采集分系统与可视化分系统

数据格式定义

2017/3/6 V1.0

参与单位：电子所、成电

# 1 数据格式定义

## 1.1 原始采集数据特点分析

（1）不能获取推文的所有字段

采集程序所使用的stream api 与rest api原始数据字段不同，甚至同为rest api，search/tweets api与statuses/user\_timeline api的原始数据字段也不同（search/tweets原始字段中有metadata，而statuses/user\_timeline没有）。同时，官方文档没有给出全部字段，如：在media字段中，source\_user\_id字段没有给出，但是采集过程中有些推文出现。

（2）用户需求

* 可视化分系统实时展示采集结果
* 保留采集到的原始数据所有字段

## 1.2 查询接口

采集分系统将采集的原始数据导入MongoDB数据库，ElasticSearch检索引擎对MongoDB数据库中推文及账户表指定字段进行同步，可视化分系统通过ElasticSearch检索引擎获取相关数据信息。ElasticSearch及可视化分系统所使用的推文及账户指定字段格式见（1.3实时采集及ES同步接口），MongoDB数据库存储的原始字段结构如下：

（1）推文表字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 字段说明 | 类型 | 备注 |
| contributors | 对这条推文有贡献的用户 | dic | 不常用 |
| truncated | 是否截断（是否超过140个字） | bool |  |
| text | 推文内容 | string |  |
| is\_quote\_status | 这条推文是否为引用推文 | bool |  |
| in\_reply\_to\_status\_id | 若为回复推文，则此字段为回复推文的id | int64 | 回复推文指用户转推文时，加上了自己的评论 |
| id | 推文ID | int64 |  |
| favorite\_count | 推文被点赞数 | int32 |  |
| entities | 对推文的解析（这个对分析推文十分有帮助） | dic | 包括标签信息、媒体信息、符号信息、链接，推文提及的人 |
| quoted\_status\_id | 引用推文的ID | int64 |  |
| retweeted | API授权用户是否转发了本条推文 | bool |  |
| coordinates | 地理信息，经纬度 | dic |  |
| timestamp\_ms | 推文创建时间UTC（毫秒）（仅出现在statuses/filter） | string |  |
| quoted\_status | 引用的推文信息 | dic | 一条推文 |
| source | 发布本推文的客户端 | string |  |
| in\_reply\_to\_screen\_name | 若为回复推文，则此字段为回复对象的screen\_name | string |  |
| id\_str | 推文ID字符串 | string |  |
| display\_text\_range | 表示推文的长度（从0个字节到第几个字节）（仅出现在statuses/filter） | list\_int32 |  |
| retweet\_count | 推文被转发的数目 | int32 |  |
| in\_reply\_to\_user\_id | 若为回复推文，则此字段为回复对象的用户id | int64 |  |
| favorited | API授权用户是否点赞 | bool | 指的是采集这条推文的API，是否点赞 |
| user | 推文的用户信息 | dic |  |
| geo | 地理位置信息，若无，则用'coordinates': 地理信息，经纬度 | dic |  |
| in\_reply\_to\_user\_id\_str | 若为回复推文，则此字段为回复对象的用户id字符串 | string |  |
| possibly\_sensitive | 若此字段为真，则表示链接中可能有更多的消息 | bool |  |
| lang | 使用的语言 | string |  |
| extended\_tweet | 推文的扩展信息（仅出现在statuses/filter） | dic | 拥有完整的推文（感觉比较重要） |
| created\_at | 推文创建时间 | string | 原始 |
| quoted\_status\_id\_str | 引用推文的ID字符串 | string |  |
| filter\_level | 这表示过滤水平（low,medium,none（仅出现在statuses/filter） | string |  |
| in\_reply\_to\_status\_id\_str | 若为回复推文，则此字段为回复推文的id字符串 | string |  |
| place | 推文相关地点 | dic | 不一定是这条推文的来源地 |
| retweeted\_status | 转推的推文 | dic |  |
| current\_user\_retweet | 最近用户的转推 | dic | 包括转推文的ID |
| metadata | 在search/tweets中出现 | dic |  |
|  |  |  |  |
| created\_at\_bj | 北京时区时间戳（加8小时） | long | 增加字段 |
| captured\_time | 程序采集推文时间 | integer |
| classification | 发这条推文的人的分类 | string |
| useful | 是否有用(默认true) | bool |

（2）账户表字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 字段说明 | 类型 | 备注 |
| id | 账户ID | Int64 | account\_id |
| follow\_request\_sent | 如果该用户为私密账号（protected为true），是否已经向其发送关注请求 | bool |  |
| has\_extended\_profile | 是否有扩展用户信息 | bool |  |
| profile\_use\_background\_image | 是否使用上传图片作为背景 | bool |  |
| default\_profile\_image | 未上传头像，使用默认头像 | bool |  |
| profile\_background\_image\_url\_https | 背景图片https链接 | string |  |
| verified | 是否被认证 | bool |  |
| translator\_type | 翻译者类型 | string | 不确定是干什么的，已知有regular |
| profile\_text\_color | 文本内容颜色 | string |  |
| profile\_image\_url\_https | 头像图片https链接 | string |  |
| profile\_sidebar\_fill\_color | 侧边栏填充颜色 | string |  |
| entities | 用户提供的描述信息 | Object(字典型) |  |
| followers\_count | 关注者数目 | int32 |  |
| profile\_sidebar\_border\_color | 侧边栏边界颜色 | string |  |
| id\_str | 账号ID字符串 | string |  |
| profile\_background\_color | 背景颜色 | string |  |
| listed\_count | 有该账号的公共列表的数目 | int32 | 越多，表示越被人关注 |
| status | 最近所发推文 | Object（字典型） | 可以认为是推文 |
| is\_translation\_enabled | 是否可翻译 | bool |  |
| utc\_offset | 与GMT/UTC的时间偏移量（秒数） | int32 |  |
| statuses\_count | 发布推文数量 | int32 |  |
| description | 账号描述 | string |  |
| friends\_count | 该账号关注的用户数目 | int32 |  |
| location | 地理位置 | string | 用户自己填的 |
| profile\_link\_color | 链接颜色 | string |  |
| profile\_image\_url | 头像图片链接 | string |  |
| following | API授权用户是否关注 | bool | 指的是用来采集账号的API是否关注了这个账号 |
| geo\_enabled | 是否允许对其发送的推文进行位置标注 | bool |  |
| profile\_banner\_url | 横幅图片链接 | string |  |
| profile\_background\_image\_url | 背景图片http链接 | string |  |
| screen\_name | 账号唯一辨识名 | string |  |
| lang | 使用的语言 | string |  |
| profile\_background\_tile | 背景图片显示时是否进行拼接 | bool |  |
| favourites\_count | 账号点赞推文的数量 | int32 |  |
| name | 昵称 | string |  |
| notifications | 当该用户发送推文时，进行消息提醒 | bool |  |
| url | 用户信息（profile）的url链接 | string |  |
| created\_at | 账号创建时间 | string |  |
| contributors\_enabled | 是否授权其他用户以本账号名义发送推文 | bool |  |
| time\_zone | 时区 | string | 用户自己标记的 |
| protected | 是否为私密账号，是否对其所发推文进行保护 | bool | 如果这个字段为true，则无法采集推文 |
| default\_profile | 用户信息使用默认主题设置 | bool |  |
| is\_translation | 是否是twitter翻译社区的一分子 | bool |  |
|  |  |  |  |
| created\_at\_bj | 北京时区时间戳（加8小时） | long | 增加字段 |
| alive | 是否存活 | bool |
| account\_update\_time | 账号更新时间 | long |
| classification | 账号类别 | string |
| useful | 是否有用 | bool |
| avatar\_captured | 是否抓取过头像 | bool |
| banner\_captured | 是否抓取过背景 |  |
| captured\_followers\_count | 已经抓取的粉丝数量 | int32 |
| captured\_friends\_count | 已经抓取的朋友数量 | int32 |
| captured\_tweet\_count | 已采集的历史推文的数量 | Int32 |

（3）粉丝和关注列表（friend\_ids、followers\_ids）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 字段说明 | 类型 | 备注 |
| id | 账户ID | int64 | account\_id |
| screen\_name | 账号唯一辨识名 | string |  |
| friends\_ids | 关注账户 | list\_long |  |
| followers\_ids | 粉丝账户 | list\_long |  |

## 1.3 实时采集及ES同步接口

采集分系统将处理后的实时信息数据按如下表结构传输：

（1）推文使用字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 字段说明 | 类型 | 备注 |
| text | 推文内容 | string |  |
| in\_reply\_to\_status\_id | 若为回复推文，则此字段为回复推文的id | long | 回复推文指用户转推文时，加上了自己的评论 |
| id | 推文ID | long |  |
| favorite\_count | 推文被点赞数 | int32 |  |
| retweeted | API授权用户是否转发了本条推文 | bool |  |
| source | 发布本推文的客户端 | string |  |
| in\_reply\_to\_screen\_name | 若为回复推文，则此字段为回复对象的screen\_name | string |  |
| retweet\_count | 推文被转发的数目 | int32 |  |
| favorited | API授权用户是否点赞 | bool | 指的是采集这条推文的API，是否点赞 |
|  |  |  |  |
| quoted\_status:{  text  } | 原始推文 | string | 嵌入字段 |
| retweeted\_status:{  text  } | 原始推文 | string |
| retweeted\_status:{  quoted\_status:{  text  }  } | 原始推文 | string |
| user: {  id  } | 账户ID | long |
| user: {  screen\_name  } | 账户名 | string |
| user: {  name  } | 昵称 | string |
| created\_at\_bj | 北京时区时间戳（加8小时） | long | 增加的字段 |
| captured\_time | 程序采集推文时间 | long |
| classification | 发这条推文的人的分类 | string |
| useful | 是否有用(默认true) | bool |

（注：原始推文3个只会取唯一一个，如果quoted\_status:{text}存在则取这个，如果不存在，则取retweeted\_status:{quoted\_status:{text}},如果都不存在，取retweeted\_status:{text}）。

（2）账户使用字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 字段说明 | 类型 | 备注 |
| id | 账户ID | int64 | 以前是account\_id |
| followers\_count | 关注者数目 | int32 |  |
| listed\_count | 有该账号的公共列表的数目 | int32 | 越多，表示越被人关注 |
| statuses\_count | 发布推文数量 | int32 |  |
| description | 账号描述 | string |  |
| friends\_count | 该账号关注的用户数目 | int32 |  |
| location | 地理位置 | string | 用户自己填的 |
| screen\_name | 账号唯一辨识名 | string |  |
| favourites\_count | 账号点赞推文的数量 | int32 |  |
| name | 昵称 | string |  |
| created\_at\_bj | 北京时区时间戳（加8小时） | long | 增加字段 |
| alive | 是否存活 | bool |
| account\_update\_time | 账号更新时间 | long |
| classification | 账号类别 | string |
| useful | 是否有用 | bool |
| avatar\_captured | 是否抓取过头像 | bool |