



基于微博 API 的分布式抓取技术

陈舜华¹, 王晓彤¹, 郝志峰¹, 蔡瑞初¹, 肖晓军², 卢宇²

(1. 广东工业大学计算机学院 广州 510006; 2. 广州优亿信息科技有限公司 广州 510630)

摘要:随着微博用户的迅猛增长,越来越多的人希望从用户的行为和微博内容中挖掘有趣的模式。针对如何对微博数据进行有效合理的采集,提出了基于微博 API 的分布式抓取技术,通过模拟微博登录自动授权,合理控制 API 的调用频次,结合任务分配控制器高效地获取微博数据。该分布式抓取技术还结合时间触发和内存数据库技术实现重复控制,避免了数据的重复爬取和重复存储,提高了系统的性能。本分布式抓取技术具有可扩展性高、任务分配明确、效率高、多种爬取策略适应不同的爬取需求等特点。新浪微博数据爬取实例验证了该技术的可行性。

关键词:新浪微博;爬取策略;分布式爬取;微博 API

doi: 10.3969/j.issn.1000-0801.2013.08.025

A Distributed Data-Crawling Technology for Microblog API

Chen Shunhua¹, Wang Xiaotong¹, Hao Zhifeng¹, Cai Ruichu¹, Xiao Xiaojun², Lu Yu²

(1. School of Computers, Guangdong University of Technology, Guangzhou 510006, China;

2. Guangzhou Usease Information Technology Co., Ltd., Guangzhou 510630, China)

Abstract: As more and more users begin to use microblog, people eagerly want to dig interesting patterns from the microblog data. How to efficiently collect data from the service provider is one of the main challenges. To address this issue, a distributed crawling solution based on microblog API was present. The distributed crawling solution simulates microblog login, automatically gets authorized, and control the invoked frequency of the API with a task controller. A time trigger method with memory database was also proposed to avoid extra trivial data duplication and improve efficiency of the system. In the distributed framework, the crawling tasks can be assigned to distributed clients independently, which ensures the high scalability and flexibility of the crawling procedure. The feasibility of the crawler technology according to Sina microblog instance was verified.

Key words: Sina microblog, crawling strategy, distributed crawl, microblog API

1 引言

近年来,社交网络的发展引人注目,参考文献[1]介绍了社交网络的发展现状及趋势。目前,约有一半的中国网民通过社交网络沟通交流、分享信息,社交网络已成为覆盖用户最广、传播影响最大、商业价值最高的 Web 2.0 业务。微博作为一种便捷的媒体交互平台,在全球范围内吸

引了数亿用户,已成为人们进行信息交流的重要媒介,用户可以通过微博进行信息记录和交流、娱乐消遣以及社会交往等^[2]。

Twitter 自 2006 年由 Williams E 等人联合推出以来,发展迅猛。Twitter 作为一种结合社会化网络和微型博客的新型 Web 2.0 应用形式正风靡国外,其应用涉及商业、新闻教育等社会领域,已成为网络舆论中最具有影响力的一种^[3]。

2010 年国内微博迎来春天,微博如雨后春笋般崛起,四大门户网站均开设微博。截至 2012 年 12 月底,新浪微博注册用户数已超过 5 亿,同比增长 74%,日活跃用户数达到 4 620 万,微博用户数与活跃用户数保持稳定增长,已经成为中国最大的移动社交网络。

社交网络大数据的分析和和管理^[4]吸引了许多人进行研究,通过对微博数据的挖掘,可以进行预测用户的爱好、预警热点新闻舆论、挖掘人际关系信息、预测信息传播趋势等有趣的研究。成功地预测信息的传播可以提高市场的效益,促进商业产品的推广,也可以实时监控信息传播的新趋势。不管是在商业上还是在政治上,都迫切需要数据采集以支持各种数据分析。利用传统的网页爬虫^[5]爬取数据需要大量时间解析网页,这种方式不但处理速度较慢,而且对数据的整合需要自己进行合理转换,不适合实时数据需求的系统。而基于微博 API 的方法具有手动授权、接口限制等问题。本文以舆情系统作为背景分析数据需求,提出基于微博 API 的数据抓取技术,以新浪微博为例,解决抓取过程中授权过期、接口约束、重复爬取等问题,制定合适的爬取策略,合理分配爬取任务和 API。采用分布式的爬取框架,可扩展性高,可以根据数据爬取需求合理增加或删减爬取机,而不必更改系统的框架。

2 采集系统

2.1 采集目标

本文主要从舆情分析角度分析数据需求,包括对用户和关键字的监控、预警、敏感词分析、用户情感分析。在参考文献[6]中,微博消息的传播途径主要有:转发途径和粉丝途径,这意味着微博数据的爬取主要集中在用户关系图和转发关系图。在微博消息传播中,用户、微博消息和用户关系是直接影响微博传播力的 3 个要素。用户的行为包括对微博的转发、评论、回复、收藏等。在这里,需要爬取微博的基本信息(转发和评论);抓取用户的基本信息(好友、粉丝、用户所发的微博和用户标签)。

2.2 采集约束

在基于微博 API 爬取信息的过程中,爬取的效率跟微博 API 约束有很大关联。具体的 API 限制频率可以查看参考文献[7]。由于 API 的约束,具体的爬取策略既要满足数据需求,又要克服这些约束利用有限的接口爬取更多的信息,需要对用户接口进行合理分配,也需要在程序中合理控制接口的调用情况,设计的爬取程序需要具有良好的容

错能力,无论是 access token 的过期还是接口调用超过限制频次,在程序中都进行了很好的处理。当 access token 过期时,会自动授权,然后继续爬取;当接口调用超过限制频次时,会让程序处于休眠状态直到接口恢复调用,继续爬取。

2.3 采集准则

根据微博 API 数据爬取的特点,提出如下爬取准则:用最少的接口爬取最多的信息,尽量避免重复,可扩展性强。这里的“最少接口爬取最多信息”,是指合理安排接口的分配,尽量根据数据需求分配数据接口,避免重复爬取和存取数据。另外,爬取的规模可以根据需求增加或减少,这要求系统的可扩展性强。

3 采集系统设计

3.1 系统功能模块

(1) 公共微博爬取模块

该模块调用微博 API 的读取接口 `statuses/public_timeline`,是整个爬取系统的入口,只有在公共微博爬取一定的用户信息和微博信息后,才可以进行后续用户维度和微博维度数据的爬取。

(2) 用户维度爬取模块

该模块主要调用微博 API 的用户信息相关接口。在用户维度上,爬取信息主要是指抓取用户的关系图信息,从而建立起用户的关系图,这对信息的传播行为分析至关重要。由于微博接口的限制,每次只能爬取 200 个用户和粉丝,这个数据对于后续的研究是远远不够的,所以用户关系图需要靠微博的转发关系进行扩展,假设用户 A 转发用户 B 的微博,则其存在转发关系。

(3) 微博维度爬取模块

该模块主要调用微博 API 的微博相关接口。在微博维度上,爬取信息主要是指抓取微博的转发、评论信息,从而建立起微博的转发关系图。对于每条微博,每次调用一次 API 只能爬取一层转发关系图,当一层关系图爬取足够多时,可以根据数据库里的信息建立多层转发图。

3.2 系统框架

爬取系统的整体框架如图 1 所示。

采用了分布式的架构并行爬取数据,分成两个主要部分。①~④部分主要抓取最新的公共微博。时间触发服务器在一定时间间隔上触发客户机抓取数据并将数据存入 Hbase,将用户 ID 和微博 ID 存进内存数据库 redis 中。redis 主要有两个作用:一是进行 ID 存储,提供后续微博维

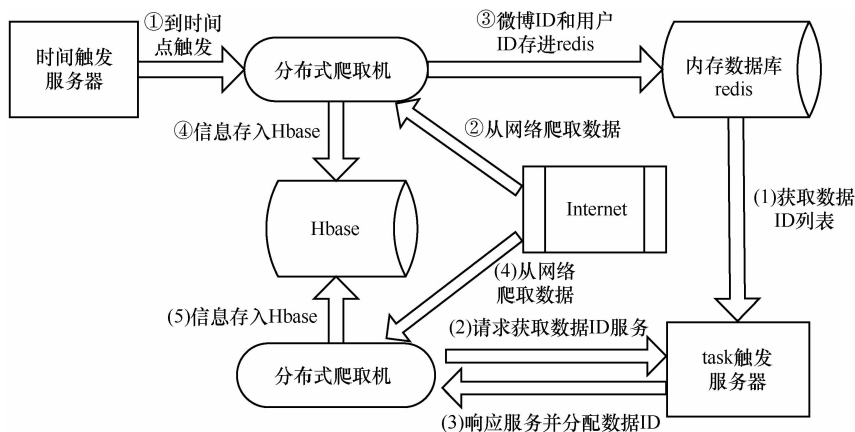


图1 爬取系统的整体框架

度和用户维度的爬取;二是进行重复控制,如果爬取到相同的微博或用户就不将信息存入 Hbase。(1)~(5)部分主要根据 redis 中用户 ID 和微博 ID 抓取其他信息,这里配置了一台 task 触发服务器,初始化时从 redis 获取 ID,每次客户机通过 thrift 请求分配 ID,task 触发服务器响应请求将 ID 传给客户机进行爬取并将爬取后的信息存入 Hbase。整个框架的任务分工明确,可扩展性强,如果想要爬取更多用户信息,只需要增加爬取客户机向 task 触发服务器请求就可以,而不需要更改整个系统的框架。

4 算法设计

对于基于 API 数据的爬取,在很大程度上受限于接口约束和授权码过期等问题,如何设计合理的算法,在遵守准则的前提下克服以上约束,从而实现数据爬取的最优,是本文要解决的问题。

4.1 算法分析

(1)重复控制

基于 API 爬取数据算法需要考虑如下几个问题。第一,申请应用时接口有限,以新浪微博为例,每个应用只有 15 个测试用户,每个用户每小时只能爬取 150 次,也由于 IP 地址每小时 1 000 次的限制,每小时只能有 7 个测试用户在一个 IP 地址上爬取 1 000 次,用有限的接口爬取更多的信息,只有对接口进行有利地分配和调度,才能充分利用有限的接口爬取更多的数据信息,因此必须对爬取操作进行重复控制操作。如果对某个接口,如公共微博的爬取接口 statuses/public_timeline,接口调用频次超过公共微博的产生量,将导致接口很快被调用完,但爬取的数据很多是重复的,既浪费了接口,爬取的数据也少,违背了用最

少的接口爬取更多数据的准则。

基于以上准则,数据爬取的重复控制需要做得很好,才能爬取更多的信息。为了避免重复爬取导致的接口浪费,应用时间触发器。

假设有 12 台机器,每台机器有编号 1、2、3..., 频次 1 000 次/h,1 h 的微博产生量如图 2 所示。

1 h 要完成访问的总频次: $p = 12 \times 1\,000$ 次/h

每台机器的间隔爬取时间: $t = 1/p = (12 \times 1\,000)$

每台机器的休眠时间: $tn = 1/1\,000$

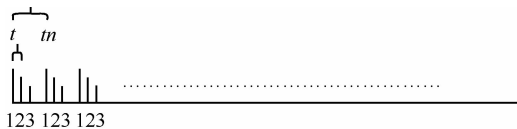
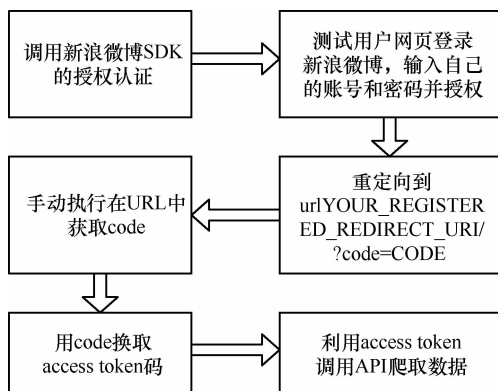


图2 机器抓取微博控制

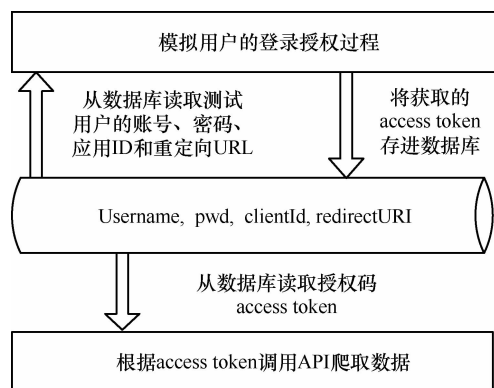
(2)模拟授权

基于微博 API 爬取数据的首要任务是进行用户授权并获得 access token,这样才可以通过微博 API 进行数据抓取。各种主流微博的授权流程基本一致,但又各有不同,仅以新浪微博为例介绍授权基本流程。根据新浪微博授权机制,用户登录授权后获得用户重定向的 URL 中的 code,然后再通过 code 获取 access token,这个过程需要手动执行。对于测试用户来说,授权码只有一天就过期,这不利于爬取程序的自动化,所以在本系统中,模拟了新浪微博的登录过程,自动解析重定向的 URL,自动获取 code 并调用微博接口获得授权码 access token,这样完成了自动授权,提高了系统的效率。OAuth 认证和新浪微博模拟登录过程可以见参考文献[8]。

模拟授权的流程如图 3 所示。



(a) 模拟手动授权



(b) 模拟自动授权

图 3 模拟授权的流程

(3) 任务分配, 接口分配

爬取数据的任务分配和接口分配是很重要的,具体哪台电脑爬取哪些信息,分配多少接口,与数据需求密切相关,接下来讨论的前提是一个应用在一个 IP 地址上的任务分配。一个应用最多有 15 个测试用户,每个测试用户有一个授权码,每个授权码每小时只能调用 150 个接口,一个 IP 地址每小时只能调用 1 000 次,所以只能用 7 个授权码分配给 3 个爬取模块,考虑到数据需求,分配给公共模块 1 个授权码、100 个接口,分配给用户维度爬取 4 个授权码、450 个接口,分配给微博维度爬取 2 个授权码、450 个接口。按照这样的比例,动态添加用户爬取机以扩展系统的爬取规模。

(4) 数据爬取策略

根据数据需求,不同方面的数据采用不同的策略,在公共微博的爬取上,采用时间触发策略,尽量使接口均匀爬取数据,也有助于数据上的重复控制;在时间触发间隔上,实现动态调整,即根据爬取公共微博客户端的数量动态计算爬取时间间隔。

在微博维度和用户维度的爬取上,采用请求响应的策略,可以使多台机器并行爬取数据而又避免重复爬取相同的微博和用户信息,而且可扩展性高,可以随时添加客户机,只要向 thrift 发送请求,就可以进行数据爬取,而不影响整个系统的运作。

4.2 算法流程

图 4 为算法流程。首先,开启 MySQL 和 redis,具体爬取任务分 3 块:公共微博、用户维度和微博维度。用户维度和微博维度都是主动调用 thrift 服务获取用户 ID 或微博 ID 进行爬取,而公共微博因为需要一定的时间跨度,需要控制服务器在一定的时间点触发普通 crawler 爬取公共微博,属于被动爬取机制。在调用 API 的过程中,出现授权码过期,系统会自动重新授权,如果遇到 API 限制,则进入休眠,直到接口调用恢复。

4.3 实验结果

采用 3 个应用、21 个授权码、3 个 IP 地址、9 台机器,实现算法对新浪微博数据一周时间的分布式爬取,实验结果见表 1。

在公共模块,可以爬取最新发布的微博,用户数比微博数少是由于同一个用户可能在一周内发布几条微博。在公共模块采取 3 台机器爬取,每台机器每小时调用 100 个接口,每次最多爬取最新的 200 条,这样理论上可以爬取 10 080 000 条,由于可能重复爬取或接口调用返回结果有时达不到 200 条,所以实验结果比理论值少。在公共模块调用的接口没有提供转发信息,这需要微博维度爬取实现转发情况的获取,由于公共微博是最新发布的,所以并没有调用评论接口进行爬取,而只是简单获取微博和用户的信息。

在用户维度爬取方面,爬取了 75 142 个用户的粉丝、好友、标签和用户发的微博;在微博维度爬取方面,爬取了 75 420 条微博的转发和评论。

从实验结果看,后续可以增加更多的机器爬取用户和微博的其他信息,平衡公共微博爬取的速度。

5 结束语

本文提出了基于微博 API 的分布式抓取技术,根据微博 API 等约束,本着用最少的接口爬取最多数据的原则,制定了具体的爬取策略,并绘制具体的数据爬取流程和系统结构。该爬取系统采用并行爬取方式,可扩展性强,只要有足够的 IP 地址和授权码,就可以扩展系统的爬取节点,而无需改动其他系统结构。相比参考文献[9]

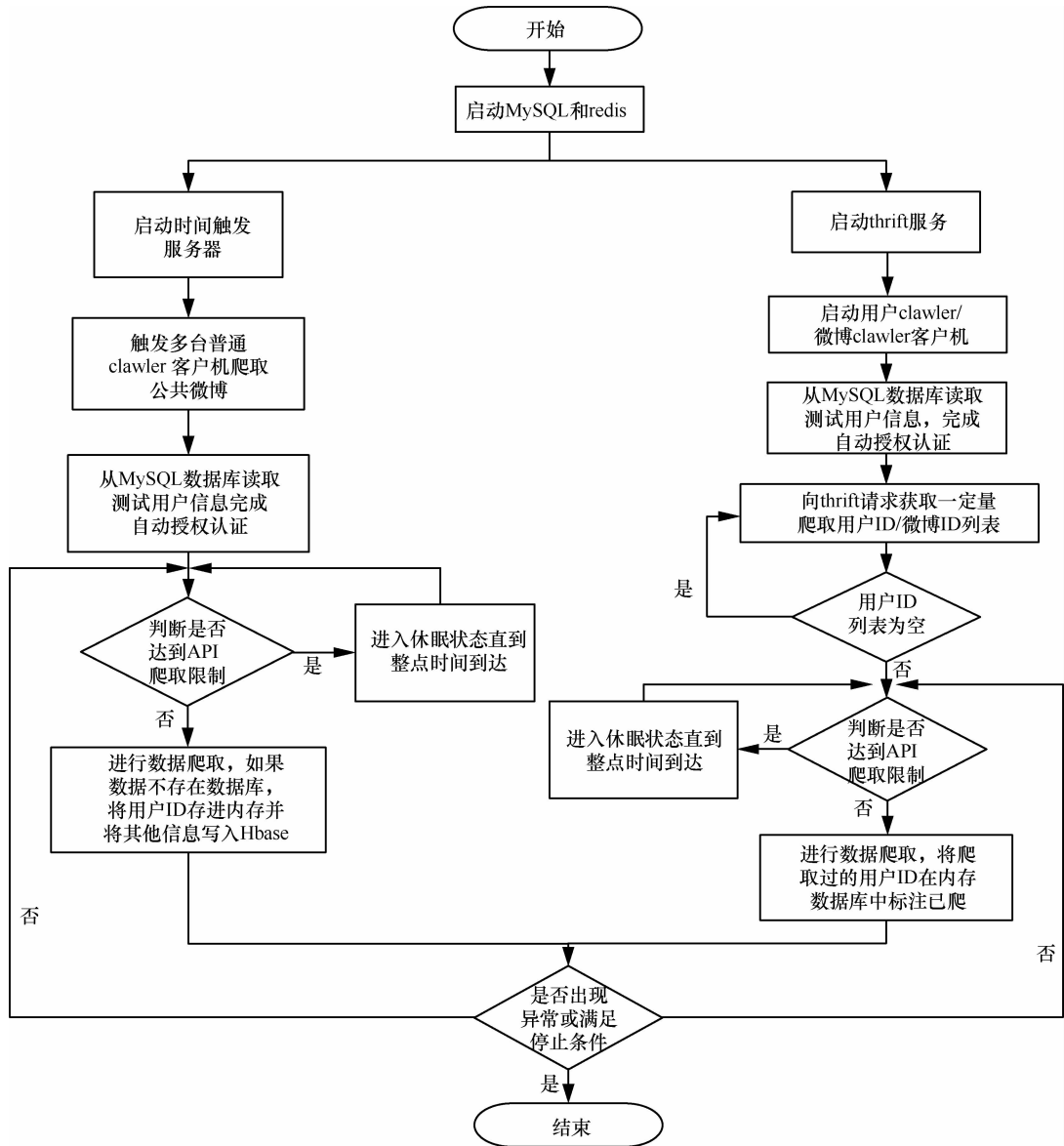


图4 爬取流程

表1 实验结果

	微博数	用户数	转发	评论	粉丝	好友
公共微博(3台机器)	9 984 524	8 825 623	×	×	×	×
用户维度爬取(3台机器)	226 801	75 142(输入)	×	×	2 204 260	1 582 644
微博维度爬取(3台机器)	75 420(输入)	×	186 059	85 619	×	×

提出的挖掘方法,爬取信息更全,可扩展性更高,用户体验更强。

参考文献

- 1 王亮. SNS 社交网络发展现状及趋势. 现代电信科技, 2009(6)
- 2 王娟. 微博客用户的使用动机与行为. 山东大学硕士学位论文
- 3 李保秀. 微博社交网络舆情监测指标体系构建. 科技广场, 2012(4)
- 4 漆晨曦. 电信企业大数据分析、应用及管理发展策略. 电信科学, 2013(3)
- 5 唐波. 网络爬虫的设计与实现. 电脑知识与技术, 2009, 11(5): 2867~2868

(下转第 155 页)

率也越来越高,严重影响无线网络质量。本项目研究并提出了一种降低无线网络互调干扰的频率优化算法:当小区频率组合满足 $935 \times 10^6 \leq f_{\min} < f_{\max} \leq 954 \times 10^6$ Hz,且最大频点号与最小频点号之差小于 65 时,则该小区不可能出现由于频率组合所导致的 5 阶互调干扰。在佛山组织开展试点区域验证,通过方案实施后,试点区域干扰水平明显下降,语音质量、质差话务等 MRR 统计均有不同程度的提升。

参考文献

- 1 郭占涛.移动通信系统中三阶互调干扰的研究和分析.移动通信,2011,35(24)
- 2 蒋江湖,吴少周,张世全.移动通信天线的无源互调干扰研究.陕西理工学院学报(自然科学版),2011,27(2)
- 3 孙智博.无线通信系统中排除三阶互调干扰频率的方法.陕西师范大学学报(自然科学版),2003,31(Z1)
- 4 杨秀.浅谈无线通信系统克服三阶互调干扰措施的演变.移动通信,2011,35(8)
- 5 张世全,傅德民,葛德彪.无源互调干扰对通信系统抗噪性能的影响.电波科学学报,2002,17(2)

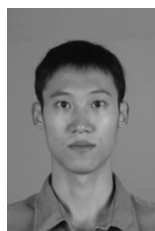
[作者简介]



胡志东,男,现就职于中国移动通信集团广东有限公司网络优化中心网优室,主要从事无线网络优化管理工作。



曾伟超,男,现就职于中国移动通信集团广东有限公司网络优化中心无线室,主要从事无线设备维护以及网络优化管理工作。



唐焯宜,男,现就职于中国移动通信集团广东有限公司网络优化中心网优室,主要从事无线网络优化管理工作。

(收稿日期:2013-06-25)

(上接第 150 页)

- 6 王晶,朱珂,汪斌强.基于信息数据分析的微博研究综述.计算机应用,2012,32(7):2027~2029,2037
- 7 新浪接口约束.<http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI>
- 8 廉捷,周欣,曹伟等.新浪微博数据挖掘方案.清华大学学报,2011,51(10):1300~1305
- 9 周鑫,彭斯俊,罗熹.基于新浪微博开放平台的用户数据挖掘.中国科技论文在线, <http://www.docin.com/p-531480044.html>
- 10 李军,陈震,黄霁崴.微博影响力评价研究.信息安全,2012,3(7)

[作者简介]



陈舜华,女,广东工业大学硕士研究生,主要研究方向为社交网络数据分析、网络爬虫等。

王晓彤,男,广东工业大学硕士研究生,主要研究方向为社交网络影响因子分析。

郝志峰,男,广东工业大学教授、博士生导师,主要从事机器学习、人工智能等研究工作。

蔡瑞初,男,广东工业大学副教授、硕士生导师,主要从事数据挖掘、机器学习、信息检索等研究工作。

肖晓军,男,博士,广州优亿信息科技有限公司高级工程师,具有多年电信行业管理工作经历,主要研究方向为大数据、数据挖掘与电信行业应用等。

卢宇,男,广州优亿信息科技有限公司工程师,主要从事大数据、机器学习等相关研发工作。

(收稿日期:2013-07-10)