Text data mining using conversation analysis

Igor Rytsarev
Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS;
Samara National Research University
Samara, Russia
rycarev@gmail.com

***Abstract*—This paper suggests an algorithm of text data mining based on conversation analysis. Natural languages are developing dynamically nowadays. New semantic units are constantly being introduced into the spoken language. In these conditions, chains of dependency graphs of semantic units are constantly being rebuilt. This paper proposes a method for identifying synonyms based on conversation analysis. The proposed method has been tested on data collected from social networks.**

Keywords—Social networks, Data Mining, Algoritms

I. INTRODUCTION

The social networks are currently undergoing a turbulent growth: every day, users send billions of messages and submit billions of comments. Their analysis has a great impact on many areas of business. For example, it is impossible to overestimate the influence of internet marketing on the promotion of goods and services. However, in order to use these mechanisms effectively, it is necessary to understand the demands of users. 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 [1-7]. Thus, the issue of determining the closeness of text units 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.

II. DATA COLLECTION FROM SOCIAL NETWORKS

The social network Vkontakte was selected as a data source for this research. 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. relevant information only.

Within this study, the developed software package was used to collect more than 5,000 posts and over 170,000 comments from the two most popular communities of the city of Samara ("Podslushano Samara" and "Uslyshano Samara").

III. DETERMINATION OF THE CLOSENESS OF TEXT UNITS BASED ON CONVERSATION ANALYSIS

Conversation analysis, i.e., the study of structures and formal properties of a language in its social and economic application, is related to all major areas of ethnic and methodological research.

Initially, the conversation analysis was intended for the study of verbal and everyday speech only, and more than that, only conversations between several interlocutors. H. Sacks, the creator of the method, attracted the attention of scientists to the fact that conversations are central for a social world.

A conversation shall necessarily be organized, it implies the existence of an order that does not need to be explained again and again during the exchange of phrases. The order is also needed for the spoken words to be clear to all the conversation participants. The conversation shows the social, interactive competence of people willing to explain their behavior and to interpret the behavior of interlocutors. Inside the local sequences of conversation, and only there, social institutions are finally "spoken into existence". As a result, the smallest and seemingly insignificant details of the conversation actually become a means of actualizing the most important social institutions.

The goal of conversationalists is to describe social practices and expectations that the interlocutors rely on when constructing their own behavior and interpreting the behavior of others.

Conversion analysis focuses on particular cases as opposed to idealization that is inevitably connected with any theoretical generalization, from the point of view of Garfinkel and Sacks. In their opinion, idealization impedes scientific development, since any typology is not much connected with the content of real cases which it is supposed to be based on. Sacks sought to develop a method of analysis that would remain at the level of primary data, raw material, specific, isolated events of human behavior. In contrast to classical sociology, he argued that the details of any spontaneous human interaction are strictly organized – to the extent that provides for their formal description.

On the basis of the above prerequisites, the peculiarities of conversation analysis can be formulated as follows. First, this method follows the data, i.e. the analysis is based on empiricism without using (possibly) predetermined hypotheses. Secondly, the smallest details of the text are considered to be an analytical resource and not an obstacle to be discarded. Third, the authors of the method are convinced that the order in organizing the details of everyday speech exists not only for researchers, but – first and utmost – for the people who construct this speech [8,9].

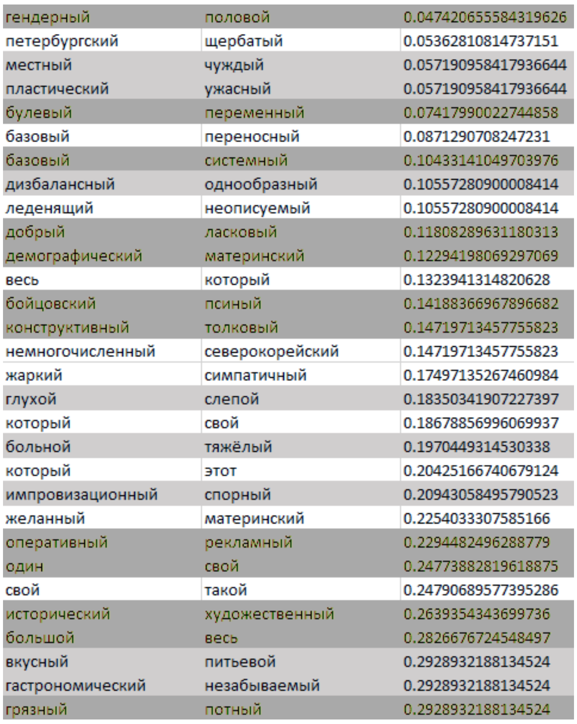
This idea formed the basis for the study. Initially, it was suggested that on a large data set two text units have similar use distance vectors V (vector that shows how the two text units relate to each other within the data, where index i (indicates the distance between units) and Vi (the number of combinations between units, V0 - total number of uses of two text units within one sentence) serve as metrics.

The data collected from the Vkontakte social networks have been pre-processed; each text unit has been brought to its normal form (the pymorphy2 package was used for this). The data was then pre-processed to extract the necessary statistics (WordCount, maximum sentence length). The next step was to create a distance matrix.

A cosine distance was used to calculate distances between two vectors:

$$distance = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} == \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

The results are shown in Figure 1.



гендерный	половой	0.047420655584319626
петербургский	щербатый	0.05362810814737151
местный	чуждый	0.057190958417936644
пластический	ужасный	0.057190958417936644
булевый	переменный	0.07417990022744858
базовый	переносный	0.0871290708247231
базовый	системный	0.10433141049703976
дизбалансный	однообразный	0.10557280900008414
леденящий	неописуемый	0.10557280900008414
добрый	ласковый	0.11808289631180313
демографический	материнский	0.12294198069297069
весь	который	0.1323941314820628
бойцовский	псиный	0.14188366967896682
конструктивный	толковый	0.14719713457755823
немногочисленный	северокорейский	0.14719713457755823
жаркий	симпатичный	0.17497135267460984
глухой	слепой	0.18350341907227397
который	свой	0.18678856996069937
больной	тяжёлый	0.1970449314530338
который	этот	0.20425166740679124
импровизационный	спорный	0.20943058495790523
желанный	материнский	0.2254033307585166
оперативный	рекламный	0.2294482496288779
один	свой	0.24773882819618875
свой	такой	0.24790689577395286
исторический	художественный	0.2639354343699736
большой	весь	0.2826676724548497
вкусный	питьевой	0.2928932188134524
гастрономический	незабываемый	0.2928932188134524
грязный	потный	0.2928932188134524

Fig. 1. The result of distance calculation between distance vectors.

The calculated distances shown in Figure 1 are filtered by the distance value (0-close, 1-far). The proposed pairs of words can be (conditionally) divided into three categories (the proposed interpretation of the results and the division is not accurate, but only the point of view of the author of the article):

- •Dark grey – the most accurate matches (40%);
- •White – the pair of words can be (conditionally) considered synonyms (33%);
- •Grey –antonyms (27%).

The results of the proposed approach suggest that it is easy to construct a graph of interchangeability of words when analyzing text data and use it to extract contextual meaning from the data set.

IV. APPLY CONVERATION ANALYSIS TO THE STATISTICAL DEFINITION OF THE AUTHOR OF A LITERARY TEXT

Conversation analysis showed good results in the problem of determining the closeness of text units and therefore a theory was proposed that it is possible to use this approach for the statistical definition of the author of a literary text.

The main idea of the study is to make multidimensional vectors that store distances between each pair of words in the text. It has been suggested that when comparing distances between pairs of words it is possible to estimate the degree of closeness of text fragments.

This study can be roughly divided into two tasks:

1. Data preparation.
2. Determining the optimal text fragment size for comparison between texts.

To check the first stage of this hypothesis it was suggested to prepare sets of text data by the sliding window method. The sliding window method is an algorithm of transformation, which allows to form a set of data from the source text, which can serve as a set for research.

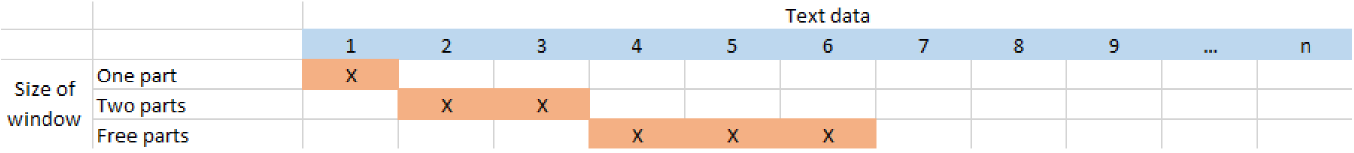
In this case, the window is understood as the size of the window containing the set of texts that are used to conduct the research. During the algorithm operation the window is shifted along subchapters of the text by one measurement unit, and each position of the window forms one text. An example of the method operation is shown in Figure 2.

The next step of the second stage of the study is to compare the data obtained with different window sizes. To do this, we took the windows that include the first element of the data set. These windows have been reduced to the same size (by excluding word sets that were not included in the smaller window). Next, the Pearson correlation coefficient of matrices was calculated. The results of calculating the Pearson correlation coefficient between different sizes of the window are shown in Table 1.

The results of the second stage allow to make an assumption that it is possible not to use the whole text, but only a part of it, when analyzing text data. It can be seen from Table 1 that the maximum correlation growth is achieved by increasing the window size to 5 units and then decreasing.

V. CONCLUSION

This paper has investigated the possibility of applying conversation analysis to social networks' text data analysis and showed that this approach is applicable to the context analysis for establishing logic chains between texts. The main problem is the interpretation of results, since the patterns can be implicit and can vary depending on the context in which text units are used. The application of the conversation analysis to the statistical definition of the author of a literary text have also been studied. This approach has shown its effectiveness. The optimal size of the data set was determined. In the future, the author plans to continue the research in this area and use the approaches based on machine learning and other NLP methods.



		Text data										
		1	2	3	4	5	6	7	8	9	...	n
Size of window	One part	X										
	Two parts		X	X								
	Free parts				X	X	X					

Fig. 2. An example of the sliding window method.

TABLE I. VALUE PEARSON CORRELATION COEFFICIENT BETWEEN DIFFERENT WINDOW SIZES

Window size	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	-	0.26	0.32	0.57	0.78	0.79	0.8	0.81	0.83	0.85
2	-	-	0.34	0.58	0.79	0.8	0.8	0.83	0.84	0.87
3	-	-	-	0.59	0.81	0.8	0.81	0.84	0.84	0.87
4	-	-	-	-	0.81	0.82	0.84	0.85	0.86	0.87
5	-	-	-	-	-	0.83	0.84	0.86	0.88	0.9
6	-	-	-	-	-	-	0.85	0.86	0.9	0.92
7	-	-	-	-	-	-	-	0.89	0.9	0.92
8	-	-	-	-	-	-	-	-	0.9	0.93
9	-	-	-	-	-	-	-	-	-	0.94
10	-	-	-	-	-	-	-	-	-	-

ACKNOWLEDGMENT

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant # 0777-2020-0017) and partially funded by RFBR, project numbers # 19-29-01135, # 19-31-90160.

REFERENCES

[1] I.A. Rytsarev, A.V. Kupriyanov, D.V. Kirsh and R.A. Paringer, “Research and analysis of messages of users of social networks using BigData technology,” CEUR Workshop Proceedings, vol. 2416, pp. 504-509, 2019.

[2] A.F.R. Araújo, V.O. Antonino and K.L. Ponce-Guevara, “Self-organizing subspace clustering for high-dimensional and multi-view data,” Neural Networks, vol. 130, pp. 253-268, 2020.

[3] D.L. Golovashkin and N.L. Kasanskiy, “Solving diffractive optics problem using graphics processing units,” Optical Memory and Neural Networks (Information Optics), vol. 20, no. 2, pp. 85-89, 2011. DOI: 10.3103/S1060992X11020019.

[4] R. Deng, “Research on the Model Construction and Development of Computer Information Acquisition System”, IOP Conference Series: Materials Science and Engineering, vol. 740, no. 1, 012143, 2020. DOI: 10.1088/1757-899X/740/1/012143.

[5] V. Sanz, A. Pousa, M. Naiouf and A. De Giusti, “Efficient Pattern Matching on CPU-GPU Heterogeneous Systems,” Lecture Notes in Computer Science, vol. 11944 LNCS, pp. 391-403, 2020.

[6] A.S. Mukhin, I.A. Rytsarev, R.A. Paringer, A.V. Kupriyanov and D.V. Kirsh, “Determining the proximity of groups in social networks based on text analysis using big data,” CEUR Workshop Proceedings, vol. 2416, pp. 521-526, 2019.

[7] I.A. Rytsarev, D.V. Kirsh and A.V. Kupriyanov, “Clustering of media content from social networks using BigData technology,” Computer Optics, vol. 42, no. 5, pp. 921-927, 2018. DOI: 10.18287/2412-6179-2018-42-5- 921-927.

[8] O.G. Isupova, “Conversion Analysis: Representation of Value,” Sociology: methodology, methods, mathematical modeling (4M), vol. 15, pp. 33-52, 2002.

[9] J. Meredith, “Conversation analysis and online interaction,” Research on Language and Social Interaction, vol. 52, no. 3, pp. 241-256, 2019.