Analyzing Cell Phone Usage in China from the Microblogging Social Media Weibo

Kai Zhang[1,*]

[1] *Department of Chemical Engineering, Columbia University, New York, New York, 10027, USA*

(Dated: April 29, 2018)

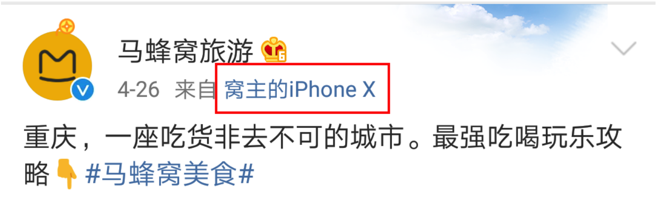
Sina Weibo, the Chinese equivalent of Twitter, is the most popular microblogging social media in China with more than two hundred million active users. Here we perform web page scraping with the Python 3 HTTP library Requests, to extract open user information from Weibo, such as location, gender, date of birth, and type of cell phone used. Besides basic demographic statistics, this work eventually will allow us to acquire useful business information, such as the market share of each types of cell phone and their geographic distribution [1].

1. INTRODUCTION

Sina **Weibo**, the Chinese equivalent of Twitter, is the most popular microblogging social media in China with more than two hundred million active users — nearly one fifth of the entire population. In addition to common user information such as location, gender and date of birth, Weibo also records the terminal of each post and opens it to the public under the user's permission (Fig. 1). Since more than 90% Weibo users are on mobile devices, this message provides a useful resource for analyzing cell phone usage among Chinese customers [2]. Here we perform web page scraping with the Python 3 HTTP library Requests, to extract open user information from Weibo. Besides basic demographic statistics, this work eventually will allow us to acquire useful business information, such as the market share of each types of cell phone and their geographic distribution. After collecting enough data, we can further apply standard machine learning classification methods such as random forest to build a recommender system of cell phone advertisement, based on customer information such as age, gender, location and social media activity.

2. DATA ACQUISITION

To visit the http address of Weibo pages, we first need a collection of users' id numbers, which are open to the public unless being hidden by the user on purpose. One automatic way to implement this task is to visit the top Weibo users (hubs) with a lot followers (fans) and extract the fans' id's from the Weibo html pages of hub users. Weibo opens a list of the first five thousand fans of each user to visitors, so we can roughly (for overlapping fans) obtain n_{id}=5000×n_{hub} user id's by visiting n_{hub} hubs. In our first round of test, we choose two hubs and obtained n_{id}=9874 user id's.



马蜂窝旅游

4-26 来自 窝主的iPhone X

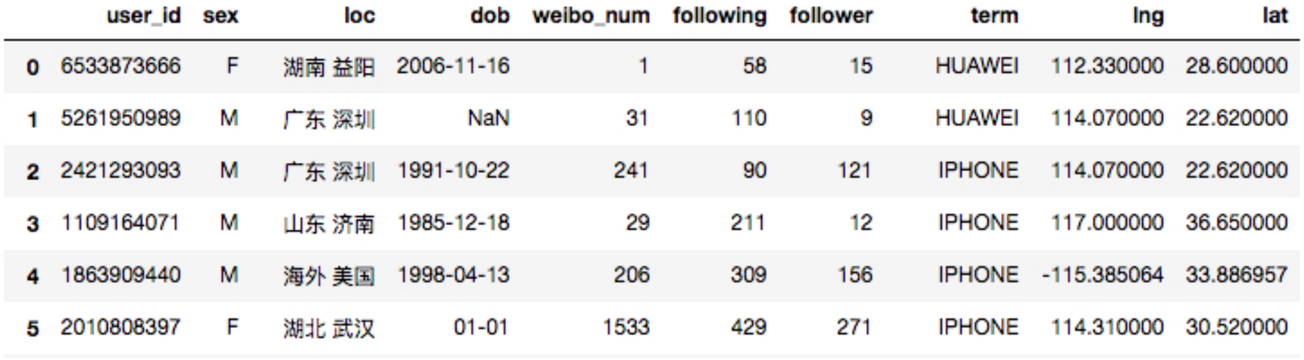
重庆，一座吃货非去不可的城市。最强吃喝玩乐攻略👇#马蜂窝美食#

FIG. 1: Weibo keeps track of the terminal of each post (red box).

We then scraped the needed information of all these n_{id}=9874 users, including user's id, gender, date of birth, location, number of posts, number of following, number of followers and type of terminal. To avoid over connection, we add a waiting time of a few minutes after every 50 visits. This puts a constraint on how many samples we can collect on each day (∼5000 per thread). For cell phone usage analysis, most terminal information is not accurate thus not useful. This happens when the terminal is on the web version, or from automatic commercial promotion (e.g. Sina movie), or just missing because many users never publish a post at all. After removing these redundant or missing samples, we manage to identify 2578 users (∼25%) with faithful cell phone information. Since one type of cell phone comes in a variety of nicknames given by the users, we need group them under a common name by matching the keywords in the nickname. For the sparse dataset we have in this first round of test, we did not distinguish between different phone models of the same manufacture.

Finally, we need to translate the location written in a string of Chinese characters into geographic data, i.e. latitudes and longitudes. The location information on Weibo is detailed up to the level of city or town for inland users but country for oversea users. Since the same location may repeat multiple times among different users, to reduce the number of search, we first create a dictionary (without repetition) of [“location”: “latitude” ,“longitude”] by scraping geographic information from Google map. The latitude and longitude of each user are then added as extra columns to the data frame by looking up the location dictionary. The resulted clean data frame has ten columns (Fig. 2).

[*] Electronic address: `kai.zhang.statmech@gmail.com`



	user_id	sex	loc	dob	weibo_num	following	follower	term	lng	lat
0	6533873666	F	湖南 益阳	2006-11-16	1	58	15	HUAWEI	112.330000	28.600000
1	5261950989	M	广东 深圳	NaN	31	110	9	HUAWEI	114.070000	22.620000
2	2421293093	M	广东 深圳	1991-10-22	241	90	121	IPHONE	114.070000	22.620000
3	1109164071	M	山东 济南	1985-12-18	29	211	12	IPHONE	117.000000	36.650000
4	1863909440	M	海外 美国	1998-04-13	206	309	156	IPHONE	-115.385064	33.886957
5	2010808397	F	湖北 武汉	01-01	1533	429	271	IPHONE	114.310000	30.520000

FIG. 2: An example of the clean dataset including user's id, gender, location (in Chinese), date of birth, number of posts, number of followings, number of followers, type of phone, longitude and latitude.

3. DATA ANALYSIS AND EARLY FINDINGS

Using the clean dataset we generated, we first analyze the market share and geographic distribution of top cell phone brands in China (Fig. 3). We find iPhone, Huawei, Oppo, Vivo and Xiaomi are the most popular five brands among Chinese customers. This agrees with most business analysis reports [3]. The geographic distribution suggests that most active cell phone Weibo users are concentrated in metropolitan areas such as Beijing, Shanghai and Hong Kong. To make the plot, border coordinates of China are used [4].

We next analyze the correlation between cell phone usage and some user features (Fig. 4). For the top five brands, gender histograms suggest that Oppo and Vivo enjoy a considerable preference among female users, by noticeably ∼30% more. Statistics of Weibo activity (number of posts, followings and followers) show that, in contrast to other phone users, iPhone users tend to have more followers that followings. These interesting indicators can eventually be harnessed to build a recommender system of cell phones using standard machine learning classification methods.

4. FUTURE WORK

The power of the analyzing method proposed in this work is currently constrained by the small dataset collected within a short period of time. In a long run, we expect to collect $5000 \times n_{\mathrm{thread}} \times n_{\mathrm{days}} \times 25\%$ faithful information of cell phone users with n_{thread} parallel threads. As the dataset increases, we can further make use of features like date of birth (age), which is not used now for most of them are either missing or unreliable (e.g. two-year old).

More data will also allow us to study geographic distribution on a finer length scale, with which we can answer questions about the popularity of each cell phone brand/model in each city or economic area. Manufactures can then adjust their market strategy based on such information.

Are iPhone users tend to follow iPhone users? With big dataset, it will also be interesting to investigate the network structure of Weibo and cell phone users.

[1] The code used in this work can be found at github https://github.com/statisticalmechanics/projectweibo.

[2] Oversea usage may also be analyzed, although Weibo is not widely used outside of China.

[3] Https://www.kantarworldpanel.com/global/News/Later-iPhone-X-Release-Hurts-Apple-Share.

[4] Http://gmt-china.org/datas/.

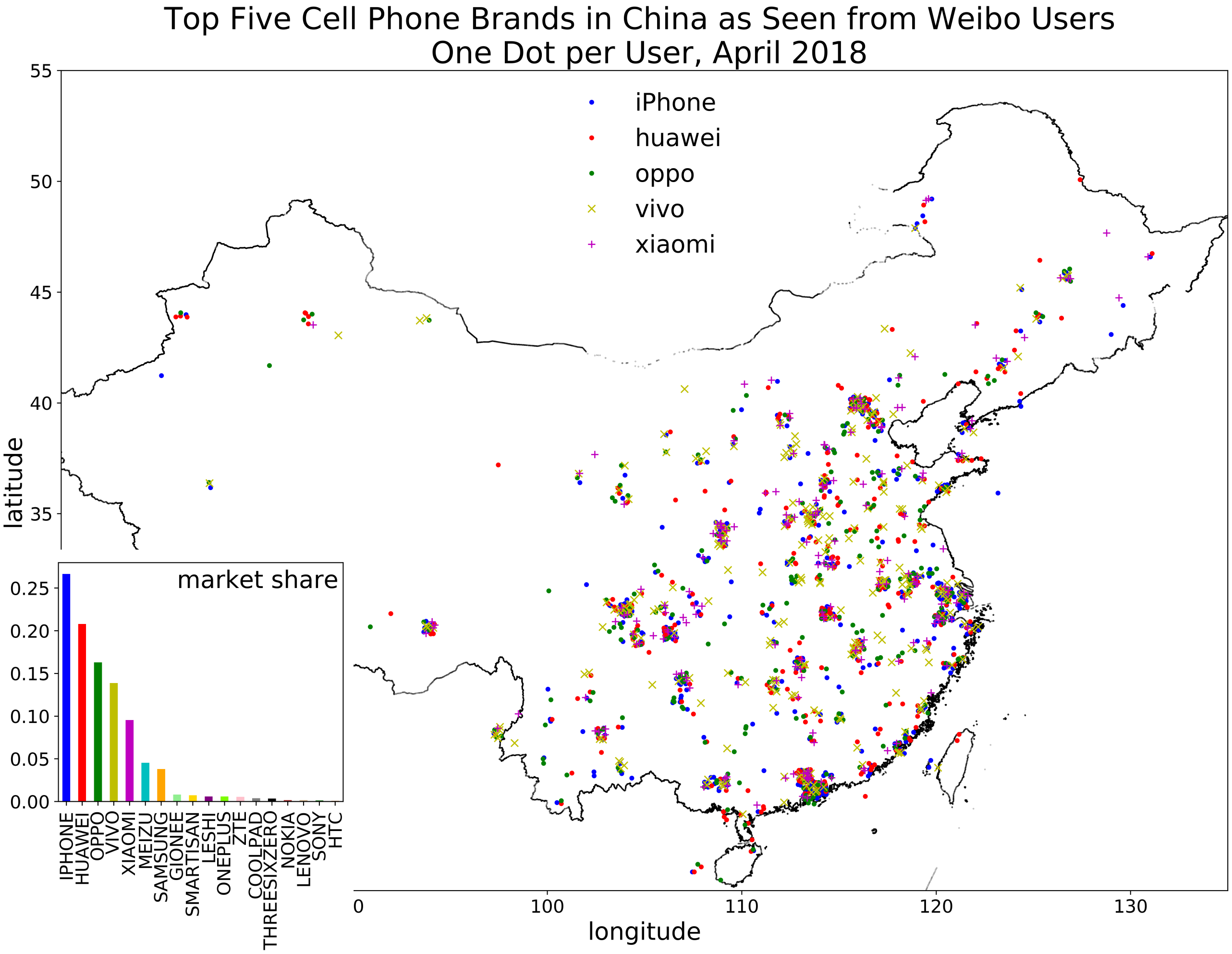


FIG. 3: The geographic distribution and histogram of top cell phone brands in China as seen from 2578 Weibo users. A random noise of 0.3 degree is added to the latitude and longitude of each dot for visualization convenience.

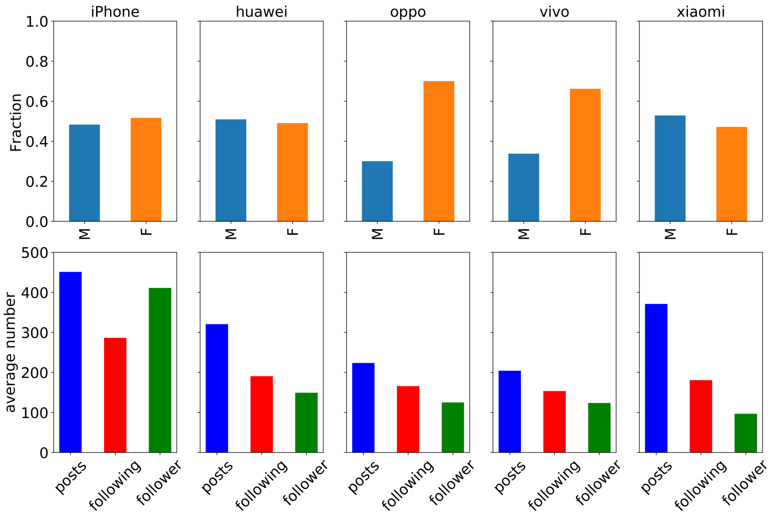


FIG. 4: User statistics of top five cell phone brands. Top: histogram of gender. Bottom: average number of posts, followings and followers.