Prediction of Information Diffusion Probabilities for Independent Cascade Model

Kazumi Saito[1], Ryohei Nakano[2], and Masahiro Kimura[3]

[1] School of Administration and Informatics, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
k-saito@u-shizuoka-ken.ac.jp
[2] Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466–8555 Japan
nakano@ics.nitech.ac.jp
[3] Department of Electronics and Informatics, Ryukoku University
Otsu, Shiga 520-2194, Japan
kimura@rins.ryukoku.ac.jp

Abstract. We address a problem of predicting diffusion probabilities in complex networks. As one approach to this problem, we focus on the independent cascade (IC) model, and define the likelihood for information diffusion episodes, where an episode means a sequence of newly active nodes. Then, we present a method for predicting diffusion probabilities by using the EM algorithm. Our experiments using a real network data set show the proposed method works well.

1 Introduction

Recently, attention has been devoted to investigating complex networks such as social, computer and biochemical networks [3,18,15,13,16]. A network can play an important role as a medium for the spread of various information [20,15]. For example, innovation, hot topics and even malicious rumors can propagate through social networks among individuals, and computer viruses can diffuse through email networks. Widely-used fundamental probabilistic models of information diffusion through networks are the *independent cascade (IC) model* [8,10,9]. The IC models can also be identified with the so-called *susceptible/infective/recoverd (SIR) model* for the spread of disease in a network [15]. Here we consider information diffusion phenomena in a given network on basis of the IC model.

The IC model needs to be provided with some adequate parameter values in advance. More specifically, the *diffusion probability* through each link in the network must be specified for the IC model in advance. However, it is usually difficult to know the diffusion probabilities through links for any real network in advance. Therefore, it is an important research issue to infer the diffusion probabilities through links from an observed data set of information diffusion.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part III, LNAI 5179, pp. 67–75, 2008.

2 Proposed Method

In this section, after explaining some preliminaries, we formalize a problem for estimating probabilities of information diffusion using a data set obtained from a directed network. Then we propose a method for estimating information diffusion probabilities.

2.1 Preliminaries

For a given directed network (or equivalently graph) $G = (V, E)$, let V be a set of nodes (or vertices) and E a set of links (or edges), where we denote each link by $e = (v, w) \in E$ and $v \neq w$, meaning there exists a directed link from a node v to a node w. For each node v in the network G, we define $F(v)$ as a set of child nodes of v as follows:

$$F(v) = \{w : (v, w) \in E\}. \tag{1}$$

Similarly, we define $B(v)$ as a set of parent nodes of v as follows:

$$B(v) = \{u : (u, v) \in E\}. \tag{2}$$

We introduce the IC model [8,10,9]. In this model, for each directed link $e = (v, w)$, we specify a real value $\kappa_{v,w}$ with $0 < \kappa_{v,w} < 1$ in advance. Here $\kappa_{v,w}$ is referred to as the *diffusion probability* through link (v, w). The diffusion process proceeds from a given initial active set $D(0)$ in the following way. When a node $v(\in D(t))$ first becomes active at time-step t, it is given a single chance to activate each currently inactive child node w, and the attempt succeeds with probability $\kappa_{v,w}$. If v succeeds, then w becomes active at time-step $t + 1$, i.e., $w(\in D(t + 1))$. If multiple parent nodes of w first become active at time-step t, then their activation attempts are sequenced in an arbitrary order, but all the attempts are performed at time-step t. Whether or not v succeeds, v will not make any further attempts to activate w in subsequent rounds. The process terminates if no more activations are possible.

2.2 Problem Formulation

Let an information diffusion episode $D = D(0) \cup D(1) \cup \cdots \cup D(T)$ be the union of ordered sets, where a set $D(t)$ consists of nodes newly made active at time-step t. For some link $e = (v, w)$, where $v \in D(t)$ and $w \in D(t + 1)$, it is probable that the node v succeeded in activating the node w at time-step t through the link e. Since we should consider other possibilities that some other nodes $v' \in D(t) \cap B(w)$ also succeeded in activating the node w, we calculate the probability that the node w becomes active at time-step $t + 1$ as follows:

$$P_w(t + 1) = 1 - \prod_{v \in B(w) \cap D(t)} (1 - \kappa_{v,w}). \tag{3}$$

Here note that if $w \in D(t + 1)$, it is guaranteed that $D(t) \cap B(w) \neq \emptyset$.

Next, let $C(t)$ denote a set of nodes made active by time-step t, i.e., $C(t) = \cup_{\tau \leq t} D(\tau)$. In the case that $v \in D(t)$ and $w \notin C(t+1)$, we know that the node v definitely failed in activating the node w through the link e. Clearly, in the case that (a) $v \in D(t)$ and $w \in C(t)$, or (b) $v \notin D$, we cannot obtain any information about the attempt with respect to the link $e = (v, w)$. Therefore, for an episode D we can define the following likelihood function with respect to $\boldsymbol{\theta} = \{\kappa_{v,w}\}$:

$$L(\boldsymbol{\theta}; D) = \left(\prod_{t=0}^{T-1} \prod_{w \in D(t+1)} P_w(t+1) \right) \left(\prod_{t=0}^{T-1} \prod_{v \in D(t)} \prod_{w \in F(v) \setminus C(t+1)} (1 - \kappa_{v,w}) \right) \quad (4)$$

Let $\{D_s : s = 1, \cdots, S\}$ be a set of S independent information diffusion episodes. Then we can define the following objective function with respect to $\boldsymbol{\theta}$.

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \log L(\boldsymbol{\theta}; D_s) \\ &= \sum_{s=1}^{S} \sum_{t=0}^{T_s - 1} \left(\sum_{w \in D_s(t+1)} \log P_w^{(s)} + \sum_{v \in D_s(t)} \sum_{w \in F(v) \setminus C_s(t+1)} \log(1 - \kappa_{v,w}) \right) \quad (5) \\ P_w^{(s)} &= 1 - \prod_{v \in B(w) \cap D_s(t_w^{(s)} - 1)} (1 - \kappa_{v,w}) \quad (6) \end{aligned}$$

Here $P_w^{(s)}$ stands for the probability that a node w becomes active at time-step $t_w^{(s)} (= t+1)$ in an episode D_s. Given D_s, the time-step $t_w^{(s)} (= t+1)$ when w becomes active is known, and is omitted in representing $P_w^{(s)}$. Then, our problem is to obtain the set of information diffusion probabilities, $\boldsymbol{\theta} = \{\kappa_{v,w}\}$, which maximizes Eq. (5).

2.3 Learning Method

Since it is rather hard to analytically maximize Eq. (5), we apply the Expectation-Maximization (EM) algorithm in order to obtain its solution.

For a link (v, w) in an episode D_s where $v \in D_s(t)$, we know an activation attempt through the link (v, w) was surely performed. If $w \in D_s(t+1)$, then the attempt through (v, w) succeeded with the probability $\widehat{\kappa}_{v,w} / \widehat{P}_w^{(s)}$, and failed with the probability $1 - (\widehat{\kappa}_{v,w} / \widehat{P}_w^{(s)})$. Here $\widehat{\kappa}_{v,w}$ stands for a current estimate of $\kappa_{v,w}$, and the value $\widehat{P}_w^{(s)}$ is calculated by using Eq. (6) and current estimates $\widehat{\boldsymbol{\theta}} = \{\widehat{\kappa}_{v,w}\}$. On the other hand, if $w \notin C_s(t+1)$, then the attempt through (v, w) failed with no doubt. Considering these cases, we have the following Q-function for S episodes.

$$\begin{aligned} Q(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}) = \sum_{s=1}^{S} \sum_{t=0}^{T_s - 1} \sum_{v \in D_s(t)} \Bigg(& \sum_{w \in F(v) \cap D_s(t+1)} \left(\frac{\widehat{\kappa}_{v,w}}{\widehat{P}_w^{(s)}} \log \kappa_{v,w} + (1 - \frac{\widehat{\kappa}_{v,w}}{\widehat{P}_w^{(s)}}) \log(1 - \kappa_{v,w}) \right) \\ & + \sum_{w \in F(v) \setminus C_s(t+1)} \log(1 - \kappa_{v,w}) \Bigg) \end{aligned} \quad (7)$$

By solving the optimality condition $\partial Q/\partial \kappa_{v,w} = 0$, we have the following new estimate of $\kappa_{v,w}$.

$$\kappa_{v,w} = \frac{1}{|S^+_{v,w}| + |S^-_{v,w}|} \sum_{s \in S^+_{v,w}} \frac{\widehat{\kappa}_{v,w}}{\widehat{P}^{(s)}_w} \tag{8}$$

Here $S^+_{v,w}$ indicates a set of episodes, each of which satisfies both $v \in D_s(t)$ and $w \in D_s(t+1)$ for the link (v,w), while $S^-_{v,w}$ denotes a set of episodes, each of which satisfies both $v \in D_s(t)$ and $w \notin D_s(t+1)$ for the link (v,w). Moreover, $|S|$ indicates the number of elements of a set S.

We can summarize our learning method based on the EM algorithm. The method repeats the following two EM steps until convergence.

step 1. Estimate each success probability $\widehat{P}^{(s)}_w$ by using Eq. (6).
step 2. Update each diffusion probability $\kappa_{v,w}$ by using Eq. (8).

We consider that such probabilities estimated by the above method can be used to a wider variety of applications.

3 Evaluation by Experiments

3.1 Network Data

We describe the details of the network data used in our experiments.

Blogs are personal on-line diaries managed by easy-to-use software packages, and they have spread rapidly through the World Wide Web [9]. They are equipped with a link creation function called a *trackback*, and so *bloggers* (i.e., blog authors) discuss various topics and establish mutual communications by putting trackbacks on each other's blogs. We treated a link created by a trackback as bidirectional for simplicity, and employed a trackback network of blogs as an example of information propagation network.

We exploited the blog "Theme salon of blogs" in the site "goo"[1], where a blogger can recruit trackbacks of other bloggers by registering an interesting theme. By tracing up to ten steps back in the trackbacks from the blog of the theme "JR Fukuchiyama Line Derailment Collision", we collected a large connected trackback network in May, 2005. This network was a directed graph of 12,047 nodes and 79,920 links.

3.2 Experimental Settings

We generated a value uniformly at random in some range $[\alpha, \beta]$ $(0 < \alpha < \beta < 1)$, then assigned this value to the diffusion probability $\kappa_{u,v}$ for any directed link (u,v) of a network; that is, $\kappa_{u,v} \in [\alpha, \beta]$. We determine the typical values of α, β for the blog network, and use them in the experiments. It is known that the IC model is equivalent to the bond percolation process that independently declares

[1] `http://blog.goo.ne.jp/usertheme/`

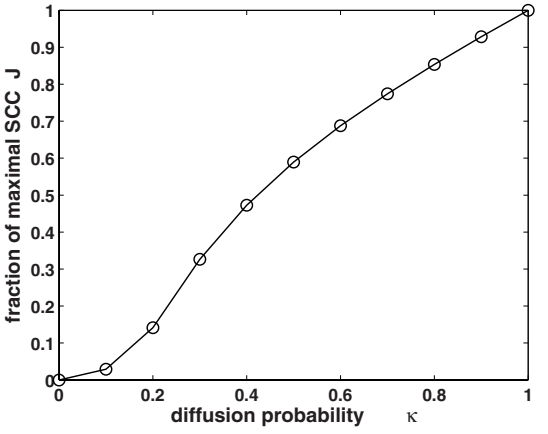

Fig. 1. The fraction J of the maximal SCC as a function of the diffusion probability κ

every link of the network to be "occupied" with probability κ [15]. Let J denote the expected fraction of the maximal strongly connected component (SCC) in the network constructed by occupied links. Note that J is an increasing function of κ. We focus on the point κ_* at which the average rate of the change of J, $dJ/d\kappa$, attains the maximum, and regard it as the typical value of κ for the network. Note that κ_* is a critical point of $dJ/d\kappa$, and defines one of the features intrinsic to the network. Figure 1 shows the values of J as a function of κ. Here, we estimated J using the bond percolation method with the same parameter value as [11]. From this figure we experimentally estimated κ_* to be $\kappa_* = 0.2$ for the blog network. Thus, we set the range such that $\alpha = 0.1$ and $\beta = 0.3$.

3.3 Experimental Results

Figure 2 shows how the learning performance changed with respect to S, where we changed S from 1 to 100 at intervals of 10 episodes. In our preliminary experiment, we set the size of the initial active set to be just one, and an initial active node is selected at random. We evaluated the learning performance by using the average absolute errors between true values and the corresponding estimated ones of diffusion probabilities, and standard deviations of the errors. As expected, the learning performance improved as the number S of episodes increased.

Figure 3 shows the distribution of T_s for the case $S = 100$. Since an initial node is selected at random, T_s is widely distributed ranging from one to 53. Here recall that there exist $79,920$ links in the network. These experimental results indicate that our approach worked well for this size of network.

4 Related Work and Discussion

There exist a large amount of work for information diffusion through networks.

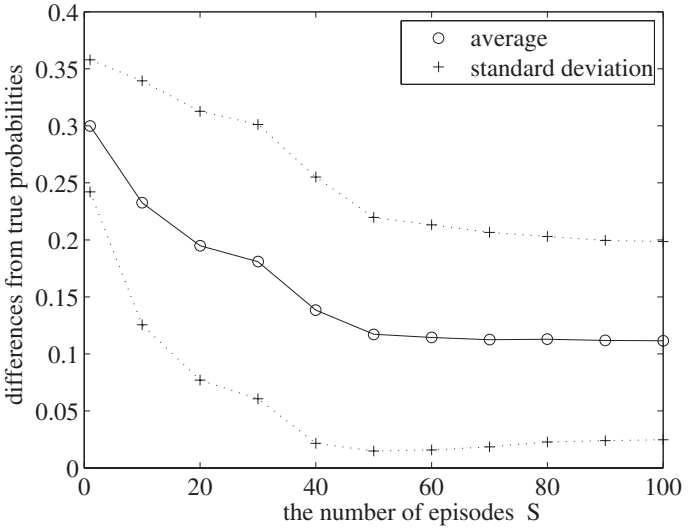

Fig. 2. How learning performance changes when the number of episodes S increases

Gruhl *et al.* [9], Adar and Adamic [1], and Leskovec, Adamic and Huberman [12] addressed the problem of tracking the propagation patterns of topics and influence through network spaces. Albert, Jeong and Barabási [2], Broder *et al.* [5], Callaway *et al.* [6], and Newman, Forrest and Balthrop [14] showed that the strategy of removing nodes in decreasing order of out-degree is effective for preventing the spread of undesirable things through networks. Balthrop *et al.* [4] studied effective "vaccination" strategies for preventing the spread of computer viruses through networks.

In contrast, a piece of information can diffuse through a social network in the form of "word-of-mouth" communication. Thus, it is important to find a limited number of influential nodes that are effective for the spread of information through a social network in terms of sociology and "viral marketing". Domingos and Richardson [7], Richardson and Domingos [17], Kempe, Kleinberg and Tardos [10], and Kimura, Saito and Nakano [11] studied a combinatorial optimization problem called the *influence maximization problem*. The problem is to extract a set of k nodes to target for initial activation such that it yields the largest expected spread of information, where k is a given positive integer. In particular, Kempe, Kleinberg and Tardos [10], and Kimura, Saito and Nakano [11] investigated the influence maximization problem for the IC model on which we focus in this paper.

Unlike discrete-time diffusion models in networks as above, Song *et al.* [19] proposed a continuous-time information diffusion model for a set of users. More specifically, they incorporated the diffusion rate that captures how efficiently the information can diffuse among the users. On the basis of their model, they predicted how likely the information will propagate from a specific sender to a specific receiver within a limited time, and applied it to a recommendation

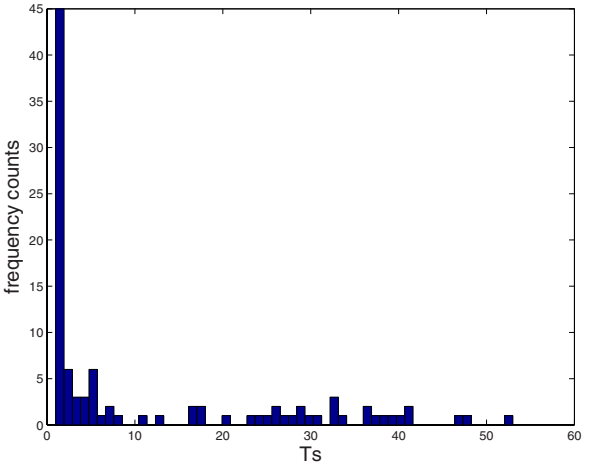

Fig. 3. Distribution of Ts for the case $S = 100$

system. They also estimated the expected time for information diffusion to a particular user, and applied it to a user ranking system. Our future work also includes applying our proposed method for the IC model to recommendation and ranking systems.

For the discrete-time IC-like model, Gruhl *et al.* [9] have already addressed a learning problem of diffusion parameters, and intuitively derived an EM-like algorithm for this purpose. Here when each parameter of geometric distribution is set such that $r_{v,w} = 1$, their model substantially reduces to our model. In this setting and our formalization, their posterior probability becomes as follows.

$$q(w|v) = \kappa_{v,w} \Big/ \sum_{u \in B(w) \cap D(t)} \kappa_{u,w}. \tag{9}$$

On the other hand, the posterior probability of our algorithm gives the following.

$$q(w|v) = \kappa_{v,w} \Big/ \left(1 - \prod_{u \in B(w) \cap D(t)} (1 - \kappa_{u,w}) \right). \tag{10}$$

We can easily see that the denominator of Eq. (9) is an approximation to that of Eq. (10) by ignoring cross terms of κ's.

5 Conclusion

We addressed the problem of predicting diffusion probabilities in complex networks. As one approach to this problem, we focused on the IC model, and defined the likelihood for multiple episodes. Then, we presented a method for predicting diffusion probabilities by using the EM algorithm. Our experiments using a

real blogroll network showed that the proposed method improved the prediction performance with the increase of episodes. In future, we plan to introduce the delay of propagation into the IC model, and extend our learning method to cope with the model extension.

Acknowledgment

This work was partly supported by the Grant-in-Aid for Scientific Research (C) (No. 20500147) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Adar, E., Adamic, L.: Tracking information epidemics in blogspace. In: Proc. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), pp. 207–214 (2005)
2. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. Nature 406, 378–382 (2000)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
4. Balthrop, J., Forrest, S., Newman, M.E.J., Williampson, M.W.: Technological networks and the spread of computer viruses. Science 304, 527–529 (2004)
5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the Web. In: Proc. of the 9th International World Wide Web Conference (WWW 2000), pp. 309–320 (2000)
6. Callaway, D.S., Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Network robustness and fragility: Percolation on random graphs. Physical Review Letters 85, 5468–5471 (2000)
7. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), pp. 57–66 (2001)
8. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12, 211–223 (2001)
9. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proc. of the 13th International World Wide Web Conference (WWW 2004), pp. 107–117 (2004)
10. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), pp. 137–146 (2003)
11. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. In: Proc. of the 22nd AAAI Conference on Artificial Intelligence (AAAI 2007), pp. 1371–1376 (2007)
12. Leskovec, J., Adamic, L., Huberman, B.A.: The dynamics of viral marketing. In: Proc. of the 7th ACM Conference on Electronic Commerce (EC-2006), pp. 228–237 (2006)
13. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proc. of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005), pp. 786–791 (2005)

14. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E 66, 035101 (2002)
15. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
16. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818 (2005)
17. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), pp. 61–70 (2002)
18. Strogatz, S.H.: Exploring complex networks. Nature 410, 268–276 (2001)
19. Song, X., Chi, Y., Hino, K., Tseng, B.L.: Information flow modeling based on diffusion rate for prediction and ranking. In: Proc. of the 16th International World Wide Web Conference (WWW-2007), pp. 191–200 (2007)
20. Watts, D.J.: A simple model of global cascades on random networks. In: Proc. of the National Academy of Sciences of the United States of America, vol. 99, pp. 5766–5771 (2002)