Followee analysis on Douban and Weibo

CSE 140 Final Project

Jianhao Zhou

1. Summary of research questions and results

Q: Do Weibo users more tend to follow the crowd than Douban users do?

A: Yes.

Q: Is there any difference (in style) between Weibo and Douban celebrities?

A: Yes.

2. Motivation and background.

Weibo and Douban are two online communities in China. Weibo is a popular micro blog service, with more than 400 million users. Douban is a much smaller website with about 60 million users. The two sites are well-known in China, but with different styles. Weibo users represent the majority of Chinese Internet users: they are easily to be surprised, provoked and enraged. Sometimes they are deemed examples of blind herd mentality, interested in any topic as long as it's popular. For example, rumors spread quickly on Weibo. Douban is difficult to define. It was first a website for book, music and movie reviews. It attracted a lot of "arty youth". Later, it turned into a comprehensive interest community. The topics on Douban are mainly arts, literatures and philosophies. Douban is usually considered an "elite community". The users somehow show superiority feeling for being the "minority". But sometimes Douban users are accused of being poseurs. So, I'd like to figure out whether the Weibo users tend to follow the crowd

more than Douban users do. Doing so could lead the question to a scientific conclusion rather than just a conjecture.

3. Dataset.

douban.txt

    A plain-text file containing about 270 thousand lines of Douban follower records

    Format:

        1000165   2327571

        (User tab followee)

    (Available at http://www.datatang.com/data/43916. Requires registration)

weibo.xml

    An xml file containing about 400 thousand lines of Weibo follower records

    Format:

        <RECORD>

        <person_id>10145</person_id>

        <guanzhu_id>10029</guanzhu_id>

        </RECORD>

        (person_id: user, guanzhu_id: followee)

    (Available at http://www.datatang.com/data/14350. Requires registration)

4. Methodology (algorithm or analysis).

Both websites uses one-way follow system.

Step1:

Read all the inputs and store the relations for each site in dictionaries, the dictionaries map from followers to followees.

Step2:

Find "popular accounts", defined by top N percent users ranked by number of followers. Use several different Ns (0.5, 1, 1.5, 2, ... , 10) to get the result.

Step3:

For every user in each site, count the number of popular accounts in his or her followees, divide the number by the user's total followees and get a percentage. I call it Herd Index (HI) based at N.

Step4:

Calculate each site's total HI, and divide by the number of users in the site's record, get the site's average HI.

Step5:

Compare the two site's HIs based at different Ns and plot the data into a graph.

Step6:

Go to the popular accounts' homepages to figure out the difference (in style) between Weibo and Douban popular accounts. I would classify those accounts into two types: commercial accounts (like news accounts, VIP accounts of real world stars, marketing accounts, etc.) and personal accounts (individuals with personal charm).

5. Results.

The average followees of Weibo users: 254.91

The average followees of Douban users: 27.73

The Herd Index at different Ns graph:

Herd Index based at N
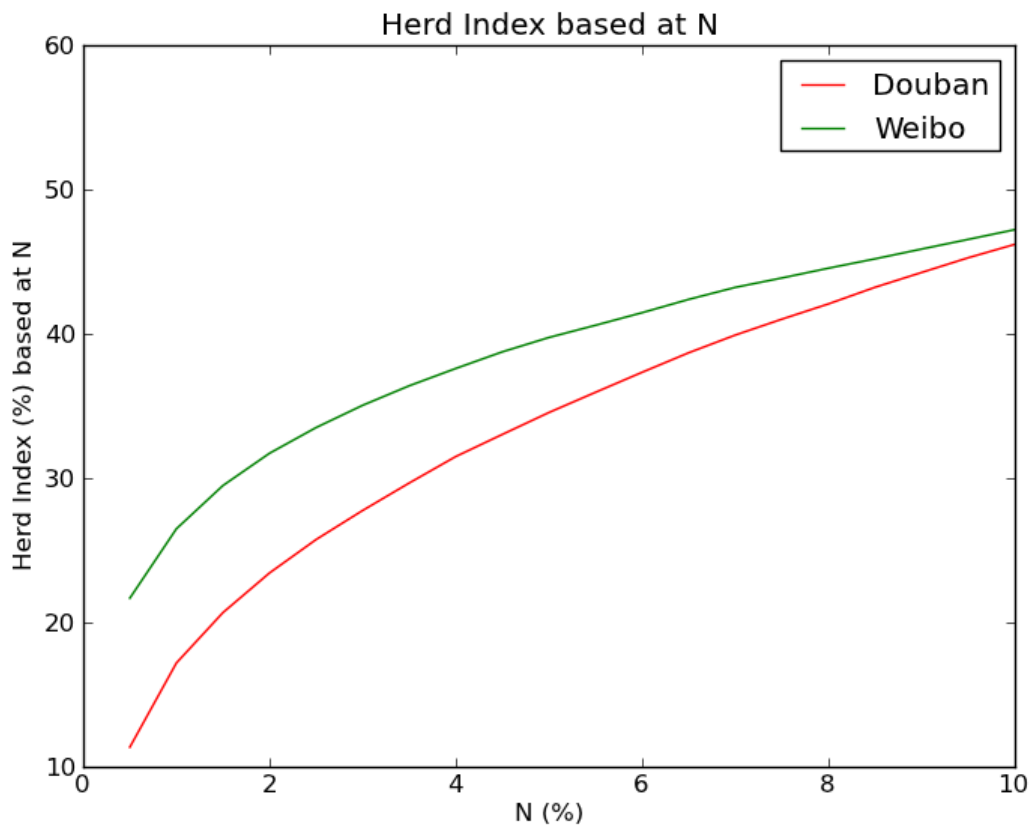
From the average number of followees, Weibo users follow much more followees than Douban users do. And Weibo has much higher Herd Index than Douban. The difference is most obvious when N is small (small N means extremely popular accounts) At N = 0.5, the Herd Index of Weibo is nearly twice much as Douban. So it's obvious that Weibo users more tend to follow the crowd.

Weibo top 10 accounts:

Breaking News: a news account

Xiaogang Feng: a Chinese director

New Weekly: a magazine

Chen Yao: actress, known as "Weibo queen", extremely active on Weibo.

Kaifu Lee: former vice president of Google, best-selling author. Regards himself as Chinese young people's "life mentor". The account is running by his team.

Degang Guo: comedian

Shiyi Pan: property developer

Kevin Tsai: television host in Taiwan

Bad Jokes Selection: sending lots of tweets everyday

Chengpeng Li: soccer commentator

None of the Douban stars are real world celerities.

Popular accounts on Weibo are mainly commercial accounts. Weibo helps a lot with exposure. Real world celebrities use it to hype. Marketing companies use it to create buzz. Therefore, Weibo is more commercial, while Douban is more personal.

6. Reproducing your results.

Run the analysis.py

Get the results:

Output in the console:

The average numbers of followees of Weibo and Douban users

The homepage links for top 10 Weibo and Douban popular users

File "plot.png" at the same directory:

The HI-N graph

7. Collaboration.

None

Reflection.

It's kind of exciting to design and solve a problem using the cool things I learned from this class.

Update: It's really amazing to see the program providing results supporting my conjecture. It's kind of like my idea was scientifically proved.