BIG DATA Related Sina Weibo Social Network Mining

Shang-Yi Ning, Zhe-Chao Huang and Peng Zhang

Abstract—Since the emergence of social network, there has been interest in mining valuable information hiding behind the frequent activities on social networks. In our Social Media Mining class, basic ideas and algorithms in social media mining are introduced, while there are a lot more interesting ideas, methods and algorithms in this realm is still for us to delve deeper.
By applying knowledge learned from class and our after-class self-guided learning along with the tools and programming languages we've mastered, we have conducted an independent and comprehensive research on Weibo- the counterpart of Twitter in China. Shown in this report are our work in mining Weibo including how we design criteria to identify 1000 BIG DATA related users in Weibo, the way we collect data using web crawler, abundant descriptive analyses of users and the follow-up relation network, literature review of user influence, the logic and rationality of a comprehensive and original model of user influence, tag cloud and improvement and extra semantic analysis of weibo text.

Index Terms—Sina Weibo, BIG DATA, Social Network, User Influence.

1 Introduction

Weibo has gained huge popularity since the first day it was lauched. As an emerging form of communication, Weibo allows users to publish brief message updates and provide "social-networking" functionality. Unlike other social network services that require users to grant friend links to other users befriending them, Weibo employs a social-networking model called "following", in which each user is allowed to choose who she wants to follow without seeking any permission. Conversely, she may also be followed by others without granting permission first. In one instance of "following" relationship, the user whose updates are being followed is called the "friend", while the one who is following is called the "follower". One of the most studied problems in Weibo is the identification of influential users. This problem is especially important, considering the high percentage of users who are often inactive or do not provide additional information. Thus, valuable information can be collected quicker once we identify influential users. Furthermore, the criteria for identifying influential users are as many as the growing number of techniques to rank them.

Our study is based on the data we collect from weibo.cn which includes tags, weibo texts, nicknames *etc*. The criteria to choose the data and how we get these data are listed in Section 2 and 3. Meanwhile, some basic analysis and their visualization results can be accessed through Section 4. Section 5 and 6 mainly focus on the process of generating the most influencial users and their shared tags. The last part displays some extra contents of research which are helpful for further study.

- *Shang-Yi Ning and Peng Zhang are from the School of Data Science, Fudan University, Shanghai, China. Email:{14300180071, 14300180051}@fudan.edu.cn.*
- *Zhe-Chao Huang is from the School of Management, Fudan University, Shanghai, China. Email:{13307130205}@fudan.edu.cn*

When using "Weibo" we are referring to the company, in contrast, we use weibo to mean a certain short text posted by users.

2 Criteria of user selection

The quality of data is definitely important for any data-driven study, thus the selection of users becomes the first and one of the biggest concerns during our study. To better complete our study centering on data-related users, we have contrived a great body of selection methods, and at last we choose the one, that we think is the best way to select these related users.

Our method starts with a idea that we want to search users that tag themselves with "big data". However, the presumption of this idea is surely questionable after deeper inspection.

Firstly, "Big data" this concept itself is too catholic, a great number of users will tag themselves with "big data", as long as they are a little bit related to "data" (Ambiguity). Also, because of the popularity of big data, we can not rule out the possibility that some people deliberately select big data as a tag in order to create a more popular social identity (Fraudulence).

To tackle with these problem, we deduced that the greater number of tags they choose that are related to big data, the smaller the possiblity of Fradulence or Ambiguity. So, we firstly identified 10 tags that have the most concurrence possibility with "big data". These 10 tags are selected through TF-IDF search through 100 pieces of paper concerning big data.

Then we used each tag respectively to conduct a "search based on tag". Each individual search returns a list of 500 users. From the search, we got a list of 5500 users and their tags before shifting out repeated users.

To select 1000 users from these 5500 candidates, we added some more criteria. It is quite a reasonable assump-

tion that more tags related to ”big data” show closer connection to ”big data”. Thus, 5500 candidates are ranked by the number of tags they hold among the 10 selected tags.

After this procedure, we found that the top 1000 users contain more than 3 tags directly in our tag list which ensures their relativity with ”big data”. For the rest of this paper, sometimes these 1000 users will be referred as target users, or the 1000 target users.

3 Implement of spider and solution of suspension

3.1 Tools and basic implement

We used the Python libraries **Beautiful Soup** and **request** to implement our spider script. All the work was done in Windows 10, Python 2.7.13, and Anaconda 4.3.21 . As we all know, weibo.com and weibo.cn, from where we can get the information we want, both need to login if you want to get more information. We used **request** to simulate login with cookies. It can send requests and cookies to the server and return the code of a web page. We can easily parse the web page and get the information we want by **Beautiful Soup**.

3.2 Account suspension and the solution

The process of getting web pages and fetching information only needs 4 to 6 lines of code. However, the biggest problem we got into is account suspension. Weibo began to make a severer limit to spidering since April 2017, which is mainly focused on cookie simulated login. Most of the codes on GitHub are out of action because of it. Before introducing the solution of this suspension, we should have a look at the principle of it. If you call requests to weibo.cn too often, say, 300 to 500 times in a quarter, your cookie will be temporarily suspended for about 10 minutes. Moreover, it is no use of changing IPs by a proxy server without changing cookies for the component Weibo suspended is the access of the cookie you used. After analysing the reason of suspension, there are several ways to prevent being suspended. First of them is to change ways of spidering, such as spidering weibo.com instead of weibo.cn, or login by analog click with Python library **selenium**. The second method is to slow down, including waiting for 2 to 3 seconds between requests and just continue calling request and wait until unblocked. Those methods can solve the problem perfectly, but the speed of spidering may reduce to around 1000 to 2000 pages per hour compared to over 5000 pages per hour with our improved strategy. The method we used in our final edition of code is combining the second method by changing cookies, which means you can change cookies if detected to be suspended. The final flow chart is shown in Figure 1.

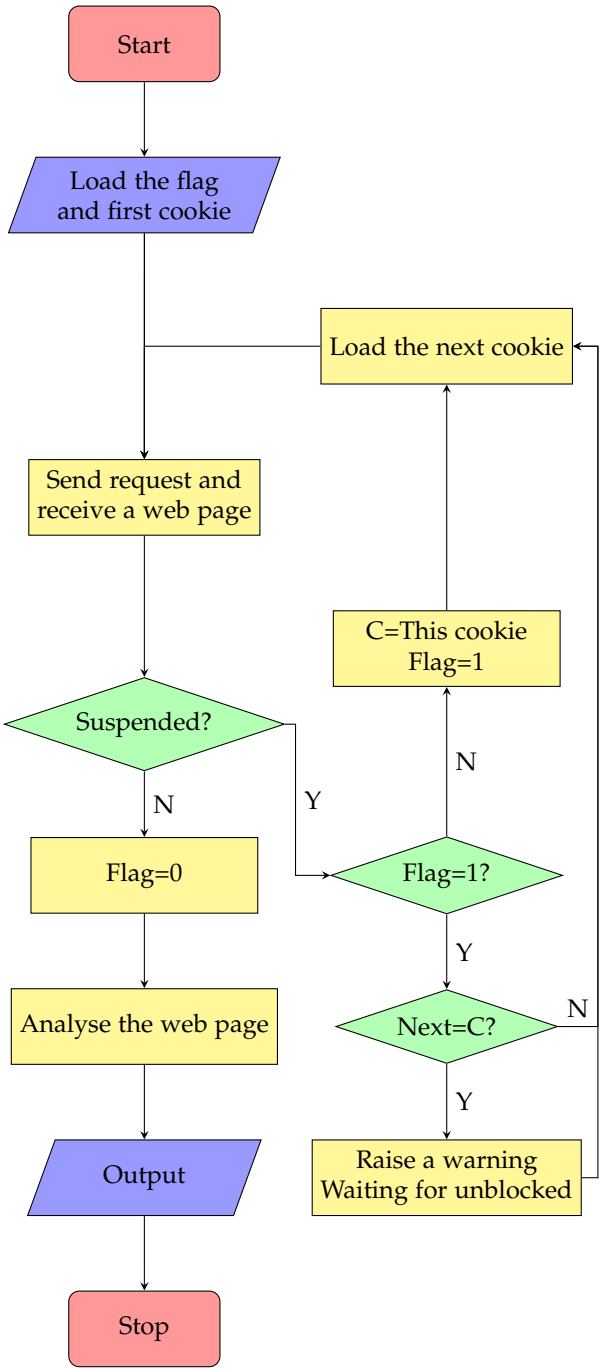

Fig. 1: Flow chart of spidering

4 Descriptive analysis of user attributes and follow-up network

4.1 Description of the Dataset

Using the criteria we designed, that is noted in Chapter 2 and the techniques mentioned in Chapter 3, we successfully got abundant data for each user we select.

We shall take a user to illustrate the items that we get in our dataset. For every specific User X among our 1000 users, we have the following information about User X, which is shown in Table 1.

4.2 Descriptive analyses of the users-based on their profiles

First of all, we conducted some descriptive analysis based on data to help us know about who are the users we selected.

Catagory	Variables	Explanation
profile	userid	the unique user id of user X assigned by weibo.com
	username	user name of user X
	gender	
	birthday	
	area	the province or city that user X lives in or comes from
	self description	self-description written by user X
	tag	the tags selected by user X
	authentication	If user X is authenticated by weibo.com (the Big V), he or she will have this information, indicating who X is in the real world.
	authentication information	A more detailed information about user X's authentication
follow relation	following list	A list of users that this user follows
	following number	number of users that user X is following
	follower number	number of users that follow user X
forward relation	forward list	A list of users whose weibos have been forwarded by this user in his most recent 200 weibo
	average forward time	average times of forward for weibo posted by user X
weibo	weibo list	the text content of the most recent 200 weibo posted by user X

TABLE 1: Information we spidered about User X

4.2.1 *Who Are They*

At the beginning we want to introduce the concept of "BIG V". If a user is a "BIG V", that means he is a public figure that many people may know, and his identification has been checked by weibo.com. After counting on our dataset, we found that 600 in our 1000 selected users are "BIG V".

In order to help readers better understand who are our targets users, we used their authentication information to draw a word-cloud, and finally got the image shown in Figure 2.

As shown in the word cloud, many of our target 1000 users are official accounts of companies, also there are also a considerable number of users being high-level management as well as scientists, professors and authors of books.

4.2.2 *Geographic Distribution of the Users*

Using Baidu echarts, we draw a map based on user profiles, getting the result shown in Figure 3. Geographically, most of them concentrate on the economically most-developed regions (*i.e.* Beijing, Shanghai, Guangdong, Zhejiang).

4.3 Descriptive analyses of the user network

Because we have already got the following list of each user, if we build an edge from the follower to the followee, we can get a directed graph based on follow-up relationships.

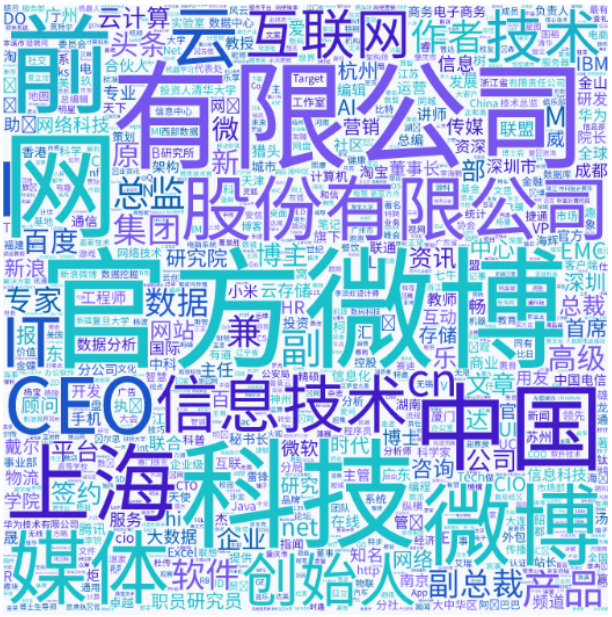

Fig. 2: Word-cloud of authentication information

Fig. 3: The map based on user profiles.

If we only keep the edges between our 1000 target users, we will end up with a directed graph with 1000 nodes and directed edges between them.

For the completeness of our study, we also spent a lot of time analyzing this directed graph.

4.3.1 *Distribution of degrees*

We firstly ranked the users in terms of degree, then we drew a scatter plot of "degree vs. rank", which is shown in Figure 4. In the graph, the node with the highest degree has more than 100 neighbours, also, more that half of our users have more than 25 neighbours. A more traditional chart of degree distribution is shown in Figure 5. From these two charts we can see that only a few users have high degrees, and most of users have low degrees. Fortunately, the network of our target users is not too sparse.

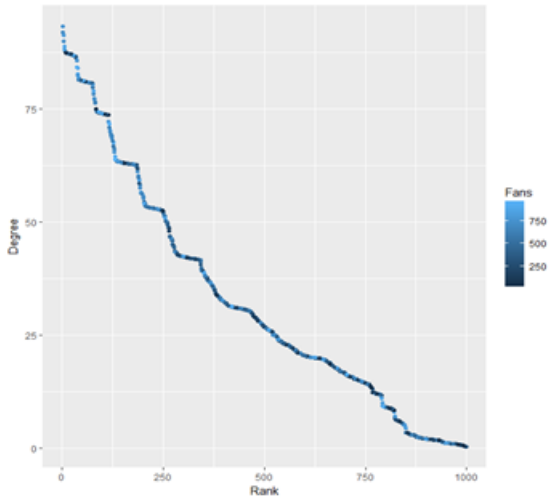

Fig. 4: The plot of "degree vs. rank".

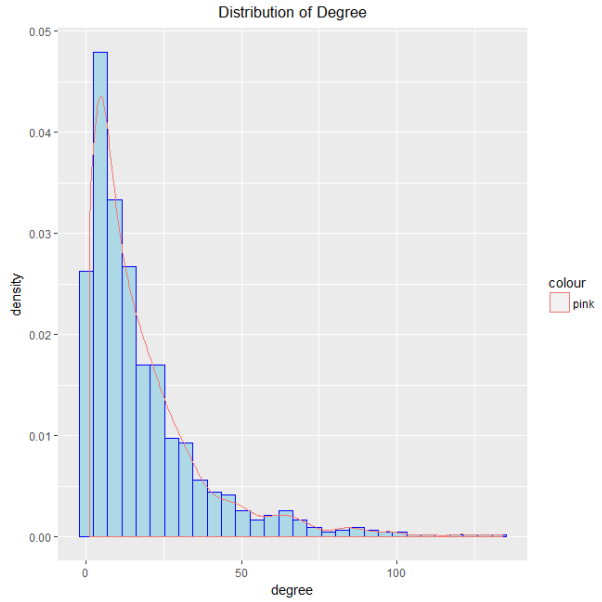

Fig. 5: The plot of "degree vs. density".

We also calculated a lot of descriptive metrics of this graph.

Average Degree :	18.100
Diameter of the Graph :	10
Size of Largest Clique :	8

5 Measuring User Influence-an integrated approach

5.1 User Influence Based on User Centrality

Because of the accessibility of data, the most intuitive way to measure user influence is to equate user influence with node centrality. In the sections above, we have already construct the follow-up network, it is not hard to calculate centrality using commonly accepted algorithm.

We will firstly calculate centrality, and then discuss the limitation of centrality to measure user influence.

5.1.1 Eigenvector centrality

Eigenvector centrality tries to generalize degree centrality by incorporating the importance of the neighbors. It is defined for both directed and undirected graphs.The formula can be listed as:

$$\lambda C_e = A^{\mathrm{T}} C_e$$

Where λ is some fixed constant. Assuming $C_e = (C_e(v_1), Ce(v_2), \ldots, C_e(v_n))^{\mathrm{T}}$ is the centrality vectors for all nodes.

The top five users are "爱可可-爱生活", "好东西传送门", "王威廉", "龙呈镖局", "我爱机器学习".

5.1.2 Betweeness centrality

Betweenness centrality is a method measuring how central v_i's role is in connecting any pair of nodes s and t.

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

where σst is the number of shortest paths from node s to t (also known as information pathways), and $\sigma st(vi)$ is the number of shortest paths from s to t that pass through v_i. The top five users are "社会网络与数据挖掘", "CSDN云计算", "中国云计算", "ZD至顶网", "数据分析精选".

5.1.3 Coreness centrality

Coreness centrality tries to solve the problem raised in eigenvector centrality when it considers directed graphs by adding a bias term to the centrality value.

$$C_{\text{Katz}}(v_i) = \alpha \sum_{j=1}^{n} A_{i,j} C_{\text{Katz}}(v_j) + \beta$$

where the first term is similar to eigenvector centrality, and its effect is controlled by constant α. The second term β, is the bias term that avoids zero centrality values. The top five users are "数据分析故事", "一起大数据", "DoNotExist", "199IT-互联网数据中心", "数据化分析".

5.1.4 Closeness centrality

In closeness centrality, the intuition is that the more central nodes are, the more quickly they can reach other nodes.Closeness centrality is defined as

$$C_c(v_i) = 1/\bar{l}_{v_i}$$

where

$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

is node v_i's average shortest path length to other nodes.The top five users are "社会网络与数据挖掘", "数据挖掘研究院", "中国云计算", "李开复", "199IT-互联网数据中心".

5.1.5 PageRank centrality

PageRank Web Search defines a centrality measure for the nodes (webpages) in a web-graph

$$C_p(v_i) = \alpha \sum_{j=i}^{n} A_{j,i} \frac{C_p(v_j)}{d_j^{\text{out}}} + \beta$$

This equation is only defined when d_{out}^{j} is nonzero.The top five users are "数据化管理", "梁斌penny", "爱可可-爱生活", "李开复", "互联网分析沙龙".

Spearman correlation coefficent Prob> $\|r\|$ under H_0: $\rho = 0$					
	betweenness	closeness	coreness	eigenvector	pagerank
betweenness	1.00000	0.69896 <.0001	0.72532 <.0001	0.83413 <.0001	0.85604 <.0001
closeness	0.69896 <.0001	1.00000	0.93030 <.0001	0.64210 <.0001	0.59409 <.0001
coreness	0.72532 <.0001	0.93030 <.0001	1.00000	0.72554 <.0001	0.66219 <.0001
eigenvector	.83413 <.0001	0.64210 <.0001	0.72554 <.0001	1.00000	0.94033 <.0001
pagerank	0.85604 <.0001	0.59409 <.0001	0.66219 <.0001	0.94033 <.0001	1.00000

TABLE 2: Spearman correlation coefficent

Kendall τb correlation coefficent Prob> $\|\tau\|$ under H_0: $\tau = 0$					
	betweenness	closeness	coreness	eigenvector	pagerank
betweenness	1.00000	0.54517 <.0001	0.57444 <.0001	0.67969 <.0001	0.71580 <.0001
closeness	0.54517 <.0001	1.00000	0.78675 <.0001	0.48411 <.0001	0.43815 <.0001
coreness	0.57444 <.0001	0.78675 <.0001	1.00000	0.58136 <.0001	0.50761 <.0001
eigenvector	.67969 <.0001	0.48411 <.0001	0.58136 <.0001	1.00000	0.80984 <.0001
pagerank	0.71580 <.0001	0.43815 <.0001	0.50761 <.0001	0.80984 <.0001	1.00000

TABLE 3: Kendall τb correlation coefficent

5.1.6 The correlation of different centrality measurement and related tests

To show that these measurements have similar results, we study the correlation of these centrality measurements.

Separately, we used two non-parametric methods to calculate the pairwise correlations, namely, Spearman correlation and kendall correlation. Results are shown in Table 2 and 3. The reason of why we use non-parametric correlation is that it only takes the rank into consideration, rather than the absolute value. Because we want to compare the rank generated by different measures, non-parametric methods are better here.

In each box, the first number is the correlation value of the two measures, the second number is the p-value of hypothesis test of correlation. Generally speaking, the results of these measures are similar.

5.2 The limitation of "Centrality as Influence"

The method of "centrality as influence" shown in the above section is maybe a plausible way to tackle with the problem. However, we must note that it is meanwhile inevitably limited, for it only takes this small network into consideration. Moreover, it is noticeable that this small network is just a sample of the whole network,and the nodes in are highly homogeneous (they are all big-data related users). We have to believe that if we only analyze the centrality, it will lead us to oversimplification.

5.3 Literature Review on "User Influence" in social network

To better approach this problem, we firstly conducted some literature review on this topic.

The first thing to mention is that our text book introduces three variables to be taken into consideration: number of followers, number of mentions and number of forwards. This is firstly published by Meeyoung Cha and Hamed Haddadi (2010), one of their findings noteworthy is that users with high indegree (which means large number of followers), does not necessarily mean they are more likely to be forwarded and mentioned. In this sense, the textbook support us with that the influence, under no circumstances could be simplified as just the topology of the G1 graphs.

Kwak H, Lee C (2010)studied the topology of twitter social network , they compare the PageRank centrality and number of followers, finding that these two measures have high correlation, but when studying the correlation with forwards and follow-up relation, they concluded that forwards and follow-up relations are not congruent when measuring influence.

Moreover, when epitomizing influence related researches, Fabi´an Riquelme divided influence measure into three parts: activity measure, popularity measure and influence measure.

Ye *et al*. analyze the propagation patterns of general messages and show how breaking news spread through Twitter. Furthermore, their team evaluate different social influences by examining their stabilities, assessments, and correlations which lead to the conclusion forwards and mentions can be stable criteria. Weng et al. push forward an extension of PageRank algorithm which is named TwitterRank. This method is sensitive to different themes of the texts and relies on the feature that the number of users and their relationships corresponds to uniform distribution.

In terms of user influence, it is not possible to reach a consensus on this topic, however, all the papers are indicating:

- User influence is a comprehensive concept, have different aspects.
- These aspects are not always congruent, in other words, they are not necessarily correlated.

Also, a question concerning how to combine these different aspects naturally emerges.

5.4 Analytical Framework of User Influence

Taking precedent researchers' ideas as initials and integrating our understanding of weibo, we proposed an analytical framework for user influence measure.

First of all, the influence of a user is reflected in these two aspects, expert recognition (or peer recognition) and social acceptance. The former is the recognition by the users who are professional or authorative in the same field, while the latter is the influence measured on a social perspective. For example, a scholar like 周志华 maybe will not have many followers, but this does not prevent he from being influential in a social network, and actually, he is respected and valued on weibo. Also, an official account may have a limited number of indegree among the big-data circle, but

	Expert Recognition	Social Acceptance
Content	Forward Network Analysis	Forward Times
User	Following Relation Analysis	Follower Number

TABLE 4: Factors impacting the influence of a user

have a greater impact on a normal Weibo user. We can not deny their influence only if they are not popular among experts.

Second, we also notice that follow a user is not completely tantamount to acknowledge the value of the weibo content of that user. Personally, we think that is the reason, as mentioned by Meeyoung Cha and Hamed Haddadi (2010), that the forward times are not highly related with number of followers. For example, we may follow "小米手机" if we are user of this mobile phone brand, but we may not like the content of this official account, not to mention taking time to forward the weibo it posts.

As a result, the influence of a user is at least impacted by these factors: Social Acceptance, Expert Recognition, Recognition of User, Recognition of Content.

Fortunately, we think these factors can somehow be revealed in some corresponding data. For example, the follow-up network of a small expert group might reveal the expert recognition of users to some extent, and the following number of a specific user might reveal social acceptance of this user. We constructed a matric to demonstrate our analytical framework. It is important to note, that we concede our framework might be not exhaustive enough, but it is at least a good explanative attempt. Also, for the variables in the box, we are not claiming they are the best or the only variables that can reveal the latent factors.

5.5 Factor Analysis of the variables

5.5.1 brief introduction of factor analysis model

Let X be with mean μ, the correlation matrix is Σ, $F_1, F_2 \dots F_m$ and $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ are unobservable random variables, and,

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

Its matrix form is $X^{p\times 1} - \mu^{p\times 1} = L^{p\times m}F^{m\times 1} + \varepsilon^{p\times 1}$.

5.5.2 Find underlying factors

We use four variables to measure the influence of a user: Pagerank$_1$ (The pagerank value of the follow-up network), pagerankk$_2$ (the pagerank value of forward relation network), lf (log value number of followers) and lr (log value of average forward times). The reason why we conduct log transformation to number of followers and forward times is that the number of followers and number of forward times of some users are hundreds times larger than those of other users. Because factor analysis is a linear model, it is more reasonable if we conduct log transformation to these two variables. And we want to state that this part is a little bit different with our presentation, because when calculating the presentation ranking, we did log-transformation to all these four variables, we later find out we have no justification to conduct log-transformation to pagerank1 and pagerank2, so we have modified it when writing this report.

	Factor$_1$	Factor$_2$
lf	0.29357	0.74961
pagerank	0.80496	0.32733
pagerank$_2$	0.91214	0.07903
lr	0.08033	0.88114

TABLE 5: Rotated Factor Pattern

Factor$_1$	Factor$_2$
1.5722765	1.45517121

TABLE 6: Variance Explained by Each Factor

From the output of SAS program, SAS detect two underlying factors. The first factor have larger load (load is a terminology in factor analysis) on pagerank1 and pagerank2, thus I want to name it Peer factor, the second factor has larger load on lf and lr, thus I want to name it Social Factor. These two factors have explained $3.02/4 = 75\%$ of the total variance, and for each variable, they have explain more than 64.8% of the variance. This result is quite good.

Also SAS program gives a matrix indicating how to calculate factor score using standardized variables as well as print the score for each factor. Score of Factor$_1$ $= -0.02293u(\text{lf}) + 0.5096u(\text{pagerank}) + 0.67436u(\text{pagerank}_2) - 0.23029u(\text{lr})U(x)$. Here is a standardizing function

$$u(x) = (x - E(x))/sd(x).$$

The best thing about factor analysis is that we can combine the scores of different factors using their explained variance as weight, to calculate an integrated score.

$$\begin{aligned} \text{Integrated Score} =& 1.5722765 * \text{Score of Factor1} + \\ & 1.451712 * \text{Score of Factor2} \end{aligned}$$

Table 8 listed the top 20 users with their integrated score.

5.5.3 Strength of our method

From Table 8, we can see academic authorities, official account of big companies, official account of research-oriented organizations, and celebrities are all included into our ranking. It is far better than just to take centrality into consideration. So the first strength of our method is that it is every comprehensive, considering much more aspects or influence, all of which are logically and empirically reasonable. The second strength of our method is that it automatically generate a way to combine these different variables. It is no like many other methods we read in related papers, those suggest to subjectively assign a combination coefficient. The third strength is that we successfully quantify the influence of each user using integrated factor score, not only provide a basis for ranking but also embody the difference in user-influence. For example, we can see from Table 8 that though

lf	pagerank	pagerank$_2$	lr
0.6481	0.7548	0.8382	0.7829

TABLE 7: Standard Scoring Coefficients

Username	Authentication	score
数据化管理	零售业数据化管理咨询顾问、培训师 微博签约自媒体	19.82111
爱可可-爱生活	知名互联网资讯博主 微博签约自媒体	13.12759
199IT-互联网数据中心	199it.com官方网站微博	12.89743
李开复	创新工场董事长兼首席执行官	11.87444
互联网分析沙龙	互联网分析沙龙 techxue.com官方微博	11.77049
雷锋网	雷锋网官方微博	8.338453
今日头条	北京字节跳动科技有限公司	8.27314
梁斌penny		8.183435
社会网络与数据挖掘	社会网络与数据挖掘自媒体微博	7.511439
CSDN云计算	这里是CSDN云计算频道官方微博，做领先的云计算技术传媒。	7.249951
小米公司	小米科技有限责任公司	7.120942
小米手机	北京小米科技旗下手机品牌小米手机	6.851038
英特尔商用频道	英特尔中国商用官方微博	6.592365
南大周志华	高等学校教师 周志华	6.371098
唐杰THU	清华大学副教授，Arnetminer创始人 唐杰	5.963739
薛蛮子	天使投资人、微博打拐发起人之一	5.942586
王威廉	UC Santa Barbara计算机科学系助理教授	5.791154
数知实验室	数聚变（北京）科技有限公司	5.667298
网易新闻客户端	网易新闻客户端官方微博	5.658789
Coding	云端开发平台 Coding.net 官方微博	5.062991

TABLE 8: Top 20 users with their integrated score

数字实验室 is more influential than 网易新闻客户端, their influence are quite close (5.667 versus 5.659).

6 Tag-cloud and improvement

6.1 Initial ideas

In order to generate the tag-cloud of each user, we first adopt the data of their tags marked by themselves to finish this work by calculating the frequency of their tags. As is shown in Figure 6, most of the tags are related to big data which indicates the property of the choices of the1000 users. Among these tags, ”云计算” and ”数据挖掘” rank the first two labels.

6.2 Problems of the initial one

However, some users chosen from tha data are actually nowhere like the one shown in their tags. For example, the user whose ID is 1750659547 displays his tags as big data, Internet, education etc, whereas he is an officer in real life and his true interests are found to be like wine, red pockets and so in Figure 7 we analyse his contents of weibo.

Fig. 6: Tags related to big data

Fig. 7: The contents of weibo of a lying user

6.3 Improvements

To prohibit this situation, we introduce TF-IDF to help us detect the underlying interests of the users based on weibo text. We take every user as a document and calculate the TF-IDF for every word in this document. Then,the Top 10 words are used to represent this user.The final tag-cloud is generated according to the Top 10 words. Though there still exist some words that seem unrelated to big data, the final results do improve a lot compared with the initial one and reflect the interests of users more accurately.

7 Extra Semantic Analysis

7.1 Web Crawler of Baidu Baike

Becasue the weibo text of our target users are featured with high professionalism and stark subject concentration, by no means could a regular word cut achieve an ideal result. To illustrate, suppose that a regular word cut definitely could not detect ”深度学习” as a word, instead it will be cut into ”深度”and ”学习”. To offset the effect of this ”preliminary cutting”, we spend sometime to build a customed word list related to ”BIG DATA”. The method is described in the following subsections.

Fig. 8: The final result of tag-cloud

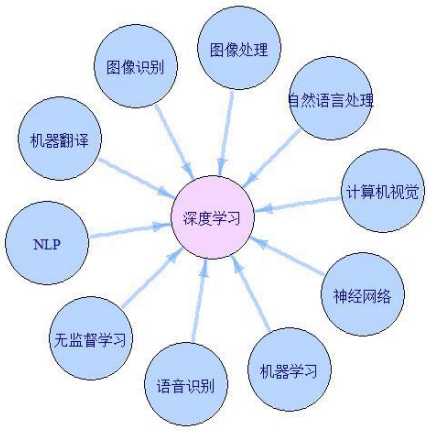

Fig. 9: Words related to big-data

7.1.1 Method of Generating customed word list

Our method relies on Baidu Baike, using Breadth-First-Search (BFS). The logic here is that if a word is used to explain a conception, then logically it is semantically related to this concept. So, first of all, we put 10 basic seed words into a priority queue. In every iteration, we get the first element of the priority queue,let's call it temporary seed word, visit the page of this word in Baidu Baike (Requests), find all the words used to describe this temporary seed word that have been collected by Baidu Baike and add them to the tail of the priority queue.

7.1.2 Our findings when generating the word list

It is necessary to underline that, when analyzing each individual word's webpage,we only find the words that appear in the main text into our queue. This is due to our finding that in other parts of the webpage unrelated words would frequently show up.

7.1.3 Results and application

After 1000 iterations, we get a word list of more than 27000 words, covering nearly every topic within "BIG DATA". **All of our semantic analysis, including the word tag mentioned in the sections above, has applied our customed word list as a complement of traditional word cut.**

7.2 Understanding the weibo context

It is interesting to find out this phenomen: on social media, the meaning of certain words might be altered moderately or even drastically. In order to better observe this phenomen and help us better understand the meaning of words in the weibo context. In order to do this we want to introduce the concept of word embedding. The general idea of word embedding is to represent a word as a vector of real numbers. A rudimentary version of word embedding is usually based on skip-gram model. Our analysis is basicly based on

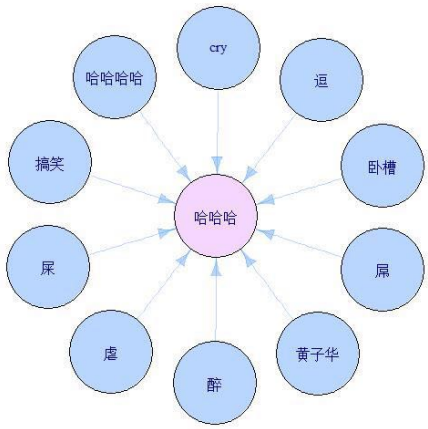

Fig. 10: Words related to "哈哈哈"

Word2Vec by Tomas Mikolov(2013). Depending on Gensim module of python we train the weibo text we get.

Part of the results we get are shown in Figure 9 and 10. In the following figure, we calculated the most similar ten words for the word in the centric word.

Figure 9 shows that our model successfully detect the similarity between big-data related words.

Figure 10 shows that actually, our training result is quite in accordance with our former assumption, that the weibo context is quite different from normal daily conservation or formal writing. For example, the word "cry" is at first glance an ant of "哈哈哈", however, in the weibo context, they are quite similar. Because on weibo, few of us will use "cry" alone it is often used in a phrase like "笑cry" and "蠢cry". And "黄子华" is basicly a famous funny guy, so it is not

Topic1	Topic2	Topic3	Topic4	Topic5
新	朋友	工具	研发	选择
时代	研究	想	目标	股票
模式	跑	照片	喵	确实
商业	学院	长	启动	采访
期待	教授	Excel	重点	哭
梦	人员	一条	职位	酷
互联网	地点	女性	简历	程序
畅谈科技未来	院校研究分享	实用工具分享	公司研发招聘	日常生活扯皮
Topic6	Topic7	Topic8	Topic9	Topic10
微信	合作	AI	图	升级
号	战略	搜索	电脑	回应
送	华为	人类	事件	内部
点	联通	社区	币	值得
链接	医疗	一篇	比特	离开
关注	协议	高考	医院	刷
点击	推动	程序员	勒索	朋友圈
满地打滚求粉	关注科技大事	AI统治力讨论	新闻热点跟进	科技公司动态

TABLE 9: The latent topics of weibo text

surprising to see "黄子华" similar to "哈哈哈".

7.3 Topic Model-LDA

Latent Dirichlet allocation (LDA) is an exmaple of topic models that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. We use LDA to discover the latent topics of the all weibo text (taking every single weibo as a document). Table 9 illustrates our result. Each column corresponds a theme and its seven most representative words with the last row of each column standing for our explaination of each topic.

7.4 Tentative study on weibo-similarity and recommendation proposal

At last, we conduct some tentative study on similarity between different weibos or Weibo users, and want to use it for recommendation.

7.4.1 two different angles of recommendation

In our study, we take two different angles to look at this question, first is to take every weibo as a document, measure the similarity between weibos, and recommend the weibo with highest similarity value. The second method is to take every user's most recent 200 weibo combined as a long document, and then measure similarity and produce recommendation based on the analysis of long documents.

We hope that on the basis of this study, we can find the underlying similar interests of different users and put it into application by making recommendation for Weibo users.

However, the results of our tentative study is far from perfect, so we would like to just describe our methodology, generalize why our results are not as perfect as expected, and propose the next step of our study according to our reflection on these tentative studies.

7.4.2 Two methods to measure similarity and analysis of strength and weakness

- **TF-IDF**: after calculating the TF-IDF value of words in different documents, we can generate a vector for each document, by measuring the similarity of these vectors we can derive the similarities between documents.
- **LDA**: LDA model is essentially a bayesian model, it can represent documents into topics and topics into words. Thus it provides us with another method to measure similarities between different documents, we have also trained our documents using this method.

However, the results of all these three methods applied individually do not work well. After reviewing the models, we think that TF-IDF based similarity is facing the problems like high dimension and sparse data. LDA model does not work well because of the wide range of topics and a possible direction for us to dig deeper is that we can introduce our word2vec result into TF-IDF calculation to curtail the number of dimensions(if we firstly conduct a cluster analysis of word2vec vectors, and treat every words in the same cluster as the same when calculating TF-IDF) , also word2vec result can be used to improve LDA model. Besides, some new methods like Doc2Vec by Thomas Mikolov(2014) are also good ideas to be considered.

8 Summary

In this survey we have collected 1000 users which we thought to be connected with "Big Data" based on the criteria set up in section 2. Moreover, we distinguished the Top 20 users who are the most influential ones among these 1000 users using several indexes. One method that needs to highlight here is that we introduced factor analysis of the variables into our survey which help us find the underlying factors of the users. The final results of Top 20 can be viewed from Table 8. To complete the tag-cloud with the data, TF-IDF was of great help to generate figure 8. In order to improve our research and for further study, we did some extra work which include employing word to vector technique to allow the machine better understand the content of Weibo text. What is more, LDA model is selected for the purpose of completing recommendation system.

Though the final results of our research may not so perfect due to limit of time, it can help researchers and developers to know the variety of centrality measures for Weibo network. This variety of criteria to identify influential actors is a proof that the concepts related with the spread of influence in social networks have not yet reached a consensus. Also, the improvement of our recommendation system can be a great topic in our further study.

9 Collaboration

- *Shang-Yi Ning* finished most work of weibo spidering, including collecting information and choosing tags and users who are our target users, collected data for nearly 1GB and finished this part of the report. He was also in charge of the art work, including drawing word-cloud, final oral report and the composing of this report.
- *Zhe-Chao Zhang* coded the web crawler of weibo user profile and Baidu Baike, he also made small contribution to assisting Shangyi using beautifulsoup. When

working on the project, He individually spent a lot of time to complete the descriptive analysis, calculating different centralities, their corelation and other graph metrics. He proposed our influence measure method after reading related paper, calculate the data for all the charts and tables shown in our report including the word cloud. He also conducted the extra semantic analysis (TF-IDF, Baike dictionary, Word2Vec, LDA etc) In terms of reporting, he is in charge of making PPT slides, presenting our results in class, writing reports abstract and chapter 2,4,5 and 7.

- *Peng Zhang* finished the visualization of our research which includes some statistical results such as the distribution of degrees and some comprehensive results such as the community detection after analyzing our data. Also, he took part in the writing of the final report, finishing the introduction part and tag-cloud part. Moreover, he took the responsibility to review the report in order to check some minor errors.

REFERENCES

[1] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In DBLP, editor, *International Conference on Weblogs and Social Media*, Washington, Dc, Usa, 2010.

[2] T. Mikolov, K. Chen, G., Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013.

[3] F. Riquelme and P. González-Cantergiani. Measuring user influence on twitter: a survey. *Information Processing and Management*, 52(5):949–975, 2016.

[4] J. Weng, E. P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In ACM, editor, *ACM International Conference on Web Search and Data Mining*, pages 261–270, 2010.

[5] S. Ye and S. F. Wu. Measuring message propagation and social influence on twitter.com. In Springer-Verlag, editor, *International Conference on Social Informatics*, volume 11, pages 216–231, 2010.