ОБНАРУЖЕНИЕ СООБЩЕСТВ АЛГОРИТМОМ ЛЕЙДЕНА

А. О. Яблонская

Белорусский государственный университет, г. Минск;
anna.yablonskaya2002@yandex.ru;
науч. рук. – А. Э. Малевич, канд. физ.-мат. наук, доц.

В статье рассматриваются методы обнаружения сообществ в графах. Подробно исследован алгоритм Лейдена разбиения графа на сообщества. Описаны примеры применения алгоритма Лейдена: рекомендательная система фильмов, взаимосвязи героев романа Виктора Гюго «Отверженные», игры по американскому футболу между различными колледжами.

Ключевые слова: графы; кластеризация; обнаружение сообществ; алгоритм Лейдена; приложения алгоритма Лейдена.

ВВЕДЕНИЕ

Одной из важных характеристик графов, представляющих реальные системы, является разбиение на сообщества, т. е. организация вершин графа в подгруппы таким образом, чтобы вершины внутри каждой группы соединялись множеством рёбер, но сравнительно мало рёбер соединяли вершины из разных групп. Обнаружение сообществ имеет большое значение в социологии, биологии и компьютерных науках.

Множество вершин графа называется сообществом, если связь вершин в данном множестве лучше, чем связь между вершинами множества и вершинами вне него. На рис. 1 вершины, выделенные одним цветом, составляют сообщества.

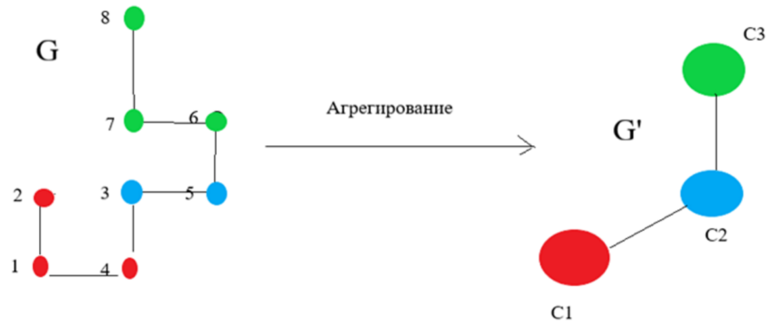

Рис. 1. Разбиение графа

Разбиение графа $\mathcal{P} = \{C_1, \ldots, C_r\}$ состоит из $r = |\mathcal{P}|$ сообществ, где каждое сообщество $C_i \subseteq \mathcal{P}$ состоит из множества узлов $V = \bigcup_i C_i$, $C_i \cap C_j = \emptyset$ для любых $i \neq j$.

Одним из способов нахождения сообществ является учёт модульности. Использование этого понятия позволяет максимизировать разницу между фактическим и ожидаем количеством рёбер в сообществе C. Существуют различные виды модульности, однако для данного подхода интересна лишь константная модель Поттса. Согласно этой модели функцию качества $\mathcal{H}(G, \mathcal{P})$ или кратко $\mathcal{H}(\mathcal{P})$ разбиения $\mathcal{P}$ на сообщества графа G запишем в виде

$$\mathcal{H}(\mathcal{P}) = -\sum_{\mathrm{C}} e_c - \gamma {n_c}^2,$$

где каждое сообщество $C \in \mathcal{P}$ состоит из e_c ребер и n_c вершин, а γ – это параметр разрешения, существенно влияющий на качество разбиения графа на сообщества.

АЛГОРИТМЫ ОБНАРУЖЕНИЯ СООБЩЕСТВ

Существуют различные алгоритмы обнаружения сообществ. Наиболее популярными являются алгоритм Лувена и алгоритм Лейдена [1]. Оба этих алгоритма основаны на модульности и рассматриваются как методы быстрого обнаружения сообществ в сети. Алгоритм Лувена имеет несколько недостатков. Во-первых, данный алгоритм может находить «плохо» связанные сообщества, т.е. сообщества, которые не связаны внутренне. Во-вторых, данный алгоритм даёт лишь одну гарантию: разбиения на сообщества, вершины которых невозможно объединить в другие сообщества. Другими словами, он гарантирует, что сообщества хорошо разделены. Все данные недостатки были устранены в алгоритме Лейдена.

Данный алгоритм состоит из трех фаз (см. рис. 2):

5. Быстрое локальное перемещение узлов. В отличие от алгоритма Лувена, данный алгоритм не посещает все узлы до тех пор, пока они не перестанут перемещаться, тем самым посещая и те узлы, которые нельзя перемещать. В процедуре быстрого локального перемещения посещаются только узлы, окрестности которых изменились.

6. Уточнение разделения. На данном этапе алгоритм находит разбиение $\mathcal{P}_{refined}$, которое является уточнением разбиения $\mathcal{P}$, полученного на первом этапе алгоритма. В процессе уточнения происходит перемещение некоторых вершин, при этом целевое сообщество выбирается случайным образом из списка подходящих сообществ-кандидатов. Степень случайности выбора сообщества определяется параметром $\theta > 0$. Затем сообщества из $\mathcal{P}$ могут быть разделены на подсообщества.

7. Агрегация сети. На этом этапе происходит агрегация графа с учётом уточнённого разбиения.

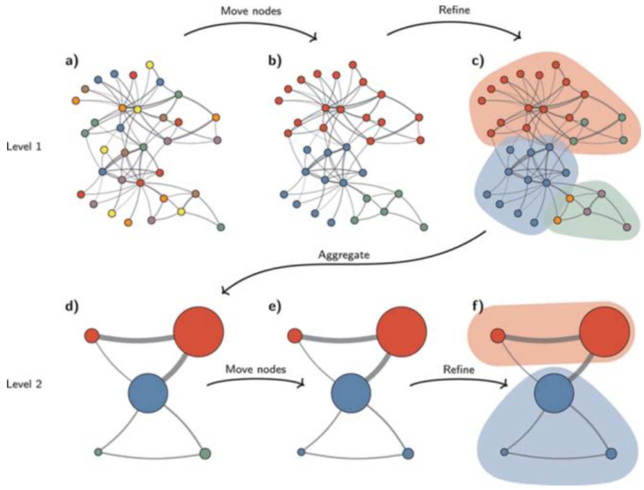

Рис. 2. Алгоритм Лейдена [1]

ПРИМЕНЕНИЕ АЛГОРИТМА

Алгоритм был протестирован на нескольких данных. Первые данные, на которых был опробован алгоритм, представляют собой рекомендательную систему фильмов [2]. Здесь поиск сообществ играет огромную роль, поскольку он даёт возможность определить интересы пользователей, а также объединить фильмы в группы по каким-либо признакам, что позволяет в свою очередь подбирать правильные рекомендации.

Второй набор данных, на котором был использован алгоритм, содержат в себе взвешенную сеть взаимосвязей героев романа Виктора Гюго «Отверженные» [3]. Этот граф является примером простого социального взаимодействия людей, что является одной из самых популярных задач, где применяется алгоритм.

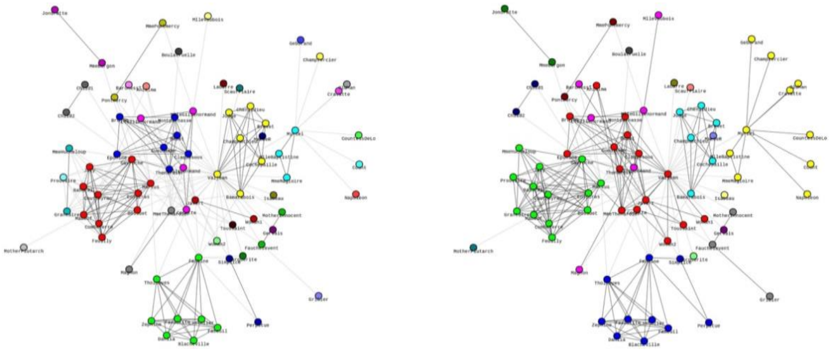

Рис. 3. Два варианта разбиения графа «Отверженные» на сообщества

Третий набор данных содержит сеть игр по американскому футболу между различными колледжами [3]. Его также можно рассматривать как граф социального взаимодействия.

Зависимость функции качества от параметра разрешения и степени случайности для всех трёх рассмотренных наборов данных представлена в таблице.

Зависимость функции качества от параметра разрешения

Фильмы		«Отверженные»			Футбол		
Параметр разрешения	Значение функции качества	Параметр разрешения	Степень случайности	Значение функции качества	Параметр разрешения	Степень случайности	Значение функции качества
1.0000	0.418	0.800	6	0.378	0.800	20	0.437
0.5000	0.598	0.800	5	0.396	0.800	5	0.412
0.0500	0.831	0.200	3	0.481	0.100	42	0.604
0.0100	0.891	0.200	7	0.489	0.400	7	0.571
0.0010	0.929	0.150	7	0.493	0.090	1	0.604
0.0005	0.936	0.110	1	0.526	0.020	42	0.400
0.0001	0.940	0.108	1	0.514	0.010	1	0.000

В результате тестирования были выявлены зависимости значений функции качества от параметра разрешения и степени случайности для разобранных примеров. Чем ниже параметр разрешения, тем выше качество разбиения. Было обнаружено также, что степень случайности незначительно влияет на качество разбиения.

Библиографические ссылки

1. *Traag V. A.* From Louvain to Leiden: guaranteeing well-connected communities / V. A. Traag, L. Waltman, N. j. van Eck // Sci Rep 9, 5233 (2019). DOI: 10.1038/s41598-019-41695-z.
2. Парсинг базы фильмов из IMDB. Сбор информации о фильмах и рекомендаций к ним / 2019. [Электронный ресурс] – Режим доступа: https://a-parser.com/resources/268/ – Дата доступа: 06.06.2022.
3. *Newman M.* Network data / 2013. [Электронный ресурс] — Режим доступа: http://www-personal.umich.edu/~mejn/netdata/ — Дата доступа: 06.06.2022.