PROCEEDINGS OF SCIENCE

Prediction of Information Dissemination Based on Passive-aggressive Algorithm

Zhanwei Tian[1]
Department of Management science and Engineering Northeast Agriculture University
Harbin,China 150036
E-mail: tianzhanwei@gmail.com

Zihang Zeng, Qiang Chen[2]
Industry engineer of laboratory
Northeast Agriculture University college of engineering
Harbin, China 150036

Prediction on information sharing is the basis for the information control and supervision. Attributes of Micro-blog users and information contain data of users' preferences, physiological characteristics and content type, etc. Based on these data, the information sharing can be predicted. The analysis modes of Micro-blog information dissemination, theory and method in respect of prediction of information sharing based on PA （*Passive-aggressive)* algorithm propose information sharing prediction model. Sina Micro-blog data have verified the model. Results show that the model has high prediction accuracy in terms of information sharing.

CENet2015
12-13 September 2015
Shanghai, China

[1]Speaker
[2]This work was financially supported by the Heilongjiang Social Science Foundation (13C010).

PoS(CENet2015)080

1. Introduction

Micro-blog has become one of the main platforms to obtain information for Internet users. Micro-blog information of the main body of information generated by ordinary users is characterized by timeliness, fragmentation and virus, and has impact of traditional media on the dissemination of public events, unexpected events[1].

Compared with the traditional Internet applications such as BBS forum, person to person network etc. making friends on website. An important feature of Micro-blog is a weak tie between "concern" relations of the users, which makes the Micro-blog information dissemination network path to be small with the high efficiency of information dissemination[2]. Information shared continuously by many users produces the fission propagation transmitted in network, and the share has become the basis of Micro-blog information dissemination mechanism. The prediction of information sharing, on the one hand, is the basis of the breadth, speed, path and influence of information dissemination to predict; on the other hand, it can mine the user's behavior patterns and provide evidence for the recommended information, accurate advertisement and marketing and e-commerce applications. Currently, researches on Micro-blog information dissemination concentrated in the field of data mining and e-commerce applications etc yet lack of theoretical research on the dissemination, share and forecast of information[3-4].

This paper firstly analyzes the mode of the Micro-blog information dissemination and share, then studies the theoretical methods of Micro-blog information share and forecast and propose sharing prediction model based on PA algorithm; finally, verifies the validity of the model utilizing Sina Micro-blog data as a case study.

2. Sharing Mode of the Micro-blog Information Dissemination

The speed of Micro-blog information dissemination is far faster than the tradit ional media and other online social networks. The information dissemination in traditional communication theory is divided into two types: the mass communication and the interpersonal communication. The mass communication information is from a source to a larger group in one-way transfer while the interpersonal communication is a transfer between two or more individuals, while Micro-blog information dissemination is a combination of two propagation modes[5] .

Compared to the traditional media or online social networks, Micro-blog has created a new mode of information dissemination. Instant messaging tools such as MSN, QQ, is the singlet spread of personal-personal, and traditional media, blogs, forums is model of personal-small diffusion. These two modes are of single dissemination, which cannot make the information form the fission propagation, but Micro-blog which forms mode of form the fission propagation mode of personal - small diffusion- small diffusion-mass just makes up for it.

"Concern" and "concerned" is the core of Micro-blog information dissemination. The user A focuses on user B if the information is produced by user B and then automatically pushed to A. Upon user A's forwarding and comments, the information further shares user A's concern, and so forth form fission propagation transmitted. The increase of concerned about the relationship between users is mainly through other interesting users found by the browsing information, specifically, the search and that recommended by the System Friends and so on. Unlike general SNS, a user concerned in Micro-blog, it does not need to go through the permit, and the users concerned does not have to respond; in this sense, this is conducive to the generation of a lot of attention relationship and increase the propagation path of the information.

Strong ties between users "concern" can be seen as a direct interaction information between persons when user as a node in the Micro-blog information dissemination network. While a large number of one-way "concern" weak ties can be seen as one-to-many information exchange between a user with and all other users. The Micro-blog system automatically pushes information generated by his attention consumer,, and user can obtain information of non-attention user by browsing directly, system recommended and search function. Information sharing between different users forms a communication network.

3. Micro-blog Information Sharing Prediction Theories

The information propagation path prediction research has important theoretical and practical significance. Prediction of information sharing is the basis of the propagation path prediction. Sharing prediction refers that the user's behavior and other information can be generated between different nodes by a known network structure, in the network so as to predict the possibility of sharing information.

Different scholars in the field of information sharing, exchange material, propagation path prediction research, such as biotechnology, social network analysis, computer network[6-7]. The main prediction methods include Markov chain and the machine self-learning algorithm based on the nodes attribute information or network topology[8].

Based on the prediction of the node attribute, the similarity between the different node attributes, the behavioral characteristics of the node and a direct communication link prediction are adopted. The results of this prediction is accurate, but the attributes of the nodes have to get ahead. It is difficult to achieve the node attributes of real information in real network, such as micro-blog users many registration information for the protection of personal privacy considerations, which is confidential. Personal information that is publicly available with a lot of false information. In addition, the behavior of the user needs to be obtained by tracking the user's tracking. Accurate information, that information can be used to predict the link for information dissemination, also needs to be judged. It is also very difficult to achieve Predictions based on node attributes doped with many subjective factors and uncertainties.

Forecast based on the network topology, is mainly to take advantage of the network path, the evolution of the structural attributes of the node and information sharing forecasts. The advantage of this method lies in that the topology of the network characteristics can be very easy to get objective factors, if the prediction is correct, the results of the prediction has a high accuracy, and the network has a sindmilar topology structure,which can be learned from each predictor selected and prediction algorithm. As to the disadvantage of this method, without considering the properties of the node, the complexity of the network are not only reflected in the complex network structure but also reflected in the complex behavior of the node.

Whether it is based on node attributes, or forecasts, based on the characteristics of the network structure are description of the existing data as authentic as possible, based on the forecast. Each method has different suitable networks. If the structural characteristics of the network is very clear, a prediction method based on the network structure, can not only reduce the computational complexity, but also obtain a better prediction result. If the network structure is complex with weighted side to side, the use of the network structure prediction, increased computational complexity exponentially, the accuracy will fall.

The result of the Micro-blog information sharing features more user-subjective emotional role, based on node attributes to predict. Micro-blog in the user's personal information, communication behavior and information related attributes to summarize node attributes to predict the characteristics of the information sharing. The difficulty here is to measure these properties and attributes the quantitative forecast information sharing quantitatively.

4. Micro-blog Information Sharing Prediction Model

Micro-blog information sharing involves many users. When there is too much data, traditional algorithms such as decision tree, association rules will be much slower. Based on PA algorithm, we construct the information sharing prediction model.

4.1 PA Algorithm

Different traditional prediction methods which based on whole data set and we can get Micro-blog's real-time information through ttime line. Upon the information release, the prediction of information sharing must be carries out immediately. Dealing with real-time prediction problems should use the online learning algorithm. We improve PA (Passive -

aggressive) algorithm by take advantage of attributes of users and information to predict information sharing[9].

Whether the sharing information is a binary classification problem or not. The basic idea of PA algorithm is that Micro-blog information producing is a continuous time sequence, in each sequence predict sharing for new information, label user with 0(not sharing) or 1(sharing). In the time series end, the result of whether the user sharing will appear, algorithm will produce instantaneous loss to judge the degree of prediction error by using new data of attributes and sharing appeared in this time series to update the existing prediction rules, in passing new rules on the next time series judgment.

Algorithm can be represented as follows: assumeis X_t an example in time series^t, expressed as a vector in vector space R^n,$y_t \in \{0, 1\}$is judge category, (X_t, Y_t) can be used as a sample. Algorithm uses a function to carry out sample classification:

$$y_{t+1} = \arg\max_{y}(w^t \bullet \phi(x_t, y_t))$$

$$w \in R^n \tag{4.1}$$

This function based on the weight vector, the algorithm's main task is through cycle learning to update weight, PA algorithm provides three strategies to update the weight, the weight can be expressed as:

$$s.t.\ell(w; \phi(x_t, y_t)) \le \xi, \xi \ge 0$$

$$w_{t+1} = \arg\min_{w} \frac{1}{2}\|w - w_t\|^2 + C\xi$$

$$\ell(w;\phi(x_t, y_t)) \tag{4.2}$$

is instantaneous loss function that can be defined as:

$$\ell(w; \phi(x_t, y_t)) = \begin{cases} 0 & \gamma(w; \phi(x_t, y_t)) > 1 \\ 1 - \gamma(w; \phi(x_t, y_t)) & others \end{cases} \tag{4.3}$$

$$\gamma(w; \phi(x_t, y_t)) = w^t \bullet \phi(x_t, y_t) - w^t \bullet \phi(x_t, y')$$

$$y' = \arg\max_{z \ne y}(w^t \bullet \phi(x_t, z))$$

$$y' \tag{4.4}$$

is the best classification category.

As to Formula (4.4), we can use Lagrange algorithm to find the optimal solution:

$$w = w_t + a^*\phi(x_t, y_t) - \phi(x_t, y')$$

$$a^* = \min(C, \frac{\ell_t}{\|\phi(x_t, y_t) - \phi(x_t, y')\|^2}) \tag{4.5}$$

l_t is t series' loss function value depending on the loss function. PA algorithm provides three methods to calculate^w . The specific process is shown in Fig. 1.

INPUT: aggressiveness parameter $C > 0$
INITIALIZE: $\mathbf{w}_1 = (0, \ldots, 0)$
For $t = 1, 2, \ldots$

- receive instance: $\mathbf{x}_t \in \mathbb{R}^n$
- predict: $\hat{y}_t = \text{sign}(\mathbf{w}_t \cdot \mathbf{x}_t)$
- receive correct label: $y_t \in \{-1, +1\}$
- suffer loss: $\ell_t = \max\{0\ ,\ 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)\}$
- update:

1. set:

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2} \quad \text{(PA)}$$

$$\tau_t = \min\left\{C\ ,\ \frac{\ell_t}{\|\mathbf{x}_t\|^2}\right\} \quad \text{(PA-I)}$$

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}} \quad \text{(PA-II)}$$

2. update: $\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$

Figure1: PA algorithm process

4.2 Prediction Modle based on PA Algorithm

Research shows that more than 50% of the Micro-blog information is published an hour after being found, regardless of whether the information's forward is influenced by time[10]. There are three peaks of users login time every day, obviously in the three peaks information has more probability by forwarded. So only to establish one global prediction model based on PA algorithm will cause more error; therefore, consider= the time factor that we construct the model as follows:

(1) Construct one global model by using the whole data to update global rules.

(2) For every hour in the day with 24 local models, use data in the hour to update local rules.

(3) When predicting, modify the original algorithm's predicting function and make results decided by the global and local rules.

$$y = \arg\max_{y}(w^t \bullet \phi(x_t, y_t) + \alpha(w^\varsigma \bullet \phi(x_\varsigma, y_\varsigma)) \tag{4.6}$$

w^t is global weight;w^ζ is local weight for a specific time hour$^\zeta$;α is the weight coefficient, which can be defined as proportion of the two parts' training sample; ; is the sample size; is the number of all sample.$\alpha=\zeta_n/N$； ζ_n The larger sample size； N is in period$^\zeta$, the more important local function role it plays.

5. Example Analysis

5.1 Data

Take Sina Micro-blog data to verify the validity of the above prediction method. For the training dataset, this paper has collected data within one hour for research. This paper randomly selects part of the original Micro-blog of Sina for research, logged on Sina Micro-blog anonymously, we can view the real-time messages through "People are saying", which have a lot of forwarded messages (the basis for judging is whether they include the "@"). After removing this part of the message, we have the original message ID and publisher ID. Due to the Sina Micro-blog restrictions, we can only get 1000 users' fans data; thus, we carry out the publisher information analysis firstly. If the publisher has more than 1,000 fans, discard the

message; otherwise we will save the message ID, message publisher ID, publisher fan ID. After one hour, use the API interface, read messages comments by message ID, get all forward user ID; then we can get information publisher and audience characteristics indicators data by the user;s ID. There is no automatic recognition method for the user and the information type data for now. It can be done only by manual identification. The final data contains 5142 information, 14,000 related forwarding users and more than 70 million non-forwarding users. Only about 2% of the fans will forward the information released by the publisher.

5.2 Prdiction Index Selection

Micro-blog information sharing process involves two main body: users and information, in which, the users can be divided into information promulgator and information audience. Combine this paper's analysis and other scholars' research results. The indexes can be used to predict information sharing as selected, shown in Table 1.

	Name	Content
Users	1. Number of fans 2. Number of following 3. Number of information 4. User type 5. VIP 6.Gend	These indexes are relevant to users' activity. Literature 11 has verified significant of them.
Information	1. Content type 2. Length 3. Tag number 4. URL number 5. Time series	Literature 12's research found these indexes are relevant to information sharing.

Table 1: Prediction Indexes Selected from Users and Information Attributes

These indicators are preliminary results in the prediction process; we will carry out cross combination to indexes, further examination, determine the best prediction indexes.

5.3 Basic Prediction

Before prediction, data should be preprocessed. Use Min-Max mode (new data = (original data - minimum) / (maximum - minimum)) to make original data's new value map into [0, 1]. Then use information's indexes and forwarding users' indexes to train forwarding study. Exploit the information's indexes and not forwarding indexes to train not forwarding study. Compared with very large scale of information will not be forward, we are more concerned the information that is sharing; thus we use the accuracy of predicting sharing to evaluate the model. With random classification, most classification (all data are predicted for forwarding), standard PA model, improved PA model are adopted to make prediction based on the sample data with prediction accuracy is shown in Table 2.

Model	Random	Most	Standard PA	Improved PA
Accuracy	2.93%	2.37%	35.5%	47.2%

Table 2: Prediction Results of Four Kinds of Classification Methods

From Table 2, it is known that prediction accuracy is very low when the random and most identification methods are used. Standard PA algorithm model can improve the accuracy to 35.5%, and the improved PA model can get 12% more than standard model.

5.4 Prediction of Different Kinds of Indexes

Considering the time series factor, the prediction accuracy is improved, which makes further thinking as to which indexes are sensitive to time; through prediction by using user and information indexes respectively. The results of prediction accuracy are shown in Table 3.

Model Index	Standard PA	Improved PA
User	37.6	38.2
Information	20.3	34.6

Table 3: Results of Prediction Accuracy by Using User and Information Indexes Respectively

From the table, it is known that improved PA model's prediction accuracy for information relevant index is much higher than users related indexes; therefore, it can be seen that information related indexes are more sensitive to the time series. It can be explained by common sense. Different users at any time may release information in the day, but some types of information may only appear during certain period, such as breakfast related information which is more likely appear in the morning.

5.5 Significant Analysis of Different Indexes

From now the prediction uses all the indexes in Table 1. Upon observation of the improved PA global model, we've find the weight of indexes is very different, as is shown in Table 4; in this sense, the indexes play different roles in the prediction.

index	weight
Number of fans	0.43
Number of following	0.22
Number of information	0.38
User type	0.14
VIP	0.27
Gend	0.31
Content type	0.19
Length	-0.17
Tag number	-0.06
URL number	-0.09

Table 4: Indexes Weight of Improved PA Model

From Table 4, there is big difference between weights of each index, we can reduce indexes, so as to improve the efficiency of model. Remove indexes whose weight absolute value are less than 0.1, 0.2, 0.3 in three separate phases, using improved PA model to predict, the accuracy is shown in table 5.

index	All	Remove <0.1	Remove<0.2	Remove <0.3
Accuracy	47.2%	45.5%	39.4%	24.8%

Table 5: Prediction Accuracy after Removing Different Indexes

From Table 5, it shows that after removing indexes weight value less than 0.2. Compared with accuracy using all indexes decreases 7.8%, and removing indexes weight value less than 0.3 decreases by 22.4%, it is reasonable to use the index to elminate the prediction of the weight of less than 0.2, then remain five indexes, the accuracy is 39.4%. Due to users' arbitrary and complexity of sharing information, accuracy of using improved PA model to predict information sharing is not high, but compared with random prediction's accuracy is less than 3%, it has increased greatly.

6. Conclusion

We analyzed Micro-blog information dissemination and sharing mode, discussed the theory and methods for predicting information sharing, used the improved PA algorithm to predict Micro-blog information sharing. From the two bodies of information dissemination users and information's attributes, we choose 11 indexes to do the prediction. The prediction accuracy of our model is higher than standard PA model. Upon analysis of the weight of all indexes, we've found indexes played very different roles in the prediction. With the indexes whose weight value are less than 0.2 are removed, 5 indexes are finally remained. If these five indexes and improved PA model are used to carry out prediction, the prediction accuracy will lbe 39.4%, which is much higher than the random prediction.

References

[1] Y. Y.Yan, M. Cheng. *Empirical Analysis of All Kinds of Social Networks and Their Relationships Formed by Information Communication among Microblog Users* [J]. Library and Information Service, 2012(3): p.28-31.(In Chinese)

[2] P. Y. Fan, H. Wang, Z. H. Jang. *Measurement of Microblogging Network* [J]. Journal of Computer Research and Development, 2012, 49(4): p.691-699.(In Chinese)

[3] M. Efron. *Information Search and Retrieval in Microblogs*[J]. Journal of the American Society for Information Science and Technology, 2011, 62(6): p.996-1008.

[4] Z. M. Liu, L. Liu. *Empirical study of sentiment classification for Chinese microblog based on machine learning* [J]. Computer Engineering and Applications, 2012, 48(1)： p.1-4(In Chinese)

[5] J. B. Walther, C.T. Carr, S.S.W. Choi. *Interaction of interpersonal, peer, and media influence sources online*[J]. A Networked Self: Identity, Community, and Culture on Social Network Sites, 2010, (17): p.246-267.

[6] H. Yu, P. Braun, M.A. Yıldırım. *High-quality binary protein interaction map of the yeast interactome network*[J]. Science, 2008, 322(5898): p.104-110.

[7] D. H.Ye,.G. P.Jiang. *Research on Virus Spreading in Multi Local World Complex Network* [J]. Computer Engineering, 2010, 36(23): p.130-132.（In Chinese)

[8] L. Y.Liu. *Link Prediction on Complex Networks* [J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): p.651-661.(In Chinese)

[9] K. Crammer, O. Dekel, J. Keshet. *Online passive-aggressive algorithms*[J]. The Journal of Machine Learning Research, 2006, (7): p.551-585.

[10] H. Kwak, C. Lee, H. Park. *What is Twitter, a social network or a news media?*[C]. ACM, 2010: p.591-600.