The Model of Information Diffusion in Social Networking Service

Ilya Viksnin, Liubov Iurtaeva, Nikita Tursukov, Alexandr Muradov
ITMO University
Saint Petersburg, Russia
wixnin@mail.ru, lyubava.infa@yandex.ru, {stepingnik, ialexmur}@gmail.com

***Abstract*—The work includes a brief overview of the new trends in the network models, review their weakness as well as the description of the new improved model. It is important to note that the results provided by the model are close as possible to the actual distribution.**

I. Introduction

Social networks are an essential part of a today life. People use them to communicate with friends and family, companies use them to interact with employees, potential and existing customers. Social networks are used to advertise products and services, to promote ideas, to post news etc. The one of an efficient way to spread information among a huge number of people is make an announcement in a social network. Sometimes this way is even better than mass media. Due to an easy and rapid spread of information through social networks, unwanted information can spread. To stop the further spread it is important to find users who deal with the information. Thus, understanding the principles of information dissemination is necessary regardless of the objectives [1].

Scientists from different fields such that sociology, economics, marketing, physics and other, are interested in the mechanisms of dissemination of ideas and opinions in social groups. The existing works can be categorized into two groups: static and dynamic analysis. The first category focuses on the changing state of social networks in time using a variety of mathematical models [2]-[7]. The second category focuses on the stability of the network [8]. That means the model does not allow to connect new users or delete existing ones. This article refers to the second category. The most detailed overview of the existing models of the impact of the dissemination of information on social networks and their applicability is presented in work.

In a paper [11] a threshold distribution model, a weighted cascade model and a independent cascade model were discussed in detail, and different algorithms of distributing information on the threshold and cascade models were compared. In a paper [12] a model of information dissemination in social network Twitter was developed and tested. This article provides a great idea to increase and correct a linear distribution model that the authors used in this paper.

The proposed model shows the dynamics of the distribution of users by analyzing their interactions, behavior and properties of information. The most popular models of information dissemination in a stable network and their weaknesses are reviewed in the second part of this paper. The third part describes the proposed model and the constructions principles. The results are provided in the fourth section. Finally, fifth part infers the effectiveness of the developed model.

This project aimed at an effective model of the spread of information in social networks. That means data about information dissemination are most close to the real distribution. To achieve this goal we analyze existing models, identify their advantages and disadvantages, develop a new model and perform experiments to evaluate its effectiveness.

A social network can be modeled as a directed graph G=(V, E), where V={v_1,v_2,…,v_n} – a vertex set of the graph, and E={e_1,e_2,…,e_m} – a set of edges of the graph. The set of vertex is set of users in social network, and the set of edges of the graph is array of binary values, which show user’s subscriptions to other users. We say that user who reposted an investigated announcement has active state. Otherwise, the user has an inactive state. Impact influence is an ability to affect the desire to repost from another vertex.

II. Advantages of invented models

At the moment widespread models are threshold models, independent cascades models, infection models and models based on cellular automata.

In the threshold model, the only possible transition from active to inactive condition. Each node is influenced by each of active neighbor and it is activated if the condition (1) is true.

$$\sum_{j=1}^{n} w_{ij} > f_i \tag{1}$$

where i — the analyzing node, j — the neighboring node, w_{ij} — the influence of active neighboring node on analyzing node, f_i — the threshold number for the analyzing node.

Thus, if the total effect of neighbors of the considered node is more than its threshold value, it will become in active state.

The principle of model cascades is that when the node becomes active, it gets a chance to activate each of its neighbors with some probability. A node becomes active if the condition (2) is true.

ISSN 2305-7254

$$w_{ij} > f_i \qquad (2)$$

Each of the neighboring active node into the queue tries to activate the considered node.

In models of infection there is analogy with the epidemic [9]. Initially, some nodes are in the active state. Each node which converted into the active state remains infectious for a fixed number of steps t. At each step t all infected nodes with a certain probability infect susceptible neighbors. After t steps the node is not contagious and not receptive to information.

In models based on cellular automata nodes can be linked by strong or weak link. Node can activate by the influence of their neighbors or external environment. The influence at strong links is more than influence at weak links. There is condition (2) of activating nodes.

$$(1\text{–}\,(1\text{–}\,\alpha)(1\text{–}\,\beta_w)^j(1\text{—}\,\beta_s)^m) > f_i \qquad (3)$$

where β_W – the probability of influence at a strong connection, β_S – the probability of influence at a weak connection, α – the probability of influence at external environment, j — the number of vertices associated with vertex by strong connectivity, m — the number of vertices associated with vertex by weak connectivity.

The described models have several disadvantages. Next it will be discussed in detail.

First, the degree of influence of each vertex are different and depend on many factors in real life. In existing models the degree of influence for all vertices is set the same or on probability distribution. In the developed model for each vertex the degree of influence is calculated individually and is based on a quantitative measure of the activity of neighbor in relation to the user.

Second, the probability of further spread depends on the relevance of this information. Relevance of information trends to reduce over time. The probability of spreading old news is less than the probability of spreading of recent news.

Thirdly, it is necessary to consider activity of the user in the social network. Some users never have been active in social networks, and therefore the probability diffusion by them is low.

III. Model

Authors chose social network “Vkontakte” for developing and testing model. In this network users can post different news. If other users like this information, they can put a mark “like” on the news. If users want to share this news with some people, they can put a mark “repost” on the news, then their friends will see this news.

Based on these disadvantages new model was developed. According to the developed model vertex i which connected with j active neighbors at time t activates if the condition (4) execute.

$$\sum_{j=1}^{n} D_{ij} * A_i * L_t > f_i \qquad (4)$$

where D_{ij} – the influence of the neighbor j to the node i, A_i – the activity of the vertex i, L_t – the relevance of the information at time t.

The coefficient D is calculated according to the following rule (5).

$$D_{ij} = \frac{Like_{ij} + Repost_{ij} + Friends_{ij}}{Like_i + Repost_i + Friends_i} \qquad (5)$$

where $Like_{ij}$ — the number of "likes" on node i from node j, $Repost_{ij}$ — the number of "repost" on node i from node j, $Friends_{ij}$ — the number of common friends, $Like_i$ — the number of all "likes" at node i, $Repost_j$ — the number of all "repost" at node i, $Friends_i$ — the number of all friends at node i.

The coefficient A is calculated according to the following rule (6).

$$A_i = \frac{Repost_i}{Repost} \qquad (6)$$

where Repost — the number of "Repost" all adjacent to the vertex i.

The coefficient L is calculated according to the following rule (7).

$$L_t = LP_{t-1} - LP_t \qquad (7)$$

$$LP_t = \frac{Like_{t-1} + Repost_{t-1}}{PosLike_{t-1} + PosRepost_{t-1}} \qquad (8)$$

$Like_t$ — the number of "likes" for news at the time t, $Repost_t$ — the number of "repost" for news at the time t, $PosLike_t$ — the possible number of "likes" for news at the time t, $PosRepost_t$ — the possible number of "repost" for news at the time t.

When calculating the coefficients weight of each of mark "likes" is equal to 2, the mark "repost" is equal to 4, friend is equal to 1.

IV. Description of the experiment

Web service was developed for checking the model performance. There is a possibility to select a record at any user of the network "VKontakte", enter the users identifier in a social network "VKontakte", the text of the record and the threshold value for nodes. The result will be a list of people who really put the stamp "repost" to the record, a list of people who made a mark "repost" according to the developed model, and the percentage of models correctness (the parameter that will show how much the actual diffusion differs from the diffusion of the developed model).

For example, we selected one record from the some user. For anonymity of real people this user was labeled like user_1. Also other users, who are in this experiment, have a personal alias like user_n, where n – a number from 2 to 10. The threshold value we chose 0,00001. After entering the required information, we received the data collected in the network "VKontakte" about people, who did a repost of this record (Table I), and people, who did repost in accordance with the principles of the model (Table II).

TABLE I. REAL DIFFUSION INFORMATION

№	People with 'repost'
1	user_2
2	user_3
3	user_4
4	user_5
5	user_6
6	user_7
7	user_8
8	user_9

TABLE II. MODEL'S DIFFUSION INFORMATION

№	People with 'repost'
1	user_2
2	user_10
3	user_7
4	user_11
5	user_5
6	user_12
7	user_13
8	user_14
9	user_15
10	user_16
11	user_17
12	user_18

Also, the real distribution and the distribution on the basis of the developed model are presented in graph form (Fig.1, Fig.2).

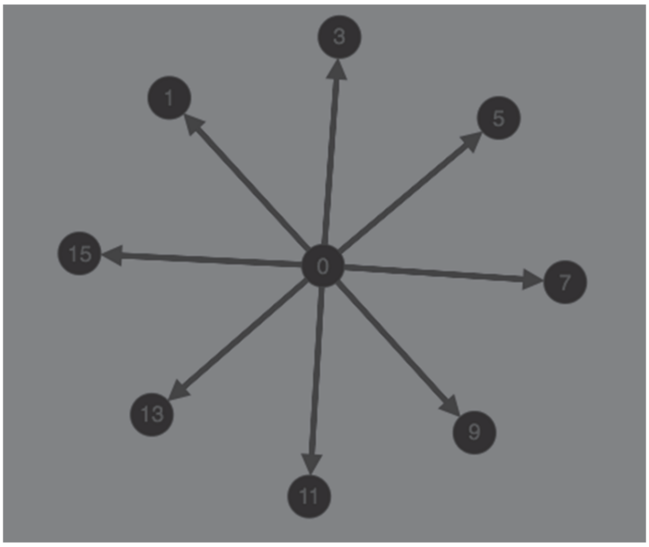

Fig. 1. Real diffusion information

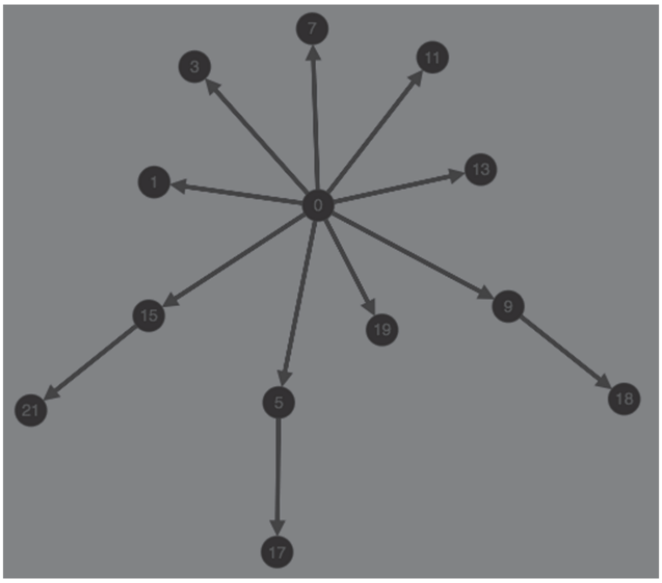

Fig. 2. Model's diffusion information

Main performance indicator is a percentage that displays how the data from table 2 differ from table 1. For computing indicator rule (9) is used.

$$p = r/R * (1 - v/V) \tag{9}$$

where r – a number right guessed users, R – a number of all users in Table I, v – a number wrong guessed people, V - a number of all users in Table II.

For choosing news r = 3, R = 8, v = 9, V = 12. Then the indicator is computed as 9,3%. The threshold value was chosen randomly. With another threshold value the result of the model and the indicator will be different.

Thus, knowing the real diffusion of the considered record, it is possible to find the most effective threshold value for this network. In future this will allow to predict the diffusion of other records with greater accuracy.

V. CONCLUSION

In this paper, we have developed a mathematical model of information diffusion in social networks. In the developed model it is possible to search the optimal thresholds for maximum efficiency.

In the future authors plan to change a network for modeling of diffusion information. Because network «VKontakte» has a lot of closed profile of users, that deteriorate a data set for modeling. Also in the network «VKontakte» people do not make a long chain of mark "repost". Listed shortcomings of the network „VKontakte" make the model less reliable.

REFERENCES

[1] F. Alvanaki et al. See what's enblogue: Real-time emergent topic identification in social media. In EDBT, pages 336–347, 2012.

[2] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in Proc. of the 13rd

international conference on World Wide Web (WWW), 2004, pp. 491–501.

[3] O. Yagan, D. Qian, J. Zhang, and D. Cochran, “Conjoining speeds up information diffusion in overlaying social-physical networks,” IEEE J. Sel. Areas Commun., vol. 31, no. 6, pp. 1038–1048, 2013.

[4] M. Kimura, K. Saito, and R. Nakano, “Extracting influential nodes for information diffusion on a social network,” in Prof. of AAAI Conference on Artificial Intelligence, 2007, pp. 1371– 1376.

[5] M. Kimura, K. Saito, and H. Motoda, “Blocking links to minimize contamination spread in a social network,” ACM Trans. Knowl. Discov. Data, vol. 3, no. 2, pp. 9:1–9:23, 2009.

[6] Y. Wang, G. Cong, G. Song, and K. Xie, “Community-based greedy algorithm for mining top-k influential nodes in mobile social networks,” in Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1039–1048.

[7] M. U. Ilyas, M. Z. Shafiq, A. X. Liu, and H. Radha, “A distributed and privacy preserving algorithm for identifying information hubs in social networks,” in Prof. IEEE INFOCOM, 2011, pp. 561–565.

[8] D. Centola, “The spread of behavior in an online social network experiment,” Science, vol. 329, no. 5996, pp. 1194–1197, 2010.

[9] M. Draief and L. Massouli. Epidemics and rumours in complex networks. Cambridge University Press, New York, 2010.

[10] M. Cha et al. Measuring user influence in Twitter: The million follower fallacy. In ICWSM, 2010.

[11] D. Kempe, J. Kleinberg and É. Tardos. “Maximizing the spread of influence through a social network”, ACM, New York, 2003.

[12] J. Yang, J. Leskovec, “Modeling Information Diffusion in Implicit Networks”, Stanford University, in Proceedings of IEEE International Conference on Data Mining, 2010.