Twitter Social Graph Analysis

Sajib Sen
Department of Computer Science
University of Memphis
Memphis,TN
Email: ssen4@memphis.edu

I. Introduction

Twitter, a microblogging service, serves as a new medium for social interactions and networking. Such as, news about flood in some area , students protest in school etc. Although there are couple of social networking services already exists, but Twitter has some definite style which evolved as well-defined markup cultures. Although the users in Twitter follow others and/or are followed, unlike Facebook, MySpace etc. the relationship between users are not reciprocated. Being a follower on Twitter means that the user can get all the news (tweets) from the user being followed. Throughout the years the common practice of using Twitter evolved a markup culture. For example retweet represent as RT, to address a user '' has been used followed by a user identifier address, and '#' followed by a word represents hashtag. Furthermore, to provide brevity in expressions for the user Twitter has allowed character limits per posting combined with these markup vocubulary.

The goal of this project is to analyze and study the topological characteristics of Twitter, such as how people are connected to each other, who are the most influential people, how people are dispersed across demographic area etc. To complete the goal I chose some dataset of Twitter graph[1] and ran in Apache Pyspark big data framework. To categorize the user based on some characteristics ranking algorithms as well as build in libraries of PySpark has been used.

This report is organized as follows. In section II, I discussed some literature work related to my work. Section III describes about the dataset I used. In section IV, a short overview of environment setup was presented. Moreover, in section V, I provided the result and analysis of the dataset. And, finally in section VI, I discussed the difficulties I found in this project.

II. LITERATURE REVIEW

There are couple of literature works related to my work. In [2] authors conducted preliminary analysis of Twitter in 2007. Their dataset covers about 76, 000 users and 1, 000, 000 posts. They find user clusters based on user intention to topics by clique percolation methods.Krishnamurthy et al. also analyze the user characteristics by the relationships between the number of followers and that of followings [3].Zhao and Rosson qualitatively investigate the motivation of using Twitter [4]. Huberman et al. reports that the number of friends is actually smaller than the number of followers or followings [5]. Jansen conducts preliminary analysis of word-of-mouth branding in Twitter [6].

III. Dataset Description

In this project I used two publicly available twitter graph dataset [1] named as *Twitter_rv.net* and *Celebrities_profile.txt*. In the first dataset authors in [1] crawled the entire twitter site and extracted 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets. This dataset size in almost 26 GB. Unfortunately, due to Twitter's new terms of services, the dataset only has edge list with anonymous user id removing all further informations. The presented format for this dataset is:

FORMAT

* USER FOLLOWER
  - USER and FOLLOWER are represented by numeric ID (integer).
  - These numeric IDs are the same as numeric IDs Twitter managed

An example of top four data in the dataset can be as below

EXAMPLE

- 12 13
- 12 14
- 12 15
- 16 17

* Users 13,14, and 15 are followers of user 12.
* User 17 is a follower of user 16

The second dataset (i.e. *Celebrities_profile.txt*) has some definite properties. At first, all the users in this dataset has more than 10,000 followers and most importantly this dataset has all the properties of the original twitter social graph dataset. The format for this dataset is as follows:

FORMAT

numeric_id, verified, profile_sidebar_fill_color, profile_text_color, protected, location, profile_background_color, utc_offset, statuses_count, description, friends_count, profile_link_color, profile_image_url, notifications, profile_background_image_url, screen_name, profile_background_tile, favourites_count, name, url, created_at, time_zone, profile_sidebar_border_color, following, gender (infered by name)

- All fields except gender are returned by user method (users/show) of Twitter API
- 6500 unique celebrities

EXAMPLE

12 False EADEAA 333333 False San Francisco 8B542B -28800 4209 Creator, Chairman and co-founder of Twitter 574 9D582E http://s3.amazonaws.com/twitter_production/profile_images/54668082/Picture_2_normal.png False http://static.twitter.com/images/themes/theme8/bg.gif jack False 614 Jack Dorsey None Tue Mar 21 20:50:14 +0000 2006 Pacific Time (US & Canada) D9B17E False m

IV. Environment Setup

To analyze both dataset I used Apache spark in Jupyter Notebook. The environment setup was as follows:

- ubuntu 18.04
- python-3.6
- pyspark
- spark-2.4.1
- hadoop-2.7
- Java Version : 1.8.0_201 (Oracle Corporation)
- Spark Version: 1.6.0
- SPARK_WORKER_MEMORY=5G
- SPARK_DAEMON_MEMORY=10G
- SPARK_WORKER_CORES=5
- SPARK_EXECUTOR_CORES=3
- SPARK_EXECUTOR_MEMORY=2G
- SPARK_DRIVER_MEMORY=2G

V. Result and Analysis

The *Celebrities_profile.txt* data was analyzed with the help of *Twitter_rv.net*. As the first dataset has all the necessary information used in twitter, I chosed to rank the unique users based on some particular time zone (Fig 1), location (Fig 2), number of followers for each user (Fig 3) and at last segregation based on gender infered by name (Fig 4) in the dataset.

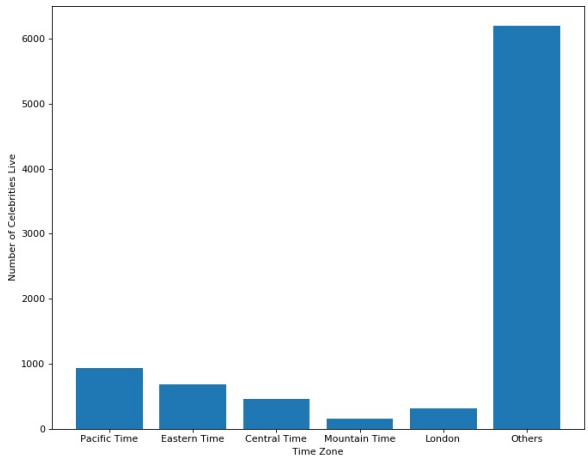

Fig. 1. Profiling celebrities based on time zone

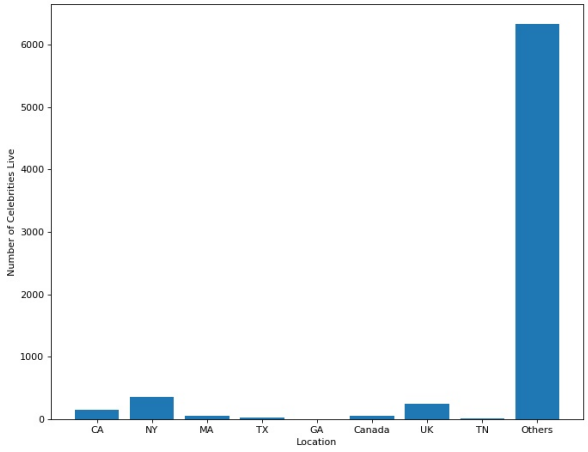

Fig. 2. Profiling celebrities based on location

numeric_id	name	number_of_followers
19058681	ashton kutcher	2615432
16409683	Britney Spears	2553668
15846407	Ellen DeGeneres	2391165
428333	CNN Breaking News	2142902
19397785	Oprah Winfrey	1908793
783214	Twitter	1854051
813286	Barack Obama	1788362
16190898	Ryan Seacrest	1739051
17461978	THE_REAL_SHAQ	1692121
25365536	Kim Kardashian	1619765

Fig. 3. Ranking celebrities based on number of followers

gender (infered by name)	count
m	2395
f	907
?	3197

Fig. 4. Profiling celebrities based on gender

VI. Conclusion

In this project, I analyzed a twitter social graph and presented the results based on some ranking for the users in the dataset. Throughout the project I faced a lots of difficulties to deal with big dataset for a single node machine. I found GraphFrame/GraphX libraries in Spark are very inefficient for graph related problem. On the otherhand, for a single node machine map-reduce using python found to be more efficient than the usual GraphFrame/GraphX libraries in Spark. However, parallelize the query processing also found to be much efficient and faster than non-parallelize algorithm.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 591–600.

[2] T. F. Java, X. Song and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *In Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007.

[3] P. G. B. Krishnamurthy and M. Arlitt, "A few chirps about twitter," in *In Proc. Of the 1st workshop on Online social networks*. ACM, 2008.

[4] D. Zhao and M. B. Rosson, "How and why people twitter: the role that micro-blogging plays in informal communication at work," in *In Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 2009.

[5] D. M. R. B. A. Huberman and F. Wu, "Social networks that matter: Twitter under the microscope," in *arXiv:0812.1045v1*, 2008.

[6] K. S. B. J. Jansen, M. Zhang and A. Chowdury, "Micro-blogging as online word of mouth branding," in *In Proc. of the 27th international conference extended abstracts on Human factors in computing systems*. ACM, 2009.