Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 5, September 2017, pp. 721-730
DOI: 10.23940/ijpe.17.05.p15.721730

A Study on the Influence Propagation Model in Topic Attention Networks

Xiao Chen[a,b,d], Jingfeng Guo[a,d,*], Kelun Tian[a], Chaozhi Fan[a], Xiao Pan[c]

[a] *College of Information Science and Engineering, YanShan University, Qinhuangdao, 066004, China*
[b] *Qian'an College, North China University of Science and Technology, Qian'an, 064400, China*
[c] *College of Economic and Management, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China*
[d] *The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao, 066004, China*

Abstract

The social networks with the complex user relations and huge amount of data and hidden information, bring new opportunities and challenges for the study of information diffusion and influence maximization. In recent years, there are more and more researches on the influence maximization of topic preference. However, most of the existing researches only take the topic as an attribute of the users, and the importance of the topic in network structure is not considered. In view of this situation, firstly, this paper constructed a new topic attention network model fusing the social relation and the topic preference. Secondly, based on connected degree of set pair and Markov random walk model, we propose the calculated method of the topic preference for users, and then mining the seed set with influence by the greedy strategy. Thirdly, we propose the calculated method of the activation probability of the user based on the user relation and the topic preference, and propose the influence maximization algorithm TAN_CELF in topic attention networks. Finally, on Dou-ban network dataset, from three metrics ISST, ISRT and ISRNT, compare with algorithm L_GAUP and CELF, the experimental results show that algorithm TAN_CELF that is proposed by this paper has a higher performance on influence scope.

Keywords: Topic preference; Connected degree; Markov random walk; Topic attention networks; Influence maximization; Information diffusion

(Submitted on March 26, 2017; Revised on June 16, 2017; Accepted on August 13, 2017)

1. Introduction

With the rapid development of Internet technology, some large-scale social network sites (such as Facebook, Micro-blog, Dou-ban etc.) gradually rise, and social networks have become an important platform for people to share, exchange and diffusion of information. Therefore, influence maximizing (short for IM) and information diffusion based on social networks attracts a lot of attention.

At present, the related researches on IM are mainly based on the relational social networks. However, the social network is a complex network that contains various entities and relations between entities. The entities not only have the natural link attributes, but also have the attributes of the entities' interest, attitude and preference to the information. Thus, the speed and scope of information diffusion in networks is not only related to the natural relation between the propagator and the receiver, but also has a more important relation with their interest and preference. If they have a common interest or acceptance to the information, it is quickly received and continues to spread, and vice versa.

Therefore, if both the user entity and the topic entity are considered in network, the network is called as topic attention network (short for TAN), as shown in Figure 1(a). In TAN, the different users have different preferences on the same topic, and the different topics have different users with diffusion influence. Therefore, the main work is to find the users with most influential, and to maximize influence on the specific topic in this paper.

A new algorithm TAN_ECLF based on TANs is proposed in this paper. It focuses on more accurate mining of seed vertices, which maximizes the diffusion of the topic. Firstly, the formal description of the TAN model is given. Secondly,

* Corresponding author.
E-mail address: jfguo@ysu.edu.cn.

based on connected degree of set pair theory and Markov random walk model, the calculated method of the topic's preference is proposed, and the seed set with influence by the greedy strategy can be obtained. Thirdly, the independent cascade model of topic attention network (short for TAN_IC) based on IC (Independent Cascade) model is given, and the calculating method of the probability that a user is activated is proposed based on the two factors of the relation between users and the user's preference to the topic. Finally, the experimental results show that the algorithm TAN_CELF can obtain a more accurate seed vertex set, so that the scope of the topic's influence is even greater.

The reminders of the paper are organized as follows. Section 2 introduces the related work. Section 3 introduces the preliminary. The new IM model of the topic attention networks is proposed formally in Section 4. The experimental settings are introduced in Section 5. Section 6 verifies the correctness and effectiveness of algorithm TAN_CELF. The paper is concluded in Section 7.

2. Related Work

The influence maximization (short for IM) problem begins with the study of viral marketing [11]. Subsequently, it is widely studied in various fields both at home and abroad. In 2003, the IM problem was first introduced to the social networks analysis by Kempe [5]. And the detailed definition and evaluation index are proposed, which provides the theoretical basis and guidelines for the following research.

At present, there are many research results on the IM in social networks, which can be divided into two categories. (1) The methods based on heuristic. The core idea is to calculate the influence of vertices according to the degree centrality and distance centrality. These methods are simple and efficient. However, due to only considering the network structure, the vertex selection is limited, and it is impossible to guarantee the maximization of information diffusion to the whole network. (2) The methods based on greedy strategy. The core idea is that the vertex with the maximum influence in each simulation propagates is added to the seed vertices set. The above methods are superior to the heuristic methods for obtaining the influence vertices, but the computational complexity is high and the efficiency is low. In view of the shortcomings in the above-mentioned, scholars have put forward a lot of optimization and improved methods [2,3,6,8,14]. In order to expand the influence scope of vertices in heuristic methods, Kimura and Saito propose shortest-path based influence cascade models and provide efficient algorithms of compute influence spread under these models [6]. In order to improve the efficiency of greedy hill-climbing algorithm, Leskovec present a CELF optimization to the original greedy algorithm based on the sub modularity of the influence maximization objective [8]. Their experimental results demonstrate that CELF optimization could achieve as much as 700 times speedup in selecting seed vertices, which is a very impressive result. In order to take into account the influence scope and efficiency, literature [14] proposes a greedy algorithm SCG (Set Covering Greedy). It is better than the method of vertex centrality algorithms in influence scope, and greedy hill-climbing algorithms in efficiency, and so on.

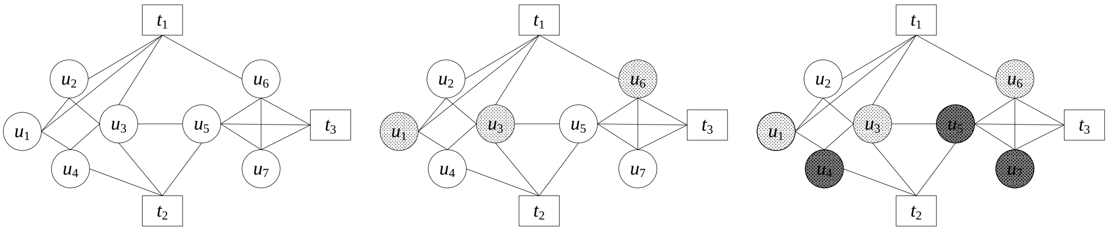
However, the existing methods neglected the fact that the topic can play an important role in influence maximization. In the real world, there are a variety of topics hidden in social networks, and they will also affect the diffusion effect and scope. The experimental results in reference [12,13] show that the different users have different preferences on the same topic, and the diffusion degree of different topic is different. That is, different topics have different influence on users and different users play a different role in the diffusion process for a specific topic. Therefore, the topic plays a vital role in the information diffusion. Subsequently, the scholars began to integrate the topic into the social network, research on the IM problem [1,4,9,10,17]. A probabilistic model to study the topic distribution and the influence of the topic aware is proposed by reference [9]. The traditional IC model is extends, and a Topic-aware Influence Cascade (short for TIC) is proposed by reference [1]. The Top-K influential vertices based on different topics were first studied in reference [17]. The IM algorithm L_GAUP based on the topic preference is proposed by reference [4] to improve the algorithm in reference [17]. The status and the impact of a topic in network structure is not concerned in the methods of the IM based on the topics above mentioned. In this paper, the status and the impact of a topic in network structure is to be concerned base on the TANs.

3. Influence Maximization of Topic Attention Networks

3.1. Topic Attention Networks

Definition 1. Topic Attention Networks (Short for TANs). The TANs is defined as a two tuple $TAN=(V,E)$. The $V=\{U,T\}$ is the vertex set, $U=\{u_1,u_2,\ldots,u_n\}$ is the user entity set, $T=\{t_1,t_2,\ldots,t_m\}$ is the topic entity set; $E=\{EU,EUT\}$ is the edge set, $EU=\{(u_i,u_j)|u_i, u_j\in U\}$ is the relation between users, $EUT=\{(u_i,t_j)|u_i\in U, t_j\in U\}$ is the relation between the user and the topic.

In TANs, let $A=\{AU,AUT\}$ is the adjacency matrix of network. Where $AU=\left(u_i,u_j\right)_{n\times n}=\begin{cases}1 & \left(u_i,u_j\right)\in EU\\ 0 & \left(u_i,u_j\right)\notin EU\end{cases}$ is user-user adjacency matrix. $AUT=\left(u_i,t_k\right)_{n\times m}=\begin{cases}1 & \left(u_i,t_k\right)\in EUT\\ 0 & \left(u_i,t_k\right)\notin EUT\end{cases}$ is user-topic adjacency matrix.



(a) (b) (c)

Figure 1. (a) Topic attention network (b) TAN with marked seed vertex (c) TAN after dissemination

For example, a TAN as shown in Figure 1(a), the user entity set is $U=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$, the topic entity set is $T=\{t_1,t_2,t_3\}$, the relation set between users is $EU=\{(u_1,u_2), (u_1,u_4), (u_2,u_3), (u_3,u_4), (u_3,u_5), (u_5,u_6), (u_5,u_7), (u_6,u_7)\}$, the relation set between users and topics is $EUT=\{(u_1,t_1), (u_2,t_1), (u_3,t_1), (u_6,t_1), (u_3,t_2), (u_4,t_2), (u_5,t_2), (u_5,t_3), (u_6,t_3), (u_7,t_3)\}$. The matrix *AU* and *AUT* are shown in Figure 2(a) and Figure 2(b).

$$AU=\begin{bmatrix}0&1&0&1&0&0&0\\1&0&1&0&0&0&0\\0&1&0&1&1&0&0\\1&0&1&0&0&0&0\\0&0&1&0&0&1&1\\0&0&0&0&1&0&1\\0&0&0&0&1&1&0\end{bmatrix} \qquad AUT=\begin{bmatrix}1&0&0\\1&0&0\\1&1&0\\0&1&0\\0&1&1\\1&0&1\\0&0&1\end{bmatrix}$$

(a) (b)

Figure 2. (a) Adjacent matrix AU (b) Adjacent matrix AUT

In TAN, $NU\left(u_i\right)_L=\begin{cases}\left\{u_j \mid \left(u_i,u_j\right)\in E\right\} & L=1\\ \left\{u_j \mid \left(u_j,u_p\right)\in E\cap u_p\in NU\left(u_i\right)_{L-1}\cap u_j\notin\bigcup_{l=1}^{L-1}NU\left(u_j\right)_l\right\} & L\geq 2\end{cases}$ is the user's neighbor set of u_i that the short path is L from u_i to u_j. Among them, each item in $NU(u_i)_1$ is adjacent with u_i directly. $CNU(u_i,u_j)_{L\cap M}=NU(u_i)_L\cap NU(u_j)_M$ is the common user neighbors set between u_i and u_j. $CNT(u_i,u_j)_{L\cap M}=NT(u_i)_L\cap NT(u_j)_M$ is the common topic neighbors set between u_i and u_j. If $L=M$, denoted as $CNU(u_i,u_j)_L$ and $CNT(u_i,u_j)_L$.

Definition 2. Connected Degree between Vertices. Given a $TAN=(V,E)$, $\forall u_i,u_j\in U$, $\forall t_k\in T$, concerned degree between u_i and u_j denoted as $\mu(u_i,u_j)$, as shown in formula (1).

$$\mu\left(u_i,u_j\right)=\frac{S_1\times w_1+S_2\times w_2+S_3\times w_3}{N}+\frac{(1)_{1\times F}}{N}i\left(t_k\right)_{F\times 1}+\frac{P}{N}j \tag{1}$$

In the formula (1), S_1 is the number of the common 1 level topic neighbor set between u_i and u_j, that is $S_1=|CNT(u_i,u_j)_1|$; S_2 is the number of the common 1 level user neighbor set between u_i and u_j, that is $S_2=|CNU(u_i,u_j)_1|$; S_3 is the link relation between u_i and u_j, if $(u_i,u_j)\in EU$, then $S_3=2$, else $S_3=0$. w_1, w_2 and w_3 respectively correspond to the weight of S_1, S_2 and S_3. Taking into account the importance of the network structure and the topics, let $w_1>w_2\geq w_3$, and $w_1+w_2+w_3=1$. The experimental results show that $w_1=0.7$, $w_2=0.15$ and $w_3=0.15$ is better.

F is the number of the common $1\cap 2$, $2\cap 1$ and 2 level topic neighbor set between u_i and u_j, that is $F=|CNT(u_i,u_j)_{1\cap 2}|+|CNT(u_i,u_j)_{2\cap 1}|+|CNT(u_i,u_j)_2|$. $(1)_{1\times F}$ is the row vector, the vector value is 1. $i(t_k)$ is the weight of the common topic vertex, consider the density structure characteristics in network, it value is quantified by clustering coefficient. $P=N\text{-}S\text{-}F$ is the number of the rest vertices, where $N=|V|$. $j=0$ is ignoring the influence of other user vertices on the connected degree.

Definition 3. Connected Degree of Vertex. Given a *TAN*=(*V*,*E*), for any vertex u_i, the connected degree of u_i is the sum of connected degree between u_i and all vertices, denoted as $\mu(u_i)$, as shown in formula (2).

$$\mu(u_i)=\sum_{j=1}^{|U|}\mu(u_i,u_j) \tag{2}$$

We still use the same example in Figure 1(a), let u_1 and u_3 as the two study objects. We can see that $NU(u_1)_1=\{u_2,u_4\}$, $NT(u_1)_1=\{t_1\}$, $NU(u_3)_1=\{u_2,u_4,u_5\}$ and $NT(u_3)_1=\{t_1,t_2\}$. Then, $CNT(u_1,u_3)_1=\{t_1\}$, that is $S_1=1$; $CNU(u_1,u_3)_1=\{u_2,u_4\}$, that is $S_2=2$; $(u_1,u_3)\notin EU$, that is $S_3=0$. Similarly, $CNT(u_1,u_3)_{1\cap 2}=\emptyset$, $CNT(u_1,u_3)_{2\cap 1}=\{t_2\}$ and $CNT(u_1,u_3)_2=\emptyset$; that is $F=1$; so $P=10-3-1=6$. Therefore, $\mu(u_1,u_3)=\frac{1\times 0.7+2\times 0.15+0\times 0.15}{10}+\frac{(1)_{1\times 1}}{10}\times\frac{1}{3}+\frac{6}{10}\times 0=0.133$. By calculating the connected degree between vertices, we can get the matrix of connected degree between user vertices, as shown in Figure 3(a). Connected degree of vertices is shown in Figure 3(b). It is seen that the vertex's connected degree reflects the influence of the vertex in the networks. The greater the value of $\mu(u_i)$, the greater influence.

$$RU=\begin{bmatrix} 0.000 & 0.125 & 0.133 & 0.110 & 0.073 & 0.083 & 0.063 \\ 0.125 & 0.000 & 0.145 & 0.092 & 0.085 & 0.082 & 0.062 \\ 0.133 & 0.145 & 0.000 & 0.147 & 0.195 & 0.165 & 0.100 \\ 0.110 & 0.092 & 0.147 & 0.000 & 0.098 & 0.072 & 0.056 \\ 0.073 & 0.085 & 0.195 & 0.098 & 0.000 & 0.210 & 0.190 \\ 0.083 & 0.082 & 0.165 & 0.072 & 0.210 & 0.000 & 0.916 \\ 0.063 & 0.062 & 0.100 & 0.056 & 0.190 & 0.196 & 0.000 \end{bmatrix} \qquad \mu(u_i)=\begin{bmatrix} 0.587 \\ 0.591 \\ 0.885 \\ 0.575 \\ 0.851 \\ 0.808 \\ 0.667 \end{bmatrix}$$

(a) (b)

Figure 3. (a) RU (b) Vertex connected degree

3.2. Influence Maximization

In TANs, the goal of IM problem is to find a user's set *S* (|*S*|=*K*) closely related with topic, the *K* users as initial vertices set (referred to as *seeds*); each item in *S* is active state, and then select a specific propagation model, make the expected number of users influenced on the topic by the *K* seeds is maximize.

Definition 4. Influence Spread of the Topics. In *TAN*=(*V*,*E*), given a topic *t* and a set *S* of *K* seeds, the number of activated vertices in *N* is called influence spread, denoted by $N(TAN,S,t_k)$, as shown in formula (3).

$$N(TAN,S,t_k)=\{u_i\mid u_i\in U,(u_i,t_k)\notin EUT\} \tag{3}$$

Definition 5. Influence Maximization in TANs. Given a *TAN*=(*V*,*E*), a topic t_k and an integer *K*, find a *K*-vertex set *S*, for any other *K*-vertex set $S^*\in TAN$, then $N(TAN,S,t_k)\geq N(TAN,S^*,t_k)$. *S* is called a seed set and each vertex in *S* is called a seed.

For example, Figure 1(a) shows a TAN with 7 users and 3 topics. Given a diffusion topic t_1 and the seeds size *K* (*K*=3), to maximize the expected number of influenced users. Firstly, we find a seeds $S=\{u_1,u_3,u_6\}$, as shown in Figure 1(b). Secondly, these seeds were applied to the selected propagation model, and the maximum diffusion was carried out. Finally, the vertex u_4, u_5 and u_7 are activated, which are not attended to t_1, as shown in Figure 1(c), that is $N(TAN,S,t_1)=3$.

3.3. Independent Cascade Model in TANs

In TAN, the different users have different behavior/preferences (i.e. direct attention or indirect attention) on the specific topic. For example, in Figure 1(a), u_3 is attention to t_1 directly, but u_5 is attention to t_1 through u_3. It can be seen that the preferences of users is different when the paths is different. Therefore, based on the paths relation, we propose the calculated method of the topic preference for users, taking account of connected degree of set pair and Markov random walk model [15,16]. The formal definition is as follows.

Definition 6. *L* Level User Neighbors Set of the Topic. Given a *TAN*=(*V*,*E*), $\forall t_k\in T$, if the short path length from t_k to u_i is *L*, then u_i is the *L* level user neighbor of t_k, the *L* level user neighbors set of t_k is denoted as $NU(t_k)_L$, as shown in formula (4).

$$NU\left(t_k\right)_L = \begin{cases} \left\{u_i \mid \left(u_i,t_k\right) \in EUT\right\} & L{=}1 \\ \left\{u_i \mid \left(u_i,t_k\right) \notin EUT, \left(u_i,u_j\right) \in EU, u_j \in NU\left(t_k\right)_{L-1}, u_i \notin \bigcup_{l=1}^{L-1} NU\left(t_k\right)_l \right\} & L{>}1 \end{cases} \tag{4}$$

Definition 7. Topic Preference. Given a $TAN=(V,E)$, $\forall t_k \in T$, $u_i \in NU(t_k)_L$, the L level neighbor user u_i's preference to t_k is denoted $Per_L(u_i,t_k)$, as shown in formula (5).

$$Per_L\left(u_i,t_k\right) = \begin{cases} \mu\left(u_i\right) \times \dfrac{1}{\mid NT\left(u_i\right)_1 \mid} & L{=}1 \\ \sum_{j \in NU\left(u_i\right)_1 \cap NU\left(t_k\right)_{L-1}} Per_{L-1}\left(u_j,t_k\right) \times \dfrac{\mu\left(u_i\right)}{\mid N\left(u_i\right)_1 \mid} & L{>}1 \end{cases} \tag{5}$$

In the formula (5), $\mu(u_i)$ is the influence of u_i in network; $|NT(u_i)_1|$ is the number of topics that is adjacent with u_i directly. $|N(u_i)_1|$ is the number of neighbors (include users and topics). When $(u_i,t_k) \in EUT$, $Per_1(u_i,t_k)$ is the 1 level neighbor user u_i's preference to t_k. In order to calculate simply, we assumed that the user has the same preference for all the directly connected topics. That is the proportion of the users' influence and the number of topics that he is concerned about. When $(u_i,t_k) \notin EUT$, $Per_L(u_i,t_k)$ is the L level neighbor user u_i's preference to t_k. In this paper, the user preference to the topic is calculated by Markov random walk model. That is the mean of L level neighbor user u_j's influence of t_k to u_i.

For example, the TAN as shown in Figure 1(a), the preference u_3 of topic t_1 is $Per_1(u_3,t_1)=\mu(u_3)/|NT(u_3)_1|=0.4425$; the preference u_5 of topic t_1 is $Per_2(u_5,t_1)= \sum_{j \in NU\left(u_5\right)_1 \cap NU\left(t_1\right)_1} Per_1\left(u_j,t_1\right) \times \dfrac{\mu\left(u_5\right)}{\mid N\left(u_5\right)_1 \mid}$ $=(Per_1(u_3,t_1)+Per_1(u_6,t_1))\times(0.851/5)=(\mu(u_3)/|NT(u_3)_1|+\mu(u_6)/|NT(u_6)_1|)\times(0.851/5)=(0.885/2+0.808/2)\times(0.851/5)=0.1441$.

Definition 8. Topic Influence. Given a $TAN=(V,E)$, $\forall t_k \in T$, if the number of $NU(t_k)_1=\{u_i|(u_i,t_k) \in EUT\}$ is M, the number of $NU(t_k)_2=\{u_p|(u_i,t_k) \in EUT, (u_i,t_p) \in EU\}$ is P,…, the number of $NU(t_k)_L=\{u_q\}$ is Q,…, the influence of topic t_k is the sum of all user's preference to t_k, that is $Per(t_k)$, as shown in formula (6).

$$Per\left(t_k\right) = \frac{\sum_1^M Per_1\left(u_i,t_k\right) + \sum_1^P Per_2\left(u_p,t_k\right) + \ldots + \sum_1^Q Per_L\left(u_q,t_k\right) + \ldots}{M + P + \ldots + Q + \ldots} \tag{6}$$

In the formula (6), the influence of topic t_k represents the preference of all users to topic t_k. Obviously, it is not feasible in actual networks. We can see L is the short path length from u_i to t_k. If L is too small, the steps of random walking are too few; if L is too large, it will increase the computational complexity. Meanwhile, due to us unknown network size; therefore, choosing a reasonable L will be helpful to calculate the topic influence. The experience range of step length $L \in [6,20]$ is given [15]. Considering the differences of the network size and density, the different optimal steps parameter needs to be set up. The steps L of random walk are equal to the steps when the community initial center vertex random walk in a metastable state [7], according to the large deviation theory. That is, L is the first times step length when the numbers of zero components of adjacent two steps are equal in transition matrix.

The TANs contain two categories entities: the users and the topics. Visible, the probability of user activation is also affected by two parts. On one hand, it is the structure, which is the closely relation between user neighbors on the broad range; on the other hand, it is the content, which is user's preferences for the topic. Therefore, based on the above two factors, we proposed the diffusion model in topic attention networks based on IC model, as follows.

Independent Cascade Model in Topic Attention Network (short for TAN_IC), the probability of user activation is denoted as $p(u_i,u_j,t_k)$, as shown in formula (7).

$$p\left(u_i,u_j,t_k\right) = w_1 \times Inf\left(u_i,u_j\right) + w_2 \times F\left(Per\left(u_i,t_k\right),Per\left(u_j,t_k\right)\right) \tag{7}$$

In the formula (7), $Inf(u_i,u_j)$ is the influence of u_i to u_j, as shown in formula (8). That is the common 1 level topic neighbors between u_i and u_j accounted for the all topic neighbors of them. $F(Per(u_i,t_k), Per(u_j,t_k))$ is the average of the preference of u_i and u_j to t_k, as shown in formula (9). For the above two parts, we give the weights w_1 and w_2. We expect the topic to have greater impact factors, so the experiment set $w_2>w_1$. The results of several experiments show that it will be the best effect when $w_1=0.4$ and $w_2=0.6$ respectively in the formula (7).

$$Inf\left(u_i,u_j\right)=\frac{CNT\left(u_i,u_j\right)_1}{|NT\left(u_i\right)_1 \cup NT\left(u_j\right)_1|} \tag{8}$$

$$F\left(Per\left(u_i,t_k\right),Per\left(u_j,t_k\right)\right)=\frac{Per\left(u_i,t_k\right)+Per\left(u_j,t_k\right)}{2} \tag{9}$$

3.4. Algorithm Description

In TANs, the key of IM problem is mining the seed vertex set *S*. Similar to most mining seed vertex algorithms, we also make the |*influenceSe*t(*S*)| as objective function. We propose the mining algorithm TAN_CELF of the seed vertex set based on the optimal CELF algorithm. The algorithm is described as follows.

Algorithm 1 TAN_CELF(AU,AUT,t_k,k)

Input: *AU*, *AUT*, the topic t_k, the scale *K* of the seed set
Output: the seed set *S*

```
1.  Initialize $S=\emptyset$, $Q=\emptyset$, $R=1000$
2.  for each edge $(u_i,u_j)\in EU$ do
3.    $p(u_i,u_j,t_k)=w_1\times Inf(u_i,u_j)+W_2\times F(Per(u_i,t_k), Per(u_j,t_k))$
4.  end for
5.  $S_{sub}=getSubSeed(G,AU,AUT)$
6.  for each $u_i\in S_{sub}$ do
7.     $u_i.ins = 0$
8.     for $j = 1$ to $R$ do
9.       $u_i.ins+=|influenceSet(S_{sub},\{u_i\})|$
10.    end for
11.    $u_i.ins = u_i.ins/R$
12.    $u_i.flag=0$
13.    Add $u_i$ to $Q$ by $u_i.ins$ in descending order
14. end for
15. while $|S|<k$ do
16.   $u_i=Q[top]$
17.   if $u_i.flag==|S|$ then
18.     $S=S+\{u_i\}$
19.     $Q=Q-\{u_i\}$
20.   else
21.     $u_i.ins = 0$
22.     for $j = 1$ to $R$ do
23.        $u_i.ins+=|\inf luenceSet(G_{res_{t_k}},S+\{u_i\})|$
24.     end for
25.      $u_i.ins=u_i.ins/R$-$|\inf luenceSet(G_{res_{t_k}},S)|$
26.     $u_i.flag=|S|$
27.     Resorted $Q$ by $u_i.ins$ in descending order
28.   end if
29. end while
30. Return $S$
```

In algorithm 1, we are mainly describes the steps of mining the seed vertex. Firstly, the variables are initialized (Line 1). Secondly, the spreading probability of the user's to topic is calculated (Line 2-4). Thirdly, the candidate set S_{sub} of seed vertex is calculated (Line 5). Fourthly, the influence scope of each vertex is calculated (Line 6-14). Finally, we iterative are mining the seed vertex (Line 15-30). Among them, the function of *getSubSeeds*(*AU*,*AUT*) is to calculate the preference of each user to the topic t_k in TAN. Where, according to the formula (5), calculate the $Per_L(u_i,t_k)$. And according to the formula (6), calculate the *Per*(t_k); which is taken as the threshold to obtain the seed vertex candidate set S_{sub}. The specific process is shown in algorithm 2.

Algorithm 2 *getSubSeeds*(*AU*,*AUT*)

```
Input: $AU$, $AUT$
Output: the candidate set $S_{sub}$
for each vertex $u_i \in U$ do
  According to the formula (5), calculate the $Per_L(u_i,t_k)$
end for
According to the formula (6), calculate the $Per(t_k)$
$S_{sub}=\emptyset$
for each vertex $u_i \in U$ do
  if $Per_L(u_i,t_k) \geq Per(t_k)$ then $S_{sub}=S_{sub} \cup u_i$
end for
Return $S_{sub}$.
```

4. Experimental Data Set and Evaluation Index

4.1. Experimental Environment and Experimental Data Set

Experiment hardware environment is Intel(R) Core(TM) i7-4760HQ CPU@1.8GHz with 8GB memory; the software environment is Windows 8 system, Java JDK 1.7, Eclipse 4.3.

We get the data set of users' film reviews from Dou-ban network. The basic structural information of the data set is shown in Table 1. There are 2253 users, 36 categories movies, which involve 29009 films, upload views (with scoring data) are 563173, and movie reviews (with rating data) are 119766. The attention network model based on the film topic is constructed, which contains 2253 user entities, 36 topic entities, 27988 edges between user entities, and 34818 edges between users and topics.

Table 1. Information of data set

Type		Number	Total
Vertex	user / topic	2253 / 36	2289
Edge	user-user / user-topic	27988 / 34818	62806

By statistics, the number of users directly related to each topic is shown in Figure 4. The number of topics when users≥1000 is 17; the number of topics when 500≤users<1000 is 7; and the number of topics when users<500 is 12. The topics when the number of users is more than 1000 have been most users' attention. The topics when the number of users less than 500 are only a small number of users concerned. In order to make a better diffusion effect, we choose the topics (500≤users<1000) as the experimental objects.

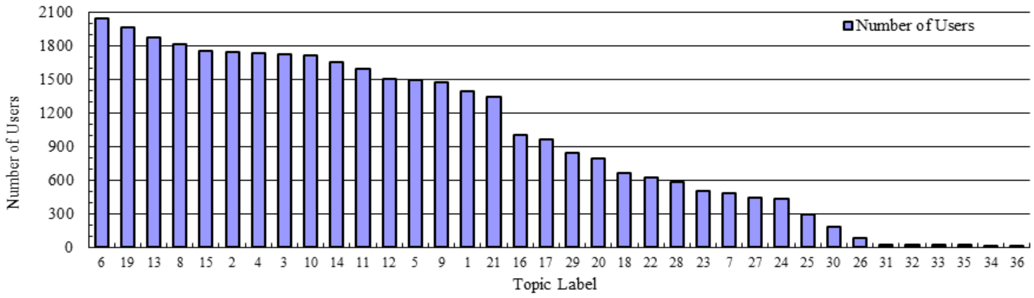


Figure 4. Users directly relate to topic

4.2. Evaluation Index

In order to evaluate the performance of the algorithm, we combine the characteristics of TANs with the evaluation index in reference [16], and put forward three metrics: ISST (Influence Spread of a Seed Set on A Specific Topic), ISRT (Influence Spread Result on A Specific Topic) and ISRNT (Influence Spread Result on A Specific Topic not Direct to the Topic).

(1) *ISST*(*S*,t_k). Similar to reference [1]. Given a specific topic t_k, based on the user's preference for t_k, *ISST*(*S*,t_k) is the influence scope value of the topic t_k by the seeds set *S*, as shown in formula (10).

$$ISST\left(S,t_k\right)=\sum\nolimits_{u_i \in seedSet(S)} Per\left(u_i,t_k\right) \tag{10}$$

(2) *ISRT*(*S*,t_k). Given a specific topic t_k, *ISRT*(*S*,t_k) is the influence scope of the topic t_k by the seeds set *S*, as shown in formula (11).

$$ISRT\left(S,t_k\right)=\sum\nolimits_{u_i \in seedSet(S)} |\,influenceSet\left(u_i\right)| \tag{11}$$

(3) *ISRNT*(*S*,t_k). Given a specific topic t_k, *ISRNT*(*S*,t_k) is the influence scope of the topic t_k by the seeds set *S*, which does not include the vertices directly associated with t_k, as shown in formula (12).

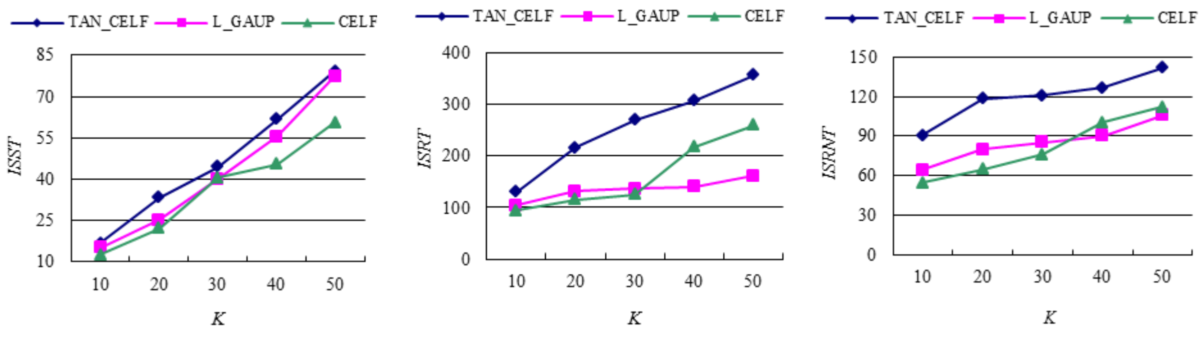
$$ISRNT\left(S,t_k\right)=\sum\nolimits_{u_i \in seedSet(S)} |\,influenceSet\left(u_i\right)| - NU\left(t_k\right) \tag{12}$$

5. Experimental Analysis

To evaluate the performance of TAN_CELF algorithm proposed by this paper, in Dou-ban data set, we implemented the CELF algorithm based on IC model and the L_GAUP algorithm based on E_IC model. In experiments, we use Monte Carlo simulation method to simulate the propagation process 1000 times, the average value of the experimental results are analyzed. The purpose of the experiment is as follows. (1) In ISST metric, compared with the traditional IM algorithm CELF and L_GAUP, it is verified that the algorithm TAN_CELF has a higher final preference influence on the specific topic. (2) In ISRT and ISRNT metric, compared with the algorithm CELF and L_GAUP, for the specific topic, the algorithm TAN_CELF can influence more users is verified. (3) For the specific topic, the correctness and effectiveness of the algorithm TAN_CELF are terrified through calculating the influence scope of the user that indirect attention to topic by the seed set S.

5.1. Experimental Results of Three Algorithms

In experiments, we randomly select a topic 20 (500≤users<1000), and set the scale of the seed set from 10 to 50. For the topic 20, in ISST, ISRT and ISRNT metric, the experimental results of three algorithms are shown in Figure 5. We can see from Figure 5, with the increase of the scale *K* of seed set *S*, the influence range of the three algorithms is gradually increased. In Figure 5(a), compare with the algorithm CELF and L_GAUP, algorithm TAN_CELF based on the influence of topic, has a higher preference influence on the specific topic. In Figure 5(b) and Figure 5(c), since the TAN_CELF algorithm using the vertex with the largest number of increments as the seed vertex, it has the greatest influence range and the better diffusion effect.



(a) (b) (c)

Figure 5. (a) ISST vs. K (b) ISRT vs. K (c) SRNT vs. K

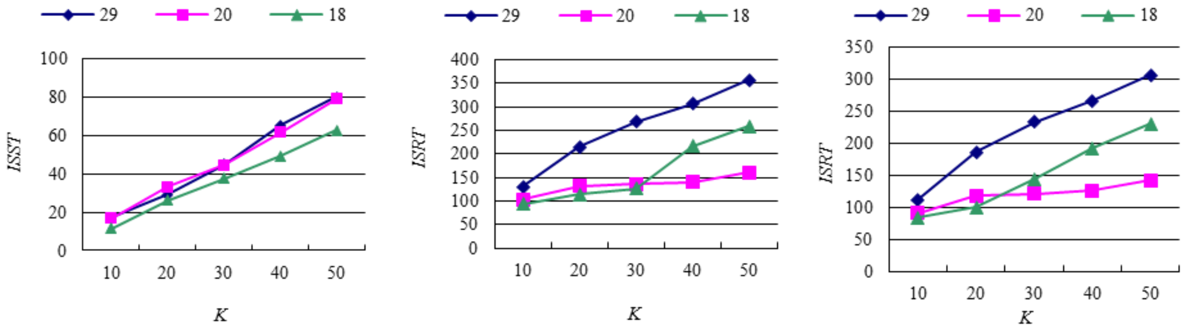
5.2. Experimental Results of Three Topics

In Dou-ban network, we randomly selected 3 movie topics, including song and dance, documentary and the west, the label of topic and the number of users with direct to the topic is shown in Table 2.

Table 2. Topic type and numbers of directly related users

Type of Topic	Songs and Dances	Documentary	The West
Label of Topic	No. 29	No. 20	No. 18
Number of users	848	790	660

In ISST, ISRT and ISRNT metric, the topic 29, 20 and 18 are propagating respectively by the TAN_CELF algorithm. The experimental results are shown in Figure 6. We can see from Figure 6, in the process of mining the seed vertex set, because the ISST, ISRT and ISRNT are proportional to each other, with the increasing of the scale of the seed vertices, the influence range of the different topic is gradually increased. The influence of the seed set *S* on a specific topic is different, and the final influence of topic 29 was significantly better than that of topic 20 and 18. When 29 is the diffusion topic, the number of users directly related to the topic 29 is more. That is, the number of vertices to the topic within one step is more, which is making the larger set of candidate seeds to meet the threshold, so the influence is better than 20 and 18.



(a) (b) (c)

Figure 6. (a) ISST vs. K (b) ISRT vs. K (c) ISRNT vs. K

6. Conclusions

In TANs, firstly, we construct the new independent cascade model of topic attention network (short for TAN_IC) based on IC model. Secondly, we propose the influence maximization algorithm TAN_CELF. Finally, in Dou-ban network data set, the influence of topic propagation is verified by experiments. Experimental results show: (1) For the same topic, with the increase of the seed vertices, the influence of the topic is gradually increased. That is, the influence scope of the users that do not pay attention to the topic significantly increased. (2) For the different topics, the same seed set has different influence range, so it is necessary to select the seed set for a specific topic. The next research goal is the efficiency of the algorithm is improved and the maximum influence achieved on Sina and Twitter and other large-scale data sets.

Acknowledgements

This work is supported by the National Science Foundation of China (No. 61472340, No. 61303017), the National Youth Science Foundation of China (No. 61602401), the National Youth Science Foundation of Hebei (No. F2017209070), the Natural Science Foundation of Hebei Province (No. F2014210068), and the Fourth Outstanding Youth Foundation of Shijiazhuang Tiedao University.

References

1. N. Barbieri, F. Bonchi, and G. Manco. “Topic-Aware Social Influence Propagation Models” in *Proceedings of the 8th International Conference on Data Mining*, pp. 81-90, Las Vegas Nevada, USA, July, 2012
2. A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. “Learning Influence Probabilities in Social Networks” in *Proceedings of the 3th International Conference on Web Search and Web Data Mining, WSDM 2010*, pp. 241-250, New York, USA, February, 2010
3. J. Guo, P. Zhang, and C. Zhou. “Personalized Influence Maximization on Social Networks” in *Proceedings of the 22th ACM International Conference on Information & Knowledge Management*, pp. 199-208, San Francisco, USA, October, 2013
4. J. F. Guo, and J. G. Lv. “Influence Maximization Based on Information Preference,” *Journal of Computer Research and Development*, no. 02, pp. 533-541, February, 2015
5. D. Kempe, J. Kleinberg, and É. Tardos. “Maximizing the Spread of Influence Through A Social Network” in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137-146, Washington, USA, August, 2003
6. M. Kimura and K. Saito. “TracTable Models for Information Diffusion in Social Networks” in *Proceedings of the 17th Knowledge Discovery in Databases: PKDD 2006*, pp. 259-271, Berlin, Germany, September, 2006

7. L. Q. Kong, and M. L. Yang. “Improvement of Clustering Algorithm FEC for Signed Networks,” *Journal of Computer Applications*, vol. 31, no. 5, pp. 1395-1399, May, 2011
8. J. Leskovec, A. Krause, and C. Guestrin. “Cost-effective Outbreak Detection in Networks” in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 420-429, San Jose, California, USA, August, 2007
9. L. Liu, J. Tang, and J. Han. “Mining Topic-Level Influence in Heterogeneous Networks” in *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010*, pp. 199-208, Toronto, Ontario, Canada, October, 2010
10. J. G. Lv and J. F. Guo. “Mining communities in social network based on information diffusion,” *Ieej Transactions on Electrical & Electronic Engineering*, vol. 11, no. 5, pp. 604-617, July, 2016
11. M. Richardson and P. Domingos. “Mining Knowledge-Sharing Sites for Viral Marketing” in *Proceedings of the 8th Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 61-70, Edmonton, Canada, August, 2002
12. K. Saito, M. Kimura and K. Ohara. “Behavioral Analyses of Information Diffusion Models by Observed Data of Social Network” in *Proceedings of the Advances in Social Computing, Third International Conference on Social Computing, Behavioral Modeling*, and Prediction, pp. 149-158, Bethesda, MD, USA, March, 2010
13. K. Saito, M. Kimura, and K. Ohara. “Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis” in *Proceedings of the 1th Advances in Machine Learning, First Asian Conference on Machine Learning*, ACML 2009, pp. 322-337, Nanjing, China, November, 2009
14. P. Vasanthi, V. Tejaswi, and P. S. Thilagam. “Time Stamp Based Set Covering Greedy Algorithm” in *Proceedings of the 21th ACM Ikdd Conference*, pp. 110-111, Sydney, Australia, August, 2015
15. B. Yang, J. Liu, and J. Feng. “On Modularity of Social Network Communities: The Spectral Characterization” in *Proceedings of the 3th Ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology*, 2008 Wi-Iat, pp. 127-133, Sydney, Australia, December, 2008
16. W. Zheng, C.K. Wang, Z. Liu and J. M. Wang. “A Multi-Label Classification Algorithm Based on Random Walk Model,” *Chinese Journal Of Computers*, vol. 33, no. 8, pp. 1418-1426, August, 2010
17. J. Zhou, Y. Zhang, and J. Cheng. “Preference-based Mining of Top-K Influential Nodes in Social Networks,” *Future Generation Computer Systems*, vol. 31, no.1, pp. 40-47, February, 2014

Xiao Chen graduated from College of Information Science and Engineering, YanShan university, China, for the degree of Master and Ph. D. She is lecturer at North China University of Science and Technology. Her research interests include graph mining, social network analysis, etc.

Jingfeng Guo graduated from College of Information Science and Engineering, YanShan university, China, for the degree of Bachelor, Master and Ph. D. He is a professor and Ph.D. supervisor at YanShan university. His research interests include database theory and application, data mining and social network analysis, etc.

Chaozhi Fan graduated from College of Information Science and Engineering, YanShan university, China, for the degree of Master. Her research interests include social networks analysis and Influence Maximization.

Kelun Tian is a Ph.D. student from the College of Information Science and Engineering, YanShan university, China. Her research interests is the social networks analysis.

Xiao Pan is an associate professor at Shijiazhuang Tiedao University, China. She was a visiting scholar in the Department of Computer Science, University of Illinois at Chicago, USA. She received her Ph.D. in Computer Science from Renmin University of China in 2010. Her research interests include data management on moving objects, location based social networks and privacy-aware computing.