数据采集程序功能

杨杰

2017.3.9

一：用户信息采集

执行文件：UserInfoCapture.bat。 配置文件：UserInfo.json。

1、输入参数：

1）userHost：连接mongodb数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

2）userDatabase ：连接mongodb的database名。类型：string，默认：“test”；

3）userCollection ：连接mongodb的collection名。类型：string, 默认：“User”；

4）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

5）screenNameList ：推特用户名字。类型：stringList, 默认：null,

实例：["bbc", "cnn"]；

6）accountIdList ： 推特用户ID。类型：intList，默认：null,

实例：[12,4567,32]。

2、输出结果：

将用户数据存入userDatabase的userCollection中。

3、执行逻辑：

如果用户输入screenNameList或者accountIdList，则只会采集这部分用户，将数据存入数据库，采完停止程序。如果上述两个参数为空，则从user数据库中没有采集过的用户,采集完整的信息。当数据库中的用户采集完毕后，停止程序。

4、注意事项：

1）每次执行前，请确认各个参数的名称。

2）参数格式为json,请严格输入json格式，如：screenNameList没有参数时是null；必须是双引号，不能使用单引号；记得每行参数后面有一个逗号；最后一个参数后面不能有逗号。

二：用户朋友ID列表采集

执行文件：UserFriendsCapture.bat。 配置文件：UserFriends.json。

1、输入参数：

1）userHost：连接mongodb用户数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

2）userDatabase ：连接mongodb的用户database名。类型：string，默认：“test”；

3）userCollection ：连接mongodb的用户collection名。类型：string, 默认：“User”；

4）userFriendsHost：连接mongodb朋友列表数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

5）userFriendsDatabase ：连接mongodb的朋友列表database名。类型：string，默认：“test”；

6）userFriendsCollection ：连接mongodb的朋友列表collection名。类型：string, 默认：“FUser”；

7）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

8）screenNameList ：推特用户名字。类型：stringList, 默认：null,

实例：["bbc", "cnn"]；

9）accountIdList ： 推特用户ID。类型：intList，默认：null,

实例：[12,4567,32]。

2、输出结果：

将朋友列表数据存入userFriendsDatabase的userFriendsCollection中。

3、执行逻辑：

如果用户输入screenNameList或者accountIdList，则只会采集这部分用户的朋友ID列表，将数据存入userFriendsCollection中，采完停止程序。如果上述两个参数都为空，则从userCollection中选取没有采集过朋友列表的用户id,采集朋友id列表,并将新的ID插入userCollection中，在userFriendsCollection中插入IDList。当数据库中的用户采集完毕后，停止程序(基本上不会停止程序，因为不断用新的用户ID插入userCollection中)。

三：用户粉丝ID列表采集

执行文件：UserFollowersCapture.bat。 配置文件：UserFollowers.json。

1、输入参数：

1）userHost：连接mongodb用户数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

2）userDatabase ：连接mongodb的用户database名。类型：string，默认：“test”；

3）userCollection ：连接mongodb的用户collection名。类型：string, 默认：“User”；

4）userFollowersHost：连接mongodb粉丝列表数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

5）userFollowersDatabase ：连接mongodb的粉丝列表database名。类型：string，默认：“test”；

6）userFollowersCollection ：连接mongodb的粉丝列表collection名。类型：string, 默认：“FUser”；

7）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

8）screenNameList ：推特用户名字。类型：stringList, 默认：null,

实例：["bbc", "cnn"]；

9）accountIdList ： 推特用户ID。类型：intList，默认：null,

实例：[12,4567,32]。

2、输出结果：

将粉丝列表数据存入userFollowersDatabase的userFollowersCollection中。

3、执行逻辑：

如果用户输入screenNameList或者accountIdList，则只会采集这部分用户的粉丝ID列表，将数据存入userFollowersCollection中，采完停止程序。如果上述两个参数都为空，则从userCollection中选取没有采集过粉丝列表的用户id,采集粉丝id列表,并将新的ID插入userCollection中，在userFollowersCollection中插入IDList。当数据库中的用户采集完毕后，停止程序(基本上不会停止程序，因为不断用新的用户ID插入userCollection中)。

四：用户头像或横幅采集：

执行文件：UserAvatarCapture.bat。 配置文件：UserAvatar.json。

1、输入参数：

1）userHost：连接mongodb用户数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

2）userDatabase ：连接mongodb的用户database名。类型：string，默认：“test”；

3）userCollection ：连接mongodb的用户collection名。类型：string, 默认：“User”；

4）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

5）hbaseThriftHost：用于连接hbase的host(注：hbase用于存储图片)。类型：string, 默认："192.168.178.212"；

6）hbaseThriftPort ：用于连接hbase的port。类型：int,默认：9090。

7）hbaseTableName ：用于连接hbase的表名。类型：string ,默认："linuxtest"；

8）picture\_type ：采集图片的类型，只有两种——avatar和banner。类型：string， 默认：avatar。

2、输出结果：

将图片存入hbase的hbaseTableName中。

3、执行逻辑：

从userCollection中取出图片链接，获取图片后发往hbase的hbaseTableName中。

五：推特特定用户实时推文采集：

执行文件：UserRealtimeTweetCapture.bat。 配置文件：UserRealtimeTweet.json。

1、输入参数：

1）userHost：连接mongodb用户数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

2）userDatabase ：连接mongodb的用户database名。类型：string，默认：“test”；

3）userCollection ：连接mongodb的用户collection名。类型：string, 默认：“ClassifiedUser”；

4）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

5）tweetHost：连接mongodb推文数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

6）tweetDatabase：连接mongodb的推文database名。类型：string，默认：“test”；

7）tweetCollection：连接mongodb的推文collection名。类型：string, 默认：“ClassifiedUserTweet”。

2、输出结果：

将采集的推文存入tweetCollection中。

3、执行逻辑：

从userCollection数据库中取出所有的用户，实时监控他们，每当他们发出一条数据，就将这条推文存入tweetCollection中。

六：根据关键字实时推文采集：

执行文件：FilterRealtimeTweetCapture.bat。 配置文件：FilterRealtimeTweet.json。

1. 输入参数：

1）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

2）tweetHost：连接mongodb推文数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

3）tweetDatabase：连接mongodb的推文database名。类型：string，默认：“test”；

4）tweetCollection：连接mongodb的推文collection名。类型：string, 默认：“FilterTweet”；

5）keywords：关键字。类型：string，实例："news, breaking news"；

6）classification：给推文分类。类型：string，实例："news"。

2、输出结果：

将采集的推文存入tweetCollection中，并发往前端实时显示。

3、执行逻辑：

根据关键字采集推文，推文中包含这些关键字，不同的关键字用逗号隔开，如果用空格隔开，那么采集的推文将包括所有的单词，如"news, breaking news"，推文中包含news或者同时包含breaking与news的推文才会采集。将这些推文存入tweetCollection中，并发往前端实时显示。

七：用户历史推文采集：

执行文件：HistoryTweetCapture.bat。 配置文件：HistoryTweet.json。

1、输入参数：

1）userHost：连接mongodb用户数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

2）userDatabase ：连接mongodb的用户database名。类型：string，默认：“test”；

3）userCollection ：连接mongodb的用户collection名。类型：string, 默认：“User”；

4）proxyList ：代理列表。类型：stringList , 默认：["192.168.178.149:9090","192.168.178.150:9090","192.168.178.151:9090"]；

5）tweetHost：连接mongodb推文数据库。默认："mongodb://192.168.178.218:27017,192.168.178.220:27017/?replicaSet=shard1"；

6）tweetDatabase：连接mongodb的推文database名。类型：string，默认：“test”；

7）tweetCollection：连接mongodb的推文collection名。类型：string, 默认：“HistoryTweet”。

8）screenNameList ：推特用户名字。类型：stringList, 默认：null,

实例：["bbc", "cnn"]；

9）accountIdList ： 推特用户ID。类型：intList，默认：null,

实例：[12,4567,32]。

2、输出结果：

将采集的推文存入tweetCollection中。

3、执行逻辑：

如果用户输入screenNameList或者accountIdList，则只会采集这部分用户的历史推文，将数据存入tweetCollection中，采完停止程序。如果上述两个参数都为空，则从userCollection中选取没有采集过历史推文的用户id,采集历史推文,采集所有用户后，结束程序。

。