УДК 519.173.5

Влияние выбора графовой метрики на качество кластеризации графов и стохастических сетей

В.С. Ивашкин
Московский физико-технический институт (государственный университет)

В современном мире существует большое количество систем и процессов, которые удобно анализировать в графовом представлении. Графами можно описывать как огромные структуры, такие как социальные сети, интернет, так и сравнительно небольшие локальные системы. Вершины обозначают объекты системы, а ребра — связи, соотношения между вершинами. Для графа общего вида практическая задача измерения расстоянии/близости между вершинами не имеет однозначного решения: она решается выбором той или иной метрики. От удачности этого выбора зависит качество решения прикладной задачи. Например, преимущество поисковой машины Google долгое время во многом определялось использованием метрики PageRank, основанной на ссылочной связности страниц.

Одной из центральных задач в анализе графов и сетей является задача кластеризации. Кластеризация в данном контексте — задача разбиения множества вершин графа на группы вершин исходя из их схожести без какой-либо предварительной разметки (т.н. «обучение без учителя»). Целью данной работы является сравнение качества кластеризации при использовании наиболее известных графовых метрик, а также определение наилучшей метрики для каждого из исследуемых типов графов.

В качестве метрик используются однопараметрические семейства Plain Walk [1], Forest [2], Communicability [3], Heat [4], их поэлементно логарифмированные версии Walk [5], Logarithmic Forest [6], Logarithmic Communicability, Logarithmic Heat, а также семейства Sigmoid Commute Time [7], Sigmoid Corrected Commute Time [7], Randomized Shortest Path [8] и Free Energy [8].

Исследование проводилось на графах, генерируемых по модели с различными вероятностями возникновения ребра снаружи и внутри кластера (модель $G(n, p_{\text{in}}, p_{\text{out}})$), а также на известных тестовых множествах (datasets). Исследована зависимость качества кластеризации от параметров метрик, а также от различных параметров генерации графов. В частности, исследована зависимость качества кластеризации от сбалансированности кластеров для различных метрик. В результате сделаны выводы о том, какую метрику стоит использовать для кластеризации графов в зависимости от их параметров.

Литература

1. *Чеботарев П.Ю., Шамис Е.В.* О мерах близости вершин графов // Автоматика и Телемеханика, 1998. Т. 59 (10) С. 113–133.
2. *Chebotarev P., Shamis E.* The forest metrics for graph vertices // Electronic Notes in Discrete Mathematics, 2002. Vol. 11. P. 98–107.
3. *Estrada, E.* The communicability distance in graphs // Linear Algebra and its Applications, 2012. Vol. 436(11). P. 4317–4328.
4. *Chung F., Yau S.T.* Coverings, heat kernels and spanning trees // Journal of Combinatorics, 1998. Vol. 6. P. 163–184.
5. *Chebotarev P.* The walk distances in graphs // Discrete Applied Mathematics, 2012. Vol. 160 (10–11). P. 1484–1500.
6. *Chebotarev P.* Studying new classes of graph metrics // Geometric Science of Information, 2013. P. 207–214.
7. *Fouss F., Saerens M., Shimbo M.* Algorithms and Models for Network Data and Link Analysis. Cambridge University Press, 2016.
8. *Kivimäki I., Shimbo M., Saerens M.* Developments in the theory of randomized shortest paths with a comparison of graph node distances // Physica A: Statistical Mechanics and its Applications, 2014. Vol. 393. P. 600–616.