This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2702162, IEEE Transactions on Knowledge and Data Engineering

Modeling Information Diffusion over Social Networks for Temporal Dynamic Prediction

Dong Li†, Zhiming Xu†, Yishu Luo†, Sheng Li†, Anika Gupta⋆
Katia Sycara⋆, Shengmei Luo‡, Lei Hu‡, Hong Chen‡
†Harbin Institute of Technology, China. {hitlidong, xuzm, lisheng}@hit.edu.cn, luo3380@gmail.com
⋆Carnegie Mellon University, USA. anikag@andrew.cmu.edu, katia@cs.cmu.edu
‡ZTE Corporation, China. {luo.shengmei, hu.lei2, chen.hong3}@zte.com.cn

ABSTRACT

How to model the process of information diffusion in social networks is a critical research task. Although numerous attempts have been made for this study, few of them can simulate and predict the temporal dynamics of the diffusion process. To address this problem, we propose a novel information diffusion model (GT model), which considers the users in network as intelligent agents. The agent jointly considers all his interacting neighbors and calculates the payoffs for his different choices to make strategic decision. We introduce the time factor into the user payoff, enabling the GT model to not only predict the behavior of a user but also to predict when he will perform the behavior. Both the global influence and social influence are explored in the time-dependent payoff calculation, where a new social influence representation method is designed to fully capture the temporal dynamic properties of social influence between users. Experimental results on Sina Weibo and Flickr validate the effectiveness of our methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: data mining

Keywords

Information diffusion; intelligent agents; model; prediction

1. INTRODUCTION

Information diffusion modeling over social networks is a critical and challenging task. Diffusion models are used to explain and simulate how information diffuses in a social network. They have a wide range of applications, including information recommendation, viral marketing, breaking news detection, and so on. The current studies on information diffusion modeling can be divided into two categories: theory-centric models and data-centric models.

Theory-centric models mainly come from epidemiology, sociology and economics. The most widely-studied diffusion models of this category are the epidemic model,the independent cascade model and the linear threshold model. These models are helpful for studying the information diffusion problems such as influence maximization problem [6, 4, 5]. However, they assume that the users in the network are passively influenced to spread information. Due to the lack of support from actual diffusion data, these models do not have the ability of diffusion prediction.

Data-centric models are usually learned from actual information diffusion data, and can be divided into macro-models and micro-models. Macro-models [9, 11] can generate diffusion cascades whose macro properties are similar to that of actual diffusion cascades. But they still can not predict the information diffusion process. This limitation is addressed by micro-models, which can predict whether a user in a social network will be activated by a piece of information. Since the information diffusion process is caused by user behavior, information diffusion prediction is actually user behavior prediction. Most micro-models [7, 10] can predict the behavior of a user, but they can not predict when the user will perform the behavior.

In this paper, we propose a novel information diffusion model (GT model) for temporal dynamic prediction. In contrast to traditional theory-centric models, the GT model regards the users in the network as intelligent agents. It can capture both the behavior of individual agent and the strategic interactions among these agents. By introducing the time-dependent payoffs, the GT model is able to predict the temporal dynamics of the information diffusion process. Different from most data-centric models, the GT model can not only predict whether a user will perform a behavior but also can predict when he will perform it. We make the following contributions in this work:

- We propose a novel information diffusion model (GT model), where, between different choices (behaviors), the user jointly considers all his interacting neighbors' choices to make strategic decisions that maximizes his payoff.
- We propose a time-dependent user payoff calculation method in the GT model by exploring both the global influence and social influence.
- We propose a new social influence representation method, which can accurately capture the temporal dynamic properties of social influence between users.
- We conduct experiments on Sina Weibo and Flickr datasets. The comparison results with closely related work indicate the superiority of the proposed GT model.

2. PROBLEM FORMULATION

A social network can be represented as $G = (V, E, T)$, where V is a set of $|V| = N$ number of users; E is the set

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507823 .

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2702162, IEEE Transactions on Knowledge and Data Engineering

of edges: a directed/undirected edge $(u, v) \in E$ represents a social tie between user u and user v; T is a function labeling each edge with the time when the social tie was created.

Definition 1. **Activation action:** An activation action can be represented as a triple (u, a, t_u), which can be interpreted as user u was activated by information a at time t_u. Let A_u be the set of all information user u adopts over all time. We record all the activation actions of all users as the action log $\Omega = \{(u, a, t_u)\}$.

Definition 2. **Information diffusion:** An information a diffuses from user u to user v iff: (i) $(u, v) \in E$; (ii) $\exists (u, a, t_u), (v, a, t_v) \in \Omega$ with $t_u < t_v$; and (iii) $T(u, v) < t_u$. We record as $diff(a, u, v, \Delta t)$, where $\Delta t = t_v - t_u$.

Definition 3. **Diffusion cascade:** For each information a, the diffusion cascade can be defined as $DC(a) = (V(a), E(a))$, where $V(a) = \{v|\exists t_v : (v, a, t_v) \in \Omega\}$ and $E(a) = \{\texttt{directed edge}(v_1, v_2)|diff(a, v_1, v_2, \Delta t)\}$.

Definition 4. **Global influence:** Given a social network, $global_v$ is used to denote the global influence of user v, indicating the influence capability of v over the whole network.

Definition 5. **Social influence:** Given two users u, v in a social network, we use $social_{uv}(t)$ to represent the influence strength of user u on user v at time t.

Since the influence strength of user u on user v varies with time, introducing the time variable t can give more accurate description of social influence.

Based on the concepts described above, we present the following problem:

Problem 1. **User payoff learning:** Given a social network G and an action log Ω, a critical task of our work is to learn the user's time-dependent payoffs for his different choices.

In general, a user's payoff contains two parts: individual payoff from his idiosyncratic preferences and social payoff from his social contacts. In this work, we mainly focus on how to learn the social payoff. We introduce the time factor into the payoff since one user may get different payoff when he adopts his friend's behavior at different time. This time-dependent user payoff enables the GT model to predict the temporal dynamics of information diffusion process.

3. THE PROPOSED MODEL

In the proposed GT model, the diffusion process unfolds in discrete time-steps t, and begins from a given initial active user set. When a user v observes a piece of information at time t, he calculates his payoffs for different choices depending on his neighbors' status so as to make strategic decision. If he adopts the information, his status becomes activated at time $t+1$. We next describe the proposed model in detail.

In a social network, we first consider the simplest situation in which a user has two possible choices, A and B, when he observes a piece of information. As an example, we can imagine the information is a tweet in Twitter, choice A is retweeting the tweet and choice B is not. For a user v facing his one neighbor u, the payoffs of his different choices are defined as:

If u and v both choose A, v gets payoff $a_{uv}(\Delta t)$.
If u and v both choose B, v gets payoff $b_{uv}(\Delta t)$.
If u choose A and v choose B, v gets payoff $c_{uv}(\Delta t)$.
If u choose B and v choose A, v gets payoff $d_{uv}(\Delta t)$.

$\Delta t = t_u - t_v$ denotes the time delay between user u and v making the choice. Based on these different choices of u and v, a payoff matrix of user v is generated as shown in Fig.1(a).

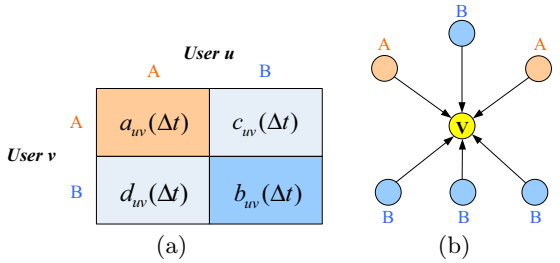


Figure 1: (a) is the payoff matrix of user v; (b)shows user v makes choice between A and B depending on all of its neighbors' choices.

Fig.1(a) is the situation on a single edge in the network, i.e. only considers one neighbor of user v. In general, the choice of v depends on all of his neighbors' choices as shown in Fig.1(b). The total payoff of v is the sum of all individual payoffs that he gets when faces each single neighbor. Here we use $N_A(v)$ to denote the set of v's neighbors who adopt choice A, and $N_B(v)$ the set of neighbors who adopt choice B. If user v adopts choice A at time t_v, he will get payoff:

$$payoff_A(v, t_v) = \Sigma_{u \in N_A(v)} a_{uv}(t_v - t_u) + \Sigma_{u \in N_B(v)} c_{uv}(t_v - t_u) \quad (1)$$

Similarly, if v adopts choice B at time t_v, he will get payoff:

$$payoff_B(v, t_v) = \Sigma_{u \in N_A(v)} d_{uv}(t_v - t_u) + \Sigma_{u \in N_B(v)} b_{uv}(t_v - t_u) \quad (2)$$

Finally, between these two choices, user v will make the decision that maximizes his payoff. Therefore, if $payoff_A(v, t_v) \geq payoff_B(v, t_v)$, user v will make choice of A at time t_v, else he will adopt choice B.

In the following, we will present a method to calculate user payoffs of his different choices in different situations, i.e. the payoff matrix in Fig.1(a). Considering two linked users, u and v, an intuitive explanation of our method is: the more payoff user v has gotten in the past by following user u's choices, the greater tendency v will have (can be regarded as: the more payoff he will get) at this time to make the same choice as user u. Based on this concept, we explore both the global influence and social influence for the payoff calculation. Global influence shows the authority of a user while social influence reflects the degree to which one user has affected another. Specifically, the greater global influence user u has and the greater social influence shown between user u and user v, the more payoff user v will get if he makes the same choice as user u. This calculation method is not limited by any specific diffusing information, thus it is applicable to different diffusing information in different social networks. Based on the description above, we define the payoff matrix of user v facing his neighbor u as

$$\begin{cases} c_{uv}(\Delta t) = d_{uv}(\Delta t) = 0 \\ a_{uv}(\Delta t) = b_{uv}(\Delta t) = global_u * social_{uv}(\Delta t) \end{cases} \quad (3)$$

We can see, if user v adopts the behavior different from user u, he gets no payoff; else if user v adopts the same behavior as u, the global influence and social influence are jointly explored to measure his payoff. Next, we will present a new social influence calculation method which can fully capture the temporal dynamics of social influence between users.

Many efforts have been made for the research of social influence. However, only the CT Model and DT Model proposed by [8] consider the time factor. CT model describes the social influence by an exponential decay function. It has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2702162, IEEE Transactions on Knowledge and Data Engineering

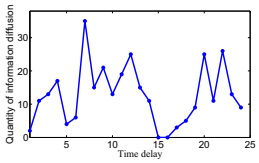


Figure 2: An example of social influence between two users on Sina Weibo dataset.

a significant drawback in that it assumes that the social influence follows an exponential distribution. DT Models set the influence at a constant value within a time window. Due to the rough simulating mechanisms, it is difficult for the CT and DT models to capture the complex dynamics of social influence between users. In this work, we propose a new method for accurate representation of the social influence. We represent the social influence function as a non-negative vector with length K, where the kth component $social_{uv}(k)$ denotes the social influence of user u on his neighbor v at time k and is defined as

$$social_{uv}(k) = \frac{|\{a|\exists \Delta t : diff(a,u,v,\Delta t) \wedge k-1 \leq \Delta t \leq k\}|}{|A_u|} \quad (4)$$

Fig.2 shows an example of the social influence represented by our method on Sina Weibo dataset. We can see that, in contrast with the CT and DT model, our proposed method can accurately capture the temporal dynamic properties of social influence between users.

Finally, we highlight that the GT model is not only applicable to the situation with two choices on one piece of information, but can also deal with the situations with more choices on multiple pieces of information.

4. ALGORITHMS

4.1 Learning Algorithms

Global Influence. In this work, we use two methods to learn the global influence of individual user:

1) Pagerank Algorithm. In [2], the pagerank algorithm is used to calculate the importance of web pages based purely on the link structure of the World Wide Web. Here, we employ it to the social networks for global influence calculation.

2) Diffusion Cascades. Diffusion cascades triggered by a piece of information that a user adopts directly indicates this user's global influence. Therefore, it is reasonable to use the average size of diffusion cascades [1] as the measurement.

Social Influence. In order to calculate the social influence, we use our proposed method as described in Section 3, which considers the influence function as a non-negative vector with length K:

$$(social_{uv}(1), social_{uv}(2), \cdots, social_{uv}(K))$$

Algorithm 1 illustrates how to calculate the element $social_{uv}(k)$ based on the diffusion cascades. For the parameter K, different value is assigned for different dataset by statistical methods which will be discussed in detail in Section 5.

4.2 Information Prediction Algorithm

Based on the GT model proposed in Section 3, we present the Algorithm 2 for predicting the information diffusion process. This algorithm focuses on the question of whether a user will perform a behavior at time t. For a user u, if he has already performed the behavior, we assume that u is active; if he has not performed the behavior yet but at least one of his neighbors did, we assume that u is inactive.

Algorithm 1 Social influence calculation.

For each social link (u,v) do
  For $k=1$ to K do
    $social_{uv}(k) = count_{uv}(k) = 0$;
  End
End
For each $diff(a,u,v,\Delta t) \in$ diffusion cascades do
  If $k-1 < \Delta t < k$ Then $count_{uv}(k)++$;
End
For each social link (u,v) do
  For $k=1$ to K do
    $social_{uv}(k) = count_{uv}(k)/|A_u|$;
  End
End

Algorithm 2 Information diffusion prediction.

For each information a in testing dataset do
  For each inactive user v
    $Active\_payoff(v) = 0$;
    $Inactive\_payoff(v) = 0$;
    For each link related with v, (u,v) do
      If u is active do
        $k = [t - t^u_{active}]$;
        $Active\_payoff(v) = Active\_payoff(v) + social_{uv}(k) * personal_u$;
      End
      If u is inactive do
        $k = [t - t^u_{inactive}]$;
        $Inactive\_payoff(v) = Inactive\_payoff(v) + social_{uv}(k) * personal_u$;
      End
    End
    If $Active\_payoff(v) \geq Inactive\_payoff(v)$
    Then v is active;
    Else v is inactive;
  End
End

5. EXPERIMENTS

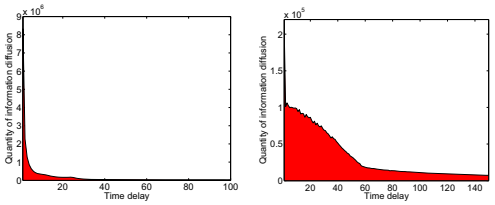
5.1 Experimental Setup

Datasets. Given the social network and action log, we evaluate the proposed model on two datasets.

- **Sina Weibo.** Sina Weibo is the Twitter of China, and now has 400 million users. Firstly, we crawled 251639 users and 4359915 edges from Sina Weibo. Then, we collected about 36 million micorblogs published by these 251639 users from 11/07/2011 to 11/28/2011. Considering the properties of Sina Weibo dataset, we set one hour as a time step.
- **Flickr.** This dataset is collected by [3]. The authors collected a total of 2.5 million users and 33 million links. They also collected 34 million favorite-markings behavior information over 11 million photos. In Flickr dataset, we regard 3 days as a time step.

For different datasets, the maximum diffusion time delay K adopts different values in social influence function. Fig.3(a) and (b) show the quantity of information diffusion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2702162, IEEE Transactions on Knowledge and Data Engineering



(a) Sina Weibo dataset (b) Flickr dataset

Figure 3: Distributions of information diffusion quantity over time delay.

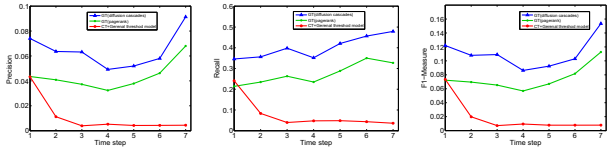


Figure 4: Prediction performances on Weibo dataset.

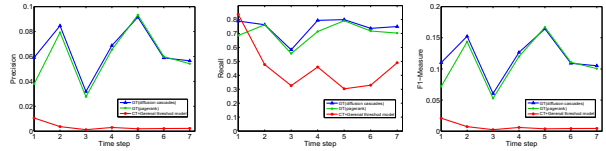


Figure 5: Prediction performances on Flickr dataset.

over time delay on Sina Weibo dataset and Flickr dataset. Both of them have a long-tail shape. In Sina Weibo dataset, 81.5% of diffusion actions are performed with the time delay less than 24 hours, so we set the parameter K at 24 (24 hours/1 hour). In Flickr dataset, 85.0% of diffusion actions are performed with the time delay less than 90 days, so we set the parameter K at 30 (90 days/3 days).

Baseline methods. We compare the proposed GT model with the closest work of [8], where two time-dependent models, CT model and DT model, are presented for capturing social influence, then they are applied together with general threshold model to predict time-dependent information diffusion. Since in [8], the CT model got better prediction performance than DT model, we here compare our proposed method with the method that combines CT model and general threshold model. For our method, we employ two kinds of methods to calculate user's global influence, pagerank and diffusion cascades, which achieve different accuracies.

Evaluation. We adopt three measurements to evaluate these tested prediction methods, which are Precision, Recall and F1-Measure.

5.2 Prediction Performance

For each information a in testing dataset, given its diffusion progress before time t $(0 : t-1)$, our goal is to predict which users will be activated by this information at time t. In our experiment, we assume that the user observed the behaviors of all his neighbors for each information.

Fig.4 and Fig.5 show the prediction performances of all the tested approaches under different measurements at 7 time steps (time 2-8) on Sina Weibo and Flickr dataset. We can see that the proposed GT model (either using pagerank or diffusion cascades for global influence calculation) can consistently achieve better performance comparing with baseline method [8]. The baseline method highly depends on the activation threshold of users which are hard to set. A same activation threshold value is assigned for all users in their work while in fact different users have different activation thresholds. Therefore, the predicting performances of [8] are uncompetitive. In contrast, our model, strategically considering all the interacting users, improves the performance dramatically.

Furthermore, we can also see that the GT model using diffusion cascades for global influence calculation achieves better prediction performance than that using pagerank. This is mainly because the pagerank method only analyzes the topology structure of network while diffusion cascades are mined from both the network structure and user behaviors, so using diffusion cascades method can get more accurate influence value than using pagerank method. These results illustrate that when our model is fed with more accurate parameters, it shows better performance in prediction task. This is a good sanity check of our model.

As shown in Fig.4 and Fig.5, the curves of the baseline method under three measurements are decreasing with time, demonstrating the loss of the prediction ability as time goes on. In contrast, our model achieves pretty better and time-independent performance. This is because in our model the prediction for a user's behavior doesn't rely on any of his neighbor's activation time. The GT model combines both his active and inactive neighbors to measure his payoffs of different choices and then gives a reliable prediction.

6. CONCLUSION

We have presented a novel information diffusion model in this paper. It regards the users in a social network as intelligent agents, and jointly considers all the interacting users to make strategic prediction. By introducing the time-dependent payoffs, the model has the capability to predict the temporal dynamics of information diffusion process. Both the global influence and social influence are explored for user payoff calculation, where the social influence representation method is newly designed for fully capturing its temporal dynamics. Experimental results have confirmed the rationality and effectiveness of the proposed model.

Acknowledgements. This work is supported by the Natural Science Foundation of China (No. 61173074), the ZTE cooperation project (No. MH20120428) and the ARO MURI Award Number W911NF0810301.

7. REFERENCES

[1] E. Bakshy, J. Hofman, W. Mason, and et al. Everyone's an influencer: Quantifying influence on twitter. In *WSDM*, pages 65–74, 2011.
[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.
[3] M. Cha, A. Mislove, and P. K. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, 2009.
[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, pages 1029–1038, 2010.
[5] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD*, pages 199–208, 2009.
[6] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.
[7] H. Fei, R. Jiang, Y. Yang, and et al. Content based social behavior prediction: a multi-task learning approach. In *CIKM*, pages 995–1000, 2011.
[8] A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, 2010.
[9] J. Leskovec, M. McGlohon, C. Faloutsos, and et al. Cascading behavior in large blog graphs. In *SDM*, pages 202–209, 2007.
[10] L. Liu, J. Tang, J. Han, and et al. Mining topic-level influence in heterogeneous networks. In *CIKM*, pages 199–208, 2010.
[11] D. Wang, Z. Wen, H. Tong, and et al. Information spreading in context. In *WWW*, pages 735–744, 2011.