和讯微博“随便看看”栏目

——调研及抓取方案设计

# 一、调研内容：

## 1.栏目更新频率

“随便看看”栏目中，微博的展现数量为固定的17页，每页30条微博的形式，足够展示一小时以内发送的微博。但是它的更新频率却灵活多变。

挑一个微博发表较为高峰的时段，晚间20:51-21:57一个多小时的时间段中，跟踪“随便看看”栏目中微博的更新状况。发现在这一小时多的时间里，共出现6次页面刷新。但每次间隔都不同，这七次的时间间隔及出现的新的微博条数分别为：8min（164条），25min（214条），8min（198条），8min（95）条，1min（62条），13min（56条）。并且每次刷新过后，第一条微博不一定是一分钟以前发送的最新微博，有可能已经是10+min以前发送过的微博。在这一个多小时内刷新的微博条数共789条。

从目前的情况来看，17页数据也就是510条微博，在仅开启一个线程的情况下只需不到10s的时间就可以抓完。所以，设定5 min为一次抓取频率较为合适，既保证了数据的实时性，又不会占用太多资源。

## 2.微博的数据特点

1. 实时性

栏目中均显示的是1小时内新产生的微博，所以保证了数据的实时性，省去了自己一一去更新每个用户的微博的步骤，可直接从这个集中的平台中就可以获取到更新内容。

但是经过调研发现，其更新的内容并不是最新的，跟每个用户微博主页的的微博更新比还是有一些延迟，比如，“随便看看”栏目中刚有一次刷新，产生了一部分新的数据，但在这部分中，可能没有用户当前产生的较新数据，而在下一次栏目的刷新过后可以找到，所以有一定延迟，但是并不影响数据的完整性。

1. 每条微博的信息量

经过调研，该栏目中，每条微博的信息量都是非常充足的。因为从其页面的源代码中可以得到：该条微博的ID，该条微博所有者的用户名及其userid，如下图面源码所示：

<div id="articelItem\_20432023" class="nr\_con">

<div id="nr\_conL\_20432023" class="nr\_conL" articleusername="金凌网络" articleuserid="20066984">

…

</div>

其中最重要的一个数据userid是从别的地方很难获得的。因为之前做抓取工作时，用户的id来源于其url，而有些用户的url中是字母形式，或字母数字混合形式，判断其应该有相应的纯数字串形式，但却未找到相对应的方法。但在该栏目中却发现了这一信息。曾在一个字母id的用户中做了这样一个实验，该用户本来的微博主页url中id为字母形式，但是将其id改为该栏目中显示的userid链接后仍为其主页。所以证明，每个用户确实有一条纯数字id形式。

并且，由以上源代码可得，抓取工作只需抓取栏目中的页面，无需进入该用户的微博主页便可得到该条微博的基本信息，故抓取效率极高。这是在新浪微博及其他微博中很少见到的有利条件，调查其他微博的滚动更新平台，均没有像这样每条微博数据的源码中包含信息量极多的。

1. 微博用户的覆盖面

首先，该栏目中，是按照微博的发表时间进行更新，而没有名人热度，转发热度等优先级的处理。所以，任何用户发表的微博都会出现在这一区域中。

根据之前的调查，发现每条微博之间根据时间前后关系ID是呈递增状态的，即新发表的微博ID是原来微博的ID+1。但是在后续的跟踪调查中发现，该栏目的微博ID并不是完全符合这一规则，如有出现ID+2或者ID+3的情况，但是发生频率并不固定。有可能某一页一次都没有发生，另一页却发生了多次。因为最初发现ID相连关系后，认为每一条微博都应该出现在该平台上，但是现在猜测会不会有微博没有出现在平台上。但是，跟踪了某些平台上用户的微博，发现除了实时性中提到的延迟，所有发表过的微博都会出现在该栏目中。故有可能那些没有出现过的微博ID不存在或由于系统问题未被使用。

关于这一问题，我也做了亲身体验，我新申请了一个用户，然后用百度分享按钮发表了一条微博，该微博在大约半小时后出现在“随便看看”栏目中，所以，从此可看出，“随便看看”栏目囊括了所有正常用户发表的微博信息。

整体而言，该栏目对用户的覆盖面还是非常广泛的，信息来源比较真实，基本包含了所有和讯用户发表的微博信息。

1. 微博来源

根据对该栏目的调查发现，微博的来源不仅仅是用户通过和讯微博这一条路经发表的，除此之外还有：博客，WAP，短信，享拍微博通，百度分享按钮，网页分享。

其中博客即和讯博客，会直接转到微博平台上；WAP和短信都是用户通过手机发表微博的途径；享拍微博通是一款基于API的多个微博平台及SNS同步更新工具。通过享拍微博通，用户可以向已注册的各个微博平台发送消息和查看各个平台接收的信息。其中包括：新浪微博、腾讯微博、搜狐微博、网易微博、嘀咕、人民网微博、人间网、开心网、人人网等十多个微博及社交网络平台。百度分享按钮是因为和讯本身就载入了百度分享的代码，在和讯的新闻等页面上可以直接方向相关文章到微博上。自己也亲身体验了这个功能。

从这方面的调研也发现，和讯微博的信息量来源较广，不一定全是用户通过微博客户端去发表的内容，所以采集到微博的同时也采集到了用户从其他地方得到的有用信息，价值更高。

## 3.和讯独有的微博产品——新飞狐



新飞狐客户端软件是国内首款大型高手互动型金融云终端，2012年4月23日上

线，由和讯历时一年半打造而成。 新飞狐同时也是国内首个金融信息开放平台，通

过开放策略吸引行业众多的产品与应用开发企业入驻，为用户提供个性化与海量的选

择。

新飞狐是一套基于“六度空间”理论实名制的社会化网络（SNS）证券投资分析和

价值评估免费软件。也是一个集聚互动社区、个人门户、即时通讯和组件平台为一体、

以中小投资者为目标用户群体的web3.0时代的SNS商务平台。软件包含有网络聊天

（IM）、投资分享、高手直播、微博、网络社区等功能。它的基本功能是通过对信息

的实时揭示（包括行情信息和资讯信息），以及对市场信息数据的统计，按照一定的

分析模型来给出数（报表）、形（指标图形）、文（资讯链接）等功能与服务。并开发

出如：技术分析、基本面分析、资讯汇集、智能选股、自动选股、联动委托交易等等

实用性功能。

和讯新飞狐主要给用户提供三种服务：首先是点评分析服务，通过相关股评专家的综合分析，给予股民更直接的投资建议和计划，尤其是社区、微博的嵌入和和讯IM应用，可以更加有效和强化这种服务。

# 二、抓取方案设计

根据之前的调研结果可知，我们要实现的较为简洁的高效率抓取工作。因为他的抓取是实时性的更新操作，所以可以使用原抓取工作中的更新模块。其基本原理是：通过比对微博ID来确定该微博是否为更新的微博；通过比对该页第一条微博是否为更新的微博来确定是否进行抓取。主要步骤为：

1. 首次抓取“随便看看”栏目1-17页的全部微博信息，并记录当前最新一条微博的ID；
2. 每5min进行一次抓取栏目第一页页面的操作，抽取当前最新微博ID；
3. 判断该ID是否与之前记录的最新一条微博ID相同；
4. 若相同则说明该5min中内栏目没有更新，转（2）；
5. 若不相同则说明栏目中的微博进行了更新，存储当前最新微博ID后进行抓取操作；
6. 从第一页开始抓取，通过比对每一页的第一条ID是否大于原最新微博ID来判断该页是否需要抓取，抓取结束后继续转（2）

流程图如下：

**首次抓取1-17页微博信息**

**并存储当前最新微博ID1**

**抓取第一页信息，抽取当前最新微博ID2**

**ID1=ID2**

**从第一页开始抓取，通过比对每一页的第一条ID是否大于原最新微博ID1来判断该页是否需要抓取**

**没5min做一次**

**Y**

N