Hidden Communities in the Russian Social Network Corpus: a Comparative Study of Detection Methods

Ivan Mamaev[1[0000-0003-3362-9131]] and Olga Mitrofanova[1[0000-0002-3008-5514]]

[1] Saint Petersburg State University, Universitetskaya emb. 11, Saint Petersburg, Russia
mamaev_96@mail.ru, o.mitrofanova@spbu.ru

Abstract. The paper presents a comparative study of different methods that help to detect hidden communities within social networks. The tested approaches were divided into three main groups: a graph-based method, a clustering method, and a hybrid method. The experiments were conducted on the Russian corpus of posts from VKontakte social network. We discuss advantages and disadvantages of all the methods, and predict the ways of their improving.

Keywords: Social Networks, Corpus Linguistics, Jaccard Index, Cluster analysis, Topic Modeling, Automatic Topic Labeling

1 Introduction

In recent decades, social networks have been actively developing through the perspective of network analysis: sociologists study relationships between users, linguists study texts on the Internet, etc. Despite significant progress in investigating social networks, one of the existing gaps is the detection of hidden communities. A hidden community can be defined as a set of users that have a lot of implicit connections. Unlike "friends" on social networks or members of a real community, users of hidden communities may not know each other, they may live in different regions of the same country, but their interests coincide. These interests may be reflected in users' posts. Using various similarity metrics, one can unite posts in a common semantic group, they forming a semantic network, which conventionally looks as follows (Figure 1).

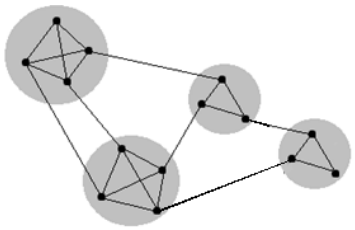

Fig. 1. An example of a semantic network

There are a great number of algorithms for analyzing the structure of social network communities. However, one of the problems is the impossibility of determining the

effectiveness, reproducibility, flexibility, and universality of a method because scholars often carry out experiments involving a single dataset and a single method. In our opinion, the most productive way is to compare several methods using one dataset. A qualitative and quantitative comparison of methods and their results can reveal the best approach that is applicable to the posts of social networks. In our study, we will analyze several algorithms, identify the structural properties of the obtained communities, and also focus on their advantages and disadvantages.

2 Related works

Contemporary algorithms for detecting hidden communities in social networks can be divided into three main groups:

- methods based on graphs;
- methods based on clustering;
- hybrid methods.

The first papers dedicated to detecting hidden communities on the Internet appeared in the early 2000s. The papers [2, 8] describe the use of the hidden Markov model and a random graph to detect communities, the accuracy of the results reached 90%. It should be noted that the approach did not include semantic information about target users which could have improved the accuracy of the method.

In [6], a two-stage algorithm HICODE (Hidden Community Detection) was developed. The first stage consists in finding the number of community layers. The first layer is a group with the strongest links, each subsequent layer has a lower degree of connections. The second stage is called the refinement stage. Its idea is in improving the quality of the layers. In this case one can obtain complete data because a stronger community structure can distort information about the structure of weaker ones. The HICODE algorithm has also been applied to hidden communities on Reddit [13]. The authors of the paper concluded that the boundaries of communities are implicit for several reasons: users of forums communicate a lot with each other, the structure of the site itself is not regulated, the constant expansion of topics on the forums leads to adjustments in the results.

With the growth of machine learning algorithms, social media researchers have begun to use cluster analysis more often. In [3, 7, 10, 12], clustering results are discussed as regards both English and Russian data. In particular, [7] provides an analysis of the relations of Kinopoisk users with topic communities in VKontakte social network. The authors used the silhouette metric to evaluate clustering performance, results showed that k-means was the most efficient algorithm.

The next step in the development of methods for detecting hidden communities was a combination of various approaches. [1] describes a hybrid method: the edges of the graph are compared based on the Jaccard index; they are united into hidden community clusters. In [14], a clustering experiment based on the semantic similarity of English Twitter posts was conducted. To identify the degree of text similarity, WordNet was used. Computational linguistics methods were also applied in [9]: researchers

described the process of detecting communities in experiments with topic modeling and automatic topic labeling.

3 Experiments

3.1 Corpus collection

All of the above methods have been tested on different datasets. In this study, to compare existing algorithms for detecting hidden communities in social networks, we used the corpus of Russian posts (8 679 402 tokens) on the VKontakte social network proposed in [9]. The choice of this corpus is connected with the fact that algorithms for detecting hidden communities based on Russian data are not fully described.

The following algorithms were selected for comparison: The Jaccard index (graph method), clustering using doc2vec (clustering method), topic modeling and automatic topic labeling (hybrid method). There is a brief description of procedures for creating models of hidden communities.

3.2 Graph-based method

The corpus of posts consists of more than 25 000 users' texts. For the convenience of analysis, we presented them in the form VK_USER_X_POST_Y, where X is a user id, Y is an ordinal number of the post. Using Python 3.7, a script was created to calculate the values of the Jaccard index. We considered it necessary to set the minimum similarity threshold to 0.75, i.e. ¾ of the content of the posts should match. Pairs of users that met these requirements were written to a .csv file. Using the Gephi application, we built a resultant graph that shows the state of hidden communities.

3.3 Clustering analysis

There are a lot of cluster analysis methods: k-means, DBSCAN, hierarchical method, etc. In our research, to deal with a large amount of data, we will take k-means from the scikit-learn library[1] due to a rather high degree of work and visualization of the implementation. To improve the quality of the obtained models, we will use a pre-trained doc2vec model of the corpus with the following parameters: the size of the context window – 5, the dimension of the vector – 100.

3.4 Hybrid method

We took the method described in [9]. It is necessary to search for the optimal number of user topics, build topic models, and generate topic labels. In contrast to the previous algorithms, in topic modeling it is necessary to implement the basic NLP procedures: tokenization, lemmatization, processing with the help of a stop-list, adding

[1] https://scikit-learn.org/stable/

bigrams and trigrams. Topic labeling is used to improve the interpretability of topics, the procedures are carried with the help of word2vec, RuWordNet[2], the Russian National Corpus[3] and a Frequency Dictionary of Contemporary Russian[4] by O.N. Lyashevskaya and S.A. Sharov. The Gephi application was also used to visualize the graph.

4 Results and Evaluation

4.1 Models of hidden communities

There are three models of hidden communities created with the help of different algorithms (Figures 2–4).

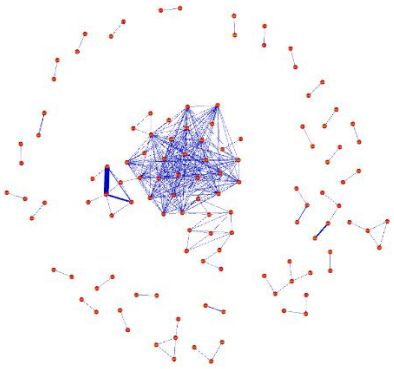

Fig. 2. Hidden communities created with the help of the graph method

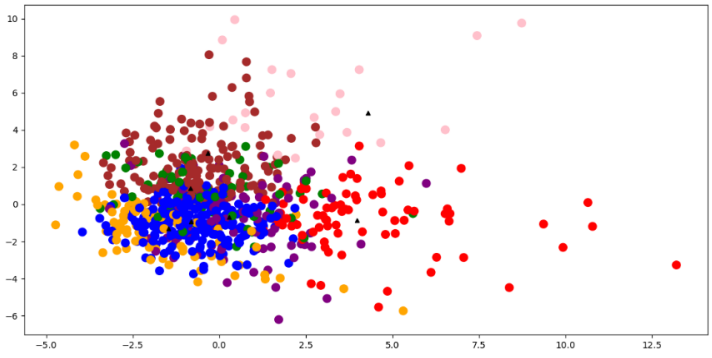

Fig. 3. Hidden communities created with the help of the cluster analysis

[2] https://ruwordnet.ru/ru
[3] https://ruscorpora.ru/new/index.html
[4] http://dict.ruslang.ru/freq.php

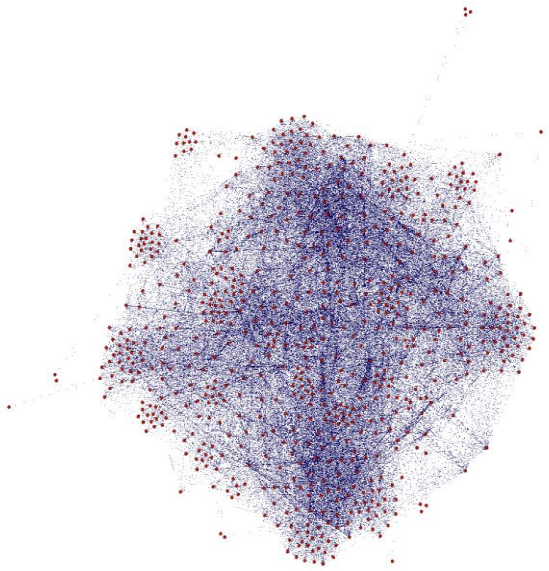

Fig. 4. Hidden communities created with the help of the hybrid method

4.2 Graph method

Using the Jaccard index, 344 pairs of semantically close posts were detected, they being grouped into 34 communities. Figure 2 shows a fragmentary structure of the relationships that exist in the corpus. The first obvious result of graphical analysis is the ability to select several subsets: a giant component (a great number of central nodes of the graph that are interconnected) and stand-alone nodes. The presence of a large number of intersections indicates close relationships within a given semantic field.

It should be noted that the names of topical hidden communities are not displayed when they are formed, as a result, further analysis is carried out manually. For instance, users 696481, 39371, and 2500528 are members of the central hidden community. After analyzing the posts, it was found that, despite the difference in topics, they all have one common topic – online conferences. When carrying out further analysis, we also identified that other users of this community also had information about online meetings.

As for fragmented communities, they have more stable social topics. Users 981916, 118634, 72972 and 1417120 are united by a concert topic.

The situation is similar for users 62729 and 4583: they are united by the topic of health. At the same time, users 1361373 and 892245 are also interested in healthcare, but they were not in the health community.

The above factors may indicate the following peculiarities when working with graph methods.

1. Graph methods partially allow you to track trends in society. In particular, it applies to online meetings caused by the coronavirus pandemic: user 696481 held remote meetings dedicated to religious topics, and user 2500528 commented on distance education.
2. If the life in the society had been calmer, the final representation could have been more fragmented, users of the big community having no connections. This indicates the impossibility of taking into account semantic variations of posts.

4.3 Cluster analysis

The k-means method analyzes the sets of given objects and creates k optimal groups, but the nature of these groups is not known in advance. As a result, the clustering process must be repeated a number of times with different parameters in order to find the most stable variant of hidden communities. 7 groups of hidden communities were detected. For a more detailed analysis, the graphic data were labeled corresponding to user ids (Figure 5).

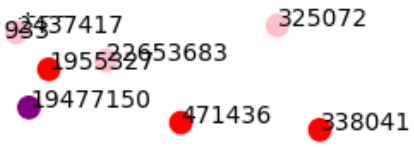

Fig. 5. Example of data labels

As we have already noted, the k-means method does not require much time; it is suitable for preliminary procedures, after which more powerful algorithms are required. At the same time, we can note some shortcomings of this algorithm when analyzing hidden communities. First, the algorithm is outlier sensitive: i.e. theoretically, those users whose topics do not coincide with others, will be assigned to any class. User 1415502 is keen on subcultures, but other users from the corpus do not write posts on the same topic. At the same time, using the k-means method, it was observed that this user and user 167175 are in the topical community dedicated to games.

It should also be noted that the algorithm cannot cope with the task in which objects can belong to different topical groups. Unlike the graph method, all users can belong to the only hidden community. So, for instance, the posts of user 1955327 are dedicated to books and health, although in Figure 5 the text of the user were assigned to the only topical community. A possible solution to the problem is to combine different clustering methods. In particular, the DBSCAN algorithm will take into account the “noise” when constructing models, while the c-means algorithm – a fuzzy clustering method, which is the improvement of the k-means method – assigns texts to different clusters with a certain probability. It is an equivalent to some elements of the model based on graphs when one node can be included in several communities.

4.4 Hybrid method

When working with the hybrid approach, there are also advantages and disadvantages. A lot of algorithms for topic modeling and their variations have already been successfully adapted to Russian corpora of social networks [4, 5, 11].
It should also be noted that, while improving the interpretation of models, we used a double ranking algorithm (Google PageRank and ipm in a Frequency Dictionary of Contemporary Russian by O.N. Lyashevskaya and S.A. Sharov) to avoid low-frequency topic labels. In the final model of communities, users can belong to different topic communities, or have unique interests that do not intersect with the interests of other people.

The approach allows taking into account the various interests of a user, so some people are members of several hidden communities. For instance, 820468, 14846, and 243903 are interested in linguistics, while 62729, 4583, 1361373, and 892245 pay attention to health. Mind that the graph-based method also united 62729 and 4583 in the same community, but 1361373 and 892245 were not in it.

At the same time, we faced certain difficulties. For instance, texts on social networks have some peculiarities of spelling of words or their graphical representation, they making us improve algorithms of tokenization and normalization or edit texts manually (remove diacritical symbols, correct misspellings, etc.).

Moreover, approaches for automatic topic labeling in Russian and English do not have a gold standard that allows assessing the quality of the developed methods. We used Google Forms and 10 independent experts to assess the candidates: 0 is for irrelevant labels, and 1 is for relevant ones (Figure 6, Table 1).

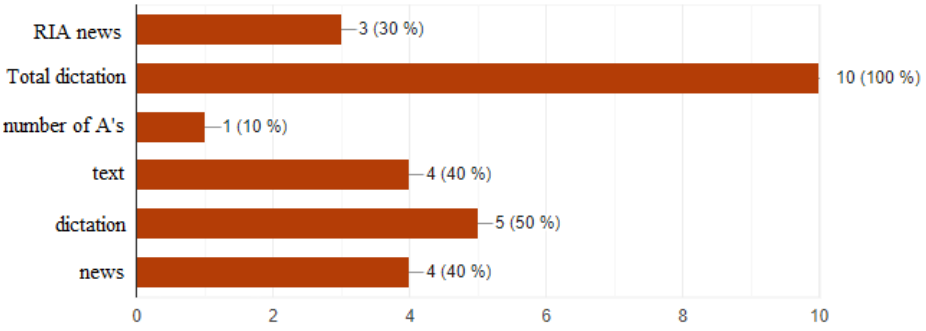

Fig. 6. Comparative diagram of candidates for user 3854113

Table 1. Results of an expert assessment of candidates for topic labels of user 3854113

Topic of user 3854113: *год, диктант, тотальный, язык, акция, два, вид, новость, человек, жизнь (year, dictation, total, language, action, two, kind, news, person, life).*			
Bi- and trigrams	*РИА новости (RIA news)*	*тотальный диктант (total dictation)*	*число пятёрок (number of A's)*
Assessing results	3	10	1
Unigrams	*текст (text)*	*диктант (dictation)*	*новости (news)*
Assessing results	4	5	4

In the example above, assessors are sure that bigram total dictation is the best one, while number of A's is the worst one.

There was also a need to unify different labels that are in the same semantic field: for instance, when we found adjective and noun modifiers in collocations, we chose the first one as a dominant collocation: student organizations, organizations of students – student organizations. It is also a time-consuming process.

Further studies can involve the development of the gold standard, which will significantly save time for building models of hidden communities, and a semantic analyzer that will unify different lexical-semantic variants of labels.

Below we present a summary of the methods based on some parameters (Table 2).

Table 2. Comparison of methods

Parameters	**Graphs**	**Clustering**	**Hybrid methods**
Is the method time-saving?	Yes	Yes	No
Can users belong to several communities?	Yes	Depending on the algorithm	Yes
Are lexical and semantic features taken in account?	No	No	Yes
Is it possible to find out the name of hidden communities?	No	No	Yes

5 Summary

With the development of social networks, the focus and scope of users' interaction are getting expanded. The analysis of the corpus of posts provides an insight into the structure of hidden communities existing in any online platform. The discovery of hidden communities is used in biology, sociology, as well as computational linguistics.

In this paper, we have used a corpus of VKontakte posts and made a detailed review of some approaches to community detection in social networks. The obtained results of comparison show that, unfortunately, the state-of-the-art approaches are far from perfect: some have difficulties with analyzing linguistic properties of texts, and

others are rather time-consuming. We may expect that the manual creation of hidden community models of the corpus will have certain differences compared to our results. Nonetheless, the combination of the methods, described in previous sections, can still be used for preliminary experiments.

We hope to develop the research in the following ways:

- improving the doc2vec model of the corpus for obtaining more precise results on clustering;
- combining various clustering methods for choosing the most optimal one;
- expanding the size of the already existed corpus including posts of other Russian social networks in order to detect latent links between a great number of users;
- improving basic NLP procedures to simplify training topic models;
- developing and assessing a gold standard for topic labeling procedures in the Russian segment of topic modeling.

References

1. Ahn, Y., Bagrow, J., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature, 466, 761–764 (2010).
2. Baumes, J., Goldberg, M., Magdon-Ismail, M., Wallace, W.: Discovering Hidden Groups in Communication Networks. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.). Intelligence and Security Informatics. ISI 2004, Lecture Notes in Computer Science, 3073, 378–389. Springer, Berlin, Heidelberg (2004).
3. Bedi, P., Sharma, C.: Community detection in social networks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(3), 1–22 (2016).
4. Bodrunova, S., Blekanov, I., Kukarkin, M.: Topic modeling for Twitter discussions: Model selection and quality assessment. In: Proceedings of the 6th SGEM International Multidisciplinary Scientific Conferences on SOCIAL SCIENCES and ARTS SGEM2018, Science and Humanities, 207–214. STEF92 Technology Ltd., Sofia, Bulgaria (2019).
5. Bodrunova, S., Blekanov, I., Kukarkin, M.: Topics in the Russian Twitter and relations between their interpretability and sentiment. In: Sixth International Conference on Social Networks Analysis, Management and Security, 549–554 (2019).
6. He, K., Li, Y., Soundarajan, S., Hopcroft, J.: Hidden community detection in social networks. Inf. Sci., 425, 92–106 (2018).
7. Khlopotov, M., Startseva, N., Makarenko, A.: Analysis of movie lovers' preferences and their thematic communities in social networks. The Eurasian Scientific Journal, 11(2), 1–11 (2019).
8. Magdon-Ismail, M., Goldberg, M., Wallace, W., Siebecker D.: Locating Hidden Groups in Communication Networks Using Hidden Markov Models. In: Chen, H., Miranda, R., Zeng, D.D., Demchak, C., Schroeder, J., Madhusudan, T. (eds.) Intelligence and Security Informatics. ISI 2003, Lecture Notes in Computer Science, 2665, 126–137. Springer, Berlin, Heidelberg (2003).
9. Mamaev, I., Mitrofanova, O.: Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus. In: Filchenkov, A., Kauttonen, J., Pivovarova, L. (eds.) Artificial Intelligence and Natural Language. AINL 2020, Communications in Computer and Information Science, 1292, 17–33. Springer, Cham (2020).

10. Mishra, N., Schreiber, R., Stanton, I., Tarjan, R.: Clustering Social Networks. In: Bonato, A., Chung, F.R.K. (eds.) Algorithms and Models for the Web-Graph. WAW 2007, Lecture Notes in Computer Science, 4863, 56–67. Springer, Berlin, Heidelberg (2007).
11. Nagornyy, O., Koltsova, O.: Redefining media agendas: topic problematization in online reader comments. Media and Communication, 7(3), 145–156 (2019).
12. Rytsarev, I., Kupriyanov, A., Kirsh, D., Liseckiy, K.: Clustering of social media content with the use of BigData technology. Journal of Physics Conference Series, 921–927 (2018).
13. Salz, D., Benavides, N., Li, J.: Hidden Community Detection in Online Forums. CS224W: Machine Learning with Graphs, 1–10 (2019).
14. Singh, K., Shakya, H., Biswas, B.: Clustering of People in Social Networks based on Textual Similarity. In: Perspectives in Science, 570–573 (2016).