Application of principal component analysis to identify semantic differences and estimate relative positioning of network communities in the study of social networks content

I A Rytsarev$^{1,2}$, R A Paringer$^{1,2}$, A V Kupriyanov$^{1,2}$ and D V Kirsh$^{1,2}$

$^1$Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
$^2$Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

e-mail: rycarev@gmail.com

Abstract. In the paper, we propose an approach to the analysis of social groups and their relative positioning based on the identification of semantic differences in texts presented in the form of frequency dictionaries. The initial textual data was obtained by collecting records of thematic Internet communities. To collect entries, we implemented a specialized software module for downloading and analyzing posts as well as comments from open communities of interest in the social network VKontakte. The developed algorithm of frequency dictionary compilation evaluates the characteristics of data collected from social networks. For keywords identification, we propose a new approach based on the analysis of word frequency distribution, using methods for dimension reduction of feature spaces. The presented algorithm using the principal component analysis allowed to assess the significance of words by coefficients of the linear transformation. Along with the keywords, we identified semantic differences of social network communities and estimated their relative positioning in the transformed feature space.

1. Introduction

Development of the newest information and industrial technologies, dynamics of its wide implementation influence all the personal and public activities in everyday life. Nowadays we can see rapid transformation of interaction ways between people, social groups and organizations.

Virtual network communities can be described as social groups in terms of social psychology (there are common interests and activities in such groups, awareness of membership, sense of collective “us”, and possibility of direct personal contact between members). At the same time, these communities have specific features caused by its virtual, digital nature. Participation in most of virtual communities is not bounded by age, gender, social status or territory. These communities have free membership (there are no entrance fees, membership cards, formal obligations). Members of the communities are able to enter any of it or leave it at any moment as they wish. It allows people to choose content according to their interests and motives. Besides, in most cases it is not necessary to be a member of virtual community to see its content.
Communication via Internet is so wide-scaled that it requires development of new methods to study and analyze it [1]. The following features of data from social networking services (SNS) create certain methodological difficulties:

- subjective selectivity and social desirability of the content published by users;
- possibility to delete the content previously published by users;
- existence of fake accounts and bots producing unpersonal, unified, automatic content;
- features of SNS itself, its possibilities and interface implementations that bound and induce certain activities of users at the same time, thus creating a discrepancy between users’ behaviour in virtual and real life communities;
- autocompleting of forms and choices “by default”, often made without awareness, probably not consistent with real choice or opinion of users;
- ethical issues concerning the use of personal information and the right of a user to delete it permanently.

Nevertheless, the ability to quantify and classify digital footprints of nearly endless complexity, high speed of information processing (even real-time), its verifiability, transparency, and low financial costs create new opportunities and benefits. However, we have to invent new methods to gather and analyze such data to avoid difficulties mentioned above [2].

The content published by users in SNS, and digital footprints allow revealing and taking into account personal psychological characteristics (human values, emotions, mood, self-regulation strategies and motives), essential to economic, social and political decision making [3]. Moreover, SNS make it possible to monitor these characteristics and their dynamics in real-time. There have been already developed trading algorithms that use as input variables actual social mood calculated upon recent posts in SNS (namely, Twitter, LiveJournal or Facebook) [4, 5], and these algorithms proved to be very effective in a real stock market. Similar psychological characteristics influence not only economic decisions, but political and social decisions and behaviour as well.

However, prediction of psychological characteristics of SNS users is still much less precise than prediction of social or demographic characteristics. For instance, in one of the largest up-to-date studies by Kosinski et al. social and demographic characteristics (such are e.x., sex or sexual orientation) of SNS users were predicted with high accuracy, while error rate for personality traits was relatively high. Nevertheless, even such not so precise results can be useful to optimize search requests and personal ads on webpages. If we want to lower the error rate in diagnostics and prognosis, we have to go deeper and analyze digital footprints not as behaviour markers only. We have to follow new directions and research into meaning and semantics of posts published by users, and invent new mathematical and statistical methods to conduct such analysis.

These studies demonstrate potential benefits of analysis of big data obtained from SNS. However, they are the first steps in a long journey. The models proposed in these and other studies include generalized characteristics of very large samples and communities only. It is necessary to develop more precise models to lower error rates in prognosis and decision making. To achieve this, we have to take into account subjective semantics and psychological features of particular social communities and groups in SNS. One of the tools to achieve this goal is the «Social Sonar» created at Samara University. It is based on new methods of analysis of group dynamics in SNS. In this case, group dynamics refers to all the processes of social groups’ life cycle: its creation, functioning, development, stagnation, involution, and vanishing. The present study proposes a new method of modelling semantic similarities and differences between SNS communities. This method allows raising the efficiency and precision of such analysis.

2. Data collection and preprocessing
We selected the social network VKontakte (VK) for the analysis of virtual communities as it is one of the most popular Russian-speaking social media in the Internet. Key features of VK are its extensive accessibility and the absence of strictly defined subjects. In addition, the social network freely provides an application programming interface for writing external services.
Retrieving data from VK is possible only with standard utilities provided by the developers of the network. Therefore, the first stage of the study was to get access to the social network servers using the provided API. The following steps are necessary to obtain the personal access keys:

1. Visit the developer section and follow instructions of the wizard to create a new application.
2. Get the application ID and the secret key to establish a connection to the network servers.
3. In the app settings, configure permissions to work with communities and records.

The next stage of our work was the development of a software module for data collection. The implementation was carried out using high-level programming language Python and the scripting library for VK. This library is created by third-party developers and has many methods based on the official VK programming interface. After importing the module into the project environment, it creates an authorization object which stores all authorization information (login, password, App ID, authorization key). Henceforth, all interactions with the social network services are performed using this object.

As a matter of understanding, here are some terms used in the upcoming paragraphs. A record is any text message in the community. A post is a record that carries information about one or several events, it often impels you to initiate a discussion on the topic. Comments are entries below the post that reflects the reaction of a specific user to the post or another participant’s comment.

The stage of collecting records from the wall (a screen that display in real time all the information about the topic of your selection) consists of two procedures: a collection of posts and a collection of comments.

The procedure of collecting posts calls the special method of VK API that returns a list of posts for a selected user or community, defined by a unique user ID or community address.

The procedure of collecting post’s comments is carried out in a similar way. The main differences are the method called and the set of initial parameters: to receive comments, you must specify a unique identifier not only for the community, but also for the post.

The use of the two procedures described above allows to collect the information from any open VK community. However, there is a software limitation that restricts the number of requests to the social network servers per day, which significantly complicates the collection of information. In particular, third-party libraries have a daily limit of 2500 requests and can receive no more than 100 records per request.

It should be noted, the record’s properties contain the date of publication. To reduce the number of requests, it makes sense to filter requested records by the predefined time interval. Comparing the time in the record with the selected period, you can get only necessary records corresponding to the interval of interest.

3. Algorithm for frequency dictionary compilation
A record is a special data structure, containing both text fields (the content) and metadata, containing supportive information such as identifiers of the record, its author, time of publication, etc.

Since the basis of the conducted research is the compilation of a frequency dictionary, the developed algorithms analyze primarily the text fields of records. Potentially, the metadata of records can also be taken into account and used for processing, since it stores lots of personalizing information.

After retrieving the content of the record, the textual fragment is split into separate words. In this case, a word is a sequence of alphabetic characters, separated by spaces, numbers or punctuation marks. Since the entries in social networks are full of typos and errors, the spelling check is the crucial step for the correct evaluation of a word’s frequency. To perform the spell-checking procedure we used Yandex “Speller” service.

After performing the proposed steps, the number of occurrences of each unique word $w$ in the whole text data is described by the equation:

$$count(w) = \sum_{s_i \in S} c(w, s_i), \quad c(w, s_i) = \begin{cases} 1 & w = s_i \\ 0 & otherwise \end{cases}$$

where $S = \{s_i\}$ is the text data split into words.
The final number of word’s occurrences is divided by the total number of entries, and thus the evaluated word’s frequency is recorded in the frequency dictionary. The scheme of the developed algorithm is presented in Figure 1.

4. Algorithm for semantic differences detection
We developed a matching algorithm for the identification of textual data semantic differences on the basis of dictionaries frequency analysis. The concept of the algorithm is to apply existing techniques for dimension reduction of the feature space in order to rank words. The principal component analysis technique [17] was used in this article.

The initial data for the algorithm are frequency dictionaries. Each dictionary can be represented by a vector of words. For two dictionaries, it is possible to form a new single feature that will separate these dictionaries in the best way. In the principal component analysis, such feature is formed by multiplying the values of the initial space by the corresponding vector of coefficients. The values of the vector coefficients used to form a new feature can be used as an estimate of words’ contribution to the formed feature. Thus, it is possible to make a list of words that have made the greatest contribution to the feature’s formation. The list of words compiled in this way describes a feature that provides the dictionaries separation and describes their semantic difference.

The first step of the algorithm is data normalization. The values of the word frequencies were normalized in order to obtain dictionary frequencies in the interval [0;1] according to the following formula:

\[ y(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \]

where \( x_{\min} \) – minimum value of the vector, \( x_{\max} \) is the maximum value of the vector.

The next step of the algorithm is the application of the principal component analysis. Principal component analysis is one of the most widely used methods to reduce the dimension of data with the lesser loss of information. The calculation of principal components is usually reduced to the calculation of eigenvectors and eigenvalues of the covariance matrix, describing original data. By definition, the covariance of two features \( X_i \) and \( X_j \) is calculated as follows:

\[ \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_iX_j] - \mu_i\mu_j, \]
where $\mu_i$ – mean value of the $i$-th feature.

The covariance matrix is a symmetrical matrix, where the diagonals represents the feature’s dispersions and outside the diagonals are the covarations of the corresponding pairs of features. The Rayleigh relation [18] implies that the maximum variation of the dataset will be achieved along the eigenvector of this matrix corresponding to the maximum eigenvalue. Therefore, the main components onto which we project the data are simply eigenvectors of the corresponding eigenvalues of the matrix. The values of eigenvector elements are the desired estimated coefficients for the formation of a new feature describing semantic differences in the best way. To obtain a visual interpretation of the relative positions of the dictionaries in the resulting semantic space, we need to multiply the word frequencies by the corresponding eigenvector.

5. Results

To study the performance of the developed algorithms, we selected communities of similar areas of interest. All the selected communities unite the residents of Samara and Samara region. Three of the five groups are open communication platforms for students and teachers of two largest Universities in Samara. Due to the existing restriction on data collection within the framework of this study, we introduced an additional criterion for records filtering based on the time of their publication (from June 1 till September 30, 2018). The list of the selected communities is the following:

I. "Podslushano v Samarskom Universitete".
II. "Podslushano Samarskiï Universitet" (old name: "Podslushano v SamGU").
III. "Podslushano v SamGTU ".
IV. "Usly`shano Samara".
V. "Podslushano Samara".

Table 1. Amount of posts and comments.

| Month          | Type | Community |
|----------------|------|-----------|
|                | I    | II        | III       | IV        | V          |
| June 2018      | Post | 29        | 141       | 304       | 361        | 1296       |
|                | Comment | 147     | 333       | 662       | 3948       | 54364      |
| July 2018      | Post | 23        | 150       | 292       | 359        | 1356       |
|                | Comment | 182     | 430       | 978       | 3860       | 66429      |
| August 2018    | Post | 39        | 268       | 313       | 454        | 1375       |
|                | Comment | 238     | 564       | 991       | 3526       | 89785      |
| September 2018 | Post | 99        | 236       | 481       | 382        | 1228       |
|                | Comment | 415     | 375       | 1043      | 3070       | 49253      |
| Total          | Post | 190       | 795       | 1390      | 1556       | 5255       |
|                | Comment | 982     | 1702      | 3674      | 14404      | 259831     |

Table 2. A fragment of the frequency dictionary.

| Original | Words | Translation | Community |
|----------|-------|-------------|-----------|
|          | I     | II          | III       | IV        | V          |
| мне      | to me | 0.032       | 0.025     | 0.022     | 0.024      | 0.032      |
| меня     | for me | 0.030      | 0.023     | 0.016     | 0.019      | 0.026      |
| только   | only  | 0.024       | 0.031     | 0.036     | 0.033      | 0.026      |
| просто   | simply | 0.016      | 0.020     | 0.023     | 0.022      | 0.019      |
| будет    | will be | 0.015     | 0.008     | 0.006     | 0.009      | 0.012      |
| может    | can  | 0.014       | 0.009     | 0.010     | 0.007      | 0.013      |
| люди     | people | 0.011      | 0.007     | 0.006     | 0.007      | 0.009      |
| автор    | author  | 0.011     | 0.008     | 0.012     | 0.011      | 0.010      |
| такие   | such  | 0.010       | 0.005     | 0.004     | 0.005      | 0.008      |
| время   | time  | 0.010       | 0.007     | 0.007     | 0.007      | 0.008      |
| детей   | children | 0.008    | 0.007     | 0.007     | 0.008      | 0.007      |
| человек | person | 0.007      | 0.012     | 0.006     | 0.010      | 0.007      |
| тогда   | then  | 0.007       | 0.005     | 0.003     | 0.003      | 0.006      |
After collecting records of the selected communities for the predefined period of time, in accordance with the algorithm of frequency dictionary compilation, we selected texts of the numerous records and split them into separate words. Then, the spell-checking was performed and the number of posts and comments for each of the studied communities was calculated (Table 1). The frequency dictionaries obtained as a result of the algorithm are partially presented in Table 2.

Applying the principal component method to monthly frequency dictionaries allows to estimate the relative positioning of network communities in the formed feature space. A graphical representation of the results is shown in Figure 2. The June figures was taken as basis to provide the uniform coordinate system for all points. The color interpretation for communities is the following: I – yellow point, II – red point, III – turquoise point, IV – purple point, V – blue point.

![Graphical representation of results](image)

**Figure 2.** Relative position of communities by months: a) June; b) July; c) August; d) September.

6. **Conclusion**

Based on the analysis of the results obtained, it can be concluded that communities in social networks are not static, but transforming objects, thus they need to be analyzed as dynamical development processes. Particularly, the developed method allows to estimate relative positioning of network communities and trace their changes in time. It should also be noted, the method can show the dynamics of “leakage” users from one community to another depending on the time of year.

The further research should focus on the development of technique for automatic filtering of frequency dictionaries, as well as approaches to the analysis of text messages based on new frequently used words or expressions.

7. **References**

[1] Rytsarev I A, Kozlov D D, Kravtsova N S, Kupriyanov A V, Liseckiy K S, Liseckiy A K, Paringer R A and Samykina N Yu 2018 Application of the principal component analysis to
detect semantic differences during the content analysis of social networks CEUR Workshop Proceedings 2212 262-269

[2] Kosinski M, Matz S C, Gosling S D, Popov V and Stillwell D 2015 Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines American Psychologist 70(6) 543-556 DOI: 10.1037/a0039210

[3] Schwartz H A, Eichstaedt J C and Kern M L 2013 Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach PLoS ONE 8(9) DOI: 10.1371/journal.pone.0073791

[4] Jolliffe I T 2002 Principal Component Analysis (Springer)

[5] Wasserman L 2005 All of Statistics: A Concise Course in Statistical Inference (Springer)

[6] Rytsarev I A, Kirsh D V and Kupriyanov A V 2018 Clustering of media content from social networks using BigData technology Computer Optics 42(5) 921-927 DOI: 10.18287/2412-6179-2018-42-5-921-927

[7] Rytsarev I A, Kupriyanov A V, Kirsh D V and Liseckiy K S 2018 Clustering of social media content with the use of BigData technology Journal of Physics: Conference Series 1096 DOI: 10.1088/1742-6596/1096/1/01208

[8] Kropotov Y A, Proskuryakov A Y and Belov A A 2018 Method for forecasting changes in time series parameters in digital information management systems Computer Optics 42(6) 1093-1100 DOI: 10.18287/2412-6179-2018-42-6-1093-1100

[9] Mikhaylov D V, Kozlov A P and Emelyanov G M 2016 Extraction of knowledge and relevant linguistic means with efficiency estimation for the formation of subject-oriented text Computer Optics 40(4) 572-582 DOI: 10.18287/2412-6179-2016-40-4-572-582

Acknowledgments
This work was financially supported by the Russian Foundation for Basic Research under grant # 19-29-01135, # 18-37-00418, # 17-01-00972 and by the Ministry of Science and Higher Education within the State assignment to the FSRC “Crystallography and Photonics” RAS No. 007-GZ/Ch3363/26 (theoretical results).
This paper contains results of the project Geohub and multimodal multi-structure storages, which is carried out within the framework of the Program “Center of Big Data Storage and Analysis” of the National Technology Initiative Competence Center. It is supported by the Ministry of Science and Higher Education of the Russian Federation according to the agreement between the M.V. Lomonosov Moscow State University and the Foundation of project support of the National Technology Initiative from 11.12.2018, No 13/1251/2018.