Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2019, №2, Том 11 / 2019, No 2, Vol 11 https://esj.today/issue-2-2019.html

URL статьи: https://esj.today/PDF/90ITVN219.pdf

Ссылка для цитирования этой статьи:
Телевной А.Д., Иванов С.Е. Гибридный метод кластеризации данных при анализе социальных сетей // Вестник Евразийской науки, 2019 №2, https://esj.today/PDF/90ITVN219.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

For citation:
Televnoy A.D., Ivanov S.E. (2019). Hybrid method of data clustering for analyzing social networks. *The Eurasian Scientific Journal*, [online] 2(11). Available at: https://esj.today/PDF/90ITVN219.pdf (in Russian)

УДК 004.62

ГРНТИ 20.23.17

Телевной Андрей Дмитриевич
ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, Россия
Аспирант
E-mail: adtelev@mail.ru

Иванов Сергей Евгеньевич
ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, Россия
Доцент
Кандидат физико-математических наук
E-mail: serg_ivanov@corp.ifmo.ru

Гибридный метод кластеризации данных при анализе социальных сетей

Аннотация. Объектом исследования являются данные агентов социальных сетей (как пользователей, так и онлайн-сообществ). Предметом исследования являются методы кластеризации данных социальных сетей. Особое внимание уделяется анализу научных публикаций, посвященных уже существующим методам кластеризации данных. Сформулированы особенности проведения кластерного анализа данных в социальных сетях, которым не соответствует большинство используемых исследователями методов кластеризации. Результаты применения «гибридного» метода кластеризации визуализированы программными средствами языка программирования R с помощью графа Coreness, также был построен структурный профиль одного из онлайн-сообществ социальной сети «ВКонтакте».

Произведен анализ научных публикаций в области кластеризации данных социальных сетей. В результате анализа публикаций по теме исследования, авторами был сделан вывод о том, что большинство исследователей социальных сетей пользуется уже разработанными методами кластеризации, которые не могут считаться специализированными для анализа социальных сетей, так как не отражают характерных особенностей исследований данной области.

Предпочтительный выбор методов, основанных на мере центральности, объясняется тем, что мера центральности может рассматриваться не только как отражение потенциальной активности объекта социальной сети в графе, но и как возможность управления передачей информации (через частоту прохождения кратчайших путей между парами вершин в связанном графе).

Сформулирован алгоритм предложенного «гибридного» метода кластеризации данных, представлены характеристики используемых метрик центральности. Осуществлено моделирование структурных показателей выбранного для исследования онлайн-сообщества, а также представлены показатели количественного расчета метрик центральности, выбранных для исследования.

Авторами был предложен «гибридный» метод кластеризации данных, обладающий такими преимуществами, как возможность самостоятельного определения пользователем количества кластеров и оптимизация вычислительных ресурсов при анализе социальных сетей, что является важным достоинством с учетом активно возрастающих объемов данных.

Ключевые слова: анализ социальных сетей; онлайн-сообщество; метрика центральности; метод кластеризации; кластеризация данных; графовая модель; структурный профиль онлайн-сообщества; особенности кластерного анализа

Введение

В настоящее время активность в социальных сетях стала неотъемлемой частью жизни для огромного количества людей по всему миру. Для многих пользователей уже вошло в привычку неоднократное ежедневное проявление той или иной активности, будь то общение с друзьями, просматривание ленты новостей, публикация разнообразного контента и т. д. Логично, что при значительном увеличении количества данных в социальных сетях, увеличиваются и требования к используемым математическим и программным аппаратам, позволяющим анализировать подобные данные.

Одной из ключевых задач при анализе социальных сетей является задача кластеризации данных. Кластерный анализ данных может решать различные задачи: от выявления маркетологами референтных групп пользователей до упрощения доступа к информации путем ее тематической группировки. Кластеризация – это процесс разделения исследуемого множества объектов на группы «похожих» элементов, называемые кластерами.

На сегодняшний день существует множество алгоритмов кластеризации, однако нельзя уверенно сказать, что существует их единственная классификация. В своей работе мы предлагаем остановиться на классификации алгоритмов, включающей в себя три основные группы:

1. Алгоритмы квадратичной ошибки.
2. Алгоритмы иерархической кластеризации.
3. Алгоритмы теории графов.

Анализируя различные публикации по исследованию вопроса кластеризации данных в социальных сетях, был сделан вывод о том, что большинство исследователей берет в качестве основы уже созданный ранее, неспециализированный для социальных сетей метод, после чего поэтапно улучшает его, с учетом специфики решения своей конкретной задачи.

Среди публикаций отечественных авторов стоит отметить работу Чеснокова В.О. и Ключарёва П.Г., в которой рассматривается метод получения сообществ (кластеров) социального графа ближайшего окружения пользователя социальной сети, то есть графа, вершинами которого являются друзья пользователя, с диаметром, равным двум и с центром в вершине, соответствующей пользователю. В результате данной работы было разработано программное обеспечение, позволяющее собрать данные из социальной сети и провести анализ соответствующего социального графа. Также произведена оценка кластеризации с помощью

трёх внутренних метрик: ожидаемой плотности, индекса силуэтов и нормализованной гамма-статистики Хуберта [3].

Также можно выделить статью [2], посвященную анализу социальных сетей, представленных в виде графа. В работе приводятся подходы к моделированию распределений социальных сетей, а также алгоритмы, используемые для поиска сообществ. Использование описанных в статье методов и подходов позволяет производить классификацию сегментов социальной сети на кластеры, а также находить элементы, представляющие наибольший интерес (например, пользователей, чье существенное влияние распространяется сразу на несколько независимых сообществ). По мнению авторов, при нахождении степенного распределения вершин графа, описывающего социальную сеть, можно осуществлять моделирование социальных сетей с заданным распределением.

Стоит также упомянуть работы авторов [4; 7], посвященные изучению центральности узлов во взвешенных сетях. Результатом работы [7] является обобщение основных показателей центральности узла (степень, близость, промежуточность) с точки зрения не только веса, но и количества связей.

Выбор подходящего метода кластеризации для сетевых данных априори может быть неприятным и запутанным процессом. Для решения этой проблемы автор статьи [5] опирается на апостериорный подход, разработанный Гриммером и Кингом (2011), который одновременно сравнивает сотни возможных методов кластеризации через краткую и интуитивно понятную визуализацию. Общий метод был адаптирован к контексту социальных сетей, расширен дополнительными функциями визуализации, предназначенными для повышения интерпретируемости и описания ее принципиального использования с изложением этапов выбора класса методов для сравнения.

Статья Рыцарева И.А. посвящена изучению проблемы классификации аккаунтов социальных сетей на основе пользовательского медиаконтента. Автор справедливо формулирует одни из главных трудностей на пути решения проблемы: гетерогенность контента и колоссальные объемы данных. Результатом работы является предложенный подход на основе текстового аннотирования с использованием технологии BigData, который подкреплен наглядными результатами исследования [8].

Успешная попытка построения мультиагентной системы для кластеризации групп социальной сети по различным критериям предпринята авторами статьи [9]. Авторы справедливо отмечают, что недостатком кластерных методов является невозможность изменения количества кластеров на последующих этапах кластеризации, что может привести к некорректному результату в случае плохого начального разбиения данных. Одним из достоинств данной работы стоит отметить использование различных видов метрик для репрезентативности полученных в ходе исследования результатов.

В статье [6] авторы представляют новый метод кластеризации для обнаружения таких групп и предлагают использовать его для анализа некоторых классических социальных сетей. Данный метод имеет две отличительные особенности: недвоичное иерархическое дерево и особенность перекрывающейся кластеризации. Недвоичное иерархическое дерево намного меньше двоичных деревьев, построенных большинством традиционных методов, благодаря чему происходит четкое выделение значимых кластеров, что значительно сокращает дальнейшие ручные усилия по выбору кластера.

В работе авторов Силина М. и Панова М. [10] приведен обзор и экспериментальное сравнение базовых алгоритмов кластеризации графов как на реальных, так и на сгенерированных наборах данных. Уделяется особое внимание анализу методов оценки качества кластеризации.

Стоит отметить, что кластерный анализ данных в социальных сетях имеет свои специфические особенности, связанные, прежде всего, с формой организации социальных сетей. Как правило, социальные сети представляются исследователями в виде связанных графов, узлы которых являются социальными объектами (пользователи, атрибуты пользователей, сообщества и т. д.), а ребра – социальными связями между ними. Таким образом, для кластерного анализа социальных сетей можно использовать методы кластеризации, основанные на теории графов. Однако, мы можем расширить ассортимент возможных используемых методов путем проведения процедуры нормализации графа. Результатом нормализации графа станет матрица с наборами данных, в которой связь одной вершины с другой просто является ее дополнительной характеристикой.

Цель работы

Целью данной работы являлась разработка «гибридного метода кластеризации», позволяющего усреднять результаты используемых методов кластеризации и оптимизировать вычислительные затраты на создание модели.

Базовые положения

Для улучшения процесса кластеризации данных в социальных сетях авторами предлагается обратить особое внимание на такую метрику связанного графа, как центральность. В качестве базовых метрик для предлагаемого «гибридного метода кластеризации» были выбраны следующие метрики центральности: Betweenness Centrality, Status Centrality, Closeness Centrality, PageRank Centrality.

Предпочтительный выбор методов, основанных на мере центральности, объясняется тем, что мера центральности может рассматриваться не только как отражение потенциальной активности объекта социальной сети в графе, но и как возможность управления передачей информации (через частоту прохождения кратчайших путей между парами вершин в связанном графе).

С целью оптимизации модели также предполагается осуществлять кластеризацию на кластеры фиксированных размеров, определяемых исследователем вручную.

Необходимость усреднения результатов нескольких методов кластеризации объясняется желанием получения более репрезентативной картины, нежели при использовании единственного метода, который так или иначе будет усиливать или занижать важность определенных узлов графа согласно соответствующему ему алгоритму кластеризации.

Алгоритм применения «гибридного метода кластеризации»:

1. Построение списка весов вершин путем последовательного применения вышеуказанных метрик центральности связанного графа.
2. Нормализация максимального значения из построенного списка путем деления значения веса каждой вершины на максимальный вес.
3. Вычисление среднего квадратичного значения веса вершины графа из списка нормализованных весов, полученных в п. 2.
4. Преобразование списка вершин путем упорядочивания построенного перечня по возрастанию весов вершин.
5. Кластеризация итогового списка на кластеры фиксированного размера.

В рамках данного исследования было проанализировано онлайн-сообщество социальной сети «ВКонтакте» – «Университет ИТМО Аспирантура» (https://vk.com/club_aspirantov).

В качестве программной среды для вычисления метрик центральности были выбраны RStudio языка программирования R и пакет «igraph», сбор данных из социальной сети осуществлялся с помощью пакета «vkR», для вычислительных расчетов были задействованы пакеты «dplyr», «DescTools».

В таблице 1 представлены рассчитанные структурные показатели для онлайн-сообщества «Университет ИТМО Аспирантура» (по состоянию на апрель 2019 года). Предложенный перечень структурных показателей описан в диссертационной работе Ю.Г. Рыкова «Структура социальных связей в виртуальных сообществах: сравнительный анализ онлайн-групп социальной сети «ВКонтакте» [1].

Таблица 1

Значение структурных показателей для сообщества

№	Структурные показатели	Значение структурного показателя
1	Чистая полная численность	2930
2	Доля активных участников (коэффициент Жаккара)	0.402
3	Доля связанных пользователей	0.812
4	Доля изолятов (от чистой полной численности)	0.188
5	Доля вершин в максимальной связанной компоненте	0.989
6	Доля ребер в максимальной связанной компоненте	0.152
7	Плотность сети «дружбы»	0.0036
8	Плотность сети «дружбы» без изолятов	0.0055
9	Значение модулярности	0.415
10	Средняя геодезическая дистанция сети «дружбы»	3.555
11	Диаметр сети «дружбы»	9
12	Среднее значение степени центральностей сети «дружбы» для полной сети.	10.588
13	Средний коэффициент кластеризации сети «дружбы»	0.275
14	Индекс Джини для распределения степеней центральности в сети «дружбы» без изолятов	0.526
15	Индекс Джини для распределения центральности посредничества в сети «дружбы» без изолятов	0.763
16	Централизация графа по степени центральности без изолятов	0.069
17	Централизация посредничества без изолятов	0.056
18	Доля участников, создающих контент в группе в виде постов, комментариев или сообщений в обсуждениях.	0.221
19	Доля участников, не создающих контент в группе, но ставящих «лайки» к контенту группы	0.393
20	Доля пассивных участников, не «лайкающих» и не создающих контент	0.964

Составлено автором

В таблице 2 представлены использованные в рамках данной работы характеристики метрик центральности.

Таблица 2

Характеристика метрик центральности

Метрика центральности	Формула вычисления метрики
Closeness Centrality	Среднее расстояние от вершины v_i до соседних вершин $\sum_j d_{i,j}/k$, где d – матрица расстояний между вершинами графа, k – количество вершин, с которыми связана v_i.

Метрика центральности	Формула вычисления метрики
Betweenness Centrality	Центральность по посредничеству – $\sum_{N_s \neq V_i} \sum_{N_i \neq V_i} \frac{1}{2} \frac{\sigma_{v_s v_t}(v_i)}{\sigma_{v_s v_t}}$, где $\sigma_{v_s v_t}$ – число кратчайших путей между v_s и v_t, $\sigma_{v_s v_t}(v_i)$ – число кратчайших путей, проходящих через v_i. Стандартизуется делением на максимально возможную величину – (n-1)(n-2)/2.
PageRank Centrality	Важность веб-страницы (узла в графе) определяется как $x_i = \alpha \sum_j a_{ij} \frac{x_j}{k_j^{out}} + \beta$, где α и β – константы, k_j^{out} – количество ребер, исходящих из вершины j (ссылки со страницы j), которое равно 1 при отсутствии исходящих ссылок
Eigen Centrality	Относительная оценка центральности вершины может быть определена как: $X_v = \frac{1}{\lambda} \sum_{t \epsilon M(v)} x(t) = \frac{1}{\lambda} \sum_{t \epsilon G} a_{v,t} x_t$, где $M(v)$ – множество соседей v, а λ – константа.

Составлено автором

В таблице 3 представлены количественные результаты расчета метрик центральности.

Таблица 3

Расчет метрик центральности

Метрика центральности	Соответствующая функция в R	Значение метрики
Closeness Centrality	Centr_clo()	0.004
Betweenness Centrality	Centr_betw()	0.056
PageRank Centrality	Page_rank()	1
Eigen Centrality	Centr_eigen()	0.957

Составлено автором

На рисунке представлен граф Coreness онлайн-сообщества после применения «гибридного метода кластеризации».

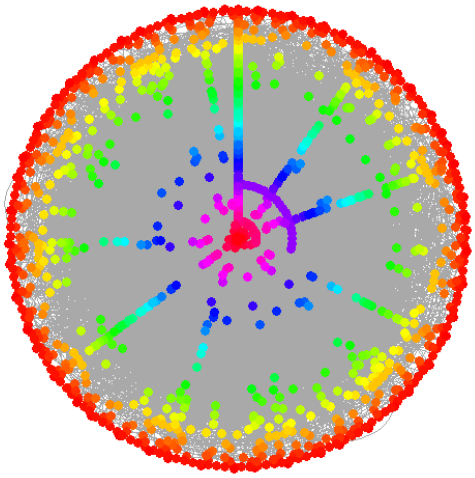

Рисунок. *Граф Coreness онлайн-сообщества (составлено автором)*

Выводы

По результатам данной работы был предложен «гибридный» метод кластеризации данных социальных сетей, а также построен структурный профиль одного из онлайн-сообществ социальной сети «ВКонтакте».

Возможность самостоятельного определения количества кластеров исследователем является объективным преимуществом предложенного «гибридного» метода кластеризации.

Предложенный метод кластеризации позволяет оптимизировать ресурсы при анализе социальных сетей, что является важным достоинством с учетом активно возрастающих объемов данных.

ЛИТЕРАТУРА

1. Рыков Ю.Г. Структура социальных связей в виртуальных сообществах: сравнительный анализ онлайн-групп социальной сети «ВКонтакте»: дис. ... кандидата социологических наук: 22.00.04 / Рыков Юрий Георгиевич; [Место защиты: Нац. исслед. ун-т "Высш. шк. экономики"]. – Москва, 2016. – 196 с.
2. Хотилин М.И., Благов А.В. Визуальное представление и кластерный анализ социальных сетей // Самара: Материалы Международной конференции и молодежной школы «Информационные технологии и нанотехнологии». 2016 г. С. 1067–1072.
3. Чесноков В.О., Ключарёв В.Г. Выделение сообществ в социальных графах по множеству признаков с частичной информацией // Москва: Журнал «Наука и Образование» МГТУ им. Н.Э. Баумана. 2015 г. Выпуск 9. С. 188–199.
4. Melville P., Mooney R., Nagarajan R. Content-Boosted Collaborative Filtering for Improved Recommendations // University of Texas, USA: Материалы конф. / AAAI-02, Austin, TX, USA, 2002. – 2002. – P. 187–192.
5. Choosing a Clustering: An A Posteriori Method for Social Networks. Samuel D. Pimentel. Journal of Social Structure. Vol. 15. 2014.
6. A new clustering method and its application in social networks. Peixin Zhao, Cun-Quan Zhang. Journal Pattern Recognition Letters. Vol. 32. 2011. Pages 2109–2118.
7. T. Opsahl, F. Agneessens, J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. – Social Networks, 2010.
8. Рыцарев, И.А. Кластеризация медиаконтента из социальных сетей с использованием технологии BigData / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921–927.
9. Охапкина Е.П., Охапкин В.П. Подходы к кластеризации групп социальной сети // Компьютерные исследования и моделирование. – 2015. – Т. 7, № 5. – С. 1127–1139.
10. Силин И., Панов М. Обзор и экспериментальное сравнение алгоритмов кластеризации графов // Сочи: Сборник трудов 39-й междисциплинарной школы-конференции ИППИ РАН «Информационные технологии и системы 2015». 2015 г. С. 1042–1059.

Televnoy Andrey Dmitrievich
Saint-Petersburg national research university of information technologies, mechanics and optics, Saint-Petersburg, Russia
E-mail: adtelev@mail.ru

Ivanov Sergey Evgenevich
Saint-Petersburg national research university of information technologies, mechanics and optics, Saint-Petersburg, Russia
E-mail: serg_ivanov@corp.ifmo.ru

Hybrid method of data clustering for analyzing social networks

Abstract. The subject of research is the data of social network agents (both users and online communities). The subject of research is the methods of clustering data of the social networks. The special attention is paid to the analysis of scientific publications devoted to already existing methods of data clustering. The features of cluster analysis activity of data in the social networks are formulated, to which the most clustering methods used by researchers do not correspond. The results of applying the “hybrid” clustering method are visualized by the software tools of the R programming language using the Coreness graph. A structural profile of one of the online communities of the In-Contact social network was also built.

The analysis of scientific publications in the field of clustering data in the social networks was carried out. As a result of the analysis of publications on the research topic, the authors concluded that the majority of social network researchers use already developed clustering methods that cannot be considered specialized for analyzing social networks, since they do not reflect the characteristic features of research in this field.

The preferred choice of methods based on a measure of centrality is explained by the fact that the measure of centrality can be viewed not only as a reflection of the potential activity of a social network object in the graph, but also as a controllability of information transfer (through the frequency factor of the shortest paths between pairs of vertices in a connected graph).

The algorithm of the proposed “hybrid” data clustering method is formulated and the characteristics of the useful centrality metrics are presented. The simulation of structure coefficients selected for the study of the online community has been carried out, and indicators of the quantitative calculation of the centrality metrics selected for the study are presented.

The authors proposed a “hybrid” data clustering method with such advantages as the ability for the user to independently determine the number of clusters and the optimization of computational resources when analyzing social networks, which is an important advantage in view of actively increasing data volumes.

Keywords: social network analysis; online community; centrality metrics; clustering method; data clustering; graph model; structural profile of online community; features of cluster analysis