Topic-Aware Influence Maximization in Large Recommendation Social Networks

Jinghua Zhu(✉), Qian Ming, and Nan Wang

School of Computer Science and Technology, Heilongjiang University, Harbin, China
zhujinghua@hlju.edu.cn

Abstract. Influence maximization (*IM*) is a problem of finding several influential individuals in a social network so that their influence spread is maximized under certain propagation model. In recommendation social network such as Douban, information diffuses with multiple origins: internal and external influence. Furthermore, pairs of individuals usually have different influence strength on different topics, information, ideas and rumors etc. In this paper, we focus on the topic-aware *IM* problem for large recommendation social networks. We propose a novel TSID propagation model to formulate the multiple topics diffusion in recommendation social networks. We propose TIP algorithm to solve the influence maximization problem under TSID propagation model. Our experiment results show that TSID model can well depict the mix information propagation process in recommendation social network, the TIP algorithm has competitive response time and influence spread.

Keywords: Influence maximization · Topic-aware Recommendation social network

1 Introduction

Recently, large social networks have sprung up, social network is not only important medium to exchange information, make friends, but also important business platform. Businesses can choose a small part of influential people in social networks, through to provide them with free products, to make them through social networks recommend the product to their friends or family, reaching the largest scope of products with "word of mouth".

The information spread is affected by many factors, including the impact probability between users, the user's preference for information and the impact of the web site to the user. For example: whether or not the user to accept the product will not only be affected by their friends, but also by the push message of web site impact. The web site can be used to get more information through the home page news, sending messages, reminding message and other forms, and the user may be able to accept these messages, then further recommend to their friends. Another one example: to spread different products or ideas in the web site, because the user's different preferences for different types, so the spread effect between the same user is different, propagation process will

G. Sun and S. Liu (Eds.): ADHIP 2017, LNICST 219, pp. 195–203, 2018.
https://doi.org/10.1007/978-3-319-73317-3_24

inevitably be influenced by user preference. Just as women pay more attention to cosmetics than men, cosmetics marketing should be more dependent on women.

In recent years, many algorithms have been proposed to solve the influence maximization problem, although some algorithms take into account the topic, but did not consider the impact of the web site itself to the user. The web site is an important influence, with the users to promote the spread, can be spread in a number of local areas, and therefore spread faster than the traditional spread process. Considering the social network user's preference, combining with the impact of the site can make the selection more accurate, the spread process can better fit the actual situation. Therefore, this paper extends the SID (Super Influencer Diffusion) diffusion model, and proposes the TSID (Topic-aware Super Influencer Diffusion) diffusion model, which can deal with the topic-aware influence propagation. Based on the TSID model, this paper proposes a TIP (Topic-aware Influence Path) algorithm, according the current activated node transmission, fast calculation the influence of node based on propagation influence path selecting the node with largest marginal gain as seed with greedy thought is inactive node set. This adaptive selection method can faster and broader the spread influence.

To evaluate TIP algorithm, we choose movie reviews in the Douban network as the data set. Douban is a famous domestic social networking sites, services including project recommendation, making friends, comment which is the core service. Because a large number of user provide ratings and reviews, the Douban score has important reference in the minds of users. The experimental results show that the TIP algorithm is more extensive than the existing algorithms, and the time efficiency is high.

The main contributions of this paper are as follows:

- An extended TSID diffusion model is proposed, and the formulas for calculating the internal and external influence probabilities are given.
- The TIP algorithm is proposed, which can adaptively determine the seed set in inactive nodes according to the current communication.
- The comparative experiments on the data set and the result show that the TIP algorithm has a greater impact on the transmission range and less running time.

2 Related Works

Kempe et al. [1] first propose discrete optimization method for the influence maximization, they present a greedy hill-climbing approximation algorithm. Goyal et al. [2] exploited simple paths between neighbor nodes to estimate the influence propagation probability. Lu et al. [3] propose algorithm to get the influence in a range of four hops, they also propose an approximation algorithm to compute influence in the range of at least five hops. You et al. [4] find that under certain incentives, one would build new relationship in the social network to promote the process of information propagation. The above algorithms can deal with the traditional influence maximization problem, but they did not take into account the factors affecting the transmission process.

Barbieri et al. [5] propose TIC model and TLT model to solve the topic-aware influence maximization problem. Zhou et al. [6] propose a two-step mining algorithm

GAUP. Guo and Lv [7] propose EIC model which is based on users' activities and preference and L_GAUP algorithm. Chen et al. [8] find that the majority of seed nodes under multiple topics are derived from the composition of the topic set and they propose C-Greedy algorithm. Zhu et al. [9] propose structural hole based influence maximization algorithm. Chen et al. [10] establish a maximum influence tree to approximate the computational power of the topic based algorithm.

3 Diffusion Model and Problem Statement

3.1 TSID Model

Niu et al. [11] investigate the information diffusion of Douban network and propose SID model. In this paper, we extend SID model to TSID model which is suitable for topic aware influence maximization problem for recommend social network.

In TSID model, there are lots of common user nodes and one super node. Common nodes have two states: active or inactive. Inactive node can be activated by active node at least once, all common nodes can only change state from inactive to active, the opposite is not allowed. The propagation process of TSID model is similar to traditional IC model: Initially, the super node is active and the common nodes are inactive. Then propagation begins. In each time step, the active nodes would influence their neighbors. The super node infect all nodes with probability P_{ex}; other active nodes will infect their neighbor nodes with probability P_{in}. The propagation end when all nodes don't change states anymore.

As shown in Fig. 1, the active nodes (in red) can influence their neighbors with probability $P_{in,}$ and the super node can also influence inactive nodes with probability P_{ex}.

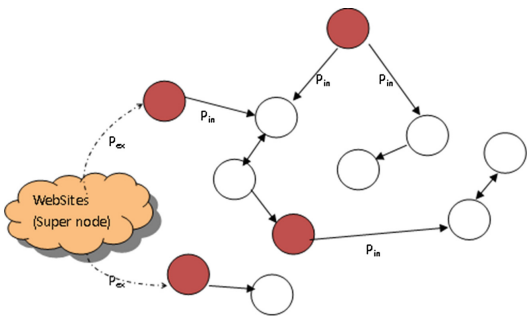

Fig. 1. The TSID model. (Color figure online)

3.2 Influence Probability

(1) ***Internal influence probability***

In this paper, P_{in} represents the internal influence probability among users. In this paper, we consider the influence probability not only is related with influence

frequency, but also is correlated with users' similarity. We define the internal influence probability from user u to v as the following formula:

$$p_{uv}^{in} = b \times W_{uv} + (1 - b) \times S_{uv} \tag{1}$$

In the above formula, the internal influence probability p_{uv}^{in} is the weighted linear combination of original uniform probability W_{uv} and the similarity S_{uv} of u and v. The similarity S_{uv} is measured by similar activities of u and v.

(2) *External influence probability*

The external influence probability P_{ex} represents the environment impact from the super influencer such as the website itself. In this paper, we define the external influence probability from super node g on user u as the following formula:

$$p_{gu}^{ex} = c \times A_u + (1 - c) \times D_u \tag{2}$$

In the above formula, c is the harmonic factor. A_u and D_u represent the activeness and dependence of user u. As shown in formula (3), A_u is the number of films that user u has watched divided by the average number of films that have been seen by all users. D_u can be measured by the portion of online time of user u as shown in formula (4).

$$A_u = |S_u| / \overline{|S|} \tag{3}$$

$$D_u = T_u^{on} / T_u^{all} \tag{4}$$

(3) *Topic aware influence probability*

Now, we incorporate the topic mixtures in the diffusion model. Under a specific topic vector t, user u has an influence on user v with probability p_{uv}^t. This probability contains two parts: internal/external probability and user preference correlation under the given topic. As shown in formula (5), if u is a super node, it will use the external probability to active v; otherwise, it will active v with the internal probability. C_u^t represents the preference of user u for topic t. F(C_u^t, C_v^t) represents the preference similarity of user u and v about topic t which can be estimated by the difference between the arithmetic mean of C_u^t and C_v^t and the standard deviation.

$$p_{uv}^t = \begin{cases} a \times p_{uv}^{in} + (1 - a) \times F(C_u^t, C_v^t) \, (u \neq g) \\ a \times p_{gv}^{ex} + (1 - a) \times F(C_g^t, C_v^t) \, (u = g) \end{cases} \tag{5}$$

3.3 Topic-Aware *IM* Problem

Given a graph $G = (V, E, W)$, the recommendation social network graph is described as $G^s = (V^s, E^s, W)$, here $V^s = \{g\} \cup V$, g represents the super node. The edges set $E^s = E \cup E'$, here E′ = {(g, v_i)} represents the influence from super node g to common nodes.

Given the recommendation social network G^s, topic distribute vector t, and budget k, ***topic-aware IM problem*** is to choose k seeds, the information propagate from these

seeds and the influence spread can be maximized. To get set $S^* = S^*(k,t)$, and $S^* = \arg\max_{|S| \le k, S \subseteq V}(\sigma(S \cup \{g\}, t))$, $\sigma(S \cup \{g\}, t)$ is the influence spread of S and super node g under topic t.

4 Topic-Aware Influence Maximization Algorithm TIP

4.1 Influence Spread

To compute the influence spread of common node u, we first analyze the ways that u will influence the other nodes. Nodes can be activated directly by user u if there is an edge between them. Nodes can also be activated indirectly by u if there exists path between them. The path from u to v ($v \neq u$) is $p_{u \to v} = \langle u, v_2, \ldots, v_m = v \rangle (m \geq 2)$. The influence probability of path p is:

$$pro(p) = \prod_{i=1}^{m-1} w(v_i, v_{i+1}) \tag{6}$$

After searching the paths from u, we calculate the influence probability of u on nodes which can be affected by it. Let $\sigma(u)$ be the number of nodes that can be influenced by u. The paths set $Path_{T \to v} = \{p | p = <u, \ldots, v>, u \in T\}$ contains all the paths starting from u. Accordingly, the paths set from node set T $\subseteq$ V to node v is $Path_{u \to T}$. The influence spread of node u is represented as $O_u = \{v | <\ldots, v> \in Path_{u \to v}\}$. The approximate influence of node u is

$$\sigma(u) = 1 + \sum_{v \in O_u} \sigma^v(u) \tag{7}$$

In the above formula, 1 is the influence of node u itself. $\sigma^v(u)$ is the probability of node u on the specific node $v \in O_u$. Given the path set $Path_{u \to v}$, the influence of u on v is the complement of all paths are failure, the formula is as follows:

$$\sigma^v(u) = 1 - \prod\nolimits_{p \in Path_{u \to v}} (1 - pro(p)) \tag{8}$$

The marginal influence of node u is represented as $MI(u) = \sigma(S \cup \{u\}) - \sigma(S)$. The marginal influence of u depends only on the sum of the influence from u to $v \in O_u \cup \{u\}$ as shown in the following formula:

$$MI(u) = 1 + \sum_{v \in O_u \cup \{u\}} MI^v(u) \tag{9}$$

Here $MI^v(u)$ is marginal influence from u to v, it can be computed as follows:

$$MI^v(u) = \sigma^v(S \cup \{u\}) - \sigma^v(S) \tag{10}$$

4.2 TIP Algorithm

In the TSID model, the super node can begin to active the user before the seed is selected, nodes that have been activated by the super node cannot be used as the seeds candidate. Super nodes can activate multiple nodes at the same time, these nodes may be far away, they start from different regions of the network at the same time, can quickly affect a wider range of network users.

The pseudo code of TIP algorithm is as follows:

Input: $G=(V^s, E^s, W)$, k, path threshold θ

Output: S seed nodes set

1) $S \leftarrow \phi$, $L \leftarrow \phi$
2) while $|S|<k$
3) for every $v \in V$
4) if (v is activated)
5) $L \leftarrow L \cup \{v\}$
6) endfor
7) for each $v \in V$
8) calculate the marginal influence $MI(v)$ of v
9) endfor
10) $s = \arg\max_{v \in V} MI(v)$
11) $S \leftarrow S \cup \{s\}$; $L \leftarrow L \cup \{s\}$
12) endwhile

return S

5 Experiments

5.1 Dataset

We use Douban network as our data. Douban data contains 485853 nodes and 4409997 edges. Contrast algorithms include C-Greedy and L_GAUP.

5.2 Experiment Results and Analysis

We do experiments to find the optimal threshold value θ trading off between the accuracy and efficiency. Figure 2 shows the effect of the threshold on the accuracy of the algorithm, Fig. 3 shows the impact on the time of the implementation.

As can be seen from Fig. 2, with the decrease of threshold, the influence spread will be increased. But when threshold θ = 1/320, the curve of growth slowed down significantly, that means the stable propagation point is here. Figure 3 shows the running time as the threshold decreases, it first increase a little, later at threshold θ = 1/640 the running time increases quickly. According to the above two figures, TIP algorithm can get the best compromise at θ = 1/320.

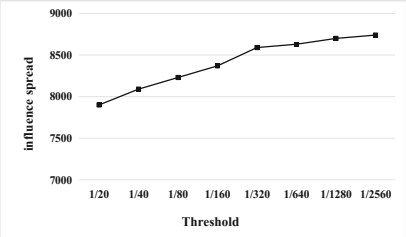

Fig. 2. Influence spread VS θ

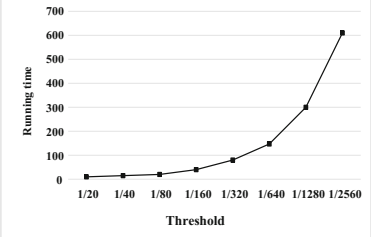

Fig. 3. Running time VS θ

We compare the accuracy and efficiency of TIP, G-Greedy and L_GAUP algorithm by varying the topic distribution. Figures 4 and 5 are respectively the influence spread and execution time of the algorithms. Figures 6 and 7 are the spread and time on multiple topics.

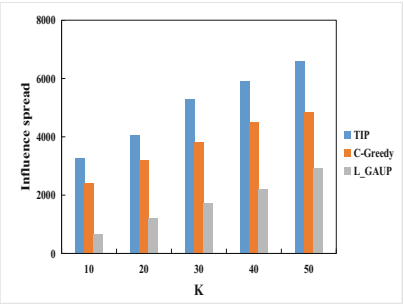

Fig. 4. Running time vs. K with one topic

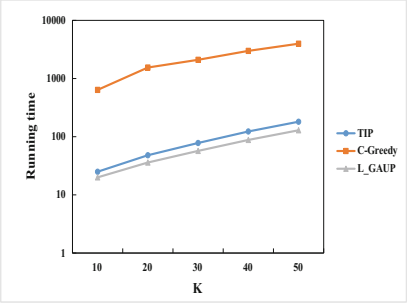

Fig. 5. Running time vs. K with one topic

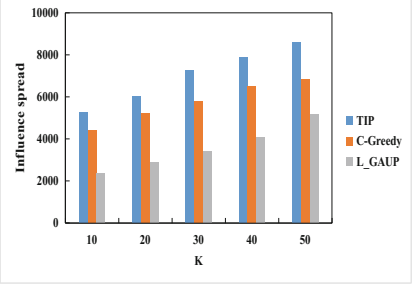

Fig. 6. Influence spread with multiple topic

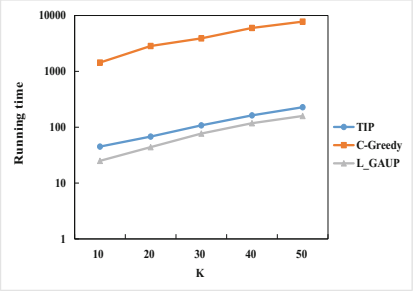

Fig. 7. Running time with multiple topic

As can be seen from Figs. 4 and 5, the spread of TIP algorithm is the largest, this is because it takes into account the impact of the super node. Figures 6 and 7 are the results of the execution time of three algorithms. TIP is two times faster than C-Greedy. We can see that TIP is more effective than traditional algorithm, it can not only ensure the time efficiency but also has a greater influence spread.

6 Conclusion

We propose a novel TSID propagation model in online recommendation social networks. TSID takes into account three impact factors during information diffusion: external influence from website, internal influence of pair wise individuals, and individuals' preference for topics. We induce the external and internal propagation probability in TSID model. Then we propose TIP algorithm to solve this problem by exploiting simple propagation path. The experiment results show that TSID model can well describe mixed topic information propagation for recommendation social networks and TIP performs well in terms of influence spread and response time.

Acknowledgment. This work was supported by the National Science Foundation of China (61632010, 61100048, 61370222), the Natural Science Foundation of Heilongjiang Province (F2016034), the Education Department of Heilongjiang Province (12531498).

References

1. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146, Washington, D.C., USA. ACM (2003)
2. Goyal, A., Wei, L., Lakshmanan, L.V.S.: SIMPATH: an efficient algorithm for influence maximization under the linear threshold model. In: 11th International Conference on Data Mining, pp. 211–220. IEEE Computer Society, Washington, D.C., USA (2011)
3. Lu, Z., Fan, L., Wu, W., et al.: Efficient influence spread estimation for influence maximization under the linear threshold model. Comput. Soc. Netw. **1**(1), 1–19 (2014)
4. You, Q., Hu, W., Wu, O.: Influence maximization in human-intervened social networks. In: 24th International Conference on Social Influence Analysis, IJCAI, pp. 9–14, Buenos Aires, Argentina (2015)
5. Barbieri, N., Bonchi, F., Manco, G.: Topic-aware social influence propagation models. In: 5th ACM International Conference on Web Search and Data Mining, pp. 81–90, Brussels, Belgium, New York, USA (2012)
6. Zhou, J., Zhang, Y., Cheng, J.: Preference-based mining of top-K influential nodes in social networks. Future Gener. Comput. Syst. **31**, 40–47 (2014)
7. Guo, J.F., Lv, J.G.: Influence maximization based on preference. J. Comput. Res. Dev. **52** (02), 533–541 (2015)
8. Chen, W., Lin, T., Yang, C.: Real-time topic-aware influence maximization using preprocessing. In: Thai, M., Nguyen, N., Shen, H. (eds.) CSoNet 2015. LNCS, vol. 9197, pp. 1–13. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21786-4_1
9. Zhu, J., Yin, X., Wang, Y., Li, J., Zhong, Y., Li, Y.: Structural holes theory-based influence maximization in social network. In: Ma, L., Khreishah, A., Zhang, Y., Yan, M. (eds.) WASA 2017. LNCS, vol. 10251, pp. 860–864. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60033-8_73
10. Chen, S., Fan, J., Li, G.: Online topic-aware influence maximization. Proc. VLDB Endow. **8** (6), 666–677 (2015)

11. Niu, J., Wang, D., Stojmenovic, M.: How does information diffuse in large recommendation social networks? IEEE Netw. **30**(4), 28–33 (2016)
12. Kim, J., Kim, S.K., Yu, H.: Scalable and parallelizable processing of influence maximization for large-scale social networks? In: 29th International Conference on Data Engineering, pp. 266–277. IEEE Computer Society, Washington, D.C. (2013)
13. Liu, X., Liao, X., Li, S., et al.: On the shoulders of giants: incremental influence maximization in evolving social networks. Comput. Sci. (2015)