How Many Clusters? An Entropic Approach to Hierarchical Cluster Analysis

Sergei Koltcov, Vera Ignatenko(✉), and Sergei Pashakhin

National Research University Higher School of Economics, 55/2 Sedova Street, St. Petersburg 192148, Russia
{skoltsov,vignatenko,spashahin}@hse.ru

Abstract. Clustering large and heterogeneous data of user-profiles from social media is problematic as the problem of finding the optimal number of clusters becomes more critical than for clustering smaller and homogeneous data. We propose a new approach based on the deformed Rényi entropy for determining the optimal number of clusters in hierarchical clustering of user-profile data. Our results show that this approach allows us to estimate Rényi entropy for each level of a hierarchical model and find the entropy minimum (information maximum). Our approach also shows that solutions with the lowest and the highest number of clusters correspond to the entropy maxima (minima of information).

Keywords: Hierarchical clustering · Rényi entropy · Number of clusters · User profiles · Online social networks

1 Introduction

The importance of information as a resource in modern society is growing significantly due to the high speed of dissemination and importance for decision-making. At the same time, online social networks (OSN) increasingly become more critical infrastructure in the process of disseminating information. On the one hand, networks represent the environment for the distribution of information; on the other hand, networks themselves generate information capable of affecting significantly economic and political preferences of people. The political turmoils of recent years in various countries (the Arab Spring, the Occupy Wall Street movement, the Ukrainian crisis), the apparent imbalance in news coverage on various online platforms (i.e. the US presidential elections), generation of numerous fake informational events and their explosive distribution through social networks demonstrate the need for a clear understanding of the information transmission and transformation processes.

In the study of news dissemination through OSN, networks should be considered as complex social systems (complex systems), requiring the use of various methodologies. There are many models of news spread which account for network topology [7], the role of 'influential users' [19] and the topical component of

K. Arai et al. (Eds.): SAI 2020, AISC 1230, pp. 560–569, 2020.
https://doi.org/10.1007/978-3-030-52243-8_40

the distributed messages [2]. However, one of the critical factors in news spread through OSN is a set of social attributes of users, such as gender, age, political preferences or religious affiliation [14]. Thus, when analyzing the distribution of information through OSN, it is necessary to solve the problem of estimating the influence of users' social attributes on the depth and speed of dissemination. This problem can be solved either by constructing regression models [14,31], or by including user features in a unified probabilistic framework [4]. However, despite the importance of adding user attributes to a model for transmitting information over OSN [12], the inclusion of a large set of features in probabilistic models is a big problem due to their extreme heterogeneity.

Another solution is to cluster users on their features and reduce them to one variable of 'user similarity'. Accordingly, 'user similarity' can replace the many user features in probabilistic models of information dissemination. However, clustering of OSN users by their socio-demographic characteristics with classic models such as K-means, C-means or hierarchical model, despite the developed techniques [23,26], causes problems, as it is necessary to determine the right number of clusters. Moreover, our experience shows that such techniques as the gap statistic [29], the jump method [27] or the elbow method [18] are unable to find the optimal number of clusters on large user data from OSN. These methods are developed on relatively small datasets and involve expensive in terms of time and memory computations, which is a critical issue with large data. Moreover, these approaches still require human judgment as to where is the optimal number given a set of measures. Therefore, it is necessary to develop other techniques for determining the optimal number of clusters.

In the framework of this work in progress, we consider the direction of 'network thermodynamics' [8], which allows one to organize data clustering, or rather, determine the number of clusters, based on the thermodynamic formalism [25,32]. In this paper, an entropy approach is proposed for determining the optimal number of clusters for profile data of OSN users with the classical hierarchical clustering method. In other words, rather than developing a new algorithm of hierarchical clustering, we use classic algorithms and aim at determining the optimal number of clusters (i.e., the optimal cut off) of a hierarchical solution.

The distinctiveness of the hierarchical method is in the construction of a hierarchical structure (dendrogram) of folded clusters. Here, at the highest level of a hierarchy, all nodes are assigned to one cluster, and at the lowest level, each element is a separate cluster. Hence, one can determine the entropy of two borderline situations and organize a search for the number of clusters inside these boundaries.

2 Background

The study of complex systems with methods of statistical physics is a leading stream in the network analysis research. Here one can distinguish several areas, each with specific goals and tasks. One is the area of network modeling such as Erdös-Rényi, Bollobás-Riordan, Watts-Strogatz models and other [6,11,22]. However, the other two areas are more relevant to our problem.

The second area studies clustering models of network structures, where researches develop metrics for graph partitioning [13]. For instance, when dealing with large in terms of nodes and edges networks, researchers describe a network with methods of statistical physics such as annealing models for modularity optimization [8,15] or with thermodynamic formalism [9,32]. Additionally, the concept of entropy, as in classic Gibbs-Shannon or Rényi-Tsallis definition based on deformed logarithm, could be found in the literature of network analysis [24,28]. This area could be referenced to as the 'network thermodynamics' [8].

Another area is closely related to the two already mentioned and involves models of hierarchical cluster analysis. Such clustering procedures attempt to restore the structure as a dendrogram, or one may say that such procedure is the sequential merging of smaller clusters into increasingly larger. One feature of the hierarchical approach to data clustering is the formation of parent-child relations, where parents are merged child clusters. In such a structure, the top level has all nodes in one cluster, and the bottom level has each node in a separate cluster.

When hierarchical clustering is applied to small data, where dendrogram is no larger than ten levels, the analysis is not so problematic. However, when data consists of several thousand or more units, the problem of choosing a dendrogram cut (the number of clusters) becomes complicated. For hierarchical clustering, the standard approach is to manually examine the dendrogram and try to put a cut-off line so that the distance distributions below the line are more heterogeneous than the distributions above the line. However, this approach is often ambiguous as it relies on human judgment. A solution to this problem could be found with the thermodynamic formalism from non-extensive statistical physics.

We ground our approach in the following works. The first [25] is proposing to search for the free energy minimum in data clustering. However, this criterion is developed only for the K-means type of algorithms. The second work [28] shows that the Tsallis entropy obtained with q-deformed Stirling formula may be used to describe hierarchical statistical systems where each level has its value of the Tsallis entropy. Such a description allows for exploring the hierarchical structure using the Tsallis entropy. As for hierarchical cluster analysis, Gibbs-Shannon entropy was used for evaluating solutions in [1,10].

Thirdly, we build on the work of Olemskoi, who proposed using the concept of internal energy to describe a hierarchical system, which allows us to determine the free energy of the entire hierarchical system, as well as at each level [24]. However, unlike Olemskoi, who considers the transition from level to level in a hierarchical tree in terms of a diffusion process on branching trees, we propose to view the transition process as a process of hierarchical clustering which is characterized by the measure of the Rényi entropy. The Tsallis entropy could be obtained from the Rényi entropy with simple transformations [3].

A similar use of the deformed Rényi entropy is considered in [20,21] for clustering of large document collections. The tests showed that the minimum of the Rényi entropy corresponds to the human choice of the number of clusters. At the same time, the maximum of entropy corresponds the lowest and the largest

numbers of clusters (from one-two to hundreds and more). In such cases, the Rényi entropy becomes larger as the distribution of features becomes uniform. However, these approaches have not been adapted for hierarchical models.

3 An Entropic Approach

Based on the discussed works, we formulate an entropic approach for determining the optimal level (the number of clusters) in hierarchical clustering.

We start from the proposition by Beck that information is related to entropy in the following way: $S = -I$ [5]. Thus, information maximum corresponds to entropy minimum. Next, we consider a set of objects (nodes) as a statistical system. At the starting point, such a system is characterized by entropy maximum (information minimum) because at the initial state each object belongs to a separate cluster. Next, we consider a number of clusters as a temperature of such system which is a function of a level in hierarchical clustering. Given that, the hierarchical clustering procedure transforms a system from the state of maximum entropy to the state of the entropy minimum by changing the number of clusters (temperature). Therefore, the optimal clustering for large and heterogeneous data would be at the state of the entropy minimum for a system.

In the framework of hierarchical clustering, one can find two borderline situations: (1) All objects belong to one cluster. Such clustering has minimal information value, and, correspondingly, such solution has large entropy. (2) Each object is a unique cluster where the probability that a particular object belongs to a cluster is constant. In this case, as it is a uniform distribution, entropy is also large.

A hierarchical clustering procedure constructs a hierarchical tree, where each level has a certain number of clusters. Each cluster may contain a different number of objects N_{ik}, where k is a cluster on level i. However, the total number of elements on each level always equals the total number of system elements N. We define the probability of elements in cluster k on level i as follows:

$$p_{ik} = \frac{N_{ik}}{N}.$$

If each cluster contains the same number of elements, we obtain a uniform distribution. Notice that we also obtain a uniform distribution on the lowest level, when each element is a cluster. Therefore, we introduce a threshold $1/N$ and investigate obtained distributions with respect to this initial uniform distribution.

Correspondingly, one can describe each level i of a dendrogram with following variables: (1) The total number of clusters K_i on level i. (2) The total number of elements with probability over the threshold $p_{ik} > \frac{1}{N}$ of level i, namely, $M_i = \sum_k N_{ik} \cdot \mathbb{1}_{\left(\frac{N_{ik}}{N} - \frac{1}{N}\right)}$, where the step function $\mathbb{1}_{(\cdot)}$ is defined by $\mathbb{1}_{(x-y)} = 1$ if $x \geq y$ and $\mathbb{1}_{(x-y)} = 0$ if $x < y$. (3) The sum of high probabilities $\tilde{P}_i$, i.e., probabilities larger than $1/N$, namely, $\tilde{P}_i = \sum_k p_{ik} \cdot \mathbb{1}_{(p_{ik} - \frac{1}{N})}$.

We can measure all these variables in the process of data clustering. With these values, one can determine internal energy and Gibbs-Shannon entropy at a given level in the following way:

$$E_i = -\ln\left(\frac{\tilde{P}}{K_i}\right),$$

$$S_i = \ln\left(\frac{M_i}{N}\right).$$

With Gibbs-Shannon entropy and internal energy, one can define free energy and Rényi entropy for each level of a hierarchy. Free energy of a hierarchical level i is expressed as

$$F_i = E_i - K_i S_i.$$

And Rényi entropy of level i can be expressed as follows [5]:

$$S_i^R = \frac{F_i}{1-q},$$

where $q = \frac{1}{K_i}$ is a deformation parameter. Thus, our approach allows us to estimate the process of hierarchical clustering from a perspective of behaviour of Rényi entropy under transition between levels, i.e., to estimate the dependence of entropy on the number of clusters.

The process of clustering begins with minimum information (maximum Rényi entropy) and also ends with maximum Rényi entropy. Hence, minimum Rényi entropy (information maximum) is located somewhere in between these maxima. Particular data features will define the location of the global minimum and a set of local minima.

4 Experiment

We test our approach on data of user-profiles from the leading Russian OSN *Vkontakte* (VK). We collected the data through VK API [30]. Then, we anonymized user data, i.e., names, surnames and IDs were deleted to avoid the possibility of revealing real users. The dataset includes digital traces of user activity such as numbers of likes, posts, reposts, comments; indicators of subscribing to one or more pages from 12 national news channels publishing news on VK; as well as user stated political beliefs (one of eight). In total, the dataset has 47 user attributes of a total 50,000 users. Our attempts to cluster this dataset with K-means and C-means while searching for the optimal number of clusters with gap statistics, jump and silhouette methods were unsuccessful.

On a machine with 64 GB RAM and i7-6700 CPU @ 3.40 GHz (four cores), we were unable to run hierarchical clustering and to test our approach on more massive datasets since the algorithm has time complexity $O(n^2)$ and uses $O(n^2)$ memory, where n is the number of samples. However, the hierarchical clustering of our data on the mentioned machine takes about 8 h (28,798 s).

We test our approach in two stages. First, we conduct hierarchical clustering using scipy.cluster.hierarchy Python package [16] with the 'complete' method of calculating the distance between newly formed clusters [17], namely, the distance between clusters u and v is expressed as $d(u, v) = \max_{i,j}(\text{dist}(u[i], v[j]))$, where 'dist' refers to Euclidean distance, $u[i]$ and $v[j]$ are objects contained in cluster u and cluster v, correspondingly. In each iteration, we select and merge two or more clusters with the smallest distance. This stage produces a hierarchy of clusters which can be visualized in the form of the dendrogram (Fig. 1). One can see how manual analysis of such dendrogram could be problematic.

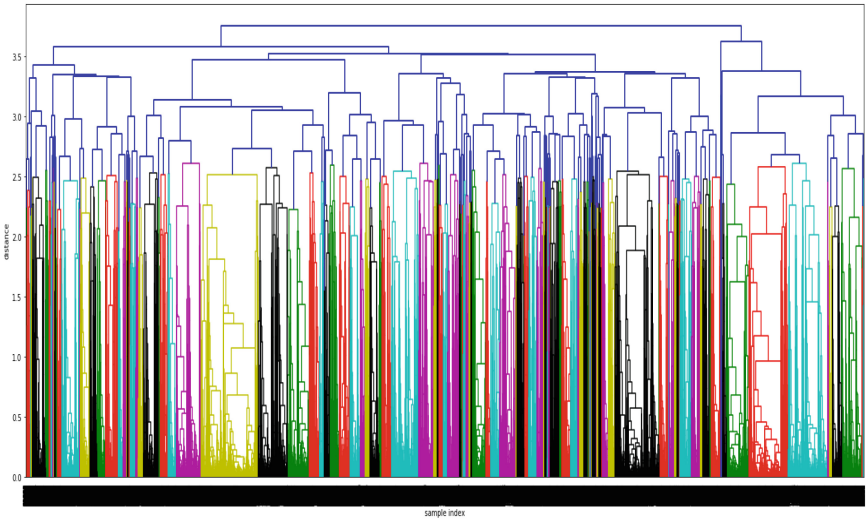

Fig. 1. The dendrogram of clustering 50,000 OSN user profiles.

Then, we calculate the number of obtained clusters on each level of the hierarchy and the number of users in each cluster. Here, all users belong to the same cluster on the upper level of the hierarchy, and each user belongs to a separate cluster on the bottom level of the hierarchy, i.e., the lowest level contains 50,000 clusters. Then, we compute Gibbs-Shannon entropy, internal energy, free energy and Rényi entropy.[1] Finally, we will consider our approach valid if (1) it will show a clear entropy minimum (a maximum of information) and (2) the entropy maxima (the minima of information) will correspond to the borderline states.

[1] An example of calculations in Python is available here: https://github.com/hse-scila/entropic-approach-hierarchical-clustering.

5 Results and Conclusion

Figure 2 and 3 show two mutually opposite processes present during hierarchical clustering of social media users. The first process is the decrease of Gibbs-Shannon entropy with a rising number of clusters, which means that the equilibrium state corresponds to the minimum of a given entropy. The equilibrium corresponds to the state when each user is a separate cluster.

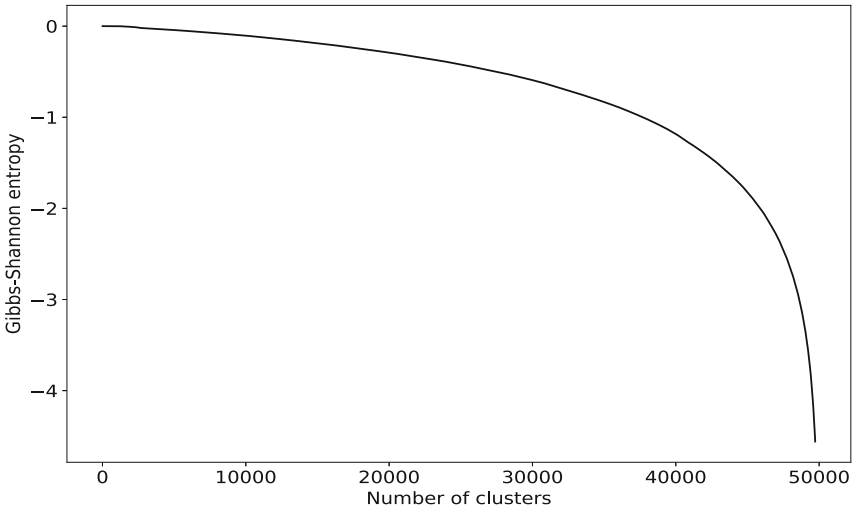

Fig. 2. Distribution of Gibbs-Shannon entropy over the number of clusters.

The second process is the increase of internal energy with a rising number of clusters (Fig. 3). The difference between these two processes has an area where they balance each other (Fig. 4). In this area, the Rényi entropy has its minimum value. Hence, the minimum of the Rényi entropy corresponds the maximum of information of a hierarchical model. For this dataset, the minimum of the Rényi entropy lies at the 1,281 clusters or 50,000 users could be grouped in 1,281 clusters (Fig. 4). In the machine learning terms, the left branch of the Rényi entropy indicates underfitting while the right branch to overfitting. Thus, the minimum of the Rényi entropy indicates the optimal parameters of hierarchical clustering.

In this work, we propose a criterion of finding the optimal number of clusters for hierarchical clustering, using entropic formalism with deformed Rényi entropy where the parameter of deformation is the number of clusters. This approach could be used for such algorithms as Infinite Mixture Models with Nonparametric Bayes and the Dirichlet Process with various implementations (Chinese restaurant process, stick-breaking algorithm).

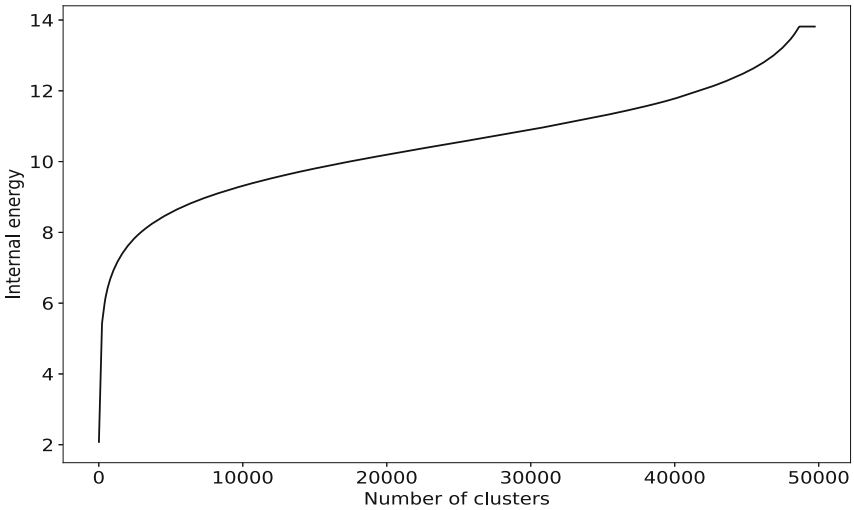

Fig. 3. Distribution of internal energy over the number of clusters.

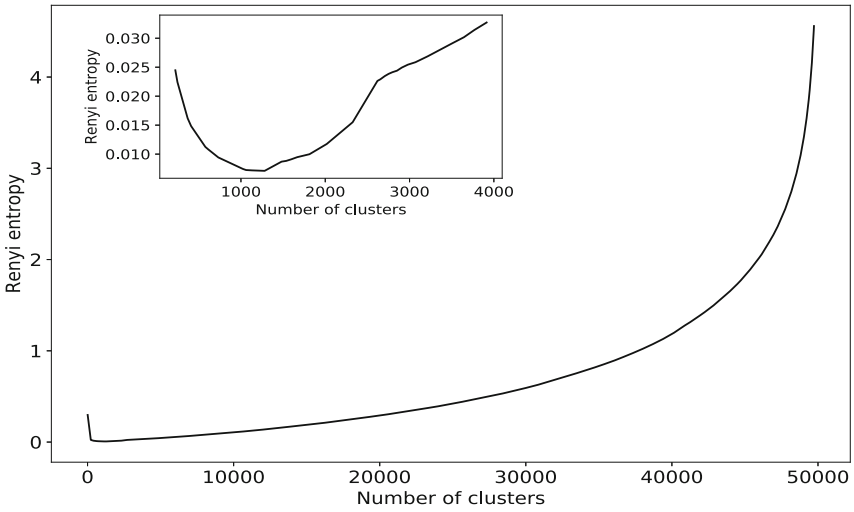

Fig. 4. Distribution of Rényi entropy over the number of clusters.

In further, we plan to test our approach with large synthetic data with a predefined number of clusters. One potential area of further testing is to consider if various combinations of user features affect the global Rényi minimum location. Another direction is to consider other than Euclidean distances to asses their fitness for hierarchical clustering of common types of data from OSN.

Acknowledgments. The reported study was funded by RFBR according to the research project No 18-011-00997 A.

References

1. Aldana-Bobadilla, E., Kuri-Morales, A.: A clustering method based on the maximum entropy principle. Entropy **17**(1), 151–180 (2015)
2. AlSumait, L., Barbará, D., Domeniconi, C.: On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM 2008, pp. 3–12, Washington, DC, USA. IEEE Computer Society (2008)
3. José, A., Balogh, S., Hernández, S.: A brief review of generalized entropies. Entropy **20**(11), 813 (2018)
4. Bao, Q., Cheung, W.K., Liu, J.: Inferring motif-based diffusion models for social networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, pp. 3677–3683. AAAI Press (2016)
5. Beck, C.: Generalised information and entropy measures in physics. Contemporary Phys. **50**(4), 495–510 (2009)
6. Bollobás, B., Riordan, O.M.: Mathematical results on scale-free random graphs. In: Bornholdt, S., Schuster, H.G. (eds.) Handbook of Graphs and Networks: From the Genome to the Internet, 1st edn, pp. 1–34. Wiley, Weinheim (2003)
7. De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, S.K., Xie, L., Kelliher, A.: How does the data sampling strategy impact the discovery of information diffusion in social media? In: ICWSM (2010)
8. Dehmer, M., Emmert-Streib, F.: Analysis of Complex Networks: From Biology to Linguistics. Wiley, Hoboken (2009)
9. Dehmer, M., Emmert-Streib, F., Chen, Z., Li, X., Shi, Y. (eds.): Mathematical Foundations and Applications of Graph Entropy. Wiley, Weinheim (2016)
10. Elayat, H., Murphy, B., Prabhakar, N.: Entropy in the hierarchical cluster analysis of hospitals. Health Serv. Res. **13**(4), 395–403 (1978)
11. Erdős, P., Rényi, A.: On the evolution of random graphs. In: The Structure and Dynamics of Networks, pp. 38–82. Princeton University Press, Princeton (2011)
12. Fogués, R.L., Such, J.M., Minguet, A.E., García-Fornes, A.: Open challenges in relationship-based privacy mechanisms for social network services. Int. J. Hum. Comput. Interaction **31**, 350–370 (2015)
13. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
14. Guille, A., Hacid, H.: A predictive model for the temporal dynamics of information diffusion in online social networks. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012 Companion, pp. 1145–1152. ACM, New York (2012)
15. Guimerà, R., Nunes Amaral, L.A.: Functional cartography of complex metabolic networks. Nature **433**(7028), 895–900 (2005)
16. Hierarchical clustering (scipy.cluster.hierarchy)—SciPy v1.3.1 Reference Guide
17. Hierarchical clustering (scipy.cluster.hierarchy.linkage)—SciPy v1.3.1 Reference Guide
18. Ketchen, D., Shook, C.: The application of cluster analysis in strategic management research: an analysis and critique. Strategic Manage. J. **17**, 441–458 (1996)
19. Kitsak, M., Gallos, L., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H., Makse, H.: Identification of influential spreaders in complex networks. Nat. Phys. **6**(11), 888–893 (2010)

20. Koltcov, S.: Application of rényi and tsallis entropies to topic modeling optimization. Phys. A: Stat. Mech. Appl. **512**, 1192–1204 (2018)
21. Koltcov, S., Ignatenko, V., Koltsova, O.: Estimating topic modeling performance with sharma-mittal entropy. Entropy **21**(7), 660 (2019)
22. Newman, M.E.J.: Models of the small world. J. Stat. Phys. **101**(3), 819–841 (2000)
23. O'Donovan, F.T., Fournelle, C., Gaffigan, S., Brdiczka, O., Shen, J., Liu, J., Moore, K.E.: Characterizing user behavior and information propagation on a social multimedia network. In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)
24. Olemskoi, A.: Synergetics of Complex Systems: Phenomenology and Statistical Theory [Sinergetika slozhnyh sistem. Fenome-nologiya i statisticheskaya teoriya]. KRASAND, Moscow (2009)
25. Rose, K., Gurewitz, E., Fox, G.C.: Statistical mechanics and phase transitions in clustering. Phys. Rev. Lett. **65**(8), 945–948 (1990)
26. Rytsarev, I.A., Kupriyanov, A.V., Kirsh, D.V., Liseckiy, K.S.: Clustering of social media content with the use of BigData technology. J. Phys. Conf. Ser. **1096**, 012085 (2018)
27. Sugar, C.A., James, G.M.: Finding the number of clusters in a dataset: an information-theoretic approach. J. Am. Stat. Assoc. **98**(463), 750–763 (2003)
28. Suyari, H., Wada, T.: Scaling property and the generalized entropy uniquely determined by a fundamental nonlinear differential equation. arXiv (2006)
29. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. Royal Stat. Soc. Ser. B (Statistical Methodology) **63**(2), 411–423 (2001)
30. VK API guide. https://vk.com/dev/manuals
31. Wang, Y., Zhang, Z.-M., Peng, Z.-H., Duan, Y.-Y., Gao, Z.-Q.: A cascading diffusion prediction model in micro-blog based on multi-dimensional features. In: Barolli, L., Zhang, M., Wang, X.-A. (eds.) Advances in Internetworking, Data & Web Technologies, pp. 734–746, Springer, Cham (2018)
32. Zhang, Q., Li, M., Deng, Y.: A new structure entropy of complex networks based on nonextensive statistical mechanics. Int. J. Modern Phys. C **27**(10), 1650118 (2016)