Predict post spreading in online social networks based on independent cascade model

O V Yakovleva[1,2], I V Rudakov[1] and Y V Stroganov[1]

[1]Bauman MSTU, ul. Baumanskaya 2-ya, 5/1, Moscow, 105005, Russian Federation

[2]E-mail: olg-199774@mail.ru

Abstract. Today, online social networks play an important role in the information spreading. Based on the social influence effect, posts can be spread across the network by repost mechanism. This paper addresses the problem of using independent cascade model for repost dynamics prediction in online social networks. A spreading post probability between two connected users is presented as function of preassigned features and coefficients fitted by the EM algorithm. This approach combines ideas of previously proposed methods. The list of considered features is composed and it includes users, their relationships and posts features. The iteratively solved by the Newton's method system of equation is derived for independent cascade model to obtain coefficients of a probability function. The experiments with the Sina Weibo dataset show that the EM algorithm more accurately estimates unknown coefficients of probability function that the logistic regression method. Since actual values of diffusion probabilities are unknown for a real network, a problem of their estimation is considered as a binary classification problem to determine accuracy of various methods. The independent cascade model tuned by proposed method of diffusion probability estimation was verified. The experimental results show than the model can commendably predict the reposting dynamics for a given post.

1. Introduction

An online social network (OSN) is a web-service that allows its users to create a personal account, explicitly connect to other users thus creating social relationships, and publish messages, called posts, to share various kinds of information, such as review of any product or service, political views, ideas, opinions, etc [1].

Based on the social influence effect, posts can be spread across the network through the principle of informational cascade [1] by repost mechanism. If a large amount of users has adopted a post, i.e. have made a repost of this post, it is considered viral. Modeling the post propagation process is useful for estimation whether the post will become viral. Also, it allows defining a group of users who will be interested in one or another content. In addition, users in OSN form communities [2] and such modeling can be helpful to detect information transmission paths from one community to another. Finally, information diffusion models are used to tackle the influence maximization problem, which asks to find a set of the most influential users in the network [3]. This analysis of OSN can be carried out for marketing, advertising and business purposes [4] or in the case of destructive information diffusion (promotion of radical ideas, suicidal content, drug distribution, incitement to ethnic and religious hatred, etc.) when a government needs to assess the situation [5,6].

2. Background and related work

An OSN can be formally represented by a directed graph $G(V, E)$, where a set of nodes $V =$

$\{v_1, ..., v_N\}$ are users and a set of edges E are relationships. If $(v_i, v_j) \in E$, the user v_j is called the follower of user v_i and the user v_i is called the friend of user v_j.

There are explanatory and predictive information diffusion models. Predictive models can be graph and non-graph based and proposed to predict spreading process in OSN by learning from past diffusion traces. The spreading process is characterized by two aspects: its structure, i.e. the spreading cascade that transcribes who reposted whom, and its temporal dynamics, i.e. the evolution of the diffusion rate which is defined as the amount of users that adopts the post over time [1].

This paper deals with the graph based Independent Cascade (IC) model [7]. In the IC model nodes can have two states: active (it means that the node has already adopted the post), and inactive (the node is unaware of the post or it is aware, but it did not adopt it). The process runs in discrete steps and starts from a set of initially activated nodes. In each step, an active node tries to influence each of its inactive followers. Regardless of its success, the same node will never get another chance to activate the same inactive follower. The success of node u in activating the node v depends on the propagation probability associated to the edge $e = (u, v)$. The process terminates when no further node gets activated.

There is the asynchronous extension of IC model named AsIC [8]. The difference between AsIC and IC models is that the first one unfolds a diffusion process along continuous time axis, and the second one is discrete. The AsIC model also requires a time-delay parameter on each edge of the graph.

Estimation of information diffusion probabilities for a real OSN is quite a challenge. In addition, verification of such methods also causes difficulties, since actual values of probabilities are unknown.

Saito *et al.* [7] proposed a discrete diffusion probability function, the values of which were learned by the Expectation Maximization (EM) algorithm. This method requires only a several past spreading cascades and does not need to calculate any node, edge or post features. It does not scale for large graphs since each edge must be present in training sample for accurate results. Therefore, this method can't be used in practice since real online social networks contain millions of users and billions of connections.

In the paper "Learning diffusion probability based on node attributes in social networks" [9], Saito *et al.* addressed the problem of learning parameters for AsIC model. A diffusion probability and a time-delay are considered as continuous functions of the *J*-dimensional vector of node attributes. This approach allows training the model in one small subgraph $G' \subset G$ and then predicting the spreading of new cascades in any other subgraph. According to the research of Saito *et al.* [7], objective function is introduced and maximized by the EM algorithm. However, authors only experimented with synthetic data and didn't provide a practical solution.

Guille *et al.* [10] also dealt with AsIC model. They solved the problem of estimating a diffusion probability as a classification problem for which the input data were semantic, social, and temporal features. That is, the diffusion probability is defined as the probability that the edge $e = (v,w)$ belongs to the "successful activation" class. To solve the classification problem, the logistic regression method is used. The advantages of Guille's approach include the simplicity and the high model training speed, since there is no need to consider all attempts of successful and unsuccessful activations in training cascades, as well as high prediction accuracy, confirmed by practical experiments. However, learning by logistic regression method contravenes the IC model assumption that any of the active friends could have influence on the newly activated node.

3. Notations

Let $G = (V, E)$ be a directed graph without self-links, where V and E ($\subset V \times V$) stand for the sets of all the nodes and edges, respectively. For each node $v \in V$, let $F(v)$ be the set of all the nodes that have edges from v, i.e., $F(v) = \{u \in V; (v,u) \in E\}$, and let $B(v)$ be the set of all the nodes that have edges to v, i.e., $B(v) = \{u \in V; (u,v) \in E\}$.

Now let's look at the diffusion process of some post. Let $D(t)$ be a set of nodes newly made active at time-step t, where $D(0)$ is a set of initially activated nodes. The diffusion episode is denoted by $D = D(0) \cup D(1) \cup ... \cup D(T)$.

Let $C(t)$ be a set of nodes made active by time-step t, i.e. $C(t) = \bigcup_{\tau \le t} D(\tau)$.

Next, let there be a training sample of S independent diffusion episodes. Then $D_s(t)$ equals $D(t)$ and $C_s(t)$ equals $C(t)$ for an episode with number $s = 1, ..., S$.

Let M be a set of all posts in OSN. A set of all posts written by user u is denoted by M_u. A set of posts written by user u and contained the mentions of other users is denoted by $\mathcal{D}_u \subset M_u$. Then, let M_u be a set of users who are mentioned in the posts of a user u and let $t\mathrm{M}^u$ be a set of all the posts which have mentioned a user u. Each post m is associated with the set of keywords K_m.

4. Proposed method

According to Guille's method [10], a diffusion probability of post m from newly activated node v to its still inactive follower w is denoted as a simple and smooth function of the feature vector and the coefficient vector as in equation (1). However, unlike Guille's method [10], the coefficient vector is learned by EM algorithm, according to Saito [7,9], not by logistic regression method. In addition, the list of features is changed relative to Guille's method [10].

$$p_{v,w,m} = p(\boldsymbol{x}_{v,w,m}, \boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T \boldsymbol{x}_{v,w,m})} \quad (1)$$

where $\boldsymbol{x}_{v,w,m}$ is the *(J+1)*-dimensional vector, first element of which is set to 1, i.e. $x_{v,w,m,0} = 1$, and others elements are composed of the *J* normalized features of user v, user w, relationship $e = (v,w) \in E$ and post m; $\boldsymbol{\theta}^T = (\theta_0, ..., \theta_J)$ is transposed coefficient vector, where $\theta_0 = const$.

4.1. Features

So it is proposed to consider the features of users (nodes), relationships (edges) and posts. The list of features described below was obtained by the following way. All features from Guille's method [10] were taken without changes, each of boolean type features was replaced on similar a real type one and several features were added by the authors. Then the best features were selected by the recursive feature elimination algorithm.

For user, five features are captured.

- User's activity as in equation (2). This feature characterizes user's ability to post and is computed as the average amount of posts, including reposts, emitted per hour.

$$I(u) = \begin{cases} \frac{|M_u|}{\Delta t}, & \text{if } |M_u| > 1 \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

where Δt is the difference between the earliest and latest posts' times measured in hours.

- The ratio of posts with mentions as in equation (3). This feature is taken without changes from the paper "A predictive model for the temporal dynamics of information diffusion in online social networks" [10] and provides an idea about the ability of a given user to distribute a content to other users.

$$dTR(u) = \begin{cases} \frac{|\mathcal{D}_u|}{|M_u|}, & \text{if } |M_u| > 0 \\ 0, & \text{Otherwise} \end{cases} \quad (3)$$

Last three user features characterize his/her popularity and socialization. They are determined by the capacities of appropriate sets.

- Number of user's followers.
- Number of user's friends.
- Number of posts which have mentioned a user.

For a relationship, i.e. for a pair of connected users, six features are considered:

- Social homogeneity by mentions as in equation (4). This feature is also taken without changes from the paper "A predictive model for the temporal dynamics of information diffusion in online social networks" [10] and for users u_1 and u_2 it reflects the similarity of the sets of users they talk to. It is computed with the Jaccard's coefficient.

$$HM(u_1,u_2) = \frac{|\mathrm{M}_{u_1} \cap \mathrm{M}_{u_2}|}{|\mathrm{M}_{u_1} \cup \mathrm{M}_{u_2}|} \tag{4}$$

- Homogeneity by posts. This feature reflects the similarity of users' interests. It is computed with the Jaccard's coefficient. There are three cases when post m_1 and post m_2 are equals: (i) m_1 is repost of m_2, (ii) m_2 is repost of m_1, (iii) m_1 and m_2 both are reposts of another post m_3.
- Social homogeneity by followers. This index describes homogeneity of users' social circles. It is computed with the Jaccard's coefficient.
- Social homogeneity by friends. This index is similar to the previous one, but it describes another aspect of users' social circles homogeneity.
- The mentioning rate of the second user by the first one as in equation (5). This feature is introduced instead of the similar boolean type one in "A predictive model for the temporal dynamics of information diffusion in online social networks" [10] and it reflects a measure of social relation between users.

$$rM(u_1,u_2) = \begin{cases} \dfrac{|\mathrm{M}_{u_1} \cap t\mathrm{M}^{u_2}|}{|\mathrm{M}_{u_1}|}, \ if \ \left|\mathrm{M}_{u_1}\right| > 0 \\ 0, \ \text{Otherwise} \end{cases} \tag{5}$$

- Correlation of temporal activity vectors as in equation (6). According to Guille's method [10] a day is considered as a partition of 6 blocks of 4 hours. So the fraction of tweets the user emitted during each block is computed and a 6-dimensional vector noted V is filling: $\sum_{i=0}^{5} V^i$. For two users coefficient of such vectors correlation shows whether users are online at the same time.

$$C = \sum_{i=0}^{5} V_{u_1}^i * V_{u_2}^i \tag{6}$$

Finally, one post feature is presented:

- User's interest in the topic of post as in equation (7). It is computed as the ratio of the number of user posts with at least one keyword of given post to the total number of user posts. This feature replaces the similar boolean type one described by Guille in "A predictive model for the temporal dynamics of information diffusion in online social networks" [10].

$$hK(u,m) = \begin{cases} \dfrac{|\{m: K_m \cap K_u \neq \varnothing\}|}{|M_u|}, \ \text{if } |M_u| > 0 \\ 0, \ \text{Otherwise} \end{cases} \tag{7}$$

where $K_u = \bigcup_{m \epsilon M_u} K_m$

Before fitting coefficients, the all feature values are normalized by maximum value in the training dataset. Normalization coefficients will be saved to divide on them subsequently computed when modeling features values.

4.2. Coefficient fitting method

According to Saito [7,9], the system of equation (8) can be derived for IC model. The method of learning coefficients is described below. Firstly, initial approximation of vector $\boldsymbol{\theta}$ is chosen. Then the next two steps repeat until convergence:

1. The values of probabilities $\hat{p}_{v,w,m_s}$ for each of posts m_s is computed by formula (1) for current approximation of vector $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$
2. Computed values $\hat{p}_{v,w,m_s}$ are substituted into equation (8) and the system equation (8) is solved by Newton's method for obtaining the next approximation of vector $\boldsymbol{\theta}$ as a result.

$$\sum_{s=1}^{S}\sum_{t=0}^{T_s-1}\sum_{v\in D_s(t)}\left(\sum_{w\in F(v)\cap D_s(t+1)}\left(\frac{\hat{p}_{v,w,m_s}}{\hat{P}_w^{(s)}}-p_{v,w,m_s}\right)\boldsymbol{x}_{v,w,m_s}-\sum_{w\in F(v)\setminus C_s(t+1)}p_{v,w,m_s}\boldsymbol{x}_{v,w,m_s}\right)=0 \quad (8)$$

where $p_{v,w,m_s}=p\left(\boldsymbol{x}_{v,w,m_s},\boldsymbol{\theta}\right)$ and $\hat{P}_w^{(s)}$ is calculated by the formula (9).

$$\hat{P}_w^{(s)}=1-\prod_{v\in B(w)\cap D_s(t)}(1-\hat{p}_{v,w,m_s}) \quad (9)$$

5. Experimental Evaluation

5.1. Dataset

For the experimental evaluation, the Weibo-Net-Tweet dataset gathered by Jing Zhang *et al.* [11] was used. The 65 diffusion episodes with sizes ranging from 50 to 600 reposts were selected from the total dataset. For each selected episode the spreading cascade was restored relies on the assumption that a user is influenced by a last friend who reposted the post before him. Then, the obtained cascades were split into training and testing ones in proportion 75:25, respectively.

The examples of unsuccessful activation attempts are also needed to train and test the model, therefore, for each spreadable post, author's followers who did not adopt this post were randomly selected. Their number was equal to the number of post adopters to equalize probabilities of making type I and type II errors (the ratio of successful and unsuccessful attempts was obtained experimentally). It is considered that a type I error ("false positive" finding) was made on time step t, if the probability computed by the formula (10) is more than 0.5, and the user w did not make a repost. Reciprocally, a type II error ("false negative" finding) is to see the probability less than 0.5 when the repost is made by the user w. The condition of equality of "false positive" and "false negative" finding probabilities must be observed to avoid underestimating or overestimating a diffusion episode size when modeling.

$$P_w=1-\prod_{v\in B(w)\cap D(t)}(1-p\left(\boldsymbol{x}_{v,w,m},\boldsymbol{\theta}\right)) \quad (10)$$

5.2. Accuracy estimate

Since actual values of diffusion probabilities in real social networks are unknown, to assess accuracy of various methods of learning coefficient vector $\boldsymbol{\theta}$ in equation (1), we will assume that a probability is determined correctly if neither a type I error nor a type II error was made. The results of testing the EM algorithm and the logistic regression method are illustrated in Table 1.

Table 1. Learning method accuracy with the training: testing ratio 75:25.

Learning method	Accuracy
EM-algorithm	0.75
Logistic regression	0.71

So model trained by the EM algorithm provides better accuracy than model trained by the logistic regression method. The accuracy of the logistic regression method is lower than obtained in "A predictive model for the temporal dynamics of information diffusion in online social networks" [10] which can be explained by incompleteness of the dataset used to compute the features.

5.3. Modeling diffusion process

For each of the 15 test posts, a spreading cascade was obtained by IC model. Results for two various

posts are shown on figure 1. In the case of first post, the number of reposts maximally increases in step 2 that means author's followers are mainly adopters. In the case of second post, only a minor proportion of adopters are influenced by author, but in the step 2 there are one or more influential users who made their followers adopt the post. In both cases predicted number of reposts is close to real one in each model time steps.

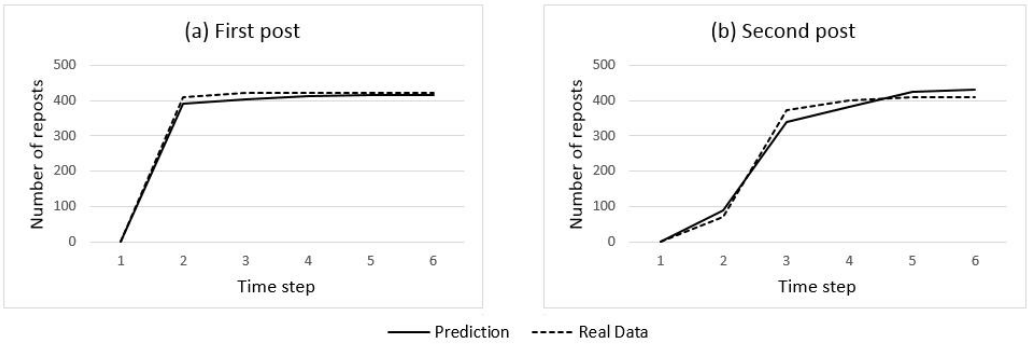

Figure 1. The number of reposts as a function of time step for two test posts.

6. Conclusion and future work

This paper addresses the problem of setting up and using IC model for obtaining a spreading cascade [1] of post in OSN. A method of learning diffusion probability is proposed. It combines the ideas of Guille [10] and Saito [9] methods. A diffusion probability is considered as a function of known features and unknown coefficients. The list of features of users, their relationships and post was composed. According to Saito [7,9], unknown coefficients are estimated with the EM algorithm. The experimental results validate the ability of IC model tuned by proposed method to correctly predict the reposting dynamics for a given post.

In future we plan to analyze post text [12] to increase an accuracy of estimation diffusion probabilities. In addition, we plan to develop a method of time-delay diffusion estimation and research the AsIC model. Finally, we need to compare an accuracy of the EM algorithm and the gradient boosting algorithm.

References

[1] Guille A, Hacid H, Favre C and Zighed D 2013 Information diffusion in online social networks: a survey *ACM SIGMOD Record* **42** pp 17-28
[2] Chesnokov V 2017 Overlapping community detection in social networks with node attributes by neighborhood influence *Models, Algorithms, and Technologies for Network Analysis* ed V Kalyagin, A Nikolaev, P Pardalos and O Prokopyev (Cham: Springer) vol 197 pp 187-203
[3] Yang W, Brenner L and Giua A 2019 Influence maximization in independent cascade networks based on activation probability computation *IEEE Access* **7** pp 13745-13757
[4] Chernova V, Tretyakova O and Vlasov A 2018 Brand marketing trends in russian social media *Media Watch* **9** pp 397-410
[5] Okhapkina E, Okhapkin V and Kazarin O 2017 The cardinality estimation of destructive information influence types in social networks *Proc. of 16th European Conf. on Cyber Warfare and Security* (Ireland: University Colledge Dublin) pp 282-287
[6] Okhapkina E, Tarasov A and Kazarin O 2017 Adaptation of information retrieval methods for identifying of destructive informational influence in social networks *Proc. of 31st Int. Conf. on Advanced Information Networking and Applications Workshops* (Taipei: Tamkang University) vol 1 pp 87-92
[7] Saito K, Nakano R and Kimura M 2008 Prediction of information diffusion probabilities for independent cascade model. *Knowledge-Based Intelligent Information and Engineering Systems* ed I Lovrek I, R J Howlett and L C Jain (Berlin, Heidelberg: Springer) pp 67-75
[8] Saito K, Ohara K, Kimura M and Motoda H 2013 Learning asynchronous-time information

diffusion models and its application to behavioral data analysis over social networks *J. Comput. Eng. and Inf.* **1** pp 30-57

[9] Saito K, Ohara K, Yamagishi Y, Kimura M and Motoda H 2011 Learning diffusion probability based on node attributes in social networks *Proc. of 19th Int. Symp. On Methodologies for Intelligent Systems* (Warsaw) pp 153-162

[10] Guille A and Hacid H 2012 A predictive model for the temporal dynamics of information diffusion in online social networks *Proc. Int. Workshop on Mining Social Network Dynamics* (Lyon) pp 1145-1152

[11] Zhang J, Liu B, Tang J, Chen T and Li J 2013 Social influence locality for modeling retweeting behaviors. *Proc. of 23st IJCAI* (Bejing) pp 2761-2767

[12] Kanev A, Cunningham S and Valery T 2017 Application of formal grammar in text mining and construction of an ontology *Proc. of 2017 Internet Technologies and Applications* (Wrexham) pp 53-57