大，也可认为此时间段内用户在线人数最多。那么在此时间段内推送的消息将有较大的可能性被关注且迅速传播。

模拟登录后通过网络爬虫进行数据获取，克服了反复调用 API 接口出现的数据量波动问题，在数据的完整性方面也大幅度提高，效率平稳。但是在只需要部分特定数据时，在数据的后期处理方面相比于调用 API 接口更繁琐。针对这一问题，对网络爬虫的改进及数据的筛选流程需要进一步进行研究。

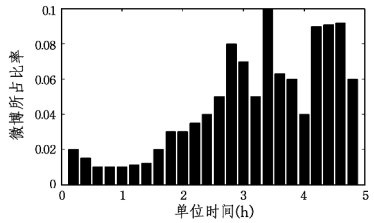

图 2 单位时间内总微博统计比例

3 基于采集器的微博数据获取

目前，现有的采集器多应用于网页信息数据的采集，对于微博，设计正确的采集规则，其也可应用于微博用户数据的获取。实验得出：对微博页面不规则版块的数据采集存在一定误差，但可高效采集页面规则板块的数据。本实验采用八爪鱼网页采集器[17]对新浪微博数据进行获取和研究。

3.1 采集规则的设定

通过网页采集器采集微博数据时，需根据数据需求设定对应的采集规则。设定采集规则的过程中最重要的一步是设计工作流程，这决定能否快速采集到所需微博用户的相关数据。

以采集登录者所关注用户及其用户粉丝微博数据为例，依据数据需求设计工作流程图，如图 3 所示。

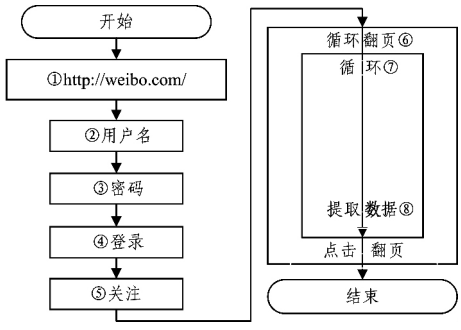

图 3 用户关注数等数据采集流程图

步骤①－步骤④实现微博登陆，步骤⑤－步骤⑧通过两次循环实现对所关注用户及其用户粉丝微博数据的成功提取。根据数据需求可修改步骤⑥－步骤⑧的工作流程图，增加循环层数或者改变提取内容(如提取 url、id 等)，从而实现对用户数据的全面采集。

3.2 数据基本分析

以登录用户为采集起点，根据不同采集规则的设定，通过用户好友或者关注关系，循环采集 3 次得到 15000 多名微博用户的数据，其中包括用户注册地址、关注数、粉丝数、微博数、用户标题、性别、注册时间、url 及 id 信息等。将采集得到的用户数据进行导出整理，存入数据库以备数据分析。

通过采集器采集到的微博用户数据部分示例如表 2 所列。计算分析该部分实验数据中用户注册微博的时间，大约 70％的用户是在 2010－2012 年注册，表明微博在这几年发展迅速；而在 2009 年注册的用户中大约 99％为＋V 用户，说明在微博初期，注册使用的几乎全为认证用户，通过分析这些认证用户信息可知，注册用户中多为目前微博最活跃且影响力相对较高的名人，其中娱乐界占据较大比例，可推断出微博是可供宣传并有高影响力[20]的优势平台。

表 2 用户综合数据获取示例表

昵称	地址	关注	粉丝	微博	性别	时间	id
林依晨 Ariel	台湾 台北市	121	11690963	640	女	2010-08-02	1785075474
Dandy Weng	广东 广州	23	18953	1334	男	2011-01-28	1938682371
周茜茜	江苏 无锡	225	156	184	女	2010-11-24	1873529751
炉鱼风尚烤鱼	浙江 杭州	80	2696	62			3558243235
淘宝褚霸	浙江 杭州	2000	56214	5519	男	2011-01-07	1915508822

采集器通过设定相应的采集规则对微博数据进行获取，没有 API 接口的调用限制，单位时间的数据采集量相对平稳，速率高，但网速的变动对其影响较大。由于是根据规则获取数据，数据的完整性会受到影响。实验过程中，准确且简捷地设定采集规则是获取大量数据的关键。

4 实验分析

本文提出了通过微博 API 接口、网络爬虫以及设定网页采集器规则来实现微博数据获取的方案，针对数据需求的不同，可选择相应的方案。通过实验可知，若需要实时采集当前最新公共微博，可选择通过微博开放平台提供的 API 接口快速获取；若需要获取批量用户的微博、粉丝、性别等页面简单的字段数据，可通过网页采集器快速、高质量地获取；若需分析大量微博数据内容，可基于模拟登录的方式，通过网络爬虫全面、完整地获取。实验过程中采集到的数据分为用户信息和微博信息，其中用户信息结构和微博信息结构如表 3 所列。

表 3 数据信息结构表

用户信息结构	微博信息结构
用户 id	微博 id
用户昵称	发布者 id
用户地址	发布者昵称
用户性别	微博内容
用户注册时间	转发数
是否为认证用户	评论数
粉丝数、微博数、关注数	微博发表时间

各微博数据采集方案均有特色，获取大数据量的微博数据时，采集方案的融合策略更有利于数据采集的实施。API 和网络爬虫方案的融合明显提升了数据采集效率；但对于采

集器和网络爬虫方案的融合，经多次实验，其采集效率均高于API和网络爬虫方案的融合。因此，采集微博数据时，可根据数据需求设定采集规则来获取用户 id，根据 id 生成唯一的 url 地址，再通过网络爬虫进行微博数据获取，提高了采集速率及采集内容的完整性。采集器和网络爬虫的融合策略的程序流程图如图 4 所示（其中根据需要采集的数据量的大小设定循环次数，在此流程图中设计两层为例）。表 4 是 1 h 内各采集方案平均采集用户信息的条数，可见采集方案的融合策略在单位时间内采集微博数据的效率有较为显著的提高，较为实用。

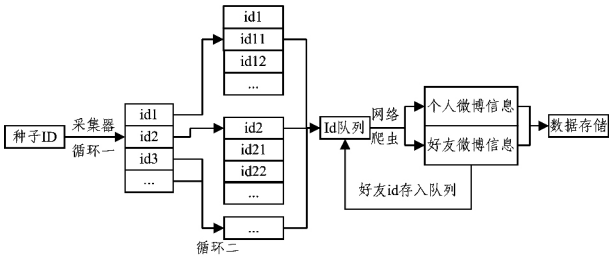

图 4　采集器与网络爬虫融合流程图

表 4　1h 内微博采集条数

采集方案	微博数
API	8100
网络爬虫	8332
采集器	8500
API 与网路爬虫融合	10030
采集器与网络爬虫融合	12790

若要研究分析用户微博数据，快速完整地获取大量数据是前提。在选择实验方法时必须考虑数据获取过程中时间的耗损问题，这可能与程序运行时的网络带宽及数据处理总量多少等因素有关。实验中记录采用 API、网络爬虫、采集器、API 与网络爬虫融合、采集器与网络爬虫融合这 5 种方法连续数小时进行采集所获取的用户数据的数量，通过分析对比得出，长时间数据采集情况下网络爬虫与采集器的融合是最稳定且最高效的采集方案。具体实验结果如图 5 所示，其中横轴为连续采集的时间，纵轴为单位时间内采集处理用户信息的数量。

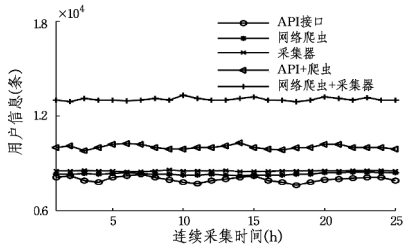

图 5　连续采集性能对比图

对各微博数据采集方案获取的数据进行综合整理分析，记录获取同等量的数据所耗的时间并进行对比，得出各种采集方案的总体性能对比结果，如表 5 所列。

表 5　采集方案的总体性能对比

	API	网络爬虫	采集器	API 与网路爬虫	采集器与网络爬虫
速率	低	中	高	高	较高
完整性	中	高	中	高	较高

此外，研究微博用户行为特征一直存在着重要意义。用户行为特征可通过多方面进行深入探究，就现有的实验结论，可通过微博发布内容、用户兴趣爱好等方面分析总结。文中通过对所获取的实验数据进行实验分析，发现用户微博活跃时间也可作为用户行为特征体现的一个重要因素。对两类用户所发布的微博进行大量采集，并对微博发布时间进行统计。实验结果如图 6 所示，可见在用户活跃时间段存在较大的差异。图中实线标线为微博推送治愈系消息类的用户，其主要发布美文美句或哲理性话语；可看出其主要活跃时间在 20 点到次日 2 点。图中虚线标线为微博推送广告消息类的用户，可见其活跃时间较为正常，在晚间几乎零活动，其活动主要集中于 8 点到 17 点。可见用户的微博活跃时间可体现用户本身的行为特征。

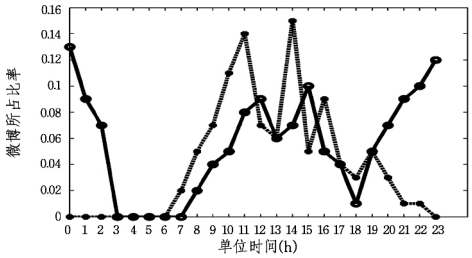

图 6　单位时间内微博统计比例

结束语　随着近几年各种社交网络的迅猛发展，尤其是微博已经发展成一个重要的社会化媒体，从微博中获取新闻时事、人际交往、自我表达、社会分享等内容已成为日常生活中的一部分。文中以新浪微博为研究对象，提出了通过微博 API 接口、模拟登录下基于网络爬虫以及基于采集器 3 种方案实现对微博数据的采集。对于不同的数据需求，实验得出相匹配的方案，且针对各种方案均存在的优缺点，提出对方案进行融合的策略，实验结果显示，融合策略更加快速高效地实现了对数据的采集。从微博数据中可以进行有效的分析和预测、分析网络事件的传播规律以及单个用户的影响力等，这些结果对于引导一些正确舆论、寻找最有潜力用户进行消息传播、宣传等有一定的价值。本文结果进一步反应出通过智能算法进行大数据分析会得到较大的收获。

参 考 文 献

[1] WANG Yuan-zhuo, JIN Xiao-long, CHENG Xue-qi. Network Big Data: Present and Future[J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138. (in Chinese)
王元卓，靳小龙，程学旗. 网络大数据：现状与展望[J]. 计算机学报，2013，36(6)：1125-1138.

[2] 中国互联网信息中心. 第 35 次中国互联网络发展状况报告[EB/OL]. [2015-02-03]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201502/t20150203_51634.htm.

[3] 新浪科技. 2015 年第一季度财务报告[EB/OL]. [2015-05-15]. http://tech.sina.com.cn/i/2015-05-15/doc-iavxeafs7518570.shtml.

[4] STAM K M, CAMERON G T, STAM A, et al. Stam Sociometric Attractiveness on Facebook[J]. Proceedings of the International Conference on Information Manag, 2014, 6(6): 180-188.

[5] ALLOWAY T P, ALLOWAY R G. The impact of engagement with social networking sites (SNSs) on cognitive skills[J]. Computers in Human Behavior, 2012, 28(5): 1748-1754.

[6] DING Zhao-Yun, JIA Yan, ZHOU Bin. Survey of Data Minging for Microblogs[J]. Journal of Computer Research and Develop-

ment,2014,51(4):691-706. (in Chinese)
丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展,2014,51(4):691-706.

[7] LI D,NIU J,QIU M,et al. Sentiment analysis on Weibo data [C]// 2014 IEEE Computing,Communications and IT Applications Conference (ComComAp). IEEE,2014:249-254.

[8] LIAN Jie,ZHOU Xin,LIU Yun. SINA microblog data retrieval [J]. J T sing hua Univ(Sci & Tech),2011,51(10):1300-1305. (in Chinese)
廉捷,周欣,刘云. 新浪微博数据挖掘方案[J]. 清华大学学报(自然科学版),2011,51(10):1300-1305.

[9] HUANG Yan-wei,LIU Jia-yong. Study on Sinamicroblog Data Acquisition Technology[J]. Information Security and Communication Security,2013(6):71-73. (in Chinese)
黄延炜,刘嘉勇. 新浪微博数据获取技术研究[J]. 信息安全与通信保密,2013(6):71-73.

[10] YAO Ke. Open API:Sina micro Bo way? [J]. Internet World,2010(8):71-72. (in Chinese)
姚科. 开放 API:新浪微博必经之路?[J]. 互联网天地,2010(8):71-72.

[11] LI X,XIE Y,LI C,et al. Analyzing the public events' influence via open microblogging APIs[C]// 2012 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE,2012:84-90.

[12] SUN Xiao,YE Jia-qi,TANG Chen-yi,et al. Method of Sina microblogging big data grabbing based on multi-strategy and its application[J]. Journal of Hefei University of Technology,2014,37(10):1210-1215. (in Chinese)
孙晓,叶嘉麒,唐陈意,等. 基于多策略的新浪微博大数据抓取及应用[J]. 合肥工业大学学报(自然科学版),2014,37(10):1210-1215.

[13] YAO Feng. Improvement of Base64 Encoding/Decoding Algorithm in Java[J]. Computer Applications and Software,2008,25(12):164-165. (in Chinese)
姚峰. Java 平台中 Base64 编码/解码算法的改进[J]. 计算机应用与软件,2008,25(12):164-165.

[14] SUN Qing-yun,WANG Jun-feng,ZHAO Zong-qu,et al. A Microblog Data Collection Method Based on Simulated Login Technology[J]. Computer Technology and Development,2014,24(3):6-10. (in Chinese)
孙青云,王俊峰,赵宗渠,等. 一种基于模拟登录的微博数据采集方案[J]. 计算机技与发展,2014,24(3):6-10.

[15] DANGRE A,WANKHEDE V,AKRE P,et al. Design and Implementation of Web Crawler[J]. International Journal of Computer Science & Information Technolo,2014,5(1):921-922.

[16] SHEN D,WANG H,CAO J,et al. The Design and Implement of High Efficient Incremental Microblogging Crawler[C]// 2012 Fourth International Conference on Multimedia Information Networking and Security (MINES). IEEE,2012:537-540.

[17] VASILE A I,PAVALOIU B,CRISTEA P D. Building a specialized high performance web crawler[C]// 2013 20th International Conference on System,Signals and Image Processing (IWSSIP). IEEE,2013:183-186.

[18] WANG Ye. The design and implementation of the theme crawler based on the breadth first[D]. Shanghai:Fudan University,2011. (in Chinese)
王桦. 基于广度优先的主题爬虫的设计与实现[D]. 上海:复旦大学,2011.

[19] LIAN Jie. Research on social network data mining based on user characteristics [D]. Beijing:Beijing Jiaotong University,2013. (in Chinese)
廉捷. 基于用户特征的社交网络数据挖掘研究[D]. 北京:北京交通大学,2013.

[20] LIU J,CAO Z,CUI K,et al. Identifying Important Users in Sina Microblog[C]// 2012 Fourth International Conference on Multimedia Information Networking and Security (MINES). IEEE,2012:839-842.

(上接第 246 页)

[2] HASTIE T,TIBSHIRANI R,FRIEDMAN J. The Elements of Statistical Learning [M]. Data mining,Interface and Prediction New York,Springer,2001.

[3] HANCOCK P J B,BURTON A M,BRUCE V. Face processing:human perception and principal components analysis [J]. Memory and Cognition,1996,24(1):26-40.

[4] MISRA J,SCHMITT W,et al. Interactive Exploration of Microarray Gene Expression Patterns in a Reduced Dimensional Space [J]. Genome Research,2012,12(7):1112-1120.

[5] SHEN Hai-peng,HUANG Jian-hua. Sparse principal component analysis via regularized low rank matrix approximation [J]. Journal of Multivariate Analysis,2008,99(6):1015-1034.

[6] JOLLIFFE I T,UDDIN M. The Simplified Component Technique:An Alternative to Rotated Principal Components [J]. Journal of Computational and Graphical Statistics,2000,9(9):689-710.

[7] JOLLIFFE I T,TRENDAFILOV N T,et al. A Modified Principal Component Technique Based on the LASSO [J]. Journal of Computational and Graphical Statistics,2003,12(3):531-547.

[8] ZOU H,HASTIE T,et al. Sparse principal component analysis [J]. Journal of Computational and Graphical Statistics,2006,15:265-286.

[9] WITTEN D M,et al. A penalized matrix decomposition,with applications to sparse principal components and canonical correlation [J]. Biostatistics,2009,10(3):515-534.

[10] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society,1996,58(1):267-288.

[11] ZOU H. The adaptive lasso and its oracle properties [J]. Journal of the American Statistical Association,2006,101(476):1418-1429.

[12] ALLEN G I,GROSENICK L,et al. the A Generalized Least-Square Matrix Decomposition [J]. Journal of the American Statistical Association,2014,109(505):145-159.

[13] QI Xin,LUO Rui-yan,ZHAO Hong-yu. Sparse principal component analysis by choice of norm[J]. Journal of Multivariate Analysis,2013,114(2):127-160.