UDC 519.17, 519.237

Implementation, Efficiency Analysis and Quality Evaluation of Clustering Algorithms for Graph Models of Social Networks

M. S. Ionkin, M. V. Ogneva

Michael S. Ionkin, orcid.org/0000-0002-4726-8245, Saratov State University, 83, Astrakhanskaya Str., Saratov, Russia, 410012, msionkin@gmail.com
Marina V. Ogneva, orcid.org/0000-0002-9828-7681, Saratov State University, 83, Astrakhanskaya Str., Saratov, Russia, 410012, ognevamv@gmail.com

The article deals with the community detection problem (the clustering problem) for undirected graphs. The clustering (grouping together of similar objects) is one of the fundamental tasks in the data analysis. This task is applied in a wide range of areas: image segmentation, marketing, anti-fraud, forecasting, text analysis and much more. At the moment, there is no universal and effective solution of this problem. There are several dozens of methods and there are many modifications of them which group objects that are as similar as possible to each other. The article describes algorithms for solving this task, presents their asymptotic complexity estimates, traditional metrics and quality functionals needed to evaluate the results of their work. The authors propose a solution to the problem which is the opposite of the resolution limit problem (algorithms find communities that are quite small in relation to the entire graph). The authors implemented the Smart Local Moving algorithm which is an improvement of the well-known Louvain algorithm. Performed an experimental comparison of the considered algorithms efficiency on large sparse graphs containing several hundreds of thousands of vertices and edges which corresponding to real data from YouTube, Amazon, Live Journal. The comparative analysis was performed on these three “impersonal” data sets with a previously known division into communities (ground-truth communities), as well as on a data set with all available information about the vertices (users) from the social network “Vkontakte”. The communities found by different algorithms on the same data set were also compared with each other. The authors examined such characteristics as the execution time of algorithms, values of modularity and normalized mutual information.

Key words: clustering, community detection, graph models, data analysis.

References

1. Aggarwal C. C., Charu C., Reddy C. K. *Data clustering. Algorithms and applications.* New York, CRC Зкуыы, 2014. 652 p.
2. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review. *ACM Computing Surveys*, 1999, vol. 31, no. 3, pp. 264–323. DOI: 10.1145/331499.331504.
3. Newman M. E. J. Detecting community structure in networks. *The European Physical Journal B – Condensed Matter and Complex Systems*, 2004, vol. 38, no. 2, pp. 321–330. DOI: 10.1140/epjb/e2004-00124-y.
4. Leskovec J., Rajaraman A., Ullman J. *Mining of massive datasets.* 2nd ed. Cambridge Univ. Press, 2014. 511 p.
5. Fortunato S. Community detection in graphs. *Physics Reports*, 2010, vol. 486, iss. 3, pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

6. *Information and analytical resource dedicated to machine learning, pattern recognition and intelligent data analysis*. Available at: http://www.machinelearning.ru (accessed 12 February, 2017) (in Russian).
7. Rosvall M., Axelsson D., Bergstrom C. T. The map equation. *The European Physical Journal – Special Topics*, 2009, vol. 178,iss. 1, pp. 13–23. DOI: 10.1140/epjst/e2010-01179-1.
8. Pons P., Latapy M. Computing communities in large networks using random walks. *Computer and Information Sciences – ISCIS 2005*, 2005, pp. 284–293. DOI: 10.1007/11569596_31.
9. Raghavan U. N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 2007, vol. 76, iss. 3, pp. 036106. DOI: 10.1103/PhysRevE.76.036106.
10. Clauset A., Newman M. E. J., Moore C. Finding community structure in very large networks. *Phys. Rev. E*, 2004, vol. 70, iss. 6, pp. 066111. DOI: 10.1103/PhysRevE.70.066111.
11. Girvan M., Newman M. E. J. Community structure in social and biological networks. *Proc. National Academy of Sciences*, 2002, vol. 99, no. 12, pp. 7821–7826. DOI: 10.1073/pnas.122653799.
12. Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, vol. 2008, no. 10, pp. P10008. DOI: 10.1088/1742-5468/2008/10/P10008.
13. Waltman L., Eck N. J. A smart local moving algorithm for large–scale modularity–based community detection. *The European Physical Journal B*, 2013, vol. 86, no. 11, pp. 471. DOI: 10.1140/epjb/e2013-40829-0.
14. Romano S., Bailey J., Nguyen V., Verspoor K. Standardized mutual information for clustering comparisons : one step further in adjustment for chance. *Proc. 31st International Conference on Machine Learning*. Beijing, China, PMLR, 2014, vol. 32, no. 2, pp. 1143–1151. Available at: http://proceedings.mlr.press/v32/romano14.pdf (accessed 25 April, 2017).
15. Haykin S. *Neural Networks: A Comprehensive Foundation*. Singapore, Pearson Education (Singapore) Pte Ltd., 1998. 842 p. (Russ. ed. : Moscow, Publ. House Williams, 2006. 1104 p.)
16. Fortunato S., Barthelemy M. Resolution limit in community detection. *Proc. National Academy of Sciences*, 2007, no. 104, pp. 36–41. DOI: 10.1073/pnas.0605965104.
17. Traag V. A., Dooren P. V., Nesterov Y. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E*, 2011, vol. 84, iss. 1, pp. 016114. DOI: 10.1103/PhysRevE.84.016114.
18. *Stanford Large Network Dataset Collection*. Available at: https://snap.stanford.edu/data (accessed 25 April, 2017).

Cite this article as:
Ionkin M. S., Ogneva M. V. Implementation, Efficiency Analysis and Quality Evaluation of Clustering Algorithms for Graph Models of Social Networks. *Izv. Saratov Univ. (N. S.), Ser. Math. Mech. Inform.*, 2017, vol. 17, iss. 4, pp. 441–451 (in Russian). DOI: 10.18500/1816-9791-2017-17-4-441-451.