Potential of the dynamic approach to data analysis

Vera Orlova[1,*], *Vyacheslav* Goiko[2], *Yulia* Alexandrova[2], and *Evgeny* Petrov[2]

[1]Tomsk State University of Control Systems and Radioelectronics, 634050, Lenin Avenue, 40, Tomsk, Russia
[2]Tomsk State University, 634050, Lenin Avenue, 36, Tomsk, Russia

Abstract. Explores the potential of a dynamic data analysis approach to study user behavior in social networks. Currently, information appears on social networks that allows differentiating user groups by their activity within the technical capabilities of a particular social network. The description of the information field of Tomsk is presented, a brief analysis is given. A dynamic approach to the study of user behavior, the structure of nodes and connections of social networks makes it possible to identify the rate of growth or decrease in the size of the network, the redistribution of connections between groups. There are four main stages in the analysis of social networks: 1) data collection; 2) selection of data for analysis; 3) selection and application of the analysis method; and 4) drawing conclusions. To obtain a complete picture of the information field of the Tomsk region, posts for 2019 were unloaded from all regional communities. All posts were classified based on training sample and specialized machine learning algorithm.

1 Intoduction

A huge amount of heterogeneous data (text, sound, images, etc.) is published on the Internet on a daily basis, which are generated by traditional, postal services and social networks through stationary computers and mobile devices to solve the communication needs of the population. In addition, the information transmitted provides an excellent opportunity for in-depth analysis, forecasting, analytics using artificial intelligence and big data. Technical improvements of networks allow, in addition to text information, to transmit voice and video information, as well as to leave a formalized opinion about the open part of the correspondence, to leave text comments.

Currently, there is a period of explosive growth of Internet services for various purposes (social networks (Facebook, Telegram, Twitter, VK, etc.), online stores (Ali-Express, Amazon, etc.), scientific communities (Researchgate, Linkelnd, etc.)). These Internet services occupy an important place in the communication of various groups of the population of many countries, but still do not solve any narrow tasks for which narrowly focused services are created, including most of the social services. On the other hand, there

* Corresponding author: orlov5508@rambler.ru

is an attachment of certain groups of the population to a certain service, which forces them to have several profiles in various services to communicate with different groups of the population. The functions of each population group in social networks are significantly different and therefore, the detection of profiles belonging to one person in several services (including social networks) allows you to get information about the direction of communication. Such information is in demand in many practical tasks of information retrieval, telecommunications companies, recommendation systems, etc.

Among the applied research, we note the work of S. Keisler, L. Sproul (2001), M. Castells (2009), who devoted their work to the study of various aspects of social networks. A significant contribution to the study of new possibilities of information communication technologies was made by Russian scientists: Ya.N. Zasursky (1999), D.A. Gubanov, D.A. Novikov (2010, 2014), Leshchenko (2011). Social media is a mass media space where user identification markers such as interests, preferences, moods, and integration vectors are found. The use of digital technology products is becoming the norm in everyday life, therefore, it is necessary to develop tools for studying the impact and consequences on people's behavior, the structure of their value orientations, and behavioral strategies.

At present, the dynamics of social events is quite high, and in this regard, social networks allow the events of one country and even a person to be turned into information for the whole world. This makes it possible to use this information not only for scientific, but also commercial purposes, which has made it possible to achieve significant progress in the field of social network analysis. However, most of the well-known work focuses on studying static situations in social networks or assessing dynamics on a global scale (for example, the spread of Covid disease is an example). In recent years, the availability of large dynamic datasets of social networks has grown significantly, which has fueled interest in developing automated approaches for analyzing temporal events in social networks.

A dynamic approach to the study of user behavior, the structure of nodes and connections of social networks makes it possible to identify the rate of growth or decrease in the size of the network, the redistribution of connections between groups, etc. A quantitative measure for assessing these indicators allows you to determine the pattern of changes and, accordingly, build predictive situations for the formation of certain connections in social networks. It is clear that in order to identify the dynamics of changes, it is important to assess the time intervals that determine a clearly marked change. The development of techniques for visualizing the network structure at the current time and comparison with past time intervals provides an opportunity for a more accurate understanding of trends.

Dynamic analysis of social networks (DAS) is an emerging area where there is significant potential for research and development of analytical software applications. DAS is aimed at analyzing the behavior of social networks at different time scales [1], detecting repetitive patterns [2], the structure of the community (formation, development, existence or dissolution) [3].

Traditional analysis of social network data is performed on a series of nodes and edges [4], usually obtained from metadata about interactions between several network participants, without actually analyzing the content of these interactions (messages). For these purposes, you can use information from the databases described above (see Table 1), or from the current data set obtained by the corresponding programs. If there is such a possibility (open social networks), then it is possible to combine metadata with data of the informational content of each message. Further, using the above software products, you can proceed to the analysis of data from social networks. The analysis allows you to get a feature that describes the behavior of network subjects (users and groups), their moods, as well as changes in a particular trend over time. In addition, having historical data about the network, it becomes possible to analyze its dynamics, as well as predict the hidden

relationships that exist in the data set. Clustering of communities based on behavior over time can be carried out by analyzing only metadata or joint analysis with the content of messages [5].

Since the analysis of social networks focuses on the study of connections and their dynamics, but also to a certain extent allows you to assess user behavior. There are four main stages in the analysis of social networks: 1) data collection; 2) selection of data for analysis; 3) selection and application of a method of analysis; and 4) drawing conclusions. In order to identify and investigate patterns that arise in the network, first the selection of groups of people must be made. The ability to analyze each network node (especially for large and heterogeneous networks) is significantly limited by the available computing resources and therefore it is necessary to choose from the general sample, at the initial stage, only a representative group of users for further analysis. [6] The next step in the analysis of social networks is to choose the most suitable analysis method, and there are three main approaches here: 1) full network or methods that are based on collecting and researching data across the entire network (each user and his posts). This approach gives the best analysis results, but is the most time consuming and difficult to collect complete data; 2) snowball methods, when the work starts with one local user or a small sample of users. For each user of the network, some or all of their connections with other network participants must be found and classified, and the process ends when there are no new links for classification or after the number of specified iterations of the algorithm ends. This approach works well for analyzing a highly connected group of users in a large network, but it has several weaknesses. One of them is associated with the choice of an isolated group member, according to some parameters (communication time, number of messages, type of messages, etc.), which cannot be considered passive and weeded out at the second stage of analysis, and secondly, it is necessary to look for the most active by the connections of the user (or group), which can hardly be found by this. method. Second, if the first actor is not properly selected, the method can lead to nothing like that. Choosing a user by chance gives an incomplete picture of the entire network; 3) the egocentric method allows you to explore only one user and the flow of communication with his environment. This method provides useful information only for the local network and its impact on the user [7].

Case Tomsk

To obtain a complete picture of the information field of the Tomsk region, posts for 2019 were unloaded from all regional communities. All posts were classified on the basis of a training sample and a specialized machine learning algorithm, which were previously compiled and developed for the tasks of the Center for Applied Analysis of Big Data in order to be used to analyze textual content of social networks. One of the categories is "garbage", which includes advertising, spam, various contests and sweepstakes, these posts were deleted and were not included in the sample. Unique informative posts were left for further analysis. An array of data in social networks was analyzed in 2019 (posts of communities of the city of Tomsk). In 2019, 51,553 unique and meaningful posts were uploaded. Using machine learning, the posts were also clustered into three spheres of society: social, economic and political.

Most of all, in 2019, Tomsk residents were interested in the social sphere (86%), in two there was significantly less interest - 8% - economic, 6% - political.

The overwhelming majority of the attitude to events covered in posts was defined as neutral (75%). In a negative way, 15% of posts were noted, in a positive way - 10%.

The respondents demonstrate the distribution of sentiment by thematic categories, and we see that each of them is characterized by a repetition of the general structure of sentiment. At the same time, we note that the largest share of negativity is inherent in the political sphere.

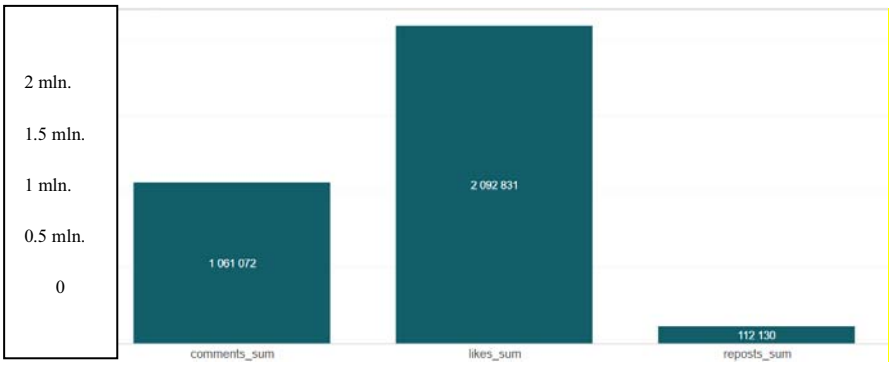

Fig. 1. The sum of comments, likes and reposts on posts of communities in Tomsk.

In terms of posts, the topic “Security” is confidently leading. The second and third places were taken by the topics "Infrastructure" and "Housing and Utilities", respectively. Combining this with the data that most posts were neutral, we can say that people were calmly interested in some mundane issues related to everyday life in the city.

The number of garbage posts is 2 times less than informative, which allows you to get a more detailed picture of what was of interest to Tomsk residents in 2019.

In regional communities of the Tomsk region, posts for 2019 were selected for further research. All posts were classified on the basis of a training sample and a specialized machine learning algorithm, which were previously compiled and developed for the tasks of the Center for Applied Analysis of Big Data in order to be used to analyze textual content of social networks. One of the categories is "junk", which includes advertising, spam, various contests and sweepstakes. All posts in this category ("trash") were deleted and did not participate in further analysis, and unique informative posts were left for further analysis. After receiving a list of informative and unique posts, a semantic analysis of texts was performed using the PolyAnalyst text analytics tool. Key concepts and individual keywords were identified and extracted. With the help of the tool "Relationship of terms" semantic connections between words are built and presented in the form of graphs [8].

Tomsk and Tomsk Oblast are “united” by the common term “source” for them, which, however, is used here in a variety of senses, both as, for example, “source of power” and as “source of fire”.

On the graph of connections of the most frequently used words, the words “man” and “woman” are clearly visible, here we note that a man is mentioned 3,654 times, and a woman - 2,485 times. Upon a detailed examination of these clusters, we see that in posts with the mention of the word “man "Often there are words:" police "," police officers "," committing a crime "," regional Ministry of Internal Affairs "," court ", from" positive associations - "real (man)": Also, with the help of the PolyAnalyst text analytics tool, the so-called “entities” - persons and organizations - are highlighted [9].

The next graph demonstrates the greatest positiveness in the coverage of the Safe and High-Quality Roads program in the region in the social network for the formation of a positive media image of the authorities in the region [10].

Among the TOP-100 posts, the topic of the export of timber by the Chinese from the region stands out, and this issue was raised as an independent one, for example, “The court banned a Chinese company from logging in 3 districts of the Tomsk region”, and in the context of other events: “In the Tomsk region, where the Chinese every day Thousands of cubic meters of timber are being taken out, and a worker of an electrical engineering plant was fined 15 thousand rubles for bringing a Christmas tree home to children [11]. "

Residents of the region publish photo reports about the sites of deforestation, taken with a quadrocopter, the question of forest fires that are "unprofitable" to extinguish for officials, etc. is raised, there are publications on the topic of deforestation criticizing the regional authorities [12].

Thus, socially active users try to draw attention to this topic. However, it is worth noting the fact of injecting false information - the publication of photographs of felling and forest fires passed off as the Tomsk region, because social networks are also a currently available tool for manipulating public opinion. In a large number of materials from users of the social network "Vkontakte" there are complaints about the provision of medical care (ambulances, hospitals). Tomsk residents note the incompetence of the medical staff, unwillingness to provide assistance, rude attitude, etc [13].

Among the publications on a political topic, the greatest resonance was caused by reports about the rally in Tomsk "He is not our king", about the behavior of Tomsk residents in the presidential elections, posts criticizing the existing government. The economic theme in the materials is defined rather conditionally, since it is intertwined with other categories, for example, with the economic consequences of the export of timber from the region [14].

One of the problems with traditional social media analysis is that often only the relationships between the participants are considered, not what they actually send each other messages about. This does not take into account the frequency of transmission of messages (for example, several times a day, a week, or another period of time). Often approaches ignore information about the direction of messages, i.e. how many messages were sent by participant A for participant B and how many times, participant B answered A. However, note that all this information is required for different areas of research, for example, highlighting message topics: family, scientific, technical, etc. In addition, message flows can have multiple topics for the same participants. A difficult problem is when, for example, two people are not friends on a social network, but they have mutual friends, so they may recognize each other after some time of communication, or may not know if the list of applicants is large enough [15].

The modern telecommunications environment is built in such a way that it records the actions of its users in detail. This information is spatio-temporal and redundant in its content. The use of this information is becoming more and more active in a variety of applications in tourism, retail, online stores, etc. Therefore, the urgent task is to develop new information and software technologies for efficiently extracting useful information from a significant amount of data arising in the course of finding users in social networks.

This article Supported by RFBR 20-011-31154 opn.

References

1. K. Rost, L. Stahel, B.S. Frey, *Digital Social Norm Enforcement: Online Firestorms in Social Media* (2016) https://doi.org/10.1371/journal.pone.0155923
2. A.R. Rico, The University of Texas at Austin Fans of Columbine shooters Eric Harris and Dylan Klebold DOI: https://doi.org/10.3983/twc.2015.0671
3. E. Ferrara, Zeyao Yang, Measuring Emotional Contagion in Social Media (2015) https://doi.org/10.1371/journal.pone.0142390
4. K. Samson, Trust as a mechanism of system justification (2018) https://doi.org/10.1371/journal.pone.0205566

5. D. Antonakaki, D. Spiliotopoulos, Ch.V. Samaras, P. Pratikakis, S. Ioannidis, P. Fragopoulou, Social media analysis during political turbulence (2017) https://doi.org/10.1371/journal.pone.0186836
6. K. Okamura, S. Yamada, Adaptive trust calibration for human-AI collaboration (2020) https://doi.org/10.1371/journal.pone.0229132
7. P.R. Ward, L. Mamerow, S.B. Meyer, Interpersonal Trust across Six Asia-Pacific Countries: Testing and Extending the 'High Trust Society' and 'Low Trust Society' Theory (2014) https://doi.org/10.1371/journal.pone.0095555
8. T. Haesevoets, Ch.R. Folmer, A. Van Hiel, Is Trust for Sale? The Effectiveness of Financial Compensation for Repairing Competence- versus Integrity-Based Trust Violations (2015) https://doi.org/10.1371/journal.pone.0145952
9. J. Gereke, M. Schaub, D. Baldassarri, Ethnic diversity, poverty and social trust in Germany: Evidence from a behavioral measure of trust (2018) https://doi.org/10.1371/journal.pone.0199834
10. M. van der Linden, M. Hooghe, T. de Vroome, Extending trust to immigrants: Generalized trust, cross-group friendship and anti-immigrant sentiments in 21 European societies (2017) https://doi.org/10.1371/journal.pone.0177369
11. G. Bente, T. Dratsch, K. Kaspar, *Cultures of Trust: Effects of Avatar Faces and Reputation Scores on German and Arab Players in an Online Trust-Game. The PLOS ONE Staff. Correction: Cultures of Trust: Effects of Avatar Faces and Reputation Scores on German and Arab Players in an Online Trust-Game.* PLOS ONE 9(8), e107075 (2014) https://doi.org/10.1371/journal.pone.0107075
12. L.M. PytlikZillig, Ch.D. Kimbrough, E. Shockley, A longitudinal and experimental study of the impact of knowledge on the bases of institutional trust (2017) https://doi.org/10.1371/journal.pone.0175387
13. L. McFarland, D.A. McFarland, K. Lewis, A. Goldberg, *Sociology in the Era of Big Data: The Ascent of Forensic Social Science,* American Sociologist **47**, 1, 12–35 (2015)
14. K. Okamura, S. Yamada, *Adaptive trust calibration for human-AI collaboration*, Journal PLOS ONE **15**, 2, e0229132 (2020) https://doi.org/10.1371/journal.pone.0229132
15. F.S. Sabatini, *Online Social Networks and Trust.* Social Indicators Research **4**, 1–32 (2018)