Information Diffusion Prediction Model in Online Social Networks

By

Tu JiaChen, Paul

(1630005050)

A Final Year Project thesis (STAT4004)
submitted in partial fulfillment of the requirements
for the degree of

Bachelor of Science (Honours)
in Statistics

at

BNU-HKBU
UNITED INTERNATIONAL COLLEGE

November, 2019

DECLARATION

I hereby declare that all the work done in this Project is of my independent effort. I also certify that I have never submitted the idea and product of this Project for academic or employment credits.

____

Tu JiaChen, Paul
(1630005050)

Date:________________________

Information Diffusion Prediction Model in Online Social Networks

Tu Jia Chen, Paul

Science and Technology Division

Abstract

The online social media have become the most important ways for people to conduct social activities. Information flashes through the complex links between users on the Internet. There have already been many researchers who have noticed this phenomenon and conducted a lot of relative researches about online social network. Making good use of online social networks can help governors supervise the users' opinions, implement Internet marketing, block the spread of rumors, and so on. These advantages continuously appeal scientists to explore the dissemination of information. In this thesis, we first extracted information from the real world data and defined several new attributes to better represent the data. Then, we resampled the data and applied machine learning techniques to train the models to predict the information diffusion. Eventually, we got several meaningful factors to explain the dynamics of this phenomenon.

Keywords: online social network, information diffusion, principal component analysis, neural network

Preface

Acknowledgement

First of all, I must appreciate my supervisor Dr. Ping He. It is her continued patience, timely enlightenment, and constructive criticism that have helped me a lot in accomplishing the research ans thesis. Also, I would like to express my gratitude to Dr. Yuhui Deng who have spent plenty of time teaching me knowledge about machine learning techniques. Then, I am indebted to my parents who are always supporting and encouraging me. Moreover, I should thank to all the other teachers and classmates who always accompanying with me. Finally, special thanks should go to three seniors: Mr. Yuanqi Li, Mr.Linchen Weng, and Mr. Jianyi Zhang who have helped me a lot at the beginning of this thesis.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Literature Review 2

1.3 Definitions 5

1.4 Assumptions 5

2 Methodology **7**

2.1 Data Description 7

2.2 Data Preprocessing 9

2.3 Data Balancing 11

2.3.1 Synthetic Minority Oversampling Technique 12

2.4 Factor Analysis 12

2.5 Artificial Neural Network 13

3 Experiment and Result **14**

3.1 Statistical Test 14

3.2 Experiment 17

3.2.1 Model 1 17

3.2.2 Model 2 18

3.2.3 Factor Analysis 20

4 Advantages and Improvements **24**

4.1 Advantages 24

4.2 Improvements 25

Bibliography **26**

List of Tables

3.1 Levene's Test Result of Similarity 15

3.2 Levene's Test Result of Time Gap 15

3.3 T-test Result of Similarity 16

3.4 T-test Result of Time Gap 16

3.5 Table 1 of Loadings 20

3.6 Table 2 of Loadings 21

3.7 Summary of FA 22

List of Figures

3.1 Plot of loss in Model 1 18
3.2 Plot of loss in Model 2 19

Chapter 1

Introduction

1.1 Background

In recent years, due to the popularity of smart devices and internet, the ways of social activities have evolved and the online social networks (OSNs) are playing a vital role in the information diffusion. A large number of Internet users tend to send, receive, accept, and share information through various online social platforms such as Sina Micro-blog, Twitter and Facebook. Users can post a short paragraph, pictures or videos at a time. Then, all their followers can see the content. By this way, information propagation is much faster and more complicated on this kind of online social platform. However, before this revolution, people can only acquire information passively from traditional media like television, radio and newspaper. Although these new social media have made the social activities more convenient in our life, they also have caused some unexpected problems, for example, spread of rumor and harmful information which give rise to bad social impacts. Therefore, having a good understanding of dynamics of information diffusion in the OSNs is essential for not only inspecting the opinions on the Internet, but also cutting off the propagation of rumor and harmful information as soon as possible.

As the complexity of relationship between people in the real society, a complete social network may be very complex with enormous size. The whole structure of information propagation in OSNs can be regarded as a network by linking all the related nodes which stand for users together. The edges which stand for the following relationship in the network structure can be marked as directive ones pointing from followers to following people. A source message appears at a node and then, spreads along the directed edges and eventually reaches other nodes in the network. Researchers can easily find the influence range of a given event. However, the underlying information of spreading properties and paths in the social network has more research value. Knowing more about the factors that motivate the information diffusion between users is necessary in providing convincing explanation for the people's social characteristics. Once we can predict the information diffusion activities between each pair of users in the OSNs, the information diffusion process can be completely simulated with the structure of the network.

1.2 Literature Review

For decades, with respect to the diffusion process, there have already been plenty of related work. At the very beginning, in 1927, Kermack and Mckendrick [5]proposed the SIR (susceptible Infected Removed) model applied in epidemiology. This model shows a general way which is easy for researchers to understand and to describe the mechanism of the spread of disease. It divides the people into three states which are susceptible, infected, and removed. Here, susceptible represents the people who are not infected but have a probability to be infected; infected represents the infected people; removed represents the people who have recovered from the disease and cannot be infected later. The three kinds of people may change their states with a given infecting rate or recovering rate. They also proposed the SIS (Susceptible Infected Susceptible) model [6] which allows the removed people in the SIR model to become susceptible again. These two models are the classic models of epidemiology. Researchers then find that the spread of disease is similar with the information diffusion and therefore, they apply this epidemic model in the study of information diffusion. But this kind of model has a shortcoming that it simulates the process on the macroscopic level rather than considering the microscopic distinctions between different relationships and the network structure of the population. Based on the network structure of the population, IC (Independent Cascades) model [1] and LT (Linear Threshold) model [2] are two representative models. For both models, the diffusion process proceeds iteratively in a synchronous way along a discrete time-axis [4]. The IC model supposes that for each edge of the network, there exists distinct diffusion rate and for each iteration, each of the newly activated nodes (infected nodes) conducts an infecting action to try to active (infect) non-activated nodes (susceptible nodes) according to the rate. Different form the IC model which concentrates on senders, the LT model mainly focuses on receivers by assigning distinct influence degrees on different edges and setting thresholds for receivers to accept the information from their neighbors. It will check all the non-activated nodes once by comparing the overall influence degree from their activated neighbors and the thresholds of themselves. The iterations of these two models will stop when there are no newly activated nodes. These graph-based models are more realistic since they remove the assumption that each individual connects to all the others in the population. Wang et al. [8] proposed a RWSIR (relative weight SIR) model based on SIR model and IC model. Comparing with the SIR model, RWSIR model not only takes the OSNs into consideration but also assumes that different relationships owns different affecting rates. This approach distinguishes the relationships between online users by dividing users into authoritative nodes and normal nodes using the topological properties of the network. The infecting probability is well defined by both information senders and information receivers.

The above models describe and predict the information diffusion by regarding the process as discrete one. To make the diffusion model closer to reality, Saito et al. [7] proposed asynchronous independent cascades (AsIC) model and asynchronous linear threshold (AsLT) model that both apply a continuous time-axis and a time-delay parameter to the diffusion model. For AsIC model, when a sender is about to infect a receiver under a given diffusion probability, the receiver will be infected after a period of time whose value is randomly selected from exponential distribution with a parameter δ. Similarly, in AsLT model, when the influence degree reaches a receiver's threshold, the receiver will be activated δ time later. Guille et al. [3] further proposed the Time-Based Asynchronous Independent Cascades (T-BaSIC model based on the AsIC meta-model and defines the temporal dynamics of information diffusion in OSNs. The parameters in this model are determined by three dimensions: semantics, social, and time. They defined several attributes to describe the users' characteristics which contribute to the diffusion probability and apply machine learning techniques to infer the diffusion probability.

1.3 Definitions

Definition 1.1 *Interest domain*
A user's interest domain is a set of key words extracted from the user's forwarded messages. A user's interest domain is written as:

$$\boldsymbol{I} = \{\boldsymbol{i}_1, \boldsymbol{i}_2, ..., \boldsymbol{i}_n\}, n = 1, 2, ...$$

Example 1.1 *If User $\boldsymbol{A}$ with interest domain $\boldsymbol{I}_{\boldsymbol{A}}$ has forwarded a event(message) Event*1 *with key words set $\boldsymbol{I} = \{\boldsymbol{i}\}$ sent from others, the User $\boldsymbol{A}$'s interest domain will become $\boldsymbol{I}_{\boldsymbol{A}} \cup \boldsymbol{I}$.*

Definition 1.2 *Jaccard similarity coefficient*
Given two sets $\boldsymbol{A}$ and $\boldsymbol{B}$, Jaccard similarity coefficient is defined as:

$$\mathcal{J}(\boldsymbol{A}, \boldsymbol{B}) = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A} \cup \boldsymbol{B}|}$$

When set $\boldsymbol{A}$ and set $\boldsymbol{B}$ are both empty, $\mathcal{J}(\boldsymbol{A}, \boldsymbol{B}) = 0$.

1.4 Assumptions

Assumption 1.1 *All the users who have reposted at least one message from a source user are regarded as fans of the source user.*

Assumption 1.2 *Every fan of a source user has read all the messages posted from the source user.*

Assumption 1.3 *Every fan of a source user do not receive any other same type information from other users.*

Assumption 1.4 *All the content reposted by a user is what this user is truly interested in.*

Chapter 2

Methodology

2.1 Data Description

In Sina Micro-Blog, the system will assign an unique user identification number to each account to distinguish it from others. Also, each message sent from any account owns an unique message identification number. If a user reposts a piece of information sent by an other user, the system will record the message identification number of the message been reposted as parent. Original messages do not have a parent message and therefore, the parent message identification number is recorded as null. Any user can only set their gender as male or female in this system. User A can follow some other users (e.g. user B) to receive messages sent from user B. Then, user A becomes a follower of user B and user B becomes a friend of user A. If user B also follows user A, they are regarded as bi-followers of each other. For a message, it can be reposted, commented, and liked. Post and repost actions are regarded as two types of events in the data set. The data record the number of messages posted, comments of every message, the time each event happens and number of repost events.

The dataset we used in this research is the Sina Micro-Blog rumor data set which is crawled from the Sina Micro-Blog platform. This data set contains 4663 different events which consists of 2312 rumor events and 2351 non-rumor events. Each data file starts at a source message posted by a source user and records all the forwarding events. Each data file is stored as a .json format file which contains users' and events' detailed information. All the primary attributes are listed below:

User's information:

1. Uid: user's identification number.
2. Gender: user's gender
3. Favourites_count: number of user's favourites
4. Bi_followers_count: number of user's bi-directional followers
5. Friends_count: number of user's friends
6. Followers_count: number of user's followers
7. Statuses_count: number of user's posts

Event's information:

1. Reposts_count: number of reposts
2. Text: text content of event
3. Mid: identification number of message
4. Parent: identification number of original message
5. Comments_count: number of comments
6. T: time stamp

2.2 Data Preprocessing

The objective of data preprocessing is to construct a data set which represent the relationship between source users and fans with all the related attributes of users. The data should contain both positive data (repost events happen) and negative ones (repost events did not happen) to support the training of prediction model using machine learning techniques. We extracted all the primary information mentioned above from the original data set. Then, we further introduced several new variables which describe some other facts of either the repost relationship between source user and reposters or the attributes of all reposters:
1. Similarity between event's key words and user's interest domain
We believe that the topics of the original post texts have an important influence on the users' repost action. Therefore, we applied 'jieba' python module which is a well-behaved Chinese text processing tool of text analysis to extract keywords. The types of keywords we are interested in are noun, person name, place name, verb, and other proper noun. We extracted several most important keywords from each original post of a source user and assigned them to each reposter to build the interest domain of reposters. This help to depict the characteristics of each user. After that, we can compare each user's interest domain with the topics of a specific event and use the Jaccard similarity coefficient to indicate the degree of correlation.
2. Time gap of repost action
The time gap between the time of a message posted and the time of a user reposting the message is believed to reflect a user's activity. An relatively active user is more likely to participate in the repost events.

3. Mean of basic information
For those users who did not repost a message sent from a source user, we recorded this event as negative data. To depict the characteristics of these users, we use the mean value of the all attributes (including time gap of repost action) of the users in the positive data.

The whole data extraction procedure is listed as below:
1. For each source message poster, we list all the events in order of time. These events are divided into two parts. For the first part, about 75 percent of former events are selected to train the characteristics of fans. The rest events which compose the second part are used to validate and produce the positive and negative data we need.
2. The first part of the events is used to be historical information. Using this part of the events, all the users who have reposted at least one message are regarded as fans of the original poster. In each event, the key words extracted from text are used to train the interest domain of reposters. All the other basic attributes are extracted and taken average to represent the overall characteristics of reposters.
3. The second part of the events is used to validate and get the data set we want so far. For each new event, and for all the recorded reposters, we check that whether each user has reposted every event or not. We first extract several keywords from the event and calculate similarity by comparing them with each user's interest domain. After that, if a reposter reposted a event, we record the information of source user and up-to-date information of this reposter. Time gap can also be calculated. Next, We use the new information of this reposter to update its mean of basic information and interest domain. Otherwise, if a user did not repost a event,

we record the information of source user and historical information of this user.

2.3 Data Balancing

From the dataset, on the perspective of the message posted by source user, we find that for each post of a source user, only a small proportion of his or her fans will eventually repost the message. Likewise, on the fans' perspective, most of the fans only reposted few messages from the source user. In the real case, there indeed exists a lot of inactive fans which will have a negative effect on the prediction model. In the data set we got, the proportion of positive data is only about 5 percent or even lower. This kind of case will have a significant effect on the training process of prediction model since the model will be more likely to classify a piece of data to the major class of data set.

There are two basic ideas about dealing with unbalancing data: random undersampling and random oversampling. The random undersampling method is to randomly reduce the number of majority class samples while the random oversampling method is to randomly increase the number of minority class samples. However, they both have disadvantages respectively. Random undersampling will cause a loss of information of data while random oversampling will cause duplicity of minority samples which may lead to over-fitting problem in classifier.

2.3.1 Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) is a popularly used method to increase the amount of minority class data. Comparing to the simple random oversampling, it generates data instead of duplicating minority class data. To a certain degree, it has less overfitting effect. The generating process of SMOTE is based on the distance between data points in minority class. The detailed steps are listed below:

1. For each minority class sample $\boldsymbol{x}_n$, we calculate its distances to all the other samples in the same class and get K nearest samples (neighbors) $\{x^i_{neighbor}\}, i = 1, 2, ..., k$. The calculation method of distance can be Euclidean distance, Manhattan distance etc.

2. Determine the multiple $\boldsymbol{N}$ of scale to expand. Then, for each data, randomly select $\boldsymbol{N}$ samples from the $\boldsymbol{K}$ neighbors with replacement.

3. For data $\boldsymbol{x}_n$ and one of its selected neighbors $\boldsymbol{x}^i_{neighbor}$, generate a new minority sample $\boldsymbol{x}_{new}$:

$$\boldsymbol{x}_{new} = \boldsymbol{x}_n + rand(0, 1) \times (\boldsymbol{x}_n - \boldsymbol{x}^i_{neighbor})$$

where rand(0,1) is a random number between 0 and 1.

2.4 Factor Analysis

Factor analysis (FA) is one of the most widely-used multivariate analysis methods. The purpose of FA is to extract the underlying common factors which reflect the correlations between variables and reduce data dimension. The variables in data are representations of the common factors which are unobserved in real life. Those common factors are usually useful in explaining the real world problems.

2.5 Artificial Neural Network

Artificial neural network (ANN) is a information processing system which processes the input information like the neurons in human brain. ANN has great advantages in processing nonlinear data. ANN consists of input layer, hidden layer, and output layer. Each layer may contain many neurons and all the nodes in two adjacent layers are highly connected. Information travel from the input layer and get into the next hidden layer. The information are weighted summed at each neuron and activated. Then, all the processed information continuously diffuse along the links between neurons. Eventually, information get to the output layer, activated and outputted. Single layer perceptron and multilayer perceptron (MLP) are two types of ANN. In our research, a MLP classifier with 3 hidden layers which consist of 20, 15, and 10 neurons respectively is applied. The activation function is set as tanh function. Quasi-Newton method families are used as optimizer in the classifier.

Chapter 3

Experiment and Result

3.1 Statistical Test

To test whether the variables 'similarity' and 'time gap' reflect significant difference between positive data and negative data, we conduct Levene Test and student's t test to test the difference between the means of these variables in two groups.
We first divide the data into two groups according to response variable. Then, we conduct Levene's Test to test the homogeneity of variance of the two variables respectively. The results are shown in the tables below.

H_0: The variance of similarity in two groups are the same.
H_a: The variance of similarity in two groups are different.

H_0: The variance of time gap in two groups are the same.
H_a: The variance of time gap in two groups are different.

Levene's Test For Homogeneity of Variance		
Degree of Freedom	F value	p-value
1	63902	<2.2e-16

Table 3.1: Levene's Test Result of Similarity

Levene's Test For Homogeneity of Variance		
Degree of Freedom	F value	p-value
1	48.831	2.794e-12

Table 3.2: Levene's Test Result of Time Gap

From the Table 3.1, the test statistic is equal to 63902 with p-value less than 2.2e-16. Under a significance level of 95%, we reject the null hypothesis and conclude that the variance of similarity between two groups are different.

From the Table 3.2, the test statistic is equal to 48.831 with p-value equals to 2.794e-12. Under a significance level of 95%, we reject the null hypothesis and conclude that the variance of similarity between two groups are different.

Next, we compare the means of variables similarity and time gap in two groups respectively. The results are shown in the tables below.

H_0: The difference of mean of similarity in two groups is equal to 0.
H_a: The difference of mean of similarity in two groups is not equal to 0.

H_0: The difference of mean of time gap in two groups is equal to 0.
H_a: The difference of mean of time gap in two groups is not equal to 0.

Welch Two Sample t-test	
Degree of Freedom	13395
t value	-24.105
p-value	<2.2e-16
mean in group 0	0.0059
mean in group 1	0.0160
95 percent confidence interval	$[-0.0109, -0.0093]$

Table 3.3: T-test Result of Similarity

Welch Two Sample t-test	
Degree of Freedom	7939
t value	7.999
p-value	1.421e-15
mean in group 0	27292.43
mean in group 1	16270.21
95 percent confidence interval	$[8321.388, 13723.052]$

Table 3.4: T-test Result of Time Gap

From Table 3.3, the t-test statistic is equal to -24.105 with p-value less than 2.2e-16. Under a significance level of 95%, we reject the null hypothesis and conclude that there is significant difference between the means of variable similarity in the two groups.

From Table 3.3, the t-test statistic is equal to 7.999 with p-value equals to 1.421e-15. Under a significance level of 95%, we reject the null hypothesis and conclude that there is significant difference between the means of variable time gap in the two groups.

Therefore, from the conclusions mentioned above, the two newly introduced variables in our data set can well reveal the difference of positive data and negative data.

3.2 Experiment

In this part, we build several different models and explain the results.

3.2.1 Model 1

We first build a MLP classifier. The data used in this model is acquired after only conducting simple random undersampling on majority class data. The data have 13480 rows and 18 columns. Among them, 6740 are positive and the other 6740 are negative. Maximum iteration times are set as 10 thousand. The result is as follows:

Training Set		Prediction	
		1	0
Real Case	1	4865	203
	0	500	4542

Testing Set		Prediction	
		1	0
Real Case	1	1451	221
	0	306	1392

In the training set, the accuracy of prediction for ’not repost’ (repost=0) is 90.0833% and for ’repost’ (repost=1) is 95.9945%. In the testing set, the accuracy of prediction for ’not repost’ (repost=0) is 81.9788%

and for 'repost' (repost=1) is 86.7823%.

The model fits the training data very well with over 90 percent accuracy. However, the accuracy of predicting the testing set decreased about 10 percent. The overall accuracy is good but this model is not generalized enough.

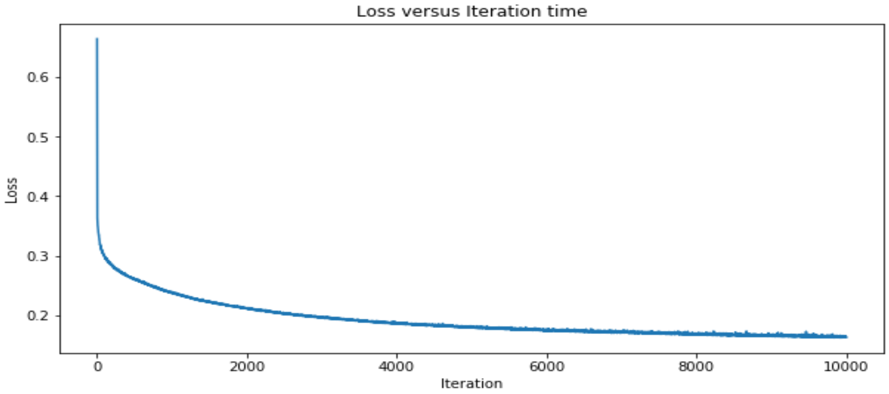

Figure 3.1: Plot of loss in Model 1

3.2.2 Model 2

We use the data resampled using SMOTE to train a MLP classifier again. SMOTE generates some more positive samples into the data that may help build a more generalized prediction model. Simple random undersampling was conducted on the majority class data and 20 thousand negative data were sampled. Then, SMOTE is conducted on the minority class data to generate more positive data. Totally, the new data set has 40 thousand rows. Among them, a half is positive data and the other half is negative ones. The result is as follows:

Training Set		Prediction	
		1	0
Real Case	1	13941	1030
	0	1192	13837

Testing Set		Prediction	
		1	0
Real Case	1	4518	511
	0	504	4467

In the training set, the accuracy of prediction for ’not repost’ (repost=0) is 92.0687% and for ’repost’ (repost=1) is 93.1200%. In the testing set, the accuracy of prediction for ’not repost’ (repost=0) is 89.8612% and for ’repost’ (repost=1) is 89.8389%.
The model fits the training data very well with over 90 percent accuracy. As for the testing data, this model predicts better then the first model above. The generalization of this model is better than the last one.

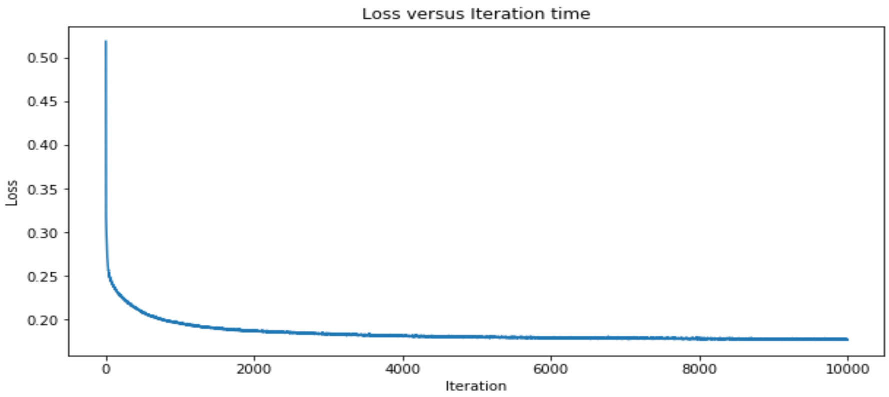

Figure 3.2: Plot of loss in Model 2

3.2.3 Factor Analysis

We conducted Factor Analysis on our data set. After getting 10 factors and rotation using variance maximum method, all the loadings are gotten and listed in the two tables below.

Loadings Factor Variables	1	2	3	4	5
1	-0.956			0.193	
2	-0.956			0.190	
3				-0.975	
4	0.947			0.155	
5	0.835			0.454	
6	0.397			0.208	
7	0.335				
8	-0.107	-0.118	-0.910		
9			-0.921		
10		-0.901			
11			-0.194		0.158
12					0.987
13	-0.319				
14		-0.937			
15		-0.829			
16					
17					

Table 3.5: Table 1 of Loadings

Loadings Factor Variables	6	7	8	9	10
1			-0.102	-0.174	
2			-0.103	-0.175	
3				-0.190	
4			0.125	0.244	
5			0.133	0.279	
6				0.841	
7				0.901	
8					
9					0.102
10					
11					0.964
12					0.147
13			-0.936	-0.108	
14					
15					
16	0.999				
17		0.998			

Table 3.6: Table 2 of Loadings

The eigen values, variance contribution rates, and cumulative variance contribution rates of all the 10 factors are listed in the table below. The ten factors can well represent our data with explanation of over 90 percent variance.

Factor	Eigen value	Variance proportion	Cum. variance proportion
1	3.813	0.224	0.224
2	2.397	0.141	0.365
3	1.752	0.103	0.468
4	1.314	0.077	0.546
5	1.004	0.059	0.605
6	1.001	0.059	0.664
7	1.000	0.059	0.722
8	0.947	0.056	0.778
9	1.765	0.104	0.882
10	0.979	0.058	0.939

Table 3.7: Summary of FA

Explaination of Loadings

From the result, we can explain the factors according to the relatively high loadings in each factor.

Factor 1: Source user's number of posts and favorite posts have high positive loadings on this factor; source user's number of bi-directional followers and friends have high negative loadings on this factor. The factor may be explained as reposter's likelihood of posting.

Factor 2: Reposter's number of followers, comments, and reposts have high negative loadings on this factor. The factor may be explained as reposter's likelihood of receiving and reposting.

Factor 3: Reposter's number of bi-directional followers and friends have high negative loadings on this factor. The factor may be explained as simplicity of reposter's social circle.

Factor 9: Source user's number of comments and reposts have high positive loadings on this factor. The factor may be explained as influence of source user.

Each of the rest factors has only one strongly related variable which has a high loading on it. That is, the rest variables(factors) are already important factors that not strongly related with other variables.

Chapter 4

Advantages and Improvements

4.1 Advantages

Our model performed well in predicting the repost actions in Sina Microblog platform with about 90 percent of accuracy. Also, it is the application of SMOTE to generate minority class data which is relatively rare in real case that makes our model generalized when predicting. Since the data set we used in training part is balanced, our prediction model could also used to estimate the number of reposts of a message sent from a source user. Lastly, we applied FA on the data and got some meaningful factors which could further be used to better evaluate the users' characteristics in OSNs.

4.2 Improvements

Our model still has many aspects to be improved:

1. Due to lack of data with complete network, our model did not take the structure of online social networks into consideration. Many features of topology may have a positive effect on the prediction model of information diffusion.

2. The data we use are not complete ones. Once we could get more historical data about each user, we could better depict users' characteristics according to more historical repost information. For example, based on more complete historical data, we could get a more accurate interest domain for every user.

3. The key words extracted from the content are still not accurate enough to represent the topics of original message. An additional classification model could be applied to classify all kinds of messages with predetermined labels.

Bibliography

[1] J. Goldenberg, B. Libai, and E. Muller. Talk of the Network : A Complex Systems Look at the Underlying Process of Word-of- Mouth Author (s): Jacob Goldenberg , Barak Libai and Eitan Muller Published by : Springer Stable URL : http://www.jstor.org/stable/40216600 JSTOR is a not-for-profit serv. 12(3):211–223, 2001.

[2] Mark Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[3] Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*, pages 1145–1152, 2012.

[4] Adrien Guille, Hakim Hacid, C Favre, and Djamel Abdelkader Zighed. Information Diffusion in Online Social Networks : A Survey. *SIGMOD record, ACM*, 42(2):17–28, 2013.

[5] W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A:*

Mathematical, Physical and Engineering Sciences, 115(772):700–721, 1927.

[6] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics-II. The problem of endemicity. *Bulletin of Mathematical Biology*, 53(1-2):57–87, 1932.

[7] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. Selecting information diffusion models over social networks for behavioral analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6323 LNAI(PART 3):180–195, 2010.

[8] Jin Long Wang, Fang Ai Liu, and Zhen Fang Zhu. An information spreading model based on relative weight in social network. *Wuli Xuebao/Acta Physica Sinica*, 64(5), 2015.