Determining the proximity of groups in social networks based on text analysis using big data

A S Mukhin[1], I A Rytsarev[1,2], R A Paringer[1,2], A V Kupriyanov[1,2], 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: andrey63ru@mail.ru

Abstract. The article is devoted to the definition of such groups in social networks. The object of the study was selected data social network Vk. Text data was collected, processed and analyzed. To solve the problem of obtaining the necessary information, research was conducted in the field of optimization of data collection of the social network Vk. A software tool that provides the collection and subsequent processing of the necessary data from the specified resources has been developed. The existing algorithms of text analysis, mainly of large volume, were investigated and applied.

1. Introduction

Currently, social networks are booming: every day their users send billions of messages and leave millions of comments under the relevant posts. The analysis of such content is of great importance for many areas of business. For example, it is impossible to overestimate the impact of Internet marketing on the promotion of goods and services. However, clear understanding of user requests is essential to use these mechanisms effectively. The source of such information can be the materials published by users of social networks, as well as the shares and reposts by users and the entire communities. Thus, the issue of determining the proximity of groups in the social network Vkontakte using the BigData technology, considered in this paper, is certainly a relevant objective and a task of great scientific importance in the field of data analysis.

Data processing from social networks is very popular now. For example, in the article [1] proposes a text normalization with deep convolutional character level embedding (Conv-char-Emb) neural network model for SA of unstructured data. This model can tackle the problems: (1) processing the noisy sentence for sentiment detection (2) handling small memory space in word level embedded learning (3) accurate sentiment analysis of the unstructured data. In the article [2], authors introduce SS3, a novel supervised learning model for text classification that naturally supports these aspects. SS3 was designed to be used as a general framework to deal with ERD problems. In the article [3], authors propose a nonparametric model (NPMM) which exploits auxiliary word embeddings to infer the topic number and employs a “spike and slab” function to alleviate the sparsity problem of topic-word distributions in online short text analyses. NPMM can automatically

decide whether a given document belongs to existing topics, measured by the squared Mahalanobis distance. In the article [4], examine the long-term relationship between signals derived from nine years of unstructured social media microblog text data and financial market developments in five major economic regions. Employing statistical language modeling techniques we construct directional sentiment metrics. In the article [5], the authors propose a background clustering technology for discussion. Compared with the traditional methods, background future clustering keeps the constrains caused by data sparseness and spatio-temporal dependence off, and can be used for unpredictable activities discovery

2. Social network data collection

The social network Vkontakte was selected as a data source for this study [6]. The reasons for this choice are as follows:

- the network provides open access to its data (no restrictions on accessing the server data);
- Vkontakte is the most popular social network in Russia and the fifth most popular social network in the world;
- Vkontakte is a full-fledged social network (unlike Twitter and Instagram, which are microblogs) allowing to create thematic communities, which are particularly interesting for this study.

As part of this study, a Python software package was developed, containing an authorization module, a data collection module, and a filtration module. This software package allows to collect data and filter them to take the relevant information only.

Within this study, the developed software package was used to collect more than 8,000 posts and over 280,000 comments on them from the two most popular communities of the city of Samara ("Podslushano Samara" and "Uslyshano Samara") and from the community of the Samara University students ("Podslushano Samarsky Universitet").

Streaming data obtained from social networks contains a lot of service information. Only the relevant data is important for further analysis; therefore, it is necessary to separate the service information from the relevant data. The software package pre-processing module structures the collected data and filters the relevant and the service fields.

3. Determination of the proximity of groups using BigData technology

To determine the proximity of groups, several metrics for the comparison of word indexes were considered: Euclidean distance, city-block distance and Mahalanobis distance [7, 8]. The Euclidean distance was chosen, since it is most suitable for this experiment according to the following criteria:

1) It is the most widely used and universal metric;
2) The Euclidean distance is calculated based on the original, not the standardized data.

To calculate this metric, attribute vectors were formed between the groups by combining two word indexes or more into a common one [9]. Weight was assigned to each word in the word index, thus each group took the form of a vector of attributes (words) with own weights. In this paper, it was decided to use the word frequency count as the weight [10, 11].

Such an approach for calculating the weights of words in word indexes using traditional methods and technologies requires huge computational resources and takes a long time when the volume and the number of analyzed word indexes increases, so it was decided to use BigData technology and computational clusters for this purpose [12]. At this stage, an algorithm involving MapReduce technology was developed, that rejected non-informative parts of the word index (words consisting of less than three or more than fifteen characters) and also counted the frequency of words in the text. As a result, three word indexes were developed, the elements of which had their own weights, one of them is presented in Fig. 1.

At the next step, it was decided to use two word indexes (of the groups "Podslushano Samara" and "Uslyshano Samara") to get a common word index, and to use the other word index (of the

group “Podslushano Samarsky Universitet”) for test counting. The common word index consisted of overlapping words with the weights recalculated according to formula 2.

```
('что', 1780)
('пожалуйст', 1078)
('анон', 881)
('карт', 625)
('помог', 584)
('ребят', 570)
('сказал', 568)
('можн', 489)
('есл', 401)
('был', 368)
('только', 303)
('андрей', 294)
('скажит', 291)
('город', 234)
('наход', 231)
('самар', 228)
('потер', 217)
('спасиб', 213)
('марат', 207)
('человек', 207)
('подскаж', 199)
('когда', 197)
('сказал', 195)
('говорил', 191)
('теперь', 184)
('очень', 183)
```

Figure 1. Part of the word index developed for the group “Podslushano Samara”.

$$g(g_1, g_2) = \left(\frac{g_1 + g_2}{n}\right) \quad (1)$$

where: $g(g_1, g_2)$ is the weight of the word in the common word index; g_1 is the weight of the word in the first word index; g_2 is the weight of the word in the second word index.

The last step was to calculate first the distances between the resulting word index and the groups it was based on, and then the distance between the resulting word index and the test group. The Euclidean formula of distance between the two groups was used to measure this value.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

The results are provided in the Table 1.

Table 1. The results of calculating the distances between the groups.

Title	Euclidean distance
“Podslushano Samara”	188.32
“Uslyshano Samara”	173.11
“Podslushano Samarsky Universitet”	165.98

Based on the results, we can conclude that the distances between the first two groups are very close. This implies that the common word index is compiled quite accurately and reflects the context of the messages in the said groups well. After analyzing the distance to the test group, one can notice the proximity of all three values, but it is also clear that this distance has doubled as compared to the other two. It can be assumed that this group is slightly different from the other two.

4. Conduct research for five groups

In the next step, it was decided to add two more communities to the already analyzed groups and conduct additional research. The distribution of groups with their number, the number of subscribers, the number of posts and the number of comments below them is presented in Table 2.

Table 2. Analyzed communities and their quantitative indicators.

Group number	№1	№2	№3	№4	№5
Count of subscribers at the time of receiving data	217 679	92 024	150 587	60 225	10 465
Count of posts	32 485	10 957	20 783	8 572	1 034
Count of comments	3 898 212	1 949 106	133 188	171 444	22 748
Count of words	40 167 816	17 541 954	14 650 780	1 371 552	107 650

In order to show the applicability of the proposed method for calculating distances between groups using a common dictionary, we calculated the distances between groups without using a common dictionary and with its use. Table 3 presents the results of calculations without a common dictionary. Table 4 presents the results of calculating the Euclidean distances between all pairs of groups and the templates of common dictionaries built on their basis.

Table 3. Euclidean distance calculation results for all five groups without using a common dictionary.

	№1	№2	№3	№4	№5
№1	-	-	-	-	-
№2	513.363	-	-	-	-
№3	571.324	603.66	-	-	-
№4	644.413	863.041	867.504	-	-
№5	701.423	727.723	689.51	974.583	-

Table 4. The results of the calculation of the Euclidean distances between groups and their common vocabulary.

	№1	№2	№3	№4	№5
№1	-	-	-	-	-
№2	188.32 173.11	-	-	-	-
№3	195.55 365.98	219.66 386.85	-	-	-
№4	296.01 542.91	288.95 589.46	272.03 541.34	-	-
№5	193.5 390.71	201.74 402.43	225.43 453.11	167.21 467.21	-

Comparing the obtained results, we can notice that the distances for calculations using a common dictionary are less than calculations without it. From this we can assume that the use of the method of finding a common dictionary for calculating Euclidean distances is justified and applicable for solving the problem posed.

5. The study of the dependence of the volume of the general dictionary used to calculate the Euclidean distance between groups

We investigate at what volume of a general dictionary the results of determining the degree of similarity of groups among themselves give the most informative readings. To do this, we carry out an experimental calculation of the Euclidean distances for groups numbered 1 and 2 between them

and their common vocabulary by changing the dimensions of the common vocabulary. For the study, we will choose the size of the dictionary equal to the greatest number of unique words for the second group (18.948 words), the small size of the general dictionary (300 words) and several intermediate values. Table 5 shows the results of calculations of this experiment.

After analyzing the results obtained, it can be noted that for the anomalously large and, on the contrary, anomalously small size of the general dictionary, the results turned out to be as non-informative as possible. Most likely this is explained by the fact that with a small dictionary, for the most part, only the most common words that do not carry more information and are approximately equally found in the texts of both groups, for the maximum size of a common dictionary, the situation is fundamentally opposite, tk. Many rare words come into account that are found only in one of the groups, and therefore the results show such an abnormally large scatter. When analyzing the results produced for the intermediate sizes of the general dictionary, it is seen that the values of the distances cease to have strong leaps relative to each other when using a common dictionary of about 3/5 of the amount of unique words for the group with the highest number. Such a result is due to the fact that with such a volume the most non-informative words and words are cut off.

Table 5. The results of the calculation of the Euclidean distances between groups and their common vocabulary.

Dictionary size	Distance to group number 1	Distance to group number 2
18948	349.92	306.92
14000	278.21	249.96
11000	188.32	173.11
9000	195.44	181.06
2000	161.32	155.78
300	60.66	61.05

6. Conclusion

Within the framework of this study, a set of software modules was developed allowing to determine the distance between the communities of the social network Vkontakte. As a result of the work, a common word index was compiled, on the basis of which the degrees of proximity between 3 communities were determined. In the future, the results of the work can be used to develop algorithms for determining the proximity of larger groups and communities using the BigData technology.

7. References

[1] Arora M, Kansal V 2019 Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis *Social Network Analysis and Mining* **9(1)** 12

[2] Burdisso S G, Errecalde M and Montes-y-Gómez M 2019 A text classification framework for simple and effective early depression detection over social media streams *Expert Systems with Applications* **133** 182-197

[3] Chen J, Gong Z and Liu W 2019 A Nonparametric Model for Online Topic Discovery with Word Embeddings *Information Sciences*

[4] Groß-Klußmann A, König S and Ebner M 2019 Buzzwords build Momentum: Global Financial Twitter Sentiment and the Aggregate Stock Market *Expert Systems with Applications*

[5] Zhu C, Du J 2018 Background feature clustering and its application to social text *Information Processing Letters* **136** 44-48

[6] Xu X 2007 Scan: a structural clustering algorithm for networks *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* 824-833

[7] Rytsarev I A, Kozlov D D, Kravtsova N S, Kupriyanov A V, Liseckiy K S, Liseckiy S 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
[8] 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(1)**
[9] 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
[10] Mikhaylov D V, Kozlov A P and Emelyanov G M 2016Extraction of knowledge and relevant linguistic means with efficiency estimation for the formation of subject-oriented text sets *Computer Optics* **40(4)** 572-582 DOI: 10.18287/2412-6179-2016-40-4-572-582
[11] 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(1)** DOI: 10.1088/1742-6596/1096/1/01208
[12] 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

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).