ALGORITHM FOR ESTIMATING THE TIME OF POSTING MESSAGES ON VKONTAKTE ONLINE SOCIAL NETWORK

Gribanova Ekaterina, Savitsky Alexandr

Faculty of Control Systems
Tomsk state university of control systems and radioelectronics
e-mails: geb@asu.tusur.ru, bourbon7850@gmail.com
Russian Federation

Abstract: An algorithm for estimating the time of posting messages in online social network communities is presented. We will consider the formation of a comprehensive rating of the time of posting messages, including both user characteristics (periods of their individual activity, number of friends, number of communities, number of likes, reposts, comments) and the number of posts published by other participants.

Key words: online social networks, posting messages, posting times.

1. INTRODUCTION

Today online social networks are popular platforms for the promotion of goods and services [1, 2], dissemination of social and political information. Social networks include a large number of participants; therefore, a competently determined method of posting information can result in its wide dissemination. And the information published by participants about themselves, entry into communities of interest help to identify the target audience.

One of the most important factors determining the number of times a message is viewed is the time it was posted. When choosing the moment of posting a message, it is necessary to analyze a number of indicators. The best time for posting a message is considered to be the moment when a larger number of users are online, as in this case they are likely to see this information. In addition, when determining the moment of posting, a number of other parameters should be taken into account, such as, for example, the number of messages posted by other participants on the social network.

In Russia the most popular social networks are Facebook, VKontakte, Twitter, and Odnoklassniki. At the beginning of 2019, social media in Russia had the following monthly coverage: VK – 38.1 million people, Odnoklassniki – 23.8 million people, Facebook – 22.3 million people, Twitter – 8.4 million people. In November 2019, users published 556,380 posts on VK, 119,470 posts – on Odnoklassniki, 52,859 posts – on Facebook, 32,214 posts – on Twitter. VKontakte social network covers the largest audience; in this paper a study will be performed using the data from this network.

2. LITERATURE REVIEW

We should mention the following existing scientific works in the field of studying the time of posting messages on social networks. A method was proposed in [3]; according to it each user gets a certain rating depending on the number of subscribers, the time elapsed since the time of posting the last message and interaction with other users, and the best time is determined from the point of view of the maximum total rating.

Article [4] describes the study, which collected data on 37 thousand Brazilian users of Orkut, MySpace, Hi5, and Linkedln social networks. The data were collected using a social media aggregator, which is designed to manage social media contents via a single interface. The data set contains 4,894,924 HTTP requests, including timing, HTTP status, IP address, user IDs, cookies, incoming and outgoing traffic. According to the information received over 12 days, the largest number of participants having the online status is observed at 3PM. The paper also analyzes other types of user activities, such as viewing followers' pages, sending messages, time on site, etc. The authors argue that the information presented in the paper is useful for developers of social media management services (e.g., for improving a web design) and advertising agencies.

Paper [5] considers the issue of setting the problem of determining the time interval during which the posted message will have the largest number of views and also discusses systems for automatic posting of information on social networks (BuzzLike, Sociate, Feedman, EcoTime). The authors developed a system to collect data on published posts using the VK API techniques, analyzed the collected data to determine the best time for posting, and published posts at a given point in time. The authors formulated an optimization problem to determine the best time for posting: maximizing the user response rate in relation to a post (the number of likes and comments) for a given time interval. This approach, however, does not take into account that users' responses can be determined to a greater extent by a post itself and its ability to arouse interest in the target audience, rather than time for posting.

Article [6] presents the results of a correlation analysis of user involvement depending on various factors to determine the best strategy for posting messages on Facebook network. In particular, results were obtained according to which the greatest involvement is from 8PM to 10PM. The data from the Tufts University Hirsh Health Sciences Library page were used in the study. The authors' main purpose is to make this page more interesting to users and, accordingly, increase their activity. As a result, a number of conclusions were made. In particular, a higher posting frequency correlates with a larger number of likes and a higher engagement rate. The analysis also shows that the engagement rate in the summer season does not significantly decrease compared to the academic year.

The authors of [7] consider a game model, in which advertisers compete with each other determining the time of posting.

There are also works examining the dissemination of information on a social network depending on the structure of the network, characteristics of individual participants (influence, opinion, etc.) [8–11]. For example, work [12] presents the results of a study in the field of forecasting the speed and degree of information dissemination on Twitter social network using the regression model. Work [13] is concerned with the study of the influence of participants in the course of content distribution on Second Life social network. Article [14] analyzed the influence of Twitter users using three indicators: the number of subscribers, retweets and mentions by other users.

In addition, there are studies and recommendations of analytical agencies (Sprout Social, Popsters, Latergramme). However, such studies are general in nature and do not take into account specific communities whose members may live in different cities and have a different pace of living depending on their social status (students, employees, holidaymakers, etc.). In contrast, studies of individual groups and pages [4, 6] are more specific. The findings can be useful for communities whose target audiences are similar to the group under study.

So, the existing works are based on the use of user characteristics to determine the best time (online status, publication of messages, etc.). In this paper, unlike the existing ones, the formation of a comprehensive rating is considered. It includes both user characteristics (including periods of their individual activity) and the publication of messages by other participants.

Based on the developed model, it is necessary to analyze the social media community to choose the best time for posting in order to increase the number of views and, therefore, the engagement rate, which determines the further development of the group (the author of the paper is the administrator of this community).

In connection with the need to process a large amount of information in order to obtain results, it was also necessary to develop a program that would automate the process of information collection and construction of a rating of the time of posting messages.

3. MESSAGE POSTING TIME ALGORITHM

Messages posted by communities and participants to which the user is subscribed are displayed in his/her news feed. As part of the work a survey on topic

"Do you manage to view the news feed and how many communities are you in?" was conducted in the VKontakte community under consideration, in which 211 people took part. The following results were obtained:

- "I browse the entire news feed" – 37.44%
- "I browse only part of the news feed" - 48.82%.
- "I do not browse the news feed" - 2.84%.

Based on the results obtained it can be concluded that most participants view only part of the news feed, so posting a message at the time when users are online or will be online in the near future will increase the likelihood of viewing (otherwise the message may be displaced by newer ones and not read).

At the same time community members are distinguished by potential opportunities for further dissemination of information. So, the following characteristics of the participants were considered:

- Number of friends (x_1). The more friends the participant has, the more people he can convey the information received.
- Activity (x_2). The higher the user's activity, the greater the likelihood that he/she will share information with subscribers and friends on his page. The number of likes, reposts and comments for the last month on the user's page was used in this work as an indicator of activity.
- Number of communities (x_3). If a person is subscribed to a large number of communities, then the likelihood that the message will remain unnoticed by him/her and be displaced by others increases, if it was posted long before the participant got the online status.

To obtain an integral rating of the participant *u*, it is necessary to normalize the values and perform their linear convolution:

$$y_1(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

where $y_1(x)$ is the normalized value of *x*.
x_{max}, x_{min} are the maximum and minimum values of *x*, respectively.

The integral rating of the participant *u* will be calculated by the formula:

$$u = k_1 \cdot y_1(x_1) + k_2 \cdot y_1(x_2) + k_3 \cdot y_1(x_3) \quad (2)$$

where k_1, k_2, k_3 are coefficients of importance of indicators x_1, x_2, x_3 set by the expert, respectively.

The participant can hide the data of his/her page; in this case the rating *u* is equal to zero.

When assessing the possibility of reading a message, one can use the participant status: it is believed that the message will be viewed with a higher probability, if at the time of its publication the person is online. However, this approach is simplified because a person can view the message even if he/she will be online some time after publication. With an increase in the time interval from the time of message posting to the time when the participant gets the online status, the likelihood of viewing the message decreases, because it may be displaced by others in his/her news feed. To form an indicator based on the periods of activity of the participant, the numerical value that characterizes the distance to the point in time when the participant will be online shall be determined. Let's consider an example of calculating this indicator using the source data contained in Table 1. The status of the participant takes two values: 1 - the participant is online, 0 - the participant is offline. Let's assign the participant at each moment in time the number equal to the sum of one and the number of periods until the participant gets the online status (Table 2). The average value of the indicator of location relative to the online status is determined as its total value for a few days under study divided by the number of days. Based on the calculated value you can determine the inverse value g characterizing the share of one period in the total value of the indicator:

$$g = \frac{1}{p} \tag{3}$$

where p is an indicator of the average number of periods before the online status.

The resulting value will be in the range from 0 to 1. The value will be equal to 1, if the user is currently online; the longer it takes for the user to become online, the closer the value will be to zero.

Table 1. Initial data on the status of participants

Id	**12:00 a.m.**	**12:05 a.m.**	**12:10 a.m.**	**12:15 a.m.**	**…**	**11:55 p.m.**
62845	*0*	*0*	*1*	*1*	…	*1*
67654	*1*	*0*	*0*	*1*	…	*0*
87980	*0*	*1*	*0*	*1*	…	*1*

Table 2. Indicators of the number of periods before the participant gets the online status

Id	**12:00 a.m.**	**12:05 a.m.**	**12:10 a.m.**	**12:15 a.m.**	**…**	**11:55 p.m.**
62845	*3*	*2*	*1*	*1*	…	*1*
67654	*1*	*3*	*2*	*1*	…	*2*
87980	*2*	*1*	*2*	*1*	…	*1*

Finally, the total rating of the participant is calculated by the formula:

$$r = g(1+u) \tag{4}$$

The total community rating is determined as the sum of ratings of all community members:

$$Q = \sum_{i=1}^{n} r_i \tag{5}$$

where n is the number of community members.

This rating is calculated for participants at time points with a given interval.

When posting a message, not only a high total rating at the current point of time is important. This rating shall remain high for some time. So, in the next coming periods other participants may log in and view the message. To take this condition into account, it is possible to calculate the general characteristics of the total rating for several time intervals with the assignment of weights to each moment in time. For example, the calculation of such a characteristic with a proportional linear decrease in the importance of the following values will look as follows:

$$H = \sum_{j=1}^{s} \frac{1}{j} \sum_{i=1}^{m} Q_i \tag{6}$$

where m is the number of time intervals with which the indicator is studied within one hour.

s is the number of hours in the future that will be taken into account when calculating the rating.

So, coefficient 1 will correspond to the rating calculated at the moment under study, ½ - calculated at the next hour, etc. As it moves away from the current moment, the coefficient will decrease, however, future periods will be taken into account when forming the rating for the current period.

When choosing the time of posting a message, the speed of updating the news feed, where messages of participants and communities to which the user is subscribed are posted, also plays a significant role. If the speed of the news feed is very high, the posted message may be displaced by other posts and the user may not read it. To determine the speed of news feed updating, you need to set the interval b (for example, 5 minutes) and determine how many messages were posted during this interval (x_4). The resulting value is normalized by the formula:

$$y_2(x) = \frac{x_{\max} - x}{x_{\max} - x_{\min}} \tag{7}$$

The value of the total rating (6) is normalized by the formula (1) and a linear convolution of two indicators is performed: the normalized characteristic of the total rating and the normalized value of the number of published messages. As a result an integral rating of the moment of time I will be obtained:

$$I = k_4 \cdot y_1(H) + k_5 \cdot y_2(x_4) \tag{8}$$

where k_4, k_5 are the coefficients of importance of indicators $y_1(H)$, $y_2(x_4)$ established by the expert, respectively.

So, the algorithm for estimating the time of posting messages will include the following steps:

Step 1. Collection of data on the number of messages posted in the news feed.

Step 2. Collection of id and status of community members.

Step 3. Collection of id, the number of friends, the number of communities, the number of likes, reposts and comments for the last month on the user's wall.

Step 4. Calculation of shares of individual periods of user activity (based on the data obtained at step 2)

Step 5. Calculation of the total rating of community members (based on the data obtained at steps 3 and 4).

Step 6. Processing the results and obtaining the integral indicator *I* of the time the message was posted using linear convolution (based on the data obtained at steps 1 and 5).

Step 7. Determination of the moment for which the value of indicator *I* is maximum.

The first three steps are aimed at collecting the data necessary for the following steps and may be performed in any order. Steps 4, 5, 6, 7 should be performed sequentially.

To implement the algorithm, a program was developed using C# language. It uses VKontakte API methods, such as communities. getMembers, users.get, wall.get, newsfeed.get, and execute. The data obtained at the steps are recorded, stored and processed in an Excel document.

When developing systems that interact with social media, developers face a problem of data transfer security. For example, hackers may steal sensitive information (messages, passwords) through hidden data collection [4]. This paper considers the collection of open source data which are available for viewing by social media members.

4. EXPERIMENTAL RESULTS AND DISCUSSION

To conduct an analysis of the community, information about members of VKontakte social network community “Golden time Anime” was used. At the time of collecting user data, there were 8,135 members in the community. Data on subscribers were collected from October 11, 2019 to October 18, 2019. The program automatically collected data about community members every 5 minutes.

Figure 1 shows the average share of community members who are online. You can see that in the morning this share is minimum, and in the evening and at night it is maximum. The results of this study differ from similar indicators for other social media (Facebook, Twitter, etc.). According to studies by special agencies (e.g., Sprout Social), users of these social media are most active from 8 a.m. up to 4 p.m. According to Rusability agency, users of VK are most active between 5 p.m. and 6 p.m. These data also do not strongly correlate with the results for the group under study.

Figure 2 shows the average number of messages posted in the news feed (data collection interval is 5 minutes). It can be noted that in the first minutes of every hour there is a sharp increase in the number of messages posted in the news feed, which may be due to automatic posting of messages by communities at the beginning of each hour. Therefore, we can conclude that posting a message in the first minutes of each hour can lead to its displacement by other posts in the feed. These results are consistent with the recommendations from established analytical agencies (e.g., Popsters): they claim that bulk posting of ads takes place at the beginning of each hour. Therefore you should not publish your posts at that time. So, the task of determining the best time to post a message can include a restriction on the time left till the beginning of the next hour:

$$\begin{aligned} & I(t_i) \to \max \\ & (hour(t_i)+1) - t_i \geq w. \end{aligned} \tag{9}$$

t_i is the i-th moment of time,
$hour(t_i)$ - the function that returns the hour corresponding to the time point t_i.
w is the minimum value until the beginning of the next hour.

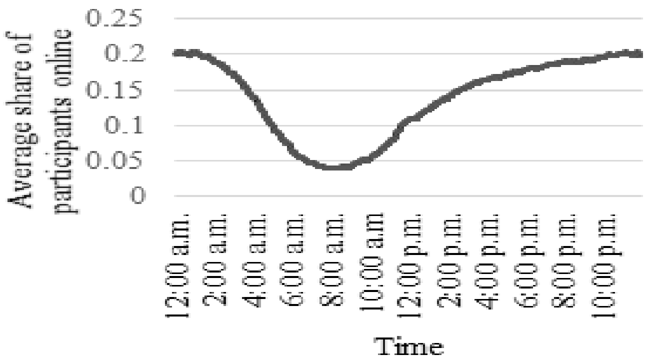

Fig. 1. The average share of online community members

Moreover, it is seen from Figure 2 that the maximum number of messages is also posted in the evening and at night.

The number of posted messages for other values of the interval b (from 10 minutes to 55 minutes in increments of 5 minutes: 10, 15, 20, etc.) was also calculated.

Ten best values of the integral indicator and the corresponding time points for different values of the time interval b are presented in Table 3. In this case equal importance coefficients were used for linear convolution ($k_1=k_2=k_3=0.33$, $k_4=k_5=0.5$), the number of periods taken into account in the future is 4 hours ($s=4$).

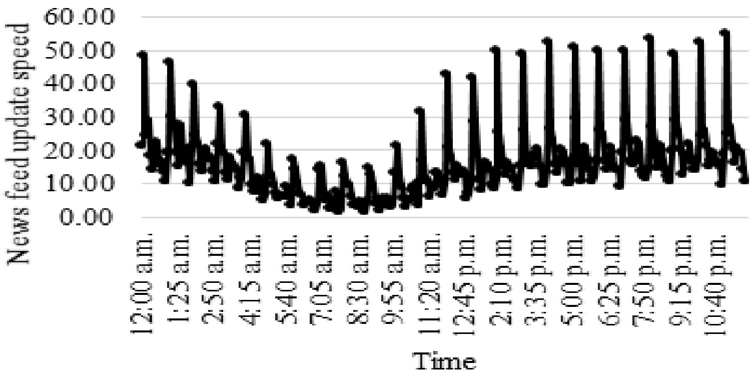

Fig. 2. Average number of posted messages

Figure 3 shows a graph of changes in the integral indicator at *b*= 10 min.

Table 3. Values of the integral indicator

b, minutes	***I***	t_i
5	*0.911*	*8:50 p.m.*
10	*0.871*	*8:45 p.m.*
15	*0.938*	*8:40 p.m.*
20	*0.816*	*8:35 p.m.*
25	*0.787*	*8:35 p.m.*
30	*0.738*	*8:25 p.m.*

If it is required to determine the point in time with another ratio of normalized values, you can change the values of importance coefficients. So, if the characteristic of the total average rating of participants is of higher priority than the average number of posted messages, then it is possible to set such values of the importance coefficients to make the condition satisfied $k_4 > k_5$.

Another method presupposes solving the inverse problem by indicating coefficients of relative priority for changing indicators [15]. Let's consider the use of this apparatus for discrete data. So, for *b*=10 min normalized values of characteristics of areas and the number of posted messages corresponding to the best value of the integral indicator are equal to $h = y_1(H)$ = 0.992, $f = y_2(x_4)$= 0.750, respectively (the corresponding time point is 8:45PM).

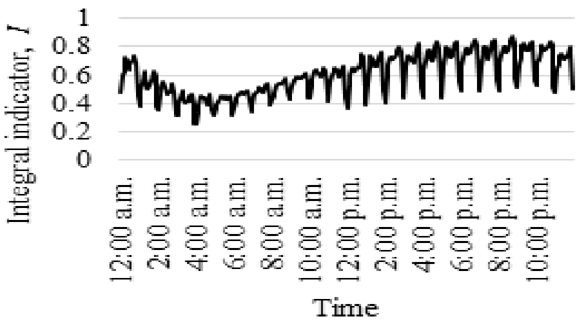

Fig. 3. Graph of changes in the integral indicator (b=10 min.)

Let it be necessary to change normalized values in such a way as to improve the rating of the news feed update speed due to rating decrease. The total change in normalized values should be –0.015, while the relative priority coefficients will be 0.7 and 0.3 (for the participants and the speed of news feed updating, respectively). Then the system will look as follows [15]:

$$\begin{cases} \frac{\Delta h}{\Delta f} = -\frac{0.7}{0.3} \\ (h + \Delta h) + (f + \Delta f) = 1.742 - 0.015 \end{cases}$$

The solution to the system will be $\Delta h = -0.026$, $\Delta f = 0.011$. Then the new values of normalized values will be:

$$h^* = h + \Delta h = 0.966,$$

$$f^* = f + \Delta f = 0.761.$$

After that the time point t_i is determined for which the deviation of the obtained normalized values from the initial values is minimum:

$$\sqrt{\left(h(t_i) - h^*\right)^2 + \left(f(t_i) - f^*\right)^2} \to \min \quad (10)$$

The minimum distance value equal to 0.026 corresponds to time point 7:45PM.

5. CONCLUSIONS

The article proposes an algorithm for estimating the time of posting messages in communities of an online social network. In contrast to the existing works, the formation of a comprehensive rating is considered in this paper. It includes periods of individual activity of participants and their individual characteristics (number of friends, communities, posts on the wall, etc.), as well as the number of messages published by other participants on the social network. Estimation models based on the use of individual characteristics (for example, the number of users online) do not take into account all the factors that form the decision to post a message, however, determination of a comprehensive rating also has difficulties associated with the need to process a large amount of data and attract expert information (in this paper the information was determined based on the experience of community administration). A linear convolution was applied as one of the methods for solving the multicriteria optimization problem to get the integral characteristic.

The algorithm was implemented using C# language; Microsoft Excel spreadsheet processor was used to store data about subscribers and results. An example of estimating the time of posting messages in VKontakte social network community was considered, as well as the solution of the inverse problem using discrete data with the help of a modified inverse calculation method.

The administrator of the online social media community used the developed algorithm to estimate the real time for posting in the group and choose the best time for posting. The presented numerical results can also be useful in choosing the time of posting messages. So, according to the data received, a large number of messages are published at the beginning of each hour, so this posting time is not optimum. For the community under study high activity is observed from 9PM to 2AM. However, this indicator will differ for different communities, as the time of online activity of participants will depend on their location and lifestyle.

REFERENCES

[1] Goyal, S., J. Gagnon. Social networks and the firm. *Revista de Administracao*, **51**, 2016, pp.240–243.

[2] Gribanova, E.B., I. Shirenkov, A. Katasonova. Econometric models for evaluation of marketing activities` indicators of social network. *Proceedings of the conference «Trends of Technologies and Innovations in Economic and Social Studies»*, Tomsk, June 2017, pp.227–233.

[3] Booth, J.A., J. Howard, A. Rankin. System and Methods for Generating Optimal Post Times for Social Networking Sites. *United States Patent*, US 0275348 A1, 2013, pp.1–8.

[4] Benevenuto, F., T. Rodrigues, M. Cha, V. Almeida. Characterizing User Behavior in Online Social Networks. *Proceedings of the 9th ACM conference on Internet measurement*, Chicago, November 2009, pp.49–62.

[5] Chumak, A.A, S. Ukustov, A. Kravets, J. Voronin. Social Networks Message Posting Support Module. *World Applied Sciences Journal*, **24**, 2013, pp.191–195.

[6] Houk, K.M., K. Thornhill. Using Facebook Page Insights Data to Determine Posting Best Practices in an Academic Health Sciences Library. *Journal of Web Librarianship*, **7:4**, 2013, pp.372–388.

[7] Altman, E., N. Shimkin. Strategic Posting Times over a Shared Publication Medium. *International Conference on NETwork Games COntrol and OPtimization*, Trento, October 2014, pp.1-9.

[8] Bhagat, S., A. Goyal, L. Lakshmanan. Maximizing product adoption in social. *Proceedings of the 5-th ACM International Conference on Web Search and Data Mining*, Washington, February 2012, pp.603–612.

[9] Bakshy, E., J. Hofman, W. Mason, D. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. *Proceedings of the 4-th International Conference on Web Search and Web Data Mining*, Hong Kong, February 2011, pp.1–10.

[10] Chen, W., A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, W. Wei, Y. Wang, Y. Yuan. Influence maximization in social networks when negative opinions may emerge and propagate. *Proceedings of the 2011 SIAM International Conference on Data Mining*, Arizona, April 2011, pp.379–390.

[11] Gribanova, E.B., I. Logvin, I. Shirenkov. Algorithm for evaluating the marketing activities of the online social network Vkontakte based on the cascade model of information dissemination. *Proceedings of TUSUR*, **3**, 2018, pp.66–73.

[12] Yang, J., S. Counts. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, May 2010, pp.355-358.

[13] Bakshy, E., B. Karrer, L. Adamic. Social influence and the diffusion of user-created content. *Proceedings of the tenth ACM conference on Electronic commerce*, Stanford, July 2009, pp.325–334.

[14] Cha, M., H. Haddadi, F. Benevenuto, K. Gummadi. Measuring user influence in Twitter: the million follower fallacy. *Proceedings of the fourth international AAAI conference on weblogs and social media*, Washington, May 2010, pp.10–14.

[15] Odintsov, B.E. *Inverse calculations in the formation of economic decisions*, Finance and Statistics, Moscow, 2004.

Information about the authors:

Gribanova Ekaterina - Associate Professor of the Department of Automated Control System Tomsk state university of control systems and radioelectronics. Areas of Scientific Research: economic modeling, information technology, optimization methods, inverse tasks.

Savitsky Alexandr – Master student of the Department of Automated Control System Tomsk state university of control systems and radioelectronics. Areas of Scientific Research: social network analysis, information technology.

Manuscript received on 29 November 2019