Modeling of Posting Behavior in Social Media

Kateryna Kononova[1], Anton Dek[2], Maksym Shpakovych[3]

[1]Department of Economic Cybernetics and Applied Economics, V. N. Karazin Kharkiv National University, e-mail: kateryna.kononova@karazin.ua
[2]Department of Economic Cybernetics and Applied Economics, V. N. Karazin Kharkiv National University, e-mail: dektox@gmail.com
[3]Department of Applied Mathematics, Kharkiv National University of Radioelectronics, maksym.shpakovych@gmail.com

Abstract. The aim of the research is a modeling of online social networks (OSNs) users posting behavior. The research is based on the data retrieved from two most popular Russian-speaking OSNs – Vkontakte and Odnoklassniki. The article explores the distribution of the users, their posts, friends, and groups. To check the common hypothesis that content creators are also the main channel of information propagation, we have applied Kohonen maps. Users clustering allowed identifying the types of their posting behavior; in particular, we distinguished "writers" from "propagators" and "readers". The research has shown that the cluster of "writers" was the smallest in both OSNs; however, these users generated the main content. Next in number was a cluster of "propagators", contained users with the largest number of reposts. Among those who are actively interested on a given topic, the most numerous group was "readers", however, the absolute majority of users stayed "indifferent"; they could be considered as a potential audience. Obtained results can be used for Green Technologies promoting in society enforcing their wide implementing.

Keywords: users posting behavior, online social networks, generation and dissemination of information, clustering, neural networks, Kohonen maps.

1 Introduction

Despite high expectations and a huge amount of publications on environmental safety, which is one of the main aims of National Agendas in many countries, it should be admitted that the general attention to green technologies decreases in recent years (Figure 1).

The previous study (Kononova et al., 2016) showed that there was no significant correlation between indices of Green Technologies and Environmental Sustainability at the macro level. Nevertheless, it was observed at the micro level, emphasizing the importance of Green Technologies promotion not only on the state level but also in society. Taking into account the fact that more than a quarter of the world population are users of several online social networks (OSNs), they offer one of the most

Proceedings of the 8th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2017), Chania, Greece, 21-24 September, 2017.

effective channels of opinion forming. However, to find the optimal (shortest and cheapest) way to promote any idea via OSN, its structure and users posting behavior should be studied in details.

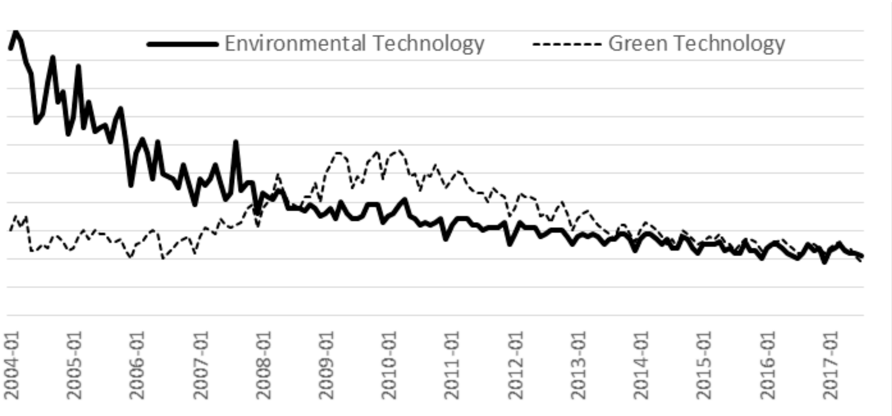

Fig. 1. Google trends on "Environmental Technology" and "Green Technology" search queries, 2004-2017 (Google, 2017)

Nowadays numerous researchers deal with agents behavior, studying its properties, motivation and significant factors. However, despite the growing number of scientific publications in this field, relevant research is fragmented and there is not yet a holistic understanding of agents behavior. At the same time, big social data accumulation in recent years has formed a separate interdisciplinary research direction – Social Media Mining, which studies social networks structures, users' profiles and their behavior. Content, generated by users of social networks, is an important source of information, which could be effectively used to identify implicit patterns of users posting behavior.

The research includes the data about users' profiles, their posts, and comments retrieved from two the most popular Russian-speaking social networks Vkontakte (VK) and Odnoklassniki (OK). The purpose of the study is modeling of users posting behavior with the data collected from the OSNs on a given topic (which is not environmental issues, but more popular one). To achieve it, the following tasks were set:

- to download the users' profiles residing in the target regions;
- to download the posts from users' pages;
- to recognize the posts on a given topic using preliminary developed dictionaries;
- to identify active on a given topic users;
- to identify and describe the types of their posting behavior.

This paper is organized as follows: in the next Section we explain the previous studies on social streams; in Section 3, we present the methodology of data collection and preparation; in Section 4 you can find frequency analysis results; the posting behavior of VK and OK users is described in Sections 5 and 6; and finally, we discuss the results and conclude the paper in Section 7.

2 Related works

Online social networks have attracted considerable scientific attention. Earlier works (Kumar et al., 2003, Gruhl et al., 2004, Adamic et al., 2005) have focused on the information propagation behavior in blog-sphere and studied the information epidemics using classical Diffusion of Innovation model. Kumar studied the "burstiness" of blogs analyzing the evolving link structure. Gruhl focused on the propagation of topics from one blog to the next based on the text of the weblog rather than its hyperlinks. Leskovec (Leskovec et al., 2008) analyzed networks structure of about 45 thousand blogs and 2.2 million postings and offered a model of network evolution.

Guo (Guo et al., 2009) analyzing posting behavior has shown strong daily and weekly patterns; however, for re-posting, the temporal patterns have not been observed. They also distinguished two groups of users: steadily posting in the network, and inactively posting (the rest ones posted occasionally). Bamsuk (Bamsuk, 2012) also investigated temporal characteristics of posting behavior making comparisons blogosphere vs. Twitter, commercial blogs vs. non-commercial blogs. Benevenuto (Benevenuto et al., 2010) used clickstream data from a social network aggregator to compare user behavior across different online social networks.

To understand user behavior, Papagelis, (Papagelis et al., 2011) investigated the causality between individual behavior and social influence by observing the diffusion of innovations among social peers. Liu (Liu et al., 2010) predicted user's interest based on click behavior. Assuming that user behavior is mainly influenced by three factors: breaking news, posts from social friends and user's intrinsic interest, Xu (Xu et al., 2012) proposed a mixture topic model to analyze users posting behavior. Roman (Roman et al., 2012) presented a stochastic model based on decision-making psychology to describe content posting dynamics on OSNs.

However, analyzing posting behavior, most of the studies are based on analytical models focusing on the ways of users connections, the structure of the networks and how it evolved over time. Driving by initial assumptions about users posting behavior they insist on a sensible but doubtable hypothesis that content creators are the same users who play an important role on information propagation. Our study, going from the data, offers a bit different view on this issue separating those who write and those who deliver the information to the majority of readers.

3 Data collection and preparation

The research includes the data retrieved from two most popular Russian OSNs – Vkontakte and Odnoklassniki concerning a specified topic (political issues in cross-border regions). VK and OK are relatively similar social networks, both providing the following data in an accessible form:

- user's ID;
- geolocation;
- the list of the user's friends;

– the list of the user's groups;
– user's posts.

All profiles and posts of those users, who indicated the target geolocation, were downloaded from these two networks. The sample includes 248k Vkontakte profiles and 238k Odnoklassniki profiles. First, both datasets were cleaned from uninformative posts:

– which did not contain textual information (only links, pictures, audio, etc.);
– that language did not match the analyzed one (to identify posts in Russian, the *poliglot* library of *python* was used).

Then, using the *tokenizer* module of the *nltk* library of *python*, we have split the sentences into separate words and removed extraneous characters (punctuation marks, emoticons, etc.). With the *pymorphy2* library, the words were lemmatized; all the letters in the words were converted to a lower case. As a result, we have two datasets (for VK and OK), which meet the following requirements:

– textual format;
– the language of the posts is Russian;
– the words are in a single word form;
– all extraneous characters are deleted;
– all words are written in lower case.
–

4 Frequency characteristics of profiles and posts

After datasets preliminary preparation, we have found and calculated the posts containing the words from a pre-compiled thematic dictionary (Table 1).

Table 1. A fragment of the VK dataset

Id	Num of friends	Num of groups	Num of posts	Num of friends' posts
195269137	49	3	992	1439
187704426	53	10	949	62
17241807	193	1	925	593
35833523	332	1	789	726
906761	300	3	773	291
23419701	273	4	754	55
51258489	536	10	735	754

Analysis of the datasets frequency characteristics has shown that more than 40% of VK users and 25% of OK users did not have friends putting under doubts the relevance of their profiles (Figure 2). However, in general, the distribution density of friends is described by Gaussian low; it indicates the relevance of the sample.

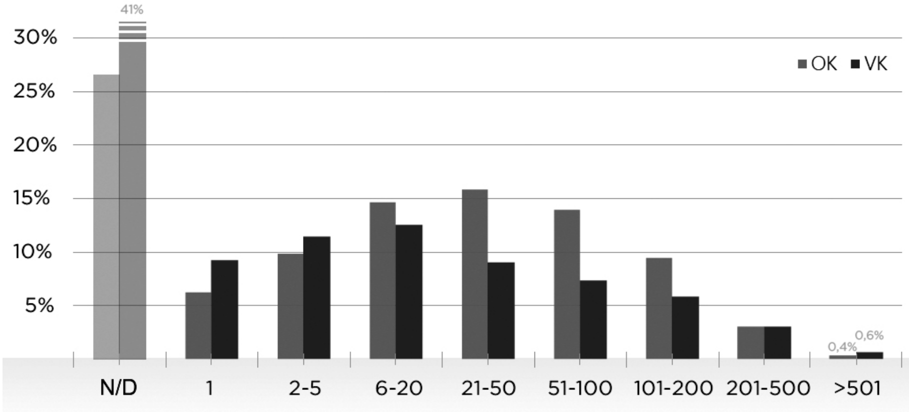

Fig. 2. Distribution of users' friends

Almost 60% of both OSNs users were not subscribed to the groups specializing on a given topic (Figure 3).

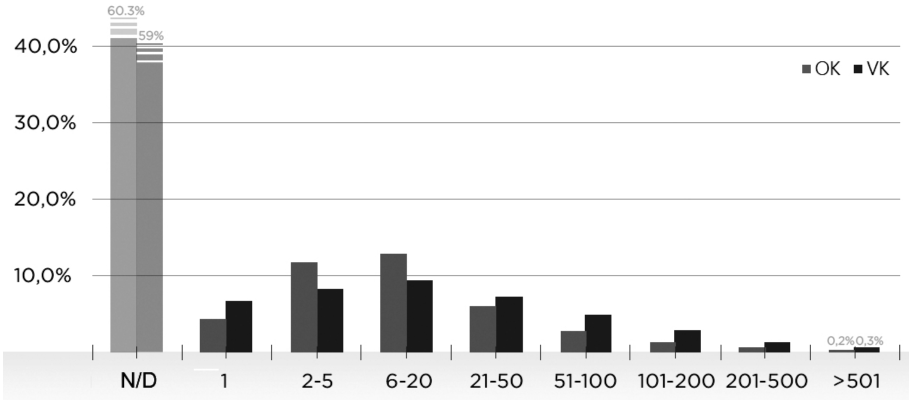

Fig. 3. Distribution of users by thematic groups

At the same time, about a quarter of a percent of users participated in more than 500 thematic groups; it allows making an assumption about the non-random nature of their behavior.

The thickening tail of the posts distribution leads to considerations about the engagement of users who have written more than 500 posts on a given topic (almost 5% of VK users, Figure 4).

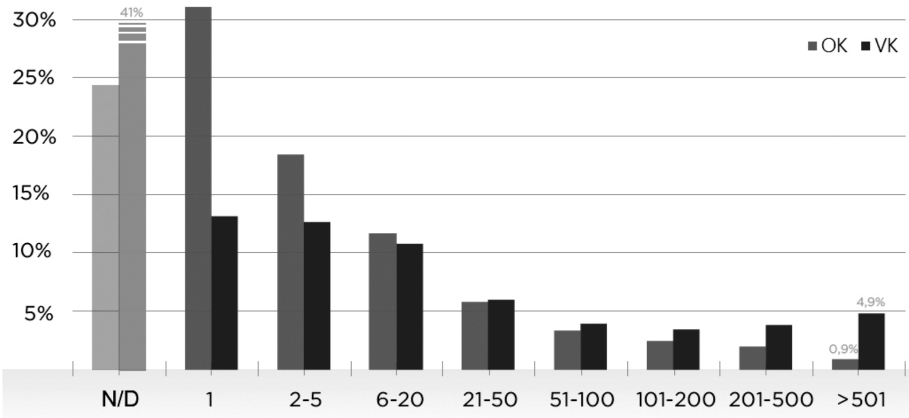

Fig. 4. Distribution of posts on a given topic

In general, a preliminary study has shown that 10% of users generate 70% of the content (Figure 5). Taking into account that some users were posting 5-10 times per day, it could be assumed that posting was kind of work for them.

The preliminary research has allowed concluding that the distributions of posts and thematic groups were non-random, initiating a more detailed study of users posting behavior.

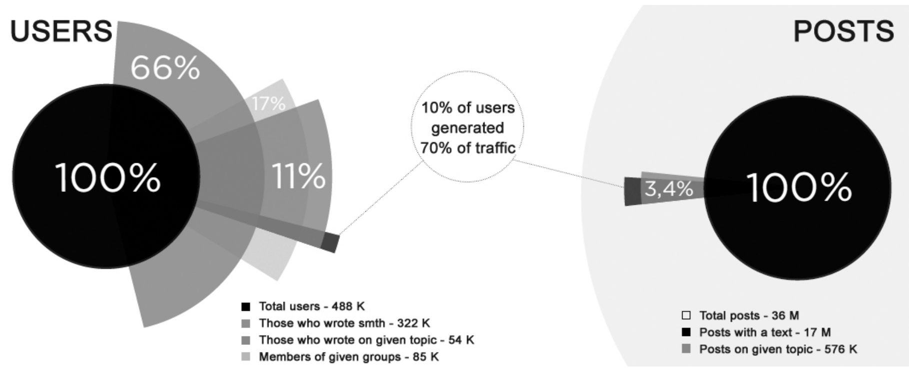

Fig. 5. Summary statistics for VK and OK users

5 Posting behavior of VK users

To identify and describe the types of posting behavior we have decided to use a neural network approach, in particular, Kohonen maps, which allow revealing hidden regularities in the data. Maps are very easy to interpret, because

- each map is a visualization of one of the user's parameters;
- each hexagonal cell contains a certain, in general not the same, number of users;

- the users location is the same at the all maps;
- the color of the cell corresponds to the value of clustering parameter (the scale is indicated at the bottom of each map).

After a series of experiments with VK sub-sample (formed from users who had more than 200 posts), the following set of maps was obtained (Figure 6).

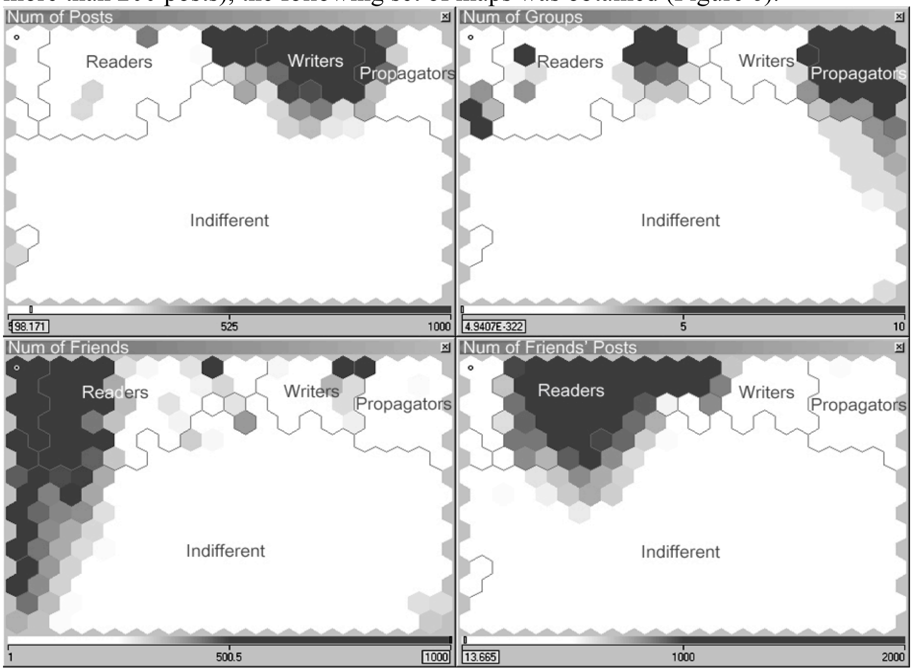

Fig. 6. Set of Kohonen maps, which are describing the behavior of VK users

The analysis of VK clusters cores (Table 2) allowed describing the following types of users posting behavior: “writers”, “propagators”, “readers” and “indifferent”.

Table 2. Characteristics of VK clusters

Clusters	Num of friends	Num of groups	Num of posts	Num of friends' posts
Writers (81 users, 6.9%)	509	10	500	84
Propagators (83 users, 7.1%)	160	70	272	160
Readers (113 users, 9.7%)	436	20	208	1455
Indifferent (890 users, 76.3%)	130	10	107	119

Let us consider the features of the clusters. Seven percent of users got to the cluster of “writers”. Although this cluster is the smallest, its users have generated the main content on the given topic. In addition, they more often then other left comments when reposting. It is interesting that these users have the largest number of friends, which could be explained by opposite reasons – either their writings are supported by the community or they expand the channels of information diffusion themselves.

Next in number is a cluster of “propagators” (also about 7%); these are users with the largest number of reposts from thematic groups. Unlike other users, they have few friends focusing on the concentrated content collection from thematic groups and its further reposting.

Among the people who are actively interested on a given topic the most numerous group is “readers” (about 10%) whose news feed is full of friends’ posts. Unlike “propagators”, they are focused on the consumption of information, rather than on it distribution.

However, the majority of VK users (more than 76%) were indifferent to the subject and did not show any activity at all.

Graphically identified types of behavior are shown in Figure 7.

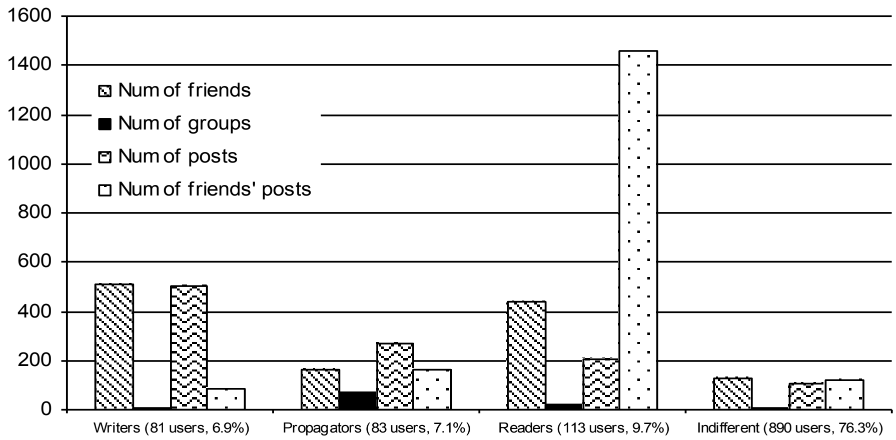

Fig. 7. Average VK clusters characteristics

6 Posting behavior of OK users

The neural network of the same architecture testing on OK dataset has shown comparable results. There are also clusters of “writers”, “propagators”, “readers” and “indifferent” users. This indicates the stability of the identified types of posting behavior. A set of Kohonen maps built on the OK dataset is shown in Figure 8.

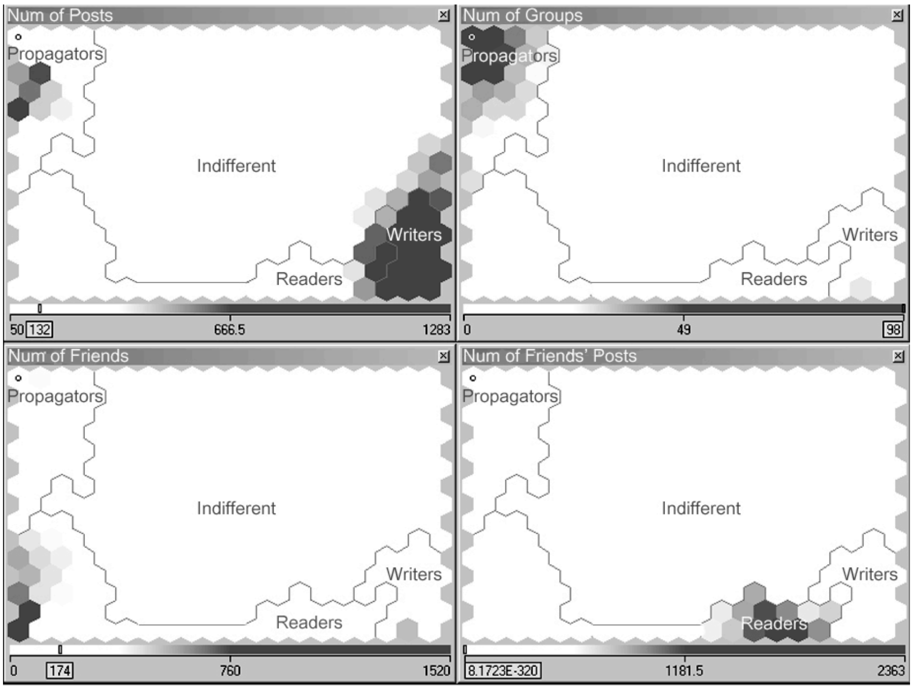

Fig. 8. Set of Kohonen maps, which are describing the behavior of OK users

The analysis of OK cluster cores (Table 3) shows that in general, the characteristics of the detected clusters are similar on both OSNs. The only significant difference was observed in the number of user's friends: in VK the biggest number of friends had "writers" and "readers", then in OK this indicator took the maximum for "propagators".

Table 3. Characteristics of OK clusters

Clusters	Num of friends	Num of groups	Num of posts	Num of friends' posts
Writers (8 users, 2.9%)	121	2	505	31
Propagators (17 users, 6.2%)	455	26	101	60
Readers (28 users, 10.1%)	83	1	225	990
Indifferent (223 users, 80.8%)	73	1	127	20

Graphically identified types of OK users posting behavior are shown in Figure 9.

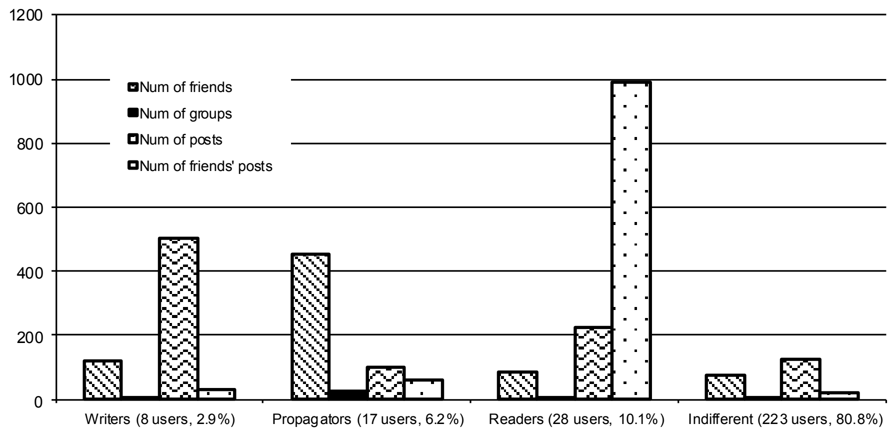

Fig. 9. Average OK clusters characteristics

7 Conclusions

Analysis of frequency characteristics of 248k VK profiles and 238k OK profiles has shown that more than 40% of VK users and 25% of OK users did not have friends; about a quarter of a percent of users took part in more than 500 thematic groups. In addition, the distribution of posts had the fat tail, identifying the users who had written more than 500 posts on a given topic. These facts actualized the need for a more detailed study.

Experimenting with different architectures has allowed creating Kohonen neural networks of the same structure (Figure 10) to identify users posting behavior on both OSNs (VK and OK).

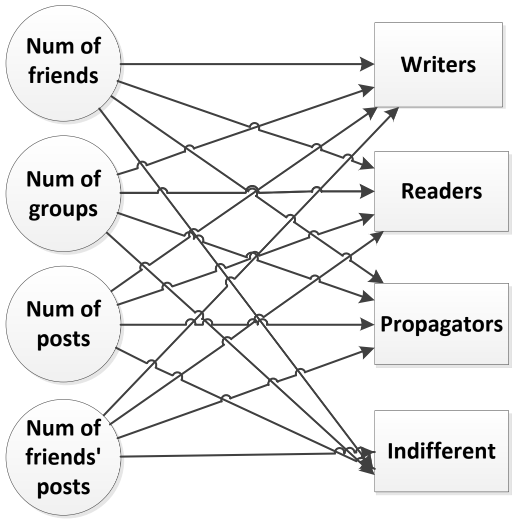

Fig. 10. Kohonen neural network, which is describing users posting behavior

Comparing the clustering results on both social networks, we note that the identified types of users behavior, namely: “writers”, “propagators”, “readers” and “indifferent”, have proved to be stable. In addition, we distinguished “writers” from “propagators” and “readers” not proving the common hypothesis that content creators are also the main channel of information propagation.

It was shown that on both OSNs the cluster of “writers” was the smallest, however, its users had generated the main content on a given topic (the number of their daily posts allows making an assumption about their bias); in addition, they more often than others left comments when reposting. The main identifying criterion of this cluster was the number of posts generated by one user.

Next in number was a cluster of “propagators”, which contained the users with the largest number of reposts from thematic groups. The main clustering criterion here was the number of thematic groups subscribed by the user.

Among those who were actively interested on a given topic, the most numerous group were “readers” whose news feed was full of relevant posts. The main clustering criterion here was the number of user friends’ posts on the given topic.

However, the absolute majority of both OSNs users were indifferent to the topic; they could be considered as a potential audience (Figure 11).

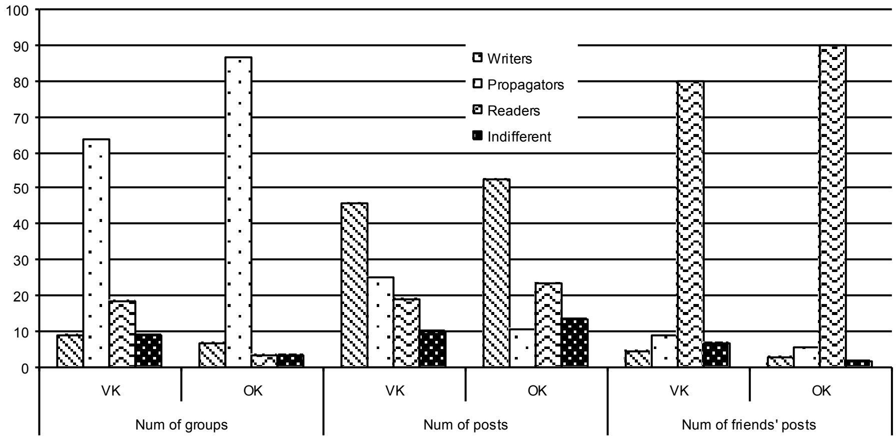

Fig. 11. VK and OK clusters characteristic

Experimental data and obtained conclusions can be used both in the theoretical analysis (substantiation of behavioral axioms and hypotheses) and in research of applied problems related to the posting behavior of social network agents for the development of effective mechanisms for the information flows formation. It could be of help in Green Technologies promoting in society enforcing their wide implementing.

References

1. Adamic L., and Glance N. (2005). *The political blogosphere and the 2004 U.S. election: Divided they blog*. Retrieved from: https://pdfs.semanticscholar.org/1197/ 1e428132ade5439f77eea258140302865ad7.pdf
2. Bumsuk L. (2012). *A temporal analysis of posting behavior in social media streams.* Retrieved from: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/ paper/viewFile/4741/5094
3. Benevenuto F., Magno G., Rodrigues T., and Almeida V. (2010). *Detecting Spammers on Twitter.* Retrieved from: http://www.decom.ufop.br/fabricio/ download/ceas10.pdf
4. Google Trends (2017). Retrieved from: https://trends.google.com/trends/explore? date=all,all&geo=,&q=Environmental%20Technology,Green%20Technology
5. Gruhl D., Guha R., Liben-Nowell D., and Tomkins A. (2004). *Information diffusion through blogspace.* Retrieved from: http://people.csail.mit.edu/dln/papers/ blogs/idib.pdf

6. Guo L., Tan E., Chen S., Zhang X., and Zhao Y. (2009). *Analyzing patterns of user content generation in online social networks.* Retrieved from: https://cs.gmu.edu/~sqchen/publications/kdd09.pdf
7. Kononova K., Kovpak E. (2016). *Green ICTs: impact on the environmental sustainability.* International Journal of Sustainable Agricultural Management and Informatics. Inderscience, Vol. 2, No. 2/3/4, pp. 95-109.
8. Kumar R., Novak J., Raghavan P., and Tomkins A. (2003). *On the bursty evolution of blogspace.* Retrieved from: http://www.disco.ethz.ch/lectures/fs12/ seminar/paper/Barbara/32.pdf
9. Leskovec J., Backstrom L., Kumar R., and Tomkins A. (2008). *Microscopic evolution of social networks.* Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.141.6919&rep=rep1&type=pdf
10. Liu J., Dolan P., and Pedersen E. R. (2010). *Personalized news recommendation based on click behavior.* Retrieved from: http://cs.northwestern.edu/~jli156/IUI224-liu.pdf
11. Papagelis M., Murdock V., and van Zwol R. (2011). *Individual behavior and social influence in online social systems.* Retrieved from: http://www.cs.toronto.edu/ ~papaggel/docs/papers/all/HT11-Individual-Behavior-and-Social-Influence-in-Online-Social-Systems.pdf
12. Roman P. E., Gutierrez M.E., and Rios S.A. (2012). *A model for content generation in On-line social network.* Retrieved from: https://www.researchgate.net/ publication/233897861
13. Xu Z., Zhang Y., Wu Y. and Yang Q. (2012). *Modeling User Posting Behavior on Social Media.* Retrieved from: http://yaowu.co/docs/sigir12.pdf