Поширення інформації в соціальних мережах має безліч потенційних практичних застосувань, таких як онлайн-маркетинг, електронне державне управління та прогнозування великих соціальних подій. Тому моделювання поширення інформації є критично важливим завданням як для розуміння механізму поширення, так і для кращого управління ним. Метою даного дослідження є з'ясувати, які чинники можуть впливати на людей при прийнятті інформації, якою обмінюються в соціальній мережі. В даному дослідженні традиційна незалежна каскадна модель поширення інформації розширюється дискретними часовими кроками. Запропонована модель може включати три різних джерела впливу поширення: вплив користувача на користувача, користувацькі уподобання в контенті і зовнішній вплив. Зокрема, ці джерела впливу кількісно виражені в реальні значення ймовірності поширення. Щоб розрахувати вплив користувача на користувача, прийнято та розширено модель передачі зараження відповідно до ролі користувача, що поширює контент. Користувацькі уподобання в контенті, які вимірюють співвідношення між користувацькими уподобаннями і прийнятим контентом, розраховується на основі тематичної моделі. Зовнішній вплив виявляється на часовому кроці поширення, визначається кількісно і включено в нашу модель для наступного часового кроку поширення шляхом застосування і вирішення логістичної функції. Крім того, процес поширення інформації характеризується побудовою дерева прийняття інформації, а масштаб поширення визначається кількісно шляхом прогнозування кількості заражених вузлів. Встановлено, що ці джерела впливу, особливо зовнішній вплив, відіграють значну роль в поширенні інформації і в кінцевому підсумку впливають на форму і розмір каскаду поширення. Модель перевірена як на штучних, так і на реальних масивах даних. Експериментальні результати підтверджують перевагу запропонованого методу в порівнянні з попередніми моделями з точки зору точності прогнозування

Ключові слова: поширення інформації, соціальна мережа, незалежна каскадна модель, ймовірність поширення

UDC 004.08
DOI: 10.15587/1729-4061.2018.150295

A METHOD FOR DETERMINING INFORMATION DIFFUSION CASCADES ON SOCIAL NETWORKS

Nguyen Viet Anh
PhD, Senior researcher
Department of Data Science and Application
Institute of Information Technology
Vietnam Academy of Science and Technology
18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam, 100000
E-mail: anhnv@ioit.ac.vn

Duong Ngoc Son
Master of Science
Technical Department of Security
Ministry of Public Security
44 Yet Kieu, Hanoi, Vietnam, 100000

Nguyen Thi Thu Ha
PhD Lecturer
Department of Commerce
Vietnam Electric Power University
235 Hoang Quoc Viet, Hanoi, Vietnam, 100000
E-mail: bhhantt@epu.edu.vn

S. Kuznetsov
Doctor of Science, Professor, Chief Scientist
Department of Information Systems
Ivannikov Institute for System Programming of the RAS
Alexander Solzhenitsyn str., 25, Moscow, Russia, 109004
Professor
Departments of System Programming
Lomonosov Moscow State University
GSP-1, Leninskie Gory, Moscow, Russia, 119991
Moscow Institute of Physics and Technology
Institutskiy lane, Dolgoprudny, Russia, 141701
Higher School of Economics
Myasnitskaya str., 20, Moscow, Russia, 101000
E-mail: kuzloc@ispras.ru

Nguyen Tran Quoc Vinh
PhD, Dean
Faculty of Information Technology
The University of Da Nang –
University of Science and Education
459 Ton Duc Thang, Lien Chieu, Da Nang, Vietnam, 550000
E-mail: ntquocvinh@ued.udn.vn

1. Introduction

Online social networks have become one of the most efficient communication platforms over the last two decades with high socio-economic impacts. This fact has motivated a large amount of recent research. Different problems are currently studied, including network modelling, social network annotation, community detection, user recommendation,

link prediction, and information diffusion [1, 2]. Information diffusion is a key social network analysis with many potential real-world applications. For example, it can be used for predicting large social events such as the Arab Spring, for improving the recommendation performance of products and services, and for maximizing advertising effects to individuals. Today, it is possible to collect massive online social network data, which allows us to study information diffusion mechanisms in much finer details than ever before.

An information diffusion process occurs when a piece of information flows from one individual to another in a network. To understand the underlying structure of a diffusion, it is compulsory to construct the diffusion cascade, which requires knowing activation decisions of people and also the social ties, over which these activations occur. Actually, the problem has been studied in areas like epidemiology for decades [3]. However, the problem of information diffusion on online social networks has become much more complex because networks now are often very large, and moreover, there is a large diversity in users profiles and the uncertainty in their behaviors making the existing methods less accurate and efficient. Thus, there is a need for methods that better approximate the mechanism of information diffusion on social networks in a more efficient manner.

2. Literature review and problem statement

A lot of existing studies adopt the Linear Threshold model (LT model) [4–7] and Independent Cascade model (IC model) [8–10] to predict diffusion cascades. In the former, an acceptance threshold θ and an aggregation function f are associated with each user (or node). In the traditional version of the LT model, the function f is the sum of the weights of edges from active neighbours of a user v to v, and v becomes active if this sum exceeds the threshold θ. The IC model, on the other hand, uses activation probabilities for edges instead of acceptance thresholds for nodes. An active node independently tries to activate its inactive neighbours and succeeds with specified probabilities.

Some studies extend the LT model and IC model by considering more factors such as time decay and user profile. In [11], the authors introduce ASIC (Asynchronous IC), which models the influence between an active user and his/her neighbours using an exponential probability distribution based on the delay of influence. In [12], the ASIC model is further extended with user profile information. The method in [13] considers different probability distributions for the delay of the influence: exponential, power law and Rayleigh distributions. Several other extensions includes the topic sensitive IC model [14], influence models considering positive and negative opinions [15], influence model considering friend and foe relationships [16].

Some studies consider the cascade prediction task as a regression problem [17, 18] or a classification problem [19, 20]. These studies first identify a number of features that may correlate with the dependent variable that can be the cascade size or activation probabilities. After that, they learn a regression model or a classifier to estimate the value of the dependent variable.

Many other studies apply PageRank algorithm variations to rank user influence according to the network structure. Kwak et. al. [21] rank users by applying PageRank on follower following graph in Twitter. The problem of this approach is that network structure is relatively static compared to the activities of users in social networks. To deal with the problem, some studies have tried to include additional information in their models, for example, Topic-Sensitive PageRank [22] and TwitterRank [23] are able to compute per topic influence ranks.

Most of the previous studies aim to predict the volume of aggregate activation (or the size of the cascade), a closely related but a different task from the one addressed in this paper. The work [18] estimates the total number of up-votes on Digg stories. The work [24] estimates the total hourly volume of news phrases. The work [25] estimates total daily hash-tag use. Some work in this line of research estimates the cascade size through sampling [5, 26, 27]. These works often use a fixed number of samples to estimate the expected size of the cascade. The problem here is that there is no single sample size that fits all kinds of networks.

Another closely related task with the one addressed in this paper is the problem of influence maximization. The goal of influence maximization is to find a seed set of users who can activate cascades for maximizing the diffusion of information in the social network. Domingos and Richardson [4, 28] address this problem using Markov random fields. The work [5] models the problem as a discrete optimization problem and proved its NP-hardness for both the LT and IC models.

Most of the previous works studying the problem of influence maximization or predicting cascade size assume the diffusion probabilities between users are given as inputs. Instead, in this paper we address the problem of predicting these probabilities

3. The aim and objectives of the study

The aim of this study is to understand underlying information diffusion mechanisms on online social networks. More specifically, the study aims at predicting the adopting probability of a user when he/she is exposed to the content.

To accomplish the aim, the following objectives have been set:

– to carefully and thouroghly review related works regarding the information diffusion problem;

– to formally formulate the problem and other definitions;

– to solve the problem theoretically based on real-world observations and analysis;

– to prove the accuracy and efficiency of the proposed method experimentally by comparing with other state-of-the-art algorithms on real-world datasets.

4. The proposed method

Quantifying the probability that a user will adopt a piece of content on an online social network is a very challenging task. In general, the exact mechanisms driving users to take actions are unknown because they are diversified and vary according to individuals. Influence between users is a highly subjective area and the quantification of influence depends strongly upon domains. Moreover, an influential user in a particular domain cannot remain influential forever. This requires constant evaluation of their diffused behaviors and contents. In this paper, a method is proposed that allows

quantification of the probability that a user takes action with regard to a content being diffused on a social network. The method takes into account the history of user-user interaction and the user-content preference for every user in the network. The method also considers and quantifies external influence that will lead to a more accurate forecasting model of information diffusion.

4. 1. Problem Formulation

An online social network is often represented by a graph, where nodes are users and edges are relationships between users. The relationship can be either directed (one-directional or two-directional) or undirected. More precisely, it depends on whether the network allows connecting in an unilateral (e. g. following manner in Twitter) or bilateral (e. g. friendship manner in Facebook) model. In a social network, users often publish contents to share or forward various kinds of information, such as ideas, political opinions, etc.

Formally, we represent a social network as a graph $G=(V, E, T)$ where V is the set of nodes which represent users, E is the set of edges which represent relationships among users, $E=(u, v)|u,v\in V,\ u\neq v$ and $T: E\rightarrow N$ is a time labeling function which labels each edge with the time-stamp showing the time when the relationship between two users is created.

Given a content c that is being diffused over a social graph G, a node (a user) $u\in V$ may have one of the following states with regard to c. Node u is in the *unaware* state with regard to c if no friends of u are *activated* to c. Node u gets exposed (in *exposure* state) to c if some friends of u are activated to c and u is not. Node u is activated (in *active* state) to the content c if u does some social action (e.g. like, share, comment) associated to the content c.

Table 1

An example of action log

User		Content	Time
User	0	A	t0
User	1	A	t1
User	1	B	t2
User	2	A	t3
User	3	B	t4
User	3	C	t5

Now some constraints about the activation mechanism of the diffusion are given according to the IC model:

– when a node $u\in V$ becomes active, say at time t, then it has only one chance to activate each currently inactive neighbor v;

– active nodes are only temporarily influential (contagious) for some time steps; this means that once a node has made all its attempts, it becomes non-influential (non-contagious) but still remains active;

– this activation successfully happens with probability p_{uv}, independently of all other neighbors' attempts;

– in case more than one contagious node try to activate a same inactive node, the order in which they will make their attempts is random.

The triple $<u, c, t>$ is used to represent that user u is activated to a content c at time t. It is also confirmed in the study that the user u propagated c at time t. Using this notation, we are given an *action log* as a triple $H=(U, C, T)$ which contains users' historical activities over a social network. Table 1 gives an example of action log.

Problem Definition. Given graph G (V, E, T), content c, current time t, action log H, the problem is to estimate the probability that a node is activated to c at time t.

4. 2. Measuring Pair-wise User Influence

To measure user influence between users on a social network, we adopt and modify the disease transmission model given in [3]. Consider a pair of users who are connected, one of whom u is active and the other v is exposed to a content c. Suppose that the average rate of *content-transmission* from u to v is r_{uv}, and that the active user remains influential for a time π_u. According to disease transmission mechanism in [3], the probability $1-\tau_{uv}$ that the content will not be transmitted from u to v is:

$$1-\tau_{uv}=\lim_{\delta_t\rightarrow 0}(1-r_{uv}\delta_t)^{\pi_u/\delta_t}=e^{-r_{uv}\pi_u}, \quad (1)$$

and the probability of transmission is:

$$\tau_{uv}=1-e^{-r_{uv}\pi_u}. \quad (2)$$

The above equation applies for continuous time. In our model, discrete time steps rather than continuous time are used, in which case instead of taking the limit in Eq. (1) we simply set $\delta_t=1$, resulting in

$$\tau_{uv}=1-(1-r_{uv})^{\pi_u}, \quad (3)$$

where π_u is measured in time steps. In general, r_{uv} and π_u are not the same between users, so the probability of transmission is also different. Furthermore, the rate r_{uv} is not symmetric and thus the probability of transmission in either direction might not be the same. In this paper, for simplicity, it is assumed that π_u is identical for all users and its optimal value will be chosen from a set of concrete time steps, {1, 2,..., 10}.

In disease transmission research, the disease transmission rate is often drawn from appropriate distributions. In this paper, this approach is not followed. It is observed that the pairwise influence can be defined based on social ties and historical interactions between users, which are given in the action log H. Given two nodes u and v, we assume that the influence from u to v is only propagated through the edge (u, v) in G. We do not consider influence through intermediate nodes between u and v.

It can be seen that the transmission rate r_{uv} is proportional to the number of times pieces of content are successfully diffused form u to v. However, in the IC model, there may be other nodes that also try to activate v. Therefore, influence must be "shared" among these nodes each time v is activated. Let $C_{u\cdot v}$ be the set of contents that activated u since the time u and v becoming friend on the social network and let $C_{u\rightarrow v}\subseteq C_{u\cdot v}$ be the set of contents that activated both u and v and that u is activated before v. The sets $C_{u\cdot v}$ and $C_{u\rightarrow v}$ can be easily obtained by scanning the action log H. The average content-transmission rate from u to v is given by:

$$r_{uv}=\frac{\sum_{c\in C_{u\rightarrow v}}\frac{1}{|S_v^c|}}{|C_{u\cdot v}|}, \quad (4)$$

where S_v^c is the set of neighbor's nodes of v that become active before v regarding the content c. Eq. (4), however, does not differentiate between *primary* and *secondary* diffusions. Primary diffusion means that the node that diffused the content is also the node that created the content. Secondary diffusion means otherwise. It is observed that influence through primary diffusion and secondary diffusion between two nodes can be very different and can severely affect the accuracy of diffusion models. For example, on online social networks, many people willingly adopt information created by their friends but reluctantly adopt information created by a strange person. This reflects the fact that almost all diffusion cascades are very small containing merely friends of the source node and large cascades are extremely rare [29, 30].

The set $C_{u,v}$ mentioned above is divided into two subsets: $C_{u,v}^{\rho}$ is the set of contents created by the set of contents that created by u and $C_{u,v}^{\xi}$ contains the others. Similarly, $C_{u\to v}$ is divided into $C_{u\to v}^{\rho}$ and $C_{u\to v}^{\xi}$, where $C_{u\to v}^{\rho}$ is the set of contents that created by u and activated v, and $C_{u\to v}^{\xi}$ contains the others. The transmission rate from u to v is now given by:

$$r_{uv}=\begin{cases}\dfrac{\left|C_{u\to v}^{\rho}\right|}{\left|C_{u,v}^{\rho}\right|}, & \text{for primary diffusion from } u \text{ to } v,\\ \dfrac{\sum_{c\in C_{u\to v}^{\xi}}\dfrac{1}{\left|S_v^c\right|}}{\left|C_{u,v}^{\rho}\right|}, & \text{otherwise.}\end{cases} \quad (5)$$

Note that for primary diffusion the influence from u to v is not "shared" with any other nodes as reflected in Eq. (5).

4. 3. Measuring User-Content Preference

It is intuitive that personal characteristics or habits of people can be closely related to their behavior. This leads us to consider topic-based influence measure in our problem setting.

Given a set of topics $C=\{C_1, C_2, ..., C_m\}$, where C_i is the class label of the i-th topic. The probability that a content c belongs to topic C_i is determined by:

$$\delta_i^c = P(C_i)\prod_{k=1}^{n} P(x_k \mid C_i), \quad (6)$$

where δ_i^c is the probability that the content c belongs to topic C_i, $P(C_i)$ is the probability of the topic C_i, $P(x_k \mid C_i)$ is the probability that attribute x_k belongs to the topic C_i and that x_k is an attribute of content c. In natural language processing, the attributes of documents are usually determined by words or phrases, which have been pre-identified.

It is assumed in the study that the preferences (or interests) of a user for topics are independent. Based on events that a user v took action in the past, the interest level of v for the topic C_i can be measured, denoted by $\rho_{v,i}$, as:

$$\rho_{v,i} = 1-\prod_{h=1}^{l}(1-\delta_i^h), \quad (7)$$

where l is the number of contents that v has adopted, and δ_i^h is the probability that the h-th content, which activated v, belongs to the topic C_i as given in Eq. (6).

When a content c is being spread in a social network, the topic-based probability that a user v will be activated by c is measured by the similarity between the topic of c and the preference of v on that topic. The Cosine similarity measure is applied to determine the similarity between c and the preference of v on the topic of c, denoted by μ_v^c, as:

$$\mu_v^c = \frac{\sum_{i=0}^{m}\rho_{v,i}\delta_i^c}{\sqrt{\sum_{i=0}^{m}(\rho_{v,i})^2}\sqrt{\sum_{i=0}^{m}(\delta_i^c)^2}}, \quad (8)$$

where $\rho_{v,i}$ is the interest level of v for the topic C_i as given in Eq. (7).

To sum up, user-user influence and user-content preference are combined to get the probability, which indicates whether a user u activates a user v with regard to a content c as:

$$p_{uv} = \alpha(\tau_{uv}) + (1-\alpha)\mu_v^c, \quad (9)$$

where $\alpha \in [0,1]$, τ_{uv} is the influence measure of u on v, given in Eq. (5), and μ_v^c is given in Eq. (8) above. It is noted that, unlike user-content preference, influence measure between users is independent with the content (Fig. 1).

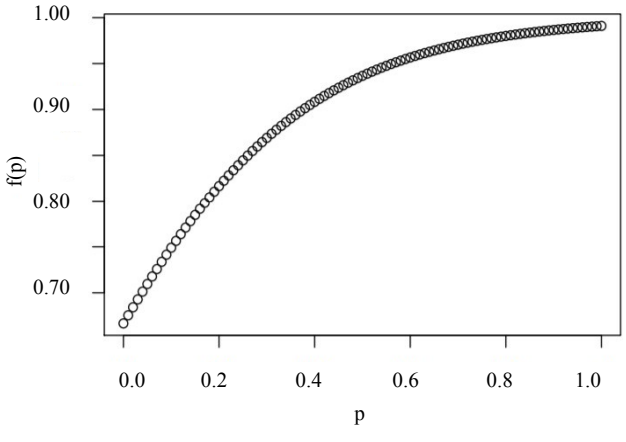

Fig. 1. Shape of the logistic function

4. 4. Measuring External Influence

Here in the model of this study, the effect of external influence in information diffusion is considered and quantified. Most of the existing works consider only internal influence among nodes and assume the absence of any other external information. This assumption, however, does not hold in general. For example, an external rumour or mass media like newspapers and televisions can easily reach people on social networks and eventually affect their actions toward an event or information. Basically, it is very hard to capture and study the effects of external influences let alone to forecast before it actually happens. This may be one of the reasons why until now there are very few studies dealing directly with this problem. The works [31, 32] also consider how external trends can affect their model, but the problem setting and the method used in their papers are very different from ours.

If the growing (virality) of a cascade is low, it may appeal only to a small group of closed people, and thus there may be no external influence or the influence is very small and can be neglected. On the contrary, if we witness fast growing stages of diffusion, it can be concluded that external influence does exist, and as reported in [18], their influence can still last even after a long time.

It is assumed that when external influence exists, every inactive node in the social network receives the same addi-

tional amount of influence, denoted as Δ_{ext}. The combined influence between two nodes u and v can be simply defined as the sum of p_{uv} and Δ_{ext}. However, this simple method has a problem: the sum value can be greater than 1. To force the value to fall between 0 and 1, it is mapped through a logistic function, which is bounded between 0 and 1 as:

$$f(p_{uv}+\Delta_{ext})=\frac{1}{1+A\exp^{-B(\rho_{uv}+\Delta_{ext})}}. \quad (10)$$

In this paper, A=1/2 and B=4 are chosen. The shape of the corresponding logistic function is shown in Fig. 1.

Suppose that in the diffusion time step t, we observed k new activated nodes. Our task now is to estimate Δ_{ext}. Let S be the set of contagious nodes, and let V_S be the set of nodes in V that exposed to S at a time step t. We denote by S_v the set of contagious neighbours of a node v. Since in the IC model, each of contagious neighbours tries to activate v independently, we define the activation function for each exposed node as:

$$\gamma(v)=\max_{u\in S_v} p_{uv}. \quad (11)$$

The expected number of activated nodes to the set S of contagious nodes is given by:

$$I(S)=\sum_{v\in V_S}\gamma(v). \quad (12)$$

Since we already know p_{uv}, and by setting $I(S)$=k, we can easily estimate the value of Δ_{ext} by solving Eq. (12).

4. 5. Diffusion tree construction

Here a method of constructing the diffusion tree for a particular piece of content is represented. Each node of the tree corresponds to a user who has adopted the content, and each edge links a user to another user called its "parent". Tree construction would be relatively straightforward if we knew exactly which user caused other users to adopt the information. Unfortunately, however, users adopt content using a variety of unknown mechanisms, which complicate the construction task.

The set of potential parents of a node v which adopted the content is denoted by S^v. The single most likely parent from the set of all potential parents of a given adoption is identified as the one that has the highest probability to influence v as shown in Eq. (13):

$$parent(v)=\arg\max_{u\in S^v}(p_{uv}). \quad (13)$$

After each adoption has been identified as either a root or a child of another node, the cascade of information diffusion is constructed in the form of a diffusion tree as shown in Algorithm 1.

Algorithm 1 generates the most likely diffusion tree (the tree with the highest probability). The probability of the generated tree T is given by:

$$P(T)=\prod_{(u,v)\in E(T)} p_{uv}. \quad (14)$$

Note that the probability of an exact tree is very small according to Eq. (14). However, it is still the tree that is closest to the "ground truth" – the real diffusion tree.

Algorithm 1 Diffusion Tree Construction

Given: graph G=(V, E, T), action log H, first active node v_0 at time t_0, tree T=∅

1: **set** node v_0 as the root of the tree T;
2: t=0;
3: **at each** time step $t \geq 0$ **do**;
4: $A_t \leftarrow$ set of active nodes at time t;
5: **for each** contagious node $u \in A_t$ **do**
6: $V_u \leftarrow$ set of inactive nodes $\in V$ which are neighbours of u;
7: **for each** node $v \in V_u$ **do**
8: **calculate** the diffusion probability p_{uv} from u to v according to Eq. (9);
9: **for each** newly active node v **do**
10: **find** the most likely parent u of v according to Eq. (13);
11: **make** a link from u to v;
12: **stop** if there are no more contagious nodes;
13: **Output** the diffusion tree T;

5. Experiments

In this section, the effectiveness of our proposed method for cascade prediction on online social networks is evaluated. The datasets are described, then evaluation metrics used for evaluation are defined, and finally evaluation results are discussed.

Datasets. Both synthetic and real-world datasets are used. For the synthetic dataset, the same method for generating the data as described in [33–35] is adopted. Specifically, Kronecker generator [36] is used to generate graphs which mimics the structural properties and the information diffusion traces in real-world networks [37]. The generated graphs are edge-directed, with core-periphery structure. 10 graphs with parameters [0.9 0.5; 0.5 0.3] are generated. The generated graphs consist of about 8,192 nodes and 25,600 edges. A K-dimensional topic distribution for each node of a graph is sampled from a K-dimensional symmetric Dirichlet distribution. This is done by assigning to each node j a uniformly distributed random variable $\theta_j \in (0, 1]^K$, which is the parameter of the Dirichlet distribution. Since the entries of θ are less than one, the generated contents are more focused on a small subset of topics. For cascade generation, content from the Dirichlet distribution is sampled first, then the discrete-time independent cascade is applied to generate a set of 5,000 cascades.

Two real-world datasets are also used. The first one is a large meme dataset [38], which traces the spread of memes across 1,700 popular media sites and blogs [39]. The dataset classifies memes per topic, and assigns each meme m to an information cascade t_m, which is a record of times when sites mentioned meme m. The second real-world dataset is named Tencent Weibo dataset, which is released by KDD Cup 2012 [40]. 4,000 and 1,000 cascades are used for training (learning diffusion probabilities) and testing, respectively, for all three datasets.

Metrics. It is very hard to evaluate diffusion models based on diffusion probabilities between users because we cannot compare estimated probabilities directly with exact probabilities. In previous works, various measures have been used to compare diffusion models. Many studies use measures like cascade size (or the number of users involved) [21, 41, 42]. Some studies use metrics with regard to shape

patterns like frequencies [21, 42] or correlations of shapes to events [21]. Metrics for temporal aspects like the time lag between messages [43] are also used. In this paper, we use the measure of cascade size to evaluate our proposed model. Given the diffusion probabilities between users, close approximations of exact cascade size by repeatedly simulating the cascade process and sampling the cascade size at each diffusion time step can be computed.

The performance of our proposed method with several existing ones is compared. The first model to compare, called IC model, is based on the Independent Cascade Model, which assigns the diffusion probability of a content to be simply the prior diffusion probability. The IC model does not consider user-user influences and content influences. For a pair of node u and v, the diffusion probability is calculated as $p_{uv}=1/d_{in}(v)$, where $d_{in}(v)$ denotes in-degree of node v, as in [44, 45]. The second model, called UI Model is a user interaction model, which considers the user-user influences.

Table 2

Comparing performance of methods in estimating diffusion size

Dataset	Model	Avg. Rel. Error (%) (estimated diffusion size)	Avg. Rel. Error (%) (most likely tree size)
Synthetic	IC	33.53	37.75
	UI	24.74	29.67
	RM	25.12	–
	Proposed Method	22.31	26.53
Meme	IC	35.56	39.21
	UI	27.34	31.67
	RM	27.65	–
	Proposed Method	22.15	27.22
Tencent Weibo	IC	36.42	40.32
	UI	28.46	32.92
	RM	29.27	–
	Proposed Method	24.19	29.37

We compare our method with the UI method given in [46], which measures the relatedness between nodes in a graph using the theory of random walk with restart. Finally, we compare our method with a regression method, called RM model, which estimate cascade size based on regressing on user-based features, content-based features, and time-based features [47].

To compare the methods, we adopt the relative error which shows how far the estimated diffusion size from the "ground truth". The relative error of estimated diffusion is computed as follows:

$$\left|\frac{\hat{I}(S)}{I(S)}-1\right|*100\,\%, \tag{15}$$

where $\hat{I}(S)$ is the estimated diffusion size of the seed set S by the method, and $I(S)$ is the ground truth for S. In our experiments, S has only 1 node.

The study also aims to evaluate the methods using the most likely generated diffusion tree (the tree with the highest probability). Because the RM model can only estimate the diffusion size, it is excluded from the test. For fair comparison, we prune the tree using a threshold θ; any edge having diffusion probability below θ will be discarded from the trees. In this experiment, it sets θ=50 %, and the size of the estimated trees is compared with that of the actual diffusion tree. The experimental results are reported in Table 2. In all experiments, the value of α in Eq. (9) is set to 0.5. As can be seen in the table, our proposed method outperforms all other methods in all datasets.

The experimental results above do not tell us about the effect of external trends affecting diffusion probabilities. External influence can speed up the process of information diffusion and make much larger cascades that other factors such as user-user and user-content interactions are unable to explain. It can be safely hypothesized that in small cascades, external trends have very small or even no influence. We examine the effect of including external influence into our model by using the top 5 % of the largest cascades from the Meme and Tencent Weibo datasets. The experimental results are given in Table 3.

As can be seen in Table 3, when estimating influence probabilities where external influence exists, existing methods like IC, UI and RM cannot take into account this factor and therefore, face with significant declination in prediction accuracy. The proposed method, on the contrary, still works well with just a slight drop in prediction accuracy.

Table 3

Comparing performance of methods in estimating diffusion size where external influence exists

Dataset	Model	Avg. Rel. Error (%) (estimated diffusion size)	Avg. Rel. Error (%) (most likely tree size)
Meme	IC	42.60	47.32
	UI	34.47	37.65
	RM	32.53	–
	Proposed Method	23.86	28.82
Tencent Weibo	IC	43.24	46.12
	UI	35.61	37.89
	RM	29.87	–
	Proposed Method	26.10	30.17

However, our method has a weakness. It estimates the external influence in a time step and assumes that the influence still affects and remains the same for the next time step. Therefore, it cannot be used for early prediction and requires constant updating of real-world diffusion data. To deeper analyse and to improve our method for external influence is a topic of interest for our future research.

6. Discussion of the research results of method

We have demonstrated the effectiveness of our method with extensive experiments on both artificial and real-world datasets. The good experimental results obtained in section 5 are due to the following facts:

1) The proposed model can corporate different sources of diffusion influence, namely, user-user influence, user-content preference, and external influence. As we all know, mechanisms that drive users to adopt information are very diversified. If we consider only one mechanism and neglect the others we will miss important information which can contribute greatly to better understanding of diffusion probabilities. Network topology, history of interaction activities,

content preference are among the techniques used in this paper to achieve our goal.

2) For user-user influence, our model can differentiate between two types of active nodes, namely the one that created the content and the one that just acts as intermediary. It is observed that influence through two types of nodes can be very different and can severely affect the accuracy of diffusion models. Despite its interest, no one to the best of our knowledge has considered and incorporated this fact in their work.

3) The proposed model can quantify external influence in information diffusion. This may be the most important contribution of our paper. Our method works in a sequential time-step manner. If a big viral is diagnosed in the current time step, its influence will diffuse to the next step. Most of existing works consider only internal influence among nodes and assume the absence of any other external information will face with difficulty when forecasting big information virals where external influence does exist.

By capturing different factors into the model, our proposed method can be easily extended to other social network analysis tasks like influence maximization, recommendation system, trust propagation, etc.

However, several challenges remain. The proposed model requires history interaction activities of users for computing user-user influence and user-content preference and therefore does not work well with users who are new in the social network. Secondly, the method for incorporating external influence can only work for the next time step if data of the previous step are known. Therefore, it cannot be used for early prediction and requires constant update of real-world diffusion data, a really hard task in practice. Moreover, in this paper, for the sake of simplicity and fast computation, some restrictions or suppositions have been made. This can hinder the accuracy of the prediction.

7. Conclusions

1. The problem of predicting diffusion probability on online social networks under the popular IC model is addressed. The proposed model incorporates user-user influence, user-content preference and external influence in a unified framework which ensures the capability of capturing the true mechanisms of information diffusion. To do so, the network topology, interaction activities among users, and interaction activities between users and diffused contents have been used.

2. Different properties of the information diffusion process are investigated and exploited in the proposed model. The efficient disease transmission model is adopted and extended to make accurate pair-wise influence prediction. Moreover, influence through different types of diffusing nodes is quantified. Topic-based analysis technique is used to characterize user habits on sharing favorite contents. Finally, external influence is diagnosed and quantified for better prediction of diffusion probabilities.

3. The learned diffusion probabilities, which are the outputs of the model, are used to construct the most likely diffusion tree and to estimate the size of diffused nodes. The diffusion tree can help to understand the diffusion process in a more intuitive and visualized way. The diffusion size or the scale of the information diffusion is especially useful for predicting important social events.

4. Extensive experiments on both artificial and real-world datasets have been conducted. It is shown experimentally that the proposed model improves the accuracy of predicting cascade size significantly in comparison with other state of the art methods. The experimental results confirm the advantages of the proposed method in predicting diffusion probabiliłties and show that the method works well in both cases when external influence exists and when not.

Acknowledgment

This research is funded by the project “Building a System for Prediction and Management of Information Spreading in Social Networks in Vietnam” under grant VAST01.01/17-18.

References

1. McCulloh I., Armstrong H., Johnson A. Social Network Analysis with Applications. Wiley Publishing, 2013.
2. David B. K., Alan G., Fernando J. V. Z. Online social network analysis: A survey of research applications in computer science. URL: https://arxiv.org/pdf/1504.05655.pdf
3. Newman M. E. J. Spread of epidemic disease on networks // Physical Review E. 2002. Vol. 66, Issue 1. doi: https://doi.org/10.1103/physreve.66.016128
4. Domingos P., Richardson M. Mining the network value of customers // Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining – KDD ‘01. doi: https://doi.org/10.1145/502512.502525
5. Kempe D., Kleinberg J., Tardos É. Maximizing the spread of influence through a social network // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining – KDD ‘03. 2003. doi: https://doi.org/10.1145/956755.956769
6. Topic Propagation Model Based on Diffusion Threshold in Blog Networks / Cao Y., Shao P., Li L., Cao Y. // 2011 International Conference on Business Computing and Global Informatization. 2011. doi: https://doi.org/10.1109/bcgin.2011.142
7. Analysis of information diffusion for threshold models on arbitrary networks / Lim S., Jung I., Lee S., Jung K. // The European Physical Journal B. 2015. Vol. 88, Issue 8. doi: https://doi.org/10.1140/epjb/e2015-60263-6
8. Saito K., Nakano R., Kimura M. Prediction of Information Diffusion Probabilities for Independent Cascade Model // International Conference on Knowledge-Based and Intelligent Information and Engineering Systems KES 2008: Knowledge-Based Intelligent Information and Engineering Systems. 2008. P. 67–75. doi: https://doi.org/10.1007/978-3-540-85567-5_9
9. Lee W., Kim J., Yu H. CT-IC: Continuously Activated and Time-Restricted Independent Cascade Model for Viral Marketing // 2012 IEEE 12th International Conference on Data Mining. 2012. doi: https://doi.org/10.1109/icdm.2012.40

10. Yang W., Brenner L., Giua A. Computation of Activation Probabilities in the Independent Cascade Model // 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT). 2018. doi: https://doi.org/10.1109/codit.2018.8394923
11. Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis / Saito K., Kimura M., Ohara K., Motoda H. // Asian Conference on Machine Learning ACML 2009: Advances in Machine Learning. 2009. P. 322–337. doi: https://doi.org/10.1007/978-3-642-05224-8_25
12. Learning Diffusion Probability Based on Node Attributes in Social Networks / Saito K., Ohara K., Yamagishi Y., Kimura M., Motoda H. // International Symposium on Methodologies for Intelligent Systems ISMIS 2011: Foundations of Intelligent Systems. 2011. P. 153–162. doi: https://doi.org/10.1007/978-3-642-21916-0_18
13. Gomez-Rodriguez M., Balduzzi D., Scholkopf B. Uncovering the temporal dynamics of diffusion networks // Proceedings of the 28 th International Conference on Machine Learning. Bellevue, 2011. URL: http://snap.stanford.edu/class/cs224w-readings/rodriguez11diffusion.pdf
14. Barbieri N., Bonchi F., Manco G. Topic-Aware Social Influence Propagation Models // 2012 IEEE 12th International Conference on Data Mining. 2012. doi: https://doi.org/10.1109/icdm.2012.122
15. Influence Maximization in Social Networks When Negative Opinions May Emerge and Propagate / Chen W., Collins A., Cummings R., Ke T., Liu Z., Rincon D. et. al. // Proceedings of the 2011 SIAM International Conference on Data Mining. 2011. P. 379–390. doi: https://doi.org/10.1137/1.9781611972818.33
16. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships / Li Y., Chen W., Wang Y., Zhang Z.-L. // Proceedings of the sixth ACM international conference on Web search and data mining – WSDM '13. 2013. doi: https://doi.org/10.1145/2433396.2433478
17. Bakshy E., Karrer B., Adamic L. A. Social influence and the diffusion of user-created content // Proceedings of the tenth ACM conference on Electronic commerce – EC '09. 2009. doi: https://doi.org/10.1145/1566374.1566421
18. Szabo G., Huberman B. A. Predicting the popularity of online content // Communications of the ACM. 2010. Vol. 53, Issue 8. P. 80. doi: https://doi.org/10.1145/1787234.1787254
19. Prediction of retweet cascade size over time / Kupavskii A., Ostroumova L., Umnov A., Usachev S., Serdyukov P., Gusev G., Kustarev A. // Proceedings of the 21st ACM international conference on Information and knowledge management – CIKM '12. 2012. doi: https://doi.org/10.1145/2396761.2398634
20. Jenders M., Kasneci G., Naumann F. Analyzing and predicting viral tweets // Proceedings of the 22nd International Conference on World Wide Web – WWW '13 Companion. doi: https://doi.org/10.1145/2487788.2488017
21. What is Twitter, a social network or a news media? / Kwak H., Lee C., Park H., Moon S. // Proceedings of the 19th international conference on World wide web – WWW '10. 2010. doi: https://doi.org/10.1145/1772690.1772751
22. Haveliwala T. H. Topic-sensitive PageRank // Proceedings of the eleventh international conference on World Wide Web – WWW '02. 2002. doi: https://doi.org/10.1145/511446.511513
23. TwitterRank / Weng J., Lim E.-P., Jiang J., He Q. // Proceedings of the third ACM international conference on Web search and data mining – WSDM '10. 2010. doi: https://doi.org/10.1145/1718487.1718520
24. Yang J., Leskovec J. Modeling Information Diffusion in Implicit Networks // 2010 IEEE International Conference on Data Mining. 2010. doi: https://doi.org/10.1109/icdm.2010.22
25. Ma Z., Sun A., Cong G. On predicting the popularity of newly emerging hashtags in Twitter // Journal of the American Society for Information Science and Technology. 2013. Vol. 64, Issue 7. P. 1399–1410. doi: https://doi.org/10.1002/asi.22844
26. Sketch-based Influence Maximization and Computation / Cohen E., Delling D., Pajor T., Werneck R. F. // Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management – CIKM '14. 2014. doi: https://doi.org/10.1145/2661829.2662077
27. Lucier B., Oren J., Singer Y. Influence at Scale // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '15. 2015. doi: https://doi.org/10.1145/2783258.2783334
28. Richardson M., Domingos P. Mining knowledge-sharing sites for viral marketing // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '02. 2002. doi: https://doi.org/10.1145/775047.775057
29. Can cascades be predicted? / Cheng J., Adamic L., Dow P. A., Kleinberg J. M., Leskovec J. // Proceedings of the 23rd international conference on World wide web – WWW '14. 2014. doi: https://doi.org/10.1145/2566486.2567997
30. Goel S., Watts D. J., Goldstein D. G. The structure of online diffusion networks // Proceedings of the 13th ACM Conference on Electronic Commerce – EC '12. 2012. doi: https://doi.org/10.1145/2229012.2229058
31. Myers S. A., Zhu C., Leskovec J. Information diffusion and external influence in networks // Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '12. 2012. doi: https://doi.org/10.1145/2339530.2339540
32. Wu D., Li C., Lau R. Y. K. Topic Based Information Diffusion Prediction Model with External Trends // 2015 IEEE 12th International Conference on e-Business Engineering. 2015. doi: https://doi.org/10.1109/icebe.2015.15
33. Scalable topic-specific influence analysis on microblogs / Bi B., Tian Y., Sismanis Y., Balmin A., Cho J. // Proceedings of the 7th ACM international conference on Web search and data mining – WSDM '14. 2014. doi: https://doi.org/10.1145/2556195.2556229
34. Uncover topic-sensitive information diffusion networks / Du N., Song L., Woo H., Zha H. // In AISTATS. 2013. P. 229–237.

35. Modeling cascade formation in Twitter amidst mentions and retweets / Pramanik S., Wang Q., Danisch M., Guillaume J.-L., Mitra B. // Social Network Analysis and Mining. 2017. Vol. 7, Issue 1. doi: https://doi.org/10.1007/s13278-017-0462-1
36. Kronecker graphs: An approach to modeling networks / Leskovec J., Chakrabarti D., Kleinberg J., Faloutsos C., Ghahramani Z. // Journal of Machine Learning Research. 2010. Issue 11. P. 985–1042.
37. Krongen: Kronecker graphs graph generator. URL: https://github.com/snap-stanford/snap/tree/master/examples/krongen
38. Structure and dynamics of information pathways in on-line media. URL: http://snap.stanford.edu/infopath
39. Gomez-Rodriguez M., Leskovec J., Krause A. Inferring Networks of Diffusion and Influence // ACM Transactions on Knowledge Discovery from Data. 2012. Vol. 5, Issue 4. P. 1–37. doi: https://doi.org/10.1145/2086737.2086741
40. KDD Cup 2012, Track 1. URL: https://www.kaggle.com/c/kddcup2012-track1/data
41. Everyone's an influencer / Bakshy E., Hofman J. M., Mason W. A., Watts D. J. // Proceedings of the fourth ACM international conference on Web search and data mining – WSDM '11. 2011. doi: https://doi.org/10.1145/1935826.1935845
42. Rise and fall patterns of information diffusion / Matsubara Y., Sakurai Y., Prakash B. A., Li L., Faloutsos C. // Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '12. 2012. doi: https://doi.org/10.1145/2339530.2339537
43. Information cascades in social media in response to a crisis / Hui C., Tyshchuk Y., Wallace W. A., Magdon-Ismail M., Goldberg M. // Proceedings of the 21st international conference companion on World Wide Web – WWW '12 Companion. 2012. doi: https://doi.org/10.1145/2187980.2188173
44. Chen W., Wang C., Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '10. 2010. doi: https://doi.org/10.1145/1835804.1835934
45. Tang Y., Shi Y., Xiao X. Influence Maximization in Near-Linear Time // Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data – SIGMOD '15. 2015. doi: https://doi.org/10.1145/2723372.2723734
46. Social influence locality for modeling retweeting behaviors / Zhang J., Liu B., Tang J., Chen T., Li J. // Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. 2013. P. 2761–2767.
47. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network / Suh B., Hong L., Pirolli P., Chi E. H. // 2010 IEEE Second International Conference on Social Computing. 2010. doi: https://doi.org/10.1109/socialcom.2010.33