



УДК 519.237

ВЫДЕЛЕНИЕ СООБЩЕСТВ В ГРАФОВЫХ МОДЕЛЯХ НА ОСНОВЕ КЛАСТЕРНОГО АНАЛИЗА

В.И. Ложечка, loje4kavica@yandex.ru, А.Н. Целых, e-mail: ant@sfedu.ru

Инженерно-технологическая академия Южного федерального университета (ИТА ЮФУ),
Таганрог

Исследования, проведенные в данной статье, направлены на использование методов кластерного анализа применительно к социальным сетям, представленным в виде графовых моделей. Для выделения в социальном графе различного рода сообществ используются методы кластерного анализа, а эффективность кластеризации оценивается с помощью математических критериев качества. Выявление закономерностей в исходных данных с использованием методов интеллектуального анализа осуществляется на основе математического моделирования. В статье рассмотрены основные принципы функционирования алгоритмов, приведен их сравнительный анализ, описаны преимущества и недостатки каждого из используемых в работе алгоритмов.

Ключевые слова: кластеризация, модульность, сетевой граф, оптимальный алгоритм кластеризации, Label propagation, Fast-greedy.

THE EXTRACTION OF COMMUNITIES IN GRAPH MODELS BASED ON THE CLUSTER ANALYSIS

Lozhechka V.I., Tselykh A.N.

Academy for Engineering and Technologies of Southern Federal University,
Taganrog

The research presented in this paper is dedicated to the use of cluster analysis methods in relation to social networks, presented in the form of graph models. Cluster analysis methods are used to identify different types of communities in a social graph, and the effectiveness of clustering is evaluated using mathematical quality criteria. Identification of regularities in the source data using the methods of intellectual analysis is carried out on the basis of mathematical modeling. The article describes the basic principles of algorithms, a comparative analysis, the advantages and disadvantages of each of the algorithms used in the research.

Keywords: clustering, modularity, network graph, optimal clustering algorithm, Label propagation, Fast-greedy.

Современные достижения в области технологий и вычислительной мощности позволяют осуществлять сбор и анализ огромного количества реальных данных. Сбор таких данных позволяет понять структуру и работу реальных систем и найти некие закономерности в данных. Ряд реальных проблем, связанных с раскрытием организационных принципов, можно представить в виде сложной сети с вершинами, которые представляют собой различные объекты, и ребрами, представляющими отношения между объектами. Социальные сети, рассматриваемые в данной статье, это типичная сложная сеть, которую можно рассматривать как соответствующие сообщества, где сообщество будет содержать пользователей, которые более тесно связаны. Социальные сети не ограничиваются социальными сетями, работающими онлайн, такими как «ВКонтакте» или «Facebook». Другими примерами социальных сетей являются сети людей, которые сотрудничают друг с другом, соавторство, а также сети общения между людьми, такие как телекоммуникации и электронная почта, но в данной статье будет рассмотрена социальная сеть «ВКонтакте».

Многие типы реальных наборов данных могут быть смоделированы с помощью графов. Графики могут представлять объекты данных, а также отношения



между объектами и предоставлять мощный математический инструмент для представления отношений в данных. Например, социальная сеть - это график, связывающий людей, которые контактируют или взаимодействуют друг с другом; узлы и ребра представляют людей и взаимодействия между людьми соответственно.

Чтобы понять структуры и функции, скрытые в крупномасштабных графах, используется кластерный анализ графов, который является одним из наиболее важных методов в различных областях исследований, таких как интеллектуальный анализ данных. Кластер можно рассматривать как группу узлов, которые плотно связаны внутри группы, но в то же время слабо связаны с другими сообществами. Обнаружив скрытые кластерные структуры в крупномасштабных графах, можно понять характеристики и взаимосвязи узлов, образующих граф. В данной статье мы рассмотрим некоторые из алгоритмов кластерного анализа, применительно к небольшой сети, созданной на основании данных из социальной сети «ВКонтакте».

Алгоритм кластеризации *optimal*. Алгоритм кластеризации *optimal* [1] основан на вычислении оптимальной структуры сообществ графа с точки зрения максимально оценки модульности. При использовании данного алгоритма кластеризации необходимо учитывать, что размер и структура кластеров при оптимальной кластеризации зависят от общего количества каналов связи. Вычисления, необходимые для оптимального разбиения графа на сообщества, выполняются путем преобразования максимизации модульности в задачу целочисленного программирования и последующего вызова библиотеки GLPK для ее решения.

Алгоритм кластеризации *optimal* может работать как со взвешенными сетями, так и с сетями без указания веса ребер. Если у графа есть такой атрибут ребра, как вес, то он используется по умолчанию. Так же при использовании алгоритма можно задать игнорирование существующих весовых коэффициентов.

Необходимо обратить внимание, что оптимизация модульности является NP-полной задачей, и все известные алгоритмы для нее имеют экспоненциальную временную сложность. Это означает, что алгоритм имеет большую вычислительную сложность, поэтому при работе с большими графами следует избегать его применения. Алгоритм быстро обрабатывает графики с количеством вершин в пределах 50, но если граф имеет несколько сотен вершин, то качество кластеризации ухудшится, а время работы увеличится.

Label.propagation. Подходы в кластеризации, основанные на распространении меток, могут восполнять недостающие связи между объектами на основании информации о соединениях, предоставленной сетью. На таком подходе основан алгоритм кластеризации *Label.propagation* [2].

Работа алгоритма состоит в следующем. Пусть изначально узел x имеет j соседей, а на итерации t данный узел имел метку $C_x(t)$. Тогда работа алгоритма заключается в выполнении следующих шагов:

1. Инициализируется метка для каждого узла ($t = 0, C_x(0) = x$)/
2. Значение t принимается равным одному и произвольно выбирается порядок обновления меток.
3. В установленном порядке обновляются все метки узлов



4. Если каждый узел сети имеет метку равную как минимум половине соседних меток, то процесс разделения сети останавливается. При этом все узлы, имеющие одинаковые метки, помещаются в один кластер.

5. Если же условие п.4 не выполняется, то устанавливается $t = t + 1$ и происходит переход к пункту 3.

Процесс обновления метки на шаге 4 является асинхронным. При синхронном обновлении узел x на t –й итерации обновляет свою метку на основе меток своих соседей на $(t - 1)$ -й итерации. Синхронное обновление в среднем более стабильно, но сходится гораздо медленнее, чем асинхронное обновление.

Fast-greedy. Алгоритм кластеризации Fast-greedy [3] основан на иерархическом подходе, но при этом он идет сверху вниз, а не снизу-вверх. Данный алгоритм пытается жадно оптимизировать функцию качества, называемую модульностью. Первоначально каждая вершина принадлежит отдельному сообществу, и сообщества объединяются итеративно, так что каждое объединение является локально оптимальным, т.е. дает наибольшее увеличение или наименьшее уменьшение текущего значения модульности. Алгоритм останавливается, когда больше невозможно увеличить модульность. Метод кластеризации Fast-greedy является довольно быстрым. Он обычно используется для первоначального анализа, потому что у него нет параметров для настройки. Но данный алгоритм имеет предел разрешения. Это значит, что сообщества ниже заданного порогового размера (в зависимости от количества узлов и ребер) всегда будут объединяться с соседними сообществами.

Экспериментальное исследование. В данной статье мы исследовали работу описанных ранее алгоритмов кластеризации на созданном нами графе. Изначально было необходимо сформировать граф. Для этого была сформирована матрица смежности, которая представляла собой связи друзей между собой. Если два человека имели дружескую связь, то она обозначалась единицей в матрице смежности, в ином случае записывался 0. Все данные использованные для формирования графа были взяты из социальной сети «ВКонтакте». Так как граф получился сравнительно небольшого размера, то можно применить все алгоритмы кластеризации, описанные нами.

Для оценки качества кластеризации мы использовали два важных параметра, а именно модульности и количество кластеров.

С точки зрения сетевой науки проблема кластеризации определяется как «обнаружение сообщества». Популярным методом обнаружения сообщества является оптимизация модульности. Модульность функции качества должна быть максимизирована для обеспечения хорошей кластеризации сети. Модульность разбиения g определяется как

$$Q(g) = \frac{1}{2m} \sum_{i,j=1}^N \left(\omega_{ij} - \gamma \frac{d_i d_j}{2m} \right) \delta(g_i, g_j),$$

где m - общее количество ребер в графе, ω_{ij} - элемент матрицы, g_i, g_j - метка вершины (номер сообщества, к которому относится вершина), а $\delta(g_i, g_j)$ это символ Кронекера, равный одному в случае если $g_i = g_j$ и нулю, если $g_i \neq g_j$.

Количество кластеров является важным фактором алгоритмов кластеризации. В некоторых методах, существует заранее назначенное количество кластеров, которое служит явным ограничением в функции качества. Так как в статье



проводится анализ социальной сети, то истинное количество кластеров не может быть известно заранее, поэтому используются алгоритмы, не требующие фиксированного задания количества выделяемых сообществ.

Помимо указанных параметров алгоритмы сравнивались по количеству времени, которое им понадобилось для разбиения графа на сообщества. Этот параметр возможно не так важен в исследовании данной сети, но он позволяет увидеть, как вычислительная сложность влияет на время работы алгоритма.

Полученные результаты описанных параметров можно увидеть в следующей таблице.

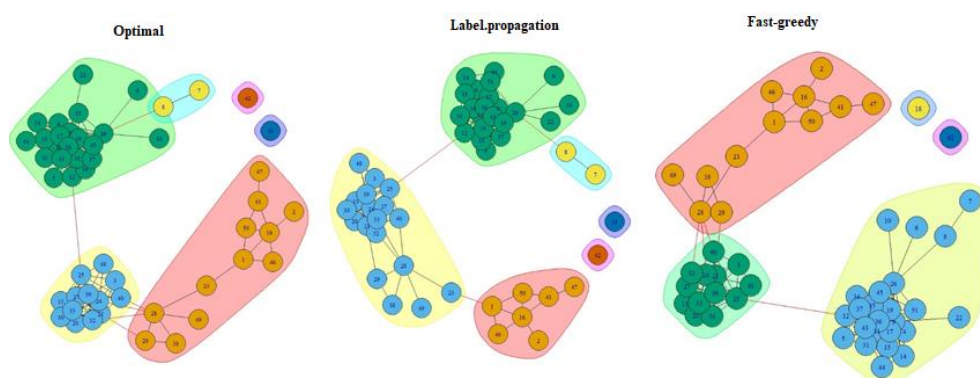
Таблица 1

Параметры кластеризации

	Optimal	Label.propagation	Fast-greedy
modularity	0.4995711	0.4960421	0.4954049
membership	6	6	5
system.time	1.77	0.01	0

Необходимо заметить, что данные в таблице, обозначающие время работы каждого из алгоритмов (system.time), приведены с округлением до сотых.

Разделение созданной сети на кластеры представлено на рисунке 1.

**Рис. 1 – Результаты разбиения графа на сообщества**

Заключение. Представление некоего набора данных в виде графа и последующий его анализ, включающий кластеризацию и отыскание зависимостей, является актуальной задачей в области интеллектуального анализа данных. Исследуемые в статье алгоритмы кластеризации позволили выявить сообщества, на основании связи людей друг с другом.

В ходе исследования реализованы алгоритмы кластеризации в программной среде RStudio и применены к реальным данным из социальной сети «Вконтакте». В результате анализа полученных результатов можно заметить, что алгоритм optimal произвел деление графа на сообщества лучше других, что можно понять по значению модульности. Но в тоже время, данный алгоритм выполнял разбиение дольше других, это затрудняет его применение к большим сетям. Важно заметить, что алгоритм Fast-greedy, хоть и выделил всего 5 сообществ, но справился с задачей быстрее двух других алгоритмов.



Список цитируемой литературы

1. Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, Dorothea Wagner: On Modularity Clustering, IEEE Transactions on Knowledge and Data Engineering 20(2):172-188, 2008
2. Wenjun Wang, W. Nick Street. Finding Hierarchical Communities in Complex Networks Using Influence-Guided Label Propagation Department of Management Sciences University of Iowa.
3. Thomas Aynaud, Jean-Loup Guillaume. Static community detection algorithms for evolving networks. WiOpt'10: Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, May 2010, Avignon, France. pp.508-514, 2010

© В.И.Ложечка, А.Н. Целых, 2019