# 微博系统数据抓取设计

# 一、总体目标

在现有微博采集、解析系统上，通过增加新的接口、功能来支持新版微博系统的数据采集要求，功能目标包括：

1、支持统一维护（讯库系统、微博系统）采集关键词信息，增加关键词的优先级，按照优先级采集数据；

2、提供微博API（普通API）抓取数据功能。根据业务数据设置，控制每个客户使用API的情况。

3、提供付费接口的功能，并结合业务数据设置实现对接口的使用情况进行控制。

4、微博解析功能进一步完善，增加转发微博，记录数据来源（主要是付费接口来源），准确解析微博的转发数、评论数。**增加人民微博的解析功能。**

5、提供统一的应用接口，包括：提交抓取任务、输出抓取数据等。

# 二、架构设计



数据抓取系统架构

**解析服务**：提供微博网页数据的解析功能。（已有，完善解析服务，增加转发URL，准确解析数量信息。

**对外WebService**：发布到外网，组成分布式数据采集系统的中枢（已有）。

**智能代理SmartAgent**：分布式采集终端（已有）。

**数据存储**：负责中转数据（已有，完善新功能）

**API抓取、调度服务**：新增功能，负责使用API抓取数据，调度与API相关的任务。

**内部WebService**：提供与内部系统交互接口。

# 三、详细设计

**微博数据抓取资源分配**

|  |  |  |
| --- | --- | --- |
| **待抓取项** | **资源分配** | **说明** |
| 关键词搜索 | 网页爬取 | 1、新浪微博：不登陆直接搜索；  2、腾讯微博：需维护登陆信息，带Cookie搜索；  3、人民微博：需维护登陆信息，带Cookie搜索； |
| 微博账号信息 | API | 1、新浪微博：获取用户信息（每次一个账号）  <http://open.weibo.com/wiki/2/users/show>  2、腾讯微博：获取批量用户的个人资料  <https://open.t.qq.com/api/user/infos> |
| 微博账号的粉丝数、微博数 | API | 1、新浪微博：批量获取用户的粉丝数、关注数、微博数，<http://open.weibo.com/wiki/2/users/counts>  2、腾讯微博：获取批量用户的个人资料  <https://open.t.qq.com/api/user/infos> |
| 微博的转发数、评论数 | API | 1、新浪微博：批量获取指定微博的转发数评论数  <http://open.weibo.com/wiki/2/statuses/count>  2、腾讯微博：转播数或点评数（批量）  <https://open.t.qq.com/api/t/re_count> |
| @我的微博 | API | 1、新浪微博：获取最新的提到登录用户的微博列表，即@我的微博  <http://open.weibo.com/wiki/2/statuses/mentions>  2、腾讯微博：获取@当前用户的最新微博  <https://open.t.qq.com/api/statuses/mentions_timeline> |
| 微博账号发布的微博（原发、转发） | API | 1、新浪微博：获取某个用户最新发表的微博列表（限制为当前授权账号）  <http://open.weibo.com/wiki/2/statuses/user_timeline>  2、腾讯微博：获取某用户的最新微博（不限制账号）  https://open.t.qq.com/api/statuses/user\_timeline |
| 微博的转发微博列表 | API | 1、新浪微博：获取指定**原创微博**的转发微博列表  <http://open.weibo.com/wiki/2/statuses/repost_timeline>  实现上需要考虑，只有原创微博有必要使用该接口。  2、腾讯微博：获取单条微博的转发或评论列表  https://open.t.qq.com/api/t/re\_list |
| 账号的粉丝列表 | API | 1、新浪微博：获取用户的粉丝列表（授权账号）  <http://open.weibo.com/wiki/2/friendships/followers>  非授权账号的粉丝列表通过网页爬取。  2、腾讯微博：获取用户的粉丝列表（无账号限制）  <https://open.t.qq.com/api/friends/user_fanslist> |
| 发微博 | API | 1、新浪微博：  发布一条新微博（原创）  <http://open.weibo.com/wiki/2/statuses/update>  转发一条微博  <http://open.weibo.com/wiki/2/statuses/repost>  对一条微博进行评论  <http://open.weibo.com/wiki/2/comments/create>  statuses/repost 转发一条微博信息  statuses/destroy 删除微博信息  statuses/update 发布一条微博信息  statuses/upload 上传图片并发布一条微博  2、腾讯微博：  发表一条微博信息 t/add  发表一条带图片的微博 t/add\_pic  用图片URL发表带图片的微博 t/add\_pic\_url  发表带心情的微博 t/add\_emotion  发表音乐微博 t/add\_music  发表带视频内容的微博 t/add\_video  发表带视频、音乐、图片等内容的微博 t/add\_multi  上传一张图片 t/upload\_pic  转发一条微博 t/re\_add  删除一条微博 t/del  评论一条微博 t/comment  3、人民微博：WebService |

人民微博：只有发微博API。其他微博、账号相关信息只能通过网页爬取。

API频率限制信息：

新浪微博：

<http://open.weibo.com/wiki/%E6%8E%A5%E5%8F%A3%E8%AE%BF%E9%97%AE%E9%A2%91%E6%AC%A1%E6%9D%83%E9%99%90>

腾讯微博：

<http://wiki.open.t.qq.com/index.php/API%E8%B0%83%E7%94%A8%E6%9D%83%E9%99%90>

# 四、接口设计

## 网页爬虫设计

### 1、解析服务

#### 接口定义

通过原有SAWebService提供解析服务接口，

WebService地址：<http://agent.xunku.org:8080/axis2/services/AgentService>

|  |  |  |  |
| --- | --- | --- | --- |
| /\*\*  \* 解析搜索界面  \* **@param** url 完整搜索url  \* **@param** content 采用utf-8编码并且使用gzip压缩后的网页内容  \* **@return**  \*/  **byte**[] parseSearchPage(String url,**byte**[] content);  返回值示例：  **url** http://s.weibo.com/weibo/%25E6%25B5%25B7%25E5%25B0%2594&Refer=index  currentPage 当前页码  maxPageNum 最大页码  dataList 微博数组。   |  | | --- | | {"dataList":[{"mid":"3695665680578000","uid":"","ucode":"315826888","uname":"后知后觉股市直播","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404160703","content":"后知后觉股市直播：【巴菲特收购国际核暖通设备供应商】相关空调(龙头)概念股：格力电器（000651）、松芝股份（002454）、美菱电器（000521）、美的集团（000333）、青岛海尔（600690）、海信科龙（000921）。\n ","comtCount":"2","relayCount":"4","url":"http://weibo.com/2244291964/AE8Hu2qmA","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/2244291964/50/5631169867/1","transferUrl":""},{"mid":"3695658764187852","uid":"","ucode":"chuangyejia","uname":"创业家杂志","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404153934","content":"创业家杂志：互联网思维大热，用互联网思维改造传统产业的案例四处迸发。传统产业中，张瑞敏以互联网再造海尔生态圈；TCL寄望引入互联网思维实现战略转型突破；李书福要带动吉利汽车融入互联网大潮…去哪儿、58同城等中间页扎堆IPO，成为李彦宏互联网思维如何从观念引导扩展为产业影响的背书。http://t.cn/8siFCxZ\n ","comtCount":"14","relayCount":"99","url":"http://weibo.com/1642482194/AE8wkhzs0","relayContent":"","verify":false,"imageUrl":"http://tp3.sinaimg.cn/1642482194/50/5662088971/1","transferUrl":"http://weibo.com/1642482194/AE8wkhzs0"},{"mid":"3695580066815333","uid":"","ucode":"peterlynchfan","uname":"付羽\_投资笔记","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404102651","content":"付羽\_投资笔记：2014年的行情很难做，2013年全年都不断有很想买的股票持续挖掘出来，鲁泰榨菜三全海尔冷链，但是2014年过去几个月了，还真没有什么很想买的股票，都是比较勉强买点试试的。。\n ","comtCount":"37","relayCount":"20","url":"http://weibo.com/1880937880/AE6tosAYJ","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/1880937880/50/5685710527/1","transferUrl":""},{"mid":"3695679585041388","uid":"","ucode":"3928744205","uname":"海尔商用冷柜","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404170218","content":"海尔商用冷柜：#上海展会进行时# 2014国际手工冰激凌大师赛现场，海尔商用冷柜总监安澜为获奖选手颁奖。\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/3928744205/AE93Ul9uI","relayContent":"","verify":false,"imageUrl":"http://tp2.sinaimg.cn/3928744205/50/22863902943/1","transferUrl":"http://weibo.com/3928744205/AE93Ul9uI"},{"mid":"3695677462971071","uid":"","ucode":"itnumber1","uname":"中国网一管家","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404165352","content":"中国网一管家：【海尔电商公布O2O战略】海尔旗下电商平台日日顺（RRS.COM）公布了其O2O战略，围绕家电、家装、家具、家饰、水家电、婴童六大市场。日日顺还谋求海尔线下3万5千家加盟店转型物流服务商，并与家居广场打通，配合8万人的服务兵团与9万台送装一体货车，帮助平台商家从工厂到消费者的全链路价值体现。 \n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/2707216535/AE90ucsUv","relayContent":"","verify":false,"imageUrl":"http://tp4.sinaimg.cn/2707216535/50/5629522161/1","transferUrl":"http://weibo.com/2707216535/AE90ucsUv"},{"mid":"3695673965521707","uid":"","ucode":"2937591795","uname":"Galangal","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404163958","content":"Galangal：#四驱兄弟##脑洞# 因为我对乔妹是真爱所以想给她换个礼服裙什么的，不过穿上之后可能会不高兴，因为人家是高速型赛车手（没有逻辑）。以及我真的是这么多年后才知道海尔家有城堡，真的是来自德国的小王子，不枉我当年苏过你和烈兄贵的CP！（还是没有逻辑）\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/2937591795/AE8UQnarN","relayContent":"","verify":false,"imageUrl":"http://tp4.sinaimg.cn/2937591795/50/40015454380/0","transferUrl":"http://weibo.com/2937591795/AE8UQnarN"},{"mid":"3695668570453247","uid":"","ucode":"3893440564","uname":"统帅玩定制-Commander队","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404161832","content":"统帅玩定制-Commander队：大家造吗？我们的微博开通啦我们是海尔统帅电器的Commander队～由黄佳红、王佩云、吴苑、章彬慧四个活力四射的美女组成，守护冰箱，看我玩转定制，我们只为冰箱代言~我们参加的是：海尔集团#统帅电器玩转定制# 大赛！求关注求赞！@统帅电器玩转定制\n ","comtCount":"3","relayCount":"5","url":"http://weibo.com/3893440564/AE8M91TUr","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/3893440564/50/5691565196/0","transferUrl":"http://weibo.com/3893440564/AE8M91TUr"},{"mid":"3695664422640915","uid":"","ucode":"hrbiomedical","uname":"海尔生物医疗","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404160203","content":"海尔生物医疗：#爱?生活#你来拍砖，我们送奖，小伙伴们，海尔生物医疗” 不要“拍砖”要“垒砖”——全流程开放日活动”火热进行中，赶快码字赢话费啦，so easy!\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/3462404802/AE8Fsb51p","relayContent":"","verify":false,"imageUrl":"http://tp3.sinaimg.cn/3462404802/50/5670389831/1","transferUrl":"http://weibo.com/3462404802/AE8Fsb51p"},{"mid":"3695659871696909","uid":"","ucode":"2565902904","uname":"守护那片海MI","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404154358","content":"守护那片海MI：承载着儿时记忆的@海尔 兄弟动画片，今天再复习一下。再重温那5点下课就跑着回家看动画片的快乐。打雷要下雨，雷欧。下雨要打伞，雷欧。天冷穿棉袄，雷欧，雷诶欧，天热扇扇子。智慧就是，这么简单。\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/2565902904/AE8y777rv","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/2565902904/50/5661659158/1","transferUrl":"http://weibo.com/2565902904/AE8y777rv"},{"mid":"3695652824986711","uid":"","ucode":"qdaustar","uname":"青岛澳星","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404151558","content":"青岛澳星：年度最佳美国EB5项目—亨德森酒店项目最新进展？如何解析英国投资移民政策修改法案？2014澳洲移民最新发展趋势？新西兰创业移民重新开放，EOI打分机制怎样操作？以上所有的问题，只要您4月12日来海尔洲际酒店参加澳星《美国/英国/澳洲/新西兰移民新政移民讲座》就能全部解决！\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/2618697587/AE8mKkVgP","relayContent":"","verify":false,"imageUrl":"http://tp4.sinaimg.cn/2618697587/50/5662081400/0","transferUrl":"http://weibo.com/2618697587/AE8mKkVgP"},{"mid":"3695650837324572","uid":"","ucode":"3961967384","uname":"南京档案","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404150804","content":"南京档案：【南师附小成立全省首家少年科学院】#南京历史上的今天 #1993年4月4日，江苏省第一家少年科学院在南京师范大学附属小学成立。2002年，南师附小六年级学生石尔秋潜心研究的“电脑红绿灯”，在“中国少年海尔科技奖”的评选中获得江苏省唯一的一等奖，成为全国１０名一等奖中的第１名。@南京教育\n ","comtCount":"0","relayCount":"1","url":"http://weibo.com/3961967384/AE8jxuJsg","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/3961967384/50/40043181351/1","transferUrl":"http://weibo.com/3961967384/AE8jxuJsg"},{"mid":"3695648953686531","uid":"","ucode":"2627319344","uname":"上海红酒交易中心SHWEX","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404150035","content":"上海红酒交易中心SHWEX：#中心快讯#近期，辛辛那提大学的凯利?海尔姆斯破译了一份古罗马埃及时期的葡萄园雇佣合同，这份合同写在一张12X8.5厘米的纸莎草纸上，记录了一位名叫Flavius的男子同意在埃及一个村庄附近的葡萄园当守卫，直至葡萄收获运输完毕。全文阅读：http://t.cn/8str0ij\n ","comtCount":"1","relayCount":"0","url":"http://weibo.com/2627319344/AE8gvft2b","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/2627319344/50/40045407789/0","transferUrl":"http://weibo.com/2627319344/AE8gvft2b"},{"mid":"3695646551792977","uid":"","ucode":"3582004075","uname":"皇帝崛起","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404145102","content":"皇帝崛起：公布答案：1.柯南 2.蜡笔小新 3.圣斗士星矢 4.大头儿子小头爸爸 5.机器猫 6. 千与千寻 7.蓝精灵 8.阿拉蕾 9.铁壁阿童木 10.海尔兄弟 11.乱码1/2 12.哪吒闹海 13.白雪公主 14.恐龙特急克塞号 15.奥特曼 16. 犬夜叉 17.没头脑和不高兴 18.大力水手 19.葫芦娃 20.忍者神龟。你猜出来了没有\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/3582004075/AE8cD7wqZ","relayContent":"","verify":false,"imageUrl":"http://tp4.sinaimg.cn/3582004075/50/5674451692/0","transferUrl":""},{"mid":"3695646290946275","uid":"","ucode":"jmwbong","uname":"中国万邦","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404145000","content":"中国万邦：【转型已成大势 不只是传统企业】关于企业转型的说法已经很多了，但是2014年的行业转型重点仍旧不会是传统企业，除了海尔这样的科技型制造业有其先瞻，大部分传统制造业还会继续在繁冗的组织架构中消耗，很... --发布到微刊《最新娱乐新闻》http://t.cn/8sisbGj\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/3907412424/AE8cd3Yav","relayContent":"","verify":false,"imageUrl":"http://tp1.sinaimg.cn/3907412424/50/40050963232/1","transferUrl":"http://weibo.com/3907412424/AE8cd3Yav"},{"mid":"3695645694773227","uid":"","ucode":"hzcntrlaircndtn","uname":"海尔中央空调","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404144738","content":"海尔中央空调：#悬浮的力量#海尔中央空调致力于融合最前沿磁悬浮科技，将最节能、环保、高效的家用电器带进更多家庭，让每一个人都能享受美好生活！\n ","comtCount":"17","relayCount":"43","url":"http://weibo.com/3664108177/AE8bfk1Jx","relayContent":"◆◆\n @艾肯空调制冷网：【图集：解密海尔中央空调互联网战略之产品交互体验】3月25日。海尔中央空调2014网络化战略发布暨交互平台创新峰会在武汉举行。在此次峰会上，海尔中央空调再一次强调了交互是其互联网战略的核心。http://t.cn/8siU3xR @海尔中央空调\n 转发(44)|评论(6)\n 4月3日 15:56 来自360浏览器超速版\n ","verify":false,"imageUrl":"http://tp2.sinaimg.cn/3664108177/50/5678767769/1","transferUrl":"http://weibo.com/2213879332/ADZcN9CuL"},{"mid":"3695645057148663","uid":"","ucode":"shreine","uname":"上海瑞芙臣医疗美容","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404144505","content":"上海瑞芙臣医疗美容：【瑞芙臣爱笑会议室】一闺蜜，个性比较好强，胸小。一天其男友要求一起共浴。她沉思片刻认真的说：“我们俩脱了衣服只剩内裤站在一起，就是海尔兄弟，你还有感觉吗？”男友默默的一个人走向的浴室……#瑞芙臣年轻化#\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/3877344166/AE8adtZH1","relayContent":"","verify":false,"imageUrl":"http://tp3.sinaimg.cn/3877344166/50/40048629846/0","transferUrl":"http://weibo.com/3877344166/AE8adtZH1"},{"mid":"3695644239678761","uid":"","ucode":"zhilengwang","uname":"中国制冷网微博","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404144151","content":"中国制冷网微博：【海尔“爆冷”参展 你准备好了吗？】海尔作为全球制冷空调暖通行业屈指可数的企业，也将重拳出击，那么在今年的制冷展（W3展馆 D35），素来受热捧的海尔小兄弟，将有怎样的“爆冷”的举动？http://t.cn/8siF1ff @海尔广州 @海尔\n ","comtCount":"0","relayCount":"2","url":"http://weibo.com/2280853670/AE88TEBT3","relayContent":"","verify":false,"imageUrl":"http://tp3.sinaimg.cn/2280853670/50/5607222831/0","transferUrl":""},{"mid":"3695643828577291","uid":"","ucode":"3969236073","uname":"中家家电商城","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404144013","content":"中家家电商城：海尔AMT抗菌防霉窗垫，抗菌率高达99%，防霉等级达到最高零级标准。有效抑制窗垫在长期使用过程中孳生的霉菌，避免衣物与内桶的交叉污染，有效呵护全家人的健康。http://t.cn/8siFE8X\n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/3969236073/AE88ezZlp","relayContent":"","verify":false,"imageUrl":"http://tp2.sinaimg.cn/3969236073/50/40049306037/1","transferUrl":"http://weibo.com/3969236073/AE88ezZlp"},{"mid":"3695641017979011","uid":"","ucode":"chinaiol","uname":"产业在线","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404142903","content":"产业在线：【海尔“爆冷”参展 你准备好了吗？】在今年的制冷展（W3展馆D35），素来受热捧的海尔小兄弟，将有怎样的“爆冷”的举动？ \_产业在线 http://t.cn/8siF4jz \n ","comtCount":"0","relayCount":"0","url":"http://weibo.com/2279829263/AE83HxtHJ","relayContent":"","verify":false,"imageUrl":"http://tp4.sinaimg.cn/2279829263/50/5670638358/0","transferUrl":""},{"mid":"3695639655471680","uid":"","ucode":"loverockets","uname":"LOVE\_火箭","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404142337","content":"LOVE\_火箭：《麦克海尔：卡南还在努力弄明白该如何打比赛》卡南在最近三场比赛中进入到了轮换阵容当中，谈到卡南的表现麦克海尔表示，卡南还在努力弄明白该如何去打比赛。虽然他在进攻端展示出了一些天赋，但也经历了成长的烦恼，在这三场比赛中他21投6中，场均贡献7.3 http://t.cn/8sikszs \n ","comtCount":"12","relayCount":"1","url":"http://weibo.com/1983655155/AE81vmXqU","relayContent":"","verify":false,"imageUrl":"http://tp4.sinaimg.cn/1983655155/50/1298873092/1","transferUrl":"http://weibo.com/1983655155/AE81vmXqU"}],"currentPage":1,"maxPageNum":50} |   /\*\*  \* 解析账号主页  \* **@param** url 账户主页url  \* **@param** content 采用utf-8编码并且使用gzip压缩后的网页内容  \* **@return**  \*/  **byte**[] parseHomePage(String url,**byte**[] content);  返回值示例：  url http://weibo.com/haierexpo  weiboPageUrl 查看更多微博的链接（当前账号）  fansListPageUrl 当前账号的粉丝列表链接。  followListPageUrl 当前账号的关注列表链接。  user 当前账号的信息。  dataList 微博数组。   |  | | --- | | {"weiboPageUrl":"http://weibo.com/p/1006061728875812/weibo","pageId":"1006061728875812","fansListPageUrl":"http://weibo.com/1728875812/fans","followListPageUrl":"http://weibo.com/1728875812/follow","user":{"uid":"1728875812","ucode":"haierexpo","uname":"海尔","domainid":1,"verify":true,"image":"http://tp1.sinaimg.cn/1728875812/180/5669966281/1","addr":"","sex":"n","userType":"p","fansCount":97879,"followCount":231,"weibocount":5917},"dataList":[{"mid":"3695628015638933","uid":"1728875812","ucode":"haierexpo","uname":"海尔","domain":"weibo.com","domainName":"新浪微博","postDate":"20140404133723","content":"#海尔兄弟新形象# 征集活动仍在持续进行中！今日推荐：来自@Sino\_Shih 的作品~帅气的琴岛、天真的海尔，再加上科技感十足的小机器人“海宝”，维持原形象特征的同时又加入了创新，是不是让你眼前一亮？周边设计中的本子和T恤大爱，完全可以直接制作有木有！！！@海尔兄弟官方微博 @站酷网 ","comtCount":"2","relayCount":"5","url":"http://weibo.com/1728875812/AE7IJnEWx","relayContent":"","verify":true,"imageUrl":"http://tp1.sinaimg.cn/1728875812/180/5669966281/1","transferUrl":""},{"mid":"3695578355594371","uid":"1728875812","ucode":"haierexpo","uname":"海尔","domain":"weibo.com","domainName":"新浪微博","postDate":"20140403120852","content":"海尔兄弟官方微博#晒卖萌得玩偶# 进行中！快来参加~\\(≧▽≦)/~//@海尔兄弟官方微博: #晒卖萌得玩偶#今天活动仍在继续，想要海尔兄弟第一批玩偶的，快来晒出你或者你身边的各种萌人萌物吧！~记得写话题和@海尔兄弟官方微博 另外按要求发完微博请自觉给我们发私信哦！~ ","comtCount":"0","relayCount":"0","url":"http://weibo.com/1907706047/ADXIj9j3I","relayContent":"@海尔兄弟官方微博 \n 不少盆友反馈今天收到了海尔兄弟愚人节的私信~额这个…… 其实关注对了官微，愚人节每天都过……那咱们就一起自娱自乐、卖萌到底吧！晒出你的卖萌照，加上#晒卖萌得玩偶#并且@海尔兄弟官方微博 就能获得第一批海尔兄弟的最新版玩偶\u0026amp;互粉！！（对！就是置顶微博叫大家投票的那个！）来比比谁更萌~\n ","verify":true,"imageUrl":"http://tp1.sinaimg.cn/1728875812/180/5669966281/1","transferUrl":""},{"mid":"3695414462995811","uid":"1728875812","ucode":"haierexpo","uname":"海尔","domain":"weibo.com","domainName":"新浪微博","postDate":"20140403093218","content":"太！全！了！！！这必须是今天最铁粉~大家快来围观！！！官博君居然到现在才看到，官博君有罪必须粉这位用户大大！ ","comtCount":"7","relayCount":"2","url":"http://weibo.com/1727860695/ADWGL4wmP","relayContent":"@曦曦已是资深懒癌晚期患者 \n #发真人照以证清白#钢丝粉哦！全套家电有木有～哦对了还有个海尔的手机～@海尔 @海尔 @海尔 嘎，cc君～托你的福～～@爱打Dota的海贼王 |发真人照以证清白\n ","verify":true,"imageUrl":"http://tp1.sinaimg.cn/1728875812/180/5669966281/1","transferUrl":""},{"mid":"3695391205256822","uid":"1728875812","ucode":"haierexpo","uname":"海尔","domain":"weibo.com","domainName":"新浪微博","postDate":"20140403214118","content":"这可真的是我们家超早期的产品之一~妹纸去看神马电影吖？带我带我~\\(≧▽≦)/~已粉！！！ ","comtCount":"3","relayCount":"0","url":"http://weibo.com/1980245395/AE1sFhaPB","relayContent":"@董小翠爱吃菜\_ \n #发真人照以示清白#@海尔 超级古老的洗衣机有木有 我们去看电影吧 |发真人照以示清白 我在这里:|工商银行数据中心停车场\n ","verify":true,"imageUrl":"http://tp1.sinaimg.cn/1728875812/180/5669966281/1","transferUrl":""},{"mid":"3695382523179539","uid":"1728875812","ucode":"haierexpo","uname":"海尔","domain":"weibo.com","domainName":"新浪微博","postDate":"20140403210335","content":"这……海尔多年来都是各大学校的教材案例吖！我们会为占据大家的教科书\u0026amp;论文继续努力的！ヾ(≧へ≦)〃[嗯!] 已粉！ ","comtCount":"7","relayCount":"1","url":"http://weibo.com/1727938213/AE1dlxXcd","relayContent":"@chiyo就是吃油 \n @海尔 据说今天在清扫僵尸粉 但是由于我住校就不PO昂贵周边图了（。 但是真爱粉是把你当研究课题来做的 我和几个同学做了海尔的跨文化传播研究！你要论文和PPT吗！（。 于是附图是PPT和论文的部分截图！\n ","verify":true,"imageUrl":"http://tp1.sinaimg.cn/1728875812/180/5669966281/1","transferUrl":""}]} |   /\*\*  \* 解析粉丝列表页面  \* **@param** url 账号粉丝页url  \* **@param** content 采用utf-8编码并且使用gzip压缩后的网页内容  \* **@return**  \*/  **byte**[] parseFansPage(String url,**byte**[] content);  返回值示例：  nextPage 下一页链接（需要处理unicode转义字符），可以使用该链接继续抓取下一页数据进行解析。  homeUser 主账号的信息。  fansList 粉丝账号的数组。   |  | | --- | | {"nextPage":"http://weibo.com/p/1002061191965271/follow?pids\u003dPl\_Official\_LeftHisRelation\_\_32\u0026page\u003d3#place","homeUser":{"uid":"1191965271","ucode":"lifeweek","uname":"三联生活周刊","domainid":1,"verify":true,"image":"http://tp4.sinaimg.cn/1191965271/180/5662870402/1","addr":"","sex":"n","userType":"p","fansCount":9031301,"followCount":521,"weibocount":11462},"fansList":[{"uid":"2824643483","ucode":"williamakers","uname":"WilliamAkers","domainid":1,"verify":false,"image":"http://tp4.sinaimg.cn/2824643483/50/5635217836/1","addr":"海外 美国","sex":"m","userType":"p","fansCount":1868,"followCount":1,"weibocount":295},{"uid":"2004713437","ucode":"2004713437","uname":"张可悟","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/2004713437/50/5610346794/0","addr":"北京 东城区","sex":"f","userType":"p","fansCount":170,"followCount":85,"weibocount":962},{"uid":"3637225275","ucode":"3637225275","uname":"Small-South","domainid":1,"verify":false,"image":"http://tp4.sinaimg.cn/3637225275/50/5671351243/1","addr":"其他","sex":"m","userType":"p","fansCount":18,"followCount":5,"weibocount":2},{"uid":"1761984404","ucode":"1761984404","uname":"洁尘","domainid":1,"verify":false,"image":"http://tp1.sinaimg.cn/1761984404/50/5671931531/0","addr":"四川 成都","sex":"f","userType":"p","fansCount":25776,"followCount":816,"weibocount":8695},{"uid":"1677914550","ucode":"21cake","uname":"21cake","domainid":1,"verify":false,"image":"http://tp3.sinaimg.cn/1677914550/50/39999038040/0","addr":"北京","sex":"f","userType":"p","fansCount":168249,"followCount":21,"weibocount":1548},{"uid":"2102697324","ucode":"xuepenghere","uname":"薛芃","domainid":1,"verify":false,"image":"http://tp1.sinaimg.cn/2102697324/50/40030561637/0","addr":"北京 海淀区","sex":"f","userType":"p","fansCount":684,"followCount":487,"weibocount":745},{"uid":"1699540307","ucode":"cnrzgzs","uname":"中国之声","domainid":1,"verify":false,"image":"http://tp4.sinaimg.cn/1699540307/50/5690693210/1","addr":"北京","sex":"m","userType":"p","fansCount":6074231,"followCount":573,"weibocount":40333},{"uid":"3363206842","ucode":"mediaservices","uname":"媒体微博小秘书","domainid":1,"verify":false,"image":"http://tp3.sinaimg.cn/3363206842/50/5678612355/1","addr":"北京 东城区","sex":"m","userType":"p","fansCount":101881,"followCount":302,"weibocount":2474},{"uid":"1782394737","ucode":"myv1","uname":"第一视频","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/1782394737/50/22870647161/1","addr":"北京 朝阳区","sex":"m","userType":"p","fansCount":1354078,"followCount":624,"weibocount":7589},{"uid":"3367309494","ucode":"3367309494","uname":"老虎苍蝇一起打\_","domainid":1,"verify":false,"image":"http://tp3.sinaimg.cn/3367309494/50/40021651924/0","addr":"北京","sex":"f","userType":"p","fansCount":63608,"followCount":23,"weibocount":40},{"uid":"2720241830","ucode":"parkerpenchina","uname":"ParkerPen派克笔","domainid":1,"verify":false,"image":"http://tp3.sinaimg.cn/2720241830/50/40036796504/1","addr":"上海","sex":"m","userType":"p","fansCount":14554,"followCount":390,"weibocount":2634},{"uid":"2083844833","ucode":"sohuexpress","uname":"搜狐新闻客户端","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/2083844833/50/22868821480/0","addr":"北京 海淀区","sex":"f","userType":"p","fansCount":796133,"followCount":683,"weibocount":19487},{"uid":"2607294791","ucode":"2607294791","uname":"为人民服务ING正传","domainid":1,"verify":false,"image":"http://tp4.sinaimg.cn/2607294791/50/5650385032/1","addr":"河南 洛阳","sex":"m","userType":"p","fansCount":7530,"followCount":548,"weibocount":248},{"uid":"3517534461","ucode":"3517534461","uname":"燃-Ran-","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/3517534461/50/5665646230/0","addr":"北京 东城区","sex":"f","userType":"p","fansCount":1440,"followCount":2,"weibocount":20},{"uid":"1219548974","ucode":"sudanno1","uname":"苏丹NO1","domainid":1,"verify":false,"image":"http://tp3.sinaimg.cn/1219548974/50/40050903241/0","addr":"北京 朝阳区","sex":"f","userType":"p","fansCount":35713,"followCount":2125,"weibocount":5678},{"uid":"1740754260","ucode":"ouraudistyle","uname":"奥迪汇","domainid":1,"verify":false,"image":"http://tp1.sinaimg.cn/1740754260/50/5681094851/1","addr":"吉林 长春","sex":"m","userType":"p","fansCount":564215,"followCount":751,"weibocount":10809},{"uid":"1821935413","ucode":"beijingwenyishenghuo","uname":"文艺生活周刊","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/1821935413/50/40051295188/0","addr":"北京","sex":"f","userType":"p","fansCount":33745,"followCount":746,"weibocount":10124},{"uid":"2335590665","ucode":"gefhsy","uname":"GE医疗中国","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/2335590665/50/22870639760/0","addr":"北京 朝阳区","sex":"f","userType":"p","fansCount":52344,"followCount":365,"weibocount":966},{"uid":"1833141401","ucode":"geinchina","uname":"GE中国","domainid":1,"verify":false,"image":"http://tp2.sinaimg.cn/1833141401/50/5662445273/1","addr":"上海 浦东新区","sex":"m","userType":"p","fansCount":80757,"followCount":980,"weibocount":1977}]} | |

#### 数据含义

##### 微博数据

|  |
| --- |
| **{"mid":"3695391205256822" ——微博ID**  **,"uid":"1728875812" —— 账号ID**  **,"ucode":"haierexpo" —— 账号编码（个性域名或者UID）**  **,"uname":"海尔" —— 账号名称**  **,"domain":"weibo.com" ——域名**  **,"domainName":"新浪微博" ——信息源名称**  **,"postDate":"20140403214118" ——发布日期**  **,"content":"这可真的是我们家超早期的产品之一~妹纸去看神马电影吖？带我带我~\\(≧▽≦)/~已粉！！！ ","comtCount":"3","relayCount":"0","url":"http://weibo.com/1980245395/AE1sFhaPB","relayContent":"@董小翠爱吃菜\_ \n #发真人照以示清白#@海尔 超级古老的洗衣机有木有 我们去看电影吧 |发真人照以示清白 我在这里:|工商银行数据中心停车场\n " ——微博内容**  **,"verify":true ——账号是否认证**  **,"imageUrl":**[**http://tp1.sinaimg.cn/1728875812/180/5669966281/1**](http://tp1.sinaimg.cn/1728875812/180/5669966281/1) **——账号头像**  **,"transferUrl":"" ——转发原微博url**  **}** |

##### 账号数据

|  |
| --- |
| **{"uid":"2083844833" ——账号id**  **,"ucode":"sohuexpress" ——账号编码（个性域名或者uid）**  **,"uname":"搜狐新闻客户端" ——账号名称（screenname）**  **,"domainid":1 ——信息源id（讯库中：1-新浪微博；2-腾讯微博；3-网易微博；4-搜狐微博；5-人民微博）**  **,"verify":false ——是否认证**  **,"image":**[**http://tp2.sinaimg.cn/2083844833/50/22868821480/0**](http://tp2.sinaimg.cn/2083844833/50/22868821480/0) **——头像**  **,"addr":"北京 海淀区" ——地域**  **,"sex":"f" ——性别（m-男，f-女，n-未知）**  **,"userType":"p" ——类型（p-个人用户；c-机构用户）**  **,"fansCount":796133 ——粉丝数量**  **,"followCount":683 ——关注数量**  **,"weibocount":19487 ——微博数量**  **}** |

### 2、抓取服务(提供给内部使用)

WebService地址：待定

#### 接口定义

|  |
| --- |
| /\*\*  \* 提交微博账号抓取任务  \* **@param** domainid 微博ID：1=新浪微博，2=腾讯微博；5=人民微博；  \* **@param** homeUrl 微博账号主页地址  \* **@param** crawltype 类型：1=关注；2=粉丝；3=微博；  \* **@param** token token  \* **@return**  \*/  String submitUserTask(**int** domainid,String homeUrl,**int** crawltype,String token);  /\*\*  \* 获取账号抓取任务的结果  \* **@param** token  \* **@return**  \*/  String getUserTaskResult(String token);  /\*\*  \* 获取微博来源  \* **@param** appsource  \* **@return**  \*/  **long** addWeiboAppSource(String appsource);  /\*\*  \* 根据id获取名称  \* **@param** sourceid  \* **@return**  \*/  String getWeiboAppSource(**long** sourceid); |

返回结果的json串参考[解析服务](#_1、解析服务)。

#### 接口说明

目前只提供了新浪微博、人民微博的部分网页爬虫功能，腾讯微博使用API即可满足要求。

任务的调度需要外部提供，抓取服务不提供调度功能。

● 新浪微博：提供如下抓取服务，

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 抓取服务 | 需要登录 | 其他限制 | 对应接口 | 备注 |
| 微博账号首页微博抓取 | 不需要 | ◆尽量分散抓取IP；  ◆控制频率 | 爬虫：  解析：parseHomePage | 比如：<http://weibo.com/lifeweek> |
| 微博搜索 | 不需要 | ◆尽量分散抓取IP；  ◆控制频率 | 爬虫：  解析：parseSearchPage | 搜索关键词 |
| 微博账号粉丝列表抓取 | 需要 | ◆尽量分散抓取IP；  ◆控制频率；  ◆维护多个微博账号的登录Cookie；  ◆最多只能抓取10页数据； | 爬虫：  解析：parseFansPage | 比如：  <http://weibo.com/1728875812/fans>  url=<http://weibo.com/uid/fans> |

●人民微博：提供如下抓取服务，

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 抓取服务 | 需要登录 | 其他限制 | 对应接口 | 备注 |
| 微博搜索 | 需要 | ◆尽量分散抓取IP；  ◆控制频率 | 爬虫：  解析：parseSearchPage | 搜索关键词  <http://t.people.com.cn/searchV3.action?searchInput=%BA%A3%B6%FB> |

●

# 五、维护设计